Online Dual-Network-Based Adaptive Dynamic Programming for Solving Partially Unknown Multi-Player Non-Zero-Sum Games With Control Constraints

PENGDA LIU\(^1\), HUAGUANG ZHANG\(^1,2\), (Fellow, IEEE), CHONG LIU\(^1\), AND HANGUANG SU\(^1\)

\(^1\)College of Information Science and Engineering, Northeastern University, Shenyang 110004, China
\(^2\)State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110004, China

Corresponding author: Pengda Liu (duozckchlz@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61627809, Grant 61433004, and Grant 61621004; and in part by the Liaoning Revitalization Talents Program under Grant XLYC1801085.

ABSTRACT In this article, a novel online method for multi-player non-zero-sum (NZS) differential games of nonlinear partially unknown continuous time (CT) systems with control constraints is developed based on neural networks (NN). The issue of multi-player NZS games with saturated actuator is elaborately analyzed and the unknown dynamics model is learned by applying identifier NN. Different from using the standard identifier-actor-critic framework of adaptive dynamic programming (ADP), the proposed method uses only identifier networks and critic networks for all the players to solve the coupled Hamilton-Jacobi (HJ) equations for multi-player NZS games, which could effectively simplify the algorithm and save computing resources. Moreover, a tuning law which utilizes the gradient descent method is designed for each critic network. Meanwhile, to remove the requirement for the initial stabilizing control, a novel stability term is designed to ensure the system stability during the training phase of the critic NN. By the means of Lyapunov approach, it is proven that the system states, the critic network weight estimation errors and the obtained control are all uniformly ultimately bounded (UUB). Finally, two numerical examples are simulated to illustrate the validity of the developed method for multi-player NZS games with control constraints.

INDEX TERMS Adaptive critic designs, adaptive dynamic programming, control constraints, multi-player, non-zero-sum games.

I. INTRODUCTION
The theories with respect to differential games have received more and more attentions since it was firstly studied in [1]. With the efforts of worldwide scholars, differential games theories, which are closely linked to our daily life nowadays, have been widely used in economics, sociology and many other domains [2]–[7]. Three indispensable elements, i.e., players, control policies and performance functions, jointly build the footstone of differential games theory. In multi-player NZS games, the key is to obtain a cluster of optimal control policies called Nash equilibrium for each player to pursue the minimization of their own performance function. In the meantime, the system stability should be ensured [8]. For the linear NZS games, solving the issue is equivalent to figuring out coupled algebraic Riccati equations. While for the nonlinear NZS games, it’s hard to obtain the analytic solutions of the coupled HJ equations. Based on this situation, numerous intelligent methods have been designed by scholars to approximately attain the solutions.

During recent years, ADP methods have been widely utilized to solve nonlinear optimal control issues. The related algorithms could expediently get the approximate optimal control schemes with the help of NN and get rid of the “curse of dimensionality” which is induced by conventional dynamic programming approaches [9]–[12]. Due to the merit of ADP, many algorithms have been proposed to address the optimal control issues. To figure out the optimal tracking issue, a novel data—driven ADP method was presented [13]. Owing to the use of recurrent NN, the method only required...
The getatable input—output data. In [14]–[17], on the basis of an actor–critic framework, different ADP algorithms were proposed to seek the optimal control schemes. Instead of the actor–critic architecture employed by the above algorithms, a few algorithms utilized only critic NN to cope with the optimal control issues [18]–[20]. Generally, this sort of framework could simplify the algorithm structure and effectively save computing resources.

On account of the trouble of coping with the coupled HJ equations, the issues of NZS games have become a challenging research area for scholars. Moreover, in many cases, it is often tough to acquire the actual model of the system dynamics [21]. Therefore, many scholars have employed various methods to learn the unknown system dynamics [22]. For instance, T−S fuzzy models can be employed to learn the dynamics for that it is powerful in approximating nonlinear systems [23]. In [24], fuzzy logic models, which are similar to T−S fuzzy models, were combined with ADP to solve the NZS games issues. For the unknown multi—input system, a three—layer NN identifier, reinforcement learning scheme and NZS game theory were utilized together to solve the optimal tracking control issue [25]. In [26], a model network was designed to identify the system whose dynamic is not known. Critic networks and actor networks were then employed to approximately learn the value functions and control policies for every players. The algorithm presented in [27] used identifier-critic framework to address the issue of NZS games when the dynamics were unknowable. A novel tuning law was adopted to increase the convergence speed of the controlled system. The integral reinforcement learning (IRL) technology, which is a crucial method of implementing ADP, was used to deal with the unknown dynamics for the NZS games [28], [29]. Recently, event-triggered mechanism has been widely employed to save transmission bandwidths and computing resources [30], [31]. This technology was also combined with ADP to obtain control schemes for every players in NZS games [32].

In actual physical systems, it’s often inevitable that the saturation nonlinearity phenomena of actuators severely hurt system performance. Hence the control issues with constraints have attracted intensive attentions. The presented method in [33] used identifier-critic framework to figure out the Hamilton-Jacobi-Isaacs (HJI) equations for unknown systems with constrained-input. A term was additionally designed to remove the need for the initial stabilizing control. In [34], by the means of IRL, a data-based method was proposed to seek the solutions when the dynamics of the system with constrained-input were completely unknown. An algorithm based on event-triggered mechanism was proposed to solve CT systems with constrained-input [35]. The method in [36] also utilized event-triggered mechanism to solve nonlinear $H_{\infty}$ control issues with constraints. In [37], an IRL method was used to figure out the optimal control policies for players in NZS games with saturated actuator. Nevertheless, this method employed both actor NN and critic NN, which perplexed the algorithm and aggravated the computing burdens. Besides, initial stabilizing control policies were required.

In general, for the existence of the coupled terms, it’s hard to attain the control schemes for every players. To the best of our knowledge, there exist few algorithms dealing with multi-player NZS games with control constraints. Inspired by the existing methods, in this article, an online dual-network-based ADP method is developed to solve the multi-player NZS games with control constraints. The contributions of this work can be summarized as follows. In the proposed online method which could online control the system in real time without using the historical data, identifier-critic framework is designed. The identifier NN are used to estimate the system dynamics and the critic NN are utilized to approximate the solutions of the coupled HJ equations. With this architecture the algorithm can be simplified and the computing burdens are lightened. By introducing an additional term to the update laws of the critic NN, the requirement of initial stabilizing control policies is relaxed. Furthermore, the stability of the multi-player NZS system is demonstrated by Lyapunov theory. Finally, simulation results verify this method.

The remainder of this article is structured as follows. Section II formulates the issue of multi-player NZS games with control constraints. Identifier NN and critic NN are constructed to learn the unknown internal dynamics and solve the coupled HJ equations in Section III. Besides, for critic NN, a novel stabilizing term is added to the update law to get rid of the need of an initial stabilizing control. Section IV demonstrates the UUB of the states, the critic weight estimation errors and the obtained controllers by Lyapunov approach. Two numerical examples are shown to verify the effectiveness of the presented method in Section V. Finally, Section VI provides the conclusion.

**Notations:** $\mathbb{R}$ represents the set which includes all real numbers, $\mathbb{R}^n$ is the $n$-dimensional Euclidean space and $\mathbb{R}^{n\times m}$ the space of all real matrices. $\mathcal{E}$ denotes a compact set which contains the origin and $C^q(\mathcal{E})$ denotes the class of functions that have continuous $q$-th derivative. $\mathbb{N}$ denotes the set $\mathbb{N} = \{1, \ldots, N\}$, where $N \geq 2$ is a positive integer. $I_{n\times n}$ denotes the unit matrix whose dimensionality is $n$. And the $\| \cdot \|$ represents the Euclidean norm of a vector or a matrix. $\nabla(\cdot) \triangleq \partial(\cdot)/\partial x$ is taken for denoting the gradient operator.

