IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 29, 2022, accepted May 12, 2022, date of publication May 17, 2022, date of current version May 27, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175828

Integral Reinforcement Learning for Tracking in a Class of Partially Unknown Linear Systems With Output Constraints and External Disturbances

CHUNBIN QIN[1], **JINGUANG WANG**[1], **XIAOPENG QIAO**[1], **HEYANG ZHU**[1], **DEHUA ZHANG**[1], **AND YONGHANG YAN**[2]

[1]School of Artificial Intelligence, Henan University, Zhengzhou, Henan 450000, China
[2]School of Computer and Information Engineering, Henan University, Zhengzhou, Henan 450000, China

Corresponding author: Yonghang Yan (yanyonghang@henu.edu.cn)

This work was supported in part by the Youth Backbone Teachers in Colleges and Universities of Henan Province under Grant 2018GGJS017, and in part by the Science and Technology Research Project of the Henan Province under Grant 222102240014.

ABSTRACT In this paper, the H_∞ tracking control problem of partially unknown linear systems with output constraints and disturbance is studied by the reinforcement learning (RL) method. Firstly, an augmented system is established based on the reference trajectory dynamics and target system dynamics, and a special cost function is established to realize asymptotic tracking. In addition, the barrier function (BF) is used to transform the augmented system, and the output constraints is realized simultaneously by minimizing the quadratic cost function of the transformed system. Using only the obtained data and part of the system dynamics, the optimal control strategy and the worst disturbance strategy are obtained by using the integral reinforcement learning (IRL). Rigorous stability analysis shows that the proposed method can make the trajectory of system states converge, and the output of the control strategy can make the tracking error asymptotically stable. Finally, a simulation example is conducted to verify the effectiveness of the proposed algorithm.

INDEX TERMS Barrier function, H_∞ tracking control, integral reinforcement learning, output constraints.

I. INTRODUCTION

Due to the constraints in the practical applications, output constraints exist widely in the controlled system, such as the rotation angle of robot arm [1], [2], the driving speed of autonomous vehicle [3]–[5], etc. When designing controllers for such controlled systems, output constraints can be a great hindrance. On the other hand, the unknown system dynamics and the influence of external disturbances are also the factors that must be considered when designing such controllers. Modern control theory such as H_∞ control method and integral reinforcement learning (IRL) method have received considerable attention in solving the problems of unknown system dynamics and external disturbances [6]–[10]. However, these methods can not satisfy the condition of output constraints when solving the above problems. So it is still a challenging problem to design controllers for partially unknown linear systems with output constraints and external disturbances. In this paper, a new adaptive control method is proposed to solve the H_∞ tracking control problem of partially unknown linear systems under the condition of satisfying the output constraints.

For the optimal control problem, it usually depends on solving a complex Hamilton-Jacobi-Bellman (HJB) equation, which is a very difficult problem to solve with the traditional mathematical tools. In the past few decades, reinforcement learning (RL) [11]–[14] was also known as adaptive dynamic programming (ADP) or approximate dynamic programming. The advantage of the adaptive dynamic programming is that the neural network (NN) can be used to approximate the optimal cost function in the optimal regulation problem, so it is widely used to solve the optimal control problem [15]–[17]. The concept of adaptive dynamic programming was first proposed by Werbos in 1977 [18]. Murray *et al.* developed an adaptive dynamic programming algorithm for optimal control of continuous time affine nonlinear systems [19], and gave a complete proof of its main theorem in [20]. Lewis *et al.* [21] proposed a synchronous policy iterative algorithm based on

The associate editor coordinating the review of this manuscript and approving it for publication was Zhiguang Feng.

an actor-critic network to solve the optimal control solution of the nonlinear system with known dynamics, and gave a proof of convergence. These methods require completely knowable system dynamics and do not take into account the influence of external disturbances. On the basis of [21], [22] proposed an online adaptive control algorithm [23] based on policy iteration (PI) to solve the continuous time two-person zero-sum game with infinite horizon cost for nonlinear systems with external disturbances. In [24], a non-strategic reinforcement learning method was used to solve the H_∞ tracking control problem for completely unknown continuous time systems. An integral reinforcement learning method based on value iteration (VI) was proposed to design H_∞ controllers for continuous time nonlinear systems [25]. An online model-free integral reinforcement learning algorithm based on neural network was proposed to solve the H_∞ optimal tracking control problem with finite horizon for completely unknown nonlinear continuous systems, in which the disturbance and constrained control input [26] were considered [27]. Adaptive output feedback neural tracking control for a class of uncertain switched multiple input multiple output nonlinear systems with non-strict feedback delays was studied in [28]. However, in the case of output constraints, the existing methods to solve external disturbances and unknown system dynamics often fail to get the desired results.

In order to solve the output constraint problem, Tee *et al.* [29] proposed an barrier Lyapunov function (BLF) by combining Lyapunov analysis with barrier function. Based on the results of Tee, Ren *et al.* [30] proved that the boundedness of BLF was safe for adaptive neural control of a class of output feedback nonlinear systems with unknown dynamics. In [31], the output constraint adaptive control problem in nonlinear stochastic systems was considered, and the influence of output constraints on control performance was overcomed. In [32], the barrier Lyapunov function design was extended to pure feedback systems with full-state constraints. For a class of nonlinear state constrained time-varying delay systems with unknown control coefficients, Li *et al.* [33] proposed an adaptive tracking control method. The adaptive control problem for a class of stochastic nonlinear systems with unknown control gain and complete state constraints is studied in [34]. Yang *et al.* [35], [36] solved the zero-sum and non-zero-sum game problem based on the barrier function, transforming the punishment for violating state constraints into the change of system state. However, the safety control problem of safety-critical systems with unknown system dynamics and external disturbances is rarely studied.

In this paper, a novel integral reinforcement learning method is proposed to solve the H_∞ tracking control problem of partially unknown continuous time linear systems with output constraints and external disturbances. The main contributions of this paper are as follows:

- In this paper, a H_∞ tracking controller satisfying the output constraints is designed under the condition of unknown dynamics and external disturbances. The stability of the transformed system can be expressed as satisfying the output constraints of the original system by using the barrier function transformation.
- A new integral reinforcement learning method is designed to obtain the solution of H_∞ tracking control problem online. The proposed algorithm only uses the obtained data and part of the system information, and the system can be partially unknown.
- It is proved that the proposed method can make the original system satisfy the output constraints under the condition of stable transformation system, and the output of the control strategy can make the tracking error asymptotically stable.

