# Robust Approximate Dynamic Programming for Nonlinear Systems With Both Model Error and External Disturbance

Jie Li, Ryozo Nagamune, *Senior Member, IEEE*, Yuhang Zhang, and Shengbo Eben Li, *Senior Member, IEEE*

***Abstract*— Model error and external disturbance have been separately addressed by optimizing the definite $H_\infty$ performance in standard linear $H_\infty$ control problems. However, the concurrent handling of both introduces uncertainty and nonconvexity into the $H_\infty$ performance, posing a huge challenge for solving nonlinear problems. This article introduces an additional cost function in the augmented Hamilton–Jacobi–Isaacs (HJI) equation of zero-sum games to simultaneously manage the model error and external disturbance in nonlinear robust performance problems. For satisfying the Hamilton–Jacobi inequality in nonlinear robust control theory under all considered model errors, the relationship between the additional cost function and model uncertainty is revealed. A critic online learning algorithm, applying Lyapunov stabilizing terms and historical states to reinforce training stability and achieve persistent learning, is proposed to approximate the solution of the augmented HJI equation. By constructing a joint Lyapunov candidate about the critic weight and system state, both stability and convergence are proved by the second method of Lyapunov. Theoretical results also show that introducing historical data reduces the ultimate bounds of system state and critic error. Three numerical examples are conducted to demonstrate the effectiveness of the proposed method.**

***Index Terms*— Approximate dynamic programming (ADP), Hamilton–Jacobi–Isaacs (HJI) equation, robust performance, uncertain nonlinear systems.**

## I. Introduction

MODEL error and external disturbance are very common in robust control problems [1], [2]. For linear dynamics, robust stabilization and disturbance attenuation problems can be described as standard $H_\infty$ control problems and solved by convex optimization methods based on Riccati equation [3]. Nevertheless, robust performance problems considering both the model error and external disturbance are generally nonconvex. Classical approaches, including D–K iteration [4] and bilinear matrix inequalities [5], heuristically optimize the $H_\infty$ performance with uncertainty by solving a series of convex optimization problems iteratively, but they are incompetent to cope with nonlinear problems.

Recently, the relationship between the $L_2$ induced norm of nonlinear systems and Hamilton–Jacobi inequality has been investigated to solve nonlinear robust performance problems [6], [7]. However, the derived Hamilton–Jacobi–Isaacs (HJI) equations are nonlinear partial differential equations, whose analytical solutions are intractable to obtain [8], [9]. To tackle the curse of dimensionality problem, approximate dynamic programming (ADP) methods are introduced to solve Hamilton–Jacobi equations numerically [10], and neural networks are employed to approximate the solution [11], [12]. Nowadays, ADP has been widely employed in optimal control problems, such as chemical process control [13] and vehicle control [14]. The research on handling external disturbance or model error has attracted more and more attention [15], [16], [17].

From the view of a zero-sum game, Abu-Khalaf et al. [18] proposed a two-loop policy iteration to solve disturbance attenuation problems, where control and disturbance policies were updated asynchronously. A simultaneous policy update paradigm with only one iterative loop was proposed to simplify the training process [19], [20], while the stability analysis of the dynamic system was not fully discussed. Then, an online approximator (OLA)-based framework was developed to apply the gradient descent method to learn the critic solution of the HJI equation with synchronously updating policies [21]. Radially unbounded Lyapunov functions were introduced in learning objective [22], [23] to ensure the stability of the evolving system and relax the necessity of initial stabilizing controller, which was tricky to find [24]. However, the persistency of excitation (PE) condition was required in the above algorithms to guarantee the persistent learning of the critic network, which also led to sample inefficiency [25]. To eliminate the PE condition, historical states were also utilized in tuning laws, where the requirement of the initial stabilizing controller was just the limitation [26], [27]. In summary, it is rather challenging to design a method that can remove the requirements of the PE condition and initial stabilizing controller and ensure the boundedness of the system state and weighting error.

For investigating nonlinear systems with model errors, the optimal robust guaranteed cost control problem was solved via constructing an uncertainty-related cost within the scope of

Manuscript received 31 August 2022; revised 22 May 2023; accepted 8 November 2023. Date of publication 28 November 2023; date of current version 8 January 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB2502901; and in part by the National Natural Science Foundation of China under Grant U20A20334, Grant 52221005, and Grant 52072213. *(Corresponding author: Shengbo Eben Li.)*

Jie Li, Yuhang Zhang, and Shengbo Eben Li are with the School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China (e-mail: jie-li18@mails.tsinghua.edu.cn; zhang-yh19@mails.tsinghua.edu.cn; lishbo@mail.tsinghua.edu.cn).

Ryozo Nagamune is with the Department of Mechanical Engineering, The University of British Columbia, Vancouver, BC V6T1Z4, Canada (e-mail: nagamune@mech.ubc.ca).

Digital Object Identifier 10.1109/TNNLS.2023.3335138

ADP techniques [28], [29]. Similar approaches were employed in tracking problems to acquire a desired tracking performance by ensuring an adequate level of cost [24]. To reinforce the training stability of the online system, a stabilizing term was considered in the learning criterion [30], [31]. Nevertheless, a probing noise composed of different sine waves was added to control inputs to satisfy the PE condition. Event-triggered $H_\infty$ control has also attracted attention, where the event-based mechanism was combined with critic learning to analyze the minimal sample interval time [32], [33]. Similarly, a probing noise was needed to satisfy the PE condition [32], [33], or initial stabilizing controllers were required and obtained via trial and error [34]. Beyond that, few studies considered both the model error and external disturbance.

In this work, we employ an uncertainty-conditioned cost (UCC) function to resolve nonlinear robust performance problems and bridge the above research gap. The robust controller generated from the augmented HJI equation is first demonstrated to fulfill disturbance attenuation performance for any considered model errors. Then, a critic learning method is developed to approximate the augmented HJI solution, where the requirements of the PE condition and initial stabilizing controller are relaxed. The system state and critic weighting error are finally proved to be uniformly ultimately bounded (UUB). The main contributions of this article are as follows.

1) Based on the zero-sum game scheme, whose original intention is to deal with disturbance, we further introduce a UCC function to the augmented HJI equation to cope with model errors. The conditions required for the UCC function to satisfy the Hamilton–Jacobi inequality in nonlinear robust control theory for all considered model errors are built. Theoretical analysis and experimental results prove that the closed-loop system with model errors, driven by the derived controller, achieves disturbance attenuation performance for any bounded perturbations.
2) We propose a critic learning method, which simultaneously removes the requirements of the PE condition and initial stabilizing controller by experience replay and Lyapunov stabilizing term, to solve the augmented HJI solution. The bounded stability of the closed-loop system and the boundedness of the critic weight error are proved by the Lyapunov extension theorem. Besides, introducing historical data is theoretically and experimentally validated to reduce the ultimate error bound of critic weight.

The rest of this article is organized as follows. In Section II, the nonlinear robust performance problem is illustrated. Section III derives an augmented HJI solution by designing a UCC function. A numerical algorithm along with its stability and convergence analysis is presented to approximately solve the augmented HJI equation in Section IV. Three numerical examples are shown in Section V to demonstrate the effectiveness of the developed method. Section VI concludes this work.

*Notation:* $\mathbb{R}$ denotes the set of real numbers. $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ denote the Euclidean space of $n$-dimensional real vectors and the space of $n \times m$ real matrices, respectively. $\|X\|$ denotes the norm of the vector or matrix $X$. The transpose operation of $X$ is denoted by $X^\top$. The $L_2$-norm of the signal $w(t)$ defined in $[0,\infty)$ is expressed as $\|w\|_2 \triangleq (\int_0^\infty w^\top(t)w(t)dt)^{1/2}$, and $w \in L_2[0,\infty)$ if $\|w\|_2 < \infty$. $I_n$ stands for an $n \times n$ identity matrix. For the symmetric matrix $X$, $\sqrt{X}$ represents the square root operation on the matrix, and $\lambda_{\min}(X)$ denotes the minimal eigenvalue. Besides, the gradient of $J(x)$ with respect to the vector $x$ is represented via $\partial J(x)/\partial x$ or $\nabla J(x)$.

## II. Problem Statement and Preliminaries

### A. Robust Performance Problem for Nonlinear Systems

Consider the following nonlinear plant:

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

with state $x \in \mathcal{X} \subseteq \mathbb{R}^n$ and control input $u \in \mathbb{R}^m$, where $\mathcal{X}$ is a compact set containing the origin. The functions $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$, and $k : \mathbb{R}^n \to \mathbb{R}^{n\times q}$ are known vector-valued or matrix-valued functions. Assume $f(x_0) = 0$ for some $x_0 \in \mathcal{X}$, and especially $f(0) = 0$. As in other literature, we assume that $f$, $g$, and $k$ are Lipschitz continuous on $\mathcal{X}$ and that the uncertain system (1) is controllable for any model errors $\Delta f(x)$ given below.

The uncertainty of the system comes from two aspects, namely external disturbance and model error. The disturbance signal $w \in \mathbb{R}^q$, and the perturbation function

$$\Delta f(x) = E_f(x)\delta_f(x) \tag{2}$$

represents the internal uncertainty caused by modeling error or parameter perturbation. $E_f(x) \in \mathbb{R}^{n\times p}$ is a known matrix-valued function and $\delta_f(x) \in \mathbb{R}^p$ is an uncertain vector-valued function. The set of perturbation functions is given by

$$\Omega_f \triangleq \{\Delta f \mid \Delta f = E_f(x)\delta_f(x),\ \delta_f(x_0) = 0 \\ \|\delta_f(x)\| \le m_f(x) \quad \forall x \in \mathcal{X}\} \tag{3}$$

where the known scalar-valued function $m_f(x)$ gives the boundary of $\delta_f(x)$. When $\Delta f(x) = 0$, the system is referred to as the nominal system.

To formulate the performance of the closed-loop system, we set $z \in \mathbb{R}^{n+m}$ as the performance output

$$z \triangleq \begin{bmatrix} \sqrt{Q}x \\ \sqrt{R}u \end{bmatrix} \tag{4}$$

and the square of its norm $\|z\|^2 = z^\top z = x^\top Qx + u^\top Ru$, where $Q \ge 0$ and $R > 0$ are symmetric matrices with $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$. Suppose that the system (1)–(4) is zero-state observable for all $\Delta f \in \Omega_f$. When the external disturbance $w$ acts on the system, it will affect the performance output $z$. The following definition characterizes the attenuation performance of the system against external disturbances.

*Definition 1 (Disturbance Attenuation):* For any disturbances $w \in L_2[0,\infty)$, if the $L_2$-gain of a system is less than or equal to $\gamma > 0$, that is,

$$\int_0^\infty \|z\|^2\, dt \le \gamma^2 \int_0^\infty \|w\|^2\, dt \tag{5}$$

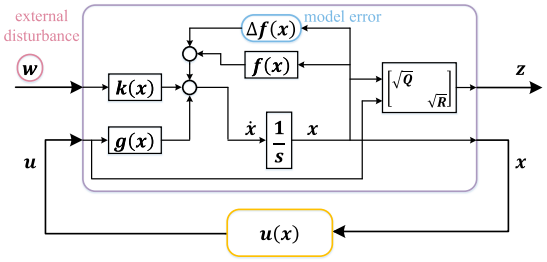


Fig. 1. Feedback system with an uncertain nonlinear system and a state feedback controller.

then the system is said to satisfy the disturbance attenuation performance with attenuation level $\gamma$.

The disturbance attenuation performance is actually the performance of concern in robust performance problems. Model error not only affects the asymptotic stability of the free system, but also affects the robustness of disturbance attenuation performance, that is, robust performance.

*Definition 2 (Robust Performance):* For any perturbations $\Delta f \in \Omega_f$, if the system (1) meets the disturbance attenuation performance with attenuation level $\gamma$ and the free system with $w = 0$ is asymptotically stable, then the system (1) is said to accomplish the robust performance with attenuation level $\gamma$.

Now, the nonlinear robust performance problem, to be considered in this article and represented by the block diagram in Fig. 1, is formulated as follows. From the perspective of optimization objectives, robust performance problems require achieving the goals of robust stabilization problems and disturbance attenuation problems simultaneously, which poses a huge challenge for controller designing.

