




# Model-free incremental adaptive dynamic programming based approximate robust optimal regulation

Cong Li<sup>1</sup>  | Yongchao Wang<sup>1</sup> | Fangzhou Liu<sup>1</sup>  | Qingchen Liu<sup>2</sup>  | Martin Buss<sup>1</sup>

<sup>1</sup>Chair of Automatic Control Engineering, Technical University of Munich, Munich, Germany

<sup>2</sup>Chair of Information-Oriented Control, Technical University of Munich, Munich, Germany

## Correspondence

Fangzhou Liu, Chair of Automatic Control Engineering, Technical University of Munich, Theresienstr. 90, 80333 Munich, Germany.  
Email: fangzhou.liu@tum.de

## Abstract

This article presents a new formulation for model-free robust optimal regulation of continuous-time nonlinear systems. The proposed reinforcement learning based approach, referred to as incremental adaptive dynamic programming (IADP), utilizes measured input-state data to allow the design of the approximate optimal incremental control strategy, stabilizing the controlled system incrementally under model uncertainties, environmental disturbances, and input saturation. By leveraging the time delay estimation (TDE) technique, we first use sensor data to reduce the requirement of a complete dynamics, where input-state data is adopted to construct an incremental dynamics which reflects the system evolution in an incremental form. Then, the resulting incremental dynamics serves to design the approximate optimal incremental control strategy based on adaptive dynamic programming, which is implemented as a simplified single critic structure to get the approximate solution to the value function of the Hamilton–Jacobi–Bellman equation. Furthermore, for the critic neural network, experience data are used to design an off-policy weight update law with guaranteed weight convergence. Rather importantly, we incorporate a TDE error bound related term into the cost function, whereby the unintentionally introduced TDE error is attenuated during the optimization process. The proofs of system stability and weight convergence are provided. Numerical simulations are conducted to validate the effectiveness and superiority of our proposed IADP, especially regarding the reduced control energy expenditure and the enhanced robustness.

## KEYWORDS

incremental adaptive dynamic programming, reinforcement learning, robust optimal regulation, time delay estimation

## 1 | INTRODUCTION

Reinforcement learning (RL) provides a mathematical formulation for learning-based control strategies and has shown superior performance in multiple scenarios, such as humanoid robotics,<sup>1</sup> unmanned aerial vehicles,<sup>2</sup> and autonomous driving.<sup>3</sup> Although the distinguishable model-free feature of RL overcomes the difficulty of applying traditional

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *International Journal of Robust and Nonlinear Control* published by John Wiley & Sons Ltd.

model-based control methods to the unknown (or hardly modeled) plants, the rigorous system stability analysis is not provided in most of the related works, see the works of Recht,<sup>4</sup> Buşoniu et al.,<sup>5</sup> and the references therein. A system without stability guarantee is potentially dangerous.<sup>6</sup> Recently, synchronous adaptive dynamic programming (ADP),<sup>7-9</sup> where actor and critic neural networks (NNs) update simultaneously in real-time, emerges as a promising control-theoretic RL subfield featured for available system stability proofs. However, its provided stability proof compromises the attractive model-free feature of RL since a mathematical form of dynamics is required to present the rigorous system stability analysis. Even though the required explicit knowledge of dynamics could be avoided by using add-on techniques such as NNs,<sup>10-12</sup> fuzzy models,<sup>13</sup> Gaussian process (GP),<sup>14</sup> or observers,<sup>15</sup> the accompanying identification processes further increase computational complexity and parameter tuning efforts. This motivates us to develop a novel, computationally simple RL-based control strategy, which exhibits both a model-free feature and a provable system stability guarantee, to accomplish the robust optimal stabilization of continuous-time nonlinear systems.

Many attempts have been conducted to enhance the robustness of synchronous ADP. To approximately solve the robust optimal stabilization problem of a completely known dynamics perturbed by an unknown but bounded additive disturbance, existing synchronous ADP related approaches are mainly divided into two categories: the  $H$ -infinity control method formulated as a zero-sum game,<sup>16</sup> and the transformed optimal control method through a well-designed utility function.<sup>17</sup> However, both methods require accurate model information to construct the corresponding Hamilton–Jacobi–Issac (HJI) or Hamilton–Jacobi–Bellman (HJB) equations. Moreover, the utilized worst-case disturbance related terms in cost functions usually result in conservative control policies that lead to reduced performance. In addition, the transformed optimal control method<sup>17</sup> requires the knowledge of the disturbance, for example, the disturbance bound. To obviate the requirement of an accurate drift dynamics, by using the defined integral reinforcement, integral RL is developed to allow the design of a partially model-free approximate optimal control strategy.<sup>18</sup> However, the complete knowledge of the input dynamics is still demanded. A further step to get rid of model information is to utilize NN based approximation schemes such as radial basis function neural networks (RBFNNs),<sup>11</sup> and recurrent neural networks (RNNs),<sup>12</sup> where the dynamics is approximated by a linear weighting of handpicked basis sets. Although the model-free control is achieved based on the universal approximation ability of NNs, it is not trivial to get a high-quality approximated model based on an additionally introduced weight update law. The control strategy designed based on inaccurate approximated models might lead to performance degradation or even instability. Moreover, the effectiveness of these plug-in methods<sup>10-12</sup> mentioned above highly relies on prior knowledge. For example, the center and width of each chosen radial basis function are determined a priori by considering the whole working space of the investigated system.<sup>11</sup> The aforementioned high reliance on prior experience also exist in the fuzzy model based work<sup>13</sup> to avoid using model information. In addition, GP<sup>14</sup> or observer<sup>15</sup> based methods are also widely used to deal with model uncertainties and/or environmental disturbances. Although efficient, these methods<sup>14,15</sup> suffer from high computation complexity and parameter tuning efforts. The counterpart to our mainly focused synchronous ADP is the so-called iterative ADP,<sup>19-21</sup> which sequentially updates actor and critic NNs (i.e., one NN is tuned, and the other holds constant). Although this method enjoys the model-free property for discrete-time systems, however, its extension to continuous-time systems entails challenges in proving system stability and ensuring that the algorithm is online and model-free.<sup>10</sup>

Among the aforementioned robust synchronous ADP related works, either complete<sup>16,17</sup> or partial model knowledge<sup>18</sup> is required. The desired model-free control is accomplished by introducing additional techniques, such as NNs,<sup>10-12</sup> fuzzy models,<sup>13</sup> GP,<sup>14</sup> or observers,<sup>15</sup> where the dynamics is required to be identified online explicitly. Unlike these computation-intensive approaches, time delay estimation (TDE)<sup>22,23</sup> is a fundamentally different mechanism to design model-free control strategies, where an incremental dynamics constructed by time-delayed signals is used to reflect the system evolution of the controlled plant incrementally without introducing any online identification processes. However, despite TDE's promising robustness feature and beneficial computation simplicity, the optimality property of TDE based methods remains to be investigated. Besides, the implementation of TDE unintentionally introduces the TDE error, which denotes the gap between the real system and the constructed incremental dynamics. Although the boundness property of this TDE error is analyzed in traditional TDE related works,<sup>22,23</sup> its influence on the controller performance is overlooked. A fundamental problem about addressing the TDE error has yet to be properly established. The idea of using sensor data to facilitate the model-free approximate optimal control strategy originates in works<sup>24-26</sup> where the Taylor series expansion based incremental control technique is used to reduce dependence on the explicit knowledge of dynamics. However, no system stability is presented in related works.<sup>24-26</sup> Besides, although identifying a global system model is avoided, a recursive least square method is still required to identify local system transition matrices, which introduces additional computation load.

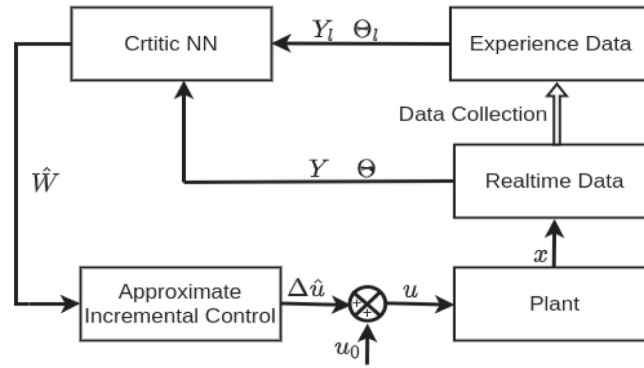


FIGURE 1 Schematic of incremental adaptive dynamic programming

This article proposes an alternative approach to achieve model-free control with guaranteed stability and optimality, as summarized in Figure 1. This is accomplished by first leveraging TDE to get an equivalent incremental dynamics (no explicit model knowledge but measured input-state data is used) to the investigated system. Thereby, we sidestep the online identification process, as well as its accompanying computation complexity and parameter tuning efforts. Then, the resulting incremental dynamics serves as a basis to allow the design of the model-free approximate optimal incremental control strategy. Furthermore, current and experience data are used to support the online NN weight learning of the critic agent. The contribution of this work is summarized as follows.