**II. PROBLEM FORMULATION**

Consider the system of the nonlinear system in an affine form as follows:

$$
\dot{x} = f(x) + \sum_{j=1}^{N} g_j(x)u_j(x),
$$

where $x \in \mathcal{E} \subset \mathbb{R}^n$ is the state, $u_j \in \Delta \subset \mathbb{R}^m$ is the control input of the $j$–th player, $\Delta = \{u_j \in \mathbb{R}^m : |u_{jp}| \leq \alpha_j, p = 1, \ldots, m\}$, and $\alpha_j > 0$ denotes the constraint bound. $f(x) \in \mathbb{R}^n$, $g_j(x) \in \mathbb{R}^{n \times m}$ and $f(0) = 0$.
Assumption 1: ∀j ∈ ℕ, the system function f(x) and g_j(x) are all locally Lipschitz on Ξ, and system (1) is controllable and observable.

Assumption 2: ∀j ∈ ℕ, g_j(x) is bounded and there exists a positive constant g_{jM} such that ∥g_j∥ ≤ g_{jM}.

Define the performance index for player i as

\[ J_i(x_0, u_1, \ldots, u_N) = \int_0^\infty \sigma_i(x, u_1, \ldots, u_N) d\tau, \]

where the utility function \( \sigma_i(x, u_1, \ldots, u_N) = x^\top Q_i x + \sum_{j=1}^N C_j(u_j) \).

The key point to solve NZS games is to seek the Nash equilibrium point solution if for any stabilizing policy pair \((u_i^*, \ldots, u_N^*)\) we have

\[ J_i^*(J_i(u_1^*, \ldots, u_N^*)) = J_i(u_1^*, \ldots, u_N^*, u_i^*) \leq J_i(u_1^*, \ldots, u_i^*, \ldots, u_N^*). \]

Definition 1 indicates that without the reasonable adjustments of the policies of all the other players, the player i can hardly obtain its optimal performance.

For any stabilizing control policy pair \(u = \{u_1, \ldots, u_N\}\), the value function can be defined as

\[ V_i(x) = \int_0^\infty \left( 2 \sum_{j=1}^N \alpha_j \int_0^{u_j} \tanh^{-1}(\mu_j/\alpha_j) R_{ij} d\mu_j + x^\top Q_i x \right) d\tau. \]

Assume that the function (5) is continuous and differentiable, then the differential equivalent for the i-th player can be described as

\[ 0 = \sigma_i(x, u_1, \ldots, u_N) + (\nabla V_i)^\top f(x) + \sum_{j=1}^N g_j(u_j), \]

\[ V_i(0) = 0. \]

Then we have the Hamiltonian function

\[ H_i(x, u_1, \ldots, u_N) = (\nabla V_i)^\top f(x) + \sum_{j=1}^N g_j(u_j) + \sigma_i(x, u_1, \ldots, u_N). \]

Moreover, the optimal value function and optimal control for the i-th player are given as

\[ V_i^*(x, u_1, \ldots, u_N) = \min_{u_i} \int_0^\infty \left( x^\top Q_i x + 2 \sum_{j=1}^N \alpha_j \int_0^{u_j} \tanh^{-1}(\mu_j/\alpha_j) R_{ij} d\mu_j \right) d\tau, \]

\[ u_i^* = -\alpha_i \tanh\left( \frac{1}{2\alpha_i} R_{ii}^{-1} g_i \nabla V_i^* \right). \]

Then we obtain the coupled HJ equations of system (1) as

\[ H_i(x, \nabla V_i^*, u_1^*, \ldots, u_N^*) = (\nabla V_i^*)^\top f(x) + \sum_{j=1}^N \alpha_j g_j \tanh(M_j^*(x)) \]

\[ = (\nabla V_i^*)^\top f(x) - (\nabla V_i^*)^\top \sum_{j=1}^N \alpha_j g_j \tanh(M_j^*(x)) \]

\[ + 2 \sum_{j=1}^N \alpha_j \int_0^{u_j} \tanh^{-1}(\mu_j/\alpha_j) R_{ij} d\mu_j \]

\[ + x^\top Q_i x, \quad V_i^*(0) = 0. \]

Denote \( M_j^*(x) = [M_1^*(x), \ldots, M_p^*(x), \ldots, M_{jm}^*(x)] \) with \( M_j^*(x) \in R^m \) and \( R_{ij} = \text{diag}(r_{ij}, \ldots, r_{ij}, \ldots, r_{ij}) \), \( 1 \leq p \leq m \). We have

\[ 2 \sum_{j=1}^N \alpha_j \int_0^{u_j} \tanh^{-1}(\mu_j/\alpha_j) R_{ij} d\mu_j \]

\[ = \sum_{j=1}^N \alpha_j \sum_{p=1}^m r_{ij} \ln[1 - \tanh^2(M_{jp}^*(x))] \]

\[ + 2 \sum_{j=1}^N \alpha_j^2 \sum_{p=1}^m r_{ij} M_{jp}^*(x) \tanh(M_{jp}^*(x)). \]

Then, the coupled HJ equation (10) is denoted as

\[ H_i(x, \nabla V_i^*, u_1^*, \ldots, u_N^*) = \sum_{j=1}^N \alpha_j \sum_{p=1}^m r_{ij} \ln[1 - \tanh^2(M_{jp}^*(x))] \]

\[ + 2 \sum_{j=1}^N \alpha_j^2 \sum_{p=1}^m r_{ij} M_{jp}^*(x) \tanh(M_{jp}^*(x)) \]

\[ + (\nabla V_i^*)^\top f(x) - (\nabla V_i^*)^\top \sum_{j=1}^N \alpha_j g_j \tanh(M_j^*(x)) \]

\[ + x^\top Q_i x = 0, \quad V_i^*(0) = 0. \]
Note that it’s intractable to tackle the equation (12) for the existence of the coupled terms and partial derivatives. Therefore, in what follows, identifier-critic NN framework is developed to approximately solve the coupled HJ equations.

III. MULTI-PLAYER LEARNING FOR NZS GAMES WITH CONTROL CONSTRAINTS USING IDENTIFIER NN AND CRITIC NN

In this section, for the purpose of solving the multi-player NZS games with control constraints, an online adaptive method based on NN is proposed. Identifier-critic framework is employed to approximately solve the HJ equations. Besides, for critic NN, by designing an additional stabilizing work is employed to approximately solve the HJ equations.

A. IDENTIFIER DESIGN

By constructing NN-based ideal identifier, (1) could be rewritten as

\[ \dot{x} = Sx + W_f^T \gamma_f(x) + \epsilon_f + \sum_{j=1}^{N} g_j u_j, \]

where \( W_f \in \mathbb{R}^{N_f} \) is the ideal weight vector, \( \gamma_f(x) \in \mathbb{R}^{N_f} \) the activation function set, \( S \in \mathbb{R}^{n \times n} \) the Hurwitz matrix and \( \epsilon_f \in \mathbb{R} \) denotes reconstruction error. \( N_f \) represents the number of neurons for identifier NN.

In practice, when employing identifier NN to learn the unknown internal dynamics, we have

\[ \hat{x} = S\hat{x} + \hat{W}_f^T \gamma_f(\hat{x}) + \sum_{j=1}^{N} \hat{g}_j \hat{u}_j, \]

where \( \hat{x} \) is the acquired state vector and \( \hat{W}_f \) the acquired weight.