*Problem 1 (Nonlinear Robust Performance Problem):* For a given uncertain nonlinear system (1) with external disturbance $w$ and perturbation set $\Omega_f$ described in (3), design a state feedback controller $u(x)$, such that for any perturbations $\Delta f \in \Omega_f$, the closed-loop system achieves preset attenuation level $\gamma$ and the free system with $w = 0$ is asymptotically stable.

### B. Preliminary

In this section, we will review an existing result in nonlinear robust control theory, as a preliminary for the proofs of our main results. Let us consider a simple nonlinear system with the only disturbance signal as the input vector

$$\begin{cases} \dot{x} = f(x) + k(x)w \\ z = h(x) \end{cases} \tag{6}$$

where $h$ is a nonlinear output function. The following lemma gives sufficient conditions for the system to comply with the disturbance attenuation performance with attenuation level $\gamma$.

*Lemma 1:* Consider the system (6) and let $\gamma > 0$. Assume that the system (6) is zero-state observable. If there exists a differentiable function $V(x) \geq 0$ with $V(x_0) = 0$, which fulfills the Hamilton–Jacobi inequality

$$(\nabla V(x))^\top f(x) + \frac{1}{4\gamma^2}(\nabla V(x))^\top k(x)k^\top(x)\nabla V(x) + z^\top z \leq 0 \tag{7}$$

for all $x \in \mathcal{X}$, then the function $V(x)$ meets the following dissipation inequality:

$$(\nabla V(x))^\top (f(x) + k(x)w) \leq -(\|z\|^2 - \gamma^2 \|w\|^2) \quad \forall w \in \mathbb{R}^m \tag{8}$$

and the $L_2$ gain of the system (6) is less than or equal to $\gamma$.

*Remark 1:* Some coefficients in the Hamilton–Jacobi inequality (7) are adjusted compared with Theorem 2 in [6]. As a result, the factors on both sides of the dissipation inequality (8) are the same, and the function $V(x)$ directly matches the integral of the cost function. A similar "completing the squares" skill can be exploited for Lemma 1, whose proof is omitted since it is a straightforward modification of that in [6].

Lemma 1 means that the nonlinear system (6) is bounded-input bounded-output (BIBO) stable. In addition to the disturbance attenuation performance, the asymptotic stability of the free system with $w = 0$ can be obtained from the assumption of zero-state observability through LaSalle's invariance principle [6], [7]. To comply with robust performance, a controller will be designed such that the closed-loop system satisfies the Hamilton–Jacobi inequality (7) regardless of parameter perturbation.

## III. Nonlinear Robust Performance Problem via Augmented HJI Solution

In this section, we will first construct a zero-sum game based on the nominal system with an additional cost function in Section III-A. Then, we will reveal how the cost function is devised to accomplish robust performance. With the UCC function, an augmented HJI equation will be derived, which is a nonlinear partial differential equation with respect to the value function. Finally, a robust controller can be extracted by estimating the solution to the augmented HJI equation.

### A. Zero-Sum Game of the Nominal System

Given the nominal system corresponding to the nonlinear uncertain system (1)

$$\dot{x} = f(x) + g(x)u + k(x)w \tag{9}$$

the value function of the initial state $x = x(0)$ is defined as

$$V(x) \triangleq \int_0^\infty (l(x(\tau), u(\tau), w(\tau)) + \Gamma(x(\tau)))d\tau \tag{10}$$

where

$$l(x, u, w) \triangleq \|z\|^2 - \gamma^2 \|w\|^2 = x^\top Qx + u^\top Ru - \gamma^2 w^\top w. \tag{11}$$

Note that $\Gamma(x)$ is an additional cost function that needs to be designed for robust performance. The method of settling the additional cost function will be given in Section III-B. Then a zero-sum game is formulated as

$$V^*(x) = \min_{u(\cdot)} \max_{w(\cdot)} \int_0^\infty (l(x(\tau), u(\tau), w(\tau)) + \Gamma(x(\tau)))d\tau \tag{12}$$

where $V^*(x)$ is the optimal value function and also the Nash value of the zero-sum game.

The key to resolving the zero-sum game is to solve the correlative Hamilton–Jacobi–Isaacs (HJI) equation [9]. First, define the Hamiltonian of the control problem as

$$H(x,u,w,\nabla V(x)) \triangleq l(x,u,w)+\Gamma(x) + (\nabla V(x))^{\top}\big(f(x)+g(x)u+k(x)w\big). \tag{13}$$

By applying the dynamic programming principle, the HJI equation is obtained as [9]

$$\min_{u}\max_{w} H\left(x,u,w,\nabla V^*(x)\right)=0. \tag{14}$$

Given the optimal value function $V^*(x)$, on the basis of two stationarity conditions, we have the optimal control policy $u^*(x)$ and the worst-case disturbance policy $w^*(x)$

$$\begin{aligned} u^*(x) &= \arg\min_{u} H\left(x,u,w,\nabla V^*(x)\right) \\ &= -\frac{1}{2}R^{-1}g^{\top}(x)\nabla V^*(x) \qquad (15)\\ w^*(x) &= \arg\max_{w} H\left(x,u,w,\nabla V^*(x)\right) \\ &= \frac{1}{2\gamma^2}k^{\top}(x)\nabla V^*(x). \qquad (16)\end{aligned}$$

Substituting the optimal control policy $u^*(x)$ and the worst-case disturbance policy $w^*(x)$ into the HJI equation (14), a nonlinear partial differential equation with respect to the optimal value function $V^*(x)$ with $V^*(x_0)=0$ is derived

$$x^{\top}Qx+\Gamma(x)+(\nabla V^*(x))^{\top}f(x) + \frac{1}{4}(\nabla V^*(x))^{\top}M(x)\nabla V^*(x)=0 \tag{17}$$

where

$$M(x)\triangleq \frac{1}{\gamma^2}k(x)k^{\top}(x)-g(x)R^{-1}g^{\top}(x). \tag{18}$$

So far, we have derived the basic form of the HJI solution for the nominal system (9). After the additional cost function $\Gamma(x)$ is determined to take into account the robustness against the uncertainty $\Delta f$ in Section III-B [see (24)], we will get the full expression of the augmented HJI equation at the end of this section [see (25)].

### B. Augmented HJI Equation With UCC Function

Given an attenuation level $\gamma$, the primary HJI equation (14) with $\Gamma(x)=0$ constructs a necessary and sufficient condition for the solution of disturbance attenuation problem [9]. By means of Lemma 1, we will find that incorporating an additional cost function related to model uncertainty $\Delta f$ contributes to achieving robust performance for the augmented HJI solution (19). The following Theorem 1 deduces the condition that the additional cost function needs to meet and gives a family of expressions for $\Gamma(x)$. For the convenience of writing, some functions $E_f(x)$, $\delta_f(x)$, and $m_f(x)$ will be abbreviated as $E_f$, $\delta_f$, and $m_f$, respectively.

*Theorem 1:* Suppose the nonlinear system (1)–(4) is zero-state observable for all $\Delta f\in\Omega_f$. If there exists a positive semidefinite function $\Gamma(x)\geq 0$, such that the augmented HJI equation admits a continuously differentiable positive semidefinite solution $V^*(x)\geq 0$ $(V^*(x_0)=V^*(0)=0)$

$$x^{\top}Qx+\Gamma(x)+(\nabla V(x))^{\top}f(x) + \frac{1}{4}(\nabla V(x))^{\top}M(x)\nabla V(x)=0 \tag{19}$$

and the additional cost function $\Gamma(x)$ satisfies

$$(\nabla V(x))^{\top}\Delta f(x)\leq \Gamma(x) \quad \forall \Delta f\in\Omega_f \ \forall x\in\mathcal{X} \tag{20}$$

then the nonlinear state feedback controller

$$u(x)=-\frac{1}{2}R^{-1}(x)g^{\top}(x)\nabla V(x) \tag{21}$$

makes the closed-loop system achieve robust performance.

*Proof:* With the given state feedback controller (21), the closed-loop dynamic system can be represented as

$$\dot{x}=f_{\mathrm{cl}}(x)+k(x)w \tag{22}$$

where

$$f_{\mathrm{cl}}(x)\triangleq f(x)+\Delta f(x)-\frac{1}{2}g(x)R^{-1}(x)g^{\top}(x)\nabla V(x)$$

and the square of the norm of the performance output becomes

$$z^{\top}z=x^{\top}Qx+\frac{1}{4}(\nabla V(x))^{\top}g(x)R^{-1}g^{\top}(x)\nabla V(x).$$

To use the conclusion in Lemma 1, substituting the above dynamic system (22) into (7) gives

$$\begin{aligned} &(\nabla V(x))^{\top}f_{\mathrm{cl}}(x)+\frac{1}{4\gamma^2}(\nabla V(x))^{\top}k(x)k^{\top}(x)\nabla V(x)+z^{\top}z \\ &= x^{\top}Qx+(\nabla V(x))^{\top}(f(x)+E_f\delta_f)+\frac{1}{4}(\nabla V(x))^{\top}M\nabla V(x) \\ &= (\nabla V(x))^{\top}E_f\delta_f-\Gamma(x)\leq 0\end{aligned}$$

where (19) and (20) are used for the last equality and inequality, respectively. Due to Lemma 1, the $L_2$-gain of the system is less than or equal to $\gamma$ for all $\Delta f\in\Omega_f$.

To show the asymptotic stability of the free system with the derived controller (21)

$$\dot{x}=f_{\mathrm{cl}}(x) \tag{23}$$

choose the solution $V(x)\geq 0$ of (19) as a Lyapunov candidate. According to the above derivation process, we have

$$\begin{aligned}\dot{V}(x) &= (\nabla V(x))^{\top}f_{\mathrm{cl}}(x) \\ &\leq -z^{\top}z-\frac{1}{4\gamma^2}(\nabla V(x))^{\top}k(x)k^{\top}(x)\nabla V(x) \\ &\leq 0.\end{aligned}$$

As a result, the case when $\dot{V}(x)=0$ leads to $\sqrt{Q}x=0$, $\sqrt{R}u=0$, and $\dot{x}=f(x)+\Delta f(x)$. Recalling the zero-state observability assumption, $x(t)\equiv 0$. Based on LaSalle's invariance principle [6], [7], the equilibrium $x_0=0$ of the free system (23) is asymptotically stable for all $\Delta f\in\Omega_f$. □

Since the additional cost function $\Gamma(x)$ is related to model uncertainty or model error, it is named the UCC function. Inspired by [28], the following corollary shows a general way to derive a family of UCC functions $\Gamma(x)$ with a parametric function $\lambda(x)$.

*Corollary 1:* The family of UCC functions

$$\Gamma(x) \triangleq \frac{m_f^2(x)}{\lambda^2(x)} + \frac{\lambda^2(x)}{4}(\nabla V(x))^\top E_f(x) E_f^\top(x) \nabla V(x)$$

with any scalar parametric functions $\lambda(x) > 0$ satisfies the following condition:

$$(\nabla V(x))^\top \Delta f(x) \le \Gamma(x) \quad \forall \Delta f \in \Omega_f \quad \forall x \in \mathcal{X}.$$

*Proof:* Substituting the expression of $\Gamma(x)$ into the above inequality gives

$$\begin{aligned}
&(\nabla V(x))^\top E_f \delta_f - \Gamma(x) \\
&\quad = -\frac{\lambda^2(x)}{4}(\nabla V(x))^\top E_f E_f^\top \nabla V(x) \\
&\qquad + (\nabla V(x))^\top E_f \delta_f - \frac{m_f^2}{\lambda^2(x)} \\
&\quad = -\left\| \frac{\lambda(x)}{2} E_f^\top \nabla V(x) - \frac{1}{\lambda(x)} \delta_f \right\|^2 - \frac{m_f^2 - \delta_f^\top \delta_f}{\lambda^2(x)} \\
&\quad \le 0.
\end{aligned}$$

□

Although $\lambda(x)$ is an adjustable function to form a sufficient condition to make (20) hold, it is usually fixed as $\lambda(x) \equiv 1$ in optimal robust guaranteed cost control problems [28]. In the following content, the UCC function $\Gamma(x)$ is instantiated as

$$\Gamma(x) = m_f^2(x) + \frac{1}{4}(\nabla V(x))^\top E_f(x) E_f^\top(x) \nabla V(x). \quad (24)$$

Based on nonlinear robust control theory, the instantiated UCC function makes the controller solved by the augmented HJI equation achieve robust performance for all considered model errors, but the primary HJI solution only achieves the preset disturbance attenuation performance for the nominal system. In theory, this reflects the advantage of the UCC function.