1. We develop a novel RL augmented control approach, which is called IADP, that enjoys both model-free feature and guaranteed closed-loop system stability. More importantly, IADP accomplishes a significant reduction in the control energy expenditure, which enables it to be favorable to energy-limited platforms.
2. Under an optimization framework, performance indexes regarding state deviations and control energy expenditures are considered. Thus, we endow TDE based methods with the optimality property. Besides, by incorporating a TDE error bound related term into the cost function, we novelly attenuate the TDE error during an optimization process.

The remainder of this article is organized as follows: problem formulation of the robust stabilization problem, problem transformation to the optimal incremental control problem, and problem equivalence proofs are provided in Section 2. Thereafter, we present the approximate optimal solution in Section 3. Numerical simulation results shown in Section 4 demonstrate the effectiveness and superiority of IADP. Finally, Section 5 concludes this work.

*Notations:* Throughout this article,  $\mathbb{R}$  ( $\mathbb{R}^+$ ) denotes the set of real (positive) numbers;  $\mathbb{R}^n$  is the Euclidean space of  $n$ -dimensional real vector;  $\mathbb{R}^{n \times m}$  is the Euclidean space of  $n \times m$  real matrices;  $I_{m \times m}$  represents the identity matrix with dimension  $m \times m$ ;  $\lambda_{\min}(M)$  and  $\lambda_{\max}(M)$  are the minimum and maximum eigenvalues of a symmetric matrix  $M$ , respectively;  $\text{diag}(a_1, \dots, a_n)$  is the  $n \times n$  diagonal matrix with the value of main diagonal as  $a_1, \dots, a_n$ ; The  $i$ th entry of a vector  $x = [x_1, \dots, x_n]^T \in \mathbb{R}^n$  is denoted by  $x_i$ , and  $\|x\| = \sqrt{\sum_{i=1}^n |x_i|^2}$  is the Euclidean norm of the vector  $x$ ; The  $ij$ th entry of a matrix  $D \in \mathbb{R}^{n \times m}$  is denoted by  $d_{ij}$ , and  $\|D\| = \sqrt{\sum_{i=1}^n \sum_{j=1}^m |d_{ij}|^2}$  is the Frobenius norm of the matrix  $D$ . For notational brevity, time-dependence is suppressed without causing ambiguity.

## 2 | PROBLEM FORMULATION

Considering the following continuous-time control-affine nonlinear system:

$$\dot{x} = f(x) + g(x)u(x) + d(t), \quad (1)$$

where  $x \in \mathbb{R}^n$ ,  $u(x) \in \mathbb{R}^m$  are system states and inputs, respectively.  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $g(x) : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$  are continuous and locally Lipschitz drift and input dynamics, respectively.  $d(t) \in \mathbb{R}^n$  represents a bounded time-varying external disturbance. Assume that no knowledge of dynamics (1) is available except for the dimensions of system states and inputs.

The main objective of this article is to tackle the robust stabilization problem of the highly uncertain dynamics (1) that operates in a disturbed environment, which is formulated as Problem 1.

**Problem 1.** Design a control strategy  $u(x)$  such that the system (1) perturbed by a bounded disturbance  $d(t)$  is stable under input saturation  $\mathbb{U}_j = \{u_j \in \mathbb{R} : |u_j| \leq \beta\}$ ,  $j = 1, \dots, m$ , where  $\beta \in \mathbb{R}^+$  is a known saturation bound.

*Remark 1.* Although the explicit form of the controlled plant (1) is provided here, which is introduced for the analytical purpose and facilitates the controller design as well as the stability analysis in the following sections, our developed control approach relies on neither model parameters nor environmental information.

## 2.1 | Incremental dynamics

The highly uncertain dynamics (1) cannot be directly used to design a controller to solve Problem 1. Therefore, based on measured input-state data, this section leverages the TDE technique to get an incremental dynamics that is an equivalent of (1). This formulated incremental dynamics reflects the system response of the controlled plant (1) incrementally without using explicit model parameters, or preceding identification procedures. Here, the attempt to relieve dependence on the accurate knowledge of dynamics departs from existing works where additional computation-intensive tools such as NNs,<sup>10-12</sup> fuzzy models,<sup>13</sup> GP,<sup>14</sup> or observers<sup>15</sup> are required to address model uncertainties and/or environmental disturbances. The constructed incremental dynamics in this section serves as a basis for the development of the desired model-free control strategy and the rigorous closed-loop system stability analysis in the following sections.

Before proceeding, the following assumption is provided to facilitate the formulation of an incremental dynamics.

**Assumption 1** (8). The input dynamics  $g = [g_1, g_2, \dots, g_m]$  is bounded, and its columns  $g_1, g_2, \dots, g_m \in \mathbb{R}^n$  are linearly independent. The function  $g^\dagger = (g^\top g)^{-1} g^\top : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$  is bounded and locally Lipschitz continuous.

*Remark 2.* Assumption 1 is common in ADP related works.<sup>8,27</sup> Here,  $g(x)$  is assumed to be full column rank such that its pseudo inverse  $g^\dagger$  could be expressed as a simple algebraic formula (the inverse of  $g^\top(x)g(x)$  exists). The introduced  $g^\dagger$  is used to extend the TDE method usually applied to the Euler–Lagrange equation<sup>22,23</sup> to the control-affine nonlinear system (1). Note that it is a common assumption that the input dynamics  $g$  is bounded. This property is widely observed in many physical systems, such as robot manipulator systems,<sup>28</sup> vehicle dynamics,<sup>29</sup> and aircraft models<sup>30</sup> fulfill such a property.

To get the incremental dynamics, we start with introducing a constant matrix  $\bar{g} \in \mathbb{R}^{n \times m}$  and multiply  $\bar{g}^\dagger$  on the dynamics (1),

$$\bar{g}^\dagger \dot{x} = \bar{g}^\dagger f(x) + \bar{g}^\dagger g(x)u(x) + \bar{g}^\dagger d(t) = H(x, \dot{x}) + u(x), \quad (2)$$

where  $H(x, \dot{x}) = (\bar{g}^\dagger - g^\dagger(x))\dot{x} + g^\dagger(x)f(x) + g^\dagger(x)d(t) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ . It is a lump term that embodies all the unknown model knowledge (i.e.,  $f(x)$ ,  $g(x)$ ) as well as external disturbances (i.e.,  $d(t)$ ).

Then, with a sufficiently high sampling rate, based on the TDE technique,<sup>22,23</sup> the unknown  $H(x, \dot{x})$  in (2) could be estimated by time-delayed signals as

$$\hat{H}(x, \dot{x}) := H(x_0, \dot{x}_0) = \bar{g}^\dagger \dot{x}_0 - u_0, \quad (3)$$

where  $x_0 = x(t-L)$ ,  $u_0 = u(x(t-L))$ .  $L \in \mathbb{R}^+$  is the delay time which is chosen as one or several sampling periods in practical digital implementations. Given that the smallest achievable  $L$  in digital devices is the sampling period,<sup>31</sup> thus we finally take the delay time  $L$  to be the same as the sampling period to get an accurate estimation of  $H(x, \dot{x})$  in (3). In other words,  $x_0$ ,  $u_0$  are the values of states and inputs at the previous sampling period.

Finally, by substituting (3) into (2), we get the incremental dynamics as

$$\Delta \dot{x} = \bar{g} \Delta u + \bar{g}^\xi \xi, \quad (4)$$

where  $\Delta \dot{x} = \dot{x} - \dot{x}_0 \in \mathbb{R}^n$  and  $\Delta u = u(x) - u_0 \in \mathbb{R}^m$  are incremental states and control inputs, respectively.  $\xi = H(x, \dot{x}) - \hat{H}(x, \dot{x}) \in \mathbb{R}^m$  denotes the so-called TDE error, which is proved to be bounded as given in Lemma 1. Here, with a predefined  $\bar{g}$ , the measured input-state data (i.e.,  $\dot{x}$ ,  $\dot{x}_0$ ,  $u$ , and  $u_0$ ) are adopted to reflect the system response in an incremental way without using model or environmental information.

*Remark 3.* The so-called sufficiently high sampling rate, which is a prerequisite for estimating the unknown  $H(x, \dot{x})$  by reusing past measured input-state data, can be chosen as the value that is larger than 30 times the system bandwidth.<sup>31,32</sup> In this setting, a digital control system can be regarded as a continuous system so that  $H(x, \dot{x})$  in (2) does not vary significantly during the sampling period. Thus, the TDE error  $\xi$  in (4) is sufficiently small.

*Remark 4.* The TDE technique, which is usually used in the robotic field,<sup>22,23</sup> is extended to the continuous-time control-affine nonlinear system (1) in this section. From a practical perspective, the applied TDE technique enables us to switch from the requirement of accurate mathematical models to sensor capabilities of providing accurate measurements of  $\Delta\dot{x}$  (constructed from  $\dot{x}$  and  $\dot{x}_0$ ) and  $\Delta u$  (constructed from  $u(x)$  and  $u_0$ ). Though the derived incremental dynamics (4) suffers a practical utility problem given that state derivatives, or even partial state variables are not directly measurable for certain cases, authors argue that state derivative estimation techniques,<sup>33,34</sup> numerical differential techniques,<sup>35</sup> or state observer<sup>36</sup> could help. These potential solutions mentioned above deviate from the main objective of this article and thus remain as future works.

