Adaptive Dynamic Programming for Control: A Survey and Recent Advances

Derong Liu, *Fellow, IEEE*, Shan Xue, Bo Zhao, *Senior Member, IEEE*, Biao Luo, *Senior Member, IEEE*, and Qinglai Wei, *Senior Member, IEEE*

***Abstract*—This article reviews the recent development of adaptive dynamic programming (ADP) with applications in control. First, its applications in optimal regulation are introduced, and some skilled and efficient algorithms are presented. Next, the use of ADP to solve game problems, mainly nonzero-sum game problems, is elaborated. It is followed by applications in large-scale systems. Note that although the functions presented in this article are based on continuous-time systems, various applications of ADP in discrete-time systems are also analyzed. Moreover, in each section, not only some existing techniques are discussed, but also possible directions for future work are pointed out. Finally, some overall prospects for the future are given, followed by conclusions of this article. Through a comprehensive and complete investigation of its applications in many existing fields, this article fully demonstrates that the ADP intelligent control method is promising in today's artificial intelligence era. Furthermore, it also plays a significant role in promoting economic and social development.**

***Index Terms*—Adaptive critic designs (ACDs), adaptive dynamic programming, approximate dynamic programming, intelligent control, learning control, neural dynamic programming, neuro-dynamic programming, optimal control, reinforcement learning (RL).**

I. Introduction

ARTIFICIAL intelligence and machine learning have attracted widespread interests in recent years. Especially in the past ten years, the research on machine learning has developed rapidly, making it one of the most important cutting-edge research fields in artificial intelligence. For decades, there have been multiple classification methods for machine learning according to the emphasis on different aspects. Generally speaking, machine learning is divided into reinforcement learning (RL), supervised learning, and unsupervised learning. In particular, RL is a learning technique based on statistics and dynamic programming. It uses the reward/punishment signal feedback from the environment/system as input, and learns in a "trial-and-error" manner. In the field of control, RL can effectively solve the "curse of dimensionality" problem in dynamic programming, so it is also called adaptive/approximate dynamic programming (ADP). ADP is an intelligent control method that has aroused extensive interest in academia and industry since it was proposed. In order to help this intelligent control technique be well understood, the origin of ADP, its structures, and the development of learning algorithms are described, respectively.

Manuscript received November 27, 2020; accepted December 1, 2020. Date of publication December 24, 2020; date of current version January 12, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0100203; in part by the National Natural Science Foundation of China under Grant 62073085; and in part by the Guangdong Introducing Innovative and Enterpreneurial Teams of "The Pearl River Talent Recruitment Program" under Grant 2019ZT08X340. This article was recommended by Associate Editor K. G. Vamvoudakis. *(Corresponding author: Derong Liu.)*

Derong Liu is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: derong@gdut.edu.cn).

Shan Xue is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: shan.xue0807@foxmail.com).

Bo Zhao is with the School of Systems Science, Beijing Normal University, Beijing 100875, China (e-mail: zhaobo@bnu.edu.cn).

Biao Luo is with the School of Automation, Central South University, Changsha 410083, China, and also with Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: biao.luo@hotmail.com).

Qinglai Wei is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qinglai.wei@ia.ac.cn).

Digital Object Identifier 10.1109/TSMC.2020.3042876

A. Origin of the Name of ADP

The acronym "ADP" stands for either "adaptive dynamic programming" or "approximate dynamic programming." The term adaptive dynamic programming was probably mentioned for the first time in 1975 in The Quarterly Journal of Economics by a paper studying optimal solutions for consuming depletable natural resources [1]. Subsequently, adaptive dynamic programming was formally studied in 1977 for inventory control [2], [3]. Another work in 1976 also mentioned it for fault detection [4]. For almost 20 years since, the study of adaptive dynamic programming continued but rarely for control applications. Until 1995 in a paper by Barto *et al.* [5], they introduced the so-called "adaptive real-time dynamic programming," which was closely related to the topic reviewed in this article. Furthermore, in 2002, Murray *et al.* [6] developed an adaptive dynamic programming algorithm for optimal control of continuous-time affine nonlinear systems, with the complete proof of its main theorem given later in [7].

Dynamic programming has been considered as a useful tool for inventory control problems and researchers have realized in very early days the difficulty of solving dynamic programming problems to obtain exact solutions. Thus, approximate solutions were sought using adaptive dynamic programming [2], [3] or using approximate dynamic programming formulation [8]. In terms of control applications, the term approximate dynamic programming for optimal control was presented by Werbos in 1987 [9].

Back in 1977, Werbos introduced an approach [10] which was later named adaptive critic designs (ACDs) [11]–[15]. Werbos [11] classified ACD into three classes, namely, heuristic dynamic programming (HDP), dual heuristic programming (DHP), and globalized DHP (GDHP). By the time when he published the 1992 book chapter [13], Werbos has used the three terms "ACD," "approximate dynamic programming," and "RL" interchangeably. The main research results in RL can be found in the book by Sutton and Barto [16] and the references cited therein. Even though both RL and ADP/ACD provide approximate solutions to dynamic programming, research in these two directions has been somewhat independent [17] in the past. The most famous algorithms in RL are the temporal difference algorithm [18] and the Q-learning algorithm [19], [20]. Compared to ADP/ACD, the area of RL is more mature and has a vast amount of literature (see [16], [21]–[23]).

Another related popular term is under the acronym "NDP" which stands for "neuro-dynamic programming" [24], [25] or "neural dynamic programming" [26]–[28]. NDP has been used interchangeably with ADP by researchers to describe approximate approaches for dynamic programming problems in control applications [24]–[32]. The very first book in NDP/ADP was published in 1996 by Bertsekas and Tsitsiklis [25] which systematically and fully explained the methodology for optimal control, intelligent control, and operations research.

When considering approximate solutions to optimal control problems, suboptimal control has also been used in the literature for this purpose. Under this term, we can also find some papers studying optimal control problems by approximating dynamic programming solutions, e.g., [33]–[37].

In this article, we will use ADP to represent "adaptive dynamic programming," "approximate dynamic programming," "neuro-dynamic programming," "neural dynamic programming," "ACDs," as well as "RL" [38]–[41].

B. Development of the Structures of ADP

Neural networks (NNs) have been a popular choice of function approximation structures for the implementation of ADP algorithms, even though many other structures can also be used, such as lookup table [42] and fuzzy logic [43]. The present review will consider NNs as a tool for the implementation of ADP algorithms. Among the three classes, HDP, using a critic NN to approximate the value function of a dynamic system, is a structure with less computational burden, and is the most studied by scholars. The critic network in DHP approximates the derivatives of the value function with respect to the state variables [14], [15], [44], [45]. The critic network in GDHP combines HDP and DHP [14], [15], [46]. It approximates the value function and the derivatives of the value function, and is a structure with the largest amount of computation. But its approximation accuracy is high. Adding the control policy to the inputs of the critic network obtains the corresponding action-dependent (AD) forms. In addition, ADHDP is also called Q-learning [11], [13], [47]–[50].

In recent years, He *et al.* [51]–[56] proposed a goal-representation online ADP method, which added a reference network to the traditional critic–actor network. Later, Zhong *et al.* [53] provided a theoretical analysis of this method. Ni *et al.* [54] also compared this method with ADHDP and SARSA learning to explain the performance of this new method. This new ADP structure has been applied to problems, such as maze navigation [54], vehicle pole balance design [55], inverted pendulum balance control [51], and industrial large-scale complex process control [53]. Direct HDP was presented in [57] and the convergence guarantee was provided for this design in [58]. The kernel-ADP structure in [59], which combined the learning and generalization abilities of sparse kernel machines with the approximation abilities of NNs, was proposed and applied to the control of ball–plate and inverted pendulum. It was shown that kernel-HDP and kernel-DHP have better performance than HDP and DHP in both empirical and theoretical aspects. The ADP method based on NN requires manual selection of a group of activation functions. However, these activation functions have no clear physical meaning in actual equipments and there is no unified and intuitive selection method. The quality and suitability of the selected activation functions cannot be known. On the other hand, fuzzy approximators can use the prior knowledge of the equipment to approximate the unknown variables easily and reasonably. Therefore, the controller design method based on the fuzzy hyperbolic model was integrated with adaptive control, and then fuzzy ADP was proposed by Tang *et al.* [43]. Aiming at the large amount of computation in the traditional NN approximation method, a global ADP method was proposed in [60]. This method can make the system achieve global asymptotic stability, which is different from the previous ADP-based control methods.

C. Development of ADP Learning Algorithms

As an important subject of machine learning, iterative algorithms have incomparable advantages in the optimal control of complex systems [61]–[69]. The iterative methods in ADP consist of two components: 1) the policy evaluation component using the critic network and 2) the policy improvement component using the actor network. According to the different implementation methods of the two components, scholars put forward different ADP learning algorithms, in which value iteration (VI) and policy iteration (PI) are the basis of these algorithms. The implementation methods of these two methods are described in detail in [62]–[69]. It is worth mentioning that the success of PI depends on the initial admissible control policy, which makes the system stable in the learning process, and thus has been used by many researchers [70]–[75]. Although VI is not restricted by this condition, the stability of the system cannot be guaranteed in general. Because of this, it is not recommended to carry out the iterative VI algorithm online, which limits its applications in industry [76]–[83]. Wei *et al.* [78] proved that the iterative VI can converge to the optimal value by choosing some positive semidefinite function as the initial value function and the stability of the system can

also be guaranteed. It is a meaningful work to prove that the value function iteration starting from an arbitrary value also guarantees stability.

The convergence proof of VI can be found in [76], [84], and [85]. In [79], VI was used to solve optimal tracking control problems of discrete-time systems. Later, in [83], it is proved that if iteration starts with an initial stable control policy, then any control policy generated by VI will also stabilize the system. Compared with the traditional VI, local VI reduces the computational burden, and its value function and control law can be updated in a subset of the state space rather than the entire state space [80]. Its admissibility, termination, and convergence were analyzed in [81] and [82], respectively.

Murray *et al.* [6] explained that the PI algorithm for continuous-time systems can obtain the optimal solution of the Hamilton–Jacobi–Bellman (HJB) equation theoretically. Later, Liu and Wei [72] first described the stability and convergence of the algorithm in discrete-time systems and provided a specific method to obtain the initial admissible control. However, the NN approximation error was not considered in [72]. Then, it was shown in [86] and [87] that the iterative value function eventually converges to the neighborhood of the optimal solution in the presence of NN approximation errors. The integral RL (IRL) algorithm [88], [89] adopts an integral Bellman equation on the basis of traditional PI, which makes policy evaluation relax the requirement of system dynamics. An improved learning algorithm using policy gradient in policy improvement was adopted in [90] and [91].

Some improved algorithms have been derived from the integration of VI and PI, such as generalized PI algorithm [92], [93], generalized VI [94]–[96], and multistep policy evaluation [97], [98]. VI and PI were used as special cases of generalized PI algorithms. Therefore, most ADP methods can be considered as generalized PI algorithms [92], [93]. The development of generalized PI algorithms for continuous-time systems can be found in [99]. The results of [94] and [96] show that the generalized VI algorithm can start from any positive semidefinite function. Using the multistep scheme for policy evaluation, Luo *et al.* [97], [98] developed a multistep HDP algorithm to realize the tradeoff between PI and VI. In [100], an adaptive RL method was developed. It relaxes the requirement for initial admissible control and speeds up the convergence of the VI by integrating VI and PI with a balancing parameter.

D. Structure and Symbols of This Article

It is noted that most of the existing ADP literature is biased toward discrete-time systems [62], [63], [69], [101], [102]. As an alternative, in this article, continuous-time systems are taken as examples to illustrate the development of ADP in control problems in recent years, and a brief analysis is made for discrete-time systems. Depending on the type of problems addressed, the remainder of this article is arranged as follows. The basic optimal regulation problems are given in Section II, including optimal state regulation in Section II-A, optimal output regulation in Section II-B, and optimal tracking control in Section II-C. The game theory is introduced in Section III, mainly aimed at nonzero-sum games. Later, ADP methods for large-scale systems are described in Section IV. The future perspectives and the conclusions of this article are presented in Sections V and VI, respectively. The nomenclature of this article is described in Table I.

TABLE I
NOMENCLATURE

Notation	Meaning
$\mathbb{R}$	The set of all real numbers
$\mathbb{R}^n$	n-dimensional real vector
$\mathbb{R}^{n\times m}$	$n \times m$ real matrix
x	The state vector of the system
u	The control vector of the system
$F(\cdot), f(\cdot), g(\cdot)$	The system functions
A, B	The system matrices
$L(\cdot)$	The utility functions
Q, R	Positive definite symmetric matrices
∇V	The gradient of V
$\|\cdot\|$	The absolute value
$\|\cdot\|$	The norm
A^{T}	The transpose of matrix or vector A
$\hat{A}$	The estimate of the unknown quantity A
$\widetilde{A}$	The estimation error, i.e., $\widetilde{A} = A - \hat{A}$
α	Constants in the interval $[0, 1]$
$\beta, \eta, \mathcal{D}$	Positive numbers
$\mathcal{W}$	Weight matrix of NN
ϕ	Activation function of NN
ε	Reconstruction error of NN

II. ADP for Optimal Regulation Problems

Optimal regulation problems include optimal state regulation, optimal output regulation, and optimal tracking control. Optimal output regulation and optimal tracking control are more in line with actual engineering needs, and they can be converted into optimal state regulation. In fact, the zero trajectory tracking problem is the regulation problem.

A. ADP for Optimal State Regulation

The optimal state regulator is to keep the state near the equilibrium and to maximize the value function of the system. It is needed for, e.g., temperature control and pressure control in industrial processes. When the state of the dynamic system deviates from the equilibrium state, it is particularly important to design an effective controller to stabilize the system. An equilibrium state can be transformed into the zero state, so the zero state is usually regarded as the system equilibrium state for convenience.

