Nonlinear Dyn (2023) 111:11671–11683
https://doi.org/10.1007/s11071-023-08419-5

ORIGINAL PAPER

Adaptive critic design for nonlinear multi-player zero-sum games with unknown dynamics and control constraints

Yu Huo · Ding Wang · Junfei Qiao · Menghua Li

Received: 12 November 2022 / Accepted: 16 March 2023 / Published online: 12 April 2023

Abstract In this paper, a novel optimal control scheme is established to solve the multi-player zero-sum game (ZSG) issue of continuous-time nonlinear systems with control constraints and unknown dynamics based on the adaptive critic technology. To relax the requirement of system dynamics, a neural network-based identifier is applied to reconstruct the unknown multi-player ZSG system. Then, by developing a new nonquadratic function, the associated Hamilton-Jacobi-Isaacs (HJI) equation of the constrained ZSG is derived. Moreover, an adaptive critic framework is constructed to approximate the optimal cost function. Meanwhile, the strategy sets of optimal control and the worst disturbance are estimated by utilizing the single-critic network, respectively. After that, a modified critic weight updating mechanism with experience replay technique is introduced to relax the requirement of the persistence of excitation condition. Theoretically, by employing the Lyapunov stability theorem, the uniform ultimate boundedness stability of the ZSG system state and the critic network weight approximation error are proved. Finally, a representative example is simulated to validate the efficacy of the constructed framework.

Keywords Adaptive critic designs · Adaptive dynamic programming (ADP) · Control constraints · Experience replay (ER) · Multi-player zero-sum games (ZSGs) · System identification

Y. Huo · D. Wang · J. Qiao (✉) · M. Li
Faculty of Information Technology, The Beijing Key Laboratory of Computational Intelligence and Intelligent System, The Beijing Institute of Artificial Intelligence, and The Beijing Laboratory of Smart Environmental Protection, Beijing University of Technology, Beijing 100124, China
e-mail: adqiao@bjut.edu.cn

Y. Huo
e-mail: HuoYu@emails.bjut.edu.cn

D. Wang
e-mail: dingwang@bjut.edu.cn

M. Li
e-mail: limenghua@emails.bjut.edu.cn

1 Introduction

Differential game theory has emerged as an advantageous instrument for theoretical and applied study in the control field [1–4]. For complex real-world systems, many of these consist of multiple controllers which could be deemed as some relevant players in games [5]. When the differential game is employed to the control issues, it essentially translates to the multi-controller optimal control problems [6,7]. Generally, game problems are distinguished into zero-sum games (ZSGs) [8] and non-ZSGs [9]. Furthermore, game problems depend on solving the Hamilton-Jacobi (HJ) or the Hamilton-Jacobi-Isaacs (HJI) equations in the optimal control issue which are intractable or impossible to solve [10–14]. Therefore, many scholars have shown different approximate approaches to tackle this difficulty. Specially, the adaptive dynamic programming

(ADP) technology was introduced to address HJ or HJI equations for varieties of game problems [15–18].

For non-ZSG issues, an off-policy scheme was established to deal with the multi-player non-ZSG, while the system dynamics was not required [19]. By employing an ADP-based mechanism, each player could gain the performance index and control. Zhao et al. [20] designed a dual actor-critic scheme. And there is proof to verify control policies have reached the Nash equilibrium for the non-ZSG. In [21], an online ADP-based model-free control structure was proposed to handle the multi-player non-ZSG problem for discrete-time unknown systems.

With regard to ZSG problems, the H_∞-constrained control issue is transformed into a two-player ZSG problem, which pointed a key direction for H_∞ control problems [22]. After that, under the scheme of ADP, a critic network was designed to deal with the HJI equation. In [23], the linear two-player ZSG were investigated based on an adaptive online learning architecture, which was utilized to approximately solve the modified game algebraic Riccati equation via online data. Yazidi et al. [24] developed a pioneering mechanism that is capable of converging to the mixed Nash equilibrium by solving two-player ZSG with incomplete information. Different from [25], Song et al. [26] established an only single-critic network framework to turn the weight and solve the HJI equation without complete dynamics. The aforementioned ADP-based results were only researched for two-player ZSG problems. However, most industrial process plants are commonly controlled by multiple controllers. This means that the cost function designed for two-player ZSGs no longer applies to multi-player ZSGs. Therefore, multi-player ZSGs should attract more attention. In [27], an off-policy framework was devised for multi-player ZSG of completely unknown systems. Therewith, the iterative cost function, controls and disturbances were obtained. In [28], the single-critic mechanism was employed and the event-based structure was developed in a multi-player ZSG form to reduce the data transmission and computation. Qiao et al. [29] extended the adaptive critic mechanism to the problem of combining multi-player ZSG and optimal tracking control. Then, this work provided two cases of multi-player ZSG in the simulation stage.

Note that the most of aforementioned ADP-based frameworks require the known system dynamics, which is difficult to achieve accurately in industrial process. To overcome this disadvantage, system identification algorithms based on neural networks are utilized to reconstruct unknown system dynamics by approximate structures [30–32]. For example, Na et al. [33] utilized critic-based ADP and identifier network approaches by means of system data to online address the optimal tracking control issue. In [34], an ADP mechanism for a two-player nonlinear ZSG was designed by utilizing the identifier-critic network. In [35], an intelligent control mechanism was established by using the recurrent neural network and a unique critic network, instead of utilizing the mathematical model. Huo et al. [36] extended the results to constrained decentralized systems by utilizing the identifier-critic mechanism.

Subsequently, in order to address the problem caused by the persistence of excitation (PE) condition, the experience replay (ER) scheme was designed for nonlinear systems [37–40]. The ER scheme can effectively utilize the historical and available data simultaneously. Under the ER framework, a novel ADP-based approach was developed in [41] to approximate the Nash equilibrium for multi-player non-ZSGs with unknown drift dynamics, which also could accelerate the convergence rate of critic network weights. In [42], the critic network was developed with a new weight updating rule based on the ER method for uncertain interconnections systems. Thereafter, Zhu et al. [43] realized the optimal control of constrained-input partially unknown systems. In order to tune the critic network weight, they leveraged the ER algorithm to effectively use the record data. To relax the PE condition, the ER scheme was introduced to off-policy framework to address the optimal output regulation issue with unknown system dynamics [44].

Moreover, control constraints are considered to be wide-spread factors in practical systems due to the inherent physical properties of the actuators. As a result, the system performance is likely poor or even unstable. Thus, the developed ADP-based controller is supposed to obtain the desired performance with control constraints [45–49]. Accounting for control constraints, an adaptive critic design based on ADP was implemented for nonlinear non-ZSGs in the two-player form [50]. Further, an actor-critic architecture was proposed to approximately gain the Nash equilibrium by utilizing the real-time data. In [51], the unknown multi-player ZSG with control constraints was considered, and the observer-critic structure was established to tackle the HJI equation. Sun and Liu

[52] investigated a fixed directed graph structure for multi-agent systems with control constraints to handle the distributed differential game tracking issue. Nevertheless, control constraints for multi-player ZSGs are considered in only few studies. More importantly, the ADP-based optimal control for ZSGs was also investigated in [26,28,51] and [53]. However, the single-critic network scheme was not established in [53], the constrained control input was not considered in [26] and [28], and the weight updating rule with the ER technology was not analyzed in [51]. These works promote our research interests. Hence, this article concerns the ER-based adaptive critic design for unknown multi-player ZSGs with control constraints.

The innovations of this article can be listed as four parts.

1. This paper extends the ADP-based scheme to solve the multi-player ZSG issue for the nonlinear system. It is appropriate for both two-player ZSG problem and multi-player ZSG problem.
2. Additionally, by constructing a modified non-quadratic utility function, control constraints are considered under the multi-player ZSG situation.
3. Different from the traditional identifier-actor-critic mechanism [54], the identifier-critic scheme for all players is developed to solve the HJI equation, which can further simplify the method structure and reduce the computing cost.
4. By introducing the ER mechanism, a novel weight tuning criterion is employed and the PE condition is relaxed to an easy-checked rank condition [see Remark 3], which means an easy-to-execute scheme is designed. Moreover, the uniform ultimate boundedness (UUB) stability of the critic network weight estimation error and the multi-player system can be both guaranteed.

The outline of the article is summarized as follows. In Sect. 2, the problem description is provided. In Sect. 3, a neural network-based identifier is established to identify the system dynamics, and the stability is proved. Moreover, the single-critic network scheme is introduced with the stability analysis. In Sect. 4, one simulation example is shown. In Sect. 5, the conclusion is presented.

2 Problem statement

Consider the multi-player nonlinear ZSG system

$$\dot{x}(t) = f(x(t)) + \sum_{q=1}^{N} g_q(x(t))u_q(t) + \sum_{l=1}^{M} k_l(x(t))d_l(t), \tag{1}$$

where $x \in \mathbb{R}^n$ denotes the state; $u_q \in \mathbb{R}^{m_q}$ and $d_l \in \mathbb{R}^{w_l}$ are the constrained control inputs and the disturbance inputs, respectively. Note that $f(x) \in \mathbb{R}^n$, $g_q(x) \in \mathbb{R}^{n\times m_q}$ and $k_l(x) \in \mathbb{R}^{n\times w_l}$ are assumed unknown and Lipschitz continuous on a compact set $\Omega \in \mathbb{R}^n$ with $f(0) = 0$. Let $x(0) = x_0$ be the initial state and the system is stabilizable on Ω.