Finally, with the UCC function $\Gamma(x)$ in (24), the augmented HJI equation (19) becomes

$$\begin{aligned}
&x^\top Q x + m_f^2(x) + (\nabla V(x))^\top f(x) \\
&\quad + \frac{1}{4}(\nabla V(x))^\top \left( M(x) + E_f(x) E_f^\top(x) \right) \nabla V(x) = 0. \quad (25)
\end{aligned}$$

The augmented HJI equation (25) about the value function $V(x)$ is a novel nonlinear partial differential equation, which is different from that of the disturbance attenuation problem and the optimal robust guaranteed cost control problem. Since the analytical solution is rather difficult to identify, we will put forward a numerical ADP algorithm in Section IV to approximate its solution and extract the robust controller.

## IV. ADP-Based Algorithm for Approximating the Augmented HJI Solution

For the uncertain nonlinear system (1), a state feedback controller achieving robust performance for any perturbations can be obtained by solving a zero-sum game of the nominal system with a UCC function. However, it is rather difficult to reach the analytical solution to the augmented HJI equation (25). Thus, a numerical algorithm to solve (25) is necessary.

In light of the traditional approximate dynamic programming technique, this article exhibits an online algorithm by employing only one value network, namely the critic network. The schematic of the propounded ADP-based algorithm for attaining the robust controller is shown in Fig. 2, where the solid line denotes the signal, and the yellow dashed line indicates the backpropagation path for tuning the learned weight $\hat{\omega}_c$ of the critic network. Note that the first layer of the critic network represents the process of computing the feature vector $\sigma_c(x)$ from the state input $x$ through a fixed operation, which is different from the traditional multilayer perceptron. The algorithm framework and each block including critic network (32), control function, and disturbance function (34), the corresponding nominal system (35) with control input and disturbance input, approximate Hamiltonian (36) and stabilizing term (39) will be explained in detail with the matching equations in Sections IV-A and IV-B.

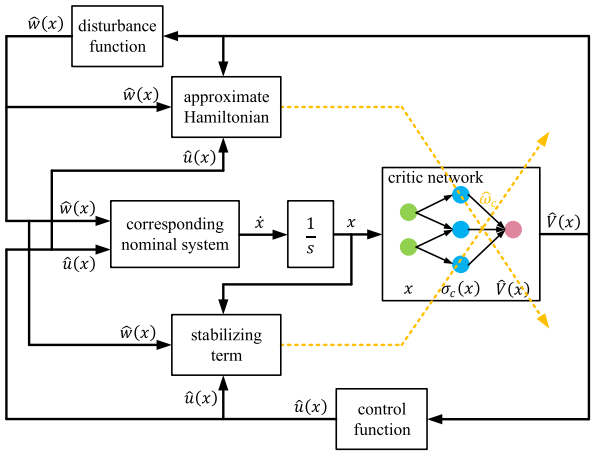


Fig. 2. Schematic of the developed algorithm.

### A. Neural Network Approximation

Noting the hypothesis in Theorem 1 that the solution $V^*(x)$ to the augmented HJI equation (19) is continuously differentiable, $V^*(x)$ can be approximated by neural networks based on the universal approximation property. To simplify the analysis, a single-layer neural network is employed

$$V^*(x) \triangleq \omega_c^\top \sigma_c(x) + \epsilon_c(x) \quad (26)$$

where $\omega_c \in \mathbb{R}^h$ is the ideal weight, $\sigma_c(x) \in \mathbb{R}^h$ is the feature function, usually consisting of polynomial basis, and $\epsilon_c(x) \in \mathbb{R}$ is the unknown approximation error. The expression of the solution varies for different dynamic systems. To our knowledge, it is tough to provide general guidelines for designing the feature function. For a specific system, experimental experience is beneficial for selecting feature functions. Then, the gradient of the value function is denoted as

$$\nabla V^*(x) = (\nabla \sigma_c(x))^\top \omega_c + \nabla \epsilon_c(x). \quad (27)$$

Assume that the weight $\omega_c$, the gradient of the feature $\nabla \sigma_c(x)$, the approximation error $\epsilon_c(x)$, and its gradient $\nabla \epsilon_c(x)$ are all bounded on the compact set $\mathcal{X}$. Recalling (15) and (16), the

optimal controller and worst-case disturbance are written as

$$u^*(x) = -\frac{1}{2}R^{-1}g^\top(x)\left(\left(\nabla\sigma_c(x)\right)^\top\omega_c + \nabla\epsilon_c(x)\right)$$
$$w^*(x) = \frac{1}{2\gamma^2}k^\top(x)\left(\left(\nabla\sigma_c(x)\right)^\top\omega_c + \nabla\epsilon_c(x)\right). \quad (28)$$

Under the optimal controller $u^*(x)$ and worst-case disturbance $w^*(x)$, the nominal system (9) can be rewritten as

$$\dot{x} = f(x) + g(x)u^*(x) + k(x)w^*(x)$$
$$= f(x) + \frac{1}{2}M(x)\left(\left(\nabla\sigma_c(x)\right)^\top\omega_c + \nabla\epsilon_c(x)\right). \quad (29)$$

For the convenience of writing in the subsequent derivation process, a matrix is introduced as follows:

$$\xi(x) \triangleq \nabla\sigma_c(x)\left(M(x) + E_f(x)E_f^\top(x)\right)\left(\nabla\sigma_c(x)\right)^\top. \quad (30)$$

Substituting the solution $V^*(x)$ approximated by the neural network (26) into the augmented HJI equation (25) yields

$$H(x,\omega_c) \triangleq x^\top Qx + m_f^2(x) + \omega_c^\top\nabla\sigma_c(x)f(x) + \frac{1}{4}\omega_c^\top\xi(x)\omega_c + e_H(x) = 0 \quad (31)$$

where

$$e_H(x) \triangleq \left(\nabla\epsilon_c(x)\right)^\top\left(f(x) + \left(M(x) + E_f(x)E_f^\top(x)\right)\times\left(\frac{1}{4}\nabla\epsilon_c(x) + \frac{1}{2}\left(\nabla\sigma_c(x)\right)^\top\omega_c\right)\right)$$

denotes the approximation error of the Hamiltonian caused by the neural network.

In our algorithm, the weight $\hat{\omega}_c$ of the estimated critic network $\hat{V}(x)$ will be trained and updated iteratively

$$\hat{V}(x) \triangleq \hat{\omega}_c^\top\sigma_c(x). \quad (32)$$

Accordingly, the gradient of the estimated critic network is denoted as

$$\nabla\hat{V}(x) = \left(\nabla\sigma_c(x)\right)^\top\hat{\omega}_c. \quad (33)$$

Moreover, the estimated state feedback controller $\hat{u}(x)$ and disturbance $\hat{w}(x)$ are expressed as

$$\hat{u}(x) = -\frac{1}{2}R^{-1}g^\top(x)\left(\nabla\sigma_c(x)\right)^\top\hat{\omega}_c$$
$$\hat{w}(x) = \frac{1}{2\gamma^2}k^\top(x)\left(\nabla\sigma_c(x)\right)^\top\hat{\omega}_c. \quad (34)$$

Under the estimated optimal controller $\hat{u}(x)$ and estimated worst-case disturbance $\hat{w}(x)$, the closed-loop dynamics of the nominal system (9) can be rewritten as

$$\dot{x} = f(x) + g(x)\hat{u}(x) + k(x)\hat{w}(x)$$
$$= f(x) + \frac{1}{2}M(x)\left(\nabla\sigma_c(x)\right)^\top\hat{\omega}_c. \quad (35)$$

Similarly, the Hamiltonian can be approximated via the estimated weight $\hat{\omega}_c$

$$\hat{H}\left(x,\hat{\omega}_c\right) \triangleq x^\top Qx + m_f^2(x) + \hat{\omega}_c^\top\nabla\sigma_c(x)f(x) + \frac{1}{4}\hat{\omega}_c^\top\xi(x)\hat{\omega}_c. \quad (36)$$

### B. Algorithm Design

The objective of our algorithm is to learn the estimated weight $\hat{\omega}_c$ of the critic network to minimize the approximation error of the Hamiltonian (36) in the augmented HJI equation. Furthermore, with the aim of ensuring stability during learning and relaxing the necessity of the PE condition, a Lyapunov stabilizing term [21], [31] mentioned in the following assumption and experience replay technique [25], [27] are also included in the algorithm design.

*Assumption 1:* Let $J_s(x)$ be a continuously differentiable, radially unbounded Lyapunov function that satisfies

$$\dot{J}_s(x) = \left(\nabla J_s(x)\right)^\top\left(f(x) + g(x)u^*(x) + k(x)w^*(x)\right) < 0. \quad (37)$$

Furthermore, suppose that there exists a positive definite matrix $\Phi \in \mathbb{R}^{n\times n}$ such that

$$\dot{J}_s(x) = -\left(\nabla J_s(x)\right)^\top\Phi\nabla J_s(x) \le -\lambda_{\min}(\Phi)\left\|\nabla J_s(x)\right\|^2. \quad (38)$$

*Remark 2:* This assumption has been applied in previous work to strengthen the stability during training and facilitate the stability analysis of closed-loop systems [21], [28]. The requirement of $J_s(x)$ being radially unbounded can be satisfied by suitably choosing a quadratic polynomial function of the state, such as $J_s(x) = 0.5\ x^\top x$.

The algorithm aims to minimize the error of the Hamiltonian (36). Hence, the primary objective is formulated as

$$E\left(x,\hat{\omega}_c\right) \triangleq \frac{1}{2}\hat{H}^2\left(x,\hat{\omega}_c\right). \quad (39)$$

To enhance the stability of the dynamic system during online learning, a stabilizing term inspired by Lyapunov stability theory is also considered [21]. Besides, for the sake of relaxing the requirement of the PE condition and improving sample efficiency, both historical states and current data [25] are utilized to achieve continuous learning and ensure convergence to a near-optimal solution. Take the derivative of the objective function with respect to the learned weight $\hat{\omega}_c$ to derive its dynamic expression

$$\dot{\hat{\omega}}_c = -\alpha_c\frac{\partial E\left(x,\hat{\omega}_c\right)}{\partial\hat{\omega}_c} - \alpha_c\sum_{i=1}^{N_b}\frac{\partial E\left(x_i,\hat{\omega}_c\right)}{\partial\hat{\omega}_c} - \alpha_s\Pi(x,\hat{u},\hat{w})\frac{\partial\dot{J}_s(x)}{\partial\hat{\omega}_c} \quad (40)$$

where $\alpha_c$ is the learning rate, $\alpha_s$ is a flexible rate to balance the steepest descent term and the stabilizing term, $\{x_i\}_{i=1}^{N_b}$ are stored historical states, $x$ is the current state, the nominal system (35) is used to derive the stabilizing term, and the unstability indicator function is defined as

$$\Pi(x,\hat{u},\hat{w}) \triangleq \begin{cases} 1, & \nabla J_s(x)^\top(f(x) + g(x)\hat{u} + k(x)\hat{w}) \ge 0 \\ 0, & \nabla J_s(x)^\top(f(x) + g(x)\hat{u} + k(x)\hat{w}) < 0. \end{cases}$$

*Remark 3:* The stabilizing term can not only relax the requirement of the initial stabilizing controller, but also stabilize the system and facilitate the subsequent stability proof [21], [28], [30]. When the system is unstable during

learning, the stabilizing term is activated through the unstability indicator function, and the learned weight evolves in the direction of stabilizing the system. On the other hand, when the system is stable, the unstability indicator function masks the stabilizing term, making the learning objective degenerate into the original target to minimize the approximation error of the Hamiltonian.

*Remark 4:* The historical states applied in online learning are sampled from a replay buffer, which is easy-to-implement. Generally speaking, the number $N_b$ of historical states needs to be large enough to ensure that the partial terms of weight gradient are linearly independent [25]. In this way, experience replay can play a similar role of persistent excitation to achieve continuous learning and guarantee that the Lyapunov derivative in the subsequent theoretical analysis is negative.