The rest of this paper is organized as follows: The linear tracking control with state constraints problem formulation are given in section 2. In section 3, the barrier transformation and traditional policy iteration algorithm are considered. In the next section, an integral RL method is proposed to obtain the optimal solution. In section 5, a numerical example is then presented to show the effectiveness of the proposed method. Finally, the conclusion of this paper is given.

II. LINEAR TRACKING CONTROL PROBLEM WITH STATE CONSTRAINTS

Considering the following linear continuous-time system,

$$\dot{x} = fx + gu + kd, \quad y = Cx, \tag{1}$$

where $x \in R^n$ is the system state, $u \in R^m \subset U$ is the control input, $d \in R^m$ is the external disturbance term, $f \in R^{n\times n}$ gives the drift dynamics of the system, $g \in R^{n\times m}$ and $k \in R^{n\times m}$, $C \in R^{p\times n}$ is the output matrix, $y \in R^{p\times 1}$ is the system output. U denotes the set of all admissible inputs. Define that every element in C is not less than zero. It is also assumed that the system (1) is stabilizable.

Assumption 1: The linear continuous-time system satifies the state constraints expressed as,

$$x_i \in (a_i, A_i), \quad i = 1 \cdots n, \tag{2}$$

where $a_i < 0$, $A_i > 0$ are the lower and upper boundaries of the system states, $a_x = [a_1; \cdots ; a_n]$ and $A_x = [A_1; \cdots ; A_n]$.

Based on the Assumption 1, we define the output constraint vectors as follows $a_y = Ca_x = [a_{y1}; \cdots ; a_{yp}]$, $A_y = CA_x = [A_{y1}; \cdots ; A_{yp}]$. The output constraints can be expressed as,

$$y_j \in (a_{yj}, A_{yj}), \quad j = 1 \cdots p. \tag{3}$$

Assumption 2: The reference output trajectory is defined as $\dot{y}_d = Fy_d$ which does not approach to zero as time goes to infinity, such as unit step, sinusoidal waveforms, etc., and the reference output trajectory y_d satisfies the output constraints (3).

In order to realize the tracking control, we first establish an augmented system according to the system (1) and reference output trajectory y_d. The augmented system state is defined as

$$\zeta = [x^T \quad y_d^T]^T. \tag{4}$$

Based on the equation (1) and the reference output trajectory $\dot{y}_d$, we can define

$$\dot{\zeta} = \begin{bmatrix} f & 0 \\ 0 & F \end{bmatrix} \zeta + \begin{bmatrix} g \\ 0 \end{bmatrix} u + \begin{bmatrix} k \\ 0 \end{bmatrix} d \equiv T\zeta + Gu + Kd. \quad (5)$$

Based on the state constraints (2) and the output constraints (3), the state constraints of the augmented system (5) can be defined as

$$\begin{aligned} a_\zeta &= [a_1; \cdots; a_n; a_{y1}; \cdots; a_{yp}], \\ &= [a_{\zeta 1}; \cdots; a_{\zeta n}; a_{\zeta n+1}; \cdots; a_{\zeta q}], \quad (6) \\ A_\zeta &= [A_1; \cdots; A_n; A_{y1}; \cdots; A_{yp}]. \\ &= [A_{\zeta 1}; \cdots; A_{\zeta n}; A_{\zeta n+1}; \cdots; A_{\zeta q}]. \quad (7) \end{aligned}$$

Note that the desired reference output trajectory y_d does not converge to zero as time goes to infinity. When the desired reference output trajectory is unstable and does not converge to zero, the feedback control will make the cost function of infinite horizon approach infinity [37]. According to Bellman's optimality principle, the cost function must be finite before the optimal feedback control strategy can be used to minimize it.

To relax the limit that the reference output trajectory must converge to zero, a discounted cost function is introduced as follows,

$$V(\zeta, u, d) = \int_t^\infty e^{-\beta(\tau - t)}(r(\zeta, u) - \gamma^2 d^T d)d\tau, \quad (8)$$

where $r(\zeta, u) = \zeta^T C_1^T Q C_1 \zeta + u^T R u$, $Q > 0$ and $R > 0$ are symmetric matrices, $C_1 = [C \quad -I]$, $\beta > 0$ is the discount factor, $\gamma > 0$ represents a bound on L_2 gain required to move from disturbance d to the cost function, that is

$$\int_t^\infty e^{-\beta(\tau - t)} r(\zeta, u)d\tau \leq \gamma^2 \int_t^\infty (e^{-\beta(\tau - t)} d^T d)d\tau, \quad (9)$$

for all $d \in L_2[0, \infty)$.

Based on the Assumption 1-2, the goal of the H_∞ tracking control problem with output constraints is to find an optimal control strategy u^* such that, the system (5) has L_2 gain less than or equal γ, the output satisfies the constraints (3) and the tracking error asymptotically stable. It can be described mathematically as

$$\begin{aligned} &\int_t^\infty e^{-\beta(\tau - t)}(\zeta^T C_1^T Q C_1 \zeta + u^T R u)d\tau \\ &\qquad \leq \gamma^2 \int_t^\infty (e^{-\beta(\tau - t)} d^T d)d\tau, \\ &s.t. \quad y_j \in (a_{yj}, A_{yj}), j = 1 \cdots p, \\ &\qquad and\ y \to y_d,\ as\ t\ increase. \quad (10) \end{aligned}$$

Unlike the previous studies, the output constraints (3) brings great difficulty to solve the optimal control strategy. This is because the proposed discounted cost function is only affected by the system state and the output reference trajectory. In the next section, we will propose a barrier transformation approach to satisfy the output constraints in (3).

Remark 1: Each element of the output matrix C is predefined so that the output constraints can be defined by the state constraints. At the same time, the constrains (3) can be satisfied by constraining the system states.