Define the cost function as

$$J(x_0, \mathscr{U}, \mathscr{D}) = \int_0^\infty h(x(t), \mathscr{U}, \mathscr{D})\mathrm{d}\tau, \tag{2}$$

where $\mathscr{U} = \{u_1, \ldots, u_N\}$ is the set of constrained control inputs, $|u_q| \le \beta_q$ with $\beta_q > 0$ being the constraint bound. $\mathscr{D} = \{d_1, \ldots, d_M\}$ is the set of disturbance inputs, $h(x(t), \mathscr{U}, \mathscr{D}) = x^\mathsf{T} Qx + U(\mathscr{U}, \mathscr{D})$ is the utility function, and $U(\mathscr{U}, \mathscr{D}) = 2\sum_{q=1}^{N} R_q \int_0^{u_q} \beta_q \rho^{-\mathsf{T}}(v/\beta_q)\mathrm{d}v - \lambda^2 \sum_{l=1}^{M} d_l^\mathsf{T} d_l$ with λ denoting the disturbance attenuation level. $Q \ge 0$ and $R_q \ge 0$ are positive symmetric matrices. Moreover, $\rho(\cdot)$ is a monotonic bounded odd function and we choose $\rho(\cdot) = \tanh(\cdot)$.

Then, the multi-player ZSG subject to (1) is defined as

$$J^*(x_0) = \inf_{u_1}\inf_{u_2}\cdots\inf_{u_N}\sup_{d_1}\sup_{d_2}\cdots\sup_{d_M} J(x_0, \mathscr{U}, \mathscr{D}), \tag{3}$$

where $J^*(x)$ denotes the optimal cost function.

For the multi-player ZSG, it seeks to attain the saddle point solution (u_q^*, d_l^*) to satisfy the inequalities

$$J(x, \mathscr{U}^*, \mathscr{D}) \le J(x, \mathscr{U}^*, \mathscr{D}^*) \le J(x, \mathscr{U}, \mathscr{D}^*), \tag{4}$$

where $\mathscr{U}^* = \{u_1^*, u_2^*, \ldots, u_N^*\}$ and $\mathscr{D}^* = \{d_1^*, d_2^*, \ldots, d_M^*\}$ indicate the sets of the optimal control strategies and the worst disturbance strategies, respectively.

Based on cost function (2), one has

$$0 = h(x, \mathscr{U}, \mathscr{D}) + (\nabla J(x))^\mathsf{T}\left(f(x) + \sum_{q=1}^{N} g_q(x)u_q + \sum_{l=1}^{M} k_l(x)d_l\right), \tag{5}$$

where $\nabla(\cdot) \triangleq \partial(\cdot)/\partial x$ denotes the gradient operator.

The Hamiltonian function is constructed as

$$
\begin{aligned}
&H\left(x, \nabla J(x), \mathscr{U}, \mathscr{D}\right) \\
&\quad = x^{\mathsf{T}} Q x + 2\sum_{q=1}^{N} R_q \int_0^{u_q} \beta_q \rho^{-\mathsf{T}}(v/\beta_q)\mathrm{d}v \\
&\quad - \lambda^2 \sum_{l=1}^{M} d_l^{\mathsf{T}} d_l \\
&\quad + (\nabla J(x))^{\mathsf{T}} \left(f(x) + \sum_{q=1}^{N} g_q(x) u_q \right. \\
&\quad \left. + \sum_{l=1}^{M} k_l(x) d_l \right).
\end{aligned} \tag{6}
$$

The associated HJI equation can be described as

$$
\min_{\mathscr{U}} \max_{\mathscr{D}} H\left(x, \nabla J^*(x), \mathscr{U}, \mathscr{D}\right) = 0. \tag{7}
$$

Then, the optimal constrained control policy and the worst disturbance strategy can be derived from the following stationary conditions

$$
\frac{\partial H\left(x, \mathscr{U}, \mathscr{D}, \nabla J^*(x)\right)}{\partial u_q} = 0, \quad q = 1, 2, \ldots, N, \tag{8}
$$

$$
\frac{\partial H\left(x, \mathscr{U}, \mathscr{D}, \nabla J^*(x)\right)}{\partial d_l} = 0, \quad l = 1, 2, \ldots, M. \tag{9}
$$

Therefore, the optimal control law and the worst disturbance law can be obtained by

$$
u_q^* = -\beta_q \tanh(B^*), \tag{10}
$$

$$
d_l^* = \frac{1}{2\lambda^2} k_l^{\mathsf{T}} \nabla J^*, \tag{11}
$$

where $B^* = (1/(2\beta_q)) R_q^{-1} g_q^{\mathsf{T}} \nabla J^*$.

Inserting (10) and (11) into (7), we can get the HJI equation expressed as

$$
\begin{aligned}
0 &= x^{\mathsf{T}} Q x \\
&\quad + 2\sum_{q=1}^{N} \left(R_q \int_0^{-\beta_q \tanh(B^*)} \beta_q \tanh^{-\mathsf{T}}(v/\beta_q)\mathrm{d}v \right) \\
&\quad + \frac{1}{4\lambda^2} \sum_{l=1}^{M} \left((\nabla J^*)^{\mathsf{T}} k_l(x) k_l^{\mathsf{T}}(x) \nabla J^* \right) + (\nabla J^*)^{\mathsf{T}} f(x) \\
&\quad - (\nabla J^*)^{\mathsf{T}} \sum_{q=1}^{N} \left(g_q(x) \beta_q \tanh(B^*) \right).
\end{aligned} \tag{12}
$$

Note that it is intractable to tackle equation (12). Generally, the traditional policy iteration (PI) scheme can be employed to overcome this bottleneck, but this scheme depends on the system dynamics. Hence, in the next section, the identifier-critic network framework is developed which can tackle the constrained multi-player ZSG issue without requiring the system dynamics.

Remark 1 Obviously, this paper considers the multi-player ZSG with control constraints. Therefore, the traditional quadratic cost function is no longer suitable for solving such issue. In this paper, the control constraint problem can be tackled by utilizing an improved non-quadratic cost function which restricts the control policies within the given bound.

3 Approximate solution for multi-player ZSGs

In this section, an identifier-critic framework based on neural networks is constructed for the multi-player ZSG problem of unknown dynamics with control constraints.

First, an identifier network is designed to relax the requirement of unknown system dynamics. Then, a single-critic network is applied and the implementation process is also given. Finally, the stability is proved by using the Lyapunov approach.

3.1 System identification

For the multi-player ZSG system dynamics is unknown, an identifier is used to reconstruct the unknown dynamics. System (1) can be reformulated by

$$
\begin{aligned}
\dot{x} &= Sx + \omega_f^{\mathsf{T}} \varphi_f(x) + \varepsilon_f(x) \\
&\quad + \sum_{q=1}^{N} \left(\omega_{gq}^{\mathsf{T}} \varphi_{gq}(x) + \varepsilon_{gq}(x) \right) u_q \\
&\quad + \sum_{l=1}^{M} \left(\omega_{kl}^{\mathsf{T}} \varphi_{kl}(x) + \varepsilon_{kl}(x) \right) d_l,
\end{aligned} \tag{13}
$$

where $S \in \mathbb{R}^{n\times n}$ is a designed matrix. $\omega_f \in \mathbb{R}^{n\times n}$, $\omega_{gq} \in \mathbb{R}^{n\times n}$, and $\omega_{kl} \in \mathbb{R}^{n\times n}$ represent the ideal weight matrices. $\varphi_f(\cdot) \in \mathbb{R}^{n}$, $\varphi_{gq}(\cdot) \in \mathbb{R}^{n\times m_q}$, and $\varphi_{kl}(\cdot) \in \mathbb{R}^{n\times w_l}$ denote the activation functions. $\varepsilon_f(\cdot) \in \mathbb{R}^{n}$, $\varepsilon_{gq}(\cdot) \in \mathbb{R}^{n\times m_q}$, and $\varepsilon_{kl}(\cdot) \in \mathbb{R}^{n\times w_l}$ are bounded reconstruction errors. The activation functions are picked as the tanh function and satisfy

$$0 \leq \varphi(x) - \varphi(y) \leq \delta(x - y), \tag{14}$$

$\forall x, y \in \mathbb{R}$ and $x \geq y$, $\delta > 0$. Based on (13), the output of the identifier network is written as

$$\dot{\hat{x}} = S\hat{x} + \hat{\omega}_f^{\mathsf{T}}\varphi_f(\hat{x}) + \sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(\hat{x})u_q + \sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(\hat{x})d_l, \tag{15}$$

where $\hat{\omega}_f \in \mathbb{R}^{n\times n}$, $\hat{\omega}_{gq} \in \mathbb{R}^{n\times n}$, and $\hat{\omega}_{kl} \in \mathbb{R}^{n\times n}$ denote the estimations of the corresponding ideal weights. Moreover, the identification error is described as

$$\tilde{x} = x - \hat{x}. \tag{16}$$

Then, the derivative of (16) can be derived as

$$\begin{aligned}\dot{\tilde{x}} &= \dot{x} - \dot{\hat{x}} \\ &= S\tilde{x} + \tilde{\omega}_f^{\mathsf{T}}\varphi_f(\hat{x}) + \omega_f^{\mathsf{T}}\left(\varphi_f(x) - \varphi_f(\hat{x})\right) + \varepsilon_f(x) \\ &\quad + \sum_{q=1}^{N}\Big(\tilde{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(\hat{x}) + \omega_{gq}^{\mathsf{T}}\left(\varphi_{gq}(x) - \varphi_{gq}(\hat{x})\right) \\ &\quad + \varepsilon_{gq}(x)\Big)u_q \\ &\quad + \sum_{l=1}^{M}\Big(\tilde{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(\hat{x}) + \omega_{kl}^{\mathsf{T}}\left(\varphi_{kl}(x) - \varphi_{kl}(\hat{x})\right) \\ &\quad + \varepsilon_{kl}(x))d_l,\end{aligned} \tag{17}$$

where $\tilde{\omega}_f = \omega_f - \hat{\omega}_f$, $\tilde{\omega}_{gq} = \omega_{gq} - \hat{\omega}_{gq}$, and $\tilde{\omega}_{kl} = \omega_{kl} - \hat{\omega}_{kl}$.