However, although an equivalent of (1) is provided in (4) without using explicit knowledge of dynamics, the unknown TDE error  $\xi$  hinders us to directly utilize (4) to design controllers. Therefore, a method will be developed to address the TDE error  $\xi$  in the next section. Before proceeding, here we first provide the theoretical analysis about the boundness property of  $\xi$ , which facilitates the method to tackle the TDE error  $\xi$  under an optimization framework in Section 2.2.

**Lemma 1.** *Given a sufficiently high sampling rate,  $\exists \bar{\xi} \in \mathbb{R}^+$ , there holds  $\|\xi\| \leq \bar{\xi}$ .*

*Proof.* Combining (2) with (3), the TDE error follows

$$\begin{aligned}\xi &= H(x, \dot{x}) - \hat{H}(x, \dot{x}) = H(x, \dot{x}) - H(x_0, \dot{x}_0) \\ &= (\bar{g}^\dagger - g^\dagger(x))(\dot{x} - \dot{x}_0) + (g_0^\dagger - g^\dagger(x))\dot{x}_0 + g^\dagger(x)f(x) - g_0^\dagger f_0 + g^\dagger(x)d(t) - g_0^\dagger d_0 \\ &= (\bar{g}^\dagger - g^\dagger(x))\Delta\dot{x} + (g_0^\dagger - g^\dagger(x))\dot{x}_0 + g^\dagger(x)(f(x) - f_0) + (g^\dagger(x) - g_0^\dagger)f_0 + g^\dagger(x)(d(t) - d_0) + (g^\dagger(x) - g_0^\dagger)d_0.\end{aligned}\quad (5)$$

Besides, based on the system (1), we get

$$\begin{aligned}\Delta\dot{x} &= f(x) + g(x)u(x) + d(t) - f_0 - g_0u_0 - d_0 \\ &= g(x)\Delta u + (g(x) - g_0)u_0 + f(x) - f_0 + d(t) - d_0.\end{aligned}\quad (6)$$

Then, substituting (6) into (5) yields

$$\begin{aligned}\xi &= (\bar{g}^\dagger - g^\dagger(x))g(x)\Delta u + (\bar{g}^\dagger - g^\dagger(x))[(g(x) - g_0)u_0 + f(x) - f_0 + d(t) - d_0] + (g_0^\dagger - g^\dagger(x))\dot{x}_0 \\ &\quad + g^\dagger(x)(f(x) - f_0) + (g^\dagger(x) - g_0^\dagger)f_0 + g^\dagger(x)(d(t) - d_0) + (g^\dagger(x) - g_0^\dagger)d_0 \\ &= (\bar{g}^\dagger g(x) - I_{m \times m})\Delta u + \delta_1,\end{aligned}\quad (7)$$

where  $\delta_1 = \bar{g}^\dagger(g(x) - g_0)u_0 + \bar{g}^\dagger(f(x) - f_0) + \bar{g}^\dagger(d(t) - d_0)$ .

For a sufficiently high sampling rate, the gap between successive states is sufficiently small. Thus, it is reasonable to assume that there exists a positive constant  $\bar{\delta}_1 \in \mathbb{R}^+$  such that  $\|\delta_1\| \leq \bar{\delta}_1$ . In addition, the bounded control input  $u$  implies that  $\|\Delta u\| \leq 2\beta$  holds. By choosing a suitable  $\bar{g}$  such that  $\|\bar{g}^\dagger g(x) - I_{m \times m}\| \leq c$  establishes, then we get

$$\begin{aligned}\|\xi\| &\leq \|\bar{g}^\dagger g(x) - I_{m \times m}\| \|\Delta u\| + \|\delta_1\| \\ &\leq c \|\Delta u\| + \bar{\delta}_1 \leq 2\beta c + \bar{\delta}_1 = \bar{\xi}.\end{aligned}\quad (8)$$

This concludes the proof. ■

*Remark 5.* By using the Taylor series expansion based incremental control technique, previous works<sup>24-26,37,38</sup> attempt to provide the incremental dynamics by offering the first-order approximation of  $\dot{x}$  in the neighborhood of  $[x_0, u_0]$ . It follows

$$\begin{aligned}
\dot{x} &= f(x) + g(x)u(x) \\
&= f_0 + g_0u_0 + \left. \frac{\partial[f(x) + g(x)u(x)]}{\partial x} \right|_{x=x_0, u=u_0} (x - x_0) + \left. \frac{\partial[f(x) + g(x)u(x)]}{\partial u} \right|_{x=x_0, u=u_0} (u - u_0) + \mathcal{H.O.T.} \\
&\cong \dot{x}_0 + F[x_0, u_0]\Delta x + G[x_0, u_0]\Delta u,
\end{aligned}$$

where  $F[x_0, u_0] = [\partial(f(x) + g(x)u(x))/\partial x]|_{x=x_0, u=u_0} \in \mathbb{R}^{n \times n}$  is the system matrix, and  $G[x_0, u_0] = [\partial(f(x) + g(x)u(x))/\partial u]|_{x=x_0, u=u_0} \in \mathbb{R}^{n \times m}$  is the control effectiveness matrix. However, the approximation error resulting from the high order term  $\mathcal{H.O.T.}$  is directly omitted without considering its influence on the controller performance. Furthermore, a recursive least square method is demanded to search for suitable gain matrices  $F[x_0, u_0]$  and  $G[x_0, u_0]$  to construct the incremental dynamics.<sup>24-26</sup> This required online identification of  $F[x_0, u_0]$  and  $G[x_0, u_0]$  introduces additional computational burden.

## 2.2 | Problem transformation to optimal incremental control

To address the unknown TDE error in the incremental dynamics (4), here we attempt to investigate the original robust stabilization problem shown as Problem 1 from an optimal control perspective, whereby the TDE error could be reflected in the performance index and further be attenuated during the optimization process. This departs from existing TDE related works<sup>22-26,37,38</sup> that directly ignore the influence of the TDE error on the controller performance. Moreover, the effort to solve Problem 1 under an optimization framework enables us to take the desired performance indexes regarding state deviations and control energy expenditures into consideration. These considered performance indexes endow the resulting TDE based model-free control strategy with guaranteed optimality.

The TDE error  $\xi$  in (4) is unknown. Thus, the available incremental dynamics to design a controller to solve Problem 1 follows

$$\Delta \dot{x} = \bar{g}\Delta u. \quad (9)$$

To attenuate the TDE error  $\xi$  that is overlooked in (9), as well as to optimize the performance of states and control inputs, we consider the cost function of (9) as

$$V(x(t)) = \int_t^\infty r(x(\tau), \Delta u(\tau)) d\tau, \quad (10)$$

where  $r(x, \Delta u) = x^\top Qx + \mathcal{W}(u_0 + \Delta u) + \bar{\xi}_o^{-2} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^+$ . The common quadratic positive definite term  $x^\top Qx$  reflects users' preference for the controller performance concerning state deviations, where  $Q \in \mathbb{R}^{n \times n}$  is a positive definite matrix. The nonquadratic positive definite control penalty function  $\mathcal{W}(u_0 + \Delta u)$ , which relates to the measured  $u_0$  and to be designed  $\Delta u$ , is introduced to enforce the control limit on  $u(x)$  based on the bounded tanh function. The explicit form of this part follows<sup>39</sup>

$$\mathcal{W}(u_0 + \Delta u) = 2 \sum_{j=1}^m \int_0^{u_{0j} + \Delta u_j} \beta \tanh^{-1}(\vartheta_j / \beta) d\vartheta_j, \quad (11)$$

where  $\vartheta_j \in \mathbb{R}^m$ . Originally, we could incorporate the quadratic TDE error bound  $\bar{\xi}^{-2}$  into  $r(x, \Delta u)$  to attenuate the TDE error  $\xi$  during the optimization process. However, according to (8) of Lemma 1, the explicit value of  $\bar{\xi}$  is unknown. Thus, we seek for a bounded  $\bar{\xi}_o^{-2}$ , where  $\bar{\xi}_o = \bar{c} \|\Delta u\|$  and  $\bar{c} \in \mathbb{R}^+$  is chosen as illustrated in Theorem 1, to replace  $\bar{\xi}^{-2}$  to accomplish the same goal. It is worth noting that the designed utility function  $r(x, \Delta u)$  here enables us to perform the optimization of incremental control inputs. This achievable optimization of incremental control inputs enables IADP to enjoy control effort reductions, which will be verified in the simulation part. This is one distinguishing feature of our proposed IADP.

*Remark 6.* Note that there exist other options to address the TDE error  $\xi$ . For example, by treating the unknown TDE error  $\xi$  in (4) as a kind of disturbance, we can introduce the widely used disturbance-observer based methods<sup>40</sup> or sliding mode



control methods<sup>41</sup> to compensate the TDE error  $\xi$ . Comparing to these add-on methods, our strategy enjoys computation simplicity.

The aforementioned settings allow us to formulate an optimal incremental control problem presented as Problem 2, whose equivalence to Problem 1 will be later proved in Theorem 1.

**Problem 2.** Given Assumption 1 and Lemma 1, consider the incremental dynamics (9), find an incremental control strategy  $\Delta u$  to minimize the cost function defined as (10).

Before proceeding to formally solve Problem 2, by following Reference 39, Definition 1, where admissible controls are defined based on (1), here we define the set of incremental control inputs that are considered admissible for Problem 2 dealt with in this section. The admissible incremental control defined in Definition 1 facilitates the following derivation of the closed-form optimal incremental control strategy.