III. PROBLEM TRANSFORMATION AND TRADITIONAL POLICY ITERATION ALGORITHM

In this section, the barrier function is used to transform the system (5) with the output constraints into a transformed system without the output constraints, that is, the H_∞ tracking control problem with the output constraints is transformed into a H_∞ tracking control problem without the output constraints. Before moving on, the following definition of the barrier function is introduced.

Definition 1: The function $b(\cdot) : R \to R$ defined on (a, A) is referred to as a barrier function, if

$$b(z; a, A) = log\frac{A(a - z)}{a(A - z)}, z \in R \quad (11)$$

where a and A are two constants satisfying $a < 0 < A$. Moreover, the inverse of the barrier function is as follows

$$b^{-1}(y; a, A) = aA\frac{e^{\frac{y}{2}} - e^{-\frac{y}{2}}}{ae^{\frac{y}{2}} - Ae^{-\frac{y}{2}}}, \quad (12)$$

with the derivative by,

$$\frac{db^{-1}(y; a, A)}{dz} = \frac{Aa^2 - aA^2}{a^2 e^y - 2aA + A^2 e^{-y}}. \quad (13)$$

Remark 2: To satisfy the output constraints, the barrier function in Definition 1 should have the following characteristics:

1) *The barrier function $b(\cdot)$ is a finite value within the state range of the constraints (a, A).*
2) *As the state approaches the constraint (a, A), $b(\cdot)$ approaches infinity, i.e., $\lim_{z \to a^+} b(z; a, A) = -\infty$, $\lim_{z \to A^-} b(z; a, A) = +\infty$.*
3) *The barrier function $b(\cdot)$ converges as the state converges.*

A. TRANSFORMED SYSTEM BASED ON BARRIER FUNCTION

Consider the following transformed system,

$$\begin{aligned} s_q &= b(\zeta_q, a_{\zeta q}, A_{\zeta q}), \quad (14) \\ \zeta_q &= b^{-1}(s_q; a_{\zeta q}, A_{\zeta q}). \quad (15) \end{aligned}$$

According to the chain rule and the equation (14), (15), we can derive the following equation,

$$\begin{aligned} \dot{s}_q &= \frac{\dot{\zeta}_q}{\frac{db^{-1}(z; a_{\zeta q}, A_{\zeta q})}{dz}|_{z = s_q}} \\ &= (T_q(\zeta) + G_q u + K_q d)\frac{a_{\zeta q}^2 e^{s_q} - 2a_{\zeta q}A_{\zeta q} + A_{\zeta q}^2 e^{-s_q}}{A_{\zeta q}a_{\zeta q}^2 - a_{\zeta q}A_{\zeta q}^2} \\ &= \bar{T}_q(s) + \bar{G}_q(s)u + \bar{K}_q(s)d, \quad (16) \end{aligned}$$

where

$$\begin{aligned}\bar{T}_q(s) &= \frac{a_{\zeta q}^2 e^{s_q} - 2a_{\zeta q}A_{\zeta q} + A_{\zeta q}^2 e^{-s_q}}{A_{\zeta q}a_{\zeta q}^2 - a_{\zeta q}A_{\zeta q}^2} \\ &\quad \times T_q([b^{-1}(s_1); \cdots ; b^{-1}(s_q)]), \\ \bar{G}_q(s) &= \frac{a_{\zeta q}^2 e^{s_q} - 2a_{\zeta q}A_{\zeta q} + A_{\zeta q}^2 e^{-s_q}}{A_{\zeta q}a_{\zeta q}^2 - a_{\zeta q}A_{\zeta q}^2} \\ &\quad \times G_q([b^{-1}(s_1); \cdots ; b^{-1}(s_q)]), \\ \bar{K}_q(s) &= \frac{a_{\zeta q}^2 e^{s_q} - 2a_{\zeta q}A_{\zeta q} + A_{\zeta q}^2 e^{-s_q}}{A_{\zeta q}a_{\zeta q}^2 - a_{\zeta q}A_{\zeta q}^2} \\ &\quad \times K_q([b^{-1}(s_1); \cdots ; b^{-1}(s_q)]). \end{aligned} \tag{17}$$

The transformed system dynamics $\dot{s} = [\dot{s}_1; \cdots ; \dot{s}_q]$ is as follows,

$$\dot{s} = \bar{T}(s) + \bar{G}(s)u + \bar{K}(s)d, \tag{18}$$

where $\bar{T}(s) = [\bar{T}_1(s); \cdots ; \bar{T}_q(s)]$, $\bar{G}(s) = [\bar{G}_1(s); \cdots ; \bar{G}_q(s)]$, $\bar{K}(s) = [\bar{K}_1(s); \cdots ; \bar{K}_q(s)]$.

Assumption 3: Assume that the transformed system (18) *has the following characteristics:*

(1).$\bar{T}(s)$ is Lipschitz continuous function and there is a constant λ_t such that $\|\bar{T}(s)\| \leq \lambda_t\|s\|$, $s \in \Omega$, where Ω is a compact set containing the origin.

(2).$\bar{G}(s)$, $\bar{K}(s)$ are bounded on Ω and there exist constants λ_g, λ_k such that $\|\bar{G}(s)\| \leq \lambda_g$, $\|\bar{K}(s)\| \leq \lambda_k$.