Let the state estimate error \( \tilde{x} = x - \hat{x} \) and the weight estimate error \( \tilde{W}_f = W_f - \hat{W}_f \), then we can derive that

\[ \dot{\tilde{x}} = S\tilde{x} + \tilde{W}_f^T \gamma_f(\tilde{x}) + W_f^T (\gamma_f(x) - \gamma_f(\tilde{x})) + \epsilon_f. \]  

Before proceeding, two mild assumptions are necessary [39].

Assumption 3: The ideal identifier weight \( W_f \) and reconstruction error \( \epsilon_f \) in the reconstructed dynamic (13) are both bounded, and it holds that \( \|W_f\| \leq W_M, \epsilon_f \leq \epsilon_M \).

Assumption 4: The activation function \( \gamma_f(\cdot) \) is Lipschitz continuous with

\[ \|\gamma_f(x) - \gamma_f(\tilde{x})\| \leq L_{\gamma_f} \|\tilde{x}\|. \]

where \( L_{\gamma_f} \) is a constant.

Theorem 1: Consider the system (1) and employ the identifier (14). When the identifier weight is tuned with

\[ \dot{\hat{W}}_f = H_{\gamma_f} (\hat{x}) \hat{x}^\top, \]

where \( H \in \mathbb{R}^{N_f \times N_a} \) is a positive definite symmetric matrix, the state estimation error \( \tilde{x} \) is UUB.

Proof: The Lyapunov function candidate is constructed as

\[ L_d = L_{d1} + L_{d2} = \frac{1}{2} \tilde{x}^\top \tilde{x} + \frac{1}{2} tr(\dot{\hat{W}}_f^T H^{-1} \dot{\hat{W}}_f). \]  

Due to (15) and the above two assumptions, we have

\[ \dot{\tilde{L}}_{d1} \leq \tilde{x}^\top (S\tilde{x} + \tilde{W}_f^T \gamma_f(\tilde{x}) + \epsilon_M) \]

\[ = -\tilde{x}^\top M_s \tilde{x} + \tilde{x}^\top \tilde{W}_f^T \gamma_f(\tilde{x}) + \tilde{x}^\top \epsilon_M, \]

(19)

where \( M_s = -S - \hat{W}_f^T \gamma_f(\tilde{x}) \). The positive definite matrix \( M_s \) can be obtained by appropriately choosing the \( S \). Due to that \( \dot{\hat{W}}_f = -\hat{W}_f \) and the tuning law (17), we have

\[ \dot{L}_{d2} = -\tilde{x}^\top \dot{\hat{W}}_f^T \gamma_f(\tilde{x}). \]  

(20)

From (19) and (20), it yields that

\[ \dot{\tilde{L}}_d \leq -\tilde{x}^\top M_s \tilde{x} + \tilde{x}^\top \epsilon_M \leq -P_{Ms} \|\tilde{x}\|^2 + \|\tilde{x}\| \epsilon_M. \]

(21)

where \( P_{Ms} = \lambda_{\min}(M_s) \). When \( \|\tilde{x}\| > \frac{\epsilon_M}{P_{Ms}} \), it holds that \( \dot{L}_d < 0 \). Based on the analysis above, \( \tilde{x} \) is UUB when the tuning law (17) is utilized.

Here, the identifier dynamics \( S\hat{x} + \hat{W}_f^T \gamma_f(\hat{x}) \) is obtained and will be utilized below.

B. ADAPTIVE CONTROLLER DESIGN

Owing to the universal approximation nature of NN, the value function for player \( i \) can be represented as

\[ V_i^*(x) = W_i^T \gamma_i(x) + \epsilon_i, \]

(22)

where \( W_i \in \mathbb{R}^{N_i} \) and \( \gamma_i \in \mathbb{R}^{N_i} \) are the ideal weight vector and the activation function set, respectively. \( N_i \) denotes the number of neurons for player \( i \) and \( \epsilon_i \in \mathbb{R} \) is reconstruction error. According to (22), the derivative of \( V_i^*(x) \) with respect to \( x \) is

\[ \nabla V_i^* = (\nabla \gamma_i)^T W_i + \nabla \epsilon_i, \]

(23)

where \( \nabla \gamma_i = \partial \gamma_i(x)/\partial x \in \mathbb{R}^{N_i \times n} \) and \( \nabla \epsilon_i \) denotes the partial derivative of \( \epsilon_i \).

For that the ideal weights are difficult to obtain, we utilize estimated critic networks to approximate \( V_i^* \)

\[ \hat{V}_i(x) = \hat{W}_i^T \gamma_i(x). \]

(24)

Accordingly, the partial derivative of \( \hat{V}_i \) is

\[ \nabla \hat{V}_i = (\nabla \gamma_i)^T \hat{W}_i. \]

(25)

From (9) and (23), \( u_i^* \) is derived as

\[ u_i^*(x) = -\alpha_i \tanh(M_i(x)) + \epsilon_{u_i^*}, \]

(26)

where \( \epsilon_{u_i^*} = -\frac{1}{2} I - \tanh^2(k_i) R_i^{-1} g_i^T \nabla \epsilon_i \) with \( I = [1, \ldots, 1]^T \in \mathbb{R}^m \) and \( k_i \in \mathbb{R}^m \) which is selected between \( M_i^*(x) \) and \( M_i(x) = \frac{1}{2m} R_i^{-1} g_i^T (\nabla \gamma_i)^T W_i \).

Similarly, the approximate control for the \( i \)-th player is presented as

\[ \hat{u}_i(x) = -\alpha_i \tanh(\hat{M}_i(x)), \]

(27)
where \( \hat{M}_i(x) = \frac{1}{2\alpha_r} R_{ii}^{-1} g_i^T (\nabla \gamma_i) \nabla_i \hat{W}_i = [\hat{M}_i(x), \ldots, \hat{M}_m(x)]^T \in \mathbb{R}^m \).

Combining (10), (12), (24) and (27), we derive the approximate HJ error as

\[
e_i \triangleq H_i(x, \hat{W}_i, \hat{u}_1, \ldots, \hat{u}_N)
= \hat{W}_i^T \nabla \gamma_i \left( f(x) + \sum_{j=1}^N g_j \hat{u}_j \right) + x^T Q_i x
+ 2 \sum_{j=1}^N \alpha_j \int_0^{\hat{\mu}_j} \tanh^{-1}(\mu_j / \alpha_j) R_{ij} d \mu_j
= \sum_{j=1}^N \alpha_j^2 \sum_{p=1}^m r_{ijp} \ln[1 - \tanh^2(\hat{M}_{jp}(x))]
+ 2 \sum_{j=1}^N \alpha_j^2 m \sum_{p=1}^m r_{ijp} \hat{M}_{jp}(x) \tanh(\hat{M}_{jp}(x))
+ \hat{W}_i^T \nabla \gamma_i \left( f(x) - \sum_{j=1}^N \alpha_j g_j \tanh(\hat{M}_j(x)) \right) + x^T Q_i x.
\] (28)

Let \( e'_i \triangleq H_i(x, \hat{W}_i, \hat{u}_1, \ldots, \hat{u}_N) = x^T Q_i x + 2 \sum_{j=1}^N \alpha_j \int_0^{\hat{\mu}_j} \tanh^{-1}(\mu_j / \alpha_j) R_{ij} d \mu_j + \hat{W}_i^T \nabla \gamma_i \left( f(x) + \sum_{j=1}^N g_j \hat{u}_j \right) \).