Nonlinear dynamic systems can generally be divided into affine and nonaffine systems, which are described by

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t),\ \ x(0) = x_0 \tag{1}$$

$$\dot{x}(t) = F(x(t), u(t)),\ \ x(0) = x_0 \tag{2}$$

respectively, where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the control, and x_0 is the initial state. Generally, it is assumed that the nonlinear dynamic system is controllable.

The general state regulator problem can be divided into the finite-time state regulator and infinite-time state regulator. The value function V of the finite-time state regulator can be written as

$$V(x(t)) = \int_t^{t_f} L(x(\nu), u(\nu))\mathrm{d}\nu = \int_t^{t_f} (L_1(x(\nu)) + L_2(u(\nu)))\mathrm{d}\nu \quad (3)$$

where $L(x, u) = L_1(x) + L_2(u)$, and $t_f > 0$ is the given terminal time. V considers the behavior of the system from the initial state to the equilibrium point, and is a compromise between terminal error, control energy, and state deviation. When t_f in (3) approaches infinity, i.e.,

$$V(x(t)) = \int_t^{\infty} L(x(\nu), u(\nu))\mathrm{d}\nu = \int_t^{\infty} (L_1(x(\nu)) + L_2(u(\nu)))\mathrm{d}\nu \quad (4)$$

the value function of the infinite-time state regulator is obtained. The optimal control problem of (1/2) can simply be stated as to determine $u(t)$ in order to minimize $V(x_0)$.

Problem 1: For affine/nonaffine systems (1/2) and finite/infinite-time value functions (3/4), the problem of optimal state regulation is to design a learning control structure, and then gradually explores the optimal control function which minimizes the value function and stabilizes the closed-loop systems (1/2).

The Hamiltonian of the systems (1/2) is designed as

$$H(x, u, \nabla V) = L(x, u) + (\nabla V)^{\mathsf{T}}\dot{x}. \quad (5)$$

Based on (5), the HJB equation is presented as

$$\min_u H\big(x, u, \nabla V^*\big) = 0 \quad (6)$$

where V^* is the optimal value of V. Then, the optimal control function can be obtained by

$$u^* = \arg\min_u H\big(x, u, \nabla V^*\big). \quad (7)$$

For affine system (1), when the control energy function $L_2(u)$ is quadratic, i.e., $L_2(u) = u^{\mathsf{T}}Ru$, (7) can further be written as

$$u^* = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)\nabla V^*. \quad (8)$$

The following definition is needed for the PI algorithm.

Definition 1 (Admissible Control) [76]: A control $u(t)$ is admissible with respect to the value function on a compact set $\Omega \in \mathbb{R}^n$ if $u(t)$ is continuous on Ω, $u(0) = 0$, $u(t)$ stabilizes the systems on Ω, and $\forall x_0 \in \Omega$, $V(x_0)$ is finite.

1) PI Algorithm: The PI algorithm for the continuous-time system is shown in Algorithm 1, which is adapted from [69].

Algorithm 1 is an iterative ADP algorithm for optimal control of the continuous-time system (1) with the value function (4) based on PI. It is adapted from algorithms for discrete-time systems in [69], where various iterative ADP algorithms are listed, including VI, PI, and generalized PI.

Algorithm 1 PI for (1) and (4) With $L_2(u) = u^{\mathsf{T}}Ru$

Step 1 Initialization:

$i = 0$.

Select an initial admissible control policy u^0.

Step 2 Evaluation:

$i = i + 1$.

The value function V^i under the control u^{i-1} is obtained according to

$$L(x, u^{i-1}) + (\nabla V)^{i\mathsf{T}}\Big(f(x) + g(x)u^{i-1}\Big) = 0. \quad (9)$$

Step 3 Improvement:

The updated control policy u^i is obtained according to

$$u^i = \arg\min_u H(x, u, \nabla V^i). \quad (10)$$

More specifically,

$$u^i = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)\nabla V^i. \quad (11)$$

Step 4 Judgment:

If preset conditions for convergence are not met, go back to **Step 2**.

Step 5 Stop:

Obtain the optimal control policy $u^* = u^i$ and the optimal value function $V^* = V^i$.

2) Improved Online Technique: The admissible control required for initialization is often difficult to obtain. This has prompted scholars to propose improved learning rules. The following general assumption supports the reinforcement of the learning process by adding an additional stabilizing term [103]–[106].

Assumption 1: Take a continuously differentiable Lyapunov function $V_R(x)$ composed of polynomials for system (1). Assume that $V_R(x)$ satisfies

$$\dot{V}_R(x) = (\nabla V_R(x))^{\mathsf{T}}\dot{x} = (\nabla V_R(x))^{\mathsf{T}}\big(f + gu^*\big) < 0. \quad (12)$$

Suppose there is a positive-definite matrix $S(x)$ such that $(\nabla V_R(x))^{\mathsf{T}}(f + gu^*) = -(\nabla V_R(x))^{\mathsf{T}}S(x)\nabla V_R(x)$ holds.

Design an additional stabilizing term added to the weight updating rules

$$\Xi(x, u) = \begin{cases} 0, & \text{if } (\nabla V_R(x))^{\mathsf{T}}(f(x) + g(x)u) < 0 \\ 1, & \text{otherwise.} \end{cases} \quad (13)$$

Remark 1: The design of (13) is based on the use of the Lyapunov function to judge whether the system is stable or not. That is to say, in the learning process of optimal control, when the dynamic system is found to be unstable, the learning process is induced to follow the negative gradient direction of the derivative of the Lyapunov function. Therefore, this improved online technique can relax the condition of initial admissible control [103], [104].

3) Integral Reinforcement Learning: Traditional policy evaluation requires the knowledge of system dynamics. In the case of unknown internal dynamics, many algorithms cannot be used directly. To deal with this problem, IRL was developed [88], [89], [107], [108].

For time t and time interval T, it can be found that (4) is equivalent to [108]

$$\begin{aligned} V(x(t-T)) &= \int_{t-T}^{t} L(x(\nu), u(\nu))\mathrm{d}\nu \\ &\quad + \int_{t}^{\infty} L(x(\nu), u(\nu))\mathrm{d}\nu \\ &= \int_{t-T}^{t} L(x(\nu), u(\nu))\mathrm{d}\nu + V(x(t)). \end{aligned} \tag{14}$$

Equation (14) is used in the policy evaluation process of IRL.

Remark 2: Because of the physical characteristics of actuators in engineering, it is necessary to explain the situation of input constraints [108]–[116]. The measurement of control energy adopts the general nonquadratic form as

$$\begin{aligned} L_2(u) &= 2\int_0^u \eta_m \mathrm{artanh}\left(\frac{\nu}{\eta_m}\right)^{\mathsf{T}}\mathrm{d}\nu \\ &= 2\sum_{j=1}^{m}\int_0^{u_j} \eta_m \mathrm{artanh}\left(\frac{\nu_j}{\eta_m}\right)\mathrm{d}\nu_j \\ &= 2\eta_m u^{\mathsf{T}}\mathrm{artanh}\left(\frac{u}{\eta_m}\right) + \eta_m^2\sum_{j=1}^{m}\ln\left(1-\frac{u_j^2}{\eta_m^2}\right) \end{aligned} \tag{15}$$

where $|u| < \eta_m$, η_m is a positive constant, and artanh$(\cdot)$ is the inverse of the hyperbolic tangent function tanh$(\cdot)$. Note that other monotonically bounded functions can also be used to deal with constrained control problems.

Remark 3: It is possible that faults occur during the execution of the actuator [117]–[122]. The general form of the actuator failure is expressed as

$$u_f(t) = \rho_f(t)u(t) + \delta_f(t) \tag{16}$$

where $0 < \rho_f(t) \leq 1$ and $\delta_f(t)$ represent the efficiency coefficient and bias fault, respectively. When $\rho_f(t) = 1$ and $\delta_f(t) = 0$, there is no fault in the actuator.

4) Off-Policy IRL: Off-policy learning is often employed when the accurate system model is unknown. Compared with on-policy learning, it uses system data generated by arbitrary control to solve the HJB equation [123]–[131]. If the system (1) is written as

$$\dot{x}(t) = f(x(t)) + g(x(t))u^{i-1} + g(x(t))\left(u(t) - u^{i-1}\right). \tag{17}$$

Then, based on (9) and (11), we have

$$(\nabla V)^{i\mathsf{T}}\left(f(x(t)) + g(x(t))u^{i-1}\right) = -L\left(x(t), u^{i-1}\right) \tag{18}$$

$$(\nabla V)^{i\mathsf{T}} g = -2u^{i\mathsf{T}}R \tag{19}$$

respectively. The derivative of the value function with respect to (17) is

$$\begin{aligned} \dot{V}^i(x(t)) &= (\nabla V)^{i\mathsf{T}}\Big(f(x(t)) + g(x(t))u^{i-1} \\ &\quad + g(x(t))\left(u(t) - u^{i-1}\right)\Big) \\ &= -L\left(x(t), u^{i-1}\right) - 2u^{i\mathsf{T}}R\left(u(t) - u^{i-1}\right). \end{aligned} \tag{20}$$

By integrating both sides of (20) on the interval $[t-T, t]$, the following equation is used in the off-policy IRL scheme:

$$\begin{aligned} V^i(x(t-T)) &= \int_{t-T}^{t}\left(L\left(x(t), u^{i-1}\right) + 2u^{i\mathsf{T}}R\left(u(t) - u^{i-1}\right)\right)\mathrm{d}\nu \\ &\quad + V^i(x(t)). \end{aligned} \tag{21}$$

It is found that the mathematical system model is not explicitly included in (21), but is actually implicit in the data measurement. In addition, the value function is evaluated using system data generated with arbitrary control policies, which increases the "exploration" ability of the learning process [89], [132].

5) Experience Replay/Concurrent Learning: The persistence of excitation (PE) condition can ensure convergence to the optimal solution, so it is necessary in the learning process. In general, traditional PE conditions are difficult or impossible to implement online. Moreover, in the learning process, the PE condition makes the algorithm require a large number of samples [133]–[136].

Definition 2 (PE Condition) [137]: At any given time t, the signal $\rho_1(t)$ is persistently excited if it satisfies

$$q_1 I \leq \int_{t-T}^{t} \rho_1(\nu)\rho_1^{\mathsf{T}}(\nu)\mathrm{d}\nu \leq q_2 I \tag{22}$$

where T, q_1, and q_2 are positive constants, and I is an identity matrix with appropriate dimensions.

By using current and past data repeatedly, experience replay/concurrent learning technique provides simplified conditions for real-time monitoring of PE. In [133], concurrent learning was used to describe the idea of adaptive control for uncertain systems. It is shown that the stability of the model reference adaptive controller can be achieved by using both current data and abundant recorded data. However, the optimality of the closed-loop system was not explained. Although experience replay was mentioned in [134], it did not prove convergence and stability. In [108], the experience replay technique was applied to the IRL algorithm.

6) Function Approximation Based on NN: NNs have been used for function approximation in the implementation of ADP algorithms.

Assumption 2: The continuous function $\rho_2(x)$ can be expressed by an NN [138]

$$\rho_2(x) = \mathcal{W}_0^{\mathsf{T}}\phi_0(x) + \varepsilon_0(x) \tag{23}$$

where $\mathcal{W}_0$, $\phi_0(\cdot)$, and $\varepsilon_0(x)$ are the weight, the activation function, and the reconstruction error of the NN.

According to Assumption 2, the optimal value function and the optimal control are expressed as

$$V^*(x) = \mathcal{W}_1^{\mathsf{T}}\phi_1(x) + \varepsilon_1(x) \tag{24}$$

$$u^*(x) = \mathcal{W}_2^{\mathsf{T}}\phi_1(x) + \varepsilon_2(x) \tag{25}$$

respectively. Employing a critic NN and an actor NN to approximate the optimal value function and the optimal control, we have

$$\hat{V}(x) = \hat{\mathcal{W}}_1^{\mathsf{T}}\phi_1(x) \tag{26}$$

$$\hat{u}(x) = \hat{\mathcal{W}}_2^{\mathsf{T}}\phi_1(x). \tag{27}$$

The weight estimation errors of the critic NN and the actor NN are given as

$$\widetilde{\mathcal{W}}_1 = \mathcal{W}_1 - \hat{\mathcal{W}}_1 \tag{28}$$

$$\widetilde{\mathcal{W}}_2 = \mathcal{W}_1 - \hat{\mathcal{W}}_2 \tag{29}$$

respectively.

Remark 4: According to the relationship between the optimal value function and the optimal control derived from the HJB equation, the actor network can be omitted for affine nonlinear systems and only a single-critic network is used [15], [59], [65], [139]. This structure is called the single network adaptive critic (SNAC) and was first developed in [140]. In this case, (25) becomes

$$u^*(x) = -\frac{1}{2}R^{-1}\left(\frac{\partial F(x,u,t)}{\partial u}\right)^{\mathsf{T}}\left(\nabla\phi_1^{\mathsf{T}}(x)\mathcal{W}_1 + \nabla\varepsilon_1^{\mathsf{T}}(x)\right)$$

and (27) becomes

$$\hat{u}(x) = -\frac{1}{2}R^{-1}\left(\frac{\partial F(x,u,t)}{\partial u}\right)^{\mathsf{T}}\nabla\phi_1^{\mathsf{T}}(x)\hat{\mathcal{W}}_1.$$

A single network can simplify the analysis and reduce the amount of calculation. However, when the complete knowledge of the system is unknown, the actor network needs to be considered.