**Definition 1** (Admissible incremental control). An incremental control  $\Delta \mu(x)$  is defined to be admissible with respect to (10) on  $\Omega \subseteq \mathbb{R}^n$ , denoted by  $\Delta \mu(x) \in \Psi(\Omega)$ , if  $\Delta \mu(x)$  is continuous on  $\Omega$ ,  $\Delta \mu(0) = 0$ ,  $\Delta u(x) = \Delta \mu(x)$  stabilizes (9) on  $\Omega$ , and  $V(x)$  is finite  $\forall x \in \Omega$ .

For any admissible incremental control policies  $\Delta u \in \Psi(\Omega)$ , using Leibniz's rule<sup>42</sup> to differentiate  $V$  in (10) yields the following relation

$$0 = r(x, \Delta u) + \nabla V^T \dot{x} = r(x, \Delta u) + \nabla V^T (\Delta \dot{x} + \dot{x}_0) = r(x, \Delta u) + \nabla V^T (\bar{g} \Delta u + \dot{x}_0), \quad (12)$$

where the operator  $\nabla$  denotes the partial derivative with regard to  $x$ , that is,  $\partial(\cdot)/\partial x$ .

Define the Hamiltonian function as

$$H(x, \Delta u, \nabla V) = r(x, \Delta u) + \nabla V^T (\bar{g} \Delta u + \dot{x}_0). \quad (13)$$

Let  $V^*(x)$  be the optimal cost function defined as

$$V^*(x) = \min_{\Delta u \in \Psi(\Omega)} \int_t^\infty r(x(\tau), \Delta u(\tau)) d\tau. \quad (14)$$

Combining with (13),  $V^*(x)$  satisfies the HJB equation

$$0 = \min_{\Delta u \in \Psi(\Omega)} [H(x, \Delta u, \nabla V^*)]. \quad (15)$$

Assume that the minimum on the right side of (15) exists and is unique.<sup>7</sup> By using the stationary optimality condition, that is,  $\partial H(x, \Delta u, \nabla V^*)/\partial \Delta u = 0$ , we get the closed-form optimal incremental control strategy as

$$\Delta u^* = -\beta \tanh \left( \frac{1}{2\beta} \bar{g}^T \nabla V^* \right) - u_0. \quad (16)$$

Then, we could construct the corresponding optimal control strategy as

$$u^* = u_0 + \Delta u^* = -\beta \tanh \left( \frac{1}{2\beta} \bar{g}^T \nabla V^* \right). \quad (17)$$

Departing from traditional ADP related works<sup>7,8</sup> where the total optimal control input  $u^*$  is directly designed, here we first get the theoretically derived incremental optimal control strategy  $\Delta u^*$  in (16), and then construct  $u^*$  based on the measured  $u_0$  and the designed  $\Delta u^*$ . This difference lies in that Problem 2 is formulated based on the incremental dynamics (9) that relates to incremental states and control inputs.

*Remark 7.* Alternatively, we could replace  $\mathcal{W}(u_0 + \Delta u)$  in  $r(x, \Delta u)$  with  $\mathcal{W}(\Delta u) = 2 \sum_{j=1}^m \int_0^{\Delta u_j} \alpha \tanh^{-1}(\theta_j/\alpha) d\theta_j$ . This enforces the constraint satisfaction of the incremental control inputs, which is denoted as  $-\alpha \leq \Delta u_j \leq \alpha$ ,  $\alpha \in \mathbb{R}^+$ ,

$j = 1, \dots, m$ . By following the aforementioned derivation processes (14)–(17), the corresponding optimal incremental control follows  $\Delta u^* = -\alpha \tanh(\frac{1}{2\alpha} \bar{g}^\top \nabla V^*)$ . Then, the resulting optimal control is  $u^* = u_0 + \Delta u^*$ . However, in this case, the control limit on  $u(x)$  cannot be addressed. Given that input saturation is common in real life and violations of it might lead to serious consequences, we prefer to incorporate (11) into  $r(x, \Delta u)$  to enforce the control limit on  $u(x)$ .

To get  $\Delta u^*$  (16) and  $u^*$  (17),  $\nabla V^*$  remains to be determined. We defer the explicit method to acquire  $\nabla V^*$  in Section 3, and focus now on the equivalence proof to show that after solving Problem 2, the resulting  $u^*$  (17) constructed from the designed  $\Delta u^*$  (16) is the robust stabilization solution to Problem 1.

**Theorem 1.** *Given Assumption 1 and Lemma 1, consider the system described by (1), if there exists a scalar  $\bar{c} \in \mathbb{R}^+$  such that*

$$\bar{\xi} < \bar{c} \|\Delta u\|, \tag{18}$$

*the system (1) is robustly stabilized by the optimal control strategy (17) with the optimal incremental control strategy (16).*

*Proof.* Given that  $V^*(x = 0) = 0$ , and  $V^* > 0$  for  $\forall x \neq 0$ ,  $V^*$  defined in (14) could serve as a Lyapunov function candidate for the stability proof. Taking time derivative of  $V^*$  along the incremental dynamics (4), which is an equivalent of the original dynamics (1), we get

$$\dot{V}^* = \nabla V^{*\top}(\Delta \dot{x} + \dot{x}_0) = \nabla V^{*\top}(\bar{g}\Delta u^* + \bar{g}\xi + \dot{x}_0) = \nabla V^{*\top}(\bar{g}\Delta u^* + \dot{x}_0) + \nabla V^{*\top}\bar{g}\xi. \tag{19}$$

According to (15) and (16), the following equations hold:

$$\nabla V^{*\top}(\bar{g}\Delta u^* + \dot{x}_0) = -x^\top Qx - \mathcal{W}(u_0 + \Delta u^*) - \bar{\xi}_o^2, \quad \nabla V^{*\top}\bar{g} = -2\beta \tanh^{-1}\left(\frac{u_0 + \Delta u^*}{\beta}\right). \tag{20}$$

Substituting (20) into (19) reads

$$\dot{V}^* = -x^\top Qx - \mathcal{W}(u_0 + \Delta u^*) - \bar{\xi}_o^2 - 2\beta \tanh^{-1}\left(\frac{u_0 + \Delta u^*}{\beta}\right)\xi. \tag{21}$$

As for  $\mathcal{W}(u_0 + \Delta u^*)$  in (21), based on the explicit form in (11) and by setting  $\zeta_j = \tanh^{-1}(\vartheta_j/\beta)$ , it follows

$$\begin{aligned} \mathcal{W}(u_0 + \Delta u^*) &= 2\beta \sum_{j=1}^m \int_0^{u_{0j} + \Delta u_j^*} \tanh^{-1}(\vartheta_j/\beta) d\vartheta_j = 2\beta^2 \sum_{j=1}^m \int_0^{\tanh^{-1}\left(\frac{u_{0j} + \Delta u_j^*}{\beta}\right)} \zeta_j(1 - \tanh^2(\zeta_j)) d\zeta_j \\ &= \beta^2 \sum_{j=1}^m \left( \tanh^{-1}\left(\frac{u_{0j} + \Delta u_j^*}{\beta}\right) \right)^2 - \epsilon_u, \end{aligned} \tag{22}$$

where  $\epsilon_u = 2\beta^2 \sum_{j=1}^m \int_0^{\tanh^{-1}\left(\frac{u_{0j} + \Delta u_j^*}{\beta}\right)} \zeta_j \tanh^2(\zeta_j) d\zeta_j$ . Based on the integral mean-value theorem, there exists a series of  $\theta_j \in [0, \tanh^{-1}\left(\frac{u_{0j} + \Delta u_j^*}{\beta}\right)]$ ,  $j = 1, \dots, m$ , such that

$$\epsilon_u = 2\beta^2 \sum_{j=1}^m \tanh^{-1}\left(\frac{u_{0j} + \Delta u_j^*}{\beta}\right) \theta_j \tanh^2(\theta_j). \tag{23}$$

Based on (20) and the fact  $0 \leq \tanh^2(\theta_j) \leq 1$ , it follows

$$\epsilon_u \leq 2\beta^2 \sum_{j=1}^m \left(\frac{u_{0j} + \Delta u_j^*}{\beta}\right) \theta_j \leq 2\beta^2 \sum_{j=1}^m \left(\tanh^{-1}\left(\frac{u_{0j} + \Delta u_j^*}{\beta}\right)\right)^2 = \frac{1}{2} \nabla V^{*\top} \bar{g} \bar{g}^\top \nabla V^*. \tag{24}$$



The definition of admissible incremental control in Definition 1 concludes that  $V^*$  is finite. Additionally, there exists  $b_{\nabla V^*} \in \mathbb{R}^+$  such that  $\|\nabla V^*\| \leq b_{\nabla V^*}$ . Thus, we could rewrite (24) as

$$\epsilon_u \leq b_{\epsilon_u} = \frac{1}{2} \|\bar{g}\|^2 b_{\nabla V^*}^2. \quad (25)$$

Then, substituting (22) and (25) into (21) yields

$$\dot{V}^* \leq -x^\top Qx - (\bar{\xi}_o^2 - \|\xi\|^2) - [\beta \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) + \xi]^2 + b_{\epsilon_u}. \quad (26)$$

By choosing  $\bar{\xi}_o = \bar{c} \|\Delta u\|$ , and  $\bar{c}$  is chosen to satisfy  $\bar{c} \|\Delta u\| > \bar{\xi}$ , where  $\bar{\xi}$  is defined in (8), the following inequality holds

$$\dot{V}^* \leq -x^\top Qx + b_{\epsilon_u}. \quad (27)$$

Thus,  $\dot{V}^* < 0$  holds if  $-\lambda_{\min}(Q)\|x\|^2 + b_{\epsilon_u} < 0$ . Finally, it concludes that states converge to the residual set

$$\Omega_x = \{x \mid \|x\| \leq \sqrt{b_{\epsilon_u}/\lambda_{\min}(Q)}\}. \quad (28)$$

The aforementioned proof means that based on the optimal cost function (14), the derived optimal incremental control policy (16) of the system (9) robustly stabilizes the system (4). Given the equivalence between (1) and (4) clarified in Section 2.1, thus the optimal control input (17), which is constructed from the designed (16), robustly stabilizes the system (1). This concludes the proof. ■

We have proved in Theorem 1 that the optimal incremental control problem clarified in Problem 2 is equivalent to the robust stabilization problem shown as Problem 1. Thus, to stabilize the highly uncertain dynamics (1) operating in a disturbed environment, the following article devotes to solving Problem 2.

