Recent Progress in Reinforcement Learning and Adaptive Dynamic Programming for Advanced Control Applications

Ding Wang, *Senior Member, IEEE*, Ning Gao, Derong Liu, *Fellow, IEEE*,
Jinna Li, *Senior Member, IEEE*, and Frank L. Lewis, *Life Fellow, IEEE*

***Abstract*—Reinforcement learning (RL) has roots in dynamic programming and it is called adaptive/approximate dynamic programming (ADP) within the control community. This paper reviews recent developments in ADP along with RL and its applications to various advanced control fields. First, the background of the development of ADP is described, emphasizing the significance of regulation and tracking control problems. Some effective offline and online algorithms for ADP/adaptive critic control are displayed, where the main results towards discrete-time systems and continuous-time systems are surveyed, respectively. Then, the research progress on adaptive critic control based on the event-triggered framework and under uncertain environment is discussed, respectively, where event-based design, robust stabilization, and game design are reviewed. Moreover, the extensions of ADP for addressing control problems under complex environment attract enormous attention. The ADP architecture is revisited under the perspective of data-driven and RL frameworks, showing how they promote ADP formulation significantly. Finally, several typical control applications with respect to RL and ADP are summarized, particularly in the fields of wastewater treatment processes and power systems, followed by some general prospects for future research. Overall, the comprehensive survey on ADP and RL for advanced control applications has demonstrated its remarkable potential within the artificial intelligence era. In addition, it also plays a vital role in promoting environmental protection and industrial intelligence.**

***Index Terms*—Adaptive dynamic programming (ADP), advanced control, complex environment, data-driven control, event-triggered design, intelligent control, neural networks, nonlinear systems, optimal control, reinforcement learning (RL).**

I. Introduction

ARTIFICIAL intelligence (AI) generally refers to the intelligence exhibited through machines that humans make. The definition of AI is very broad, starting from the legends of robots and androids in Greek mythology to the well-known Turing Test, and now to the development of various intelligent algorithms in the framework of machine learning [1]–[4]. The AI technology is gradually changing our lives, from computer vision, big data processing, intelligent automation, smart factories, etc., to other aspects.

As the hottest technology in the 21st century, AI cannot be developed without machine learning. Machine learning, as the core of AI, is the foundation of computer intelligence. Reinforcement learning (RL) [5]–[7] is one of the top three approaches of machine learning, along with supervised learning and unsupervised learning. RL emphasizes the interaction between the environment and the agent, with a focus on the long-term interaction to change its policies. Through its interactions with the environment, the agent can modify future actions or control policies based on the response to its stimulating actions.

It should be emphasized that RL does not necessarily require a perfect environment model or huge computing resources. RL is inseparable from dynamic programming [8], [9]. Traditional dynamic programming has been investigated considerably in theory, which can provide the key foundation of RL. However, this technique requires the assumption of an exact system model, which is extravagant for large-scale complex nonlinear systems. Besides, such methods are severely limited in solving Hamilton-Jacobi-Bellman (HJB) equations of nonlinear systems as the dimensionality of states and controls increase [8]. Therefore, adaptive/approximate dynamic programming (ADP) [10]–[13], a method combining RL, dynamic programming, and neural networks, was skillfully proposed.

ADP has been widely used to solve a range of optimal con-

Manuscript received June 29, 2023; accepted August 19, 2023. This work was supported in part by the National Natural Science Foundation of China (62222301, 62073085, 62073158, 61890930-5, 62021003), the National Key Research and Development Program of China (2021ZD0112302, 2021ZD 0112301, 2018YFC1900800-5), and Beijing Natural Science Foundation (JQ19 013). Recommended by Associate Editor Zidong Wang. *(Corresponding author: Derong Liu.)*

Citation: D. Wang, N. Gao, D. Liu, J. Li, and F. Lewis, "Recent progress in reinforcement learning and adaptive dynamic programming for advanced control applications," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 1, pp. 18–36, Jan. 2024.

D. Wang and N. Gao are with the Faculty of Information Technology, Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing Laboratory of Smart Environmental Protection, and Beijing Institute of Artificial Intelligence, Beijing University of Technology, Beijing 100124, China (e-mail: dingwang@bjut.edu.cn; gaon@emails.bjut.edu.cn).

D. Liu is with the School of System Design and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055, China, and also with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago IL 60607 USA (e-mail: liudr@sustech.edu.cn, derong@uic.edu).

J. Li is with the School of Information and Control Engineering, Liaoning Petrochemical University, Fushun 113001, China (e-mail: lijinna721@126.com).

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

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

Digital Object Identifier 10.1109/JAS.2023.123843

trol problems for complex nonlinear systems in unknown environments. As main algorithmic frameworks, value iteration (VI) and policy iteration (PI) have been intensively promoted. The initialization requirements for VI and PI are different. Unlike PI, which must start with an initial admissible control law, the initial control law of VI has no strict requirements. But from the iterative control point of view, PI presents a more stable mechanism. Both of these two algorithms are attracting more and more attention from the control community [14], [15]. Due to their respective properties, VI has received more attention in the discrete-time domain, while PI is more commonly applied to the continuous-time domain.

There have been many classical reviews and monographs [13], [16]–[19] that summarize and discuss ADP/RL. They have brought in the profound influence and inspiration on their successors. However, it is rare to find a paper that integrates the regulator problem, the tracking control problem, the multi-agent problem, the robustness of uncertain systems, and the event-triggered mechanism, especially with discussions on both discrete-time and continuous-time cases. In this paper, we aim to discuss recent research progress on these problems, primarily focusing on discrete-time systems while supplementing some excellent work on continuous-time systems. Fig. 1 illustrates some of the key technologies in the ADP field involved in this paper. In order to promote the further development of ADP, this paper provides a comprehensive overview of theoretical research, algorithm implementation, and related applications. It covers the latest research advances and also analyzes and predicts the future trends of ADP. This paper consists of the following parts: 1) basic background, 2) recent progress of ADP in the field of optimal control, 3) development of ADP in the event-triggered framework, 4) development of ADP in complex environments and the combination of ADP with other advanced control methods, 5) the impact of the data-driven method and RL on ADP technology, 6) typical control applications of ADP and RL, and 7) discussion of the possible future directions for ADP.

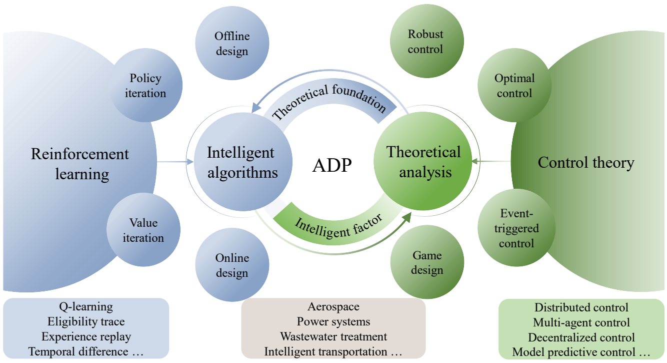


Fig. 1. Taxonomy diagram of related methods in this survey.

II. Optimal Regulation and Tracking With ADP

Generally speaking, ADP-based algorithms can be performed offline or online. In this section, we focus on the problem of optimal regulation and optimal tracking control, with a detailed overview.

A. *Offline Optimal Regulation With ADP*

1) Discrete-Time Systems: Consider the following affine discrete-time nonlinear systems:

$$x(k+1) = f(x(k)) + g(x(k))u(k), \quad k \in \mathbb{N} \tag{1}$$

where $x(k) \in \mathbb{R}^n$ is the state vector, $u(k) \in \mathbb{R}^m$ is the control vector, $\mathbb{N} = \{0, 1, 2, \dots\}$, and $\mathbb{R}$ is the set of real numbers. The system functions $f(\cdot) \in \mathbb{R}^n$ and $g(\cdot) \in \mathbb{R}^{n\times m}$ are differentiable with respect to their arguments and $f(0) = 0$. Assume that system (1) is stable on a compact set $\Omega \subset \mathbb{R}^n$.

The exact optimal feedback control law $u^*(x(k))$ is almost impossible to obtain by solving the HJB equation. For (1), Al-Tamimi *et al.* [20] proposed a VI scheme for optimal control problem. By setting $V^0(x(k)) = 0$, the iterative control policy and the iterative cost function can be solved by policy improvement

$$u^i(x(k)) = \arg\min_{u(x(k))} \left\{ U\big(x(k), u(x(k))\big) + V^i(x(k+1)) \right\} \tag{2}$$

and value function update

$$\begin{aligned} V^{i+1}(x(k)) &= \min_{u(x(k))} \left\{ U\big(x(k), u(x(k))\big) + V^i(x(k+1)) \right\} \\ &= U\big(x(k), u^i(x(k))\big) + V^i(x(k+1)) \end{aligned} \tag{3}$$

where $i \in \mathbb{N}$ is the iteration index, $U(x(k), u^i(x(k))) = x^T(k)Qx(k) + u^{iT}(x(k))Ru^i(x(k))$, and $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are both positive definite matrices. According to the mathematical induction, the iterative cost function sequence is proven to be monotonically nondecreasing by constructing an auxiliary function. Then, the convergence and optimality of VI can be proved, i.e., $\lim_{i\to\infty} V^i(x(k)) = V^*(x(k))$ and $\lim_{i\to\infty} u^i(x(k)) = u^*(x(k))$. It is noted that in offline algorithms, the approximations to the cost function and control policy are mainly achieved by neural networks or polynomials.

As we know, heuristic dynamic programming (HDP) and dual heuristic programming (DHP) are often used to implement VI. The different terms of the two structures are shown in Table I. In Fig. 2, general ADP structures are displayed, where $\hat{C}^i_{\text{out}}(x(k+1))$ and $\hat{C}^{i+1}_{\text{out}}(x(k))$ are outputs of the critic network with different iteration mechanisms, and $U_{\text{diff}}(x(k), u(k))$ is the utility function of different forms.

TABLE I
Basic Terms of the General ADP Structure

Terms	$U_{\text{diff}}(x(k),u(k))$	$\hat{C}^i_{\text{out}}(x(k+1))$	$\hat{C}^{i+1}_{\text{out}}(x(k))$
HDP	$U(x(k),u(k))$	$\hat{V}^i(x(k+1))$	$\hat{V}^{i+1}(x(k))$
DHP	$\dfrac{\partial U(x(k),u(k))}{\partial x(k)}$	$\dfrac{\partial \hat{V}^i(x(k+1))}{\partial x(k+1)}$	$\dfrac{\partial \hat{V}^{i+1}(x(k))}{\partial x(k)}$

Besides, the action-dependent and goal-representation versions are also used sometimes for these structures. Taking HDP as an example, the action-dependent HDP (ADHDP) [21] consists of three parts: the controlled object, the critic network, and the action network. It is capable of achieving optimal control without using system information. Compared

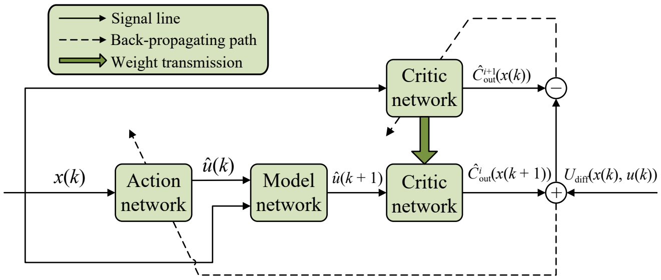


Fig. 2. The general structure of ADP.

with ADHDP, a network was added to goal-representation HDP (GrHDP) [22], which can be associated with the critic network and the action network. The goal network can generate, control, calculate, and plan more accurate system signals. It also improves the learning ability of the control system. In Table II, we show the comparison of the input and output of ADHDP with GrHDP, where

$$\begin{aligned} S(k) = \; & U\Big(x(k), u^i(x(k))\Big) \\ & + \iota U\Big(x(k+1), u^i(x(k+1))\Big) \\ & + \iota^2 U\Big(x(k+2), u^i(x(k+2))\Big) + \cdots \end{aligned} \tag{4}$$

is the internal reinforcement signal and $0 < \iota < 1$ is the discount factor. On the basis of [20], the optimal control problem was solved for nonlinear systems with control constraints by DHP [23], [24]. To overcome symmetric input constraints, Wang *et al.* [23] introduced the DHP framework involving a new nonquadratic performance index and used the data-based methods to derive efficient system models. Then, for a class of nonlinear systems with asymmetric constraints, Wang *et al.* [24] defined an innovative nonquadratic function and expanded the application scope of the DHP framework. Besides, Xu *et al.* [25] introduced a control barrier function into the utility function. Unlike [23], [24], HDP was used to solve state-constrained optimal control problems in [25].

TABLE II
COMPARISON OF THE INPUTS AND THE OUTPUTS BETWEEN ADHDP AND GRHDP

Terms	$U_{\rm diff}(x(k),u(k))$	$\hat{C}^i_{\rm in}(x(k))$	$\hat{C}^i_{\rm out}(x(k))$
ADHDP	$U(x(k),u(k))$	$x(k),u(k)$	$V^i(x(k)),V^i(x(k+1))$
GrHDP	$S(k)$	$x(k),u(k),S(k)$	$V^i(x(k))$

Compared to VI [20], monotonicity is various for iterative cost function sequence with the more general initial condition in general VI (GVI) [26]. The GVI algorithm can be initialized by a positive semi-definite function $V^0(x) = x^T\Phi x$, where Φ is a positive semi-definite matrix. Furthermore, the iterative cost function was proved to satisfy the inequality

$$\begin{aligned} \left[1+\frac{\alpha-1}{(1+\theta^{-1})^i}\right] & V^*(x(k)) \le V^i(x(k)) \\ & \le \left[1+\frac{\beta-1}{(1+\theta^{-1})^i}\right]V^*(x(k)) \end{aligned} \tag{5}$$

where $0 \le \alpha \le 1, 1 \le \beta < \infty$, and $1 \le \theta < \infty$.