For the purpose of minimizing \( e_i \), it’s requested to find \( \hat{W}_i \) to minimize the function \( E = \sum_{i=1}^N E_i = \sum_{i=1}^N \frac{1}{2 \gamma_i^2} \).

By employing the gradient descent method [14], [40], the updating laws of the critic NN can be described as

\[
\dot{\hat{W}}_i = -\beta \left[ \frac{1}{(\kappa_i + 1)^2} \cdot \frac{\partial E_i}{\partial \hat{W}_i} - \beta \hat{\kappa}_i \left( \hat{W}_i^T \hat{\kappa}_i - \frac{e'_i}{\kappa_i + 1} \right) \right],
\] (29)

where \( \kappa_i = \nabla \gamma_i \left( f(x) + \sum_{j=1}^N g_j \hat{u}_j \right) \in \mathbb{R}^{N_i}, \hat{\kappa}_i = \frac{\kappa_i}{\kappa_i + 1} \in \mathbb{R}^{N_i}, \hat{W}_i = W_i - \hat{W}_i \) and \( \beta \) represents the learning rate.

Note that by utilizing the tuning law (29) for player \( i \), an initial stabilizing control is generally needed. If the initial control is not stable, the stability of the system might not be ensured with the tuning law (29). In addition, a noise signal is usually requested to be added to the system to satisfy the persistence of excitation (PE) condition during the training process for the critic NN, which could influence the stability of the close-looped system. Therefore, based on (29), a novel tuning law is developed by Lyapunov theory to remove the requirement for the initial stabilizing control. Before proceeding, a reasonable assumption is given.

**Assumption 5:** \( \forall i \in \mathbb{N}, \) there exists a Lyapunov candidate \( L_i(x) \) which is continuously differentiable and radially unbounded. When the control policy is optimal, then \( \dot{L}_i = (\nabla L_i)^T \left( f(x) + \sum_{j=1}^N g_j \hat{u}_j \right) < 0. \) In addition, it holds that

\[
(\nabla L_i)^T \left( f(x) + \sum_{j=1}^N g_j \hat{u}_j \right) = -(\nabla L_i)^T \gamma_j(x) L_i,
\] (30)

where the matrix \( \gamma_j(x) \in \mathbb{R}^{n \times n} \) is symmetric and positive definite.

Consider a Lyapunov candidate \( L_i(x) \) for player \( i \), then we have its derivative as

\[
\dot{L}_i = (\nabla L_i)^T \left( f(x) + \sum_{j=1}^N g_j \hat{u}_j \right)
= (\nabla L_i)^T \left( f(x) - \sum_{j=1}^N \alpha_j g_j \tanh(\hat{M}_j(x)) \right).
\] (31)

Aiming at ensuring the system stability during the NN learning phase, by using the gradient descent method a novel term is developed as

\[
T_s = \frac{\beta}{2} \nabla \gamma_i R_{ii}^{-1} [ I_m - \Psi(\hat{M}_i(x))] g_i^T \nabla f(x)
- \hat{W}_i^T \nabla \gamma_i \sum_{j=1}^N \alpha_j g_j \tanh(\hat{M}_j(x))
+ \sum_{j=1}^N \alpha_j^2 m \sum_{p=1}^m r_{ijp} \hat{M}_{jp}(x) \tanh(\hat{M}_{jp}(x))
+ \frac{\beta}{2} \Omega(x, \hat{u}_1, \ldots, \hat{u}_N) \nabla \gamma_i g_i R_{ii}^{-1} [ I_m]
- \Psi(\hat{M}_i(x)) g_i^T \sum_{j=1}^N \nabla L_j,
\] (33)

where the indicator function \( \Omega(x, \hat{u}_1, \ldots, \hat{u}_N) \) is defined as

\[
\Omega(x, \hat{u}_1, \ldots, \hat{u}_N) = \begin{cases} 0, & \text{if } \nabla L_i^T \left( f(x) + \sum_{j=1}^N g_j \hat{u}_j \right) < 0, \, \forall i \in \mathbb{N}; \\ 1, & \text{otherwise}. \end{cases}
\] (34)

**Remark 1:** Consider the tuning law (33) and the indicator function (34). Both the first and the second term in (33) are developed on the basis of the gradient descent method. The former term is employed to minimize the error function \( E_i \), and the latter one is to stabilize the closed-loop system during the training phase of the critic NN. It can be seen that when
the system is unstable, then $\Omega(x, \hat{u}_1, \ldots, \hat{u}_N) = 1$. The latter term in (33) works to ensure the system stability. By this means, the requirement for an initial stabilizing control is removed.

Remark 2: Note that the purpose of the NZS games is to obtain control policies for all the players to stabilize the system and minimize their own performance index functions. Each player has their individual goal to follow. Hence $\Omega(x, \hat{u}_1, \ldots, \hat{u}_N) = 0$ holds only when the condition $\nabla L_i^T \left( f(x) + \sum_{j=1}^{N} g_i \hat{u}_j \right) < 0$ is satisfied for all the players.

Note that $\dot{\tilde{W}}_i = -\dot{\hat{W}}_i$. For the convenience of stability analysis later, from (29) and (33), we derive

$$
\dot{\tilde{W}}_i = -\beta_k \left( \tilde{W}_i^T \dot{k}_i - \frac{\epsilon_i^t}{k_i^T k_i + 1} \right)
- \frac{\beta}{\Omega(x, \hat{u}_1, \ldots, \hat{u}_N)} \nabla y_i g_i R_{ii}^{-1} M
- \Psi(M_i(x)) |g_i|^T \sum_{j=1}^{N} \nabla L_j.
$$

(35)

IV. STABILITY ANALYSIS

In this section, the stability demonstration of the system utilizing the proposed method is presented. Before proceeding, according to [14], [20], [41]–[43], some assumptions are needed.

Assumption 6: $\forall i \in \mathbb{N}$, bring the signal $\tilde{k}_i$ which is persistently excited on the time interval $[t, t + T]$, i.e., there exist positive constants $\omega_i > 0$ and $T > 0$ such that

$$
\omega_i I_{N_i \times N_i} \leq \int_{t}^{t+T} \tilde{k}_i \tilde{k}_i^T d\tau.
$$

(36)

Assumption 7: $\forall i \in \mathbb{N}$, the ideal critic NZS NN weight $W_i$ is bounded such that $\|\tilde{W}_i\| \leq W_{iM}$, where $W_{iM}$ is a positive constant. Besides, there exist positive constants $\epsilon_{iM}, \epsilon_{iM}^t, \epsilon_{iM}^n$ and $\gamma_M$ such that $\|\nabla E_i\| \leq \epsilon_{iM}, \|E_i\| \leq \epsilon_{iM}^t, \|E_{i,\|n\|}\| \leq \epsilon_{iM}^n$ and $\|\nabla y_i\| \leq \gamma_M$.

To expediently present the main theorem, according to [9], [24], a requisite lemma is stated in advance.

Lemma 1: Consider the NZS system (1). The critic tuning law and control policy for each player are given by (29) and (27), respectively. Let Assumptions 6-7 hold, then $\dot{W}_i$ is UUB for all $i \in \mathbb{N}$.