Define the error of the learned weight of the critic network as

$$\tilde{\omega}_c \triangleq \omega_c - \hat{\omega}_c \tag{41}$$

and the approximate Hamiltonian $\hat{H}(x, \hat{\omega}_c)$ can also be represented as a function of $\tilde{\omega}_c$

$$\begin{aligned}\hat{H}(x, \tilde{\omega}_c) = &\frac{1}{4}\tilde{\omega}_c^\top \xi(x)\tilde{\omega}_c - \frac{1}{2}\tilde{\omega}_c^\top \xi(x)\omega_c \\ &- \tilde{\omega}_c^\top \nabla\sigma_c(x) f(x) - e_H(x).\end{aligned} \tag{42}$$

Then, the dynamics of the weight estimation error $\tilde{\omega}_c$ is

$$\begin{aligned}\dot{\tilde{\omega}}_c = &-\alpha_c \frac{\partial E\left(x, \hat{\omega}_c\right)}{\partial \tilde{\omega}_c} - \alpha_c \sum_{i=1}^{N_b} \frac{\partial E\left(x_i, \hat{\omega}_c\right)}{\partial \tilde{\omega}_c} \\ &- \alpha_s \Pi(x, \hat{u}, \hat{w}) \frac{\partial \dot{J}_s(x)}{\partial \tilde{\omega}_c} \\ = &\left(\frac{1}{4}\tilde{\omega}_c^\top \xi(x)\tilde{\omega}_c - \frac{1}{2}\tilde{\omega}_c^\top \xi(x)\omega_c - \tilde{\omega}_c^\top \nabla\sigma_c(x) f(x) - e_H\right) \\ &\times \alpha_c \left(-\frac{1}{2}\xi(x)\tilde{\omega}_c + \frac{1}{2}\xi(x)\omega_c + \nabla\sigma_c(x) f(x)\right) \\ &- \alpha_c \sum_{i=1}^{N_b} \frac{\partial E\left(x_i, \hat{\omega}_c\right)}{\partial \tilde{\omega}_c} \\ &+ \frac{\alpha_s}{2}\Pi(x, \hat{u}, \hat{w}) \nabla\sigma_c(x) M(x) \nabla J_s(x).\end{aligned} \tag{43}$$

The stability analysis of the closed-loop system and the convergence study of the learned weight of the critic network will be presented in Section IV-C via the well-known Lyapunov theory.

*Remark 5:* The raised algorithm is an online approach that needs to interact with the system. In the learning process, the analytical expression of the system (1)–(4) is required, and the state vectors applied to learning the critic network are generated in an actual environment driven by learned or estimated policies.

### C. Stability and Convergence Analysis

In this section, by examining the dynamic properties of the integrated system consisting of the nonlinear system (1) with the inputs (34) and the tuning algorithm (40), the stability of the closed-loop system and the convergence of the weight estimation error $\tilde{\omega}_c$ will be analyzed. The following definition [21] is provided before presenting the analysis results.

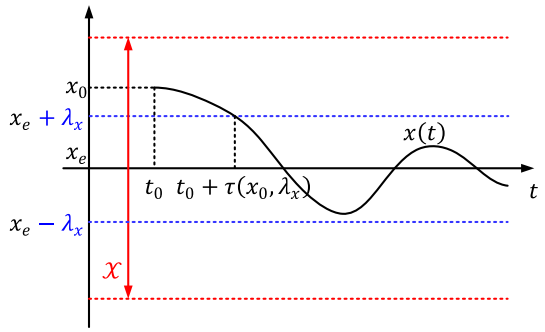


Fig. 3. Concept of UUB.

*Definition 3 (UUB):* If there is a compact set $\mathcal{X}$ so that for all $x_0 \in \mathcal{X}$, there exists a bound $\lambda_x > 0$ and a time $\tau(x_0, \lambda_x)$ such that $\|x(t) - x_e\| \leq \lambda_x$ for all $t \geq t_0 + \tau(x_0, \lambda_x)$, then the trajectory $x(t)$ around the equilibrium point $x_e$ is said to be UUB, whose concept is illustrated in Fig. 3.

Before the analysis, the following bounded assumptions of system dynamics are provided to keep consistent with previous studies [21], [28]. Besides, an assumption about the Lyapunov function is also given to characterize the stability of the optimal closed-loop dynamics.

*Assumption 2:* Some bounded assumptions about the dynamic system and neural network are given as follows.

1) In the nominal dynamics (9), matrices $g(x)$ and $k(x)$ are bounded as $\|g(x)\| \leq \lambda_g$ and $\|k(x)\| \leq \lambda_k$, where $\lambda_g$ and $\lambda_k$ are given positive constants.
2) The ideal weight $\omega_c$ in (26) is bounded by a given positive constant $\lambda_\omega$, that is, $\|\omega_c\| \leq \lambda_\omega$.
3) On the compact set $\mathcal{X}$, the terms $\nabla\sigma_c(x)$ and $\nabla\epsilon_c(x)$ in (27), and the approximation error of Hamiltonian $e_H(x)$ in (31) are bounded as $\|\nabla\sigma_c(x)\| \leq \lambda_\sigma$, $\|\nabla\epsilon_c(x)\| \leq \lambda_\epsilon$, and $|e_H(x)| \leq \lambda_e$, where $\lambda_\sigma$, $\lambda_\epsilon$, and $\lambda_e$ are given positive constants.
4) Because of the boundedness of $\nabla\sigma_c(x)$ and the state trajectory $\dot{x}$ (29) under the optimal controller and worst-case disturbance, it can be inferred that $\|\nabla\sigma_c(x)\dot{x}^*\| \leq \lambda_{\sigma x}$, where $\lambda_{\sigma x}$ is a given positive constant.

The following theorems present the main results of the stability and convergence analysis of the proposed tuning algorithm (40). The stability of the system during the implementation of the algorithm can be guaranteed by proving that the state is UUB, and the convergence can be obtained by showing that the weight estimation error is UUB [21]. First, we explore the simple case of not utilizing experience replay.

*Theorem 2:* Consider the nonlinear system (1) with both model errors and external disturbance. If the weight of the critic network is tuned by (40) without employing historical states, then the weight estimation error $\tilde{\omega}_c$ and the state $x$ of the system with the estimated policies (34) are UUB.

*Proof:* See Appendix A for the proof. □

According to the theoretical derivation of the above simple case ignoring experience replay, the stability and convergence analysis of the complete tuning law (40) is presented below.

*Theorem 3:* Follow the premise of Theorem 2. If the weight of the critic network is tuned by (40), that is, considering historical states, the weight estimation error $\tilde{\omega}_c$ and the state $x$ of the system with the estimated policies (34) are still UUB. Moreover, introducing historical data into the critic tuning law (40) can reduce ultimate bounds $\lambda_{\tilde{\omega}_c}$ and $\lambda_x$.

*Proof:* To compare with the previous theoretical analysis, adjust the Lyapunov function candidate (A.1) to

$$L\left(\tilde{\omega}_c, x\right) \triangleq \frac{1}{2\alpha_c(N_b+1)}\tilde{\omega}_c^\top \tilde{\omega}_c + \frac{\alpha_s}{\alpha_c} J_s(x). \tag{44}$$

Then, the same results of Theorem 2 can be directly obtained, because the applied zoom techniques are valid for all states.

To further explore the influence of historical data in the actual implementation, it can be observed that the first term of the Lyapunov derivative (A.3) changes from the single term of the current state to the average term of $(N_b+1)$ states. Therefore, the inequality related to $\xi(x)$ becomes

$$\frac{\left(\tilde{\omega}_c^\top \xi(x)\tilde{\omega}_c\right)}{N_b+1}\left(\tilde{\omega}_c^\top\left(\xi(x)+\sum_{i=1}^{N_b}\xi(x_i)\right)\tilde{\omega}_c\right) \geq \frac{\underline{\lambda}_\xi \underline{\lambda}_{\Sigma\xi}}{N_b+1}\left\|\tilde{\omega}_c\right\|^4$$

where the lower bound of the norm of $(\xi(x)+\sum_{i=1}^{N_b}\xi(x_i))/(N_b+1)$ has

$$\frac{\underline{\lambda}_{\Sigma\xi}}{N_b+1} \geq \underline{\lambda}_\xi.$$

The corresponding conclusion also holds for $\underline{\lambda}_\rho$. Similarly, the inequality related to $\kappa(x)$ becomes

$$\frac{\left(\tilde{\omega}_c^\top \kappa(x)\tilde{\omega}_c\right)}{N_b+1}\left(\tilde{\omega}_c^\top\left(\kappa(x)+\sum_{i=1}^{N_b}\kappa(x_i)\right)\tilde{\omega}_c\right) \leq \frac{\bar{\lambda}_\kappa \bar{\lambda}_{\Sigma\kappa}}{N_b+1}\left\|\tilde{\omega}_c\right\|^4$$

where the upper bound of the norm of $(\kappa(x)+\sum_{i=1}^{N_b}\kappa(x_i))/(N_b+1)$ has

$$\frac{\bar{\lambda}_{\Sigma\kappa}}{N_b+1} \leq \bar{\lambda}_\kappa.$$

Relevant results also hold for $\bar{\lambda}_\vartheta$ and $\bar{\lambda}_\rho$. Note that in (A.4) and (A.5), $\lambda_1 \propto \underline{\lambda}_\xi$, $\lambda_1 \propto -\bar{\lambda}_\kappa$, $\lambda_1 \propto -\bar{\lambda}_\vartheta$, $\lambda_1 \propto -\bar{\lambda}_\rho$, $\lambda_2 \propto -\underline{\lambda}_\rho$, and $\lambda_2 \propto \bar{\lambda}_\rho$. Therefore, under other unchanged conditions, employing historical states in the tuning law (40) tends to make $\lambda_1$ larger and $\lambda_2$ smaller. According to the definition of $\lambda_4$ in (A.6), $\lambda_4 \propto \lambda_2$, and $\lambda_4 \propto 1/\lambda_1$. Thus, $\lambda_4$ tends to become smaller. Finally, combined with definition formulas (A.7)–(A.10), it can be concluded that introducing historical data can reduce ultimate bounds $\lambda_{\tilde{\omega}_c}$ and $\lambda_x$. □

*Remark 6:* The reduction of the ultimate bound $\lambda_{\tilde{\omega}_c}$ of the weighting error has the potential to accelerate the convergence of critic weight [25]. Experimental results also demonstrate that introducing historical data helps reduce the ultimate error bound and speed up the convergence of the method.

Once the convergence of the critic network has been demonstrated, the same technique can be applied to the control function. The following corollary gives a simple proof of the convergence of the control policy [31].

*Corollary 2:* The estimated controller $\hat{u}(x)$ in (34) converges to a bounded neighborhood of the optimal one $u^*(x)$ in (28).

*Proof:* Based on (34) and (28), we have

$$\hat{u}(x) - u^*(x) = \frac{1}{2}R^{-1}g^\top(x)\left(\left(\nabla\sigma_c(x)\right)^\top\tilde{\omega}_c + \nabla\epsilon_c(x)\right).$$

In consequence of the result in Theorem 2 or Theorem 3, $\|\tilde{\omega}_c\| < \lambda_{\tilde{\omega}_c}$ ultimately holds. Combining assumptions 1) and 3) introduced in Assumption 2, we have

$$\left\|\hat{u}(x) - u^*(x)\right\| \leq \frac{1}{2}\left\|R^{-1}\right\|\lambda_g\left(\lambda_\sigma\lambda_{\tilde{\omega}_c} + \lambda_\epsilon\right) \triangleq \lambda_{\hat{u}}$$

which completes the proof. □

## V. Numerical Results

In this section, the proposed algorithm is compared with existing methods for robust performance in three simulation studies to verify its effectiveness. The first example considers an uncertain linear system, while the other two examples deal with uncertain nonlinear systems taken from the literature. The influence of historical data on convergence characteristics is verified in the first linear case with precise numerical solutions. Complex feature vectors are applied to the second example to explore their impact on performance. In addition, an ablation study is conducted in the third example to comprehend the superiority of the UCC function, where the developed method is directly applied to approximate the primary HJI solution.