(3).The system (18) *is controllable over the compact set Ω.*

(4).The system (18) *has the property of zero-state observability.*

The discounted cost function of the transformed system is defined as

$$J(s, u) = \int_t^\infty e^{-\beta(\tau - t)}(s^T C_1^T Q C_1 s + u^T R u - \gamma^2 d^T d)d\tau. \tag{19}$$

Based on the transformation of equation (14)(16), the H_∞ tracking control problem with output constraints has transformed into a H_∞ tracking control problem without output constraints. In other words, the goal of the H_∞ tracking control problem with output constraints becomes to find an optimal control law u^* such that, the system (18) has L_2 gain less than or equal γ, i.e.,

$$\begin{aligned}\int_t^\infty & e^{-\beta(\tau - t)}(s^T C_1^T Q C_1 s + u^T R u)d\tau \\ &\leq \int_t^\infty e^{-\beta(\tau - t)}\gamma^2 d^T d d\tau, t \geq 0. \end{aligned} \tag{20}$$

Remark 3: Because the barrier function will approach infinity at the safety constraints boundary, the reference output trajectory y_d must strictly satisfy the output constraint (3), *otherwise the transformation system states will tend to infinity during the process of tracking the reference trajectory.*

B. POLICY ITERATION ALGORITHM BASED ON SYSTEM DYNAMICS

Define the Hamiltonian for the discounted cost function (19) as

$$\begin{aligned}H(s, u, d, J) = {} & s^T C_1^T Q C_1 s + u^T R u - \gamma d^T d - \beta J(s) \\ & + J_s^T(s)(\bar{T}(s) + \bar{G}(s)u + \bar{K}(s)d), \end{aligned} \tag{21}$$

where $J_s(s)$ is the partial derivative of $J(s)$ with respect to s.

The HJB equation associated with the Hamiltonian (21) is as follows

$$\begin{aligned}s^T C_1^T Q C_1 s + u^T R u - \gamma d^T d - \beta J(s) + J_s^T(s)(\bar{T}(s) \\ + \bar{G}(s)u + \bar{K}(s)d) = 0. \end{aligned} \tag{22}$$

Given a solution $J^*(s) > 0$ to the Hamiltonian (21), the optimal control solution u^* and the worst disturbance d^* have the following stationary conditions,

$$\frac{\partial H(s, u, d, J)}{\partial u} = 0, \frac{\partial H(s, u, d, J)}{\partial d} = 0. \tag{23}$$

Then, we can get

$$u^* = -\frac{1}{2}R^{-1}\bar{G}(s)^T J_s^*, \tag{24}$$

$$d^* = \frac{1}{2\gamma^2}\bar{K}(s)^T J_s^*. \tag{25}$$

Lemma 1: Under Assumptions 1, 2 and 3, if the optimal control strategy (24) *and the worst disturbance* (25) *can solve the H_∞ tracking control problem of the transformed system* (18)*, then:*

(1) Reasonable selection of the discount factor β can ensure that the tracking error is asymptotically stable.

(2) The system states (1) *satisfies the constraint* (2) *provided that the initial state x_0 of the system* (1) *satisfies the constraint* (2).

(3) The L_2 gain condition (20) *can be guaranteed, if the performance output is designed as $s^T C_1^T Q C_1 s + u^T R u$.*

Proof: (1) Differentiating the cost function (19) *along the trajectories of the transformed system, we can get*

$$\begin{aligned}J_s^T(s)(\bar{T}(s) + \bar{G}(s)u + \bar{K}(s)d) = {} & -s^T C_1^T Q C_1 s - u^T R u \\ & + \beta J(s) + \gamma d^T d. \end{aligned} \tag{26}$$

In order to make the tracking error asymptotically stable, we define the discount factor $\beta \leq \frac{s^T C_1^T Q C_1 s + u^T R u - \gamma d^T d}{J(s)}$, then we can get

$$\dot{J}(s) = -s^T C_1^T Q C_1 s - u^T R u + \beta J(s) + \gamma d^T d \leq 0. \tag{27}$$

Therefore, the tracking error is locally asymptotically stable.

(2) Based on the equation (27)*, one can get $\dot{J}(s) \leq 0$, such that*

$$J(s(t)) \leq J(s(0)), \quad \forall t \geq 0. \tag{28}$$

As long as the initial state x_0 satisfies the constraint (2) *and the reference output satisfies* (3)*, it can be concluded that the initial cost function $J(s(0))$ is finite, thus making the cost*

```
Algorithm 1 Policy Iteration Based on System Dynamics
Initialization: Start with an admissible control policy $u_0$.
Procedure:
1. Given $u_i$, solve the cost function $J(s)$ using
    $s^T C_1^T QC_1 s + u_i^{*T} Ru_i^* - \gamma d_i^T d_i - \beta J_i^*(s) + J_{is}^{*T}(s)(\bar{T}(s)$
    $+\bar{G}(s)u_i^* + \bar{K}(s)d_i^*) = 0.$    (31)
2. Update the disturbance using
    $d_{i+1}^* = \frac{1}{2\gamma^2}\bar{K}(s)^T J_{is}^*,$    (32)
   update the control strategy using
    $u_{i+1}^* = -\frac{1}{2}R^{-1}\bar{G}(s)^T J_{is}^*.$    (33)
if $\| J^{i+1} - J^i \| \leq \epsilon$, $\epsilon$ is a select positive number.
    The learning is finished and stop the iteration solution
else
    $i = i + 1$, go to step 1 end
End Procedure
```

function $J(s(t))$ *finite. Therefore, according to the Remark 2, we can infer that*

$$x_i \in (a_i, A_i), \quad i = 1 \cdots n. \tag{29}$$

Therefore, the constraints (2) *can be satisfied.*

(3) Considering the system transformation (14) *and the constraints* (6), (7), *each element of transformation system state* $s = [b_1(\zeta_1); \cdots ; b_q(\zeta_q)]$ *is finite. Note that the optimal control input, the worst disturbance and the optimal cost function satisfy the HJB equation* (22)*. Then, as long as the perference output is designed as* $s^T C_1^T QC_1 s + u^T Ru$*, we have*

$$\begin{aligned} &H(s, u^*, d^*, J^*) \\ &\quad = 0 \Rightarrow \int_t^\infty e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u^T Ru)d\tau \\ &\quad \leq \int_t^\infty e^{-\beta(\tau-t)}\gamma^2 d^T d d\tau. \end{aligned} \tag{30}$$

This proof is completed.

The traditional model-based policy iteration algorithm is shown in algorithm 1, all the system dynamics, such as the original dynamic matrix $\bar{T}(s)$, $\bar{G}(s)$ and $\bar{K}(s)$ are essential. In practice, the unpredictability of system dynamics will make the traditional policy iteration method ineffective. In order to meet the strict requirements of the system information, the integral RL technology is applied to the tracking control design, so that the tracking control strategy can be obtained when the system dynamics is partially unknown.