### 3 | APPROXIMATE OPTIMAL SOLUTION

To solve Problem 2, this section seeks for the approximate solution to the value function of the HJB equation (15) that is hard to solve directly. Departing from common ADP related works<sup>7,8</sup> using an actor-critic structure, we introduce a single critic structure here, which decreases the computational burden and simplifies the theoretical analysis. In addition, we observe that the adopted critic NN for approximating the value function is in essence a linear approximator. This enables us to transform the critic NN weight learning problem into a parameter identification problem. Then, by further using the collected experience data to provide the sufficient excitation required for the weight convergence, we design a simple yet efficient off-policy weight update law with guaranteed weight convergence. Our approach is favorable to practical applications comparing to common methods that often directly add external noises to control inputs to meet the persistence of excitation (PE) condition required for the weight convergence,<sup>7,43</sup> which results in undesirable oscillations and additional control efforts.

#### 3.1 | Value function approximation

Based on the Weierstrass high-order approximation theorem,<sup>44</sup> for  $x \in \Omega$  with  $\Omega \subset \mathbb{R}^n$  being a compact set, the optimal value function is approximated as<sup>7</sup>

$$V^*(x) = W^{*\top} \Phi(x) + \epsilon(x), \quad (29)$$

where  $W^* \in \mathbb{R}^N$  is a weighting matrix,  $\Phi(x) : \mathbb{R}^n \rightarrow \mathbb{R}^N$  represents the activation function, and  $\epsilon(x) \in \mathbb{R}$  denotes the approximation error. The corresponding partial derivative of  $V^*(x)$  follows

$$\nabla V^*(x) = \nabla \Phi^\top(x) W^* + \nabla \epsilon(x), \quad (30)$$

where  $\nabla \Phi(x) \in \mathbb{R}^{N \times n}$ ,  $\nabla \epsilon(x) \in \mathbb{R}^n$ . As  $N \rightarrow \infty$ , both  $\epsilon(x)$  and  $\nabla \epsilon(x)$  converge to zero uniformly. Without loss of generality, the following assumption is given, which is common in ADP related works.

**Assumption 2** (7). There exist constants  $b_\epsilon, b_{\epsilon_x}, b_\Phi, b_{\Phi_x} \in \mathbb{R}^+$  such that  $\|\epsilon(x)\| \leq b_\epsilon$ ,  $\|\nabla \epsilon(x)\| \leq b_{\epsilon_x}$ ,  $\|\Phi(x)\| \leq b_\Phi$ , and  $\|\nabla \Phi(x)\| \leq b_{\Phi_x}$ .

Considering a fixed incremental control input  $\Delta u$ , inserting (30) into (15) yields

$$W^{*\top} \nabla \Phi(\bar{g} \Delta u + \dot{x}_0) + r(x, \Delta u) = \epsilon_h, \quad (31)$$

where the residual error follows  $\epsilon_h = -\nabla \epsilon^\top(\bar{g} \Delta u + \dot{x}_0) \in \mathbb{R}$ . Assume that there exists  $b_{\epsilon_h} \in \mathbb{R}^+$  such that  $\|\epsilon_h\| \leq b_{\epsilon_h}$ . By focusing on the NN parameterized (31), we rewrite it into the following linear in parameter (LIP) form

$$\Theta = -W^{*\top} Y + \epsilon_h, \quad (32)$$

where  $\Theta = r(x, \Delta u) \in \mathbb{R}$ , and  $Y = \nabla \Phi(\bar{g} \Delta u + \dot{x}_0) \in \mathbb{R}^N$ . Given that  $\Theta$  and  $Y$  could be obtained from real-time data, this formulated LIP form enables the learning of  $W^*$  to be equivalent to a parameter identification problem of an LIP system from the perspective of adaptive control, wherein  $Y$  and  $W^*$  could be treated as the known regressor matrix and the unknown parameter vector to be determined, respectively. The applied novel transformation here allows us to design a simple weight update law with guaranteed weight convergence in the subsequent section.

### 3.2 | Off-policy weight update law

By denoting the estimate of the ideal critic weight  $W^*$  in (32) as  $\hat{W} \in \mathbb{R}^N$ , then we get

$$\hat{\Theta} = -\hat{W}^\top Y, \quad (33)$$

where  $\hat{\Theta} \in \mathbb{R}$  is the estimate of  $\Theta$ . Denoting the weight estimation error as  $\tilde{W} = \hat{W} - W^* \in \mathbb{R}^N$ , and subtracting (33) from (32), we get the approximation error  $\tilde{\Theta} \in \mathbb{R}$  as

$$\tilde{\Theta} = \Theta - \hat{\Theta} = \tilde{W}^\top Y + \epsilon_h. \quad (34)$$

To achieve  $\hat{W} \rightarrow W^*$ ,  $\hat{W}$  should be updated to minimize  $E = \frac{1}{2} \tilde{\Theta}^\top \tilde{\Theta}$ . Furthermore, to guarantee the weight convergence while minimizing  $E$ , experience data are used to provide the required sufficient excitation. Finally, a simple yet efficient off-policy weight update law of the critic agent is designed as

$$\dot{\hat{W}} = -\Gamma k_c Y \tilde{\Theta} - \sum_{l=1}^P \Gamma k_e Y_l \tilde{\Theta}_l, \quad (35)$$

where  $\tilde{\Theta} = \Theta + \hat{W}^\top Y$  according to (33) and (34), which is available based on measurable  $\Theta$  and  $Y$  defined in (32).  $\Gamma \in \mathbb{R}^{N \times N}$  is a constant positive definite gain matrix.  $k_c, k_e \in \mathbb{R}^+$  are constant gains to balance the relative importance between current and experience data to the online learning process. The regressor matrix  $Y_l \in \mathbb{R}^N$  and the approximation error  $\tilde{\Theta}_l \in \mathbb{R}$  denote the  $l$ th collected data of the corresponding experience buffers  $\mathfrak{B}$  and  $\mathfrak{C}$ , respectively.  $P \in \mathbb{R}^+$  is the volume of the experience buffers  $\mathfrak{B}$  and  $\mathfrak{C}$ , that is, the maximum number of recorded data points.

Before proceeding to the guaranteed weight convergence proof based on (35), we first clarify a rank condition about the experience buffer  $\mathfrak{B}$  in Assumption 3. This rank condition serves as a richness criterion of the recorded experience data and facilitates the guaranteed weight convergence analysis in Theorem 2.

**Assumption 3.** Given an experience buffer  $\mathfrak{B} = [Y_1, \dots, Y_P] \in \mathbb{R}^{N \times P}$ , there holds  $\text{rank}(\mathfrak{B}) = N$ .

Departing from the traditional PE condition,<sup>7,43</sup> the rank condition in Assumption 3 provides an online checkable index about the data richness required for the weight convergence, which is favorable to controller designers. Assumption 3 is not restrictive, which could be easily satisfied by sequentially reusing experience data.

Based on the collected sufficient rich experience data illustrated in Assumption 3, here we provide the guaranteed weight convergence proof based on the off-policy weight update law (35).