Assumption 1 The ideal weights are bounded as

$$\|\omega_f\| \leq \bar{\omega}_f, \ \|\omega_{gq}\| \leq \bar{\omega}_{gq}, \ \|\omega_{kl}\| \leq \bar{\omega}_{kl},$$

where $\bar{\omega}_f$, $\bar{\omega}_{gq}$, and $\bar{\omega}_{kl}$ are positive constants.

Assumption 2 The reconstruction errors ε_f, ε_{gq}, and ε_{kl} are bounded by the identification error function, that is,

$$\varepsilon_f^{\mathsf{T}}\varepsilon_f \leq \gamma\tilde{x}^{\mathsf{T}}\tilde{x}, \ \varepsilon_{gq}^{\mathsf{T}}\varepsilon_{gq} \leq \gamma\tilde{x}^{\mathsf{T}}\tilde{x}, \ \varepsilon_{kl}^{\mathsf{T}}\varepsilon_{kl} \leq \gamma\tilde{x}^{\mathsf{T}}\tilde{x},$$

where γ is a constant.

Theorem 1 *Consider multi-player ZSG (1) with the system dynamics formulated by (13). The identification error $\tilde{x}$ will converge to zero when $t \to \infty$, if the weights $\hat{\omega}_f$, $\hat{\omega}_{gq}$, and $\hat{\omega}_{kl}$ are updated by*

$$\begin{aligned}\dot{\hat{\omega}}_f &= \Lambda_f\varphi_f(\hat{x})\tilde{x}^{\mathsf{T}}, \\ \dot{\hat{\omega}}_{gq} &= \Lambda_{gq}\varphi_{gq}(\hat{x})u_q\tilde{x}^{\mathsf{T}}, q = 1, \ldots, N, \\ \dot{\hat{\omega}}_{kl} &= \Lambda_{kl}\varphi_{kl}(\hat{x})d_l\tilde{x}^{\mathsf{T}}, l = 1, \ldots, M,\end{aligned} \tag{18}$$

where Λ_f, Λ_{gq}, and Λ_{kl} are symmetric positive definite matrices.

Proof Select the Lyapunov function as

$$\begin{aligned}L_3(t) &= \frac{1}{2}\tilde{x}^{\mathsf{T}}\tilde{x} + \frac{1}{2}\mathrm{tr}\left(\tilde{\omega}_f^{\mathsf{T}}\Lambda_f^{-1}\tilde{\omega}_f\right) \\ &\quad + \sum_{q=1}^{N}\frac{1}{2}\mathrm{tr}\left(\tilde{\omega}_{gq}^{\mathsf{T}}\Lambda_{gq}^{-1}\tilde{\omega}_{gq}\right) \\ &\quad + \sum_{l=1}^{M}\frac{1}{2}\mathrm{tr}\left(\tilde{\omega}_{kl}^{\mathsf{T}}\Lambda_{kl}^{-1}\tilde{\omega}_{kl}\right).\end{aligned} \tag{19}$$

Computing the time derivative of $L_3(t)$, one has

$$\begin{aligned}\dot{L}_3(t) &= \tilde{x}^{\mathsf{T}}\dot{\tilde{x}} + \mathrm{tr}\left(\tilde{\omega}_f^{\mathsf{T}}\Lambda_f^{-1}\dot{\tilde{\omega}}_f\right) + \sum_{q=1}^{N}\mathrm{tr}\left(\tilde{\omega}_{gq}^{\mathsf{T}}\Lambda_{gq}^{-1}\dot{\tilde{\omega}}_{gq}\right) \\ &\quad + \sum_{l=1}^{M}\mathrm{tr}\left(\tilde{\omega}_{kl}^{\mathsf{T}}\Lambda_{kl}^{-1}\dot{\tilde{\omega}}_{kl}\right).\end{aligned} \tag{20}$$

Observing (18) and using $-\dot{\hat{\omega}}_f = \dot{\tilde{\omega}}_f$, $-\dot{\hat{\omega}}_{gq} = \dot{\tilde{\omega}}_{gq}$, and $-\dot{\hat{\omega}}_{kl} = \dot{\tilde{\omega}}_{kl}$, we can obtain

$$\begin{aligned}\mathrm{tr}\left(\tilde{\omega}_f^{\mathsf{T}}\Lambda_f^{-1}\dot{\tilde{\omega}}_f\right) &= -\tilde{x}^{\mathsf{T}}\tilde{\omega}_f^{\mathsf{T}}\varphi_f(\hat{x}), \\ \sum_{q=1}^{N}\mathrm{tr}\left(\tilde{\omega}_{gq}^{\mathsf{T}}\Lambda_{gq}^{-1}\dot{\tilde{\omega}}_{gq}\right) &= -\tilde{x}^{\mathsf{T}}\sum_{q=1}^{N}\tilde{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(\hat{x})u_q, \\ \sum_{l=1}^{M}\mathrm{tr}\left(\tilde{\omega}_{kl}^{\mathsf{T}}\Lambda_{kl}^{-1}\dot{\tilde{\omega}}_{kl}\right) &= -\tilde{x}^{\mathsf{T}}\sum_{l=1}^{M}\tilde{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(\hat{x})d_l.\end{aligned} \tag{21}$$

Then, we have

$$\begin{aligned}\dot{L}_3(t) &= \tilde{x}^{\mathsf{T}}S\tilde{x} + \tilde{x}^{\mathsf{T}}\omega_f^{\mathsf{T}}\left(\varphi_f(x) - \varphi_f(\hat{x})\right) + \tilde{x}^{\mathsf{T}}\varepsilon_f(x) \\ &\quad + \tilde{x}^{\mathsf{T}}\sum_{q=1}^{N}\Big(\omega_{gq}^{\mathsf{T}}\left(\varphi_{gq}(x) - \varphi_{gq}(\hat{x})\right)\Big)u_q\end{aligned}$$

$$+\tilde{x}^{\mathsf{T}}\sum_{q=1}^{N}\varepsilon_{gq}(x)u_q$$
$$+\tilde{x}^{\mathsf{T}}\sum_{l=1}^{M}\Big(\omega_{kl}^{\mathsf{T}}\left(\varphi_{kl}(x)-\varphi_{kl}(\hat{x})\right)\Big)d_l$$
$$+\tilde{x}^{\mathsf{T}}\sum_{l=1}^{M}\varepsilon_{kl}(x)d_l. \tag{22}$$

Based on (14), we have

$$\begin{aligned}
\tilde{x}^{\mathsf{T}}\omega_f^{\mathsf{T}}\left(\varphi_f(x)-\varphi_f(\hat{x})\right) &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_f^{\mathsf{T}}\omega_f\tilde{x}+\frac{1}{2}\delta^2\tilde{x}^{\mathsf{T}}\tilde{x},\\
\tilde{x}^{\mathsf{T}}\omega_{gq}^{\mathsf{T}}\left(\varphi_{gq}(x)-\varphi_{gq}(\hat{x})\right) &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_{gq}^{\mathsf{T}}\omega_{gq}\tilde{x}+\frac{1}{2}\delta^2\tilde{x}^{\mathsf{T}}\tilde{x},\\
\tilde{x}^{\mathsf{T}}\omega_{kl}^{\mathsf{T}}\left(\varphi_{kl}(x)-\varphi_{kl}(\hat{x})\right) &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_{kl}^{\mathsf{T}}\omega_{kl}\tilde{x}+\frac{1}{2}\delta^2\tilde{x}^{\mathsf{T}}\tilde{x}.
\end{aligned} \tag{23}$$

Considering Assumption 2, one has

$$\begin{aligned}
\tilde{x}^{\mathsf{T}}\varepsilon_f(x) &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\tilde{x}+\frac{1}{2}\gamma\tilde{x}^{\mathsf{T}}\tilde{x},\\
\tilde{x}^{\mathsf{T}}\varepsilon_{gq}(x) &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\tilde{x}+\frac{1}{2}\gamma\tilde{x}^{\mathsf{T}}\tilde{x},\\
\tilde{x}^{\mathsf{T}}\varepsilon_{kl}(x) &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\tilde{x}+\frac{1}{2}\gamma\tilde{x}^{\mathsf{T}}\tilde{x}.
\end{aligned} \tag{24}$$