Then, by using a novel convergence analysis, Wei *et al.* [27] proved the convergence and optimality of GVI. In addition, it was shown that the termination criterion in [20] could not guarantee admissibility of the near-optimal control policy. The admissibility termination criterion was proposed as

$$V^{i+1}(x(k)) - V^i(x(k)) < U\Big(x(k), u^i(x(k))\Big). \tag{6}$$

Following [27], Ha *et al.* [28] presented a new admissibility condition of GVI described by

$$V^{i+1}(x(k)) - V^i(x(k)) < \epsilon U\Big(x(k), u^i(x(k))\Big) \tag{7}$$

where $0 < \epsilon < 1$. Besides, considering the discounted GVI, $u^i(x(k))$ was stable if $V^0(x(k)) \ge V^1(x(k))$ and the discount factor γ satisfied $\xi^i_\gamma = 1 - Q(x(k))/V^i(x(k)) < \gamma < 1$. It was also proven that if $U(x(k), u^i(x(k))) < V^i(x(k))$ and $\gamma > \xi^i_\gamma$, u^{i+b} was stable, $b \in \mathbb{R}$. Wang *et al.* [29] used GVI to address the discounted optimal regulation problem for discrete-time systems with control constraints. Kamanchi *et al.* [30] constructed a Bellman equation to apply the Newton-Raphson method to the successive relaxation VI scheme, which expanded traditional VI to the second-order iteration process. They also provided the proof of global convergence and some convincing experiments.

Despite plenty of research on VI, PI has great advantages in terms of stability guarantee of the iterative control law. Therefore, Liu and Wei [31] proposed a PI-based algorithm, where $u^0(x(k))$ was required to be admissible. $V^i(x(k))$ and $u^{i+1}(x(k))$ can be obtained by policy evaluation

$$V^i(x(k)) = U\Big(x(k), u^i(x(k))\Big) + V^i(x(k+1)) \tag{8}$$

and policy improvement

$$u^{i+1}(x(k)) = \arg\min_{u(x)} \Big\{U\Big(x(k), u(x(k))\Big) + V^i(x(k+1))\Big\}. \tag{9}$$

In [31], the convergence and stability of PI were analyzed for the first time, where the initial admissible control law was obtained by trial-and-error. The iteration process can ensure that all iterative control laws were stable. Compared with [31], the admissible control law can be obtained more conveniently in [27] and [28]. Based on [31], Liu *et al.* [15] proposed the generalized PI (GPI) for optimal control of system (1), where the convergence and optimality properties were guaranteed. In essence, VI [20] and PI [31] were both special cases of GPI. Since many systems could only be locally stabilized, in order to solve the regionality existing in discrete-time optimal control, an invariant PI method was proposed by Zhu *et al.* [32], where the suitable region for the new policy was updated.

In addition, there are some works on the combination of VI and PI. In [33], Luo *et al.* introduced an adaptive method to solve the Bellman equation, which balanced VI and PI by adding a balance factor. It is noted that the algorithm in [33] can accelerate the iterative process and do not need the initial admissible control law. To obtain the stable iterative control policy, Heydari [34] proposed stabilizing VI, where the initial admissible u^0 was evaluated to implement the VI. Based on [34], Ha *et al.* [28] developed an integrated VI method based on GVI, which was used to generate the admissible control law. In Table III, we summarize the initial conditions and monotonicity of GVI ($V^0 \le V^1$), GVI ($V^0 \ge V^1$), Stabilizing VI, and Integrated VI. Integrated VI consists of GVI ($V^0 \le V^1$) and Stabilizing VI, where GVI ($V^0 \le V^1$) provides the initial admissible control policy for Stabilizing VI. Therefore, in the following table, integrated VI only represents the monotonicity of its core component.

TABLE III
CLASSIFICATION OF VI ALGORITHMS

Method	Initial condition	Monotonicity
GVI ($V^0 \le V^1$)	$V^0(x) = x^T\Phi x$	Monotonically nondecreasing
GVI ($V^0 \ge V^1$)	$V^0(x) = x^T\Phi x$	Monotonically nonincreasing
Stabilizing VI	An admissible $u^0(x)$	Monotonically nonincreasing
Integrated VI	$V^0(x) = 0$	Monotonically nonincreasing

2) Continuous-Time Systems: Compared with discrete-time systems, most literature focuses on the PI method for continuous-time systems even though the VI strategy still can be used as in [35]. We consider the continuous-time systems

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t) \tag{10}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $f(\cdot) \in \mathbb{R}^n$, and $g(\cdot) \in \mathbb{R}^{n\times m}$ represent the state vector, the control vector, the drift dynamics, and the input dynamics, respectively. Assume that $f(0) = 0$ and the system is stabilized on the operation region. For systems in the strict feedback form with uncertain dynamics, Zargarzadeh *et al.* [36] utilized neural networks to estimate the cost function by using state measurement. In [37], a data-based continuous-time PI algorithm was proposed, where a critic-identifier was introduced to estimate the cost function and the Hamiltonian of the admissible policy. Differently from [38], the algorithm in [37] was used in continuous-time systems and did not require samples of the input and output trajectories of the system. Compared with [36], the method proposed in [37] could be extended to multicontroller systems. In addition, to release the computational burden, a novel distributed PI algorithm was established in [39]. The iterative control policy could be updated one by one. The above works [35]–[39] all focused on time-invariant nonlinear systems. Moreover, $V(\cdot)$ and $u(\cdot)$ both relied on the system state. In [40], for time-varying nonlinear systems, Wei *et al.* developed a novel PI algorithm, where the optimality and stability were discussed. It is worth noting that a mass of literatures concentrate on the progress of VI algorithms and their structures are similar to those of discrete-time systems. Bian and Jiang [41] extended VI to continuous-time nonlinear systems.

B. Online Optimal Regulation With ADP

1) Discrete-Time Systems: As mentioned in [34], differently from offline ADP, online ADP needs to be implemented through selecting the initial control policy, and improving it according to some criteria until it converges to the optimal value. Note that the key difference between offline ADP and online ADP is that the control policy generated by offline ADP keeps unchanged during the controlled stage of systems, whereas the control policy will be updated in online ADP.

First, consider online ADP for the discrete-time systems. Usually, the optimal cost function and the optimal control policy are approximated by neural networks as

$$V^*(x(k)) = \phi_c^T\sigma(x(k)) + \varepsilon_c \tag{11}$$

and

$$u^*(x(k)) = \theta_a^T\delta(x(k)) + \varepsilon_a \tag{12}$$

respectively, where ϕ_c and θ_a are the weight vectors of target neural networks, ε_c and ε_a are the bias terms, $\sigma(\cdot)$ and $\delta(\cdot)$ are the activation function vectors. The optimal cost function is estimated by the critic network

$$\hat{V}(x(k)) = \hat{\phi}_c^T\sigma(x(k)) \tag{13}$$

and the optimal control policy is estimated by the action network

$$\hat{u}(x(k)) = \hat{\theta}_a^T\delta(x(k)) \tag{14}$$

where $\hat{\phi}_c$ and $\hat{\theta}_a$ are the estimated values of ϕ_c and θ_a, respectively. Since $V^*(x(k))$ and $u^*(x(k))$ satisfy the HJB equation, we get

$$\begin{aligned} &H\Big(x(k), u^*(x(k)), V^*(x(k))\Big)\\ &\quad= V^*(x(k+1)) + U\Big(x(k), u^*(x(k))\Big) - V^*(x(k)) = 0. \end{aligned} \tag{15}$$

Substituting (13) and (14) to the HJB equation, we obtain

$$\begin{aligned} &H\Big(x(k), \hat{u}(x(k)), \hat{V}(x(k))\Big)\\ &\quad= \hat{V}(x(k+1)) + U\Big(x(k), \hat{u}(x(k))\Big) - \hat{V}(x(k)) = e_c. \end{aligned} \tag{16}$$

Define $\tilde{u}(x(k)) = \hat{u}(x(k)) + \frac{1}{2}R^{-1}g^T(x(k))\frac{\partial\hat{V}(x(k+1))}{\partial x(k+1)}$. We need to ensure that e_c and $\tilde{u}(x(k))$ both converge to zero. In order to achieve the goal, we choose suitable weights to update the control law. To relax the requirement of activation functions, Moghadam *et al.* [42] proposed an online optimal adaptive control algorithm with multi-layer neural networks, where the vanishing gradient problem was overcome. This algorithm can be extended to neural networks with an arbitrary number of hidden layers, and the weight update laws in the action-critic network can be defined as a function of temporal-difference (TD) errors without previous information of the system state.

In the above research on online ADP, the approximate optimal control policy was updated by tuning the weights of neural networks. Besides, the improved control law can be acquired by PI or VI. Since the iterative control policy obtained by PI is stable, PI is widely used in online control. However, there are also some works on updating the control policy by VI. For example, in [34], Heydari proposed an online algorithm based on stabilizing VI, where the system was controlled under different iterative policies. In [14], combining the stability condition of GVI and the concept of attraction domain, the novel online algorithm was introduced by Ha *et al.*, where the current control law was chosen by the location of the current state.

2) Continuous-Time Systems: For continuous-time systems, the principle of online ADP is similar to that of discrete-time systems. Here, we display some main progress on online methods with PI. For the weakly coupled nonlinear systems, a data-based online learning algorithm was established by Li *et al*. [43], where the original optimal control problem of the weakly coupled systems was transformed into three reduced-order optimal control problems. In [38], He *et al*. introduced a novel online PI method, where the technique of neural-network-based online linear differential inclusion was used for the first time. In addition, to solve optimal synchronization of multi-agent systems, an off-policy RL algorithm was presented in [44], where dynamic models of the agents were not required.

C. Optimal Tracking Design With ADP

1) Discrete-Time Systems: With the development of aviation, navigation, and other fields in recent years, the research interest in optimal tracking design has gradually increased within the control community. Here, we need to concentrate on the optimal tracking control problem. Define the desired tracking trajectory as

$$r(k+1) = F(r(k)),\ \ k \in \mathbb{N}. \tag{17}$$

Considering original system (1), the tracking error $e(k)$ is described as

$$e(k) = x(k) - r(k),\ \ k \in \mathbb{N}. \tag{18}$$

Assume that there exists the steady control $u_d(k)$ to make the following equation hold:

$$r(k+1) = f(r(k)) + g(r(k))u_d(k),\ \ k \in \mathbb{N}. \tag{19}$$

The objective of the optimal tracking control problem is to find the optimal control law $u(x(k))$, which can force the system output to track the reference trajectory. This can be obtained by minimizing the performance index or the cost function. Hence, the choice of the cost function is of importance without doubt. Generally, we choose the form of the cost function according to the control objective. Wang *et al.* [23] applied DHP to implement the tracking control design towards nonaffine discrete-time systems, where the discount factor was considered. After that, actuator saturation was also considered in [24]. It is noted that the form of the utility function in [45]–[47] is given by

$$U_1(e(k), u_e(k)) = e^T(k)Qe(k) + u_e^T(k)Ru_e(k) \tag{20}$$

where $u_e(k) = u(x(k)) - u_d(k)$. Since it is not convenient to calculate the reference control policy $u_d(k)$, some scholars choose other forms of utility function. For example, Kiumarsi and Lewis [48] introduced a partially model-free ADP method. In this work, an optimal tracking control of nonlinear systems with input constraints is achieved by using a discounted performance function based on the augmented system. In [49], Lin *et al*. proposed a policy gradient algorithm and used experience replay for optimal tracking design. They used the Lyapunov's direct method to prove the uniform ultimate boundedness (UUB) of the closed-loop system. The utility function in [48] and [49] is described as

$$U_2\big(e(k), u(x(k))\big) = e^T(k)Qe(k) + u^T(x(k))Ru(x(k)). \tag{21}$$

Even though the steady control is avoided in (21), it can not eventually eliminate the tracking error. To deal with this problem, Li *et al*. [50] developed a novel utility function given by

$$U_3\big(e(k), r(k), u(x(k))\big) = e^T(k+1)Qe(k+1). \tag{22}$$

The optimality of VI and PI was analyzed. In addition, Ha *et al*. [51] also analyzed the system stability of the VI algorithm for the novel utility function with a discount factor.

2) Continuous-Time Systems: There are also a few works on the continuous-time systems. In [52], Gao and Jiang solved the optimal output regulation problem by ADP and RL, where ADP was for the first time combined with the output regulation problem for adaptive optimal tracking control with disturbance attenuation. However, this approach requires partial knowledge of the system dynamics. To overcome this difficulty, in [53], the integral RL algorithm was introduced to achieve optimal online control, where the off-policy integral RL was employed to obtain the optimal control feedback gain for the first time. In addition, differently from [52], the algorithm in [53] relieved the computational burden. Then, in [54], Fu *et al*. proposed a robust approximate optimal tracking method. In order to relax the assumption that the reference signal must be continuous in continuous-time systems, a new Lyapunov function was proposed without knowing the derivative information of the tracking error.

In particular, ADP also plays a pivotal role in the optimal control of linear systems, such as the linear quadratic regulation (LQR) problem and tracking problem [55]–[60]. Generally speaking, considering optimal control for nonlinear systems, the HJB equation is usually solved to acquire the optimal control policy. However, the linear system is a special case, which has good properties. The solution of the HJB equation can be transformed into the solution of the algebraic Riccati equations, so as to obtain the exact optimal control law. In [56], Rizvi and Lin proposed an online Q-learning method based on output feedback to tackle the LQR problem. Wang *et al.* [57] developed an optimal LQR based on the discounted VI algorithm and provided a series of criteria to judge the stability of the systems. In [58], the LQR problem was solved for the continuous-time systems with unknown system dynamics and without an initial stabilizing strategy. The proposed controller was updated continuously by utilizing the measurable input-output data to avoid instability. For the same uncertain systems, Rizvi and Lin [59] proposed a model-free static output feedback controller based on RL, which avoided the influence of the exploration bias problem. In addition, researchers also pay much attention to the optimal tracking design for linear systems. For networked control systems with uncertain dynamics, Jiang *et al.* [60] developed a Q-learning algorithm to obtain the online optimal control policy based on measurable data with network-induced dropouts.

III. Event-Triggered Control With ADP

In this section, we mainly introduce the application of event-triggered technology under the ADP framework. It is discussed for discrete-time systems and continuous-time systems, respectively.

As an advanced aperiodic control method, event-triggered control plays a vital role in decreasing the computational burden, and enhancing the resource utilization rate. In short, the purpose of introducing the event-triggered mechanism is to reduce the updating times of the controller by decreasing the sampling times of the system state. Unlike the time-triggered control method, event-triggered control is designed with triggering conditions that are required to satisfy the stability of the controlled system. The control input is updated only when this triggering condition is violated. Conversely, if the triggering condition is not violated, the zero-order hold is able to keep the control input unchanged until the next event is triggered.