Remark 4: The system state of the transformation system (18) *is defined by the system* (1) *and the barrier function. The barrier function is already defined by the equation* (11)*, so the partially unknown linear systems* (1) *means that part of the transformation system is unknown.*

```
Algorithm 2 Integral RL Based Policy Iteration for Tracking Problem With Output Constraints
Initialization: Start with an admissible control policy $u_0$.
Procedure:
1. Given $u_i$, solve the cost function $J(s)$ using
    $s^T(t)P_i s(t) = \int_t^{t+\Delta t} e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u_i^T Ru_i$
    $-\gamma^2 d^T d)d\tau + e^{-\beta\Delta t}s^T(t+\Delta t)P_i s(t+\Delta t).$    (37)
2. Update the disturbance using
    $d_{i+1}^* = \frac{1}{2\gamma^2}\bar{K}(s)^T J_{is}^* = \frac{1}{\gamma^2}\bar{K}(s)^T P_i s,$    (38)
   update the control strategy using
    $u_{i+1}^* = -\frac{1}{2}R^{-1}\bar{G}(s)^T J_{is}^* = -R^{-1}\bar{G}(s)^T P_i s.$    (39)
if $\| P^{i+1} - P^i \| \leq \epsilon$, $\epsilon$ is a select positive number.
    The learning is finished and stop the iteration solution
else
    $i = i + 1$, go to step 1 end
End Procedure
```

IV. INTEGRAL RL FOR TRANSFORMED SYSTEM AND STABILITY ANALYSIS

Based on the transformation system dynamics and the traditional model-based policy iterative algorithm, the integral RL tracking control algorithm is designed for the system with partially unknown dynamics, and the tracking error can guarantee asymptotically stable under the condition of output constraints.

A. INTEGRAL RL FOR PARTIALLY UNKNOWN DYNAMICS

Based on the optimal control theory, the discounted cost function (19) can be rewritten by the positive definite quadratic function such that

$$J(s) = s^T Ps, \tag{34}$$

where P is the positive definite matrice.

For time interval $\Delta t > 0$, the cost function (19) satisfies

$$\begin{aligned} J(s(t)) = \int_t^{t+\Delta t} & e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u^T Ru \\ & -\gamma^2 d^T d)d\tau + e^{-\beta\Delta t}J(s(t+\Delta t)). \end{aligned} \tag{35}$$

Substituting equation (34) for (35), we can get

$$\begin{aligned} s^T(t)Ps(t) = \int_t^{t+\Delta t} & e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u^T Ru \\ & -\gamma^2 d^T d)d\tau + e^{-\beta\Delta t}s^T(t+\Delta t)Ps(t+\Delta t). \end{aligned} \tag{36}$$

Based on the equation (34) and (35), we use the integral reinforcement learning method to solve the H_∞ tracking control problem with the output constraints.

Theorem 1: Consider the transformation system (18), *the Hamiltonian equation* (21), *the control input* (39), *and the disturbance input* (38). *Assume that the Assumptions 1, 2 and 3 hold. The iterative control strategy* (39) *obtained from* (21) *can minimize the right side of* (37). *The iterative disturbance strategy* (38) *obtained from* (22) *can maximize the right side of* (37).

Proof: Consider the definition of the Hamiltonian (21) *that* $\frac{d^2H}{d^2u} = 2R > 0$, *so that the Hamiltonian attains a minimum in* u.

We define the first order Taylor series for $\int_t^{t+\Delta t} e^{-\beta(\tau-t)}$ $(s^T C_1^T QC_1 s + u^T Ru - \gamma^2 d^T d)d\tau$ *and* $e^{-\beta\Delta t} J(s(t+\Delta t))$ *as follows*

$$\begin{aligned}\int_t^{t+\Delta t} & e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u^T Ru - \gamma^2 d^T d)d\tau \\ &= -\frac{1}{\beta}(e^{-\beta\Delta t} - 1)(s^T C_1^T QC_1 s + u^T Ru - \gamma^2 d^T d) \\ &\quad + o(\Delta t), \qquad (40)\end{aligned}$$

$$\begin{aligned}e^{-\beta\Delta t} J_i(s(t+\Delta t)) &= e^{\beta\Delta t} J_i(s(t)) + e^{\beta\Delta t} J_{is} \\ &\quad (s(t))(\bar{T}(s) + \bar{G}(s)u^* + \bar{K}(s)d^*)\Delta t + o(\Delta t). \qquad (41)\end{aligned}$$

If the time interval Δt *is small enough, then the higher-order infinitesimal term* $o(\Delta t)$ *can be ignored. we can get*

$$\begin{aligned}\int_t^{t+\Delta t} & e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u^T Ru - \gamma^2 d^T d)d\tau \\ &= -\frac{1}{\beta}(e^{-\beta\Delta t} - 1)(s^T C_1^T QC_1 s + u^T Ru - \gamma^2 d^T d), \qquad (42)\end{aligned}$$

$$\begin{aligned}e^{-\beta\Delta t} J_i(s(t+\Delta t)) &= e^{\beta\Delta t} J_i(s(t)) + e^{\beta\Delta t} J_{is} \\ &\quad (s(t))(\bar{T}(s) + \bar{G}(s)u^* + \bar{K}(s)d^*)\Delta t. \qquad (43)\end{aligned}$$

The iterative control policy u_i *satisfies*

$$\begin{aligned}u_i &= arg \min_u H(s, u, d, J_i) \\ &= arg \min_u [s^T C_1^T QC_1 s + u^T Ru - \gamma^2 d^T d - \beta J_i(s) \\ &\quad + J_{is}^T(s)(\bar{T}(s) + \bar{G}(s)u + \bar{K}(s)d)]. \qquad (44)\end{aligned}$$

Since the iterative cost function $J_i(s)$ *is not affected by the control strategy, we can get*

$$\begin{aligned}u_i = arg \min_u [&-\frac{1}{\beta}(e^{-\beta\Delta t} - 1)(s^T C_1^T QC_1 s + u^T Ru \\ &-\gamma^2 d^T d) + e^{\beta\Delta t} J_i(s(t)) + e^{\beta\Delta t}(J_{is}^T(s)(\bar{T}(s) \\ &+\bar{G}(s)u) + \bar{K}(s)d))]. \qquad (45)\end{aligned}$$