Hence, (22) can be reconstructed as

$$\begin{aligned}
&\dot{L}_3(t)\\
&\leq\tilde{x}^{\mathsf{T}}S\tilde{x}+\frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_f^{\mathsf{T}}\omega_f\tilde{x}+\frac{1}{2}\delta^2\tilde{x}^{\mathsf{T}}\tilde{x}+\frac{1}{2}\tilde{x}^{\mathsf{T}}\tilde{x}+\frac{1}{2}\gamma\tilde{x}^{\mathsf{T}}\tilde{x}\\
&\quad+\frac{1}{2}\tilde{x}^{\mathsf{T}}\sum_{q=1}^{N}(u_q\omega_{gq}^{\mathsf{T}}\omega_{gq}\tilde{x})+\frac{1}{2}\delta^2\sum_{q=1}^{N}(u_q\tilde{x}^{\mathsf{T}}\tilde{x})\\
&\quad+\frac{(1+\gamma)}{2}\sum_{q=1}^{N}(u_q\tilde{x}^{\mathsf{T}}\tilde{x})+\frac{1}{2}\tilde{x}^{\mathsf{T}}\sum_{l=1}^{M}(d_l\omega_{kl}^{\mathsf{T}}\omega_{kl}\tilde{x})\\
&\quad+\frac{1}{2}\delta^2\sum_{l=1}^{M}(d_l\tilde{x}^{\mathsf{T}}\tilde{x})+\frac{(1+\gamma)}{2}\sum_{l=1}^{M}(d_l\tilde{x}^{\mathsf{T}}\tilde{x})\\
&=\tilde{x}^{\mathsf{T}}\Gamma\tilde{x},
\end{aligned} \tag{25}$$

where

$$\begin{aligned}
\Gamma = {}& S+\frac{1}{2}\omega_f^{\mathsf{T}}\omega_f+\frac{1}{2}\sum_{q=1}^{N}u_q\omega_{gq}^{\mathsf{T}}\omega_{gq}+\frac{1}{2}\sum_{l=1}^{M}d_l\omega_{kl}^{\mathsf{T}}\omega_{kl}\\
&+\Bigg(\frac{1}{2}+\frac{1}{2}\gamma+\frac{1}{2}\delta^2+\frac{(1+\gamma)}{2}\sum_{q=1}^{N}u_q+\frac{1}{2}\delta^2\sum_{q=1}^{N}u_q\\
&+\frac{(1+\gamma)}{2}\sum_{l=1}^{M}d_l+\frac{1}{2}\delta^2\sum_{l=1}^{M}d_l\Bigg)I_n
\end{aligned} \tag{26}$$

with I_n denoting the identity matrix. If S is reasonably chosen to let $\Gamma \leq 0$, then we have $\dot{L}_3(t) \leq 0$, and $\tilde{x}(t) \to 0$ as $t \to \infty$.

According to Theorem 1, the system dynamics can be removed. Consequently, system (1) is described by

$$\begin{aligned}
\dot{x} = {}& Sx+\hat{\omega}_f^{\mathsf{T}}\varphi_f(x)+\sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)u_q\\
&+\sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x)d_l,
\end{aligned} \tag{27}$$

□

3.2 Approximate optimal learning scheme with single-critic network

For the implementation purpose, only a single-critic network is constructed to deal with the HJI equation. The optimal cost function $J^*(x)$ is expressed as

$$J^*(x) = \omega_c^{\mathsf{T}}\varphi_c(x)+\varepsilon_c(x), \tag{28}$$

where $\omega_c \in \mathbb{R}^{n_c}$ is the ideal weight, $\varphi_c(x) \in \mathbb{R}^{n_c}$ is the activation function, n_c represents the number of neurons, and $\varepsilon_c \in \mathbb{R}$ is the reconstruction error.

The partial derivative of (28) is derived as

$$\nabla J^*(x) = (\nabla\varphi_c(x))^{\mathsf{T}}\omega_c+\nabla\varepsilon_c(x). \tag{29}$$

Then, the approximate formulation of $J^*(x)$ is written as

$$\hat{J}^*(x) = \hat{\omega}_c^{\mathsf{T}}\varphi_c(x), \tag{30}$$

where $\hat{\omega}_c$ is the estimated weight. Similarly, one has

$$\nabla\hat{J}^*(x) = (\nabla\varphi_c(x))^{\mathsf{T}}\hat{\omega}_c. \tag{31}$$

Utilizing the identification result and considering (10), (11), and (29), we have

$$u_q^* = -\beta_q\tanh\left(\frac{1}{2\beta_q}R_q^{-1}(\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq})^{\mathsf{T}}\right.$$

$$\times\left(\nabla\varphi_c^{\mathsf{T}}\omega_c+\nabla\varepsilon_c(x)\right)\Big), \tag{32}$$

$$d_l^* = \frac{1}{2\lambda^2}(\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl})^{\mathsf{T}}\left(\nabla\varphi_c^{\mathsf{T}}\omega_c + \nabla\varepsilon_c(x)\right). \tag{33}$$

In light of (31), the approximate forms of (32) and (33) are stated as

$$\hat{u}_q^* = -\beta_q \tanh\left(\hat{B}\right), \tag{34}$$

$$\hat{d}_l^* = \frac{1}{2\lambda^2}(\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl})^{\mathsf{T}}\nabla\varphi_c^{\mathsf{T}}\hat{\omega}_c, \tag{35}$$

where $\hat{B} = (1/(2\beta_q))R_q^{-1}(\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq})^{\mathsf{T}}\nabla\varphi_c^{\mathsf{T}}\hat{\omega}_c$.

Noticing the identifier-critic framework, the approximate Hamiltonian can be presented as

$$\begin{aligned}&\hat{H}\left(x, \hat{\omega}_c, \hat{u}_q^*, \hat{d}_l^*\right)\\ &= x^{\mathsf{T}}Qx+2\sum_{q=1}^{N}R_q\int_0^{\hat{u}_q^*}\beta_q \tanh^{-\mathsf{T}}(\upsilon/\beta_q)\mathrm{d}\upsilon\\ &\quad-\lambda^2\sum_{l=1}^{M}(\hat{d}_l^*)^{\mathsf{T}}\hat{d}_l^*\\ &\quad+\hat{\omega}_c^{\mathsf{T}}\nabla\varphi_c(x)\Bigg(Sx+\hat{\omega}_f^{\mathsf{T}}\varphi_f(x)+\sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)\hat{u}_q^*\\ &\quad+\sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x)\hat{d}_l^*\Bigg) \triangleq e_c.\end{aligned} \tag{36}$$

Based on the ER approach [42], we define the objective function as

$$E_c = \frac{1}{2}\left(e_c^{\mathsf{T}}e_c + \sum_{p=1}^{Z_P} e^{\mathsf{T}}(t_p)e(t_p)\right), \tag{37}$$

where $e(t_p) = h(x(t_p), \hat{u}_q^*, \hat{d}_l^*) + \hat{\omega}_c^{\mathsf{T}}\phi_p$, $\phi_p = \nabla\varphi_c(x(t_p))(Sx + \hat{\omega}_f^{\mathsf{T}}\varphi_f(x(t_p)) + \sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x(t_p))\hat{u}_q^* + \sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x(t_p))\hat{d}_l^*)$, and $p \in \{1, \ldots, Z_P\}$ is the index of the stored samples.

For the minimizing of the objective function E_c, we construct a novel critic weight tuning law based on gradient descent technique as follows

$$\begin{aligned}\dot{\hat{\omega}}_c &= -\alpha_c\left(\frac{\partial E_c}{\partial\hat{\omega}_c}\right)\\ &= -\alpha_c\phi(\phi^{\mathsf{T}}\hat{\omega}_c + h(x, \hat{u}_q^*, \hat{d}_l^*))\\ &\quad-\alpha_c\sum_{p=1}^{Z_P}\phi_p(\phi_p^{\mathsf{T}}\hat{\omega}_c + h(x(t_p), \hat{u}_q^*, \hat{d}_l^*)),\end{aligned} \tag{38}$$

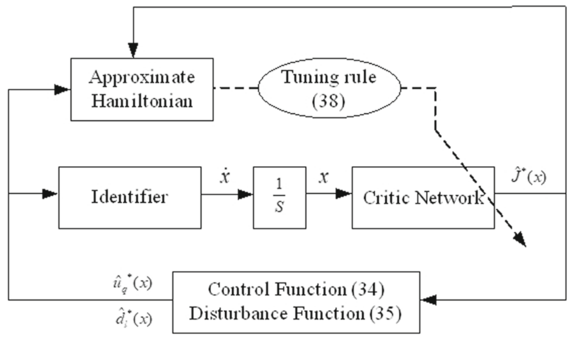

Fig. 1 Structure of the ADP-based optimal control scheme. The solid line represents the signal flow, while the dashed line denotes the neural network back-propagating path

where $\alpha_c > 0$ is the adjustable learning rate of the critic network and $\phi = \nabla\varphi_c(x)(Sx + \hat{\omega}_f^{\mathsf{T}}\varphi_f(x) + \sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)\hat{u}_q^* + \sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x)\hat{d}_l^*)$.

Remark 2 According to [55], the second term in (38) tends to relax the PE condition. Differing from the PE condition, the new condition is convenient to check during the online learning process. That is to say, the ER approach is effortless to implement by using the historical system data.

Remark 3 When using the ER approach, the new condition should be satisfied. Define $\Xi = [\varphi_c(x(t_1)), \ldots, \varphi_c(x(t_{Z_P}))]$ as the historical data matrix. Let Ξ contain numerous linearly independent elements, i.e., rank(Ξ) $=n_c$.