Proof: Construct the Lyapunov function candidate $L_0 = \frac{1}{2} \sum_{i=1}^{N} \tilde{W}_i^T \beta^{-1} \tilde{W}_i$. The derivative of $L_0$ along the trajectories obtained by the control $\hat{u}_i$ is

$$
\dot{L}_0 = -\sum_{i=1}^{N} \tilde{W}_i^T \tilde{k}_i \tilde{k}_i^T \tilde{W}_i + \sum_{i=1}^{N} \tilde{W}_i^T \tilde{k}_i \epsilon_i^t.
$$

(37)

Based on Young’s inequality, we can derive that

$$
\sum_{i=1}^{N} \frac{\tilde{W}_i^T \tilde{k}_i \epsilon_i^t}{k_i^T k_i + 1} \leq \frac{1}{2} \sum_{i=1}^{N} \tilde{W}_i^T \tilde{k}_i \tilde{k}_i^T \tilde{W}_i + \frac{1}{2} \sum_{i=1}^{N} (\epsilon_{iM}^t)^2.
$$

(38)

Hence we have

$$
\dot{L}_0 \leq -\frac{1}{2} \sum_{i=1}^{N} \omega_i \|\tilde{W}_i\|^2 + \frac{1}{2} \sum_{i=1}^{N} (\epsilon_{iM}^t)^2.
$$

(39)

Then, (39) implies that $\dot{L}_0 < 0$ when $\|\tilde{W}_i\| > \sqrt{\sum_{i=1}^{N} (\epsilon_{iM}^t)^2 / \omega_i} \equiv B_{W_0}$. Therefore, $\dot{W}_i$ is UUB with ultimate bound $B_{W_0}$.

Theorem 2: Consider the constrained-input NZS system (1). The novel critic tuning law and control policy for each player are designed by (33) and (27), respectively. Suppose all the aforementioned assumptions hold. Then, the system states and weight estimation errors of the critic NN are UUB through the condition

$$
N^2 \gamma_M^2 \eta_{RM} \eta_M^4 \epsilon_m^2 < \lambda_m \omega_m,
$$

(40)

where $N$ is the number of players, $\gamma_M = \max[\gamma_{IM}, i \in \mathbb{N}]$, $\eta_{RM} = \max[\|R_{ii}^{-1}\|, i \in \mathbb{N}]$, $g_M = \max[g_{IM}, i \in \mathbb{N}]$, $\lambda_m = \min[\lambda_{min}(Q), i \in \mathbb{N}]$ and $\omega_m$ means the minimum exploratory noise signal.

Proof: The Lyapunov function candidate is constructed as

$$
L = L_a + L_b,
$$

(41)

where $L_a = \sum_{i=1}^{N} L_i(x)$ and $L_b = \frac{1}{2} \sum_{j=1}^{N} \tilde{W}_j^T \beta^{-1} \tilde{W}_j$. $L_a$ is given in Assumption 5. Taking the derivative of $L_a$ and $L_b$ of (41), we respectively obtain

$$
\dot{L}_a = \sum_{i=1}^{N} \dot{L}_i(x) = \sum_{i=1}^{N} \nabla L_i^T \left( f(x) + \sum_{j=1}^{N} g_i \hat{u}_j \right)
= \sum_{i=1}^{N} \sum_{j=1}^{N} \nabla L_i^T \left( f(x) - \sum_{j=1}^{N} \alpha_j g_j \tanh(\hat{M}_j(x)) \right),
$$

(42)

$$
\dot{L}_b = \sum_{i=1}^{N} \tilde{W}_i^T \beta^{-1} \tilde{W}_i.
$$

(43)

According to Lemma 1, we have

$$
\dot{L}_b \leq -\frac{1}{2} \sum_{i=1}^{N} \omega_i \|\tilde{W}_i\|^2 + \frac{1}{2} \sum_{i=1}^{N} (\epsilon_{iM}^t)^2 + \Theta,
$$

(44)

where $\Theta = -\frac{1}{2} \Omega(x, \hat{u}_1, \ldots, \hat{u}_N) \sum_{i=1}^{N} \tilde{W}_i^T \nabla y_i g_i R_{ii}^{-1} [M - \Psi(M_i(x))] g_i^T \sum_{j=1}^{N} \nabla L_j]$.

Then

$$
\dot{L} \leq \sum_{i=1}^{N} \sum_{j=1}^{N} \nabla L_i^T \left( f(x) - \sum_{j=1}^{N} \alpha_j g_j \tanh(\hat{M}_j(x)) \right)
- \frac{1}{2} \sum_{i=1}^{N} \omega_i \|\tilde{W}_i\|^2 + \frac{1}{2} \sum_{i=1}^{N} (\epsilon_{iM}^t)^2 + \Theta.
$$

(45)

The following two cases are considered for the convenience of the proof.

Case 1: $\Omega(x, \hat{u}_1, \ldots, \hat{u}_N) = 0$, which means that $\nabla L_i^T \left( f(x) + \sum_{j=1}^{N} g_j \hat{u}_j \right) < 0, \forall i \in \mathbb{N}$. Due to the PE condition, $\|x\| > 0$ is guaranteed. By choosing a positive
constant \( \delta \) that satisfies \( 0 < \delta < \| \dot{x} \| \), we can derive
\[
\dot{L} \leq -\delta \sum_{i=1}^{N} \| \nabla L_i \| - \frac{1}{2} \sum_{i=1}^{N} \omega_i \| \dot{W}_i \|^2 + \frac{1}{2} \sum_{i=1}^{N} (e_i^m)^2. \tag{46}
\]
\( \forall i \in \mathbb{N}, \) when one of the following two conditions holds:
\[
\| \nabla L_i \| > \sum_{i=1}^{N} (e_i^m)^2 / \omega_i \leq \bar{B}_{\nabla L_i}, \tag{47}
\]
\[\text{or}\]
\[
\| \dot{W}_i \| > \sum_{i=1}^{N} (e_i^m)^2 / \omega_i \leq \bar{B}_{\dot{W}_i}, \tag{48}\]
it yields that \( \dot{L} < 0. \)