The residual error $e_{\mathcal{H}}$ is defined as

$$e_{\mathcal{H}} = L\big(x,\hat{u}\big) + \nabla\hat{V}(x)\big(f + g\hat{u}(x)\big) \tag{30}$$

where $\nabla\hat{V}(x) = \hat{\mathcal{W}}_1^{\mathsf{T}}\nabla\phi_1(x)$. The objective function to be minimized is

$$E_{\mathcal{H}} = \int_0^t e_{\mathcal{H}}^2(\tau)\mathrm{d}\tau. \tag{31}$$

A critic NN update rule based on least squares is given by [141]

$$\dot{\hat{\mathcal{W}}}_1 = -\beta_1\Phi\frac{\theta}{1+\eta_c\theta^{\mathsf{T}}\Phi\theta}e_{\mathcal{H}} \tag{32}$$

where $\beta_1 > 0$ is the learning rate, $\theta = \phi_1(x)(f + g\hat{u})$, and $\Phi = (\int_0^t \theta(\nu)\theta^{\mathsf{T}}(\nu)\mathrm{d}\nu)^{-1}$. An actor NN learning algorithm based on gradient is given by [141]

$$\dot{\hat{\mathcal{W}}}_2 = -\frac{\beta_2}{\sqrt{1+\theta^{\mathsf{T}}\theta}}\frac{\partial e_{\mathcal{H}}}{\partial\hat{\mathcal{W}}_2}e_{\mathcal{H}} - \frac{\beta_3\Big(\hat{\mathcal{W}}_2 - \hat{\mathcal{W}}_1\Big)}{2} \tag{33}$$

where β_2 and β_3 are positive constants, $\sqrt{1+\theta^{\mathsf{T}}\theta}$ is introduced for normalization purpose, and the second term on the right-hand side of (33) is designed to facilitate stability analysis [141]. The PE condition is needed to ensure that the signal is uniformly ultimately bounded (UUB).

Definition 3 (UUB) [108], [142]: The signal $\rho_3(t)$ is UUB on a compact set $\mathcal{P}$, if for all $\rho_3(t_0) \in \mathcal{P}$, there exist a positive bound $\mathcal{D}_b$ and a number $T_D(\mathcal{D}_b, \rho_3(t_0))$ such that $\|\rho_3(t)\| \le \mathcal{D}_b$ for all $t > t_0 + T_D$.

For unknown system dynamics, a data-driven approach was described in [143]. Reconstructing an unknown system using the following form of network:

$$\dot{x} = \mathcal{A}x + \mathcal{W}_3^{\mathsf{T}}\phi_3(z) + \varphi_3(x) \tag{34}$$

where $z = \mathcal{V}_3^{\mathsf{T}}[x^{\mathsf{T}}, u^{\mathsf{T}}]^{\mathsf{T}}$, and $\mathcal{A}$ is a constant matrix designed to stabilize the system. The corresponding identifier is chosen as

$$\dot{\hat{x}} = \mathcal{A}\hat{x} + \hat{\mathcal{W}}_3^{\mathsf{T}}\phi_3\big(\hat{z}\big) \tag{35}$$

where $\hat{z} = \mathcal{V}_3^{\mathsf{T}}[\hat{x}^{\mathsf{T}}, u^{\mathsf{T}}]^{\mathsf{T}}$. Then, we have

$$\dot{\hat{\mathcal{W}}}_3 = \beta_4\phi_3\big(\hat{z}\big)\widetilde{x}^{\mathsf{T}} \tag{36}$$

where β_4 is the learning rate. It can be proved that after sufficient learning, the identification error $\widetilde{x}$ is asymptotically stable. The system input matrix $g(x)$ is obtained by taking the partial derivative of (35) with respect to the control function u, i.e.,

$$g(x) = \hat{\mathcal{W}}_3^{\mathsf{T}}\frac{\partial\phi_3\big(\hat{z}\big)}{\partial\hat{z}}\frac{\partial\hat{z}}{\partial u}. \tag{37}$$

7) NN Learning Based on the Event-Triggering Mechanism: In the traditional network control system, the control inputs u are transmitted at a fixed sampling interval. However, in the event-triggering mechanism, they are only sampled and transmitted at the event-triggering instants $\{\tau_l\}_{l=0}^{\infty}$. Therefore, the control input sequence is expressed as $\{u(x(\tau_l))\}_{l=0}^{\infty}$. This transmission mechanism reduces communication traffic and computational load without compromising the stability of the system [144]–[154]. An event is generated by an event-triggering threshold being violated, i.e., $e_l(x) > \mathcal{T}(x)$, where $e_l(x)$ and $\mathcal{T}(x)$ are the error and the given or designed threshold of the current state, respectively. Specifically

$$e_l(x) = \begin{cases} 0, & t = \tau_l \\ x(\tau_l) - x(t), & t \in (\tau_l, \tau_{l+1}) \end{cases} \tag{38}$$

$$\mathcal{T}(x) = \frac{\big(1-\alpha_e^2\big)L_1(x) + L_2\big(\hat{u}(x(\tau_l))\big)}{\mathcal{D}_t^2} \tag{39}$$

where $0 \le \alpha_e \le 1$ and $\mathcal{D}_t > 0$ are constants. Therefore, the control input based on the event-triggering mechanism is expressed as $u(e_l(x) + x(t))$. Then, (2) and (8) become

$$\dot{x}(t) = f(x(t)) + g(x(t))u(e_l(x) + x(t)) \tag{40}$$

$$u^* = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(e_l(x) + x(t))\nabla V^*(e_l(x) + x(t)) \tag{41}$$

respectively.

Assumption 3: The control input and the closed-loop system satisfy Lipschitz continuity, i.e.,

$$\big\|u^*(x(t)) - u^*(x(\tau_l))\big\| \le \mathcal{D}_u\|e_l(x)\| \tag{42}$$

$$\|f(x) + g(x)u(e_l(x) + x(t))\| \le \mathcal{D}_x(\|x\| + \|e_l(x)\|) \tag{43}$$

where $\mathcal{D}_u$ and $\mathcal{D}_x$ are positive constants.

Under the premise that the control input satisfies Lipschitz continuity and $L_2(u) = u^{\mathsf{T}}u$, the Hamiltonian based on

the event-triggered control input satisfies the following relationship [145]:

$$\begin{aligned}\left\|H\left(x, u^*(x(\tau_l)), \nabla V^*\right)\right\| &= \left\|u^*(x(t)) - u^*(x(\tau_l))\right\|^2 \\ &\leq \mathcal{D}_u^2 \|e_l(x)\|^2 \end{aligned} \tag{44}$$

where $\mathcal{D}_u > 0$ is a constant. Since the control policy is only updated at the triggering time instant, the event-triggered closed-loop system is presented as an impulsive dynamic system, and the construction method of this hybrid system was given in [143], [145], and [154]. According to (28) and (30), we have

$$\begin{aligned} e_{\mathcal{H}} &= -\widetilde{\mathcal{W}}_1^{\mathsf{T}}\theta + \mathcal{W}_1^{\mathsf{T}}\theta + L\left(x, \hat{u}\right) \\ &= -\widetilde{\mathcal{W}}_1^{\mathsf{T}}\theta + e_{\mathcal{W}} \end{aligned} \tag{45}$$

where $e_{\mathcal{W}} = \mathcal{W}_1^{\mathsf{T}}\theta + L(x, \hat{u})$. The augmented state is defined as $\Psi = [x^{\mathsf{T}}, x^{\mathsf{T}}(\tau_l), \widetilde{\mathcal{W}}_1^{\mathsf{T}}]^{\mathsf{T}}$, so the impulsive dynamic system is

$$\begin{cases} \dot{\Psi} = \begin{bmatrix} f + g\hat{u}(x(\tau_l)) \\ 0 \\ -\beta_c \Phi \frac{\theta}{1+\eta_c \theta^{\mathsf{T}}\Phi\theta}\left(\widetilde{\mathcal{W}}_1^{\mathsf{T}}\theta - e_{\mathcal{W}}\right) \end{bmatrix}, & t \in \left[\tau_l, \tau_{l+1}\right) \\ \Psi(t) = \Psi\left(t^-\right) + \begin{bmatrix} 0 \\ e_l(x) \\ 0 \end{bmatrix}, & t = \tau_{l+1} \end{cases} \tag{46}$$

where $\beta_c > 0$ is a constant and $\Psi(t^-) = \lim_{\varsigma \to 0} \Psi(t - \varsigma)$. The event interval is defined as

$$\delta_l = \tau_{l+1} - \tau_l. \tag{47}$$

It is proved in [106] that there is a nonzero positive lower bound for the event intervals, i.e.,

$$\delta_l \geq \frac{1}{2\mathcal{D}_x} \ln\left(2\left\|\frac{\mathcal{T}(x)}{x(\tau_l)}\right\| + 1\right) > 0. \tag{48}$$

This avoids the accumulation of events, i.e., it excludes the Zeno phenomenon.

B. ADP for Optimal Output Regulation

In practical engineering control problems, the state measurement is sometimes unrealistic, so the output regulation problem in linear systems [127], [155]–[157] or nonlinear systems [158], [159] is widely concerned. It can be seen that the minimum value function is actually determined by the state vector. When the controlled system is completely observable, the output regulation problem of the system can be transformed into an equivalent state regulation problem.

Problem 2: The optimal output regulation is to design control input such that the output approaches zero when minimizing the value function.

1) Optimal Output Regulation for Linear Systems: In [155], by discretizing the unknown dynamic linear system, an optimal output-feedback control policy was proposed, which strictly shows that the detection noise in the online learning process does not affect the accuracy of the solution of the discrete Riccati equation. In the case of unknown system dynamics and disturbances, the predefined performance indices were minimized by combining output regulation and ADP in [156] to achieve tracking control and disturbance rejection. Under the relaxation of some assumptions, the conclusion was further generalized in [160]. Then, the application to grid-connected inverter system with input delay [161] shows that this method was effective to solve adaptive optimal output regulation problems.

Algorithm 2 LQR Based on IRL for System (49) and (50) (See [162])

Step 1 Initialization:
$i = 0$,
$S^i = 0$,
Select an initial stabilizing gain K^i.

Step 2 Evaluation:
$i = i + 1$,
P^i is obtained according to

$$\begin{aligned} x^{\mathsf{T}}(t)P^i x(t) = \int_t^{t+T} x^{\mathsf{T}}(\nu)(Q + (K^{i-1})^{\mathsf{T}} R K^{i-1}) \\ \times x(\nu)\mathrm{d}\nu + x^{\mathsf{T}}(t+T)P^i x(t+T). \end{aligned} \tag{51}$$

Step 3 Improvement:
The updated control policy is obtained according to

$$K^i = R^{-1}(\mathcal{B}^{\mathsf{T}}P^i + S^{i-1})\mathcal{C}^{\mathsf{T}}(\mathcal{C}\mathcal{C}^{\mathsf{T}})^{-1}, \tag{52}$$

$$S^i = RK^i\mathcal{C} - \mathcal{B}^{\mathsf{T}}P^i. \tag{53}$$

Step 4 Judgment:
If preset conditions for convergence are not met, go to **Step 2**.

Step 5 Stop:
Obtain the optimal control policy $u^* = -K^i y$ and the optimal value function $V^* = x^{\mathsf{T}}P^i x$.

Take a linear continuous-time system as an example [162]

$$\dot{x} = \mathcal{A}x + \mathcal{B}u \tag{49}$$

$$y = \mathcal{C}x \tag{50}$$

where $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are time-invariant matrices with appropriate dimensions, and y is the output vector of the system. For time interval $T > 0$ and any time t, an online learning algorithm for a suboptimal output-feedback controller based on the IRL technique is presented in Algorithm 2.

In [127], the method was applied to both linear quadratic regulator (LQR) and linear quadratic tracking (LQT) by using a discounted value function as

$$V = \int_t^{\infty} e^{-\beta_d(\nu - t)}\left(y^{\mathsf{T}}Qy + u^{\mathsf{T}}Ru\right)\mathrm{d}\nu \tag{54}$$

where β_d is a positive constant. Then, it was proved that the system state can be constructed by observing the output of a limited historical time. Using the Bellman equation, a model-free off-policy RL controller was designed in the case of the unknown system state and system dynamics. This output-feedback method is equivalent to the state-feedback control and is more robust than the static output-feedback method. Note that the ADP-based control method for the adaptive optimal output regulation of model-free linear systems under

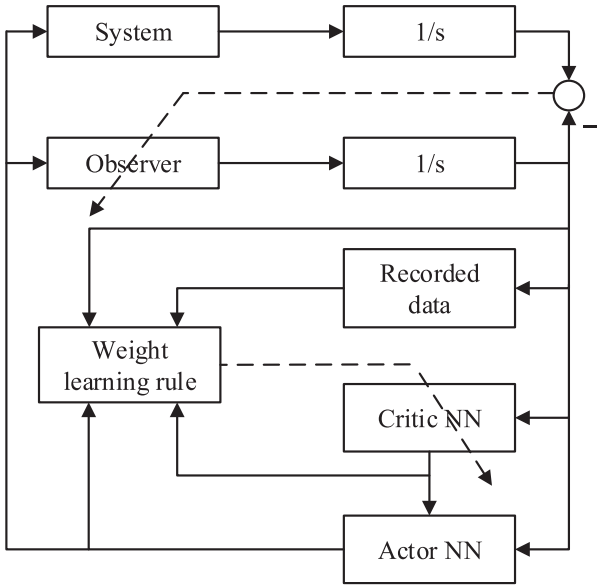

Fig. 1. Structure diagram of an observer–critic learning algorithm [159].

input saturation and the event-triggering mechanism needs to be studied in the future.

2) Optimal Output Regulation for Nonlinear Systems: Liu *et al.* proposed an optimal output regulation scheme for unknown nonlinear systems based on the observer in [158]. The critic network and the three-layer NN observer were used to obtain the optimal control law, which ensures the stability of the closed-loop system. The learning of the actor, critic, and observer in the control scheme is continuous, real time, and simultaneous. Yang *et al.* [159] relaxed PE conditions and the initial admissible control. This is the first time to design an optimal output control method based on the observer–critic structure without considering these two conditions for partially unknown affine nonlinear continuous-time systems. The learning process of this typical algorithm is shown in Fig. 1.

Future work will need to focus on online algorithms for optimal output control of nonaffine nonlinear continuous-time systems.