### A. Two-Dimensional Linear System

Consider the following mass-spring-damper system:

$$\dot{x} = (A + \Delta A)\,x + B_1 w + B_2 u \tag{45}$$

where the system matrices are given by

$$A = \begin{bmatrix} 0 & 1 \\ -\dfrac{k_0}{m} & -\dfrac{b_0}{m} \end{bmatrix}, \quad \Delta A = \begin{bmatrix} 0 & 0 \\ -\dfrac{k_0}{m}\delta_k & -\dfrac{b_0}{m}\delta_b \end{bmatrix}$$

$$B_1 = B_2 = \begin{bmatrix} 0 \\ \dfrac{1}{m} \end{bmatrix}$$

and the state vector consists of the position $x_1$ [m] and the velocity $x_2$ [m/s] as $x = [x_1, x_2]^\top$. The value of the mass $m$ [kg] as well as the nominal values of the elastic coefficient $k_0$ [N/m] and damping coefficient $b_0$ [kg/s] are

$$m = 1 \text{ kg}, \ k_0 = 3 \text{ N/m}, \ b_0 = 2 \text{ kg/s}$$

and the perturbation ranges of these parameters are

$$\delta_b = \frac{b - b_0}{b_0} \in \left[-\frac{1}{2}, \frac{1}{2}\right], \quad \delta_k = \frac{k - k_0}{k_0} \in \left[-\frac{1}{2}, \frac{1}{2}\right].$$

Therefore, the perturbation function

$$\Delta f(x) = E_f(x)\delta_f(x) = \begin{bmatrix} 0 & 0 \\ -\dfrac{k_0}{m} & -\dfrac{b_0}{m} \end{bmatrix}\begin{bmatrix} \delta_k x_1 \\ \delta_b x_2 \end{bmatrix}$$

where

$$\|\delta_f(x)\| = \left\| \begin{bmatrix} \delta_k x_1 \\ \delta_b x_2 \end{bmatrix} \right\| \leq \frac{1}{2}\|x\| = m_f(x), \quad M_f = \frac{1}{2} I_2.$$

In this example, tuning parameters for control and learning were selected as follows:

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 20 \end{bmatrix}, \quad R = 0.1, \ \gamma = 0.5.$$

A polynomial-based critic network [10]

$$\hat{V}(x) = \hat{\omega}_c^\top \sigma_c(x)$$
$$\sigma_c(x) = \left[x_1^2, x_1 x_2, x_2^2\right]^\top, \quad \hat{\omega}_c = [\omega_1, \omega_2, \omega_3]^\top$$

was employed to approximate the solution of the augmented HJI equation, where $\sigma_c(x)$ was the feature vector, and $\hat{\omega}_c$ was the learned weight. The learning rates of the critic network were chosen as $\alpha_c = 0.01$ and $\alpha_s = 0.5$. The Lyapunov function was chosen as a quadratic polynomial function, that is,

$$J_s(x) = 0.5\, x^\top x.$$

The initial state of its nominal system

$$\dot{x} = Ax + B_1 w + B_2 u \tag{46}$$

was set to $x_0 = [0.7, -0.3]^\top$. The learned weight $\hat{\omega}_c$ was initialized with zero vector, meaning that no initial stabilizing control policy was required.

During the learning process, the number of historical states $N_b = 32$. After learning for 10 s, the learned weight $\hat{\omega}_c$ converged to $[8.98808575, 1.74772668, 11.38678169]^\top$ as shown in Fig. 4. Note that, for linear systems, the augmented HJI equation degenerates into an algebraic Riccati equation

$$A^\top P + PA + Q + M_f^\top M_f - P\left(B_2 R^{-1} B_2^\top - \frac{1}{\gamma^2} B_1 B_1^\top - E_f E_f^\top\right) P = 0 \tag{47}$$

whose exact numerical solution can be obtained via MATLAB, that is, $\omega_c = [8.98808751, 1.74772708, 11.38678365]^\top$. Denote the relative error of the weight of the critic network as

$$e \triangleq \frac{\|\hat{\omega}_c - \omega_c\|_2}{\|\omega_c\|_2}. \tag{48}$$

It can be calculated that the relative error of critic weight is below one-millionth, which shows that our method converges to the optimal solution.

For further verifying the impact of experience replay on the convergence speed and ultimate error bound of critic weight, we chose different numbers of historical states in the algorithm implementation. For each parameter, the relative error of critic weight is depicted in Fig. 5. With the growth of historical states employed in the proposed critic tuning method, the convergence of critic weight is accelerating. On the other hand, as the number of historical states increases from 8 to 32, the ultimate error bound gradually decreases from $10^{-5}$ to around $10^{-7}$. The above experimental results validate the convergence analysis presented in Theorem 3.

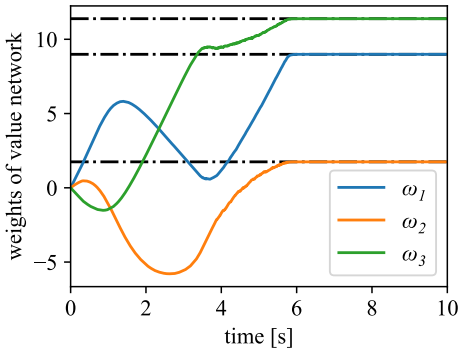


Fig. 4. Weight of the critic network for the linear uncertain system.

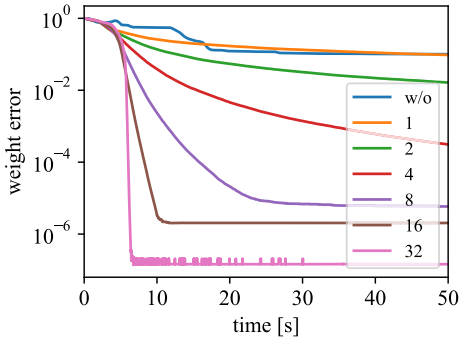


Fig. 5. Influence of the number of historical states on the convergence speed and error bound of critic weight.

To illustrate the robust performance of the raised method against model uncertainties, sinusoidal disturbance and white noise disturbance were applied to the original uncertain system (45). The actual disturbance attenuation level $\hat{\gamma}$ was calculated during the simulation as

$$\hat{\gamma} \triangleq \sqrt{\frac{\int_0^T \|z\|^2 dt}{\int_0^T \|w\|^2 dt}} \quad \forall t \in (0, T) \tag{49}$$

which represented the robustness of the control method against external disturbances. Comparison methods include as follows.

1) *LQR:* Static controller based on the linear quadratic regulator.
2) *LMI:* Robust $H_\infty$ synthesis method solved by linear matrix inequalities (LMIs) with the same performance output and preset attenuation level [35].
3) *OLA:* Online approximator-based tuning method derived from two-player zero-sum game [21].
4) *RADP:* Robust approximate dynamic programming proposed in this work.

Simulation results under different parameter perturbations are shown in Fig. 6, where the raincloud plot is employed to provide an intuitive form of data visualization. In essence, it combines scatter points, violin plots, and boxplots to provide an overview of raw data, probability distribution, and statistical inference by different quantiles and confidence intervals. Each method reports a total of 81 different sets of parameters, which are uniformly selected from the parameter perturbation space. The vertical axis is the convergent actual disturbance attenuation level $\hat{\gamma}$, that is, 50 s for sinusoidal disturbance and

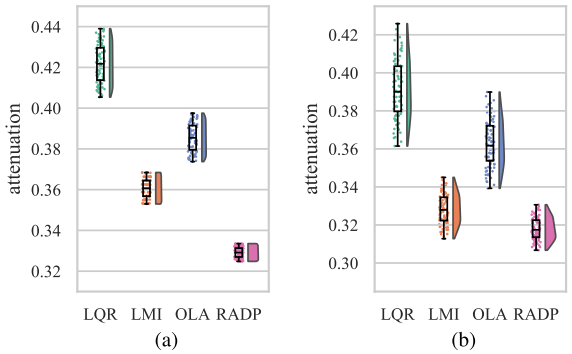


Fig. 6. Simulation comparison results of the first example. (a) Sinusoidal disturbance. (b) White noise disturbance.

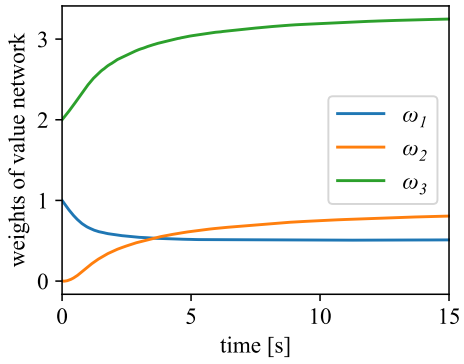


Fig. 7. Weight of the critic network for the second example.

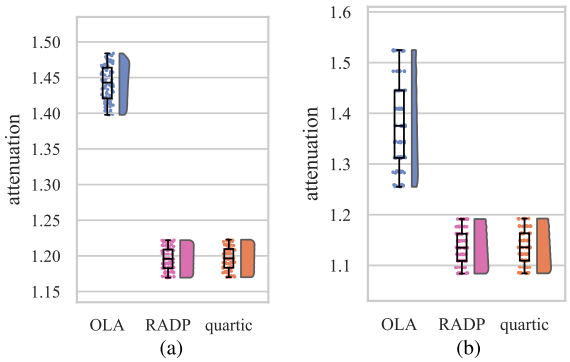


Fig. 8. Simulation comparison results of the second example. (a) Sinusoidal disturbance. (b) White-noise disturbance.

250 s for white-noise disturbance. The range of the confidence interval for different control methods reflects the degree to which the disturbance attenuation performance is affected by model error, that is, robust performance. Simulations of different disturbances demonstrate that the raised RADP method exhibits the lowest disturbance attenuation level and its attenuation performance is least affected by model error. As shown in Fig. 6(b), the actual attenuation levels under white noise disturbance for RADP and LMI are similar. Thus, it can be considered that the robust performance of the proposed method is comparable to the traditional robust $H_\infty$ synthesis method. Moreover, the compared OLA method approximately solves the original HJI equation. The comparison results show the advantages of the proposed method and UCC function.

### B. Two-Dimensional Nonlinear System

Consider the nonlinear system mentioned in [30]

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -\frac{x_1+x_2}{2} + \frac{x_2(\cos 2x_1+2)^2}{2} - \frac{x_2(\sin 4x_1+2)^2}{\gamma^2} \end{bmatrix} + \begin{bmatrix} \delta_1 x_2 \sin x_1 \\ \delta_2 x_1 \cos x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \cos 2x_1 + 2 \end{bmatrix} u + \begin{bmatrix} 0 \\ \sin 4x_1 + 2 \end{bmatrix} w \tag{50}$$

where $x = [x_1, x_2]^\top$. The perturbation function

$$\Delta f(x) = E_f(x)\delta_f(x) = I_2 \begin{bmatrix} \delta_1 x_2 \sin x_1 \\ \delta_2 x_1 \cos x_2 \end{bmatrix}$$

where $\delta_1$ and $\delta_2$ are uncertain parameters with

$$\delta_1 \in [-1, 1], \quad \delta_2 \in [-1, 1]$$
$$\|\delta_f(x)\| = \left\| \begin{bmatrix} \delta_1 x_2 \sin x_1 \\ \delta_2 x_1 \cos x_2 \end{bmatrix} \right\| \le \left\| \begin{bmatrix} x_2 \sin x_1 \\ x_1 \cos x_2 \end{bmatrix} \right\| = m_f(x).$$

In this nonlinear example, tuning parameters for control and learning were selected as follows:

$$Q = 2I_2, \quad R = 2, \quad \gamma = 2.$$

A quadratic polynomial was employed in the critic network

$$\hat{V}(x) = \hat{\omega}_c^\top \sigma_c(x)$$
$$\sigma_c(x) = \left[x_1^2, x_1 x_2, x_2^2\right]^\top, \quad \hat{\omega}_c = [\omega_1, \omega_2, \omega_3]^\top$$

where $\sigma_c(x)$ was the feature vector, and $\hat{\omega}_c$ was the learned weight. The learning rates of the critic network were chosen as $\alpha_c = 0.003$ and $\alpha_s = 0.03$. The initial state was set to $x_0 = [0.3, 0.3]^\top$. Note that the primary HJI equation of the nominal system with the utility function (11) could be solved with the help of converse optimal control method [36], whose solution $[1, 0, 2]^\top$ was employed to initialize the learned weight $\hat{\omega}_c$ to speed up the training progress. The number of historical states $N_b = 64$. After learning for 15 s, the learned weight $\hat{\omega}_c$ converged to $[0.5108, 0.8059, 3.2485]^\top$ as shown in Fig. 7.