Define the weight estimation error of the critic network as $\tilde{\omega}_c = \omega_c - \hat{\omega}_c$. Then, by taking the time derivative, we have

$$\dot{\tilde{\omega}}_c=-\alpha_c\phi\left(\phi^{\mathsf{T}}\tilde{\omega}_c-\varepsilon_H\right)-\alpha_c\sum_{p=1}^{Z_P}\phi_p\left(\phi_p^{\mathsf{T}}\tilde{\omega}_c-\varepsilon_{H_p}\right), \tag{39}$$

where $\varepsilon_H = -\nabla\varepsilon_c^{\mathsf{T}}(x)(Sx+\hat{\omega}_f^{\mathsf{T}}\varphi_f(x)+\sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)\hat{u}_q^* + \sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x)\hat{d}_l^*)$ and $\varepsilon_{H_p} = -\nabla\varepsilon_c^{\mathsf{T}}(x(t_p))(Sx + \hat{\omega}_f^{\mathsf{T}}\varphi_f(x(t_p)) + \sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x(t_p))\hat{u}_q^* + \sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x(t_p))\hat{d}_l^*)$ are the residual errors.

Based on the above discussion, the structure of the ADP-based optimal control scheme is shown in Fig. 1.

3.3 Stability analysis

In this subsection, the stability analysis of the multi-player ZSG is presented. First, the following assumption, which is used in [42,46], and [50], is provided.

Assumption 3 Denote z_g, z_k, and z_{ω_c} as positive constants. $\hat{\omega}_{gq}$, $\hat{\omega}_{kl}$, and ω_c are upper bounded as $\|\hat{\omega}_{gq}\| \leq z_g$, $\|\hat{\omega}_{kl}\| \leq z_k$, and $\|\omega_c\| \leq z_{\omega_c}$, respectively.

Assumption 4 Denote z_{ε_c}, $z_{\varepsilon_{cd}}$, z_{ε_H}, and $z_{\varepsilon_{H_p}}$ as positive constants. ε_c, $\nabla\varepsilon_c$, ε_H, and ε_{H_p} are upper bounded guaranteeing $\|\varepsilon_c\| \leq z_{\varepsilon_c}$, $\|\nabla\varepsilon_c\| \leq z_{\varepsilon_{cd}}$, $\|\varepsilon_H\| \leq z_{\varepsilon_H}$, and $\|\varepsilon_{H_p}\| \leq z_{\varepsilon_{H_p}}$, respectively.

Assumption 5 Denote z_{φ_c}, $z_{\varphi_{cd}}$, $z_{\varphi_{gq}}$, and $z_{\varphi_{kl}}$ as positive constants. φ_c, $\nabla\varphi_c$, φ_{gq}, and φ_{kl} are upper bounded guaranteeing $\|\varphi_c\| \leq z_{\varphi_c}$, $\|\nabla\varphi_c\| \leq z_{\varphi_{cd}}$, $\|\varphi_{gq}\| \leq z_{\varphi_{gq}}$, and $\|\varphi_{kl}\| \leq z_{\varphi_{kl}}$, respectively.

Theorem 2 *Consider multi-player ZSG (1) with the identifier network, developed control policy (34) and disturbance strategy (35), and single-critic network weight tuning law (38). Then, the UUB stability of the controlled system state and the critic weight estimation error is ensured.*

Proof Select the Lyapunov function as

$$L(t) = L_1(t) + L_2(t) = J^*(x) + \frac{1}{2}\tilde{\omega}_c^{\mathsf{T}}\tilde{\omega}_c. \tag{40}$$

Calculating the time derivative of $L_1(t)$ and using reconstructed system (27), one has

$$\dot{L}_1(t) = \left(\nabla J^*(x)\right)^{\mathsf{T}} \Big(Sx + \hat{\omega}_f^{\mathsf{T}}\varphi_f(x) + \sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)\hat{u}_q^* + \sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x)\hat{d}_l^*\Big). \tag{41}$$

Let

$$\varpi(u_q) = 2\sum_{q=1}^{N} R_q \int_0^{u_q} \beta_q \tanh^{-\mathsf{T}}(v/\beta_q)\mathrm{d}v. \tag{42}$$

According to [55], putting (10) in (41), one has

$$\varpi\left(u_q^*\right) = \beta_q(\nabla J^*)^{\mathsf{T}} g_q(x)\tanh\left(B^*\right) + \lambda^2\bar{R}\sum_{l=1}^{m_q}\ln\left(\bar{\mathbf{1}} - \tanh^2\left(B_l^*\right)\right), \tag{43}$$

where $B^* = [B_1^*, B_2^*, \ldots, B_{m_q}^*]^{\mathsf{T}}$ with $B_l^* \in \mathbb{R}$, $l = 1, 2, \ldots, m_q$. $\bar{\mathbf{1}}$ is a column vector having all of its elements equal to 1, and $\bar{R} = [r_1, \ldots, r_{m_q}] \in \mathbb{R}^{1\times m_q}$.

From (10)–(12) and (43), we obtain

$$\begin{aligned}(\nabla J^*(x))^{\mathsf{T}}&(Sx + \hat{\omega}_f^{\mathsf{T}}\varphi_f(x)) \\ &= -x^{\mathsf{T}}Qx - \varpi\left(u_q^*\right) - (\nabla J^*(x))^{\mathsf{T}}\sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)u_q^* \\ &\quad - \lambda^2\sum_{l=1}^{M}(d_l^*)^{\mathsf{T}}d_l^*,\end{aligned} \tag{44}$$

$$(\nabla J^*(x))^{\mathsf{T}}\sum_{l=1}^{M}\hat{\omega}_{kl}^{\mathsf{T}}\varphi_{kl}(x)\hat{d}_l^* = 2\lambda^2\sum_{l=1}^{M}(d_l^*)^{\mathsf{T}}\hat{d}_l^*. \tag{45}$$

Thus, (41) becomes

$$\begin{aligned}\dot{L}_1(t) &= -x^{\mathsf{T}}Qx - \varpi\left(u_q^*\right) \\ &\quad - \left(\nabla J^*(x)\right)^{\mathsf{T}}\sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)(u_q^* - \hat{u}_q^*) \\ &\quad - \lambda^2\sum_{l=1}^{M}(d_l^*)^{\mathsf{T}}(d_l^* - 2\hat{d}_l^*) \\ &= -x^{\mathsf{T}}Qx - \varpi\left(u_q^*\right) \\ &\quad + \beta_q\left(\nabla J^*(x)\right)^{\mathsf{T}}\sum_{q=1}^{N}(\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x)(\tanh(B^*) \\ &\quad - \tanh(\hat{B}))) - \lambda^2\sum_{l=1}^{M}(d_l^* - \hat{d}_l^*)^{\mathsf{T}}(d_l^* - \hat{d}_l^*) \\ &\quad + \lambda^2\sum_{l=1}^{M}(\hat{d}_l^*)^{\mathsf{T}}(\hat{d}_l^*).\end{aligned} \tag{46}$$

Then, utilizing (29), Assumption 3–5, and the fact that $\varpi(u_q^*)$ is positive definite [55], (46) can be rewritten as

$$\dot{L}_1(t) \leq -x^{\mathsf{T}}Qx + 2\beta_q(\omega_c^{\mathsf{T}}\nabla\varphi_c + \nabla\varepsilon_c^{\mathsf{T}})\sum_{q=1}^{N}\hat{\omega}_{gq}^{\mathsf{T}}\varphi_{gq}(x) - \lambda^2\sum_{l=1}^{M}\|d_l^* - \hat{d}_l^*\|^2 + \lambda^2\sum_{l=1}^{M}\|\hat{d}_l^*\|^2$$

$$\leq -x^{\mathsf{T}}Qx + 2\beta_q z_{\omega_c} z_{\varphi_{cd}} \sum_{q=1}^{N} z_g z_{\varphi_{gq}} + 2\beta_q z_{\varepsilon_{cd}} \sum_{q=1}^{N} z_g z_{\varphi_{gq}} + \frac{1}{4\lambda^2} z_{\varphi_{cd}}^2 \sum_{l=1}^{M} z_k^2 z_{\varphi_{kl}}^2 \left\|\hat{\omega}_c\right\|^2. \tag{47}$$

Recalling $\tilde{\omega}_c = \omega_c - \hat{\omega}_c$, we further get that

$$\begin{aligned}\dot{L}_1(t) &\leq -x^{\mathsf{T}}Qx + b_2 - 2b_1\omega_c\tilde{\omega}_c + b_1\left\|\tilde{\omega}_c\right\|^2 \\ &\leq -\lambda_{\min}(Q)\|x\|^2 - b_3\|\tilde{\omega}_c\|^2 + b_2,\end{aligned} \tag{48}$$

where $b_1 = (1/(4\lambda^2))z_{\varphi_{cd}}^2 \sum_{l=1}^{M} z_k^2 z_{\varphi_{kl}}^2$, $b_2 = 2\beta_q z_{\omega_c} z_{\varphi_{cd}} \sum_{q=1}^{N} z_g z_{\varphi_{gq}} + 2\beta_q z_{\varepsilon_{cd}} \sum_{q=1}^{N} z_g z_{\varphi_{gq}} + b_1 z_{\omega_c}^2$, and $b_3 = (1/2)b_1^2 - b_1$.