**Theorem 2.** *Given Assumption 3, the weight learning error  $\tilde{W}$  converges to a small neighborhood around zero.*

*Proof.* Consider the following candidate Lyapunov function

$$V_W = \frac{1}{2} \tilde{W}^\top \Gamma^{-1} \tilde{W}. \quad (36)$$

The time derivative of  $V_W$  follows

$$\dot{V}_W = \tilde{W}^\top \Gamma^{-1} (-\Gamma k_c Y \tilde{\Theta} - \sum_{l=1}^P \Gamma k_e Y_l \tilde{\Theta}_l) = -k_c \tilde{W}^\top Y \tilde{\Theta} - \tilde{W}^\top \sum_{l=1}^P k_e Y_l \tilde{\Theta}_l \leq -\tilde{W}^\top B \tilde{W} + \tilde{W}^\top \epsilon_{\tilde{W}}, \quad (37)$$

where  $B = \sum_{l=1}^P k_e Y_l Y_l^\top \in \mathbb{R}^{N \times N}$ , and  $\epsilon_{\tilde{W}} = -k_c Y \epsilon_h - \sum_{l=1}^P k_e Y_l \epsilon_{h_l} \in \mathbb{R}^N$ . The boundness of  $Y$  and  $\epsilon_h$  results in bounded  $\epsilon_{\tilde{W}}$ . Thus, there exists  $\bar{\epsilon}_{\tilde{W}} \in \mathbb{R}^+$  such that  $\|\epsilon_{\tilde{W}}\| \leq \bar{\epsilon}_{\tilde{W}}$ . According to Assumption 3,  $B$  is positive definite. Thus, (37) could be rewritten as

$$\dot{V}_W \leq -\|\tilde{W}\| (\lambda_{\min}(B) \|\tilde{W}\| - \bar{\epsilon}_{\tilde{W}}). \quad (38)$$

Therefore,  $\dot{V}_W < 0$  if  $\|\tilde{W}\| > \frac{\bar{\epsilon}_{\tilde{W}}}{\lambda_{\min}(B)}$ . Finally, it concludes that the weight estimation error of the critic NN will converge to the residual set

$$\Omega_{\tilde{W}} = \left\{ \|\tilde{W}\| \|\tilde{W}\| \leq \frac{\bar{\epsilon}_{\tilde{W}}}{\lambda_{\min}(B)} \right\}. \quad (39)$$

This completes the proof. ■

With a sufficiently large  $N$ ,  $\bar{\epsilon}_{\tilde{W}}$  converges to zero. Then, according to (38), we get  $\dot{V}_W \leq -\lambda_{\min}(B) \|\tilde{W}\|^2$ , that is,  $\tilde{W} \rightarrow 0$  exponentially as  $t \rightarrow \infty$ . Thus, it concludes that  $\hat{W}$  guarantees convergence to  $W^*$ .

Unlike common actor-critic structure based works,<sup>7,8</sup> the guaranteed weight convergence of  $\hat{W}$  to  $W^*$  in Theorem 2 permits us to adopt a simplified single critic structure, where the estimated critic NN weight  $\hat{W}$  could be directly used to construct the approximate optimal incremental control strategy. Therefore, based on the optimal incremental control strategy in (16), the approximate optimal incremental control strategy follows

$$\Delta \hat{u} = -\beta \tanh \left( \frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top \hat{W} \right) - u_0. \quad (40)$$

Accordingly, the approximate optimal control strategy applied at the plant (1) follows

$$\hat{u} = u_0 + \Delta \hat{u} = -\beta \tanh \left( \frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top \hat{W} \right). \quad (41)$$

*Remark 8.* From a practical perspective, our designed model-free approximate optimal incremental control strategy (40) only requires one manually tuned constant matrix  $\bar{g}$ . This feature of IADP decreases the required parameter tuning efforts comparing to existing identification based methods to fulfill model-free control strategies,<sup>10-15</sup> where multiple hyperparameters or gains need to be tuned.

Based on the off-policy weight update law (35), and the approximate optimal incremental control strategy (40) mentioned above, we provide the main conclusions of this article in Theorem 3.

**Theorem 3.** Consider the incremental dynamics (9), the off-policy weight update law of the critic NN in (35), and the approximate optimal incremental control policy (40). Given Assumptions 1–3, for a sufficiently large  $N$ , the approximate optimal incremental control policy (40) stabilizes the incremental dynamics (9), and the critic NN weight learning error  $\tilde{W}$  is uniformly ultimately bounded (UUB).

*Proof.* Consider the following candidate Lyapunov function

$$J = V^*(x) + \frac{1}{2} \tilde{W}^\top \Gamma^{-1} \tilde{W}. \quad (42)$$

By denoting  $\dot{L}_V = \dot{V}^*(x)$  and  $\dot{L}_W = \tilde{W}^\top \Gamma^{-1} \dot{\tilde{W}}$ , the time derivative of (42) reads

$$\dot{J} = \dot{L}_V + \dot{L}_W. \quad (43)$$

The first term  $\dot{L}_V$  follows

$$\dot{L}_V = \nabla V^{*\top} (\bar{g} \Delta \hat{u} + \bar{g} \xi + \dot{x}_0) = \nabla V^{*\top} (\bar{g} \Delta u^* + \dot{x}_0) + \nabla V^{*\top} \bar{g} \xi + \nabla V^{*\top} \bar{g} (\Delta \hat{u} - \Delta u^*). \quad (44)$$

Then, substituting (20) into (44) gets

$$\dot{L}_V = -x^\top Qx - \mathcal{W}(u_0 + \Delta u^*) - \bar{\xi}_o^2 - 2\beta \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) \xi - 2\beta \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) (\Delta \hat{u} - \Delta u^*). \quad (45)$$

According to (22)–(24), (45) follows

$$\dot{L}_V \leq -x^\top Qx - (\bar{\xi}_o^2 - \|\xi\|^2) - [\beta \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) + \xi]^2 + \frac{1}{2} \nabla V^{*\top} \bar{g} \bar{g}^\top \nabla V^* - 2\beta \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) (\Delta \hat{u} - \Delta u^*). \quad (46)$$

The term  $-2\beta \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) (\Delta \hat{u} - \Delta u^*)$  in (46) follows

$$-2\beta \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) (\Delta \hat{u} - \Delta u^*) \leq \beta^2 \left\| \tanh^{-1} \left( \frac{u_0 + \Delta u^*}{\beta} \right) \right\|^2 + \|\Delta \hat{u} - \Delta u^*\|^2. \quad (47)$$

By using (16), (30), and the mean-value theorem, the optimal incremental control is rewritten as

$$\Delta u^* = -\beta \tanh \left( \frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top W^* \right) - \epsilon_{\Delta u^*} - u_0, \quad (48)$$

where  $\epsilon_{\Delta u^*} = \frac{1}{2} (\mathbf{1} - \tanh^2(\eta)) \bar{g}^\top \nabla \epsilon$ , and  $\eta \in \mathbb{R}^m$  is chosen between  $\frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top W^*$  and  $\frac{1}{2\beta} \bar{g}^\top \nabla V^*$ ,  $\mathbf{1} = [1, \dots, 1]^\top \in \mathbb{R}^m$ . According to  $\|\nabla \epsilon\| \leq b_{ex}$  in Assumption 2,  $\|\epsilon_{\Delta u^*}\| \leq \frac{1}{2} \|\bar{g}\| b_{ex}$  holds. Then, by combining (40) with (48), we get

$$\Delta \hat{u} - \Delta u^* = \beta \left( \tanh \left( \frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top W^* \right) - \tanh \left( \frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top \hat{W} \right) \right) + \epsilon_{\Delta u^*}. \quad (49)$$

For simplicity, denoting  $\mathcal{E}^* = \frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top W^*$  and  $\hat{\mathcal{E}} = \frac{1}{2\beta} \bar{g}^\top \nabla \Phi^\top \hat{W}$ , where  $\hat{\mathcal{E}} = [\hat{\mathcal{E}}_1, \dots, \hat{\mathcal{E}}_m] \in \mathbb{R}^m$  with  $\hat{\mathcal{E}}_j \in \mathbb{R}, j = 1, \dots, m$ . Based on (16) and (40), the Taylor series of  $\tanh(\mathcal{E}^*)$  follows

$$\tanh(\mathcal{E}^*) = \tanh(\hat{\mathcal{E}}) + \frac{\partial \tanh(\hat{\mathcal{E}})}{\partial \hat{\mathcal{E}}} (\mathcal{E}^* - \hat{\mathcal{E}}) + O((\mathcal{E}^* - \hat{\mathcal{E}})^2) = \tanh(\hat{\mathcal{E}}) - \frac{1}{2\beta} (I_{m \times m} - \mathcal{D}(\hat{\mathcal{E}})) \bar{g}^\top \nabla \Phi^\top \tilde{W} + O((\mathcal{E}^* - \hat{\mathcal{E}})^2), \quad (50)$$

where  $\mathcal{D}(\hat{\mathcal{E}}) = \text{diag}(\tanh^2(\hat{\mathcal{E}}_1), \dots, \tanh^2(\hat{\mathcal{E}}_m))$ , and  $O((\mathcal{E}^* - \hat{\mathcal{E}})^2)$  is a higher order term of the Taylor series. By following [45, Lemma 1], this higher order term is bounded as

$$\|O((\mathcal{E}^* - \hat{\mathcal{E}})^2)\| \leq 2\sqrt{m} + \frac{1}{\beta} \|\bar{g}\| b_{\Phi_x} \|\tilde{W}\|. \quad (51)$$

Based on (50), we rewrite (49) as

$$\Delta \hat{u} - \Delta u^* = \beta(\tanh(\mathcal{E}^*) - \tanh(\hat{\mathcal{E}})) + \epsilon_{\Delta u^*} = -\frac{1}{2}(I_{m \times m} - \mathcal{D}(\hat{\mathcal{E}}))\bar{g}\nabla\Phi^T\tilde{W} + \beta O((\mathcal{E}^* - \hat{\mathcal{E}})^2) + \epsilon_{\Delta u^*}. \quad (52)$$