C. ADP for Optimal Tracking Control

For practical systems, such as unmanned aerial vehicles (UAVs) and spacecrafts, it is necessary to design controllers to track desired trajectories in an optimal manner. Therefore, the optimal tracking problem has attracted more and more attention in the control field [163], [164]. It is found that in recent years, the related work usually involves the use of augmented systems. By constructing the augmented system with tracking error and desired trajectory, the solution of the optimal tracking control problem is transformed into an optimal regulation problem [165].

Problem 3: The optimal tracking control problem is to design a control policy to make the actual output of the system track the desired trajectory and minimize the preset value function.

Consider the general value function corresponding to (1)

$$V(x(t)) = \int_t^\infty e^{-\beta_h(\nu - t)} L(x(\nu), u(\nu)) \mathrm{d}\nu \tag{55}$$

where $L(x, u) = L_1(x) + L_2(u)$ and $\beta_h > 0$ represents the discount factor. The desired reference trajectory dynamics is defined as

$$\dot{x}_h(t) = R(x_h(t)) \tag{56}$$

where $R(x_h(t))$ satisfies the Lipschitz continuity and $R(0) = 0$. The tracking error is

$$e_h(t) = x(t) - x_h(t). \tag{57}$$

According to (1) and (56), we have

$$\dot{e}_h(t) = f(x(t)) + g(x(t))u(t) - R(x_h(t)). \tag{58}$$

The augmented system state is defined as $\zeta(t) = [e_h^\mathsf{T}(t), x_h^\mathsf{T}(t)]^\mathsf{T}$, and then the augmented system dynamics are further obtained as

$$\dot{\zeta}(t) = F(\zeta(t)) + G(\zeta(t))u(\zeta(t)) \tag{59}$$

where

$$F(\zeta(t)) = \begin{bmatrix} f(x_h(t) + e_h(t)) - R(x_h(t)) \\ R(x_h(t)) \end{bmatrix}$$
$$G(\zeta(t)) = \begin{bmatrix} g(x_h(t) + e_h(t)) \\ 0 \end{bmatrix}.$$

The value function associated with (59) is

$$V(\zeta(t)) = \int_t^\infty e^{-\beta_h(\nu - t)} L(\zeta(\nu), u(\nu)) \mathrm{d}\nu \tag{60}$$

where $L(\zeta, u) = L_1(\zeta) + L_2(u)$. The Hamiltonian is given by

$$H(\zeta, u) = L(\zeta, u) - \beta_h V(\zeta) + \nabla V^\mathsf{T}(\zeta)(F(\zeta) + G(\zeta)u(t)). \tag{61}$$

The HJB equation of the tracking control problem is derived based on the Bellman principle of optimality

$$L\big(\zeta, u^*\big) - \beta_h V^*(\zeta) + \nabla V^{*\mathsf{T}}(\zeta)\big(F(\zeta) + G(\zeta)u^*(t)\big) = 0. \tag{62}$$

According to the stationarity condition, the relationship between the optimal control and the optimal value function can be obtained. For the quadratic energy function, i.e., $L_2(u) = u^\mathsf{T} R u$, we have

$$u^* = -\frac{1}{2} R^{-1} G^\mathsf{T}(\zeta) \nabla V^{*\mathsf{T}}(\zeta). \tag{63}$$

In [113], the standard form solution of the tracking control problem was first given. The disadvantage of this solution is that the feedforward and feedback parts of the control input are solved separately. By minimizing a new discounted value function of the augmented system, the two parts of the control input are obtained simultaneously.

Remark 5: In the optimal tracking control problem, it is shown that the discount factor β_h in the value function is needed. Since the control input includes feedforward control and feedback control, the feedforward control input may make the value function V unbounded when the reference trajectory

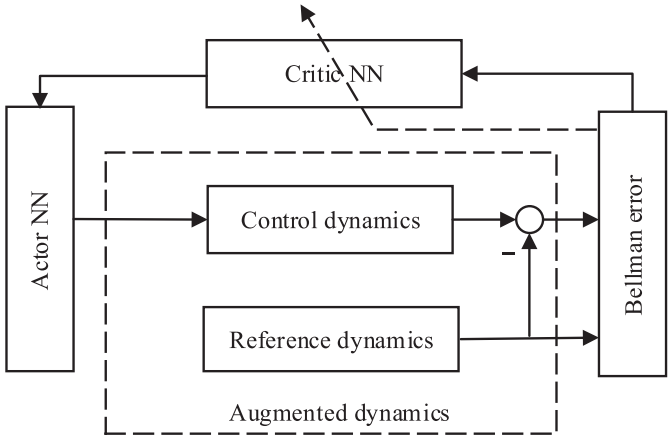

Fig. 2. Structure diagram of a tracking control learning algorithm in [166].

$x_h(t)$ does not converge to zero. The employment of the discount factor β_h ensures that the value function V is bounded, thereby effectively avoiding this problem [113].

The structure of tracking control based on a single network [166] is shown in Fig. 2. It is practically impossible to track the helicopter trajectory by using state feedback. In this view, in [103], an output-feedback controller for helicopter UAV trajectory tracking was presented by constructing an NN-based observer. For the optimal tracking control based on the event-triggered mechanism, the triggering threshold of the augmented system given in [147] is

$$\mathcal{T}_h(x) = \frac{\left(1-\alpha_h^2\right)L_1(e_h) + L_2(u)}{\mathcal{D}_h^2} \tag{64}$$

where $\alpha_h \in [0, 1]$ is a constant. The algorithm design of inverse optimization for tracking control of nonlinear nonaffine systems needs to be further studied as mentioned in [167].

Remark 6: A plenty of excellent works on optimal regulation for discrete-time systems have been presented.

1) The design of state regulator is the basis of other regulators [76], [78], [82], [168]–[174]. For example, in [78], [81], and [82], Wei *et al.* described the specific ADP algorithm of VI and local VI, respectively. In [171], the developed generalized PI algorithm can start from any positive semidefinite function, and policy evaluation and policy improvement are carried out with their own independent iteration indicators.
2) Literature, such as [161] and [175]–[178], have developed the discrete output regulator. Luo *et al.* developed multistep Q-learning to solve the optimal output control problem of model-free discrete-time aircraft in [176]. Its implementation does not require the knowledge of dynamic systems and is a data-based learning control method. Moreover, it is proved that the sequence of iterative Q functions converges to the optimal Q function. Although this multistep Q-learning method was developed in the context of helicopters, it can also be applied to other similar systems. However, when the system involves state saturation, input delay, or random disturbances, the convergence of the algorithm needs further study.
3) The tracking control problem of discrete-time systems was introduced in [98], [130], and [179]–[190]. For linear tracking control problems, PI and VI were studied in detail in [179]. Wang *et al.* first used an iterative ADP algorithm to design the finite-time optimal tracking controller for a class of discrete nonlinear systems in [180]. Input constraints and time delays were considered in [181] and [182], respectively. In [98], the convergence theory of the multistep policy evaluation method in the tracking control problem was presented. The event-triggering mechanism was applied to the suboptimal tracking control of discrete-time nonlinear systems [183]. The actor–critic off-policy algorithm was applied to solve the optimal tracking control under sudden changes [130]. It is a promising work to extend this method to the tracking control problem of nonlinear systems.

Up to this point, the introduction of ADP techniques in solving basic optimal control problems is given. It should be emphasized that the problems that ADP can solve are not limited to these.

III. ADP for Game Theory

The game theory, which is widely used in many subjects, such as mathematics and economics, is a decision-making theory in a conflicting environment. According to the agreement among the players, it can be divided into noncooperative and cooperative games, namely, zero-sum and nonzero-sum games [191], [192]. The Nash equilibrium solution of the noncooperative games opens up a new and effective way to solve H_∞ control problems (see [193]–[197]). Similarly, nonzero-sum differential games can also be solved using the ADP technique. To be specific, when there are multiple controllers in a single nonlinear system, each controller tries to minimize its own value function in the sense of Nash equilibrium. For such problems, it is needed to solve coupled algebraic Riccati equations (AREs) for linear systems or coupled Hamilton–Jacobi (HJ) equations for nonlinear dynamic systems. Similar to HJB and HJI equations, the nonlinear characteristics of coupled HJ equations makes it impossible to obtain analytic solutions directly. It is found that most of the literature on nonzero-sum games focused on continuous-time systems. Therefore, the next step is to briefly describe nonzero-sum game problems of continuous-time systems and analyze the ADP technique for solving coupled HJ equations.

Problem 4: The nonzero-sum game problem intends to design a set of admissible control policies $u^* = \{u_j^*\}_{j=0}^{\mathcal{N}}$ to minimize the corresponding value function of each player.

In general, consider the following nonzero-sum games with $\mathcal{N}$ players:

$$\dot{x} = f(x) + \sum_{j=1}^{\mathcal{N}} g_j(x)u_j,\ \ x(0) = x_0. \tag{65}$$

The set of control policies of $\mathcal{N}$ players is represented by u, i.e., $u \triangleq \{u_j\}_{j=0}^{\mathcal{N}}$. For the kth player, define its value

function as

$$V_k = \int_t^\infty L_k(x(\nu), u(\nu))\mathrm{d}\nu = \int_t^\infty (L_{k1}(x(\nu)) + L_{k2}(u(\nu)))\mathrm{d}\nu \quad (66)$$

where $L_{k2}(u) = \sum_{j=1}^{\mathcal{N}} L_{k2j}(u_j) = \sum_{j=1}^{\mathcal{N}} u_j^{\mathsf{T}} R_{kj} u_j$ and R_{kj} is a positive-definite symmetric matrix. Supposing that (66) is continuously differentiable, the Hamiltonian associated with the kth player is defined as

$$H_k(x, u, V_k) = L_{k1}(x) + L_{k2}(u) + \nabla V_k^{\mathsf{T}} \left(f(x) + \sum_{j=1}^{\mathcal{N}} g_j(x) u_j \right). \quad (67)$$

The optimal value function V_k^* satisfies

$$\min_{u_k} H_k\big(x, u, V_k^*\big) = 0. \quad (68)$$

Since the Hamiltonian has zero partial derivatives with respect to the optimal control policy, we have

$$u_k^* = -\frac{1}{2} R_{kk}^{-1} g_k^{\mathsf{T}}(x) \nabla V_k^*. \quad (69)$$

Substituting V_k and u_j into the Hamiltonian (67) with V_k^* and u_j^*, respectively, and let it equal to zero, it becomes the coupled HJ equation.

Most of the traditional solutions assume that the system dynamics is known. For example, an offline PI algorithm using the actor–critic structure was proposed in [198]. In order to reduce the computational complexity of using two networks, in [199], Liu *et al.* developed an online synchronous PI learning algorithm based on one critic network only. Inspired by [108], an online concurrent learning algorithm was proposed in [200], which was then extended to constrained nonzero-sum game problems in [114]. However, the algorithm in [114] was established based on the known dynamic model. Zhao *et al.* extended the application scope of [114] by using the NN identifier in [135]. Similar to [199], this scheme used one critic network only to approximate each player's value function and optimal control policy. The integral Q-learning algorithm, the off-policy Q-learning algorithm and the data-driven Q-learning algorithm for linear systems have been developed in [201]–[203], respectively. Song *et al.* extended the above work in [89] and developed the off-policy IRL technique. Although the work in [89] is concerned with nonlinear systems, all players have the same control input matrix $g_j(x)$, which restricts its applications. Note that the above work belongs to the scope of optimal regulation. The nonzero-sum game of optimal tracking control was first explored in [204]. Although it does not require the knowledge of dynamic models and is a data-based Q-learning scheme, it contains no convincing theoretical analysis.

According to the previous investigation, uncertainties have not attracted much attention in nonzero-sum games. However, it is nevertheless widespread in practical engineering. How to find the Nash equilibrium such that each player still minimizes the value function in the uncertain environment will be a meaningful work. As far as we know, two robust control methods were designed for two kinds of uncertainties based on the optimal control policy in [205]. Specifically, consider a single dynamic system with multiple players with uncertainties as

$$\dot{x} = f(x) + \sum_{j=1}^{\mathcal{N}} g_j(x)\big(u_j + d_j\big) \quad (70)$$

where d_j denotes the uncertainties of the jth player. It is divided into two categories

$$d_j = d_j(x),\ \big\|d_j(x)\big\| \le \beta_{d1j}\|x\| \quad (71)$$

where β_{d1j} is a positive constant corresponding to the jth player, and

$$d_j = -\beta_{d2j},\ \ \dot{\beta}_{d2j} = 0 \quad (72)$$

where β_{d2j} is an unknown constant, respectively. In [205], the data-based RL method was used to obtain the optimal control policies of (65). Then, based on the optimal control policies, a suitable modification is made to obtain the robust controller. The data-driven offline learning method in [205] analyzes the matched problem, rather than the general unmatched problem. It is worth mentioning that the nonzero-sum game is very similar to human beings' daily life. How to apply the proposed algorithm to the nonzero-sum game in life, and then bring benefits to human beings' lives is meaningful.