Next, considering (39), the derivative of $L_2(t)$ is formulated as

$$\dot{L}_2(t) = -\alpha_c\tilde{\omega}_c^{\mathsf{T}}\phi\phi^{\mathsf{T}}\tilde{\omega}_c - \alpha_c\sum_{p=1}^{Z_P}\tilde{\omega}_c^{\mathsf{T}}\phi_p\phi_p^{\mathsf{T}}\tilde{\omega}_c + \alpha_c\tilde{\omega}_c^{\mathsf{T}}\phi\varepsilon_H + \alpha_c\sum_{p=1}^{Z_P}\tilde{\omega}_c^{\mathsf{T}}\phi_p\varepsilon_{H_p}. \tag{49}$$

With the aid of the Young's inequality, we can derive the last two terms of (49) as follows:

$$\alpha_c\tilde{\omega}_c^{\mathsf{T}}\phi\varepsilon_H \leq \frac{\alpha_c}{2}\tilde{\omega}_c^{\mathsf{T}}\phi\phi^{\mathsf{T}}\tilde{\omega}_c + \frac{\alpha_c}{2}\varepsilon_H^{\mathsf{T}}\varepsilon_H, \tag{50}$$

$$\alpha_c\sum_{p=1}^{Z_P}\tilde{\omega}_c^{\mathsf{T}}\phi_p\varepsilon_{H_p} \leq \frac{\alpha_c}{2}\sum_{p=1}^{Z_P}\tilde{\omega}_c^{\mathsf{T}}\phi_p\phi_p^{\mathsf{T}}\tilde{\omega}_c + \frac{\alpha_c}{2}\sum_{p=1}^{Z_P}\varepsilon_{H_p}^{\mathsf{T}}\varepsilon_{H_p}. \tag{51}$$

Applying Assumption 3–5 and considering (50) and (51), (49) becomes

$$\dot{L}_2(t) \leq -\frac{\alpha_c}{2}\lambda_{\min}(\Phi(\phi,\phi_p))\|\tilde{\omega}_c\|^2 + \frac{\alpha_c(Z_P+1)}{2}z_{\varepsilon_H}^2, \tag{52}$$

where $\Phi(\phi,\phi_p) = \phi\phi^{\mathsf{T}} + \sum_{p=1}^{Z_P}\phi_p\phi_p^{\mathsf{T}}$.

Combining (48) with (52), one has

$$\dot{L}(t) \leq -\lambda_{\min}(Q)\|x\|^2 - b_3\|\tilde{\omega}_c\|^2 + b_2 - \frac{\alpha_c}{2}\lambda_{\min}(\Phi(\phi,\phi_p))\|\tilde{\omega}_c\|^2 + \frac{\alpha_c(Z_P+1)}{2}z_{\varepsilon_H}^2. \tag{53}$$

Therefore, (53) means $\dot{L}(t) < 0$, whenever the following inequalities hold

$$\|x\| > \sqrt{\frac{2b_2 + \alpha_c(Z_P+1)\varepsilon_H^2}{2\lambda_{\min}(Q)}} \triangleq \mathscr{D}_1 \tag{54}$$

or

$$\|\tilde{\omega}_c\| > \sqrt{\frac{2b_2 + \alpha_c(Z_P+1)\varepsilon_H^2}{2b_3 + \alpha_c\lambda_{\min}(\Phi(\phi,\phi_p))}} \triangleq \mathscr{D}_2 \tag{55}$$

with $2b_3 + \alpha_c\lambda_{\min}(\Phi(\phi,\phi_p)) > 0$. It implies that the UUB stability of x and $\tilde{\omega}_c$ is guaranteed. □

4 Simulation

In this section, we deliver a simulation of a multi-player ZSG with constrained control inputs to demonstrate the effectiveness of the established ADP-based identifier-critic framework.

Consider the multi-player ZSG described as (note: $N = 2$, $M = 1$)

$$\dot{x} = f(x) + g_1(x)u_1 + g_2(x)u_2 + k(x)d, \tag{56}$$

where

$$f(x) = \begin{bmatrix} -0.5x_1 + 0.4x_2 \\ -0.6x_1 - 0.6x_2 + 0.5x_2x_1^2 \end{bmatrix},$$

$$g_1(x) = \begin{bmatrix} 0 \\ \sin(x_1) \end{bmatrix}, g_2(x) = \begin{bmatrix} 0 \\ 2x_1 \end{bmatrix}, k(x) = \begin{bmatrix} 0 \\ x_1 \end{bmatrix}.$$

The system state $x = [x_1, x_2]^{\mathsf{T}} \in \mathbb{R}^2$ is initialized to $x_0 = [0.8, -0.8]^{\mathsf{T}}$, and $u_1, u_2 \in \mathbb{R}$ are the constrained control inputs. Let $Q = 5I_2$, $R_1 = R_2 = I$, and $\lambda = 2$. In this case, we assume the control inputs u_1 and u_2 are constrained by $|u_1| \leq 0.4$ and $|u_2| \leq 0.8$, respectively. Then $\varpi(u_1)$ and $\varpi(u_2)$ defined in the utility function are

$$\varpi(u_1) = 2R_1\int_0^{u_1}(0.4\tanh^{-1}(v/0.4))^{\mathsf{T}}\mathrm{d}v,$$

$$\varpi(u_2) = 2R_2\int_0^{u_2}(0.8\tanh^{-1}(v/0.8))^{\mathsf{T}}\mathrm{d}v,$$

respectively.

Aiming at study the unknown dynamics of (56), an identifier network is built to reconstruct system dynamics based on (15). In the system identification stage, the

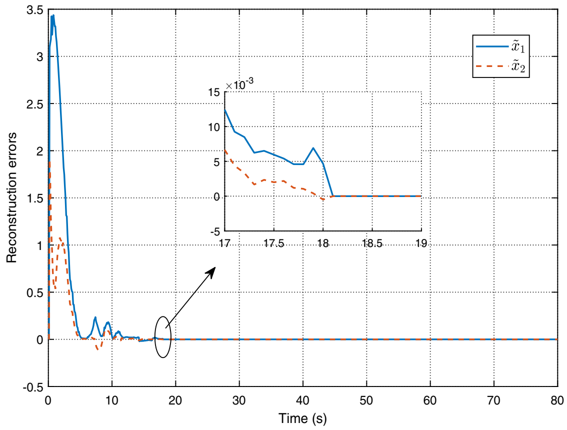

Fig. 2 Curves of reconstruction errors

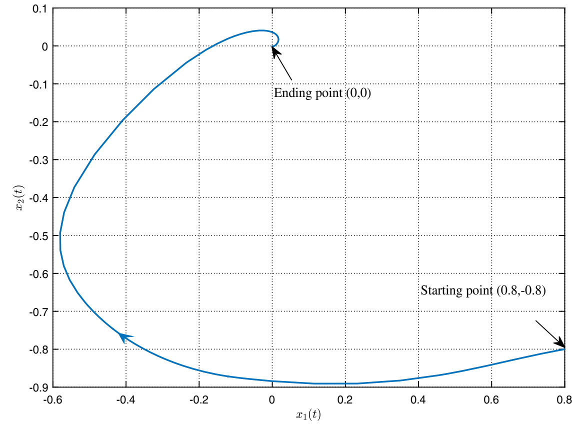

Fig. 3 State trajectories

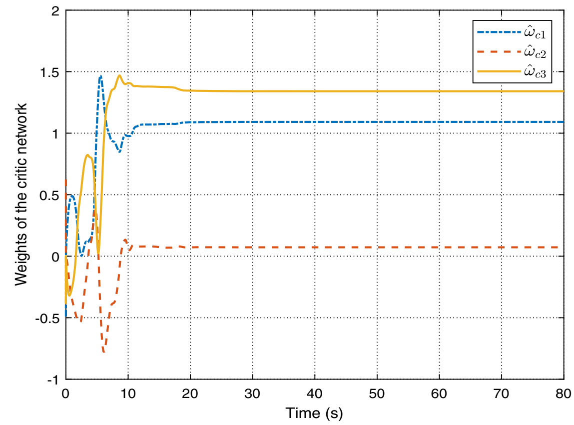

Fig. 4 Convergence curves of $\hat{\omega}_c$

initial weights $\hat{\omega}_f$, $\hat{\omega}_{gq}$, and $\hat{\omega}_{kl}$ are chosen randomly as $\hat{\omega}_f \in [-1, 1]$, $\hat{\omega}_{gq} \in [-1, 1]$, and $\hat{\omega}_{kl} \in [-1, 1]$. The identifier activation function $\varphi_f(\cdot)$, $\varphi_{gq}(\cdot)$, and $\varphi_{kl}(\cdot)$ are selected as $\varphi_f(\cdot) = \varphi_{gq}(\cdot) = \varphi_{kl}(\cdot) = \tanh(\cdot)$, and the learning matrix $S = [-1, 0; 0, -1]$. The other corresponding parameters are designed as $\Lambda_f = \Lambda_{gq} = \Lambda_{kl} = [1, 0.4; 0.1, 0.6]$.

For the proposed ADP-based approach, we select the activation function as $\varphi_c(x) = [x_1^2, x_1x_2, x_2^2]^{\mathsf{T}}$. The approximate critic network weight is $\hat{\omega}_c = [\hat{\omega}_{c1}, \hat{\omega}_{c2}, \hat{\omega}_{c3}]^{\mathsf{T}}$. The initial weight are randomly selected as $\hat{\omega}_c \in [-1, 1]$.