According to Reference 45,  $\|I_{m \times m} - \mathcal{D}(\hat{\mathcal{E}})\| \leq 2$  holds. Then, by combining (51) with (52),  $\|\Delta \hat{u} - \Delta u^*\|^2$  in (47) follows

$$\begin{aligned} \|\Delta \hat{u} - \Delta u^*\|^2 &\leq 3\beta^2 \|O((\mathcal{E}^* - \hat{\mathcal{E}})^2)\|^2 + 3\|\epsilon_{\Delta u^*}\|^2 + 3\left\|-\frac{1}{2}(I_{m \times m} - \mathcal{D}(\hat{\mathcal{E}}))\bar{g}\nabla\Phi^T\tilde{W}\right\|^2 \\ &\leq 6\|\bar{g}\|^2 b_{\Phi_x}^2 \|\tilde{W}\|^2 + 12m\beta^2 + \frac{3}{4}\|\bar{g}\|^2 b_{\epsilon_x}^2 + 12\beta\sqrt{m}\|\bar{g}\| b_{\Phi_x} \|\tilde{W}\|. \end{aligned} \quad (53)$$

Based on (20), (30), Assumption 2, and the fact that  $\|W^*\| \leq b_{W^*}$ ,  $\|\tanh^{-1}((u_0 + \Delta u^*)/\beta)\|^2$  in (47) follows

$$\left\|\tanh^{-1}\left(\frac{u_0 + \Delta u^*}{\beta}\right)\right\|^2 = \left\|\frac{1}{4\beta^2}\nabla V^{*\top}\bar{g}\bar{g}^T\nabla V^*\right\| \leq \frac{1}{4\beta^2}\|\bar{g}\|^2 b_{\Phi_x}^2 b_{W^*}^2 + \frac{1}{4\beta^2}b_{\epsilon_x}^2\|\bar{g}\|^2 + \frac{1}{2\beta^2}\|\bar{g}\|^2 b_{\Phi_x} b_{\epsilon_x} b_{W^*}. \quad (54)$$

Using (53) and (54), (47) reads

$$\begin{aligned} -2\beta\tanh^{-1}\left(\frac{u_0 + \Delta u^*}{\beta}\right)(\Delta \hat{u} - \Delta u^*) &\leq \frac{1}{4}\|\bar{g}\|^2 b_{\Phi_x}^2 b_{W^*}^2 + \frac{1}{4}b_{\epsilon_x}^2\|\bar{g}\|^2 + \frac{1}{2}\|\bar{g}\|^2 b_{\Phi_x} b_{\epsilon_x} b_{W^*} \\ &\quad + 6\|\bar{g}\|^2 b_{\Phi_x}^2 \|\tilde{W}\|^2 + 12m\beta^2 + \frac{3}{4}\|\bar{g}\|^2 b_{\epsilon_x}^2 + 12\beta\sqrt{m}\|\bar{g}\| b_{\Phi_x} \|\tilde{W}\|. \end{aligned} \quad (55)$$

Substituting (55) into (46), finally the first term  $\dot{L}_V$  follows

$$\begin{aligned} \dot{L}_V &\leq -x^T Qx - (\bar{\xi}_o^2 - \xi^T \xi) - [\beta\tanh^{-1}\left(\frac{u_0 + \Delta u^*}{\beta}\right) + \xi]^2 + \frac{3}{4}\|\bar{g}\|^2 b_{\Phi_x}^2 b_{W^*}^2 + \frac{3}{4}b_{\epsilon_x}^2\|\bar{g}\|^2 + \frac{3}{2}\|\bar{g}\|^2 b_{\Phi_x} b_{\epsilon_x} b_{W^*} \\ &\quad + 6\|\bar{g}\|^2 b_{\Phi_x}^2 \|\tilde{W}\|^2 + 12m\beta^2 + \frac{3}{4}\|\bar{g}\|^2 b_{\epsilon_x}^2 + 12\beta\sqrt{m}\|\bar{g}\| b_{\Phi_x} \|\tilde{W}\|. \end{aligned} \quad (56)$$

As for the second term  $\dot{L}_W$ , based on (35) and (37), it follows

$$\dot{L}_W \leq -\tilde{W}^T B \tilde{W} + \tilde{W}^T \epsilon_{\tilde{W}}. \quad (57)$$

Finally, as for  $\dot{J}$ , substituting (56) and (57) into (43), we get

$$\dot{J} \leq -\mathcal{A} - \mathcal{B}\|\tilde{W}\|^2 + \mathcal{C}\|\tilde{W}\| + \mathcal{D}, \quad (58)$$

where  $\mathcal{A} = x^T Qx + (\bar{\xi}_o^2 - \xi^T \xi) + [\beta\tanh^{-1}\left(\frac{u_0 + \Delta u^*}{\beta}\right) + \xi]^2$ ,  $\mathcal{B} = \lambda_{\min}(B) - 6\|\bar{g}\|^2 b_{\Phi_x}^2$ ,  $\mathcal{C} = 12\beta\sqrt{m}\|\bar{g}\| b_{\Phi_x} + \bar{\epsilon}_{\tilde{W}}$ , and  $\mathcal{D} = \frac{3}{4}\|\bar{g}\|^2 b_{\Phi_x}^2 b_{W^*}^2 + \frac{3}{2}b_{\epsilon_x}^2\|\bar{g}\|^2 + \frac{3}{2}\|\bar{g}\|^2 b_{\Phi_x} b_{\epsilon_x} b_{W^*} + 12m\beta^2$ . Let the parameters be chosen such that  $\mathcal{B} > 0$ . Since  $\mathcal{A}$  is positive definite, the above Lyapunov derivative (58) is negative if

$$\|\tilde{W}\| > \frac{\mathcal{C}}{2\mathcal{B}} + \sqrt{\frac{\mathcal{C}^2}{4\mathcal{B}^2} + \frac{\mathcal{D}}{\mathcal{B}}}. \quad (59)$$

Thus, the critic weight learning error converges to the residual set

$$\tilde{\Omega}_{\tilde{W}} = \left\{ \tilde{W} \mid \|\tilde{W}\| \leq \frac{C}{2B} + \sqrt{\frac{C^2}{4B^2} + \frac{D}{B}} \right\}. \quad (60)$$

This completes the proof.  $\blacksquare$

## 4 | NUMERICAL SIMULATION

This section conducts multiple comparative numerical simulations to validate the effectiveness and superiority of our proposed IADP, especially in terms of the reduced control energy expenditure shown in Section 4.1, and the enhanced robustness illustrated in Section 4.2. Besides, the influence of different sampling rates on IADP's performance is investigated in Section 4.3. Here, we choose the widely investigated pendulum in ADP related works<sup>19,46</sup> as a benchmark. The dynamics of the pendulum follows

$$\begin{cases} \frac{d\theta}{dt} = \vartheta + d, \\ J \frac{d\vartheta}{dt} = u - Mgl \sin \theta - f_d \frac{d\theta}{dt}, \end{cases} \quad (61)$$

where  $\theta, \vartheta \in \mathbb{R}$  denote the angle and the angular velocity of the pendulum, respectively.  $M = 1/3$  kg and  $l = 3/2$  m are the mass and length of the pendulum, respectively. Let  $g = 9.8$  m s<sup>-2</sup> be the gravity,  $J = 4/3Ml^2$  kg m<sup>2</sup> be the rotary inertia, and  $f_d = 0.2$  be the frictional factor. Here  $d$  represents an external disturbance.

### 4.1 | Validation of the reduced control effort of IADP

This section compares IADP with the zero-sum game based ADP (ZSADP)<sup>16</sup> and the transformed optimal control based ADP (TADP)<sup>17</sup> to verify the superiority of IADP regarding the reduced control effort. Note that among existing ADP related works, model-based ZSADP and TADP are the two most widely adopted methods to deal with the robust stabilization problem illustrated as Problem 1. First, to conduct convincing comparative simulations, an often used vanishing (state-dependent) disturbance in ZSADP and TADP related works<sup>16,17</sup> is deliberately chosen in Section 4.1.1 to fully show the performance of ZSADP and TADP. Then, in Section 4.1.2, except for the vanishing disturbance used in Section 4.1.1, we make a step further by additionally introducing measurement noises, non-vanishing disturbances, and sudden physical changes into the simulation environment. The conducted comparative simulations under multiple sources of uncertainties and disturbances further exemplify the advantage of our proposed IADP in terms of the reduced control effort. The numerical simulations in this subsection are conducted under a sampling rate chosen as 1000 Hz.

#### 4.1.1 | Validation under the vanishing (state-dependent) disturbance

In this subsection, by following Reference 47, the state-dependent disturbance is chosen as  $d_1 = \omega_1 \theta \sin(\omega_2 \vartheta)$ , where  $\omega_1$  and  $\omega_2$  are randomly generated within the scope  $[-\sqrt{2}/2, \sqrt{2}/2]$  and  $[-2, 2]$ , respectively. Let  $x_1 = \theta$  and  $x_2 = \vartheta$ , the original pendulum system (61) is rewritten as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = f(x) \begin{bmatrix} x_2 \\ -4.9 \sin x_1 - 0.2x_2 \end{bmatrix} + g(x) \begin{bmatrix} 0 \\ 0.25 \end{bmatrix} u + k(x) \begin{bmatrix} 1 \\ -0.2 \end{bmatrix} d_1(x) \omega_1 x_1 \sin(\omega_2 x_2). \quad (62)$$

To drive the pendulum (62) to the equilibrium point even under input saturation ( $\beta = 2$ ) and the external disturbance  $d_1(x)$ , the detailed simulation settings for IADP, ZSADP, and TADP are as follows.