To compare the effects of other feature vectors, we applied a fourth-degree (quartic) polynomial function to approximate the critic network, that is,

$$\hat{V}(x) = \hat{\omega}_c^\top \sigma_c(x), \quad \hat{\omega}_c = [\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8]^\top$$
$$\sigma_c(x) = \left[x_1^2, x_1 x_2, x_2^2, x_1^4, x_1^3 x_2, x_1^2 x_2^2, x_1 x_2^3, x_2^4\right]^\top.$$

The other implementation details remained unchanged. After learning for 15 s, the learned weight $\hat{\omega}_c$ is converged to

$$[0.5235, 0.7986, 3.2411, -0.0999, 0.0139, 0.0681, 0.1051, 0.1318]^\top.$$

To verify the robust performance of the raised method, sinusoidal and white-noise disturbances were applied to the nonlinear uncertain system (50). Repeated experiments were carried out for each method, where 81 different sets of uncertain parameters were sampled uniformly from the perturbation space. The OLA method, which approximated the HJI equation of the nominal system with the same performance output and preset attenuation level, was utilized for the comparison. The actual disturbance attenuation level $\hat{\gamma}$ in (49) was

exploited to characterize the robust performance of different controllers. Simulation results are shown in Fig. 8. It is evident that the proposed method enjoys a lower level of disturbance attenuation and is less affected by model error. Therefore, the robust performance of the controller improved by the UCC function (24) is illustrated. Additionally, the performance of the controller obtained by a quartic polynomial is similar to that obtained by a quadratic polynomial. This indicates that a relatively simple feature vector can achieve the success of the method, while a complex feature vector is not a necessary condition for the algorithm to achieve better results.

### C. Three-Dimensional Nonlinear System

Consider the nonlinear system mentioned in [28] and [30]

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ 0.1x_1 - x_2 - x_1x_3 \\ x_1x_2 - x_3 \end{bmatrix} + \begin{bmatrix} \delta x_1 \sin x_2 \cos x_3 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} w \quad (51)$$

where $x = [x_1, x_2, x_3]^\top$, $\delta \in [-1, 1]$ is an indeterminate parameter in the perturbation function

$$\Delta f(x) = E_f(x)\delta_f(x) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \delta x_1 \sin x_2 \cos x_3$$

where

$$\|\delta_f(x)\| = \|\delta x_1 \sin x_2 \cos x_3\| \le \|x_1 \sin x_2 \cos x_3\| = m_f(x).$$

In this example, tuning parameters for control and learning were selected as follows:

$$Q = 8I_3, \quad R = 5, \ \gamma = 5.$$

The critic network was approximated by

$$\hat{V}(x) = \hat{\omega}_c^\top \sigma_c(x)$$
$$\sigma_c(x) = \left[x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3\right]^\top$$
$$\hat{\omega}_c = [\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6]^\top$$

where $\sigma_c(x)$ was the feature vector, and $\hat{\omega}_c$ was the learned weight. The learning rates of the critic network were chosen as $\alpha_c = 0.03$ and $\alpha_s = 0.2$. The initial state was set to $x_0 = [2, 2, -1]^\top$, and the learned weight $\hat{\omega}_c$ was initialized with zero vector. The number of historical states $N_b = 64$. As shown in Fig. 9, after learning for 15 s, the learned weight $\hat{\omega}_c$ eventually converged to $[4.2239, 5.0494, -0.3062, 5.4908, -2.0132, 5.5768]^\top$.

To compare robust performance, sinusoidal and white-noise disturbances were applied to the original uncertain system (51) under 21 sets of uniformly sampled parameters. Comparison methods included the OLA algorithm, which incorporated the stabilizing term and solved the original HJI equation without the UCC function. Besides, the developed critic tuning method considering both stabilizing term and experience replay was also applied to the primary HJI equation without the UCC function. The actual disturbance attenuation level $\hat{\gamma}$ in (49)

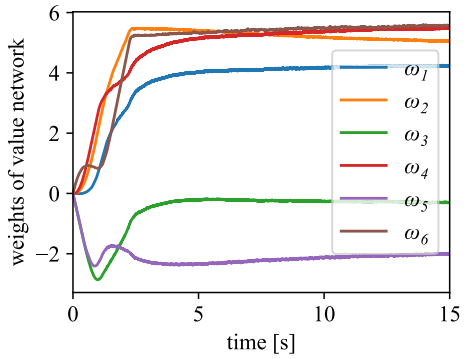


Fig. 9. Weight of the critic network for the third example.

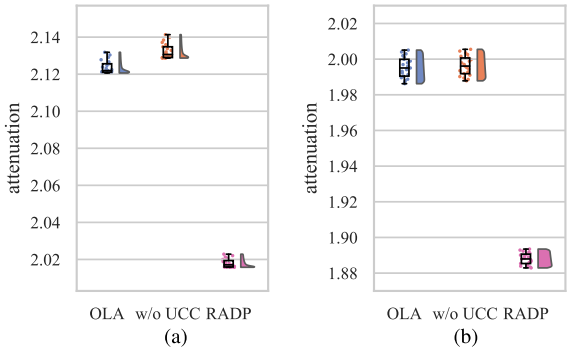


Fig. 10. Simulation comparison results of the third example. (a) Sinusoidal disturbance. (b) White-noise disturbance.

was used to characterize the robust performance of different controllers. Simulation results are shown in Fig. 10. It can be seen that even if the proposed RADP method utilizes historical data for learning, without considering the UCC function, it can only achieve robust performance similar to the OLA algorithm under different types of disturbance. This ablation study rules out the possibility of improving robust performance through the experience replay technique. On the other hand, the completed RADP method with the UCC function has a lower disturbance attenuation level and is less affected by model error under different disturbances. The results demonstrate the effectiveness of the proposed RADP approach and the superiority of the introduced UCC function.

## VI. Conclusion

A novel ADP-based algorithm, which relaxed the requirements of the PE condition and initial stabilizing controller, was put forward to solve robust performance problems of uncertain nonlinear systems. The robust performance of the obtained robust controller was achieved by constructing an augmented Hamilton–Jacobi–Isaacs (HJI) equation of the corresponding nominal system with a UCC function. The weights of the estimated critic network were tuned to approximate the solution of the augmented HJI equation, with both convergence and stability guarantees. Three numerical examples were presented to verify the efficacy and robustness of the proposed RADP algorithm.

## APPENDIX

### A. Proof of Theorem 2

*Proof:* First, choose a Lyapunov function candidate as

$$L\left(\tilde{\omega}_c, x\right) \triangleq \frac{1}{2\alpha_c}\tilde{\omega}_c^\top \tilde{\omega}_c + \frac{\alpha_s}{\alpha_c} J_s(x). \tag{A.1}$$

When applying our algorithm, the nominal system operates under the drive of the estimated optimal controller $\hat{u}(x)$ and worst-case disturbance $\hat{w}(x)$. The time derivative along the dynamics of the closed-loop system (35) and the weight estimation error (43) is

$$\begin{aligned}\dot{L}\left(\tilde{\omega}_c, x\right) &= \frac{1}{\alpha_c}\tilde{\omega}_c^\top \dot{\tilde{\omega}}_c + \frac{\alpha_s}{\alpha_c}\left(\nabla J_s(x)\right)^\top \dot{x} \\ &= \frac{\alpha_s}{\alpha_c}\left(\nabla J_s(x)\right)^\top \dot{x} \\ &\quad - \tilde{\omega}_c^\top\Bigg(-\frac{1}{4}\tilde{\omega}_c^\top \xi(x)\tilde{\omega}_c + \frac{1}{2}\tilde{\omega}_c^\top \xi(x)\omega_c \\ &\qquad + \tilde{\omega}_c^\top \nabla\sigma_c(x) f(x) + e_H\Bigg) \\ &\quad \times\left(-\frac{1}{2}\xi(x)\tilde{\omega}_c + \frac{1}{2}\xi(x)\omega_c + \nabla\sigma_c(x) f(x)\right) \\ &\quad + \frac{\alpha_s}{2\alpha_c}\Pi(x,\hat{u},\hat{w})\tilde{\omega}_c^\top \nabla\sigma_c(x) M(x)\nabla J_s(x).\end{aligned}$$

Then, the time derivative of $(1/2\alpha_c)\tilde{\omega}_c^\top\tilde{\omega}_c$ will transit from the trajectory (35) of the estimated policies to that (29) of the optimal policies. To distinguish different trials, the trajectory (29) under the optimal policies is denoted as $\dot{x}^*$. For the convenience of writing, some matrices will be introduced:

$$\begin{aligned}\vartheta(x) &\triangleq \nabla\sigma_c(x) g(x) R^{-1} g^\top(x)\left(\nabla\sigma_c(x)\right)^\top \\ \kappa(x) &\triangleq \frac{1}{\gamma^2}\nabla\sigma_c(x) k(x) k^\top(x)\left(\nabla\sigma_c(x)\right)^\top \\ \rho(x) &\triangleq \nabla\sigma_c(x) E_f(x) E_f^\top(x)\left(\nabla\sigma_c(x)\right)^\top \\ \vartheta_{\sigma\epsilon}(x) &\triangleq \nabla\sigma_c(x) g(x) R^{-1} g^\top(x)\left(\nabla\epsilon_c(x)\right)^\top \\ \kappa_{\sigma\epsilon}(x) &\triangleq \frac{1}{\gamma^2}\nabla\sigma_c(x) k(x) k^\top(x)\left(\nabla\epsilon_c(x)\right)^\top.\end{aligned} \tag{A.2}$$

According to the abbreviated matrix defined in (30), we have

$$\xi(x) = \kappa(x) - \vartheta(x) + \rho(x).$$

For the symmetric matrices $\vartheta(x)$, $\kappa(x)$, $\rho(x)$, and $\xi(x)$ mentioned above, the lower- and upper-bound operations for their norms are assumed to be $\underline{\lambda}_\vartheta > 0$, $\bar{\lambda}_\vartheta > 0$, $\underline{\lambda}_\kappa > 0$, $\bar{\lambda}_\kappa > 0$, $\underline{\lambda}_\rho > 0$, $\bar{\lambda}_\rho > 0$, $\underline{\lambda}_\xi > 0$, and $\bar{\lambda}_\xi > 0$, respectively. Note that

$$\begin{aligned}\tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x} &= \tilde{\omega}_c^\top\nabla\sigma_c(x) f(x) + \frac{1}{2}\tilde{\omega}_c^\top\vartheta(x)\tilde{\omega}_c \\ &\quad - \frac{1}{2}\tilde{\omega}_c^\top\kappa(x)\tilde{\omega}_c - \frac{1}{2}\tilde{\omega}_c^\top\vartheta(x)\omega_c + \frac{1}{2}\tilde{\omega}_c^\top\kappa(x)\omega_c \\ \tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x}^* &= \tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x} - \frac{1}{2}\tilde{\omega}_c^\top\vartheta(x)\tilde{\omega}_c \\ &\quad + \frac{1}{2}\tilde{\omega}_c^\top\kappa(x)\tilde{\omega}_c + \frac{1}{2}\tilde{\omega}_c^\top\nabla\sigma_c(x) M(x)\nabla\epsilon_c(x).\end{aligned}$$

By replacing $\tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x}$ with $\tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x}^*$ in the first term $(1/\alpha_c)\tilde{\omega}_c^\top\dot{\tilde{\omega}}_c$, the derivative of the Lyapunov candidate

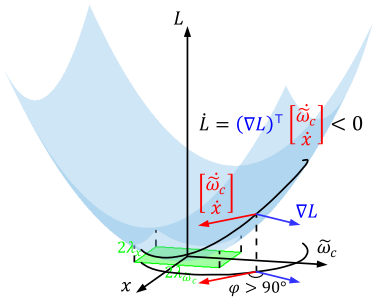


Fig. 11. Stability and convergence analysis.