A. Event-Triggered Control for Discrete-Time Systems

For discrete-time systems, the event-triggered technology has been widely used in the adaptive critic framework. In [61], Dong *et al.* used the event-triggered method to solve the optimal control problem under the HDP framework, and proved that the controlled system was asymptotically stable. An event-triggered near-optimal control algorithm was proposed for the affine nonlinear dynamics with constrained inputs in [62]. In addition, a special cost function was introduced and the system stability was analyzed. In [63], a novel adaptive control approach with disturbance rejection was designed for linear discrete-time systems. In [64], Zhao *et al.* proposed a new event-driven method via direct HDP. Then, the UUB of the system states and the weights in the control policy networks was proven. In [65] and [66], a novel event-triggered optimal tracking method was developed to control the affine system. It is worth noting that the triggering condition in these two works only acts on the time step and the updating stage of weights is not involved in the iterative process. For systems whose models are known, by using the event-triggered control approach, not only the reference trajectory can be tracked, but also the computational burden can effectively be reduced. In [67], Wang *et al.* proposed an event-based DHP method, where three kinds of neural networks were used to identify nonlinear systems, estimate the gradient of the cost function, and approximate the tracking control law. In addition, the stability of the event-based controlled system was proved by the theorem of input-to-state stability and the control scheme was applied to wastewater treatment simulation platform.

For (1), considering the role of the event-triggered mechanism, we define a monotonically increasing sequence consisting of different sampling states as $\{k_j\}_{j=0}^{\infty}$, where k_j is the jth sampling moment, $j \in \mathbb{N}$. The event-based system state signal is only updated at sampling instants: $k_0, k_1, k_2, \ldots$. In other words, for the sampled signals between k_j and k_{j+1}, the feedback control law $u(x(k))$ remains unchanged. Then, the control law can be expressed by $u(x(k)) = u(x(k_j))$, where $x(k_j)$ is the system state vector at the sampling instant k_j, $k_j \le k < k_{j+1}$. In addition, we apply the zero-order-hold to maintain the unchanged input of the event-based controller within the range $k \in [k_j, k_{j+1})$.

Then, the event-triggered error vector is defined as

$$\vartheta(k) = x(k_j) - x(k) \tag{23}$$

where $k_j \le k < k_{j+1}$. Obviously, we have $\vartheta(k) = 0$ at $k = k_j$. Then, the closed-loop form of system (1) becomes

$$x(k+1) = f(x(k)) + g(x(k))u(x(k) + \vartheta(k)). \tag{24}$$

According to Bellman's optimality principle, the optimal cost function $V^*(x(k))$ can be obtained by designing a sequence of the event-based control law $u(x(k_j))$. Therefore, $V^*(x(k))$ can be expressed by

$$V^*(x(k)) = \min_{u(\cdot)} \sum_{p=k}^{\infty} U\big(x(p), u(x(k_j))\big). \tag{25}$$

The corresponding optimal control $u^*(x(k_j))$ can be obtained by

$$u^*(x(k_j)) = -\frac{1}{2}R^{-1}\left(\frac{\partial x(k+1)}{\partial u(x(k_j))}\right)^T \frac{\partial V^*(x(k+1))}{\partial x(k+1)}. \tag{26}$$

Next, we introduce several triggering conditions commonly used in combination with adaptive critic methods.

1) Suppose there exists a positive number $\mathcal{I}$ satisfying

$$\|x(k+1)\| \le \mathcal{I}\|\vartheta(k)\| + \mathcal{I}\|x(k)\|. \tag{27}$$

In addition, the inequality $\|\vartheta(k+1)\| \le \|x(k+1)\|$ holds. By referring to [61], [62], [67], a triggering condition was designed as

$$\|\vartheta(k)\| \le \vartheta_T = \frac{1-(2\mathcal{I})^{k-k_j}}{1-2\mathcal{I}}\mathcal{I}\|x(k_j)\|, \ \ \mathcal{I} \ne 0.5. \tag{28}$$

We can obtain different levels of triggering effect by appropriately adjusting the parameter $\mathcal{I}$.

2) According to the updating method of neural networks in [64], we assume that the activation function σ_a in the action network satisfies

$$\|\sigma_a(x_1) - \sigma_a(x_2)\| \le \mathcal{P}\|x_1 - x_2\| \tag{29}$$

for all $x_1, x_2 \in \mathcal{X}$, where $\mathcal{P}$ is a positive constant and $\mathcal{X}$ is the domain of system dynamics.

Lemma 1 [64]: Let (29) hold for the nonlinear system (1). Assume the triggering condition is defined as follows:

$$\|\vartheta(k)\|^2 \le \frac{\lambda_{\min}(Q)\beta}{2\lambda_{\max}(R)\|\hat{w}_a\|^2\mathcal{P}^2}\|x(k)\|^2 \tag{30}$$

where $0 \le \beta < 1$ and $\hat{w}_a$ is the weight of the action network. $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the minimal and maximal eigenvalues of a matrix, respectively. In addition, we make the action network learning rate satisfy

$$l_a < \frac{1}{\|\sigma_a(k)\|^2} \tag{31}$$

and the critic network learning rate satisfy

$$l_c < \frac{1}{\|\sigma_c(k)\|^2}. \tag{32}$$

Then, we can declare that the event-based control input can guarantee the UUB of the controlled system.

3) The triggering condition described below can only be applied to the time-based case. For the iterative process, the traditional time-triggered method is adopted. In [65], [66], a triggering condition was defined as follows:

$$T_\gamma(x(k+1), x(k), x(k_j)) \le 0 \tag{33}$$

where

$$T_\gamma(x(k+1),x(k),x(k_j)) = \gamma\big(\Delta V^*(x(k))\big)+x^T(k)Qx(k) + u^T(x(k_j))Ru(x(k_j)) \tag{34}$$

with $\gamma > 1$. In addition, $\Delta V^*(x(k)) = V^*(x(k+1)) - V^*(x(k))$ represents the first-order difference of the optimal cost function under the time-triggered mechanism. According to the updating rule of the event-triggered mechanism, the next sampling time k_{j+1} is expressed as

$$k_{j+1} = \inf\{k|T_\gamma(x(k+1),x(k),x(k_j)) > 0, k > k_j\}. \tag{35}$$

According to the results in [65], the adjustable parameter γ plays an essential role in the event-triggered optimal control. If the main emphasis is on optimizing the cost function, γ should be chosen as small as possible. On the contrary, when considering resource utilization, γ should be chosen as large as possible. Therefore, the selection of γ should be determined according to the actual need.

B. Event-Triggered Control for Continuous-Time Systems

There are extensive studies of event-triggered control methods within the framework of ADP for continuous-time systems. In [68], Luo *et al.* designed an event-triggered optimal control method directly based on the solution of the HJB equation. In addition, the stability of the system and the lower bound on the interexecution times were proved theoretically. In [69], for a class of nonlinear multi-agent systems, novel event-triggered and asynchronous edge-event triggered mechanisms were designed for the leader and all edges, respectively. In [70], Huo *et al.* developed a decentralized event-triggered control method to aperiodically update each auxiliary subsystem. In [71], a different event-based decentralized control scheme was proposed. They used codesign strategies to trade-off control policies and triggering thresholds to simultaneously achieve optimization of subsystem performance and reduction of computational burden.

Considering the continuous-time nonlinear system (10), we assume $f+gu$ to be Lipschitz continuous on Ω that contains the origin. We assume that there exists an admissible control $u(x(t))$ and the cost function is defined as

$$V_u(x(t)) = 2\int_t^\infty \Big(Q(x(t)) + \|u(x(t))\|_R^2\Big)dt \tag{36}$$

for all $x(t) = x \in \Omega$, where $Q(x(t)) = x^T(t)Qx(t)$, $V_u(x(t)) \geq 0$ and $V_u(0) = 0$. The corresponding Hamiltonian is expressed as

$$H(x,u(x),\nabla V_u(x)) = [\nabla V_u(x)]^T(f(x)+g(x)u(x)) + Q(x) + \|u(x)\|_R^2 \tag{37}$$

where $\nabla V_u(x) = \frac{\partial V_u(x)}{\partial x}$. By using (37), differentiating (36) with respect to t yields

$$H(x,u(x),\nabla V_u(x)) = 0. \tag{38}$$

Let $V^*(x) = V_{u^*}(x)$. The optimal cost function can be expressed as

$$V^*(x) = \min_u V(x_0,u). \tag{39}$$

Next, the optimal control law under the time-triggered mechanism is defined as

$$u^*(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla V^*(x). \tag{40}$$

The event-triggered mechanism is similar to that of discrete-time systems. Therefore, we define the state as

$$\breve{x}(t) = \begin{cases} x(t), & t = t_j \\ x(t_j), & t \in (t_j, t_{j+1}) \end{cases} \tag{41}$$

for all $j \in \mathbb{N}$. The optimal control law under the event-triggered mechanism can be expressed as

$$u^*(\breve{x}(t)) = -\frac{1}{2}R^{-1}g^T(\breve{x}(t))\nabla V^*(\breve{x}(t)). \tag{42}$$

For conventional event-triggered control, the design of triggering conditions is inevitable. Next, we introduce two triggering conditions under continuous-time environments.

1) This triggering condition is established based on a reasonable Lipschitz condition.

Assumption 1 [72]: Assume that $u^*(x(t))$ has the Lipschitz property on Ω. In addition, there exists a Lipschitz constant $K_{u^*} > 0$ such that

$$\|u^*(x(t)) - u^*(\breve{x}(t))\| \leq K_{u^*}\|\vartheta_j(t)\| \tag{43}$$

for all $x(t), \breve{x}(t) \in \Omega$, where $\vartheta_j(t) = \breve{x}(t) - x(t)$ is the sampling error.

Lemma 2 [72]: Suppose that Assumption 1 holds. If the triggering condition is defined as

$$\|\vartheta_j(t)\|^2 \leq \frac{(1-2\theta_t)\lambda_{\min}(Q)}{2K_{u^*}^2}\|x(t)\|^2 = \|\vartheta_T(t)\|^2 \tag{44}$$

where $0 < \theta_t < 1/2$ and $\vartheta_T(t)$ is the triggering threshold, $u^*(\breve{x}(t))$ can force system (10) to be stable in the sense of UUB.

2) There is another triggering condition which is similar to the third one described in the discrete-time case. In [68], in order to determine the release time instant t_j, an event-triggering condition is given as follows:

$$C_\alpha(x(t),\breve{x}(t)) < 0 \tag{45}$$

where

$$C_\alpha(x(t),\breve{x}(t)) = Q(x(t)) + \|u(x(t))\|_R^2 + (1+\alpha)\Big(\nabla V^*(x(t))\Big)^T\Big(f(x(t)) + g(x(t))u(\breve{x}(t))\Big) \tag{46}$$

with $\alpha > 0$ being a constant. Then, it is proved that the controlled system was asymptotically stable by using this triggering condition.

The main purpose of the event-triggered technology is to reduce the waste of communication resources and improve computational efficiency. In recent years, networked control systems have attracted extensive attention. There is also an increasing amount of work aimed at reducing the energy consumption of network interfaces and ensuring the sustainability of networked control systems. Some related studies can be found in [73], [74].

IV. Robust Control and Game Design With ADP

In modern engineering systems, the real control plants are always affected by changes derived from the system model, external environment, and other factors. Hence, it is of great importance to attain the robust control strategy to avoid the influence of uncertainties. The problem of robust control can be turned into a problem of optimal control, which is a useful method for attaining the robust controller. However, for complex nonlinear systems, it is difficult to solve the optimal control problem. To deal with this dilemma, the ADP method is utilized. In this section, recent research progress of ADP is described, such as using ADP to solve the problem of robust control, H_∞ control, and multi-player game design. In addition, some other advanced control methods with ADP are supplemented at the end of this section.

A. Robust Control Design With ADP

By utilizing ADP, robust controllers can be designed based on the obtained optimal control strategy. Compared to traditional methods, controllers guided by ADP can not only stabilize the system, but also optimize the performance of systems. The recent work on robust control is analyzed from both discrete-time and continuous-time aspects in this section.

1) Discrete-Time Systems: We consider a class of discrete-time nonlinear systems with uncertain terms as

$$x(k+1) = f(x(k)) + g(x(k))u(k) + \Delta f(x(k)) \tag{47}$$

where $k \in \mathbb{N}$, the state $x(k) \in \mathbb{R}^n$, and the control input $u(k) \in \mathbb{R}^m$. $f(\cdot)$ and $g(\cdot)$ are differentiable with respect to its arguments and $f(0) = 0$. $\Delta f(x(k))$ is the unknown dynamics function. Considering the matched uncertainty of system dynamics, we can define $\Delta f(x(k)) = g(x(k))d(x(k))$. In addition, $d(x(k)) \in \mathbb{R}^m$ and $d(x(k))$ is upper bounded by $\|d(x(k))\| \le d_M(x(k))$ with $d_M(0) = 0$. For the uncertain system with the matched uncertainty, in order to attain robust stabilization, we need to find a state feedback control law $u(x(k))$, which can make the closed-loop system asymptotically stable for all uncertainties $d(x(k))$. A suitable cost function is designed for the corresponding nominal system. In this way, the problem of robust control can be transformed into the problem of optimal control. For the transformed optimal control problem, our goal is to acquire the feedback control law $u(x(k))$ to minimize the cost function

$$V(x(k)) = \sum_{k=0}^{\infty}\left\{\rho d_M^2(x(k)) + U\big(x(k), u(x(k))\big)\right\} \tag{48}$$

where $\rho > 0$, the utility function $U(x(k), u(x(k))) = x^T Qx + u^T Ru \ge 0$ with $U(0,0) = 0$, and Q and R are positive definite matrices. Note that the cost function (48) is different from the common form in the optimal control problem. According to Bellman's optimality principle, the cost function $V(x(k))$ satisfies the discrete-time HJB equation and can be expressed as

$$\begin{aligned} V^*(x(k)) = \min_{u(x)}\Big\{&\rho d_M^2(x(k)) + U\big(x(k), u(x(k))\big) \\ &+ V^*(x(k+1))\Big\}. \end{aligned} \tag{49}$$

Hence, the optimal control law $u^*(x)$ can be obtained as follows:

$$u^*(x(k)) = -\frac{1}{2}R^{-1}g^T(x(k))\nabla V^*(x(k+1)). \tag{50}$$

Then, by using the optimal control law $u^*(x(k))$, the discrete-time HJB equation (49) becomes

$$\begin{aligned} V^*(x(k)) = \ &\rho d_M^2(x(k)) + U\big(x(k), u^*(x(k))\big) \\ &+ V^*(x(k+1)). \end{aligned} \tag{51}$$

By choosing an appropriate utility function, robust stabilization was transformed into an optimal control problem for nominal systems [75]–[77]. In [76], the idea of solving the generalized HJB equation was employed to derive a robust control policy for discrete-time nonlinear systems subject to matched uncertainties. A neural network was used as the function approximator. In addition, Li *et al.* [77] proposed an adaptive interleaved RL algorithm to find the robust controller of discrete-time nonlinear systems subject to matched or mismatched uncertainties. An action-critic structure was given to skillfully handle experiments. The convergence of the proposed algorithm and the UUB of the system were proved. An appropriate utility function was chosen as

$$\begin{aligned} U\big(x(k), u(x(k))\big) = \ &x^T(k)Qx(k) \\ &+ u^T(x(k))u(x(k)) + \beta x(k). \end{aligned} \tag{52}$$

Note that there is a new term $\beta x(k)$ in the utility function compared to the traditional expression of $x^T(k)Qx(k) + u^T(x(k))Ru(x(k))$. Tripathy *et al.* [78] introduced a virtual input to compensate the effect of uncertainties. By defining a sufficient condition, the stable control law of the mismatched system was derived. At the same time, the stability of the uncertain system was proved. The uncertainty can be decomposed in matched and mismatched components as

$$\begin{aligned} d(x(k)) = \ &g(x(k))g(x(k))^{+}\mathfrak{S}\phi(x(k)) \\ &+ \big(I_m - g(x(k))g(x(k))^{+}\big)\mathfrak{S}\phi(x(k)) \end{aligned} \tag{53}$$

where I_m denotes the identity matrix with approximate dimensions. $g(x(k))^{+} = (g^T(x(k))g(x(k)))^{-1}g^T(x(k))$ denotes the left pseudo inverse of the matrix $g(x(k))$, and $\mathfrak{S}$ is a design matrix.