Based on (34), (42), (43), *it yields*

$$\begin{aligned}u_i = arg \min_u [\int_t^{t+\Delta t} & e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u_i^T Ru_i \\ &-\gamma^2 d^T d)d\tau + e^{-\beta\Delta t} s^T(t+\Delta t)P_i s(t+\Delta t)]. \qquad (46)\end{aligned}$$

In the same way, we can get

$$\begin{aligned}d_i = arg \max_d [\int_t^{t+\Delta t} & e^{-\beta(\tau-t)}(s^T C_1^T QC_1 s + u_i^T Ru_i \\ &-\gamma^2 d^T d)d\tau + e^{-\beta\Delta t} s^T(t+\Delta t)P_i s(t+\Delta t)]. \qquad (47)\end{aligned}$$

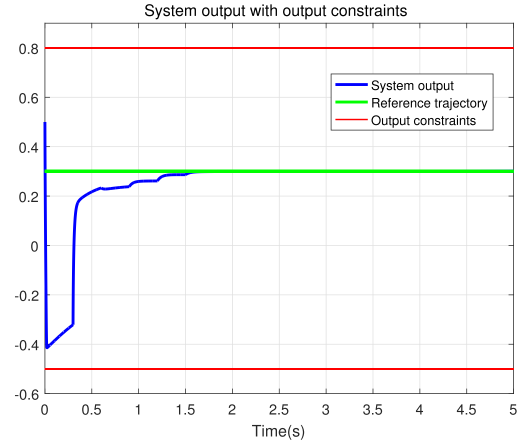

FIGURE 1. **Constrained trajectory of the tracking dynamics under the integral RL based tracking control algorithm.**

The proof is completed.

Based on the integral RL method, a new policy iterative algorithm is proposed to solve the H_∞ tracking control problem with output constraints, as shown in algorithm 2. In algorithm 2, only part of the system information is used in the iterative process. According to lemma 1 and theorem 1, the integral RL control algorithm proposed in algorithm 2 can make the tracking error locally asymptotically stable under the condition that the trajectory of the system converges.

Remark 5: Compared with the existing optimal tracking control standard solutions, the proposed method provides some advantages for solving partially unknown linear systems, which are reflected in the following aspects.

(1) In the existing strategy iteration algorithm, all the system information is repeatedly used and transformed in the solving process. The proposed algorithm only uses the obtained data and part of the system information, and the system can be partially unknown.

(2) By combining the tracking control problem with the barrier function, the tracking system and the tracking error are locally asymptotically stable under the condition of output constraints.

V. SIMULATION RESULTS

In this section, a linear example is presented to prove the validity of the proposed algorithm.

Consider the following linear system

$$\dot{x} = fx + gu + kd, \quad y = Cx,$$

where,

$$f = \begin{bmatrix} 0.5 & 1.5 \\ 2.0 & -2 \end{bmatrix}, \quad g = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \quad k = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad C = [1 \quad 0].$$

Assume that the desired output trajectory is generated by the command generator system $\dot{y}_d = 0$ with the initial value $y_d(0) = 0.3$. Define (19) as the performance function. One

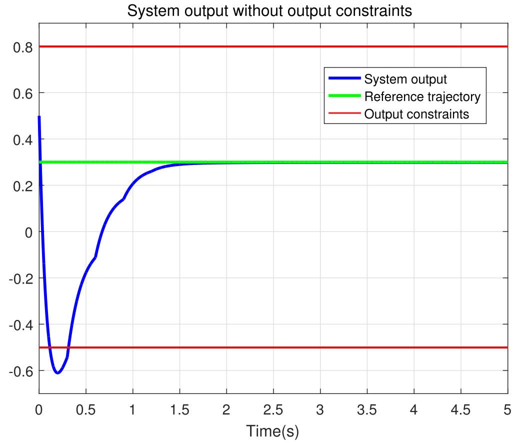

FIGURE 2. Unconstrained trajectory of the tracking dynamics under the integral RL based tracking control algorithm.

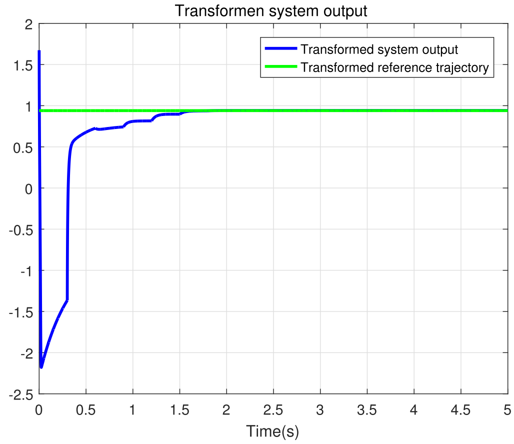

FIGURE 3. Transformed system dynamics under the integral RL based tracking control algorithm.

selects $Q = 3$, $R = 1$ and the discount factor $\gamma = 1.5$. The output constraints is defined as $y \in (-0.5, 0.8)$.

Assuming that the system drift dynamics in algorithm 2 is unknown, the data obtained from the transformed system every 0.05 seconds is used for simulation. At the same time, we make a comparison with the method in [14], which proves that the proposed method is effective for the output constraints and the tracking control problems.

Figure 1 shows the trajectory of the system output following the reference output under output constraints. Figure 2 describes the tracking control trajectory without output constraints based on integral reinforcement learning algorithm. Comparing the results of the two figures, it can be clearly seen that the proposed method can complete the tracking under the condition of ensuring the output constraints. Figure 3 shows the tracking trajectory of the transformation system, where the reference output trajectory $s_3 = 0.940$ is obtained by equation (14), which also verifies the second part of Theorem 1.

Figure 4 shows the parameter changes of matrix P in the process of iteration. The trajectory of the control strategy and disturbance strategy are shown in Figure 5. Figure 6 shows the tracking error, which obviously eventually converges to zero.