We employ the ER method with recorded data to relax the PE condition. The number of the historical data samples for the critic network is selected as 12, i.e., $Z_P = 12$. Then, the critic network learning scheme is established for 80 s with the novel critic network weight tuning law, which combines the ER technique with the standard gradient descent algorithm.

Simulation results are depicted in Figs. 2, 3, 4, 5, 6, and 7. The convergence curves of reconstruction errors of the neural network-based identifier are depicted in Fig. 2. As displayed in Fig. 2, reconstruction errors converge to a small region of origin around $t = 20$ s. It illustrates that the identifier network can well reconstruct system (56). Figure 3 shows the convergence process of system states for the ZSG with control constraints. In Fig. 3, it can be observed that system states finally converge to the equilibrium point $(0, 0)$. The convergence process of the critic network weights is displayed in Fig. 4. From Fig. 4, we can see that the critic network weights have stabilized after $t = 20$ s and their values finally converge to $\hat{\omega}_c = [1.0912, 0.0725, 1.3409]^{\mathsf{T}}$. To demonstrate the effectiveness of the proposed ADP-based learning approach, we apply the method in [28] to system (56). Then, the convergence process of the critic network weights under the method in [28] is shown in Fig. 5. By comparing Figs. 4 and 5, it is obvious seen that the developed algorithm in this paper can accelerate the convergence rate of the critic network weights. Then, the converged weights are inserted into (34) and (35) to get the approximate optimal control strategies $\{\hat{u}_1^*, \hat{u}_2^*\}$ and the approximate worst disturbance strategy $\hat{d}^*$ for nonlinear ZSG (56).

Figure 6 shows the trajectory of constrained control inputs in the control process. As illustrated in Fig. 6, it can be seen that the constrained control inputs u_1 and u_2 are effectively limited by the predetermined bound $|u_q| \leq \beta_q$ $(q = 1, 2)$ as expected, which indicates that

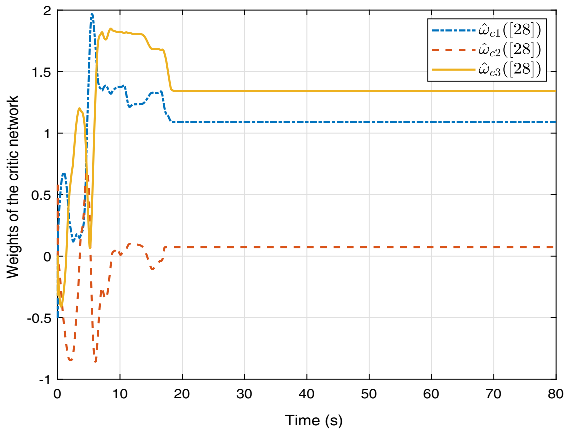

Fig. 5 Convergence curves of $\hat{\omega}_c$ under the method in [28]

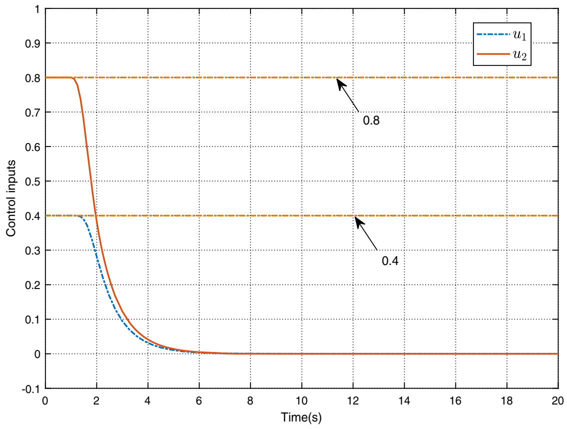

Fig. 6 Trajectories of constrained control inputs

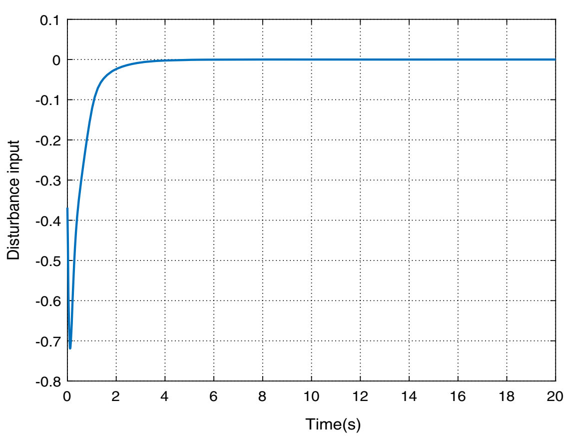

Fig. 7 Trajectory of the unconstrained disturbance input

control input signals vary within the control constraints. It proves the effectiveness of the constrained policy. Figure 7 presents the trajectory of the unconstrained disturbance input in the control process. Obviously, it is proved that the disturbance input converges to an adjustable neighborhood of the zero. Therefore, the aforementioned simulation results confirm the effectiveness of the designed ADP-based scheme with the identifier-critic form. Meanwhile, it also shows that the identifier-critic framework is applicable to the nonlinear multi-player ZSG with constrained control inputs.

5 Conclusion

In this article, the multi-player ZSG issue with unknown system dynamics and control constraints is handled by employing a novel ADP-based learning framework. Initially, the neural network-based identifier is adopted to rebuild the system dynamics by utilizing the system data. Then, we define a new non-quadratic function which addresses the control constraints and obtain the constrained HJI equation. Furthermore, a single-critic network mechanism is designed to approximately solve the constrained HJI equation. Subsequently, the novel weight tuning rule base on the ER algorithm is constructed to approach the optimal control strategies and the worst disturbance strategies. Hence, the traditional PE condition is removed via the recorded and current data. Additionally, the UUB stability of the multi-player system and the critic network weight approximation error is analyzed. After that, we demonstrate the convergence and performance of the proposed scheme through simulation studies. However, the limitation of the proposed scheme is the reconstruction error which inevitably introduced by using an identifier. In the consecutive study, how to relax the requirement of system dynamics without reconstruction errors may be investigated.

Funding This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0112302; and in part by the National Natural Science Foundation of China under Grant 62222301, Grant 61890930-5, and Grant 62021003.

Data availability statement No data were used in this paper.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

Ethical approval No conflict of interest exits in this submission, and the research work does not involve any human participants and/or animals. The manuscript is approved by all authors for publication.