In addition, in multiplayer nonzero-sum games, exploring efficient and low-cost algorithms should be paid more attention. Some literature attempted to simplify the network structure by using algorithms based on a single network. In order to further reduce the number of executions of control policies, the event-triggered mechanism was introduced into multiplayer nonzero-sum games. On one hand, the evolution of the dynamic system is affected by multiple players simultaneously. On the other hand, we need to maintain the performance of multiple players. These two reasons made it difficult to introduce the event-triggering mechanism. As far as we know, Sahoo *et al.* [206] and Mu and Wang [207] have studied this problem independently and the proposed triggering mechanisms are different. In [206], the control policy error $e_j(x(\tau_l))$ is expressed as

$$e_j(x(\tau_l)) = u_j(x(\tau_l)) - u_j. \quad (73)$$

The system based on the event-triggered control policy is written as

$$\dot{x} = f(x) + \sum_{j=1}^{\mathcal{N}} g_j(x)\big(e_j(x(\tau_l)) + u_j\big). \quad (74)$$

Define a new value function

$$V_k = \int_t^\infty L_k(x(\nu), u(\nu))\mathrm{d}\nu = \int_t^\infty (L_{k1}(x(\nu)) + L_{k2}(u(\nu)) - L_{k3}(\bar{e}(\nu)))\mathrm{d}\nu \quad (75)$$

where $\bar{e}$ is the threshold. For (75), the control policy and the threshold are considered as two players in zero-sum games,

respectively, in which the control policy is to minimize the value function and the triggering threshold is to maximize the value function. Then, an event-triggered control policy for the multiplayer nonzero-sum game based on the idea of mini-max was proposed. Although the authors elaborated that the developed scheme can avoid the Zeno phenomenon, they did not provide a theoretical analysis. How to relax the dependence on the control input matrix in the implementation process remains to be further studied. Consider the following two-player nonzero-sum game problems [207]:

$$\dot{x} = f(x) + g_1(x)u_1 + g_2(x)u_2. \tag{76}$$

Excessive triggering period results in the decrease of the approximation performance of the NN. Therefore, the authors proposed a novel event-triggering mechanism with an alarm sampling period mode. This mechanism has two triggering conditions

$$\begin{cases} \|e_l(x)\|^2 \le \mathcal{T}(x) \\ \max\{\tau_{l+1} - \tau_l\} = T_{\max} \end{cases} \tag{77}$$

where $e_l(x)$ is defined in (38), $\mathcal{T}(x)$ is the designed threshold, and $T_{\max}$ is the given alarm sampling period. When the error violates the threshold or the sampling period exceeds the alarm, the actions of the two players will be updated at the same time. The deduction of the minimum triggering interval effectively proves that the Zeno phenomenon will not occur. However, its applicability to multiplayer games is unknown. It is important to extend the proposed algorithms to general systems with constraints or unknown dynamics.

IV. ADP for Large-Scale Systems

Large-scale systems, such as power systems, transportation systems, and ecosystems, contain cross-linking terms among their subsystems. The existence of these cross-linking components increases the difficulty of the traditional centralized control design. In recent years, decentralized control methods have received much attention. The local controllers of the subsystems are designed using the local state information and further construct a decentralized controller [208]. In [209], appropriate performance indices were predefined for isolated subsystems, and it was proved that decentralized control policies can be obtained from the optimal policies of these subsystems.

Problem 5: For decentralized control problems of nonlinear systems composed of N subsystems with interconnections, the goal is to develop a set of control policies $\{u_i\}_{i=1}^N$ to stabilize the system.

Consider the nonlinear continuous-time system composed of N subsystems

$$\dot{x}_i = f_i(x_i) + g_i(x_i)u_i(x_i) + \bar{\mathcal{I}}_i(x), \quad i = 1, 2, \ldots, N \tag{78}$$

where $x = [x_1^\mathsf{T}, \ldots, x_N^\mathsf{T}]^\mathsf{T}$ is the state of the whole system composed of all interconnected subsystems, and $\bar{\mathcal{I}}_i(x)$ denotes the interconnected component of the subsystem. If $\bar{\mathcal{I}}_i(x)$ can be decomposed into

$$\bar{\mathcal{I}}_i(x) = g_i(x_i)\mathcal{I}_i(x) \tag{79}$$

then, it is called a matched interconnection term. The isolated subsystem of system (78) is given as

$$\dot{x}_i = f_i(x_i) + g_i(x_i)\bar{u}_i(x_i), \ \ i = 1, 2, \ldots, N. \tag{80}$$

The value function is given as

$$\begin{aligned} V_i &= \int_t^\infty L_i(x_i(\nu), \bar{u}_i(\nu))\mathrm{d}\nu \\ &= \int_t^\infty (L_{i1}(x_i) + L_{i2}(\bar{u}_i))\mathrm{d}\nu, \ \ i = 1, 2, \ldots, N. \end{aligned} \tag{81}$$

The Hamiltonian is defined as

$$\begin{aligned} H_i(x_i, \bar{u}_i, \nabla V_i) = {} & \nabla V_i^\mathsf{T}(f_i(x_i) + g_i(x_i)\bar{u}_i(x_i)) \\ & + L_{i1}(x_i) + L_{i2}(\bar{u}_i), \ \ i = 1, 2, \ldots, N. \end{aligned} \tag{82}$$

The optimal value function V_i^* satisfies

$$V_i^* = \min_{\bar{u}_i} V_i, \ \ i = 1, 2, \ldots, N. \tag{83}$$

Then, the HJB equation is

$$\min_{\bar{u}_i} H_i\big(x_i, \bar{u}_i, \nabla V_i^*\big) = 0, \ \ i = 1, 2, \ldots, N. \tag{84}$$

Considering $L_{i2}(\bar{u}_i) = \bar{u}_i^\mathsf{T} R_i \bar{u}_i$ and stationarity conditions, the optimal control policy for isolated subsystems becomes

$$\bar{u}_i^* = -\frac{1}{2}R_i^{-1} g_i^\mathsf{T}(x_i)\nabla V_i^*, \ \ i = 1, 2, \ldots, N. \tag{85}$$

It was shown in [210] that decentralized control $\{u_i\}_{i=1}^N$ of the interconnected system (78) can be obtained by increasing the local feedback gain proportionally. Then, the online PI algorithm was used to solve the HJB equation (84). As an improvement, the model-free online IRL algorithm was used to solve decentralized control problems of unknown interconnected systems in [211].

If (79) is not satisfied, it becomes a more general unmatched decentralized control problem [75], [212]–[216]. In [212], by introducing a bounded function $d_i(\cdot)$ to modify the value function, an unmatched interconnected large-scale system with uncertainties was considered. For unknown unmatched interconnected components, Zhao *et al.* [75] developed an NN-based local observer, which relied on the local state of isolated subsystems and the reference state of coupled subsystems. Therefore, the assumption of bounded or matched interconnections in previous methods can be relaxed. Then, a local value function with an adaptive estimation component was designed to compensate for the substitution errors. The problem considered in [213] was the same as that in [75]. However, it decomposes the interconnection terms in the following form:

$$\bar{\mathcal{I}}_i(x) = g_i(x_i)g_i^+(x_i)\bar{\mathcal{I}}_i(x) + \big(I_{n_i} - g_i(x_i)g_i^+(x_i)\big)\bar{\mathcal{I}}_i(x). \tag{86}$$

Based on (86), an auxiliary subsystem was designed. Then, it transformed the solution of the decentralized controller for the large-scale system into the optimal controller for the auxiliary subsystem. Compared with the learning structures of two networks in [213], the method in [214] used one single critic network, which simplified the network structure and did not require initial admissible control.

For complex large-scale systems, the higher the system performance requirements, the greater the likelihood of system failure in the actuator components. In order to improve system reliability, safety, and survivability in complex environments, fault diagnosis has been considered by scholars. It needs to redesign the controller to recover the system control performance as soon as possible. Therefore, fault-tolerant control has become an important and meaningful topic [217]–[219]. For the reconfigurable manipulator with random actuator fault [217], an observer and an identifier were established to detect and identify the fault, respectively, and then the fault was compensated in real time. A decentralized fault-tolerant control algorithm based on self-adjusting local feedback gain was proposed in [218] to prevent partial loss of actuator efficiency. Based on the work of Zhao *et al.* [75], time-varying actuator faults were considered in [219], and the developed algorithm assumed that these faults have known upper bounds.

Decentralized tracking control is to design a control policy that allows the actual state to track the desired trajectory [220]–[222]. The results are rare on interconnected systems with unknown dynamics. Decentralized tracking control of matched and unmatched interconnected systems was considered in [221] and [222], respectively. For the tracking control problem of system (78) satisfying (79), the decentralized control scheme was indirectly developed by solving the HJB equation of N augmented tracking subsystems in [221], and then a single network-based online ADP algorithm without initial admissible control was developed. In [222], local identifiers were used to identify dynamic models of unknown subsystems. The identification error, the substitution error, and the approximation error were compensated by an improved local value function. This method ensured the tracking control performance of the whole system. Composite learning algorithms, such as wavelet NNs and fuzzy NNs, might improve the tracking performance of large-scale nonlinear systems.

Differential games exist in many practical large-scale systems. Moreover, the existence of external disturbances may affect the stability of interconnected systems. Therefore, the study of decentralized differential games is significant. In [223], the ADP technique was first used to solve the decentralized zero-sum differential games. A class of systems studied was described by

$$\begin{aligned}\dot{x}_i = {} & f_i(x_i) + g_i(x_i)(u_i(x_i) + \mathcal{I}_{1i}(x)) \\ & + k_i(x_i)(d_i(x_i) + \mathcal{I}_{2i}(x)),\ i = 1, 2, \ldots, N. \end{aligned} \tag{87}$$

Inspired by [211], the interconnection terms are assumed to be matched and bounded. It should be noted that this assumption limits the applications of the algorithm. Presently, there is no relevant research on other decentralized differential games.

When considering the communication of large-scale systems, the event-triggered decentralized control was developed by [116], [154], and [224]. Yang and He [154] obtained the decentralized event-triggered control policy of the whole system by means of a set of optimal event-triggered control policies of the auxiliary subsystem. Then, an adaptive critic learning scheme based on experience replay was used to approximate the event-triggered HJB equation of the optimal control problem. The decentralized tracking control problem for modular reconfigurable robots was solved by Zhao and Liu by using an event-triggered ADP method [224]. In addition, the event-triggered decentralized control for large-scale systems with constrained input and external interference has been studied in [116]. It is noted that further research is needed for completely unknown large-scale practical systems.

Remark 7: For the decentralized control of nonlinear discrete-time systems, affine interconnected systems were studied in [225]. It employs the assumption of weak interconnection to turn the solution of the optimal controller of the overall system into solving the optimal controller of each subsystem.

V. Future Perspectives

Based on the above analysis, the following prospects for future research are given.

1) *Data-Based Learning Techniques:* As human being's requirements for production and life gradually improve, the system scale in real life becomes large scale. Systems such as transportation systems, power systems, telemedicine systems, chemical production systems, etc., collect a lot of data when they are running, but are difficult or cannot be described by accurate mathematical models. Observer or identifier techniques are sometimes difficult to implement in the face of these complex large-scale systems. In the era of modern measuring and computing equipments, the development of intelligent learning methods based on the acquired perfect or imperfect data has a broad development prospect. Some literature, such as [73], [204], and [226]–[229], has developed data-based techniques, but more attention should be paid to data-based learning techniques for complex, volatile, large scale, and networked control systems.
2) *Learning Techniques Based on Event-Triggering and Self-Triggering Mechanisms:* It is urgent to reduce communication traffic and computational burden for control systems with limited bandwidth. Event-triggered and self-triggered learning mechanisms are emerged to replace traditional time-triggering mechanisms to deal with these problems [106], [116], [147], [183], [206], [230]–[233]. These new mechanisms face new challenges, such as the design of the triggering mechanism and the guarantee of stability and convergence of the original systems. Therefore, more effective and skilled techniques need to be explored.
3) *Optimal Learning Techniques for Game Theory:* Most of the existing ADP methods achieve the mini-max target of the value function. However, such a goal is sometimes impractical. For example, in game theory, it attempts to design an optimization policy to achieve a balanced outcome for the benefits of multiple players [71], [191], [198], [199], [228], [234], [235]. Therefore, the use of ADP for tackling the actual game problem is promising, which can promote the development of complex human-engineered systems.

4) *Further Work for Deep RL/ADP:* Deep RL/ADP combines the perception of deep learning and the decision making of RL/ADP, and it further extends RL/ADP to problems that were previously difficult to solve [236]. This novel approach brings artificial intelligence closer to the human mindset. Since it was proposed, deep RL/ADP has made remarkable achievements in applications, such as Atari 2600 video games [237], AlphaGo [238], robotics [239], [240], vehicle classification [241], and elevator group control [242]. However, the theoretical analysis of convergence and stability of deep RL/ADP is still a problem to be solved. Therefore, further research is needed to improve the framework of the deep RL/ADP method.
5) *New Ideas for Solving the HJB Equation:* The main focus of ADP is to solve the HJB equation and avoid the curse of dimensionality, both in discrete-time and continuous-time systems. In the past few decades, methods of direct solution and iterative solution have been proposed to obtain the solution of the HJB equation, but new ideas are still needed. There are some attempts in the literature, especially for finite-horizon optimal control problems [243], [244]. This is a direction definitely worth investigating by scholars in the future.

VI. Conclusion

In this article, a comprehensive overview of the ADP-based intelligent control methods was given. First, the progress in basic optimal control problems was introduced. Some widely used learning algorithms were listed to show the progress. Then, how this intelligent control method was used to solve games and large-scale systems were analyzed, respectively. After describing its applications in various aspects, it can be seen that it has a good prospect in today's era of artificial intelligence. Some subsequent possible works are given to promote the expected further advancement of the ADP-based intelligent control methods.

References

[1] M. C. Weinstein and R. J. Zeckhauser, "The optimal consumption of depletable natural resources," *Quart. J. Econ.*, vol. 89, no. 3, pp. 371–392, Aug. 1975.

[2] S. Papachristos, "Adaptive dynamic programming and inventory control," Ph.D. dissertation, Dept. Decis. Theory, Univ. Manchester, Manchester, U.K., 1977.

[3] S. Papachristos, "Note—A note on the dynamic inventory problem with unknown demand distribution," *Manag. Sci.*, vol. 23, no. 11, pp. 1248–1251, Jul. 1977.

[4] S. Shields, "A review of fault detection methods for large systems," *Radio Electron. Eng.*, vol. 46, no. 6, pp. 276–280, Jun. 1976.

[5] A. G. Barto, S. J. Bradtke, and S. P. Singh, "Learning to act using real-time dynamic programming," *Artif. Intell.*, vol. 72, nos. 1–2, pp. 81–138, Jan. 1995.

[6] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 32, no. 2, pp. 140–153, May 2002.

[7] J. J. Murray, C. J. Cox, and R. E. Saeks, "The adaptive dynamic programming theorem," in *Stability and Control of Dynamical Systems with Applications*, D. Liu and P. J. Antsaklis, Eds. Boston, MA, USA: Birkhäuser, 2003, ch. 19.