2) Continuous-Time Systems: For continuous-time nonlinear systems, the principle of robust control with ADP is similar to that of discrete-time systems. Considering uncertainties, the continuous-time nonlinear system is defined as

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t) + \Delta f(x(t)) \tag{54}$$

and the corresponding nominal system is defined as in (10). In order to obtain the optimal feedback control law $u(x(t))$, we need to minimize the cost function

$$\begin{aligned} V(x_0) &= \int_0^{\infty}\left\{\rho d_M^2(x(\tau)) + u^T(x(\tau))Ru(x(\tau))\right\}d\tau \\ &= \int_0^{\infty} r\big(x(\tau), u(x(\tau))\big)d\tau \end{aligned} \tag{55}$$

where $\rho > 0$, and the utility function $r(x,u) \ge 0$. Compared with the normal form, it is worth noting that the cost function (55) is modified to reflect matched uncertainties. We assume the control input $u \in \Psi(\Omega)$, where $\Psi(\Omega)$ is the set of admissible control laws on Ω. Then, the nonlinear Lyapunov equation can be expressed as

$$r(x, u(x)) + (\nabla V(x))^T(f(x) + g(x)u(x)) = 0 \tag{56}$$

where $\nabla V(x) = \frac{\partial V(x)}{\partial x}$. We define the optimal cost function of the system (10) as follows:

$$V^*(x_0) = \min_u \int_0^\infty r\Big(x(\tau), u(x(\tau))\Big) d\tau. \tag{57}$$

According to (56), we define the Hamiltonian as

$$\begin{aligned} H(x,u,\nabla V(x)) = {} & \rho d_M^2(x) + u^T(x)Ru(x) \\ & + (\nabla V(x))^T (f(x) + g(x)u(x)). \end{aligned} \tag{58}$$

Considering (56)–(58), the optimal cost function satisfies the HJB equation

$$\min_u H(x,u,\nabla V^*(x)) = 0. \tag{59}$$

Hence, we can obtain the optimal control law $u^*(x) = \arg\min_u H(x,u,\nabla V^*(x))$. That is

$$u^*(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla V^*(x). \tag{60}$$

ADP-based robust control schemes can be divided into the following categories: least-squares-based transformation methods [79], adaptive-critic-based transformation methods [80], data-based transformation methods [81], robust ADP methods [82], [83], and so on. In [84], Wang proposed an adaptive method based on the recurrent neural network to solve the robust control problem. A cost function with the additional utility function was defined to counteract the effect of perturbations on the system and the stability of the relevant nominal system was proved. The application scope of the ADP method was further expanded. In [85], the robust control was transformed into an optimal tracking control problem by introducing an auxiliary system including a steady-state part and a transient part, and the stability of the transient tracking error was analyzed. Pang *et al.* [86] studied the robustness of PI for addressing the continuous-time infinite-horizon LQR problem.

B. H_∞ Control Design With ADP

In H_∞ control design, a control law is constructed for dynamical systems containing external disturbances and uncertainties. According to the principle of minimax optimality, the H_∞ control problem is usually described as two-player zero-sum differential games. In order to obtain the controller that minimizes the cost function in the worst case, we need to find the Nash equilibrium solution corresponding to the Hamilton-Jacobi-Isaacs (HJI) equation. However, for general nonlinear systems, it is hard to obtain the analytical solution of the HJI equation, which is similar to the difficulty encountered in solving the nonlinear optimal control problem. In recent years, ADP has been widely used for solving H_∞ control problems.

1) Discrete-Time Systems: Consider the following discrete-time nonlinear system with external disturbances:

$$x(k+1) = f(x(k)) + g(x(k))u(k) + h(x(k))\upsilon(k) \tag{61}$$

where $x(k) \in \mathbb{R}^n$ is the state vector, $u(k) \in \mathbb{R}^m$ is the control input, and $\upsilon(k) \in \mathbb{R}^q$ is the disturbance input. We assume $f + gu + h\upsilon$ is Lipschitz continuous on Ω containing the origin.

We define the cost function as follows:

$$V(x(k),u(k),\upsilon(k)) = \sum_{k=0}^{\infty} U(x(k),u(k),\upsilon(k)) \tag{62}$$

where $U(x(k),u(k),\upsilon(k)) = x^T(k)Qx(k) + u^T(k)Ru(k) - \ell^2\upsilon^T(k) \times P\upsilon(k)$ is the utility function. Q, R, and P are positive definite matrices. ℓ is a positive parameter.

The design objective is to find the saddle point solution $(u^*(k), \upsilon^*(k))$, such that the following Nash condition holds:

$$\begin{aligned} V^*(x(k)) &= \min_u \max_\upsilon \{V(x(k),u(k),\upsilon(k))\} \\ &= \max_\upsilon \min_u \{V(x(k),u(k),\upsilon(k))\}. \end{aligned} \tag{63}$$

Based on Bellman's optimality principle, the optimal cost function $V^*(x(k))$ satisfies the following discrete-time HJI equation:

$$\begin{aligned} V^*(x(k)) = {} & \min_u \max_\upsilon \Big\{ U(x(k),u(k),\upsilon(k)) \\ & + V^*(x(k+1)) \Big\}. \end{aligned} \tag{64}$$

Then, the saddle point solution $(u^*(k), \upsilon^*(k))$ is

$$\begin{cases} u^*(k) = -\dfrac{1}{2}R^{-1}g^T(x(k))\nabla V^*(x(k)) \\ \upsilon^*(k) = \dfrac{1}{2\ell^2}P^{-1}h^T(x(k))\nabla V^*(x(k)). \end{cases} \tag{65}$$

In [87], [88], the H_∞ tracking control problem was studied by using the data-based ADP algorithm. Hou *et al.* [87] proposed an action-disturbance-critic structure to ensure that the minimum cost function and the optimal control policy were obtained. Liu *et al.* [88] transformed the time-delay optimal tracking control problem with disturbances into a zero-sum game problem. An ADP-based H_∞ tracking control method was proposed. A dual event-triggered constrained control scheme based on DHP [89] was used to solve the zero-sum game problem and was eventually applied to the F-16 aircraft system. A disturbance-based neural network was added to the action-critic structure by Zhong *et al.* [90]. They relaxed the requirement for system information by defining a new type of the performance index. This approach extended the applicability of the ADP algorithm and was the first implementation of model-free globalized dual heuristic programming (GDHP).

2) Continuous-Time Systems: Consider a class of continuous-time nonlinear systems with external disturbances

$$\begin{cases} \dot{x}(t) = f(x(t)) + g(x(t))u(t) + h(x(t))\upsilon(t) \\ y(t) = Z(x(t)) \end{cases} \tag{66}$$

where $\upsilon(t) \in \mathbb{R}^q$ is the disturbance input, $y(t) \in \mathbb{R}^p$ is the objective output, and $h(\cdot)$ is differentiable with respect to its arguments.

In the design process of the nonlinear disturbance rejection, we should find a feedback control law $u(x)$ such that the closed-loop system is asymptotically stable and has an $\mathcal{L}_2$-gain no larger than ℓ. That is

$$\int_0^\infty \Big[\|Z(x(\tau))\|^2 + u^T R u \Big] d\tau \le \ell^2 \int_0^\infty \upsilon^T P \upsilon \, d\tau \tag{67}$$

where $\|Z(x(\tau))\|^2 = x^T(\tau)Qx(\tau)$. Note that the solution of the H_∞ control problem is the saddle point of zero-sum game theory and is denoted as a pair of laws (u^*, υ^*), where u^* and υ^* are the optimal control and the worst-case disturbance, respectively.

We define the infinite horizon cost function as follows:

$$V(x,u,\upsilon) = \int_t^{\infty} U(x(\tau),u(\tau),\upsilon(\tau))d\tau \tag{68}$$

where $U(x,u,\upsilon) = x^T Qx + u^T Ru - \ell^2 \upsilon^T P\upsilon$. As for the continuous-time case, the objective is to find the feedback saddle point solution (u^*, υ^*), such that the Nash condition

$$\begin{aligned} V^*(x_0) &= \min_u \max_\upsilon V(x_0,u,\upsilon) \\ &= \max_\upsilon \min_u V(x_0,u,\upsilon) \end{aligned} \tag{69}$$

holds, where $V^*(x_0)$ is the optimal cost. If the cost function is continuously differentiable, its infinitesimal version is the nonlinear Lyapunov equation

$$0 = U(x,u,\upsilon) + (\nabla V(x))^T (f + gu + h\upsilon) \tag{70}$$

with $V(0) = 0$. The Hamiltonian is defined as

$$\begin{aligned} H(x,u,\upsilon,\nabla V(x)) = {} & U(x,u,\upsilon) \\ & + (\nabla V(x))^T (f + gu + h\upsilon). \end{aligned} \tag{71}$$

According to Bellman's optimality principle, the optimal cost function should satisfy the HJI equation

$$\min_u \max_\upsilon H(x,u,\upsilon,\nabla V^*(x)) = 0. \tag{72}$$

Then, we obtain the optimal control law and the worst-case disturbance law as

$$\begin{cases} u^*(x) = -\dfrac{1}{2} R^{-1} g^T(x) \nabla V^*(x) \\ \upsilon^*(x) = \dfrac{1}{2\ell^2} P^{-1} h^T(x) \nabla V^*(x). \end{cases} \tag{73}$$

In practical applications, the exact system dynamics are often difficult to obtain. The identification method can also produce unpredictable errors. For continuous-time unknown nonlinear zero-sum game problems, Zhu *et al.* [91] proposed an iterative ADP method by efficiently using online data to train the neural network. In [92], a novel distributed H_∞ optimal tracking control scheme was designed for a class of physically interconnected large-scale nonlinear systems in the presence of the strict-feedback form, the external disturbance, and saturating actuators.

C. Game Design With ADP

Modern control systems are becoming more and more complex with many decision makers, who compete and cooperate with each other. As an essential theory for multiple participants to find optimal solutions, game theory is also increasingly studied in the field of control. In accordance with the cooperation pattern among the players, it can be divided into zero-sum and nonzero-sum games, or non-cooperative and cooperative games. In the zero-sum game, the players of the game are not cooperative. However, in a nonzero-sum game, there is a possibility of cooperation among the players so that each of them gets very high performance. Similarly, game theory can be combined with ADP techniques to solve optimal control problems. With the rapid development of iterative ADP, a lot of new methods have been emerged to deal with games for N players [21], [93]–[99].

1) Discrete-Time Systems: Consider a class of discrete-time systems with N players

$$x(k+1) = f(x(k)) + \sum_{j=1}^{N} g_j(x(k))u_j(k) \tag{74}$$

where $x \in \mathbb{R}^n$ is the state vector and $u_j \in \mathbb{R}^{m_j}$ with $j = 1,2,\ldots,$ N is the control input. $f(\cdot) \in \mathbb{R}^{n\times n}$ and $g_j(\cdot) \in \mathbb{R}^{n\times m_j}$ are unknown system matrices. Since there are N players, they influence each other through the system state. We define the set $\mathbb{N}^+ = \{1,2,\ldots\}$ and the complementary set of the player i is $u_{-i} = \{u_j, j \in \mathbb{N}, j \neq i\}$.

The cost function is defined as

$$V_i(x(k)) = \sum_{k}^{\infty} U_i(x(k),u_i(k),u_{-i}(k)) \tag{75}$$

where the utility function is $U_i(x(k),u_i(k),u_{-i}(k)) = x^T(k)\times$ $Q_i x(k) + \sum_{j=1}^{N} u_j^T(k) R_{ij} u_j(k)$. Q_i and R_{ij} are symmetric matrices with appropriate dimensions.

The optimal cost functions are given as

$$V_i^*(x(k)) = \min_{u_i} \sum_{k}^{\infty} U_i(x(k),u_i(k),u_{-i}(k)) \tag{76}$$

which is known as the discrete-time HJB equation. Then, we can obtain the optimal control law

$$u_i^*(k) = -\frac{1}{2} R_{ii}^{-1} g_i^T(x(k)) \nabla V_i^*(x(k+1)). \tag{77}$$

Zhang *et al.* [21] combined game theory and the PI algorithm to solve the multiplayer zero-sum game problem based on ADHDP. This method not only ensured the system to achieve stability but also minimized the performance index function for each player. Song *et al.* [93] divided the off-policy N-coupled Hamilton-Jacobi (HJ) equations into an unknown parameter part and a system operating data part. In this way, the HJ equation can be solved without the system dynamics. Therefore, this approach was very effective for solving multiplayer non-zero-sum game problems with unknown system dynamics. For the domain shift problem, Raghavan *et al.* [94] compensated for the optimal desired shift by constructing a zero-sum game and proposed a direct error-driven learning scheme.