Case ii: \( \Omega(x, \hat{u}_1, \ldots, \hat{u}_N) = 1, \) which means that
\[
\nabla L_i^T \left( f(x) + \sum_{j=1}^{N} g_j \hat{u}_j \right) \geq 0 \text{ holds for at least one player}.\]
The system (1) might be unstable and the additional term in (33) works. Then, (45) becomes
\[
\dot{L} \leq -\frac{1}{2} \sum_{i=1}^{N} \omega_i \| \dot{W}_i \|^2 + \frac{1}{2} \sum_{i=1}^{N} (e_i^m)^2 + \Gamma, \tag{49}
\]
where
\[
\Gamma = \sum_{i=1}^{N} \nabla L_i^T \left( f(x) - \sum_{j=1}^{N} \omega_j g_j \text{tanh}(\tilde{M}_j(x)) \right) - \frac{1}{2} \sum_{i=1}^{N} \tilde{W}_i^T \nabla g_j R_{ji}^{-1} \left[ L - \Psi(\tilde{M}_j(x)) \right] g_j^T \sum_{j=1}^{N} \nabla L_j.
\]
Furthermore,
\[
\Gamma = \sum_{i=1}^{N} \nabla L_i^T f(x) - \sum_{i=1}^{N} \nabla L_i^T \sum_{j=1}^{N} \omega_j g_j \left[ \text{tanh}(\tilde{M}_j(x)) \right]
+ \frac{1}{2 \omega_j} \left[ L - \Psi(\tilde{M}_j(x)) \right] R_{ji}^{-1} g_j^T \nabla L_i. \tag{50}
\]
By utilizing Taylor series expansion and the nature of hyperbolic function \( \text{tanh}(\cdot) \), from (50), we have
\[
\Gamma = \sum_{i=1}^{N} \nabla L_i^T \left( f(x) - \sum_{j=1}^{N} \omega_j g_j \text{tanh}(M_j(x)) \right)
+ \sum_{i=1}^{N} \nabla L_i^T \sum_{j=1}^{N} \omega_j g_j O(\tilde{M}_j(x) - M_j(x))^2 \tag{51}
\]
\[
= \sum_{i=1}^{N} \nabla L_i^T f(x) + \sum_{j=1}^{N} g_j u_j^* - \sum_{i=1}^{N} \nabla L_i^T \sum_{j=1}^{N} g_j \hat{u}_j
+ \sum_{i=1}^{N} \nabla L_i^T \sum_{j=1}^{N} \omega_j g_j O\left( (M_j(x) - \hat{M}_j(x))^2 \right)
\]
\[
\leq \sum_{i=1}^{N} \| \nabla L_i \| \left( 2 \sqrt{\bar{m}} \sum_{j=1}^{N} \omega_j g_j M + \sum_{j=1}^{N} g_j M \epsilon_j \right)
+ \sum_{i=1}^{N} \| \nabla L_i \| \sum_{j=1}^{N} \| R_{ji}^{-1} g^2_j \| \| \tilde{W}_j \|
- \sum_{i=1}^{N} \lambda_{\min}(Q_i) \| \nabla L_i \|^2. \tag{51}
\]
From (49) and (51), we obtain
\[
\dot{L} \leq -\frac{1}{2} \sum_{i=1}^{N} \lambda_{\min}(Q_i) \| \nabla L_i \|^2 + \sum_{i=1}^{N} \| \nabla L_i \|
- \frac{1}{2} \sum_{i=1}^{N} \lambda_{\min}(Q_i) \left( \frac{\sum_{j=1}^{N} R_{ji}^{-1} \| g_j M \| \| \tilde{W}_j \|}{\lambda_{\min}(Q_i)} \right)
+ \| \nabla L_i \| \left( \frac{\sum_{j=1}^{N} R_{ji}^{-1} \| g_j M \| \| \tilde{W}_j \|}{\lambda_{\min}(Q_i)} \right)
\]
\[
- \frac{1}{2} \sum_{i=1}^{N} \omega_i \| \dot{W}_i \|^2 + \frac{1}{2} \sum_{i=1}^{N} (e_i^m)^2
\]
\[
\leq -\frac{1}{2} \sum_{i=1}^{N} \lambda_{\min}(Q_i) \left( \frac{\sum_{j=1}^{N} R_{ji}^{-1} \| g_j M \| \| \tilde{W}_j \|}{\lambda_{\min}(Q_i)} \right)^2
\]
\[
+ \frac{N \lambda^2}{2 \lambda^2 - \sum_{i=1}^{N} \| g_i \|^2} \sum_{i=1}^{N} \| \tilde{W}_i \|^2
+ \frac{N \lambda^2}{2 \lambda^2 - \sum_{i=1}^{N} \| g_i \|^2} \sum_{i=1}^{N} (e_i^m)^2. \tag{52}
\]
where \( \chi = (2 \sqrt{\bar{m}} \sum_{j=1}^{N} \omega_j g_j M + \sum_{j=1}^{N} g_j M \epsilon_j \). Observe (52), if the condition
\[
\sum_{j=1}^{N} \| g_j M \| \| \tilde{W}_j \| < \lambda_m \omega_m
\]
then, \( \forall i \in \mathbb{N}, \) when at least one condition as follows is satisfied:
\[
\| \nabla L_i \| > \sqrt{\left( \frac{N \lambda^2}{\lambda_m^2} + \sum_{i=1}^{N} (e_i^m)^2 \right) / \lambda_m + \chi} / \lambda_m
\]
\( \triangleq \bar{B}_{\nabla L_i}, \tag{53}\)
\[\text{or}\]
\[
\| \tilde{W}_i \| > \sqrt{\left( \frac{N \lambda^2}{\lambda_m^2} + \sum_{i=1}^{N} (e_i^m)^2 \right) \left( \omega_m - \frac{N^2 \lambda^2}{\lambda_m^2} \right) \lambda_m^2 - \sum_{i=1}^{N} \| g_i \|^2} / \lambda_m^2
\]
\( \triangleq \bar{B}_{\tilde{W}_i}, \tag{54}\)
it holds that \( \dot{L} < 0. \)

Based on the analysis of the above two cases and according to the Lyapunov extension theorem, we can draw a conclusion that \( \forall i \in \mathbb{N}, \) the function \( \nabla L_i \) and the weight estimation error \( \tilde{W}_i \) are both UUB. The corresponding bounds are respectively denoted by \( B_{\nabla L_i} = \max(\bar{B}_{\nabla L_i}, \tilde{B}_{\nabla L_i}) \) and \( B_{\tilde{W}_i} = \max(\bar{B}_{\tilde{W}_i}, \tilde{B}_{\tilde{W}_i}). \) Moreover, the system state \( x \) is UUB due to the property of the function \( \nabla L_i. \)

**Remark 3:** Different from the method proposed in [37] that each player needs both critic NN and actor NN, the presented method in this article requires only critic NN for each player to tackle the coupled HJ equations. By abnegating the actor NN, identifier-critic framework is employed, which can effectively simplify the algorithm complexity and save the computing resources.

**Remark 4:** For the introduction of the indicator function \( \Omega(x, \hat{u}_1, \ldots, \hat{u}_N), \) the above analysis considers two different cases to demonstrate the UUB of the system.
states and the critic NN weight estimation errors. When 
\( \Omega(x, \hat{u}_1, \ldots, \hat{u}_N) = 1 \), it means that the system might not be stable. Then the stabilizing term is added to stabilize the system. By this means the requirement for the initial stabilizing control is removed. It’s noted that when the initial control is a stabilizing one and the added noise signal doesn’t affect the system stability, \( \Omega(x, \hat{u}_1, \ldots, \hat{u}_N) = 0 \) holds all the time during the NN training phase and the designed tuning laws for critic NN are equivalent to the standard tuning laws without the designed term. In addition, based on the analysis of the above two cases, the bound of the weight estimation error \( \hat{W}_i \) is obtained, which will be utilized in the following theorem 3.

**Theorem 3:** Consider the NZS game with control constraints of system (1). The critic NN weight tuning laws and the control schemes are given by (33) and (27). On the basis of the aforementioned assumptions and the condition (40), then the obtained control \( \hat{u}_i \) converges to \( u_i^* \) with a small bound.

**Proof:** When all the assumptions and the condition (40) hold, according to (26), (27) and the conclusion of Theorem 2, we derive

\[
\| \hat{u}_i - u_i^* \| = \| \alpha_i (\tanh(M_i(x)) - \tanh(\hat{M}_i(x))) - \varepsilon u_i^* \| \\
= \| \frac{1}{2} [U_m - \Psi(\hat{M}_i(x))] R_{\hat{u}i}^{-1} \gamma_i^\top \hat{W}_i \\
+ \alpha_i O((M_i(x) - \hat{M}_i(x))^2) - \varepsilon u_i^* \| \\
\leq 2\alpha_i \sqrt{m} + 2\| R_{\hat{u}i}^{-1} \| \gamma_i \| \hat{W}_i \| + \varepsilon u_i^* \\
\leq 2\alpha_i \sqrt{m} + 2\| R_{\hat{u}i}^{-1} \| \gamma_i \| \hat{W}_i \| + \varepsilon u_i^* \\
\equiv \varepsilon u_{ib}. \quad (55)
\]

**Remark 5:** By observing (55) we can draw that the acquired control \( \hat{u}_i \) converges to \( u_i^* \) with the small bound \( \varepsilon u_{ib} \). Hence, the acquired control signal is UUB with the tuning law (33).

**V. SIMULATIONS**

In this section, two examples are simulated to illustrate the validity of the developed method.