[8] W. H. Hausman and L. J. Thomas, "Inventory control with probabilistic demand and periodic withdrawals," *Manag. Sci.*, vol. 18, no. 5, pp. 265–275, Jan. 1972.

[9] P. J. Werbos, "Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research," *IEEE Trans. Syst., Man, Cybern.*, vol. 17, no. 1, pp. 7–20, Jan. 1987.

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

[11] P. J. Werbos, "A menu of designs for reinforcement learning over time," in *Neural Networks for Control*, W. T. Miller, R. S. Sutton, and P. J. Werbos, Eds. Cambridge, MA, USA: MIT Press, 1990, ch. 3.

[12] P. J. Werbos, "Consistency of HDP applied to a simple reinforcement learning problem," *Neural Netw.*, vol. 3, no. 2, pp. 179–189, Apr. 1990.

[13] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, D. A. White and D. A. Sofge, Eds. New York, NY, USA: Van Nostrand Reinhold, 1992, ch. 13.

[14] D. V. Prokhorov, R. A. Santiago, and D. C. Wunsch, "Adaptive critic designs: A case study for neurocontrol," *Neural Netw.*, vol. 8, no. 9, pp. 1367–1372, 1995.

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

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[17] A. G. Barto, "Reinforcement learning and adaptive critic methods," in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, D. A. White and D. A. Sofge, Eds. New York, NY, USA: Van Nostrand Reinhold, 1992, ch. 12.

[18] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, Aug. 1988.

[19] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 1989.

[20] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.

[21] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 237–285, May 1996.

[22] A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," *INFORMS J. Comput.*, vol. 21, no. 2, pp. 178–192, 2009.

[23] M. L. Littman, "Reinforcement learning improves behaviour from evaluative feedback," *Nature*, vol. 521, no. 7553, pp. 445–451, May 2015.

[24] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming: An overview," in *Proc. 34th IEEE Conf. Decis. Control*, New Orleans, LA, USA, Dec. 1995, pp. 560–564.

[25] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Belmont, MA, USA: Athena Sci., 1996.

[26] J. Si, L. Yang, and D. Liu, "Direct neural dynamic programming," in *Handbook of Learning and Approximate Dynamic Programming*, J. Si, A. G. Barto, W. B. Powell, and D. C. Wunsch, Eds. New York, NY, USA: Wiley, 2004, ch. 5.

[27] S. Chakraborty and M. G. Simoes, "Neural dynamic programming based online controller with a novel trim approach," *IEE Proc. Control Theory Appl.*, vol. 152, no. 1, pp. 95–104, Jan. 2005.

[28] D. Liu and H. Zhang, "A neural dynamic programming approach for learning control of failure avoidance problems," *Int. J. Intell. Control Syst.*, vol. 10, no. 1, pp. 21–32, Mar. 2005.

[29] D. P. Bertsekas, M. L. Homer, D. A. Logan, S. D. Patek, and N. R. Sandell, "Missile defense and interceptor allocation by neuro-dynamic programming," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 30, no. 1, pp. 42–51, Jan. 2000.

[30] P. Marbach, O. Mihatsch, and J. N. Tsitsiklis, "Call admission control and routing in integrated services networks using neuro-dynamic programming," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 2, pp. 197–208, Feb. 2000.

[31] D. Wang, C. Mu, H. He, and D. Liu, "Event-driven adaptive robust control of nonlinear systems with uncertainties through NDP strategy," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 7, pp. 1358–1370, Jul. 2017.

[32] C. Mu, D. Wang, and H. He, "Novel iterative neural dynamic programming for data-based approximate optimal control design," *Automatica*, vol. 81, pp. 240–252, Jul. 2017.

[33] M. Aoki, "On optimal and suboptimal policies in the choice of control forces for final-value systems," *IRE Trans. Autom. Control*, vol. 5, no. 3, pp. 171–178, Aug. 1960.

[34] R. Durbeck, "An approximation technique for suboptimal control," *IEEE Trans. Autom. Control*, vol. 10, no. 2, pp. 144–149, Apr. 1965.

[35] R. J. Leake and R.-W. Liu, "Construction of suboptimal control sequences," *SIAM J. Control*, vol. 5, no. 1, pp. 54–63, 1967.

[36] F.-Y. Wang and G. N. Saridis, "Suboptimal control for nonlinear stochastic systems," in *Proc. 31st IEEE Conf. Decis. Control*, Tucson, AZ, USA, Dec. 1992, pp. 1856–1861.
[37] G. N. Saridis and F.-Y. Wang, "Suboptimal control of nonlinear stochastic systems," *Control Theory Adv. Technol.*, vol. 10, no. 4, pp. 847–871, Dec. 1994.
[38] P. J. Werbos, "ADP: Goals, opportunities and principles," in *Handbook of Learning and Approximate Dynamic Programming*, J. Si, A. G. Barto, W. B. Powell, and D. C. Wunsch, Eds. New York, NY, USA: Wiley, 2004, ch. 1.
[39] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Hoboken, NJ, USA: Wiley, 2007.
[40] P. J. Werbos, "Using ADP to understand and replicate brain intelligence: The next level design," in *Proc. IEEE Int. Symp. Approx. Dyn. Program. Reinforcement Learn.*, Honolulu, HI, USA, Apr. 2007, pp. 209–216.
[41] P. J. Werbos, "Foreword—ADP: The key direction for future research in intelligent control and understanding brain intelligence," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 898–900, Aug. 2008.
[42] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Trans. Syst., Man, Cybern.*, vol. 13, no. 5, pp. 834–846, Sep./Oct. 1983.
[43] Y. Tang, H. He, Z. Ni, X. Zhong, D. Zhao, and X. Xu, "Fuzzy-based goal representation adaptive dynamic programming," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 5, pp. 1159–1175, Oct. 2016.
[44] D. Wang and D. Liu, "Neuro-optimal control for a class of unknown nonlinear dynamic systems using SN-DHP technique," *Neurocomputing*, vol. 121, pp. 218–225, Dec. 2013.
[45] M. Ha, D. Wang, and D. Liu, "Event-triggered constrained control with DHP implementation for nonaffine discrete-time systems," *Inf. Sci.*, vol. 519, pp. 110–123, May 2020.
[46] D. Liu, D. Wang, D. Zhao, Q. Wei, and N. Jin, "Neural-network-based optimal control for a class of unknown discrete-time nonlinear systems using globalized dual heuristic programming," *IEEE Trans. Autom. Sci. Eng.*, vol. 9, no. 3, pp. 628–634, Jul. 2012.
[47] M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb, "Continuous-time Q-learning for infinite-horizon discounted cost linear quadratic regulator problems," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 165–176, Feb. 2015.
[48] Q. Wei, D. Liu, and G. Shi, "A novel dual iterative Q-learning method for optimal battery management in smart residential environments," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2509–2518, Apr. 2015.
[49] Q. Wei, F. L. Lewis, Q. Sun, P. Yan, and R. Song, "Discrete-time deterministic Q-learning: A novel convergence analysis," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1224–1237, May 2017.
[50] P. Yan, D. Wang, H. Li, and D. Liu, "Error bound analysis of Q-function for discounted optimal control problems with policy iteration," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 7, pp. 1207–1216, Jul. 2017.
[51] H. He, Z. Ni, and J. Fu, "A three-network architecture for on-line learning and optimization based on adaptive dynamic programming," *Neurocomputing*, vol. 78, no. 1, pp. 3–13, Feb. 2012.
[52] Z. Ni, H. He, and J. Wen, "Adaptive learning in tracking control based on the dual critic network design," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 913–928, Jun. 2013.
[53] X. Zhong, Z. Ni, and H. He, "A theoretical foundation of goal representation heuristic dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2513–2525, Dec. 2016.
[54] Z. Ni, H. He, J. Wen, and X. Xu, "Goal representation heuristic dynamic programming on maze navigation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2038–2050, Dec. 2013.
[55] Y. Tang, J. Yang, J. Yan, Z. Zeng, and H. He, "Intelligent load frequency controller using GrADP for island smart grid with electric vehicles and renewable resources," *Neurocomputing*, vol. 170, pp. 406–416, Dec. 2015.
[56] C. Mu, Z. Ni, C. Sun, and H. He, "Data-driven tracking control with adaptive dynamic programming for a class of continuous-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1460–1470, Jun. 2017.
[57] J. Si and Y.-T. Wang, "Online learning control by association and reinforcement," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 264–276, Mar. 2001.
[58] F. Liu, J. Sun, J. Si, W. Guo, and S. Mei, "A boundedness result for the direct heuristic dynamic programming," *Neural Netw.*, vol. 32, pp. 229–235, Aug. 2012.
[59] X. Xu, Z. Hou, C. Lian, and H. He, "Online learning control using adaptive critic designs with sparse kernel machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 762–775, May 2013.
[60] Y. Jiang and Z.-P. Jiang, "Global adaptive dynamic programming for continuous-time nonlinear systems," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2917–2929, Nov. 2015.
[61] Q. Wei, D. Liu, Y. Liu, and R. Song, "Optimal constrained self-learning battery sequential management in microgrid via adaptive dynamic programming," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 2, pp. 168–176, Apr. 2017.
[62] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.
[63] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, 3rd Quart., 2009.
[64] F. L. 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.
[65] 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.
[66] Y. Zhu and D. Zhao, "Comprehensive comparison of online ADP algorithms for continuous-time optimal control," *Artif. Intell. Rev.*, vol. 49, no. 4, pp. 531–547, Apr. 2018.
[67] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, U.K.: Springer, 2013.
[68] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ, USA: Wiley, 2013.
[69] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming With Applications in Optimal Control*. Cham, Switzerland: Springer, 2017.
[70] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, Feb. 2009.
[71] K. G. Vamvoudakis and F. L. Lewis, "Online solution of nonlinear two-player zero-sum games using synchronous policy iteration," *Int. J. Robust Nonlinear Control*, vol. 22, no. 13, pp. 1460–1483, Sep. 2012.
[72] 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.
[73] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design," *Automatica*, vol. 50, no. 12, pp. 3281–3290, Dec. 2014.
[74] 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.
[75] B. Zhao, D. Wang, G. Shi, D. Liu, and Y. Li, "Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 10, pp. 1725–1735, Oct. 2018.
[76] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.
[77] D. Liu, H. Li, and D. Wang, "Error bounds of adaptive dynamic programming algorithms for solving undiscounted optimal control problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1323–1334, Jun. 2015.
[78] 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.
[79] C. Mu, C. Sun, A. Song, and H. Yu, "Iterative GDHP-based approximate optimal tracking control for a class of discrete-time nonlinear systems," *Neurocomputing*, vol. 214, pp. 775–784, Nov. 2016.
[80] Q. Wei, D. Liu, Q. Lin, and R. Song, "Discrete-time optimal control via local policy iteration adaptive dynamic programming," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3367–3379, Oct. 2017.
[81] Q. Wei, D. Liu, and Q. Lin, "Discrete-time local value iteration adaptive dynamic programming: Admissibility and termination analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2490–2502, Nov. 2017.

[82] Q. Wei, F. L. Lewis, D. Liu, R. Song, and H. Lin, "Discrete-time local value iteration adaptive dynamic programming: Convergence analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 6, pp. 875–891, Jun. 2018.

[83] 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, Sep. 2018.

[84] B. Lincoln and A. Rantzer, "Relaxing dynamic programming," *IEEE Trans. Autom. Control*, vol. 51, no. 8, pp. 1249–1260, Aug. 2006.

[85] A. Rantzer, "Relaxed dynamic programming in switching systems," *IEE Proc. Control Theory Appl.*, vol. 153, no. 5, pp. 567–574, Sep. 2006.

[86] D. Liu and Q. Wei, "Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 779–789, Apr. 2013.

[87] W. Guo, J. Si, F. Liu, and S. Mei, "Policy approximation in policy iteration approximate dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2794–2807, Jul. 2018.

[88] K. G. Vamvoudakis, D. Vrabie, and F. L. Lewis, "Online adaptive algorithm for optimal control with integral reinforcement learning," *Int. J. Robust Nonlinear Control*, vol. 24, no. 17, pp. 2686–2710, Nov. 2014.

[89] R. Song, F. L. Lewis, Q. Wei, and H. Zhang, "Off-policy actor-critic structure for optimal control of unknown systems with disturbances," *IEEE Trans. Cybern.*, vol. 46, no. 5, pp. 1041–1050, May 2016.

[90] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst., Man, Cybern, C, Appl. Rev.*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012.

[91] B. Luo, D. Liu, H.-N. Wu, D. Wang, and F. L. Lewis, "Policy gradient adaptive dynamic programming for data-based optimal control," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3341–3354, Oct. 2017.

[92] 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.

[93] Q. Wei, B. Li, and R. Song, "Discrete-time stable generalized self-learning optimal control with approximation errors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1226–1238, Apr. 2018.

[94] 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.

[95] Q. Wei, F.-Y. Wang, D. Liu, and X. Yang, "Finite-approximation-error-based discrete-time iterative adaptive dynamic programming," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2820–2833, Dec. 2014.

[96] 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.

[97] B. Luo, D. Liu, T. Huang, X. Yang, and H. Ma, "Multi-step heuristic dynamic programming for optimal control of nonlinear discrete-time systems," *Inf. Sci.*, vol. 411, pp. 66–83, Oct. 2017.

[98] B. Luo, D. Liu, T. Huang, and J. Liu, "Output tracking control based on adaptive dynamic programming with multistep policy evaluation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 10, pp. 2155–2165, Oct. 2019.

[99] J. Y. Lee, J. B. Park, and Y. H. Choi, "On integral generalized policy iteration for continuous-time linear quadratic regulations," *Automatica*, vol. 50, no. 2, pp. 475–489, Feb. 2014.

[100] 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.

[101] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish, "Reinforcement learning and optimal adaptive control: An overview and implementation examples," *Annu. Rev. Control*, vol. 36, no. 1, pp. 42–59, Apr. 2012.