2) Continuous-Time Systems: Consider the following continuous-time systems with N players:

$$\dot{x}(t) = f(x(t)) + \sum_{j=1}^{N} g_j(x(t))u_j(t) \tag{78}$$

where u_j with $j = 0,1,\ldots,N$ represents the control input. Then, we define the cost function as

$$V_k(x,u_1,\ldots,u_N) = \int_0^{\infty} \left(Q_k(x) + \sum_{j=1}^{N} u_j^T R_{kj} u_j \right) d\tau \tag{79}$$

where $Q_k(x)$ is a positive definite function and R_{kj} represents a positive definite matrix with appropriate dimensions.

Assuming that the cost function is continuously differentiable, the Hamiltonian associated with the kth player is defined as

$$H_k(x,V_k,u_1,\dots,u_N)= Q_k(x)+\sum_{j=1}^{N}u_j^T R_{kj}u_j + (\nabla V_k)^T\left(f(x)+\sum_{j=1}^{n}g_j(x)u_j\right). \quad (80)$$

The optimal cost function V_k^* satisfies

$$0=\min_{u_k} H_k(x,V_k^*,u_1,\dots,u_N) \quad (81)$$

and the optimal control law can be obtained by

$$u_k^* = -\frac{1}{2}R_{kk}^{-1}g_k^T(x)\nabla V_k^*. \quad (82)$$

Inspired by zero-sum and nonzero-sum game theory, Lv and Ren [98] proposed a solution for the multiplayer mixed-zero-sum nonlinear games. They defined two value functions containing performance indicators for zero-sum games and nonzero-sum games, respectively. The optimal strategy of each player was obtained without using the action network and the stability of the system was proved. In addition, Zhang *et al*. [99] developed a novel near-optimal control scheme for unknown nonlinear nonzero-sum differential games via the event-based ADP algorithm.

D. Other Advanced Control Methods With ADP

With the development of ADP technology, more and more advanced control methods have been improved. This section shows the application of ADP techniques in decentralized, distributed, and multi-agent systems. Meanwhile, the research progress related to the ADP/RL technique in the field of model predictive control (MPC) is displayed.

Modern control systems usually consist of several subsystems with essential interconnections. It is difficult to analyze large-scale systems by using classical centralized control techniques. Therefore, using decentralized or distributed control strategies is usually preferred to solve optimal control problems for several subsystems. Yang *et al*. [100], [101] not only studied the decentralized stability problem subject to asymmetric constraints, but also transformed the decentralized control problem into a set of optimal control problems by introducing discounted cost functions in the auxiliary subsystems. Tong *et al*. [102] developed an adaptive fuzzy decentralized control method for optimal control problems of large-scale nonlinear systems with strict-feedback form. They proposed two controllers, i.e., a feedforward controller and a feedback controller, to ensure that the tracking error of the closed-loop system converges to a small range. Without using the dynamic matrix of all subsystems, Song *et al*. [103] developed a novel parallel PI algorithm to implement the decentralized sliding mode control scheme.

In [104], taking the unknown discrete-time system dynamics into account, a local Q-function-based ADP method was introduced to address the optimal consensus control problem. Besides, a distributed PI technique was developed by the defined local Q function, which was proved to converge to the solutions of the coupled HJB equations. Fu *et al*. [105] developed a distributed optimal observer for the discrete-time nonlinear active leader with unknown dynamics. It is worth mentioning that the design of the distributed optimal observer based on ADP was developed via the action-critic framework. For the continuous-time distributed system, due to the limited transmission rate of communication channels and the limited bandwidth in some shared communication networks, time delay is an inescapable factor when dealing with the consensus problem. Therefore, in [106], for high-order integrator systems with matched external disturbances, the fixed-time leader-follower consensus problem was coped with by constructing the distributed observer.

Jiang *et al*. [107] estimated the leader's state and dynamics through an adaptive distributed observer, and used a model-state-input structure to solve the regulation equations of each follower. In addition, the stability of the system was analyzed independently. In [108], Sargolzaei *et al*. introduced a Lyapunov-based method, which reduced false-data-injection attacks in real time for a centralized multi-agent system with additive disturbances and input delays. Besides, the condition of the persistence of excitation was hard to verify. Huang *et al*. [109] redesigned the updating laws of the action and critic components to ensure the stability of the system by introducing the persistence of excitation and additional constraints. In addition, the study of tracking control of multi-agent systems has attracted significant attention due to its broad background of applications. For example, Gao *et al*. [110] first integrated ADP with the internal model principle to investigate the problem of cooperative adaptive optimal tracking control. A distributed control policy based on the data-driven technique was put forward for the leader model with external disturbances. Furthermore, the stability of its closed-loop system was also demonstrated.

MPC methods mainly solve optimal control problems with constraints [111]–[118]. There is a very similar theoretical scheme between ADP and MPC. The core of the two methods is to solve the optimal control problem and obtain the corresponding control policy. Furthermore, the control policy should be able to ensure stability. Therefore, the combination of MPC and ADP is a promising and important direction. In [112], Bertsekas pointed out the relationship between MPC and ADP. The core idea and mathematical essence of them were proposed based on PI. Dual-mode MPC has been combined with the action-critic structure to improve the performance and guarantee stability [113]. Based on these, Hu *et al.* [114] introduced a model predictive ADP method for path planning of unmanned ground vehicles at the road intersection. RL has been widely used in feedback control problems [115]. In general, the closed-loop stability with MPC is guaranteed and various MPC strategies have been proposed. However, the performance of MPC and its stability guarantee are limited by an accurate model of the system. Accurate system models are difficult to obtain in real control systems. Generally, states and actions are continuous and it is almost impossible to represent them accurately. Therefore, function approximation tools must be used [116]. Several studies combined the advantages of RL with MPC to solve optimal control problems and generated a new field [117]. Zanon and Gros [118] proposed the combination of RL and MPC to exploit advantages of both methods and then obtained an optimal and safe controller. Meanwhile, it ensured the robustness of MPC based on RL. Subsequently, the data-driven MPC using RL has become an effective approach [119].

V. Boosting ADP via Data Utilization and RL

The concept of RL appeared earlier than ADP. The work of psychologist Skinner and his followers studied how animals learn to change their behaviors according to the result of reward and punishment. The latest work in the field of RL still uses the traditional reward "r" instead of the utility function "U". RL emphasizes immediate reward over the known utility function. Although the focus of ADP is different from RL and the work is relatively independent, the ideas of many methods show that they have common roots. Werbos first combined RL with DP to build a framework that approximates the Bellman equation and proposed HDP in the 1970s. The original proposition of this approach was essentially the same as the formulation of TD in RL [6]. Similarly, ADHDP and Q-learning both employed the state-action function to evaluate the current policy [10]. Overall, ADP/RL is a class of algorithms obtained from solving optimal control problems by approximation methods.

Markov decision process (MDP) is a mathematical framework for obtaining the optimal decision in stochastic dynamic systems. As a key theory of RL, almost all RL problems can be modeled as MDPs. In this paper, MDP is denoted as follows:

$$M = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle \tag{83}$$

where $\mathcal{S}$ is the state set of the environment, $\mathcal{A}$ is the action set, $\mathcal{P}$ is the state transition probability, $\mathcal{R}$ is the reward set, and $\gamma \in (0,1]$ is the discount factor. The agent (often called controller in control theory) chooses actions to generate a trajectory sequence $\tau = \{s_0, a_0, r_0, s_1, a_1, r_1, \ldots\}$. In RL, the goal is to find the optimal policy that maximizes rewards or minimizes penalties for the agent to interact with the environment.

In the early stages of ADP/RL, the theoretical and algorithmic progress was slow due to the limitations of hardware facilities and system information. The development of system identification techniques has made it available to model nonlinear systems using data-driven methods, thereby opening up a new era of research [6], [120]–[128]. In [6], Lewis and Liu illustrated the contribution of stochastic encoder-decoder predictor and principal component analysis in modeling the world through a brain-like approach, as well as emphasized the importance of neural networks. Some model-based approaches have shown promising results. Lee and Lee [122] defined this type of methods as J-learning (based on the value function). The Bellman optimality equation can be expressed as

$$V^*(s_t) = \min_{a_t} \left\{ r(s_t, a_t) + \gamma V^*(s_{t+1}) \right\}. \tag{84}$$

Pang and Jiang [123] used the model-based method to discuss the robustness of PI for the LQR problem, and proposed an off-policy optimistic least-squares PI algorithm. They exploited the dynamical information of the system in the derivation process and incorporated the stochastic perturbations. Lu et al. [124] demonstrated the stability of closed-loop systems using optimal parallel controllers with augmented performance index functions for tracking control. They extended the practical problems to the virtual space through parallel system theory, and used methods such as neural networks to model systems and achieve optimal control.

However, this model-based learning approach can only be effective in the state space based on empirical information. The calculated control actions and performance predictions are constrained by the amount of information. Different from the model-based learning method, Q-learning proposed by Warkins and Dayan [125] used the Q function to represent the value of the action in the current state. This type of function already contains information about the system and the utility function. Compared with J-learning, it is easier to obtain control policies by using Q-learning, especially for unknown nonlinear systems. The Bellman optimality equation can be expressed as

$$Q^*(s_t, a_t) = r(s_t, a_t, s_{t+1}) + \gamma \min_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}). \tag{85}$$

Note that the above formula is described for deterministic systems. Li et al. [126] solved the optimal switching problem of autonomous subsystems and analyzed the boundedness of the approximation error in the iterative process. Jiang et al. [127] used Q-learning to improve the convergence speed of optimal policies for path planning and obstacle avoidance problems. In [95], a new off-policy model-free approach was used to study the networked multi-player game. At the same time, they achieved optimal control in systems with network-induced delays and demonstrated the convergence of the algorithm. In addition, Peng et al. [96] proposed an internal reinforce Q-learning scheme, and analyzed the convergence and system stability related to the iterative algorithm. Based on local information from neighbors, they designed a special internal reward signal to enhance the agent's ability to receive long-term information. The model-free idea applied in the field of control is only the tip of the iceberg.

TD is an RL algorithm that can learn directly from the environment without requiring the complete trajectory sequence. Sarsa and Q-learning are two classic TD algorithms. The Sarsa improves and evaluates the same algorithm (on-policy). However, Q-learning uses data sampled from other policies to improve the target policy (off-policy). Next, we introduce two accelerated methods that can be applied to TD algorithms.

The first method is experience replay which is mainly used to overcome the problems of correlated data and non-stationary distribution. It can improve data utilization efficiency [129], [130]. Pieters and Wiering [131] proposed an algorithm combining experience replay with Q-learning. The simulation results showed that the performance of the algorithm was significantly improved over the traditional Q-learning algorithm. Experience replay technique is not only used in Q-learning but also can be combined with other deep RL algorithms, which have achieved good performance in improving convergence speed and data utilization efficiency [132]. Many scholars in the field of control are inspired to combine ADP algorithm with experience replay to improve the performance of the algorithm. For discrete-time nonlinear systems, Luo *et al*. [133] designed a model-free optimal tracking controller by using policy gradient ADP designs with experience replay. It was realized based on the action-critic structure, which was applied to approximate the iterative Q function and the iterative control policy. The convergence of the iterative algorithm was established through theoretical analysis.

The second method is called eligibility traces. The traditional Q-learning is the case with only one-step estimate. If more information on traces is considered, updating the policy will be more efficacious [134]. The eligibility traces method can combine multi-step information to update unknown parameters. Eligibility traces were first introduced into the TD learning process to form an efficient learning algorithm named TD(λ) in [135]. Considering the direction of the trace, there are forward view and backward view, respectively. Although the expressions of two algorithms are different, their intrinsic essences are the same. In engineering, backward view is generally adopted for the convenience of calculation. Inspired by the field of RL, many scholars combine ADP with both forward view and backward view of eligibility traces. Compared with the traditional ADP algorithms, the performance of these algorithms has been significantly improved [136]. Al-Dabooni and Wunsch [137] proposed a forward view ADHDP(λ) algorithm by combining ADHDP with the eligibility traces and proved the UUB under certain conditions. Ye *et al*. [138] proposed a more accurate and faster algorithm by introducing backward view eligibility traces into GDHP. Meanwhile, the superiority of computational efficiency was verified by the simulation analysis.

In addition, inverse RL [139], [140] has received extensive attention in academia in recent years. This theory is able to solve inverse problems in control systems, machine learning, and optimization. Unlike methods that directly map from states to control inputs or use system identification to learn control policies, inverse RL methods attempt to reconstruct a more adaptive reward function. This reward function prevents small changes in the environment from making the policy unusable. Lian *et al*. [141] used the inverse RL method to solve the two-person zero-sum game problem, and established two algorithms according to whether the model is used or not. Overall, RL has achieved remarkable success for some complex problems [2]. RL has also attracted a lot of attention from a control point of view due to the model-free property and interaction with real-world scenarios. With the application of RL algorithms in the control field, advanced methods based on learning and environmental interaction will demonstrate more powerful capabilities in future works.

VI. Typical Applications of ADP and RL

Compared with other optimal control methods, ADP has significant advantages in dealing with complex nonlinear systems. Due to the strong ability, ADP is widely used in many fields such as wastewater treatment, smart power grid, intelligent transportation, aerospace, aircraft, robotics, and logistics.

A. Wastewater Treatment Applications

The control of the wastewater treatment process is a typical complex nonlinear control problem, and it is also one of the difficulties in the field of process control. Accompanied by a large number of interferences, biochemical reaction mechanisms are very complex. There are many factors that can influence the effect of wastewater treatment, such as the dissolved oxygen concentration and the nitrate concentration. A large part of research is based on the Benchmark Simulation Model No.1 (BSM1) platform for verification. The goal of designing the controller is to reduce energy consumption and cost as much as possible to ensure the effluent quality meets the national discharge standard and the stable operation of the device. The design framework for control of wastewater treatment plants is shown in Fig. 3.

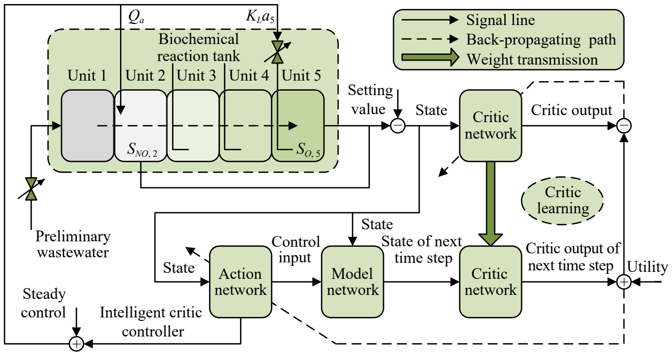


Fig. 3. The design framework for control of wastewater treatment plants.