becomes

$$\begin{aligned}\dot{L}\left(\tilde{\omega}_c, x\right) &= -\Bigg(\tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x}^* - \frac{1}{4}\tilde{\omega}_c^\top\xi(x)\tilde{\omega}_c + \frac{1}{2}\tilde{\omega}_c^\top\rho(x)\omega_c \\ &\qquad - \frac{1}{2}\tilde{\omega}_c^\top\nabla\sigma_c(x) M(x)\nabla\epsilon_c(x) + e_H\Bigg) \\ &\quad \times\Bigg(\tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x}^* - \frac{1}{2}\tilde{\omega}_c^\top\xi(x)\tilde{\omega}_c + \frac{1}{2}\tilde{\omega}_c^\top\rho(x)\omega_c \\ &\qquad - \frac{1}{2}\tilde{\omega}_c^\top\nabla\sigma_c(x) M(x)\nabla\epsilon_c(x)\Bigg) \\ &\quad + \frac{\alpha_s}{2\alpha_c}\Pi(x,\hat{u},\hat{w})\tilde{\omega}_c^\top\nabla\sigma_c(x) M(x)\nabla J_s(x) \\ &\quad + \frac{\alpha_s}{\alpha_c}\left(\nabla J_s(x)\right)^\top\dot{x}.\end{aligned} \tag{A.3}$$

Next, in accordance with the bounded conditions 1)–4) in Assumption 2, expanding all terms of the above formula and performing basic mathematical operations generates the following inequalities with respect to $\tilde{\omega}_c$ and $\nabla J_s(x)$:

$$\begin{aligned}\dot{L} &\le -\frac{1}{16}\left(\tilde{\omega}_c^\top\xi(x)\tilde{\omega}_c\right)^2 - \frac{1}{8}\left(\tilde{\omega}_c^\top\rho(x)\omega_c\right)^2 \\ &\quad + \frac{3}{8}\left(\tilde{\omega}_c^\top\kappa(x)\tilde{\omega}_c\right)\left(\tilde{\omega}_c^\top\rho(x)\omega_c\right) \\ &\quad - \frac{3}{8}\left(\tilde{\omega}_c^\top\vartheta(x)\tilde{\omega}_c\right)\left(\tilde{\omega}_c^\top\rho(x)\omega_c\right) \\ &\quad + \frac{3}{8}\left(\tilde{\omega}_c^\top\rho(x)\tilde{\omega}_c\right)\left(\tilde{\omega}_c^\top\rho(x)\omega_c\right) \\ &\quad + \left(\frac{3}{4}\tilde{\omega}_c^\top\left(\kappa(x) - \vartheta(x) + \rho(x)\right)\tilde{\omega}_c - \tilde{\omega}_c^\top\rho(x)\omega_c\right) \\ &\qquad \times\left(\tilde{\omega}_c^\top\nabla\sigma_c\dot{x}^*\right) \\ &\quad + \left(\frac{3}{8}\tilde{\omega}_c^\top\left(\kappa(x) - \vartheta(x) + \rho(x)\right)\tilde{\omega}_c - \frac{1}{2}\tilde{\omega}_c^\top\rho(x)\omega_c\right) \\ &\qquad \left(\tilde{\omega}_c^\top\vartheta_{\sigma\epsilon}(x)\right) \\ &\quad - \left(\frac{3}{8}\tilde{\omega}_c^\top\left(\kappa(x) - \vartheta(x) + \rho(x)\right)\tilde{\omega}_c - \frac{1}{2}\tilde{\omega}_c^\top\rho(x)\omega_c\right) \\ &\qquad \times\left(\tilde{\omega}_c^\top\kappa_{\sigma\epsilon}(x)\right) \\ &\quad + \frac{9}{8}\left(\tilde{\omega}_c^\top\vartheta_{\sigma\epsilon}(x)\right)^2 + \frac{9}{8}\left(\tilde{\omega}_c^\top\kappa_{\sigma\epsilon}(x)\right)^2 + \frac{\alpha_s}{\alpha_c}\left(\nabla J_s(x)\right)^\top\dot{x} \\ &\quad + 3e_H^2(x) + \frac{\alpha_s}{2\alpha_c}\Pi(x,\hat{u},\hat{w})\tilde{\omega}_c^\top\nabla\sigma_c(x) M(x)\nabla J_s(x) \\ &\le -\frac{1}{16}\left(\tilde{\omega}_c^\top\xi(x)\tilde{\omega}_c\right)^2 - \frac{1}{8}\left(\tilde{\omega}_c^\top\rho(x)\omega_c\right)^2\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{3}{16}\eta_1^2+\frac{3}{8\eta_4^2}+\frac{3}{16}\eta_8^2+\frac{3}{16\gamma^2}\eta_{12}^2\right)\left(\tilde{\omega}_c^\top\kappa(x)\tilde{\omega}_c\right)^2\\
&+\left(\frac{3}{16}\eta_2^2+\frac{3}{8\eta_5^2}+\frac{3}{16}\eta_9^2+\frac{3}{16\gamma^2}\eta_{13}^2\right)\left(\tilde{\omega}_c^\top\vartheta(x)\tilde{\omega}_c\right)^2\\
&+\left(\frac{3}{16}\eta_3^2+\frac{3}{8\eta_6^2}+\frac{3}{16}\eta_{10}^2+\frac{3}{16\gamma^2}\eta_{14}^2\right)\left(\tilde{\omega}_c^\top\rho(x)\tilde{\omega}_c\right)^2\\
&+\left(\frac{3}{16\eta_1^2}+\frac{3}{16\eta_2^2}+\frac{3}{16\eta_3^2}+\frac{1}{2\eta_7^2}+\frac{\eta_{11}^2}{4}+\frac{\eta_{15}^2}{4\gamma^2}\right)\\
&\quad\times\left(\tilde{\omega}_c^\top\rho(x)\omega_c\right)^2\\
&+\left(\frac{3}{8}\eta_4^2+\frac{3}{8}\eta_5^2+\frac{3}{8}\eta_6^2+\frac{1}{2}\eta_7^2\right)\left(\tilde{\omega}_c^\top\nabla\sigma_c(x)\dot{x}^*\right)^2\\
&+\left(\frac{9}{8}+\frac{3}{16\eta_8^2}+\frac{3}{16\eta_9^2}+\frac{3}{16\eta_{10}^2}+\frac{1}{4\eta_{11}^2}\right)\left(\tilde{\omega}_c^\top\vartheta_{\sigma\epsilon}(x)\right)^2\\
&+\left(\frac{9}{8}+\frac{3\gamma^2}{16\eta_{12}^2}+\frac{3\gamma^2}{16\eta_{13}^2}+\frac{3\gamma^2}{16\eta_{14}^2}+\frac{\gamma^2}{4\eta_{15}^2}\right)\\
&\quad\times\left(\tilde{\omega}_c^\top\kappa_{\sigma\epsilon}(x)\right)^2\\
&+3e_H^2(x)+\frac{\alpha_s}{\alpha_c}\left(\nabla J_s(x)\right)^\top\dot{x}\\
&+\frac{\alpha_s}{2\alpha_c}\Pi(x,\hat{u},\hat{w})\tilde{\omega}_c^\top\nabla\sigma_c(x)M(x)\nabla J_s(x)\\
\le&-\lambda_1\left\|\tilde{\omega}_c\right\|^4+\lambda_2\left\|\tilde{\omega}_c\right\|^2+\lambda_3^2+\frac{\alpha_s}{\alpha_c}\left(\nabla J_s(x)\right)^\top\dot{x}\\
&+\frac{\alpha_s}{2\alpha_c}\Pi(x,\hat{u},\hat{w})\left(\nabla J_s(x)\right)^\top M(x)\left(\nabla\sigma_c(x)\right)^\top\tilde{\omega}_c
\end{aligned}
$$

where

$$
\begin{aligned}
\lambda_1\triangleq&\frac{1}{16}\underline{\lambda}_\xi^2-\left(\frac{3}{16}\eta_1^2+\frac{3}{8\eta_4^2}+\frac{3}{16}\eta_8^2+\frac{3}{16\gamma^2}\eta_{12}^2\right)\bar{\lambda}_\kappa^2\\
&-\left(\frac{3}{16}\eta_2^2+\frac{3}{8\eta_5^2}+\frac{3}{16}\eta_9^2+\frac{3}{16\gamma^2}\eta_{13}^2\right)\bar{\lambda}_\vartheta^2\\
&-\left(\frac{3}{16}\eta_3^2+\frac{3}{8\eta_6^2}+\frac{3}{16}\eta_{10}^2+\frac{3}{16\gamma^2}\eta_{14}^2\right)\bar{\lambda}_\rho^2 \qquad \text{(A.4)}\\
\lambda_2\triangleq&-\frac{1}{8}\underline{\lambda}_\rho^2\lambda_\omega^2+\left(\frac{3}{8}\eta_4^2+\frac{3}{8}\eta_5^2+\frac{3}{8}\eta_6^2+\frac{1}{2}\eta_7^2\right)\lambda_{\sigma x}^2\\
&+\left(\frac{3}{16\eta_1^2}+\frac{3}{16\eta_2^2}+\frac{3}{16\eta_3^2}+\frac{1}{2\eta_7^2}+\frac{\eta_{11}^2}{4}+\frac{\eta_{15}^2}{4\gamma^2}\right)\\
&\times\bar{\lambda}_\rho^2\lambda_\omega^2+\left(\frac{9}{8}+\frac{3}{16\eta_8^2}+\frac{3}{16\eta_9^2}+\frac{3}{16\eta_{10}^2}+\frac{1}{4\eta_{11}^2}\right)\\
&\times\lambda_\sigma^2\lambda_g^4\left\|R^{-1}\right\|^2\lambda_\epsilon^2\\
&+\frac{1}{\gamma^2}\left(\frac{9}{8\gamma^2}+\frac{3}{16\eta_{12}^2}+\frac{3}{16\eta_{13}^2}+\frac{3}{16\eta_{14}^2}+\frac{1}{4\eta_{15}^2}\right)\\
&\times\lambda_\sigma^2\lambda_k^4\lambda_\epsilon^2\\
\lambda_3\triangleq&\sqrt{3}\lambda_e. \qquad \text{(A.5)}
\end{aligned}
$$

For the meaning of symbols such as $\lambda_g$, $\lambda_k$, $\lambda_\omega$, $\lambda_\sigma$, $\lambda_\epsilon$, $\lambda_e$, and $\lambda_{\sigma x}$, refer to bounded assumptions given in Assumption 2.

Note that $\lambda_1$, $\lambda_2$, and $\lambda_3$ are all positive that can be ensured by choosing constants $\{\eta_i\}_{i=1}^{15}$ appropriately.