**Example 1:** We consider a two-player linear NZS differential game of which the model is

\[
\dot{x} = \begin{bmatrix}
-0.20377 & -0.00043 \\
0.2 & -0.2
\end{bmatrix} x + \begin{bmatrix}
0 \\
0.15
\end{bmatrix} u_1 + \begin{bmatrix}
0.15 \\
0.3
\end{bmatrix} u_2, \tag{56}
\]

where \( x = [x_1, x_2]^\top \in \mathbb{R}^2 \) is the system state, \( u_1 \in \mathbb{R} \) and \( u_2 \in \mathbb{R} \) are the control policies for the two players, respectively. In addition, \( |u_1| \leq 0.4 \) and \( |u_2| \leq 0.4 \). Let \( Q_1 = 2, Q_2 = 1, R_{11} = 1, R_{12} = 2, R_{21} = 1, R_{22} = 1 \).

The activation function for the identifier NN is \( \gamma(x) = [x_1^2, x_1 x_2, x_2^2, x_1^3, x_1 x_2^2, x_2^3, x_1 x_3, x_1^2 x_2, x_2 x_3, x_1 x_2 x_3] \) and that for the i-th critic NN is chosen as \( \gamma(x) = [x_1^2, x_1 x_2, x_2^2] \) where \( i = 1, 2 \). \( S = [-0.5, 1.2; -0.3, -0.3] \) and \( H = I_8 \).

**FIGURE 1.** The evolutions of the states estimation error.

**Example 2:** In this experiment, in order to illustrate the effectiveness of the proposed method for multi-player nonlinear systems, a three-player nonlinear system is considered as

\[
\dot{x} = \begin{bmatrix}
0.15 x_2 - 0.3 x_1 \\
-0.075 x_1 - 0.075 x_2 + 0.075 x_1^2 x_2 \\
0.05 \\
0.4
\end{bmatrix} u_1 + \begin{bmatrix}
0.2 \\
0.1 \sin(x_1)
\end{bmatrix} u_2 + \begin{bmatrix}
0.02 \\
0.1
\end{bmatrix} u_3, \tag{57}
\]

where \( x = [x_1, x_2] \in \mathbb{R}^2 \) is the system state vector. And \( u_i \in \mathbb{R} \) is the control policy of player \( i \) with \( |u_i| \leq 0.6, i = 1, 2, 3 \). Choose \( Q_1 = 2, Q_2 = 1, Q_3 = 2, R_{11} = 2, R_{12} = 2, R_{13} = 1, R_{21} = 1, R_{22} = 1, R_{23} = 2, R_{31} = 3, R_{32} = 2, R_{33} = 2 \).

The activation function for the identifier is \( \gamma(x) = [x_1^2, x_1 x_2, x_2^2, x_1^3 x_2, x_1 x_2^2, x_2 x_3, x_1 x_3, x_1^2 x_2, x_2 x_3, x_1 x_2 x_3] \) and that for the
$i$–th player is selected as $y_i(x) = [x_i^2, x_1, x_2, x_3^2]^T$, $i = 1, 2, 3$. $S = [-0.2, 1.2; -0.2, -0.1]$ and $H = I_8$. The weight vector for the $i$–th critic NN is $\hat{W}_i = [\hat{W}_{i1}, \hat{W}_{i2}, \hat{W}_{i3}]^T$. The learning rate is chosen as $\beta = 0.8$. The initial state is $x_0 = [-1, 1]^T$ and the initial weights for the critic NN are randomly selected with the interval of $[-0.5, 0.5]$. The PE condition is fulfilled by a probing noise signal $s_2(t) = 3.2e^{-0.03t}[\sin(2t) \cos(2t) + \sin^3(t) \cos(t) + \sin^3(2t) \cos(0.8t) + \sin^3(-1.2t) \cos(1.5t) + \sin^3(1.12t) \cos(2.2t) + \sin^3(2.4t) \cos(2.4t) + \sin^3(t)]$. At $t = 120s$, the noises are removed.

The simulation results are presented in Figs. 6-11. The states estimation error curve of identifier is presented
in Fig. 6 to show the validity of the designed identifier. From Fig. 7, by observing the trajectories of the states, we can find that the amplitudes go down to zero fleetly when the noises are removed. Figs. 8-10 present the convergence of the weights for all of the critic NN and the weights for all networks come to constants after $t = 110s$. Fig. 11 shows the control policies during the learning phase of the critic NN.

VI. CONCLUSION

In this article, a novel online ADP method was developed to solve multi-player NZS control issue with control constraints. To tackle the constrained inputs, a number of non-quadratic terms were first utilized to constitute the performance index functions. Then, for simplifying the algorithm and easing the computing burdens, only identifier NN and critic NN were employed to approximately solve the coupled HJ equations of the partially unknown NZS games. A tuning law with a novel stability term was developed for each critic NN such that the stability of the closed-loop system was guaranteed during NN training phase and the need for the initial stabilizing control was removed. All the signals of the system were demonstrated to be UUB by utilizing Lyapunov theory. Finally, a two-player linear system and a three-player nonlinear system were simulated to verify the validity of the novel ADP method.

Our future work is to employ event-triggered mechanism in the present algorithm framework to further save communication and computing resources. Meanwhile, in order to improve the applicability and practicability of the algorithm, we will focus on investigating the present method to apply to the stochastic systems, the switched systems and other types of nonlinear systems with completely unknown dynamics.

REFERENCES

[1] R. Isaacs, Differential Games. Santa Monica, CA, USA: RAND Corporation, 1955.
[2] S. Clemhou and H. Wan, “Differential games-economic applications,” in Handbook Game Theory Econ. Amsterdam, The Netherlands: Elsevier, 1994, pp. 801–825.
[3] Z. Ni and S. Paul, “A multistage game in smart grid security: A reinforcement learning solution,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 9, pp. 2684–2695, Sep. 2019.
[4] D. Wang, H. He, C. Mu, and D. Liu, “Intelligent critic control with disturbance attenuation for affine dynamics including an application to a microgrid system,” IEEE Trans. Ind. Electron., vol. 64, no. 6, pp. 4935–4944, Jun. 2017.
[5] H. Zhang, L. Cui, and Y. Luo, “Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP,” *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 206–216, Feb. 2013.

[6] H. Zhang, J. Zhang, G.-H. Yang, and Y. Lao, “Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming,” *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 1, pp. 152–163, Feb. 2015.

[7] 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.

[8] Z. Qiang and D. Zhao, “Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics,” *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 2874–2884, Aug. 2019.

[9] 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.

[10] H. He, Z. Ni, and J. Fu, “A three-network architecture for on-line learning and optimization based on adaptive dynamic programming,” *Neurocomputing*, vol. 78, no. 1, pp. 3–13, 2012.

[11] J. Si and Y.-T. Wang, “Online learning control by association and reinforcement,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 12, no. 2, pp. 264–276, Mar. 2001.

[12] D. Liu, H. Li, and D. Wang, “Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm,” *Neurocomputing*, vol. 110, pp. 92–100, Jun. 2013.

[13] R. Song, F. L. Lewis, and Q. Wei, “Off-policy integral reinforcement learning for a class of unknown nonlinear constrained-input systems via integral reinforcement learning,” *IEEE Trans. Fuzzy Syst.*, vol. 46, no. 5, pp. 1041–1049, May 2016.

[14] A. Sahoo, H. Xu, and S. Jagannathan, “Near optimal event-triggered control of nonlinear discrete-time systems using neurodynamic programming,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 9, pp. 1801–1815, Sep. 2016.

[15] R. Song, Q. Wei, and B. Song, “Neural-network-based synchronous iteration learning method for multi-player zero-sum games,” *Neurocomputing*, vol. 242, pp. 73–82, Jun. 2017.