[102] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. 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.

[103] D. Nodland, H. Zargarzadeh, and S. Jagannathan, "Neural network-based optimal adaptive output feedback control of a helicopter UAV," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1061–1073, Jul. 2013.

[104] D. Liu, D. Wang, F.-Y. Wang, H. Li, and X. Yang, "Neural-network-based online HJB solution for optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2834–2847, Dec. 2014.

[105] D. Wang, H. He, and D. Liu, "Intelligent optimal control with critic learning for a nonlinear overhead crane system," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 2932–2940, Jul. 2018.

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

[107] V. G. Lopez and F. L. Lewis, "Dynamic multiobjective control for continuous-time systems using reinforcement learning," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2869–2874, Jul. 2019.

[108] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, Jan. 2014.

[109] D. Liu, D. Wang, and X. Yang, "An iterative adaptive dynamic programming algorithm for optimal control of unknown discrete-time nonlinear systems with constrained inputs," *Inf. Sci.*, vol. 220, pp. 331–342, Jan. 2013.

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

[111] X. Yang, D. Liu, and Y. Huang, "Neural-network-based online optimal control for uncertain non-linear continuous-time systems with control constraints," *IET Control Theory Appl.*, vol. 7, no. 17, pp. 2037–2047, Nov. 2013.

[112] X. Yang, D. Liu, and D. Wang, "Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints," *Int. J. Control*, vol. 87, no. 3, pp. 553–566, Mar. 2014.

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

[114] S. Yasini, M. B. N. Sitani, and A. Kirampor, "Reinforcement learning and neural networks for multi-agent nonzero-sum games of nonlinear constrained-input systems," *Int. J. Mach. Learn. Cybern.*, vol. 7, no. 6, pp. 967–980, Dec. 2016.

[115] H. Zhang, G. Xiao, Y. Liu, and L. Liu, "Value iteration-based H_∞ controller design for continuous-time nonlinear systems subject to input constraints," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 11, pp. 3986–3995, Nov. 2020.

[116] L. N. Tan, "Event-triggered distributed H_∞ constrained control of physically interconnected large-scale partially unknown strict-feedback systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, May 13, 2019, doi: 10.1109/TSMC.2019.2914160.

[117] B. Zhao, D. Liu, and Y. Li, "Online fault compensation control based on policy iteration algorithm for a class of affine non-linear systems with actuator failures," *IET Control Theory Appl.*, vol. 10, no. 15, pp. 1816–1823, Oct. 2016.

[118] B. Zhao, D. Liu, and Y. Li, "Observer based adaptive dynamic programming for fault tolerant control of a class of nonlinear systems," *Inf. Sci.*, vol. 384, pp. 21–33, Apr. 2017.

[119] Q. Qu, H. Zhang, R. Yu, and Y. Liu, "Neural network-based H_∞ sliding mode control for nonlinear systems with actuator faults and unmatched disturbances," *Neurocomputing*, vol. 275, pp. 2009–2018, Jan. 2018.

[120] B. Zhao, L. Jia, H. Xia, and Y. Li, "Adaptive dynamic programming-based stabilization of nonlinear systems with unknown actuator saturation," *Nonlinear Dyn.*, vol. 93, no. 4, pp. 2089–2103, Sep. 2018.

[121] H. Zhang, Y. Liang, H. Su, and C. Liu, "Event-driven guaranteed cost control design for nonlinear systems with actuator faults via reinforcement learning algorithm," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 11, pp. 4135–4150, Nov. 2020.

[122] H. Lin, B. Zhao, D. Liu, and C. Alippi, "Data-based fault tolerant control for affine nonlinear systems through particle swarm optimized neural networks," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 954–964, Jul. 2020.

[123] H. Modares, F. L. Lewis, and Z.-P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2550–2562, Oct. 2015.

[124] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Reinforcement learning solution for HJB equation arising in constrained optimal control problem," *Neural Netw.*, vol. 71, pp. 150–158, Nov. 2015.

[125] B. Kiumarsi, W. Kang, and F. L. Lewis, "H_∞ control of nonaffine aerial systems using off-policy reinforcement learning," *Unmanned Syst.*, vol. 4, no. 1, pp. 51–60, Feb. 2016.

[126] H. Modares, S. P. Nageshrao, G. A. D. Lopes, R. Babuška, and F. L. Lewis, "Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning," *Automatica*, vol. 71, pp. 334–341, Sep. 2016.

[127] H. Modares, F. L. Lewis, and Z.-P. Jiang, "Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning," *IEEE Trans. Cybern.*, vol. 46, no. 11, pp. 2401–2410, Nov. 2016.

[128] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 704–713, Mar. 2017.

[129] J. Li, B. Kiumarsi, T. Chai, F. L. Lewis, and J. Fan, "Off-policy reinforcement learning: Optimal operational control for two-time-scale industrial processes," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4547–4558, Dec. 2017.

[130] J. Škach, B. Kiumarsi, F. L. Lewis, and O. Straka, "Actor-critic off-policy learning for optimal control of multiple-model discrete-time systems," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 29–40, Jan. 2018.

[131] J. Qin, M. Li, Y. Shi, Q. Ma, and W. X. Zheng, "Optimal synchronization control of multiagent systems with input saturation via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 85–96, Jan. 2019.

[132] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for H_∞ control design," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 65–76, Jan. 2015.

[133] G. Chowdhary and E. Johnson, "Concurrent learning for convergence in adaptive control without persistency of excitation," in *Proc. 49th IEEE Conf. Decis. Control*, Atlanta, GA, USA, Dec. 2010, pp. 3674–3679.

[134] H. Xu, S. Jagannathan, and F. L. Lewis, "Stochastic optimal control of unknown linear networked control system in the presence of random delays and packet losses," *Automatica*, vol. 48, no. 6, pp. 1017–1030, Jun. 2012.

[135] D. Zhao, Q. Zhang, D. Wang, and Y. Zhu, "Experience replay for optimal control of nonzero-sum game systems with unknown dynamics," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 854–865, Mar. 2016.

[136] B. Luo, D. Liu, and H.-N. Wu, "Adaptive constrained optimal control design for data-based nonlinear discrete-time systems with critic-only structure," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2099–2111, Jun. 2018.

[137] K. J. Åström and B. Wittenmark, *Adaptive Control*. Mineola, NY, USA: Courier Corp., 2013.

[138] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Netw.*, vol. 3, no. 5, pp. 551–560, 1990.

[139] B. Zhao, D. Liu, and C. Luo, "Reinforcement learning-based optimal stabilization for unknown nonlinear systems subject to inputs with uncertain constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4330–4340, Oct. 2020.

[140] R. Padhi, N. Unnikrishnan, X. Wang, and S. N. Balakrishnan, "A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems," *Neural Netw.*, vol. 19, no. 10, pp. 1648–1660, Dec. 2006.

[141] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, Jan. 2013.

[142] H. K. Khalil, *Nonlinear Systems*. London, U.K.: Prentice-Hall, 1996.

[143] D. Wang, C. Mu, D. Liu, and H. Ma, "On mixed data and event driven design for adaptive-critic-based nonlinear H_∞ control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 993–1005, Apr. 2018.

[144] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks," *IEEE Trans. Autom. Control*, vol. 52, no. 9, pp. 1680–1685, Sep. 2007.

[145] K. G. Vamvoudakis, "Event-triggered optimal adaptive control algorithm for continuous-time nonlinear systems," *IEEE/CAA J. Autom. Sinica*, vol. 1, no. 3, pp. 282–293, Jul. 2014.

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

[147] K. G. Vamvoudakis, A. Mojoodi, and H. Ferraz, "Event-triggered optimal tracking control of nonlinear systems," *Int. J. Robust Nonlinear Control*, vol. 27, no. 4, pp. 598–619, Mar. 2017.

[148] X. Zhong and H. He, "An event-triggered ADP control approach for continuous-time system with unknown internal states," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 683–694, Mar. 2017.

[149] Y. Zhu, D. Zhao, H. He, and J. Ji, "Event-triggered optimal control for partially unknown constrained-input systems via adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4101–4109, May 2017.

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

[151] D. Wang, H. He, X. Zhong, and D. Liu, "Event-driven nonlinear discounted optimal regulation involving a power system application," *IEEE Trans. Ind. Electron.*, vol. 64, no. 10, pp. 8177–8186, Oct. 2017.

[152] X. Yang, H. He, and D. Liu, "Event-triggered optimal neuro-controller design with reinforcement learning for unknown nonlinear systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 9, pp. 1866–1878, Sep. 2019.

[153] 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.

[154] X. Yang and H. He, "Adaptive critic learning and experience replay for decentralized event-triggered control of nonlinear interconnected systems," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 50, no. 11, pp. 4043–4055, Nov. 2020.

[155] W. Gao, Y. Jiang, Z.-P. Jiang, and T. Chai, "Adaptive and optimal output feedback control of linear systems: An adaptive dynamic programming approach," in *Proc. 11th World Congr. Intell. Control Autom.*, Shenyang, China, Jun/Jul. 2014, pp. 2085–2090.

[156] 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.

[157] F. A. Yaghmaie, S. Gunnarsson, and F. L. Lewis, "Output regulation of unknown linear systems using average cost reinforcement learning," *Automatica*, vol. 110, Dec. 2019, Art. no. 108549.

[158] D. Liu, Y. Huang, D. Wang, and Q. Wei, "Neural-network-observer-based optimal control for unknown nonlinear systems using adaptive dynamic programming," *Int. J. Control*, vol. 86, no. 9, pp. 1554–1566, Sep. 2013.

[159] X. Yang, D. Liu, and Q. Wei, "Online approximate optimal control for affine non-linear systems with unknown internal dynamics using adaptive dynamic programming," *IET Control Theory Appl.*, vol. 8, no. 16, pp. 1676–1688, Nov. 2014.

[160] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, Oct. 2012.

[161] W. Gao and Z.-P. Jiang, "Adaptive optimal output regulation of time-delay systems via measurement feedback," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 938–945, Mar. 2019.

[162] L. M. Zhu, H. Modares, G. O. Peen, F. L. Lewis, and B. Yue, "Adaptive suboptimal output-feedback control for linear systems using integral reinforcement learning," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 1, pp. 264–273, Jan. 2015.

[163] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 3051–3056, Nov. 2014.

[164] C. Chen, H. Modares, K. Xie, F. L. 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.

[165] C. Li, D. Liu, and H. Li, "Finite horizon optimal tracking control of partially unknown linear continuous-time systems using policy iteration," *IET Control Theory Appl.*, vol. 9, no. 12, pp. 1791–1801, Aug. 2015.

[166] K. Zhang, H. Zhang, G. Xiao, and H. Su, "Tracking control optimization scheme of continuous-time nonlinear system via online single network adaptive critic design method," *Neurocomputing*, vol. 251, pp. 127–135, Aug. 2017.

[167] W. Gao and Z.-P. Jiang, "Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2614–2624, Jun. 2018.

[168] D. Wang, D. Liu, Q. Wei, D. Zhao, and N. Jin, "Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming," *Automatica*, vol. 48, no. 8, pp. 1825–1832, Aug. 2012.

[169] Q. Wei and D. Liu, "A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1176–1190, Oct. 2014.

[170] X. Zhong, H. He, H. Zhang, and Z. Wang, "Optimal control for unknown discrete-time nonlinear Markov jump systems using adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2141–2155, Dec. 2014.

[171] Q. Wei, D. Liu, and X. Yang, "Infinite horizon self-learning optimal control of nonaffine discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 866–879, Apr. 2015.

[172] 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.

[173] Y. Zhang, B. Zhao, and D. Liu, "Deterministic policy gradient adaptive dynamic programming for model-free optimal control," *Neurocomputing*, vol. 387, pp. 40–50, Apr. 2020.

[174] M. Liang, D. Wang, and D. Liu, "Improved value iteration for neural-network-based stochastic optimal control design," *Neural Netw.*, vol. 124, pp. 280–295, Apr. 2020.

[175] Q. Yang and S. Jagannathan, "Reinforcement learning controller design for affine nonlinear discrete-time systems using online approximators," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 377–390, Apr. 2012.

[176] B. Luo, H.-N. Wu, and T. Huang, "Optimal output regulation for model-free quanser helicopter with multistep Q-learning," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 4953–4961, Jun. 2018.

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

[178] Y. Jiang, B. Kiumarsi, J. Fan, T. Chai, J. Li, and L. 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.

[179] B. Kiumarsi, F. L. Lewis, M.-B. Naghibi-Sistani, and A. Karimpour, "Optimal tracking control of unknown discrete-time linear systems using input-output measured data," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2770–2779, Dec. 2015.

[180] D. Wang, D. Liu, and Q. Wei, "Finite-horizon neuro-optimal tracking control for a class of discrete-time nonlinear systems using adaptive dynamic programming approach," *Neurocomputing*, vol. 78, no. 1, pp. 14–22, Feb. 2012.

[181] H. Zhang, R. Song, Q. Wei, and T. Zhang, "Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 1851–1862, Dec. 2011.

[182] B. Kiumarsi and F. L. 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.

[183] Y. Batmani, M. Davoodi, and N. Meskin, "Event-triggered suboptimal tracking controller design for a class of nonlinear discrete-time systems," *IEEE Trans. Ind. Electron.*, vol. 64, no. 10, pp. 8079–8087, Oct. 2017.

[184] Y. Huang and D. Liu, "Neural-network-based optimal tracking control scheme for a class of unknown discrete-time nonlinear systems using iterative ADP algorithm," *Neurocomputing*, vol. 125, pp. 46–56, Feb. 2014.

[185] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167–1175, Apr. 2014.

[186] X. Yang, D. Liu, D. Wang, and Q. Wei, "Discrete-time online learning control for a class of unknown nonaffine nonlinear systems using reinforcement learning," *Neural Netw.*, vol. 55, pp. 30–41, Jul. 2014.