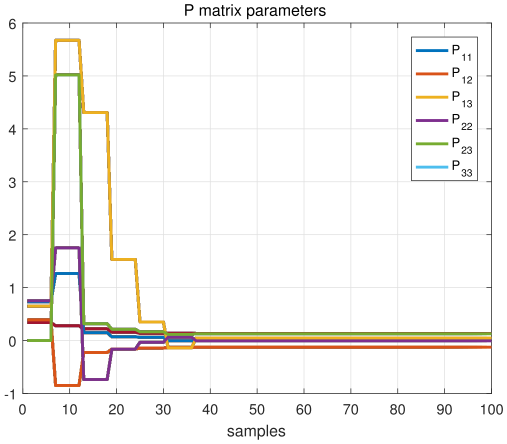

FIGURE 4. Evolutions of the P matrice during the integral RL process.

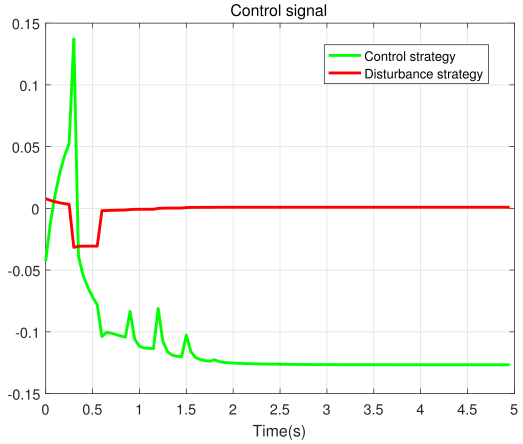

FIGURE 5. Trajectory of control strategy under integral RL process.

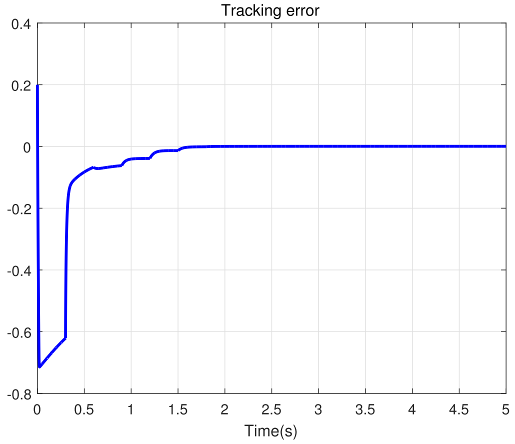

FIGURE 6. Tracking error trajectory by our integral RL method.

VI. CONCLUSION

In this paper, we studied the H_∞ tracking control problem for partially unknown linear systems with output constraints and disturbances. The asymptotic tracking and output constraint of the system were realized by building an augmented system and a reasonable system transformation. Integral reinforcement learning was used to obtain the optimal control strategy and the worst disturbance strategy online. It was proved that the proposed method can minimize the performance of the system under the influence of output constraints and disturbances. The numerical simulation example also demonstrated the effectiveness of the proposed method.

REFERENCES

[1] W. Li, A. Ames, and M. Egerstedt, "Safety barrier certificates for collisions-free multirobot systems," *IEEE Trans. Robot.*, vol. 33, no. 3, pp. 661–674, Jun. 2017.

[2] F. Ferraguti, C. Talignani Landi, S. Costi, M. Bonfè, S. Farsoni, C. Secchi, and C. Fantuzzi, "Safety barrier functions and multi-camera tracking for human–robot shared environment," *Robot. Auto. Syst.*, vol. 124, Feb. 2020, Art. no. 103388.

[3] Y. Chen, H. Peng, and J. Grizzle, "Obstacle avoidance for low-speed autonomous vehicles with barrier function," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 1, pp. 194–206, Jan. 2018.

[4] Z. Li, A. Zhou, J. Pu, and J. Yu, "Multi-modal neural feature fusion for automatic driving through perception-aware path planning," *IEEE Access*, vol. 9, pp. 142782–142794, 2021.

[5] M. Held, O. Flardh, and J. Martensson, "Optimal speed control of a heavy-duty vehicle in urban driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1562–1573, Apr. 2019.

[6] H. Zhang, C. Qin, B. Jiang, and Y. Luo, "Online adaptive policy learning algorithm for H_∞ state feedback control of unknown affine nonlinear discrete-time systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2706–2718, Dec. 2014.

[7] C. Liu, H. Zhang, S. Sun, and H. Ren, "Online H_∞ control for continuous-time nonlinear large-scale systems via single echo state network," *Neurocomputing*, vol. 448, pp. 353–363, Aug. 2021.

[8] H. Jiang, H. Zhang, Y. Luo, and X. Cui, "H_∞ control with constrained input for completely unknown nonlinear systems using data-driven reinforcement learning method," *Neurocomputing*, vol. 237, pp. 226–234, May 2017.

[9] H. Zhang, Z. Ming, Y. Yan, and W. Wang, "Data-driven finite-horizon H_∞ tracking control with event-triggered mechanism for the continuous-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 11, 2021, doi: 10.1109/TNNLS.2021.3116464.

[10] C. He, Y. Wan, Y. Gu, and F. L. Lewis, "Integral reinforcement learning-based multi-robot minimum time-energy path planning subject to collision avoidance and unknown environmental disturbances," *IEEE Control Syst. Lett.*, vol. 5, no. 3, pp. 983–988, Jul. 2021.

[11] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, p. 1054, Sep. 1998.

[12] W. Dixon, "Optimal adaptive control and differential games by reinforcement leanring principles [book review]," *IEEE Contr. Syst. Mag.*, vol. 34, no. 3, pp. 80–82, 2014.

[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.

[14] 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.

[15] 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.

[16] H. Zhang, H. Jiang, Y. Luo, and G. Xiao, "Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4091–4100, May 2017.

[17] X. Yang, B. Li, and G. Wen, "Adaptive neural network optimized control using reinforcement learning of critic-actor architecture for a class of non-affine nonlinear systems," *IEEE Access*, vol. 9, pp. 141758–141765, 2021.

[18] P. J. Werbos, "Advanced forecasting methods for global crisis warning and models of intelligence," *Gen. Syst. Yearbook*, vol. 32, pp. 25–38, 1977.

[19] 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.