[16] R. Padhi, N. Umikrishnan, X. Wang, and S. N. Balakrishnan, “A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of constrained-input systems,” *Neural Netw.*, vol. 19, no. 10, pp. 1648–1660, Dec. 2006.

[17] Q. Zhang, D. Zhao, and D. Wang, “Event-based robust control for uncertain nonlinear systems using adaptive dynamic programming,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 37–49, Jan. 2018.

[18] D. Wang, C. Mu, D. Liu, and H. Ma, “On mixed data and event driven design for Adaptive-Critic-Based nonlinear H∞ control,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 993–1005, Apr. 2018.

[19] L. Ma, N. Xu, X. Hua, and X. Zhao, “Adaptive finite-time output-feedback control design for switched pure-feedback nonlinear systems with average dwell time,” *Nonlinear Anal., Hybrid Syst.*, vol. 37, Aug. 2020, Art. no. 100908, doi: 10.1016/j.nahs.2020.100908.

[20] Y. Chang, Y. Wang, F. E. Alsaadi, and G. Zong, “Adaptive fuzzy output feedback control for nonlinear NCSs with quantization and stochastic communication protocol,” *IEEE Trans. Fuzzy Syst.*, early access, Jun. 2020, doi: 10.1109/TFUZZ.2020.3005342.

[21] H. Su, H. Zhang, Y. Liang, and Y. Mu, “Online event-triggered adaptive critic design for partially unknown networked systems,” *Neurocomputing*, vol. 368, pp. 84–98, Nov. 2019.

[22] Y. Lv, X. Ren, and J. Na, “Adaptive optimal tracking controls of unknown multi-input systems based on zero-sum game theory,” *J. Franklin Inst.*, vol. 356, no. 15, pp. 8255–8272, Oct. 2019.

[23] D. Liu, H. Li, and D. Wang, “Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics,” *IEEE Trans. Syst., Man., Cybern. Syst.*, vol. 44, no. 8, pp. 1015–1027, Aug. 2014.

[24] Y. Lv, X. Ren, and J. Na, “Online optimal solutions for multi-player nonzero-sum game with completely unknown dynamics,” *Neurocomputing*, vol. 283, pp. 87–97, Mar. 2018.

[25] R. Song, F. L. Lewis, and Q. Wei, “Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 704–712, Mar. 2017.

[26] H. Su, H. Zhang, S. Sun, and Y. Cai, “Integral reinforcement learning-based online adaptive event-triggered control for non-zero-sum games of partially unknown nonlinear systems,” *Neurocomputing*, vol. 377, pp. 243–255, Feb. 2020.

[27] J. Cai, R. Yu, B. Wang, C. Mei, and L. Shen, “Decentralized event-triggered control for interconnected systems with unknown disturbances,” *J. Franklin Inst.*, vol. 357, no. 3, pp. 1494–1515, Feb. 2020.

[28] Z.-M. Li, X.-H. Chang, and H. J. Park, “Quantized static output feedback fuzzy tracking control for discrete-time nonlinear networked systems with asynchronous event-triggered constraints,” *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Aug. 15, 2019, doi: 10.1109/TSYMC.2019.2931573.

[29] C. Mu and K. Wang, “Aperiodic adaptive control for neural-network-based nonzero-sum differential games: A novel event-triggering strategy,” *ISA Trans.*, vol. 92, pp. 1–13, Sep. 2019.

[30] X. Yang, D. Liu, H. Ma, and Y. Xu, “Online approximate solution of HJI equation for unknown constrained-input nonlinear continuous-time systems,” *Inf. Sci.*, vol. 328, pp. 435–454, Jan. 2016.

[31] X. Yang, D. Liu, B. Luo, and C. Li, “Data-based robust adaptive control for a class of unknown nonlinear constrained-input systems via integral reinforcement learning,” *Inf. Sci.*, vol. 369, pp. 731–747, Nov. 2016.

[32] L. Dong, X. Zhong, C. Sun, and H. He, “Event-triggered adaptive dynamic programming for continuous-time systems with control constraints,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1941–1952, Aug. 2017.

[33] D. Wang, C. Mu, Q. Zhang, and D. Liu, “Event-based input-constrained nonlinear H∞ state feedback with adaptive critic and neural implementation,” *Neurocomputing*, vol. 214, pp. 848–856, Nov. 2016.

[34] H. Ren, H. Zhang, Y. Wen, and C. Li, “Integral reinforcement learning off-policy method for solving nonlinear multi-player nonzero-sum games with saturated actuator,” *Neurocomputing*, vol. 335, pp. 96–104, Mar. 2019.

[35] 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, Aug. 2012.

[36] S. Xue, B. Luo, D. Liu, and Y. Li, “Adaptive dynamic programming based event-triggered control for unknown continuous-time nonlinear systems with input constraints,” *Neurocomputing*, vol. 396, pp. 191–200, Jul. 2020.

[37] H. Modares, F. L. Lewis, and M. B. Naghibi-Sistani, “Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning,” *Automatica*, vol. 50, no. 7, pp. 1780–1792, Jul. 2014.

[38] D. Wang, D. Liu, C. Mu, and Y. Zhang, “Neural network learning and robust stabilization of nonlinear systems with dynamic uncertainties,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1342–1351, Apr. 2018.

[39] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, “Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.

[40] K. G. Vamvoudakis and F. L. Lewis, “Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations,” *Automatica*, vol. 47, no. 8, pp. 1556–1569, Aug. 2011.
HUAGUANG ZHANG (Fellow, IEEE) received the B.S. and M.S. degrees in control engineering from Northeast Dianli University, Jilin City, China, in 1982 and 1985, respectively, and the Ph.D. degree in thermal power engineering and automation from Southeast University, Nanjing, China, in 1991.

He joined the Department of Automatic Control, Northeastern University, Shenyang, China, in 1992, as a Postdoctoral Fellow for two years. Since 1994, he has been a Professor and the Head of the School of Information Science and Engineering, Institute of Electric Automation, Northeastern University. His main research interests include fuzzy control, stochastic system control, neural networks-based control, and nonlinear control and their applications. He has authored or coauthored over 280 journal articles and conference papers, six monographs, and co-invented 90 patents.

Dr. Zhang was awarded the Outstanding Youth Science Foundation Award from the National Natural Science Foundation Committee of China, in 2003. He was named the Cheung Kong Scholar by the Education Ministry of China, in 2005. He was a recipient of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 2012 Outstanding Paper Award and the Andrew P. Sage Best Transactions Paper Award, in 2015. He is the E-Letter Chair of the IEEE CIS Society and the former Chair of the Adaptive Dynamic Programming and Reinforcement Learning Technical Committee on the IEEE Computational Intelligence Society. He was an Associate Editor of IEEE TRANSACTIONS ON FUZZY SYSTEMS, from 2008 to 2013. He is an Associate Editor of Automatica, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, and Neurocomputing.

CHONG LIU received the B.S. degree in electronic and information engineering from Inner Mongolia Normal University, China, in 2011, and the M.S. degree in electronic science and technology from the Changchun University of Science and Technology, Changchun, China, in 2015. He is currently pursuing the Ph.D. degree in control theory and control engineering with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His current research interests include adaptive dynamic programming, neural networks, and optimal control.

HANGUANG SU received the B.S. degree in automation control and the M.S. degree in control engineering from Northeastern University, Shenyang, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree in control theory and control engineering with the College of Information Science and Engineering.

His research interests include reinforcement learning, optimal control, fuzzy control, and adaptive dynamic programming and their applications.

* * *