Control of a single variable has been considered. For example, the online ADP scheme was proposed in [142] by using the echo state network as the function approximation tool. The high-performance control of dissolved oxygen variable in the wastewater treatment plant was realized.

To improve the efficiency of wastewater treatment, many scholars consider both the dissolved oxygen and nitrogen concentration. For example, Wang *et al*. [67] combined the DHP algorithm and the event-triggered mechanism to improve resource utilization and applied it to multi-variable tracking control of wastewater treatment. By using PI and the experience replay mechanism, Yang *et al*. [129] proposed a dynamic priority policy gradient ADP method and applied it to solve multi-variable control of wastewater treatment without the system model.

In the process of wastewater treatment, the setpoint of operating variables is generally set by manual experience. Considering the uncertain environment and disturbance factors, manual experience is often difficult to adapt to different industrial conditions, and it is difficult to balance energy consumption and water quality during operation. Many scholars have studied the optimization of the wastewater treatment process. Qiao *et al*. [143] developed an online optimization control method, which not only met the requirements of effluent water quality, but also reduced the operating cost of the system. For the setpoint of dissolved oxygen, a model-free RL algorithm [144] was proposed that could learn autonomously and actively adjust the setpoint of dissolved oxygen.

B. Power System Applications

Power system is a kind of complex nonlinear plants with multiple variables. The emergence of smart grid has opened up a new direction of power systems. The smart grid design includes renewable energy generation, transmission, storage, distribution, and optimization of household appliances, and so on.

Recently, the ADP/RL algorithm has been widely used in the field of the smart grid due to its advantages. An ADHDP method was applied to solve the residential energy scheduling problem [145], which effectively improves the power consumption efficiency. For multi-battery energy storage systems with time-varying characteristics, a new ADP-based algorithm was proposed in [146]. The robust stabilization of mismatched nonlinear systems was achieved by combining auxiliary systems and policy learning techniques under the condition of dynamic uncertainties [83]. Experimental verification was carried out on a power system. An adaptive optimal data-driven control method was presented based on ADP/RL for three-phase grid-connected inverter of the virtual synchronous generator [147]. To ensure the stable operation of smart grids with load variations and multiple renewable generations, a robust intelligent algorithm was proposed in [148]. It utilized a neural identifier to reconstruct the unknown dynamical system and derived approximate optimal control and worst-case disturbance laws. Wang *et al.* [22] proposed an ADP method with augmented terms based on the GrHDP framework. They constructed new weight updating rules by adding adjustable parameters and successfully applied them to a large power system.

C. Other Applications

The ADP method has also been applied to other fields such as intelligent transportation [149], [150], robotics [7], [51], [127], [151], aerospace [152], [153], smart homes [154], [155], and cyber security [156]–[160], among others. Liu *et al.* [149] proposed a distributed computing method to implement switch-based ADP and verified the effectiveness of the method by using two cases of urban traffic and architecture. The method divided the system into multiple agents. To avoid switching policy conflicts, a heuristic algorithm was proposed based on consensus dynamics and Nash equilibrium. Wen *et al.* [151] combined ADP with RL to propose a direct online HDP approach for knee robot control and clinical application in human subjects. For the optimal attitude-tracking problem of hypersonic vehicles, Han *et al.* [152] and Zhao *et al.* [153] developed a novel PI algorithm and an observation-based RL framework, respectively, which ensured the system stability in the presence of random disturbances. Wei *et al.* [154] proposed a deep RL method to control the air conditioning system by recognizing facial expression information to improve the work efficiency of employees. Hosseinloo *et al.* [155] established an event-based microclimate control algorithm to achieve an optimal balance between energy consumption and occupant comfort. With the widespread application of cyber-physical systems, their security issues have received wide attention. Nguyen and Reddi [156] provided a very comprehensive survey of RL technology routes for cyber security and discussed future research directions. For nonlinear discrete-time systems with event-triggered [157] and stochastic communication protocols [158], Wang *et al.* constructed different action-critic frameworks and discussed the boundedness of the error and the stability of the system based on Lyapunov theory, respectively. More and more ADP-based methods [159], [160] are focusing on improving cyber security. With the rapid development of ADP/RL, its applications will be more extensive.

VII. Summary and Prospect

ADP and RL have made significant progress in theoretical research and practical applications, showing great potential in future tasks. This paper explores the theoretical work and application scenarios by analyzing discrete-time and continuous-time systems, focusing on developing advanced intelligent learning and control. With the current complex system environment and tasks, there are still many theoretical and algorithmic problems that have not yet been solved. Through the present analysis of ADP, this paper concludes several essential directions.

1) Most of the current ADP schemes assume that the function approximation process is exact. However, with the increase of the number of network layers and iterations, the approximation error caused by the function approximator can not be ignored. In the actual iterative process, each step of the function approximator results in an approximation error that propagates to the next iteration. In other words, these approximation errors may change in future iterations, leading to the emergence of a "resonance" type phenomenon, and affecting the reliability of the solution. Therefore, both theoretical and practical applications of ADP need to consider the convergence of ADP algorithms in the presence of approximation errors in policy evaluation and policy improvement.

2) The ADP approach currently addresses mostly systems with low-dimensional states and controls. There is no effective solution for high-dimensional, continuous state and control spaces in real complex systems. With the development of RL and even deep RL, the optimal regulation and trajectory tracking for high-dimensional systems are possible using big data technology. It is important to propose ADP methods with fast convergence and low computational complexity by introducing different forms of relaxation factors.

3) It is of great importance to utilize advanced network technologies to decrease communication traffic and prolong device lifespan. The round-robin protocol, the try-once-discard protocol, the stochastic communication protocol, and the event-triggered protocol are essential in improving performance and saving resources. Based on these protocols, the combination of the ADP technology with decentralized control, robust control, and MPC is crucial in achieving optimal control while minimizing resource consumption.

4) In recent years, the study of brain science and brain-like intelligence has attracted significant interest from researchers worldwide. The optimality theory is closely related to the

study of understanding brain intelligence. Most organisms in nature want to conserve limited resources and achieve their goals in parallel optimally. It is important to consider brain-like intelligence to extend ADP and attain optimal decision and intelligent control of complex systems in an online method. To ensure the stability, convergence, optimality, and robustness of the brain-like intelligence algorithms for ADP, it still requires efforts of a large number of scholars.

5) The field of ADP has a wealth of results that can guide many systems in a theoretical sense to achieve optimal objectives. In practice, however, for a large number of nonlinear systems, abrupt changes in control inputs and the construction of dynamical systems are extremely challenging. Parallel control can be seen as a virtual reality interactive control method. It reconstructs the actual system based on the real input and output data. By combining ADP with parallel control, the control strategy will be greatly improved for real physical systems in the future.

References

[1] P. McCorduck and C. Cfe, *Machines Who Think: A Personal Inquiry Into the History and Prospects of Artificial Intelligence*. 2nd ed. New York, USA: CRC Press, 2004.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[3] K. Doya, H. Kimura, and M. Kawato, "Neural mechanisms of learning and control," *IEEE Control Syst. Mag.*, vol. 21, no. 4, pp. 42–54, Aug. 2001.

[4] D. Wang, M. Ha, and M. Zhao, "The intelligent critic framework for advanced optimal control," *Artif. Intell. Rev.*, vol. 55, no. 1, pp. 1–22, Jan. 2022.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, USA: The MIT Press, 2018.

[6] F. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, USA: Wiley, 2013.

[7] L. Kong, W. He, C. Yang, and C. Sun, "Robust neurooptimal control for a robot via adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2584–2594, Jun. 2021.

[8] R. E. Bellman, *Dynamic Programming*. Princeton, USA: Princeton University Press, 1957.

[9] F. Lewis, D. L. Vrabie, and V. L. Syrmos, *Optimal Control*. 3rd ed. Hoboken, USA: Wiley, 2012.

[10] P. Werbos, "Advanced forecasting methods for global crisis warning and models of intelligence," *Gen. Syst.*, vol. 22, pp. 25–38, Jan. 1977.

[11] D. Liu, S. Xue, B. Zhao, B. Luo, and Q. Wei, "Adaptive dynamic programming for control: A survey and recent advances," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 51, no. 1, pp. 142–160, Jan. 2021.

[12] Z.-P. Jiang and Y. Jiang, "Robust adaptive dynamic programming for linear and nonlinear systems: An overview," *Eur. J. Control*, vol. 19, no. 5, pp. 417–425, Sept. 2013.

[13] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming With Applications in Optimal Control*. Cham, Switzerland: Springer, 2017.

[14] M. Ha, D. Wang, and D. Liu, "Generalized value iteration for discounted optimal control with stability analysis," *Syst. Control Lett.*, vol. 147, p. 104847, Jan. 2021.

[15] D. Liu, Q. Wei, and P. Yan, "Generalized policy iteration adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 45, no. 12, pp. 1577–1591, Dec. 2015.

[16] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, UK: Springer, 2013.

[17] D. Wang, H. He, and D. Liu, "Adaptive critic nonlinear robust control: A survey," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3429–3451, Oct. 2017.

[18] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[19] F. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst. Mag.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

[20] A. Al-Tamimi, F. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst., Man, Cybern., Part B: Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[21] H. Zhang, H. Jiang, C. Luo, and G. Xiao, "Discrete-time nonzero-sum games for multiplayer using policy-iteration-based adaptive dynamic programming algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3331–3340, Oct. 2017.

[22] X. Wang, D. Ding, X. Ge, and Q.-L. Han, "Supplementary control for quantized discrete-time nonlinear systems under goal representation heuristic dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022. DOI: 10.1109/TNNLS.2022.3201521

[23] D. Wang, M. Zhao, M. Ha, and J. Ren, "Neural optimal tracking control of constrained nonaffine systems with a wastewater treatment application," *Neural Netw.*, vol. 143, pp. 121–132, Nov. 2021.

[24] D. Wang, M. Zhao, and J. Qiao, "Intelligent optimal tracking with asymmetric constraints of a nonlinear wastewater treatment system," *Int. J. Robust Nonlinear Control*, vol. 31, no. 14, pp. 6773–6787, Sept. 2021.

[25] J. Xu, J. Wang, J. Rao, Y. Zhong, and H. Wang, "Adaptive dynamic programming for optimal control of discrete-time nonlinear system with state constraints based on control barrier function," *Int. J. Robust Nonlinear Control*, vol. 32, no. 6, pp. 3408–3424, Apr. 2022.

[26] H. Li and D. Liu, "Optimal control for discrete-time affine non-linear systems using general value iteration," *IET Control Theory Appl.*, vol. 6, no. 18, pp. 2725–2736, Dec. 2012.

[27] Q. Wei, D. Liu, and H. Lin, "Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 840–853, Mar. 2016.

[28] M. Ha, D. Wang, and D. Liu, "Offline and online adaptive critic control designs with stability guarantee through value iteration," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13262–13274, Dec. 2022.

[29] D. Wang, M. Zhao, M. Ha, and J. Qiao, "Discounted near-optimal regulation of constrained nonlinear systems via generalized value iteration," *Int. J. Robust Nonlinear Control*, vol. 31, no. 17, pp. 8481–8503, Nov. 2021.

[30] C. Kamanchi, R. B. Diddigi, and S. Bhatnagar, "Generalized second-order value iteration in Markov decision processes," *IEEE Trans. Autom. Control*, vol. 67, no. 8, pp. 4241–4247, Aug. 2022.

[31] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Mar. 2014.

[32] Y. Zhu, D. Zhao, and H. He, "Invariant adaptive dynamic programming for discrete-time optimal control," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 50, no. 11, pp. 3959–3971, Nov. 2020.

[33] B. Luo, Y. Yang, H. N. Wu, and T. Huang, "Balancing value iteration and policy iteration for discrete-time control," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 50, no. 11, pp. 3948–3958, Nov. 2020.

[34] A. Heydari, "Stability analysis of optimal adaptive control under value iteration using a stabilizing initial policy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4522–4527, Sept. 2018.

[35] T. Bian and Z.-P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design," *Automatica*, vol. 71, pp. 348–360, Sept. 2016.

[36] H. Zargarzadeh, T. Dierks, and S. Jagannathan, "Optimal control of nonlinear continuous-time systems in strict-feedback form," *IEEE*

Trans. Neural Netw. Learn. Syst., vol. 26, no. 10, pp. 2535–2549, Oct. 2015.

[37] C. Possieri and M. Sassano, "Data-driven policy iteration for nonlinear optimal control problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7365–7376, Oct. 2023.

[38] S. He, H. Fang, M. Zhang, F. Liu, and Z. Ding, "Adaptive optimal control for a class of nonlinear systems: The online policy iteration approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 549–558, Feb. 2020.

[39] Q. Wei, H. Li, X. Yang, and H. He, "Continuous-time distributed policy iteration for multicontroller nonlinear systems," *IEEE Trans. Cybern.*, vol. 51, no. 5, pp. 2372–2383, May 2021.

[40] Q. Wei, Z. Liao, Z. Yang, B. Li, and D. Liu, "Continuous-time time-varying policy iteration," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 4958–4971, Dec. 2020.

[41] T. Bian and Z.-P. Jiang, "Reinforcement learning and adaptive optimal control for continuous-time nonlinear systems: A value iteration approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 2781–2790, Jul. 2022.

[42] R. Moghadam, P. Natarajan, and S. Jagannathan, "Online optimal adaptive control of partially uncertain nonlinear discrete-time systems using multilayer neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4840–4850, Sept. 2022.

[43] C. Li, D. Liu, and D. Wang, "Data-based optimal control for weakly coupled nonlinear systems using policy iteration," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 48, no. 4, pp. 511–521, Apr. 2018.

[44] J. Li, H. Modares, T. Chai, F. Lewis, and L. Xie, "Off-policy reinforcement learning for synchronization in multiagent graphical games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2434–2445, Oct. 2017.

[45] Q. Wei, D. Liu, and Y. Xu, "Neuro-optimal tracking control for a class of discrete-time nonlinear systems via generalized value iteration adaptive dynamic programming approach," *Soft Comput.*, vol. 20, no. 2, pp. 697–706, Feb. 2016.