References

1. Denardo, E.V.: Introduction to Game Theory. Springer, Boston (2011)
2. Vamvoudakis, K.G., Modares, H., Kiumarsi, B., Lewis, F.L.: Game theory-based control system algorithms with real-time reinforcement learning: how to solve multiplayer games online. IEEE Control Syst. Mag. **37**(1), 33–52 (2017)
3. Ni, Z., Paul, S.: A multistage game in smart grid security: a reinforcement learning solution. IEEE Trans. Neural Netw. Learn. Syst. **30**(9), 2684–2695 (2019)
4. Bidram, A., Davoudi, A., Lewis, F.L., Guerrero, J.M.: Distributed cooperative secondary control of microgrids using feedback linearization. IEEE Trans. Power Syst. **28**(3), 3462–3470 (2013)
5. Wei, Q., Li, H., Yang, X., He, H.: Continuous-time distributed policy iteration for multi-controller nonlinear systems. IEEE Trans. Cybern. **51**(5), 2372–2383 (2021)
6. Liu, D., Li, H., Wang, D.: Online synchronous approximate optimal learning algorithm for multiplayer nonzero-sum games with unknown dynamics. IEEE Trans. Syst. Man Cybern. Syst. **44**(8), 1015–1027 (2014)
7. Li, Y., Wei, C., An, T., Ma, B., Dong, B.: Event-triggered-based cooperative game optimal tracking control for modular robot manipulator with constrained input. Nonlinear Dyn. **109**(4), 2759–2779 (2022)
8. Modares, H., Lewis, F.L., Sistani, M.B.N.: Online solution of nonquadratic two-player zero-sum games arising in the H_∞ control of constrained input systems. Int. J. Adapt. Control Signal Process. **28**(3), 232–254 (2014)
9. Vamvoudakis, K.G.: Non-zero sum Nash Q-learning for unknown deterministic continuous-time linear systems. Automatica **61**, 274–281 (2015)
10. Wang, D., Ha, M., Zhao, M.: The intelligent critic framework for advanced optimal control. Artif. Intell. Rev. **55**(1), 1–22 (2022)
11. Ha, M., Wang, D., Liu, D.: Discounted iterative adaptive critic designs with novel stability analysis for tracking control. IEEE/CAA J. Automatica Sinica **9**(7), 1262–1272 (2022)
12. Li, Y., Liu, Y., Tong, S.: Observer-based neuro-adaptive optimized control of strict-feedback nonlinear systems with state constraints. IEEE Trans. Neural Netw. Learn. Syst. **33**(7), 3131–3145 (2022)
13. Wang, H., Yang, C., Liu, X., Zhou, L.: Neural-network-based adaptive control of uncertain MIMO singularly perturbed systems with full-state constraints. IEEE Trans. Neural Netw. Learn. Syst. (2021). https://doi.org/10.1109/TNNLS.2021.3123361
14. Huo, Y., Wang, D., Qiao, J.: Adaptive critic optimization to decentralized event-triggered control of continuous-time nonlinear interconnected systems. Opt. Control Appl. Methods **43**(1), 198–212 (2022)
15. Lv, Y., Na, J., Zhao, X., Huang, Y., Ren, X.: Multi-H_∞ controls for unknown input-interference nonlinear system with reinforcement learning. IEEE Trans. Neural Netw. Learn. Syst. (2021). https://doi.org/10.1109/TNNLS.2021.3130092
16. Wei, Q., Liu, D., Lin, Q., Song, R.: Adaptive dynamic programming for discrete-time zero-sum games. IEEE Trans. Neural Netw. Learn. Syst. **29**(4), 957–969 (2018)
17. Dong, B., An, T., Zhou, F., Liu, K., Li, Y.: Decentralized robust zero-sum neuro-optimal control for modular robot manipulators in contact with uncertain environments: theory and experimental verification. Nonlinear Dyn. **97**(1), 503–524 (2019)
18. Wu, H., Liu, Z.: Data-driven guaranteed cost control design via reinforcement learning for linear systems with parameter uncertainties. IEEE Trans. Syst. Man, Cybern. Syst. **50**(11), 4151–4159 (2020)
19. Song, R., Lewis, F.L., Wei, Q.: Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games. IEEE Trans. Neural Netw. Learn. Syst. **28**(3), 704–713 (2017)
20. Zhao, Q., Sun, J., Wang, G., Chen, J.: Event-triggered ADP for nonzero-sum games of unknown nonlinear systems. IEEE Trans. Neural Netw. Learn. Syst. **33**(5), 1905–1913 (2022)
21. Wei, Q., Zhu, L., Song, R., Zhang, P., Liu, D., Xiao, J.: Model-free adaptive optimal control for unknown nonlinear multiplayer nonzero-sum game. IEEE Trans. Neural Netw. Learn. Syst. **33**(2), 879–892 (2022)
22. Yang, X., He, H.: Event-driven H_∞ constrained control using adaptive critic learning. IEEE Trans. Cybern **51**(10), 4860–4872 (2021)
23. Zhao, J., Lv, Y., Zhao, J.: Adaptive learning based output-feedback optimal control of CT two-player zero-sum games. IEEE Trans. Circuits Syst.-II: Express Briefs **69**(3), 1437–1441 (2022)
24. Yazidi, A., Silvestre, D., Oommen, B.J.: Solving two-person zero-sum stochastic games with incomplete information using learning automata with artificial barriers. IEEE Trans. Neural Netw. Learn. Syst. (2021). https://doi.org/10.1109/TNNLS.2021.3099095
25. Guo, X., Yan, W., Cui, R.: Reinforcement learning-based nearly optimal control for constrained-input partially unknown systems using differentiator. IEEE Trans. Neural Netw. Learn. Syst. **31**(11), 4713–4725 (2020)
26. Song, R., Li, J., Lewis, F.L.: Robust optimal control for disturbed nonlinear zero-sum differential games based on single NN and least squares. IEEE Trans. Syst. Man, Cybern. Syst. **50**(11), 4009–4019 (2020)
27. Song, R., Wei, Q., Song, B.: Neural-network-based synchronous iteration learning method for multi-player zero-sum games. Neurocomputing **242**(14), 73–82 (2017)
28. Zhang, Y., Zhao, B., Liu, D.: Event-triggered adaptive dynamic programming for multi-player zero-sum games with unknown dynamics. Soft. Comput. **25**, 2237–2251 (2021)
29. Qiao, J., Li, M., Wang, D.: Asymmetric constrained optimal tracking control with critic learning of nonlinear multiplayer zero-sum games. IEEE Trans. Neural Netw. Learn. Syst. (2022). https://doi.org/10.1109/TNNLS.2022.3208611

30. Wei, Q., Song, R., Yan, P.: Data-driven zero-sum neuro-optimal control for a class of continuous-time unknown nonlinear systems with disturbance using ADP. IEEE Trans. Neural Netw. Learn. Syst. **27**(2), 444–458 (2016)
31. Yang, X., Zhao, B.: Optimal neuro-control strategy for nonlinear systems with asymmetric input constraints. IEEE/CAA J. Automatica Sinica **7**(2), 575–583 (2020)
32. Yang, Y., Ding, Z., Wang, R., Modares, H., Wunsch, D.C.: Data-driven human-robot interaction without velocity measurement using off-policy reinforcement learning. IEEE/CAA J. Autom. Sinica **9**(1), 47–63 (2022)
33. Na, J., Lv, Y., Zhang, K., Zhao, J.: Adaptive identifier-critic-based optimal tracking control for nonlinear systems with experimental validation. IEEE Trans. Syst. Man, Cybern. Syst. **52**(1), 459–472 (2022)
34. Xue, S., Luo, B., Liu, D.: Event-triggered adaptive dynamic programming for zero-sum game of partially unknown continuous-time nonlinear systems. IEEE Trans. Syst. Man, Cybern. Syst. **50**(9), 3189–3199 (2020)
35. Wang, D.: Intelligent critic control with robustness guarantee of disturbed nonlinear plants. IEEE Trans. Cybern. **50**(6), 2740–2748 (2020)
36. Huo, X., Karimi, H.R., Zhao, X., Wang, B., Zong, G.: Adaptive-critic design for decentralized event-triggered control of constrained nonlinear interconnected systems within an identifier-critic framework. IEEE Trans. Cybern. **52**(8), 7478–7491 (2022)
37. Zhao, D., Zhang, Q., Wang, D., Zhu, Y.: Experience replay for optimal control of nonzero-sum game systems with unknown dynamics. IEEE Trans. Cybern. **46**(3), 854–865 (2016)
38. Xue, S., Luo, B., Liu, D., Yang, Y.: Constrained event-triggered H_∞ control based on adaptive dynamic programming with concurrent learning. IEEE Trans. Syst. Man, Cybern. Syst. **52**(1), 357–369 (2022)
39. Xu, Y., Li, T., Bai, W., Shan, Q., Yuan, L., Wu, Y.: Online event-triggered optimal control for multi-agent systems using simplified ADP and experience replay technique. Nonlinear Dyn. **106**(1), 509–522 (2021)
40. Kamalapurkar, R., Reish, B., Chowdhary, G., Dixon, W.E.: Concurrent learning for parameter estimation using dynamic state-derivative estimators. IEEE Trans. Autom. Control **62**(7), 3594–3601 (2017)
41. Zhang, Q., Zhao, D.: Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics. IEEE Trans. Cybern. **49**(8), 2874–2885 (2019)
42. Yang, X., He, H.: Adaptive critic learning and experience replay for decentralized event-triggered control of nonlinear interconnected systems. IEEE Trans. Syst. Man, Cybern. Syst. **50**(11), 4043–4055 (2020)
43. Zhu, Y., Zhao, D., He, H., Ji, J.: Event-triggered optimal control for partially unknown constrained-input systems via adaptive dynamic programming. IEEE Trans. Industr. Electron. **64**(5), 4101–4109 (2017)
44. Luo, B., Yang, Y., Liu, D.: Adaptive Q-learning for data-based optimal output regulation with experience replay. IEEE Trans. Cybern. **48**(12), 3337–3348 (2018)
45. Xia, L., Li, Q., Song, R., Modares, H.: Optimal synchronization control of heterogeneous asymmetric input-constrained unknown nonlinear MASs via reinforcement learning. IEEE/CAA J. Autom. Sinica **9**(3), 520–532 (2022)
46. Zhao, B., Liu, D., Luo, C.: Reinforcement learning-based optimal stabilization for unknown nonlinear systems subject to inputs with uncertain constraints. IEEE Trans. Neural Netw. Learn. Syst. **31**(10), 4330–4340 (2020)
47. Zhao, S., Wang, J.: Robust optimal control for constrained uncertain switched systems subjected to input saturation: The adaptive event-triggered case. Nonlinear Dyn. **110**(1), 363–380 (2022)
48. Mishra, A., Ghosh, S.: Variable gain gradient descent-based reinforcement learning for robust optimal tracking control of uncertain nonlinear system with input constraints. Nonlinear Dyn. **107**(3), 2195–2214 (2022)
49. Yang, X., Zhou, Y., Dong, N., Wei, Q.: Adaptive critics for decentralized stabilization of constrained-input nonlinear interconnected systems. IEEE Trans. Syst. Man, Cybern. Syst. **52**(7), 4187–4199 (2022)
50. Mu, C., Wang, K., Sun, C.: Policy-iteration-based learning for nonlinear player game systems with constrained inputs. IEEE Trans. Syst. Man, Cybern. Syst. **51**(10), 6488–6502 (2021)
51. Zhang, S., Zhao, B., Liu, D., Zhang, Y.: Observer-based event-triggered control for zero-sum games of input constrained multi-player nonlinear systems. Neural Netw. **114**(8), 101–112 (2021)
52. Sun, J., Liu, C.: Distributed zero-sum differential game for multi-agent systems in strict-feedback form with input saturation and output constraint. Neural Netw. **106**, 8–19 (2018)
53. Zhu, Y., Zhao, D., Li, X.: Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data. IEEE Trans. Neural Netw. Learn. Syst. **28**(3), 714–725 (2017)
54. Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K.G., Lewis, F.L., Dixon, W.E.: A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. Automatica **49**, 82–92 (2013)
55. Yasini, S., Sitani, M.B.N., Kirampor, A.: Reinforcement learning and neural networks for multi-agent nonzero-sum games of nonlinear constrained-input systems. Int. J. Mach. Learn. Cybern. **7**, 967–980 (2016)