After that, the two cases of the unstability indicator function will be discussed separately. For the first case, $\Pi(x,\hat{u},\hat{w})=1$ indicates that

$$
\begin{aligned}
\dot{L}\left(\tilde{\omega}_c,x\right)\le&-\lambda_1\left\|\tilde{\omega}_c\right\|^4+\lambda_2\left\|\tilde{\omega}_c\right\|^2+\lambda_3^2\\
&+\frac{\alpha_s}{\alpha_c}\left(\nabla J_s(x)\right)^\top\left(f(x)+\frac{1}{2}M(x)\left(\nabla\sigma_c(x)\right)^\top\omega_c\right)\\
\le&-\lambda_1\left\|\tilde{\omega}_c\right\|^4+\lambda_2\left\|\tilde{\omega}_c\right\|^2+\lambda_3^2\\
&-\frac{\alpha_s}{\alpha_c}\lambda_{\min}(\Phi)\left\|\nabla J_s(x)\right\|^2\\
&+\frac{\alpha_s}{2\alpha_c}\left(\lambda_g^2\left\|R^{-1}\right\|+\frac{1}{\gamma^2}\lambda_k^2\right)\lambda_\epsilon\left\|\nabla J_s(x)\right\|.
\end{aligned}
$$

Completing the squares about $\|\tilde{\omega}_c\|^2$ and $\|\nabla J_s(x)\|$ yields

$$
\begin{aligned}
\dot{L}\left(\tilde{\omega}_c,x\right)\le&-\lambda_1\left(\left\|\tilde{\omega}_c\right\|^2-\frac{\lambda_2}{2\lambda_1}\right)^2+\lambda_4\\
&-\lambda_5\left(\left\|\nabla J_s(x)\right\|-\frac{\left(\lambda_g^2\left\|R^{-1}\right\|+\frac{\lambda_k^2}{\gamma^2}\right)\lambda_\epsilon}{4\lambda_{\min}(\Phi)}\right)^2
\end{aligned}
$$

where

$$
\begin{aligned}
\lambda_4\triangleq&\lambda_3^2+\frac{\lambda_2^2}{4\lambda_1}+\frac{\alpha_s\lambda_\epsilon^2}{16\alpha_c\lambda_{\min}(\Phi)}\left(\lambda_g^2\left\|R^{-1}\right\|+\frac{\lambda_k^2}{\gamma^2}\right)^2\\
\lambda_5\triangleq&\frac{\alpha_s}{\alpha_c}\lambda_{\min}(\Phi). \qquad \text{(A.6)}
\end{aligned}
$$

Thus, if the following inequality

$$
\left\|\tilde{\omega}_c\right\|\ge\sqrt{\frac{\lambda_2}{2\lambda_1}+\sqrt{\frac{\lambda_4}{\lambda_1}}}\triangleq\lambda'_{\tilde{\omega}_c} \qquad \text{(A.7)}
$$

or

$$
\left\|\nabla J_s(x)\right\|\ge\frac{\left(\gamma^2\lambda_g^2\left\|R^{-1}\right\|+\lambda_k^2\right)\lambda_\epsilon}{4\gamma^2\lambda_{\min}(\Phi)}+\sqrt{\frac{\alpha_c\lambda_4}{\alpha_s\lambda_{\min}(\Phi)}}\triangleq\lambda'_x \qquad \text{(A.8)}
$$

holds, we can obtain that $\dot{L}(\tilde{\omega}_c,x)<0$.

For the other case, $\Pi(x,\hat{u},\hat{w})=0$ means that $\nabla J_s(x)^\top(f(x)+g(x)\hat{u}+k(x)\hat{w})<0$. Since there is a positive number $\lambda_6$ such that $\nabla J_s(x)^\top(f(x)+g(x)\hat{u}+k(x)\hat{w})\le-\lambda_6\|\nabla J_s(x)\|$, the derivative of the Lyapunov candidate has

$$
\begin{aligned}
\dot{L}\left(\tilde{\omega}_c,x\right)\le&-\lambda_1\left\|\tilde{\omega}_c\right\|^4+\lambda_2\left\|\tilde{\omega}_c\right\|^2+\lambda_3^2-\lambda_6\frac{\alpha_s}{\alpha_c}\left\|\nabla J_s(x)\right\|\\
=&-\lambda_1\left(\left\|\tilde{\omega}_c\right\|^2-\frac{\lambda_2}{2\lambda_1}\right)^2+\frac{\lambda_2^2}{4\lambda_1}+\lambda_3^2\\
&-\lambda_6\frac{\alpha_s}{\alpha_c}\left\|\nabla J_s(x)\right\|.
\end{aligned}
$$

Therefore, if the following inequality

$$
\left\|\tilde{\omega}_c\right\|\ge\sqrt{\frac{\lambda_2+\sqrt{\lambda_2^2+4\lambda_1\lambda_3^2}}{2\lambda_1}}\triangleq\lambda''_{\tilde{\omega}_c} \qquad \text{(A.9)}
$$

or

$$\|\nabla J_s(x)\| \geq \frac{\alpha_c\left(\lambda_2^2+4\lambda_1\lambda_3^2\right)}{4\alpha_s\lambda_1\lambda_6} \triangleq \lambda_x'' \tag{A.10}$$

holds, it is guaranteed that $\dot{L}(\tilde{\omega}_c, x) < 0$.

In summary, if the inequality $\|\tilde{\omega}_c\| \geq \max(\lambda'_{\tilde{\omega}_c}, \lambda''_{\tilde{\omega}_c}) \triangleq \lambda_{\tilde{\omega}_c}$ or $\|\nabla J_s(x)\| \geq \max(\lambda'_x, \lambda''_x) \triangleq \lambda_x$ holds, one has $\dot{L}(\tilde{\omega}_c, x) < 0$. As claimed by Assumption 1, $J_s(x)$ is radially unbounded. The boundedness of $\|\nabla J_s(x)\|$ implies the boundedness of the state, that is, $\|x\|$. The above inequalities guarantee that $\dot{L}(\tilde{\omega}_c, x) < 0$ outside a compact set as shown in Fig. 11. According to the Lyapunov extension theorem, we can conclude that the weight estimation error $\tilde{\omega}_c$ and the state $x$ of the closed-loop system with the estimated policies (34) are UUB with ultimate bounds $\lambda_{\tilde{\omega}_c}$ and $\lambda_x$. □

## REFERENCES


[1] K. Li, S. E. Li, F. Gao, Z. Lin, J. Li, and Q. Sun, "Robust distributed consensus control of uncertain multiagents interacted by eigenvalue-bounded topologies," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3790–3798, May 2020.

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

[3] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, "State-space solutions to standard $H_2$ and $H_\infty$ control problems," in *Proc. Amer. Control Conf.*, Jun. 1988, pp. 1691–1696.

[4] J. C. Doyle, "Structured uncertainty in control system design," in *Proc. 24th IEEE Conf. Decis. Control*, Dec. 1985, pp. 260–265.

[5] M. G. Safonov, K. C. Goh, and J. H. Ly, "Control system synthesis via bilinear matrix inequalities," in *Proc. Amer. Control Conf.*, 1994, pp. 45–49.

[6] A. van der Schaft, "$L_2$-gain analysis of nonlinear systems and nonlinear state feedback $H_\infty$ control," *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 770–784, Jun. 1992.

[7] T. Shen and K. Tamura, "Robust $H_\infty$ control of uncertain nonlinear system via state feedback," *IEEE Trans. Autom. Control*, vol. 40, no. 4, pp. 766–768, Apr. 1995.

[8] T. Basar and P. Bernhard, *$H_\infty$ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Berlin, Germany: Springer, 2008.

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

[10] 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.

[11] D. P. Bertsekas, "Value and policy iterations in optimal control and adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 500–509, Mar. 2017.

[12] S. E. Li, *Reinforcement Learning for Sequential Decision and Optimal Control*. Berlin, Germany: Springer, 2022.

[13] 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.

[14] J. Duan et al., "Relaxed actor-critic with convergence guarantees for continuous-time optimal control of nonlinear systems," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 5, pp. 3299–3311, May 2023.

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

[16] 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.

[17] J. Li et al., "Relaxed policy iteration algorithm for nonlinear zero-sum games with application to H-infinity control," *IEEE Trans. Autom. Control*, early access, Apr. 11, 2023, doi: 10.1109/TAC.2023.3266277.

[18] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Policy iterations on the Hamilton–Jacobi–Isaacs equation for $H_\infty$ state feedback control with input saturation," *IEEE Trans. Autom. Control*, vol. 51, no. 12, pp. 1989–1995, Dec. 2006.

[19] H.-N. Wu and B. Luo, "Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear $H_\infty$ control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1884–1895, Dec. 2012.

[20] H. Modares, F. L. Lewis, and Z.-P. Jiang, "$H_\infty$ 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.

[21] T. Dierks and S. Jagannathan, "Optimal control of affine nonlinear continuous-time systems using an online Hamilton–Jacobi–Isaacs formulation," in *Proc. 49th IEEE Conf. Decis. Control (CDC)*, Dec. 2010, pp. 3048–3053.

[22] H. Zhang, L. Cui, and Y. Luo, "Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 206–216, Feb. 2013.

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

[24] X. Yang, D. Liu, Q. Wei, and D. Wang, "Guaranteed cost neural tracking control for a class of uncertain nonlinear systems using adaptive dynamic programming," *Neurocomputing*, vol. 198, pp. 80–90, Jul. 2016.

[25] 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.

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

[27] X. Yang, H. He, Q. Wei, and B. Luo, "Reinforcement learning for robust adaptive control of partially unknown nonlinear systems subject to unmatched uncertainties," *Inf. Sci.*, vols. 463–464, pp. 307–322, Oct. 2018.

[28] 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.

[29] H.-N. Wu, M.-M. Li, and L. Guo, "Finite-horizon approximate optimal guaranteed cost control of uncertain nonlinear systems with application to Mars entry guidance," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1456–1467, Jul. 2015.

[30] Y. Huang, D. Wang, and D. Liu, "Bounded robust control design for uncertain nonlinear systems using single-network adaptive dynamic programming," *Neurocomputing*, vol. 266, pp. 128–140, Nov. 2017.

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

[32] D. Wang, H. He, and D. Liu, "Improving the critic learning for event-based nonlinear $H_\infty$ control design," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3417–3428, Oct. 2017.

[33] D. Wang and D. Liu, "Learning and guaranteed cost control with event-based adaptive critic implementation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6004–6014, Dec. 2018.

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

[35] J. Choi, R. Nagamune, and R. Horowitz, "Synthesis of multiple robust controllers for parametric uncertain LTI systems," in *Proc. Amer. Control Conf. (ACC)*, 2006, pp. 3629–3636.

[36] J. Doyle, J. A. Primbs, B. Shapiro, and V. Nevistic, "Nonlinear games: Examples and counterexamples," in *Proc. 35th IEEE Conf. Decis. Control*, 1996, pp. 3915–3920.

**Jie Li** received the B.S. degree in automotive engineering from Tsinghua University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Vehicle and Mobility.

He was a Visiting Student Researcher at the Department of Mechanical Engineering, The University of British Columbia, Vancouver, BC, Canada, in 2022. His current research interests include model predictive control, adaptive dynamic programming, and robust reinforcement learning.

**Ryozo Nagamune** (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from Osaka University, Suita, Japan, in 1995 and 1997, respectively, and the Ph.D. degree from the Royal Institute of Technology, Stockholm, Sweden, in 2002.

From 2003 to 2005, he was a Post-Doctoral Researcher at the University of California at Berkeley, Berkeley, CA, USA. In 2013, he was a Visiting Researcher with the National Renewable Energy Laboratory, National Wind Technology Center, Golden, CO, USA. He has been with the Department of Mechanical Engineering, The University of British Columbia, Vancouver, BC, Canada, since 2006, where he is currently a Professor. His research interests include robust control, floating offshore wind turbines, and farm control and control of solar thermal systems.

Dr. Nagamune is the Past Chair of the IEEE Joint Chapter of Control Systems, Robotics, and Automation, and Systems, Man, and Cybernetics Societies in the Vancouver Section. He will be the General Chair of the Tenth IEEE Conference on Control Technology and Applications to be held in Vancouver, in 2026.

**Yuhang Zhang** received the B.S. degree in automotive engineering from Tsinghua University, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Vehicle and Mobility.

His current research interests include decision-making and control of autonomous vehicles and safety issues of applying reinforcement learning in autonomous driving.

**Shengbo Eben Li** (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2006 and 2009, respectively.

He is currently a Professor in the interdisciplinary field of autonomous driving and artificial intelligence with Tsinghua University. Before joining Tsinghua University, he worked at Stanford University, Stanford, CA, USA; University of Michigan, Ann Arbor, MI, USA; and UC Berkeley, Berkeley, CA, USA. He has published over 130 peer-reviewed articles in top-tier international journals and conferences. His active research interests include intelligent vehicles and driver assistance, deep reinforcement learning, optimal control, and estimation.

Dr. Li was a recipient of the Best Paper Awards (Finalists) from IEEE International Conference on Intelligent Transportation Systems (ITSC), IEEE International Conference on Unmanned Systems (ICUS), IEEE Intelligent Vehicles Symposium (IV), and Annual Learning for Dynamics and Control Conference (L4DC). He has received a number of important academic honors, including the National Award for Technological Invention of China in 2013, the National Award for Progress in Science and Technology of China in 2018, the Distinguished Young Scholar of Beijing NSF in 2018, and the Youth Science and Technology Innovation Leader from MOST China in 2020. He also serves as the AEs for *IEEE Intelligent Transportation Systems Magazine* (ITSM) and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (ITS).