[46] R. Song and L. Zhu, "Optimal fixed-point tracking control for discrete-time nonlinear systems via ADP," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 657–666, May 2019.

[47] Q. Lin, Q. Wei, and D. Liu, "A novel optimal tracking control scheme for a class of discrete-time nonlinear systems using generalised policy iteration adaptive dynamic programming algorithm," *Int. J. Syst. Sci.*, vol. 48, no. 3, pp. 525–534, May 2017.

[48] B. Kiumarsi and F. Lewis, "Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 140–151, Jan. 2015.

[49] M. Lin, B. Zhao, and D. Liu, "Policy gradient adaptive critic designs for model-free optimal tracking control with experience replay," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 52, no. 6, pp. 3692–3703, Jun. 2022.

[50] C. Li, J. Ding, F. Lewis, and T. Chai, "A novel adaptive dynamic programming based on tracking error for nonlinear discrete-time systems," *Automatica*, vol. 129, p. 109687, Jul. 2021.

[51] M. Ha, D. Wang, and D. Liu, "Discounted iterative adaptive critic designs with novel stability analysis for tracking control," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1262–1272, Jul. 2022.

[52] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4164–4169, Dec. 2016.

[53] C. Chen, H. Modares, K. Xie, F. Lewis, Y. Wan, and S. Xie, "Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4423–4438, Nov. 2019.

[54] Y. Fu, C. Hong, J. Fu, and T. Chai, "Approximate optimal tracking control of nondifferentiable signals for a class of continuous-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4441–4450, Jun. 2022.

[55] S. Mukherjee, H. Bai, and A. Chakrabortty, "Model-based and model-free designs for an extended continuous-time LQR with exogenous inputs," *Syst. Control Lett.*, vol. 154, p. 104983, Aug. 2021.

[56] S. A. A. Rizvi and Z. Lin, "Output feedback Q-learning control for the discrete-time linear quadratic regulator problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1523–1536, May 2019.

[57] D. Wang, J. Ren, M. Ha, and J. Qiao, "System stability of learning-based linear optimal control with general discounted value iteration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6504–6514, Sept. 2023.

[58] S. K. Jha and S. Bhasin, "Adaptive linear quadratic regulator for continuous-time systems with uncertain dynamics," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 3, pp. 833–841, May 2020.

[59] S. A. A. Rizvi and Z. Lin, "Reinforcement learning-based linear quadratic regulation of continuous-time systems using dynamic output feedback," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4670–4679, Nov. 2020.

[60] Y. Jiang, J. Fan, T. Chai, F. Lewis, and J. Li, "Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4607–4620, Oct. 2018.

[61] L. Dong, X. Zhong, C. Sun, and H. He, "Adaptive event-triggered control based on heuristic dynamic programming for nonlinear discrete-time systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1594–1605, Jul. 2017.

[62] M. Ha, D. Wang, and D. Liu, "Event-triggered adaptive critic control design for discrete-time constrained nonlinear systems," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 50, no. 9, pp. 3158–3168, Sept. 2020.

[63] F. Zhao, W. Gao, T. Liu, and Z.-P. Jiang, "Adaptive optimal output regulation of linear discrete-time systems based on event-triggered output-feedback," *Automatica*, vol. 137, p. 110103, Mar. 2022.

[64] Q. Zhao, J. Si, and J. Sun, "Online reinforcement learning control by direct heuristic dynamic programming: From time-driven to event-driven," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 4139–4144, Aug. 2022.

[65] J. Lu, Q. Wei, Y. Liu, T. Zhou, and F.-Y. Wang, "Event-triggered optimal parallel tracking control for discrete-time nonlinear systems," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 52, no. 6, pp. 3772–3784, Jun. 2022.

[66] Q. Wei, J. Lu, T. Zhou, X. Cheng, and F.-Y. Wang, "Event-triggered near-optimal control of discrete-time constrained nonlinear systems with application to a boiler-turbine system," *IEEE Trans. Ind. Inf.*, vol. 18, no. 6, pp. 3926–3935, Jun. 2022.

[67] D. Wang, L. Hu, M. Zhao, and J. Qiao, "Adaptive critic for event-triggered unknown nonlinear optimal tracking design with wastewater treatment applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 3926–3935, Sept. 2023.

[68] B. Luo, Y. Yang, D. Liu, and H.-N. Wu, "Event-triggered optimal control with performance guarantees using adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 76–88, Jan. 2020.

[69] H. Zhang, J. Zhang, Y. Cai, S. Sun, and J. Sun, "Leader-following consensus for a class of nonlinear multiagent systems under event-triggered and edge-event triggered mechanisms," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7643–7654, Aug. 2022.

[70] X. Huo, H. R. Karimi, X. Zhao, B. Wang, and G. Zong, "Adaptive-critic design for decentralized event-triggered control of constrained nonlinear interconnected systems within an identifier-critic framework," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7478–7491, Aug. 2022.

[71] V. Narayanan, H. Modares, and S. Jagannathan, "Event-triggered control of input-affine nonlinear interconnected systems using multiplayer game," *Int. J. Robust Nonlinear Control*, vol. 31, no. 3, pp. 950–970, Oct. 2021.

[72] X. Yang and Q. Wei, "Adaptive critic learning for constrained optimal event-triggered control with discounted cost," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 91–104, Jan. 2021.

[73] M. Bahraini, M. Zanon, A. Colombo, and P. Falcone, "Optimal control design for perturbed constrained networked control systems," *IEEE Control Syst. Lett.*, vol. 5, no. 2, pp. 553–558, Apr. 2021.

[74] M. H. Mamduhi, D. Maity, S. Hirche, J. S. Baras, and K. H. Johansson, "Delay-sensitive joint optimal control and resource management in multiloop networked control systems," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 3, pp. 1093–1106, Sept. 2021.

[75] Y. Jiang, B. Kiumarsi, J. Fan, T. Chai, J. Li, and F. Lewis, "Optimal output regulation of linear discrete-time systems with unknown

dynamics using reinforcement learning," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3147–3156, Jul. 2020.

[76] D. Wang, D. Liu, H. Li, B. Luo, and H. Ma, "An approximate optimal control approach for robust stabilization of a class of discrete-time nonlinear systems with uncertainties," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 46, no. 5, pp. 713–717, May 2016.

[77] J. Li, J. Ding, T. Chai, F. Lewis, and S. Jagannathan, "Adaptive interleaved reinforcement learning: Robust stability of affine nonlinear systems with unknown uncertainty," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 270–280, Jan. 2022.

[78] N. S. Tripathy, I. N. Kar, and K. Paul, "Suboptimal robust stabilization of discrete-time mismatched nonlinear system," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 352–359, Jan. 2018.

[79] D. M. Adhyaru, I. N. Kar, and M. Gopal, "Fixed final time optimal control approach for bounded robust controller design using Hamilton-Jacobi-Bellman solution," *IET Control Theory Appl.*, vol. 3, no. 9, pp. 1183–1195, Sept. 2009.

[80] D. Wang, D. Liu, Q. Zhang, and D. Zhao, "Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 46, no. 11, pp. 1544–1555, Nov. 2016.

[81] D. Wang and X. Xu, "A data-based neural policy learning strategy towards robust tracking control design for uncertain dynamic systems," *Int. J. Syst. Sci.*, vol. 53, no. 8, pp. 1719–1732, Jan. 2022.

[82] X. Yang, H. He, and X. Zhong, "Adaptive dynamic programming for robust regulation and its application to power systems," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5722–5732, Jul. 2018.

[83] D. Wang, "Robust policy learning control of nonlinear plants with case studies for a power system application," *IEEE Trans. Ind. Inf.*, vol. 16, no. 3, pp. 1733–1741, Mar. 2020.

[84] D. Wang, "Intelligent critic control with robustness guarantee of disturbed nonlinear plants," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2740–2748, Jun. 2020.

[85] C. Mu, Y. Zhang, Z. Gao, and C. Sun, "ADP-based robust tracking control for a class of nonlinear systems with unmatched uncertainties," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 50, no. 11, pp. 4056–4067, Nov. 2020.

[86] B. Pang, T. Bian, and Z. P. Jiang, "Robust policy iteration for continuous-time linear quadratic regulation," *IEEE Trans. Autom. Control*, vol. 67, no. 1, pp. 504–511, Jan. 2022.

[87] J. Hou, D. Wang, D. Liu, and Y. Zhang, "Model-free H_∞ optimal tracking control of constrained nonlinear systems via an iterative adaptive learning algorithm," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 50, no. 11, pp. 4097–4108, Nov. 2020.

[88] Y. Liu, H. Zhang, R. Yu, and Z. Xing, "H_∞ tracking control of discrete-time system with delays via data-based adaptive dynamic programming," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 50, no. 11, pp. 4078–4085, Nov. 2020.

[89] D. Wang, L. Hu, M. Zhao, and J. Qiao, "Dual event-triggered constrained control through adaptive critic for discrete-time zero-sum games," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 53, no. 3, pp. 1584–1595, Mar. 2023.

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

[91] Y. Zhu, D. Zhao, and X. Li, "Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 714–725, Mar. 2017.

[92] L. N. Tan, "Distributed H_∞ optimal tracking control for strict-feedback nonlinear large-scale systems with disturbances and saturating actuators," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 50, no. 11, pp. 4719–4731, Nov. 2020.

[93] R. Song, Q. Wei, H. Zhang, and F. Lewis, " Discrete-time non-zero-sum games with completely unknown dynamics," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 2929–2943, Jun. 2021.

[94] K. Raghavan, J. Sarangapani, and V. A. Samaranayake, "A game theoretic approach for addressing domain-shift in big-data," *IEEE Trans. Big Data*, vol. 8, no. 6, pp. 1610–1621, Dec. 2022.

[95] J. Li, Z. Xiao, J. Fan, T. Chai, and F. Lewis, "Off-policy Q-learning: Solving nash equilibrium of multi-player games with network-induced delay and unmeasured state," *Automatica*, vol. 136, p. 110076, Feb. 2022.

[96] Z. Peng, R. Luo, J. Hu, K. Shi, S. K. Nguang, and B. K. Ghosh, "Optimal tracking control of nonlinear multiagent systems using internal reinforce Q-learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 4043–4055, Aug. 2022.

[97] J. Li, J. Ding, T. Chai, and F. Lewis, "Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 4132–4145, Sept. 2020.

[98] 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.

[99] 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.

[100] X. Yang, Y. Zhou, N. Dong, and Q. Wei, "Adaptive critics for decentralized stabilization of constrained-input nonlinear interconnected systems," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 52, no. 7, pp. 4187–4199, Jul. 2022.

[101] X. Yang, Z. Zeng, and Z. Gao, "Decentralized neurocontroller design with critic learning for nonlinear-interconnected systems," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 11672–11685, Nov. 2022.

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

[103] J. Song, L. Y. Huang, H. R. Karimi, Y. Niu, and J. Zhou, "ADP-based security decentralized sliding mode control for partially unknown large-scale systems under injection attacks," *IEEE Trans. Circuits Syst. I: Regul. Pap.*, vol. 67, no. 12, pp. 5290–5301, Dec. 2020.

[104] W. Wang, X. Chen, H. Fu, and M. Wu, "Model-free distributed consensus control based on actor-critic framework for discrete-time nonlinear multiagent systems," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 50, no. 11, pp. 4123–4134, Nov. 2020.

[105] H. Fu, X. Chen, and M. Wu, "Distributed optimal observer design of networked systems via adaptive critic design," *IEEE Trans. Syst.*, *Man*, *Cybern.: Syst.*, vol. 51, no. 11, pp. 6976–6985, Nov. 2021.

[106] Z. Zuo, B. Tian, M. Defoort, and Z. Ding, "Fixed-time consensus tracking for multiagent systems with high-order integrator dynamics," *IEEE Trans. Autom. Control*, vol. 63, no. 2, pp. 563–570, Feb. 2018.

[107] Y. Jiang, J. Fan, W. Gao, T. Chai, and F. Lewis, "Cooperative adaptive optimal output regulation of nonlinear discrete-time multi-agent systems," *Automatica*, vol. 121, p. 109149, Nov. 2020.

[108] A. Sargolzaei, B. C. Allen, C. D. Crane, and W. E. Dixon, "Lyapunov-based control of a nonlinear multiagent system with a time-varying input delay under false-data-injection attacks," *IEEE Trans. Ind. Inf.*, vol. 18, no. 4, pp. 2693–2703, Apr. 2022.

[109] J. Huang, Z. Zhang, F. Cai, and Y. Chen, "Optimized formation control for multi-agent systems based on adaptive dynamic programming without persistence of excitation," *IEEE Control Syst. Lett.*, vol. 6, pp. 1412–1417, Jul. 2022.

[110] W. Gao, Z. Jiang, F. Lewis, and Y. Wang, "Leader-to-formation stability of multiagent systems: An adaptive optimal control approach," *IEEE Trans. Autom. Control*, vol. 63, no. 10, pp. 3581–3587, Oct. 2018.

[111] Y. Wang and S. Boyd, "Fast model predictive control using online optimization," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 2, pp. 267–278, Mar. 2010.

[112] D. P. Bertsekas, "Dynamic programming and suboptimal control: A survey from ADP to MPC," *Eur. J. Control*, vol. 11, no. 4–5, pp. 310–334, Dec. 2005.

[113] L. Beckenbach, P. Osinenko, and S. Streif, "Addressing infinite-horizon optimization in MPC via Q-learning," *IFAC-PapersOnLine*, vol. 51, no. 20, pp. 60–65, Jan. 2018.

[114] C. Hu, L. Zhao, and G. Qu, "Event-triggered model predictive adaptive dynamic programming for road intersection path planning of unmanned ground vehicle," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11228–11243, Nov. 2021.

[115] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *Proc. IEEE Conf. Decision and Control*, Miami, USA, 2018, pp. 6059–6066.

[116] D. Görges, "Relations between model predictive control and reinforcement learning," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 4920–4928, Jul. 2017.

[117] E. Bøhn, S. Gros, S. Moe, and T. A. Johansen, "Optimization of the model predictive control update interval using reinforcement learning," *IFAC-PapersOnLine*, vol. 54, no. 14, pp. 257–262, Sept. 2021.

[118] M. Zanon and S. Gros, "Safe reinforcement learning using robust MPC," *IEEE Trans. Autom. Control*, vol. 66, no. 8, pp. 3638–3652, Aug. 2021.