[187] Q. Wei and D. Liu, "Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1020–1036, Oct. 2014.

[188] Q. Wei and D. Liu, "Neural-network-based adaptive optimal tracking control scheme for discrete-time nonlinear systems with approximation errors," *Neurocomputing*, vol. 149, pp. 106–115, Feb. 2015.

[189] B. Luo, D. Liu, T. Huang, and D. Wang, "Model-free optimal tracking control via critic-only Q-learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2134–2144, Oct. 2016.

[190] B. Kiumarsi, B. AlQaudi, H. Modares, F. L. Lewis, and D. S. Levine, "Optimal control using adaptive resonance theory and Q-learning," *Neurocomputing*, vol. 361, pp. 119–125, Oct. 2019.

[191] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. New York, NY, USA: Academic, 1982.

[192] T. Mylvaganam, M. Sassano, and A. Astolfi, "Constructive ϵ-nash equilibria for nonzero-sum differential games," *IEEE Trans. Autom. Control*, vol. 60, no. 4, pp. 950–965, Apr. 2015.

[193] A. J. van der Schaft, "$L2$-gain analysis of nonlinear systems and nonlinear state-feedback H_∞ control," *IEEE Trans Autom. Control*, vol. 37, no. 6, pp. 770–784, Jun. 1992.

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

[195] T. Başar and P. Bernhard, *H_∞ Optimal Control and Related Minimax Design Problems*. Boston, MA, USA: Birkhäuser, 1995.

[196] J. B. Burl, *Linear Optimal Control: H_2 and H_∞ Methods*. Menlo Park, CA, USA: Addison-Wesley, 1998.

[197] M. Sassano and A. Astolfi, "Dynamic approximate solutions of the HJ inequality and of the HJB equation for input-affine nonlinear systems," *IEEE Trans. Autom. Control*, vol. 57, no. 10, pp. 2490–2503, Oct. 2012.

[198] K. G. Vamvoudakis and F. L. Lewis, "Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations," *Automatica*, vol. 47, no. 8, pp. 1556–1569, Aug. 2011.

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

[200] R. Kamalapurkar, J. R. Klotz, and W. E. Dixon, "Concurrent learning-based approximate feedback-nash equilibrium solution of N-player nonzero-sum differential games," *IEEE/CAA J. Autom. Sinica*, vol. 1, no. 3, pp. 239–247, Jul. 2014.

[201] K. G. Vamvoudakis, "Non-zero sum nash Q-learning for unknown deterministic continuous-time linear systems," *Automatica*, vol. 61, pp. 274–281, Nov. 2015.

[202] X. Li, Z. Peng, L. Liang, and W. Zha, "Policy iteration based Q-learning for linear nonzero-sum quadratic differential games," *Sci. China Inf. Sci.*, vol. 62, no. 5, May 2019, Art. no. 52204.

[203] W. Wang, X. Chen, H. Fu, and M. Wu, "Data-driven adaptive dynamic programming for partially observable nonzero-sum games via Q-learning method," *Int. J. Syst. Sci.*, vol. 50, no. 7, pp. 1338–1352, May 2019.

[204] H. Jiang and Y. Luo, "Data-driven approximate optimal tracking control schemes for unknown non-affine non-linear multi-player systems via adaptive dynamic programming," *Electron. Lett.*, vol. 53, no. 7, pp. 465–467, Mar. 2017.

[205] H. Jiang, H. Zhang, Y. Luo, and J. Han, "Neural-network-based robust control schemes for nonlinear multiplayer systems with uncertainties via adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 3, pp. 579–588, Mar. 2019.

[206] A. Sahoo, V. Narayanan, and S. Jagannathan, "Event-triggered control of N-player nonlinear systems using nonzero-sum games," in *Proc. 8th IEEE Symp. Series Comput. Intell.*, Bangalore, India, Nov. 2018, pp. 1447–1452.

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

[208] W. Chen and J. Li, "Decentralized output-feedback neural control for systems with unknown interconnections," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 1, pp. 258–266, Feb. 2008.

[209] A. Saberi, "On optimality of decentralized control for a class of nonlinear interconnected systems," *Automatica*, vol. 24, no. 1, pp. 101–104, Jan. 1988.

[210] D. Liu, D. Wang, and H. Li, "Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 418–428, Feb. 2014.

[211] D. Liu, C. Li, H. Li, D. Wang, and H. Ma, "Neural-network-based decentralized control of continuous-time nonlinear interconnected systems with unknown dynamics," *Neurocomputing*, vol. 165, pp. 90–98, Oct. 2015.

[212] D. Wang, D. Liu, C. Mu, and H. Ma, "Decentralized guaranteed cost control of interconnected systems with uncertainties: A learning-based

optimal control strategy," *Neurocomputing*, vol. 214, pp. 297–306, Nov. 2016.
[213] X. Yang and H. He, "Adaptive dynamic programming for decentralized stabilization of uncertain nonlinear large-scale systems with mismatched interconnections," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 8, pp. 2870–2880, Aug. 2020.
[214] X. Yang and H. He, "Adaptive critic designs for optimal control of uncertain nonlinear systems with unmatched interconnections," *Neural Netw.*, vol. 105, pp. 142–153, Sep. 2018.
[215] D. Wang, D. Liu, H. Li, H. Ma, and C. Li, "A neural-network-based online optimal control approach for nonlinear robust decentralized stabilization," *Soft Comput.*, vol. 20, no. 2, pp. 707–716, Feb. 2016.
[216] Q. Wu, B. Zhao, and D. Liu, "Adaptive dynamic programming-based decentralised control for large-scale nonlinear systems subject to mismatched interconnections with unknown time-delay," *Int. J. Syst. Sci.*, vol. 51, no. 15, pp. 2883–2898, 2020.
[217] B. Zhao and Y. Li, "Local joint information based active fault tolerant control for reconfigurable manipulator," *Nonlinear Dyn.*, vol. 77, no. 3, pp. 859–876, Aug. 2014.
[218] B. Zhao, Y. Li, and D. Liu, "Self-tuned local feedback gain based decentralized fault tolerant control for a class of large-scale nonlinear systems," *Neurocomputing*, vol. 235, pp. 147–156, Apr. 2017.
[219] D. Ye and T. Song, "Decentralized reliable guaranteed cost control for large-scale nonlinear systems using actor-critic network," *Neurocomputing*, vol. 320, pp. 121–128, Dec. 2018.
[220] B. Zhao, D. Liu, X. Yang, and Y. Li, "Observer-critic structure-based adaptive dynamic programming for decentralised tracking control of unknown large-scale nonlinear systems," *Int. J. Syst. Sci.*, vol. 48, no. 9, pp. 1978–1989, May 2017.
[221] Q. Qu, H. Zhang, T. Feng, and H. Jiang, "Decentralized adaptive tracking control scheme for nonlinear large-scale interconnected systems via adaptive dynamic programming," *Neurocomputing*, vol. 225, pp. 1–10, Feb. 2017.
[222] B. Zhao and Y. Li, "Model-free adaptive dynamic programming based near-optimal decentralized tracking control of reconfigurable manipulators," *Int. J. Control Autom. Syst.*, vol. 16, no. 2, pp. 478–490, Apr. 2018.
[223] J. Sun and C. Liu, "Decentralised zero-sum differential game for a class of large-scale interconnected systems via adaptive dynamic programming," *Int. J. Control*, vol. 92, no. 12, pp. 2917–2927, Dec. 2019.
[224] B. Zhao and D. Liu, "Event-triggered decentralized tracking control of modular reconfigurable robots through adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 67, no. 4, pp. 3054–3064, Apr. 2020.
[225] S. Mehraeen and S. Jagannathan, "Decentralized optimal control of a class of interconnected nonlinear discrete-time systems by using online Hamilton–Jacobi–Bellman formulation," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1757–1769, Nov. 2011.
[226] 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.
[227] Z. Peng, Y. Zhao, J. Hu, and B. K. Ghosh, "Data-driven optimal tracking control of discrete-time multi-agent systems with two-stage policy iteration algorithm," *Inf. Sci.*, vol. 481, pp. 189–202, May 2019.
[228] H. Jiang and H. He, "Data-driven distributed output consensus control for partially observable multiagent systems," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 848–858, Mar. 2019.
[229] H. Zhang, D. Yue, C. Dou, W. Zhao, and X. Xie, "Data-driven distributed optimal consensus control for unknown multiagent systems with input-delay," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2095–2105, Jun. 2019.
[230] S. Xue, B. Luo, and D. Liu, "Integral reinforcement learning based event-triggered control with input saturation," *Neural Netw.*, vol. 131, pp. 144–153, Nov. 2020.
[231] S. Xue, B. Luo, and D. Liu, "Event-triggered adaptive dynamic programming for zero-sum game of partially unknown continuous-time nonlinear systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 9, pp. 3189–3199, Sep. 2020.
[232] S. Xue, B. Luo, D. Liu, and Y. Yang, "Constrained event-triggered H_∞ control based on adaptive dynamic programming with concurrent learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Jun. 12, 2020, doi: 10.1109/TSMC.2020.2997559.
[233] S. Xue, B. Luo, and D. Liu, "Event-triggered adaptive dynamic programming for unmatched uncertain nonlinear continuous-time systems," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 28, 2020, doi: 10.1109/TNNLS.2020.3009015.
[234] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Syst. Mag.*, vol. 37, no. 1, pp. 33–52, Feb. 2017.
[235] Q. Wei, D. Liu, Q. Lin, and R. Song, "Adaptive dynamic programming for discrete-time zero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 957–969, Apr. 2018.
[236] D. Zhao, D. Liu, F. L. Lewis, J. C. Principe, and S. Squartini, "Special issue on deep reinforcement learning and adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2038–2041, Jun. 2018.
[237] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[238] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
[239] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 39, pp. 1–40, Apr. 2016.
[240] H. Li, Q. Zhang, and D. Zhao, "Deep reinforcement learning-based automatic exploration for navigation in unknown environment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 2064–2076, Jun. 2020.
[241] D. Zhao, Y. Chen, and L. Lv, "Deep reinforcement learning with visual attention for vehicle classification," *IEEE Trans. Cogn. Develop. Syst.*, vol. 9, no. 4, pp. 356–367, Dec. 2017.
[242] Q. Wei, L. Wang, Y. Liu, and M. M. Polycarpou, "Optimal elevator group control via deep asynchronous actor–critic learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5245–5256, Dec. 2020.
[243] F.-Y. Wang, N. Jin, D. Liu, and Q. Wei, "Adaptive dynamic programming for finite-horizon optimal control of discrete-time nonlinear systems with ϵ-error bound," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 24–36, Jan. 2011.
[244] Q. Wei and D. Liu, "An iterative ϵ-optimal control scheme for a class of discrete-time nonlinear systems with unfixed initial state," *Neural Netw.*, vol. 32, pp. 236–244, Aug. 2012.

Derong Liu (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 1994.

He was a Staff Fellow with General Motors Research and Development Center, Warren, MI, USA, from 1993 to 1995. He was an Assistant Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA, from 1995 to 1999. He joined the University of Illinois at Chicago, Chicago, IL, USA, 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, Beijing, China, in 2008, where he served as the Associate Director of the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation from 2010 to 2015. He has published 19 books.

Prof. Liu received the Faculty Early Career Development Award from the National Science Foundation in 1999, the University Scholar Award from the 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 TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS Outstanding Paper Award in 2018, the IEEE SMC Society Andrew P. Sage Best Transactions Paper Award in 2018, and the IEEE/CCA JOURNAL OF AUTOMATICA SINICA Hsue-Shen Tsien Paper Award in 2019. 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 is a Fellow of the International Neural Network Society and the International Association of Pattern Recognition.

Shan Xue received the B.E. degree in measuring and control technology and instrumentations from the Anhui University of Technology, Ma'anshan, China, in 2016, and the M.E. degree in control science and engineering from the University of Science and Technology Beijing, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree in computer science and technology with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China.

Her current research interests include adaptive dynamic programming, reinforcement learning, event-triggering mechanism, and neural networks.

Bo Zhao (Senior Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control science and engineering from Jilin University, Changchun, China, in 2009 and 2014, respectively.

He was a Postdoctoral Fellow with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 2014 to 2017. Then, he joined the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences from 2017 to 2018. He is currently an Associate Professor with the School of Systems Science, Beijing Normal University, Beijing. He has authored or coauthored over 90 journal and conference articles. His research interests include adaptive dynamic programming, robot control, fault diagnosis and tolerant control, optimal control, and artificial intelligence-based control.

Dr. Zhao was the Secretary of the Adaptive Dynamic Programming and Reinforcement Learning Technical Committee of Chinese Association of Automation (CAA) and the Secretary of 2017 the 24th International Conference on Neural Information Processing. He is an Asia–Pacific Neural Network Society Member and a CAA Member.

Biao Luo (Senior Member, IEEE) received the Ph.D. degree in control science and engineering from Beihang University, Beijing, China, in 2014.

He is currently a Professor with the School of Automation, Central South University, Changsha, China. He was an Associate Professor and an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, from 2014 to 2018. His current research interests include distributed parameter systems, intelligent control, reinforcement learning, deep learning, and computational intelligence.

Dr. Luo was a recipient of the Chinese Association of Automation Outstanding Ph.D. Dissertation Award in 2015. He serves as an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, *Artificial Intelligence Review*, *Neurocomputing*, and *Journal of Industrial and Management Optimization*. He is the Secretariat of the Adaptive Dynamic Programming and Reinforcement Learning Technical Committee, Chinese Association of Automation.

Qinglai Wei (Senior Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2002 and 2009, respectively.

From 2009 to 2011, he was a Postdoctoral Fellow with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is currently a Professor and the Associate Director. He has authored four books, and published over 80 international journal papers. His research interests include adaptive dynamic programming, neural-networks-based control, optimal control, nonlinear systems, and their industrial applications.

Prof. Wei was a recipient of the Young Researcher Award of Asia–Pacific Neural Network Society. He is the associate editor of several SCI journals.