[20] J. Murray, C. Cox, and R. E. Saeks, *The Adaptive Dynamic Programming Theorem*. Boston, MA, USA: Birkhauser, 2003.

[21] K. G. Vamvoudakis and F. L. Lewis, "Online actor critic algorithm to solve the continuous-time infinite horizon optimal control problem," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2009, pp. 3180–3187.

[22] 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, 2012.

[23] W. Qi, G. Zong, and W. X. Zheng, "Adaptive event-triggered SMC for stochastic switching systems with semi-Markov process and application to boost converter circuit model," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 2, pp. 786–796, Feb. 2021.

[24] 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.

[25] G. Xiao, H. Zhang, K. Zhang, and Y. Wen, "Value iteration based integral reinforcement learning approach for H_∞ controller design of continuous-time nonlinear systems," *Neurocomputing*, vol. 285, pp. 51–59, Apr. 2018.

[26] X. W. Mu and K. Liu, "Containment control of single-integrator network with limited communication data rate," *IEEE Trans. Autom. Control*, vol. 61, no. 8, pp. 2232–2238, Aug. 2016.

[27] H. Zhang, X. Cui, Y. Luo, and H. Jiang, "Finite-horizon H_∞ tracking control for unknown nonlinear systems with saturating actuators," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1200–1212, Apr. 2018.

[28] J. Kong, B. Niu, Z. Wang, P. Zhao, and W. Qi, "Adaptive output-feedback neural tracking control for uncertain switched MIMO nonlinear systems with time delays," *Int. J. Syst. Sci.*, vol. 52, no. 13, pp. 2813–2830, Oct. 2021.

[29] K. P. Tee, S. S. Ge, and E. H. Tay, "Barrier Lyapunov functions for the control of output-constrained nonlinear systems," *Automatica*, vol. 45, no. 4, pp. 918–927, Apr. 2009.

[30] B. Ren, S. S. Ge, K. P. Tee, and T. H. Lee, "Adaptive neural control for output feedback nonlinear systems using a barrier Lyapunov function," *IEEE Trans. Neural Netw.*, vol. 21, no. 8, pp. 1339–1345, Aug. 2010.

[31] Y.-J. Liu, S. Lu, S. Tong, X. Chen, C. L. P. Chen, and D.-J. Li, "Adaptive control-based Barrier Lyapunov functions for a class of stochastic nonlinear systems with full state constraints," *Automatica*, vol. 87, pp. 83–93, Jan. 2018.

[32] C. Wang, Y. Wu, and J. Yu, "Barrier Lyapunov functions-based adaptive control for nonlinear pure-feedback systems with time-varying full state constraints," *Int. J. Control, Autom. Syst.*, vol. 15, no. 6, pp. 2714–2722, Dec. 2017.

[33] D. Li and D. Li, "Adaptive tracking control for nonlinear time-varying delay systems with full state constraints and unknown control coefficients," *Automatica*, vol. 93, pp. 444–453, Jul. 2018.

[34] W. Su, B. Niu, H. Wang, and W. Qi, "Adaptive neural network asymptotic tracking control for a class of stochastic nonlinear systems with unknown control gains and full state constraints," *Int. J. Adapt. Control Signal Process.*, vol. 35, no. 10, pp. 2007–2024, Oct. 2021.

[35] Y. Yang, K. G. Vamvoudakis, and H. Modares, "Safe reinforcement learning for dynamical games," *Int. J. Robust Nonlinear Control*, vol. 30, no. 9, pp. 3706–3726, Jun. 2020.

[36] Y. Yang, D.-W. Ding, H. Xiong, Y. Yin, and D. C. Wunsch, "Online barrier-actor-critic learning for H_∞ control with full-state constraints and input saturation," *J. Franklin Inst.*, vol. 357, no. 6, pp. 3316–3344, Apr. 2020.

[37] K. Zhang, H. Zhang, Y. Mu, and S. Sun, "Tracking control optimization scheme for a class of partially unknown fuzzy systems by using integral reinforcement learning architecture," *Appl. Math. Comput.*, vol. 359, pp. 344–356, Oct. 2019.

CHUNBIN QIN received the B.S. degree from the School of Computer and Information Engineering, Henan University, Kaifeng, China, in 2009, and the Ph.D. degree in power electronics and power transmission from Northeastern University, Shenyang, China, in 2014.

He is currently an Associate Professor with Henan University. His current research interests include adaptive dynamic programming, neural networks-adaptive optimal control, artificial intelligence algorithm, multi-agent cooperative control, event-triggered control, reinforcement learning, safety-critical control, and their industrial applications. He was awarded the "Excellent Doctoral Dissertation Award" by the China Association for Artificial Intelligence in 2016.

JINGUANG WANG received the B.S. degree from the School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China, in 2020. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Henan University, Henan.

His current research interests include optimal control, adaptive dynamic programming, approximate dynamic programming, and robust control.

XIAOPENG QIAO received the B.S. degree from the School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China, in 2020. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Henan University, Henan.

His current research interests include optimal control, adaptive dynamic programming, approximate dynamic programming, and safety-critical control.

HEYANG ZHU received the B.S. degree from the School of Measurement and Control Technology and Instruments, Henan University Minsheng College, Kaifeng, China, in 2019. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Henan University, Henan.

His current research interests include adaptive dynamic programming and event-triggered control.

DEHUA ZHANG received the B.S. degree from the School of Automation Science and Electrical Engineering, Beijing University of Aeronautics and Astronautics, Beijing, China, in 2009, and the Ph.D. degree in control theory and control engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2013. His current research interests include nonlinear control, intelligent control, optimal control, ADPRL, SCM control, robot control, and the Internet of Vehicles.

YONGHANG YAN received the B.S. degree from the School of Information Engineering, Zhengzhou University, Zhengzhou, China, in 2004, and the M.S. and Ph.D. degrees from the School of Computer Science, Beijing Institute of Technology, Beijing, China, in 2007 and 2014, respectively.

He is currently an Associate Professor with Henan University. His current research interests include computer networks, mobile ad hoc networks, wireless sensor networks, UAV, the Internet of Things, and artificial intelligence. He served as the session chair for academic conferences for many times.