[119] X. Yang, H. Zhang, Z. Wang, H. Yan, and C. Zhang, "Data-based predictive control via multistep policy gradient reinforcement learning," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 2818–2828, May 2023.

[120] V. Kompella, R. Capobianco, S. Jong, J. Browne, S. Fox, L. Meyers, P. Wurman, and P. Stone, "Reinforcement learning for optimization of COVID-19 mitigation policies," arXiv preprint arXiv: 2010.10560, 2020.

[121] D. Wang, M.-M. Zhao, M.-M. Ha, and J.-F. Qiao, "Intelligent optimal tracking with application verifications via discounted generalized value iteration," *Acta Autom. Sinica*, vol. 48, no. 1, pp. 182–193, Jan. 2022.

[122] J. M. Lee and J. H. Lee, "Approximate dynamic programming-based approaches for input-output data-driven control of nonlinear processes," *Automatica*, vol. 41, no. 7, pp. 1281–1288, Jul. 2005.

[123] B. Pang and Z.-P. Jiang, "Robust reinforcement learning: A case study in linear quadratic regulation," in *Proc. 35th AAAI Conf. Artificial Intelligence*, 2021, pp. 9303–9311.

[124] J. Lu, Q. Wei, and F.-Y. Wang, "Parallel control for optimal tracking via adaptive dynamic programming," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 6, pp. 1662–1674, Nov. 2020.

[125] C. J. C. H. Warkins and P. Dayan, "*Q*-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, May 1992.

[126] X. Li, L. Dong, and C. Sun, "Data-based optimal tracking of autonomous nonlinear switching systems," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 1, pp. 227–238, Jan. 2021.

[127] L. Jiang, H. Huang, and Z. Ding, "Path planning for intelligent robots based on deep Q-learning with experience replay and heuristic knowledge," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 1179–1189, Jul. 2020.

[128] B. Pang, Z.-P. Jiang, and I. Mareels, "Reinforcement learning for adaptive optimal control of continuous-time linear periodic systems," *Automatica*, vol. 118, p. 109035, Aug. 2020.

[129] R. Yang, D. Wang, and J. Qiao, "Policy gradient adaptive critic design with dynamic prioritized experience replay for wastewater treatment process control," *IEEE Trans. Ind. Inf.*, vol. 18, no. 5, pp. 3150–3158, May 2022.

[130] Q. Zhang and D. Zhao, "Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 2874–2885, Aug. 2019.

[131] M. Pieters and M. A. Wiering, "Q-learning with experience replay in a dynamic environment," in *Proc. 2016 IEEE Symp. Series on Computational Intelligence*, Athens, Greece, 2016, pp. 1–8.

[132] W.-C. Jiang, K. S. Hwang, and J.-L. Lin, "An experience replay method based on tree structure for reinforcement learning," *IEEE Trans. Emerg. Top. Comput.*, vol. 9, no. 2, pp. 972–982, Apr.–Jun. 2021.

[133] B. Luo, Y. Yang, and D. Liu, "Adaptive *Q*-learning for data-based optimal output regulation with experience replay," *IEEE Trans. Cybern.*, vol. 48, no. 12, pp. 3337–3348, Dec. 2018.

[134] S. Al-Dabooni and D. C. Wunsch, "Online model-free n-step HDP with stability analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1255–1269, Apr. 2020.

[135] F. J. Pineda, "Mean-field theory for batched TD(λ)," *Neural Comput.*, vol. 9, no. 7, pp. 1403–1419, Oct. 1997.

[136] R. Yousefian and S. Kamalasadan, "Design and real-time implementation of optimal power system wide-area system-centric controller based on temporal difference learning," *IEEE Trans. Ind. Appl.*, vol. 52, no. 1, pp. 395–406, Jan.–Feb. 2016.

[137] S. Al-Dabooni and D. Wunsch, "The boundedness conditions for model-free HDP(λ)," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 1928–1942, Jul. 2019.

[138] J. Ye, Y. Bian, B. Xu, Z. Qin, and M. Hu, "Online optimal control of discrete-time systems based on globalized dual heuristic programming with eligibility traces," in *Proc. 3rd Int. Conf. Industrial Artificial Intelligence*, Shenyang, China, 2021, pp. 1–6.

[139] N. Ab Azar, A. Shahmansoorian, and M. Davoudi, "From inverse optimal control to inverse reinforcement learning: A historical review," *Annu. Rev. Control*, vol. 50, pp. 119–138, Jun. 2020.

[140] S. Adams, T. Cody, and P. A. Beling, "A survey of inverse reinforcement learning," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4307–4346, Feb. 2022.

[141] B. Lian, W. Xue, F. Lewis, and T. Chai, "Inverse reinforcement learning for adversarial apprentice games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 4596–4609, Aug. 2023.

[142] Y.-C. Bo and X. Zhang, "Online adaptive dynamic programming based on echo state networks for dissolved oxygen control," *Appl. Soft Comput. J.*, vol. 62, pp. 830–839, Jan. 2018.

[143] J. Qiao, L. Wang, and H. Han, "Optimal control for wastewater treatment process based on ESN neural network," *CAAI Trans. Intell. Syst.*, vol. 10, no. 6, pp. 831–837, Dec. 2015.

[144] F. Hernandez-del-Olmo, E. Gaudioso, and A. Nevado, "Autonomous adaptive and active tuning up of the dissolved oxygen setpoint in a wastewater treatment plant using reinforcement learning," *IEEE Trans. Syst., Man, Cybern., Part C* (*Appl. Rev.*), vol. 42, no. 5, pp. 768–774, Sept. 2012.

[145] D. Liu, Y. Xu, and X. Liu, "Residential energy scheduling for variable weather solar energy based on adaptive dynamic programming," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 36–46, Jan. 2018.

[146] Y. Zhu, D. Zhao, X. Li, and D. Wang, "Control-limited adaptive dynamic programming for multi-battery energy storage systems," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4235–4244, Jul. 2019.

[147] Z. Wang, Y. Yu, W. Gao, M. Davari, and C. Deng, "Adaptive, optimal, virtual synchronous generator control of three-phase grid-connected inverters under different grid conditions — An adaptive dynamic programming approach," *IEEE Trans. Ind. Inf.*, vol. 18, no. 11, pp. 7388–7399, Nov. 2022.

[148] 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.

[149] D. Liu, S. Baldi, W. Yu, and G. Chen, "On distributed implementation of switch-based adaptive dynamic programming," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 7218–7224, Jul. 2022.

[150] D. Liu, W. Yu, S. Baldi, J. Cao, and W. Huang, "A switching-based adaptive dynamic programming method to optimal traffic signaling," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 50, no. 11, pp. 4160–4170, Nov. 2020.

[151] Y. Wen, J. Si, A. Brandt, X. Gao, and H. Huang, "Online reinforcement learning control for the personalization of a robotic knee prosthesis," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2346–2356, Jun. 2020.

[152] X. Han, Z. Zheng, L. Liu, B. Wang, Z. Cheng, H. Fan, and Y. Wang, "Online policy iteration ADP-based attitude-tracking control for hypersonic vehicles," *Aerosp. Sci. Technol.*, vol. 106, p. 106233, Nov. 2020.

[153] S. Zhao, J. Wang, H. Xu, and B. Wang, "Composite observer-based optimal attitude-tracking control with reinforcement learning for hypersonic vehicles," *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 913–926, Feb. 2023.

[154] Q. Wei, T. Li, and D. Liu, "Learning control for air conditioning systems via human expressions," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7662–7671, Aug. 2021.

[155] A. H. Hosseinloo, A. Ryzhov, A. Bischi, H. Ouerdane, K. Turitsyn, and M. A. Dahleh, "Data-driven control of micro-climate in buildings: An event-triggered reinforcement learning approach," *Appl. Energy*, vol. 277, p. 115451, Nov. 2020.

[156] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber

security," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 3779–3795, Aug. 2023.

[157] X. Wang, D. Ding, X. Ge, and Q.-L. Han, "Neural-network-based control for discrete-time nonlinear systems with denial-of-service attack: The adaptive event-triggered case," *Int. J. Robust Nonlinear Control*, vol. 32, no. 5, pp. 2760–2779, Mar. 2022.

[158] X. Wang, D. Ding, H. Dong, and X.-M. Zhang, "Neural-network-based control for discrete-time nonlinear systems with input saturation under stochastic communication protocol," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 4, pp. 766–778, Apr. 2021.

[159] R. Liu, F. Hao, and H. Yu, "Optimal SINR-based DoS attack scheduling for remote state estimation via adaptive dynamic programming approach," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 51, no. 12, pp. 7622–7632, Dec. 2021.

[160] L. Zhang, Y. Chen, and M. Li, "ADP-based remote secure control for networked control systems under unknown nonlinear attacks in sensors and actuators," *IEEE Trans. Ind. Inf.*, vol. 18, no. 9, pp. 6003–6014, Sept. 2022.

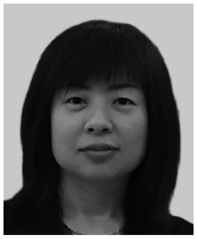
Ding Wang (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from Institute of Automation, Chinese Academy of Sciences, in 2012. He was an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He is currently a Full Professor with the Faculty of Information Technology, Beijing University of Technology.

He has authored or co-authored over 150 journal and conference papers and five monographs. His current research interests include adaptive critic control with industrial applications, reinforcement learning, and intelligent systems. Dr. Wang was selected as a Clarivate Highly Cited Researcher and Elsevier Most Cited Chinese Researcher. He is a Member of IEEE/CAA Journal of Automatica Sinica Early Career Advisory Board. He currently or formerly serves as an Associate Editor of *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *Neural Networks*, *Engineering Applications of Artificial Intelligence*, *International Journal of Robust and Nonlinear Control*, *International Journal of Adaptive Control and Signal Processing*, *Neurocomputing*, and *Acta Automatica Sinica*.

Ning Gao received the M.E. from Beijing University of Technology in 2021. He is currently pursuing the Ph.D. degree in control science and engineering with Beijing University of Technology.

His research interests include adaptive dynamic programming, reinforcement learning with industrial applications, and intelligent systems.

Derong Liu (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Notre Dame in 1994. He was a Staff Fellow with General Motors Research and Development Center, from 1993 to 1995. He was an Assistant Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, from 1995 to 1999. He joined the University of Illinois Chicago in 1999, and became a Full Professor of electrical and computer engineering and of computer science in 2006. He was selected for the "100 Talents Program" by the Chinese Academy of Sciences in 2008, and he served as the Associate Director of the State Key Laboratory of Management and Control for Complex Systems at the Institute of Automation, from 2010 to 2016. He is currently a Chair Professor with the School of System Design and Intelligent Manufacturing, Southern University of Science and Technology. He has published 13 books. He is the Editor-in-Chief of *Artificial Intelligence Review* (Springer). He was the Editor-in-Chief of the *IEEE Transactions on Neural Networks and Learning Systems* from 2010 to 2015. He received the Faculty Early Career Development Award from the National Science Foundation in 1999, the University Scholar Award from University of Illinois from 2006 to 2009, the Overseas Outstanding Young Scholar Award from the National Natural Science Foundation of China in 2008, the Outstanding Achievement Award from Asia Pacific Neural Network Assembly in 2014, the INNS Gabor Award in 2018, the IEEE SMC Society Andrew P. Sage Best Transactions Paper Award in 2018, the IEEE TNNLS Outstanding Paper Award in 2018, the IEEE/CCA Journal Automatica Sinica Hsue-Shen Tsien Paper Award in 2018, and the IEEE CIS Neural Network Pioneer Award in 2022. He has been named a highly cited researcher consecutively from 2017 to 2022 by Clarivate. He is a Fellow of the IEEE, a Fellow of the International Neural Network Society, a Fellow of the International Association of Pattern Recognition, and a Member of Academia Europaea (the Academy of Europe).



Jinna Li (Senior Member, IEEE) received the Ph.D. degree from Northeastern University in 2009. She is a Full Professor at the School of Information and Control Engineering, Liaoning Petrochemical University.

From April 2009 to April 2011, she was a postdoctor with the Lab of Industrial Control Networks and Systems, Shenyang Institute of Automation, Chinese Academy of Sciences. She was a Visiting Scholar with the Energy Research Institute, Nanyang Technological University, Singapore, the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University and the School of Electrical and Electronic Engineering, the University of Manchester, UK in Jun. 2014 to Jun. 2015, Sept. 2015 to Jun. 2016 and Jan. 2017 to Jul. 2017, respectively. Her current research interests include reinforcement learning, neural networks, and optimal operational control.

Frank L. Lewis (Life Fellow, IEEE) obtained the bachelor degree in physics/EE and the MSEE at Rice University, the M.S. degree in aeronautical engineering from Univ. W. Florida, and the Ph.D. degree at Ga. Tech. Fellow, National Academy of Inventors. Fellow IEEE, Fellow IFAC, Fellow AAAS, Fellow European Union Academy of Science, Fellow U.K. Institute of Measurement & Control. PE Texas, U.K. Chartered Engineer. Moncrief-O'Donnell Chair at the University of Texas at Arlington. UTA Charter Distinguished Scholar Professor, UTA Distinguished Teaching Professor.

Lewis is ranked as number 23 in the world of all scientists in the world and 12 in the USA in Electronics and Electrical Engineering by Research.com. Ranked number 5 in the world in the subfield of Industrial Engineering and Automation according to a Stanford University Research Study in 2021. Recognized as a Top 1% Highly Top Cited Researcher by Clarivate Web of Science every year since 2019. Fellow, National Academy of Inventors. 86 511 google citations, h-index 130. He works in feedback control, intelligent systems, reinforcement learning, cooperative control systems, and nonlinear systems. He is author of 8 U.S. patents, numerous journal special issues, 500 journal papers, 20 books, including the textbooks *Optimal Control*, *Aircraft Control*, *Optimal Estimation*, and *Robot Manipulator Control*. He received the Fulbright Research Award, NSF Research Initiation Grant, ASEE *Terman Award*, Int. Neural Network Soc. Gabor Award, U.K. Inst Measurement & Control Honeywell Field Engineering Medal, IEEE Computational Intelligence Society Neural Networks Pioneer Award, AIAA Intelligent Systems Award, AACC Ragazzini Award. He has received over $12M in 100 research grants from NSF, ARO, ONR, AFOSR, DARPA, and USA industry contracts. Helped win the US SBA Tibbets Award in 1996 as Director of the UTA Research Institute SBIR Program.