For IADP, we choose  $\bar{g} = [0, 0.1]^T$ . Its cost function is considered as

$$V_I = \int_t^\infty x^T Q x + \mathcal{W}(u_0 + \Delta u) + \bar{\xi}_o^2 d\tau, \quad (63)$$



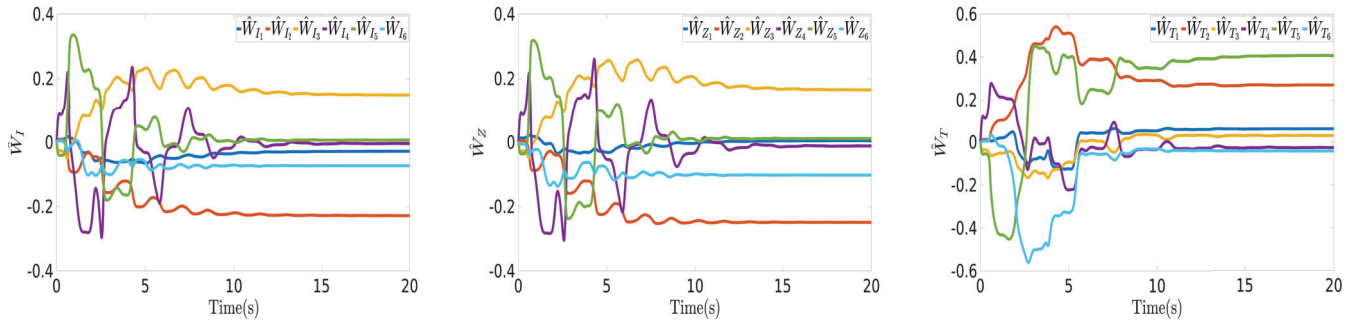


FIGURE 2 The estimated weight trajectories of IADP ( $\hat{W}_I$ ), ZSADP ( $\hat{W}_Z$ ), and TADP ( $\hat{W}_T$ ) under the disturbance  $d_1(x)$

where  $Q = I_{2 \times 2}$ ,  $\mathcal{W}(u_0 + \Delta u) = 2\beta(u_0 + \Delta u)\tanh^{-1}((u_0 + \Delta u)/\beta) + \beta^2 \log(1 - (u_0 + \Delta u)^2/\beta^2)$ , and  $\bar{\xi}_o = 2 \|\Delta u\|$ . The approximate optimal incremental control  $\Delta \hat{u}$  and the approximate optimal control  $\hat{u}$  follow (40) and (41), respectively. IADP requires neither explicit model nor environmental information except for a predefined constant matrix  $\bar{g}$ .

For ZSADP, by following Reference 16, its cost function is

$$V_Z = \int_t^\infty x^\top Qx + \mathcal{W}(u_Z) - \gamma d_Z^\top d_Z d\tau, \quad (64)$$

where  $\mathcal{W}(u_Z) = 2\beta u_Z \tanh^{-1}(u_Z/\beta) + \beta^2 \log(1 - u_Z^2/\beta^2)$ ,  $\gamma = 1$ . For this case, the approximate optimal control policy follows  $\hat{u}_Z = -\beta \tanh(\frac{1}{2\beta} g^\top \nabla \Phi^\top \hat{W}_Z)$ , and the approximate worst-case disturbance policy is  $\hat{d}_Z = \frac{1}{2\gamma^2} k^\top \nabla \Phi^\top \hat{W}_Z$ . Here  $\hat{u}_Z$  and  $\hat{d}_Z$  depend on the concert  $g(x)$  and  $k(x)$  in (62), respectively.

For TADP, according to Reference 17, the corresponding cost function follows

$$V_T = \int_t^\infty x^\top Qx + \mathcal{W}(u_T) + \rho v_T^\top v_T + l_M^2 + d_M^2 d\tau, \quad (65)$$

where  $\rho = 0.1$ . The chosen disturbance satisfies  $\|d(x)\| \leq \sqrt{2}/2 \|x\|$ . Thus,  $d_M = \sqrt{2}/2 \|x\|$  and  $l_M = 0.4\sqrt{2} \|x\|$  are chosen to address the disturbance  $d_1(x)$ . Much more details are referred to the work.<sup>17</sup> The approximate optimal control follows  $\hat{u}_T = -\beta \tanh(\frac{1}{2\beta} g^\top \nabla \Phi^\top \hat{W}_T)$ , and the approximate pseudo control follows  $\hat{v}_T = -\frac{1}{2\rho} h^\top \nabla \Phi^\top \hat{W}_T$ , where  $h = (I_{2 \times 2} - gg^\dagger)k$ . For TADP, the explicit knowledge of  $g(x)$  and  $k(x)$  in (62) is required to construct  $\hat{u}_T$  and  $\hat{v}_T$ .

The aforementioned IADP, ZSADP, and TADP all adopt the single critic structure and our developed off-policy weight update law (35). To achieve a fair comparison, simulation parameters for three methods are set as same, which is detailedly clarified as follows. To get the approximate solutions to the above value functions (63)–(65),  $\Phi(x) = [x_1^2, x_1x_2, x_2^2, x_2^3, x_1x_2^2, x_1^2x_2]^\top$  is chosen. To guarantee the weight convergence, parameters are set as  $P = 8$ ,  $\Gamma = 10^{-4}I_{6 \times 6}$ ,  $k_c = 5$ , and  $k_e = 3$ . The initial values are chosen as  $x(0) = [2, -2]^\top$ ,  $\hat{u}(0) = 0$ ,  $\hat{d}_Z(0) = 0$  (for ZSADP), and  $\hat{v}_T(0) = 0$  (for TADP). Note that to achieve a fair comparison, we also fix the values of  $\omega_1$  and  $\omega_2$  in  $d_1(x)$  as a set of randomly selected values:  $\omega_1 = -0.3906$ ,  $\omega_2 = 1.0051$ .

The critic NN weigh convergence results for IADP, ZSADP, and TADP are displayed in Figure 2. Based on our developed off-policy weight update law (35), the weight convergence is guaranteed without adding external noises to control inputs to achieve the required sufficient exploration.

The state and control trajectories of three cases are shown in Figure 3, where the pendulum is successfully driven to the equilibrium point without violating input constraints. However, regarding the peak points of state and control trajectories, the fluctuation range of IADP is smaller than ZSADP and TADP.

To reveal the superiority of IADP over ZSADP and TADP, we display their corresponding control energy expenditure  $E_u = \int_0^\infty \|\hat{u}\|^2 d\tau$ , and state deviation  $E_x = \int_0^\infty \|x\|^2 d\tau$  in Figure 4. It is observed in Figure 4 that IADP enjoys a noticeable reduction in utilized control effort, that is, energy efficiency is highly improved. This makes IADP a more suitable choice for energy-limited platforms. This significant decrease in the control effort comes from the achievable optimization of the incremental control inputs. Specifically, given  $\hat{u}(0) = 0$  and IADP prefers a small  $\Delta \hat{u}$  at each optimization step, a small  $\hat{u}$  is generated to stabilize the pendulum. Thus, we finally get a small  $E_u$ , which is a cumulative sum of quadratic

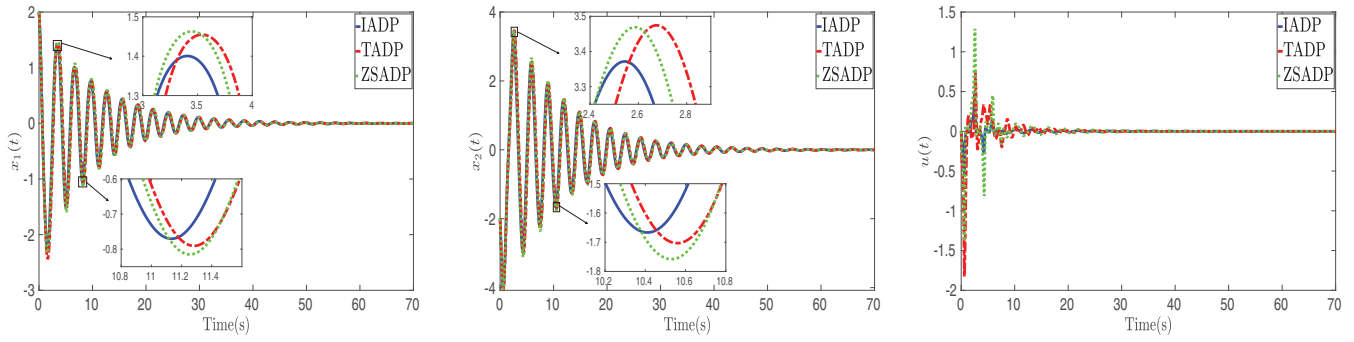


FIGURE 3 The state and control trajectories of IADP, ZSADP, and TADP under the vanishing disturbance  $d_1(x)$

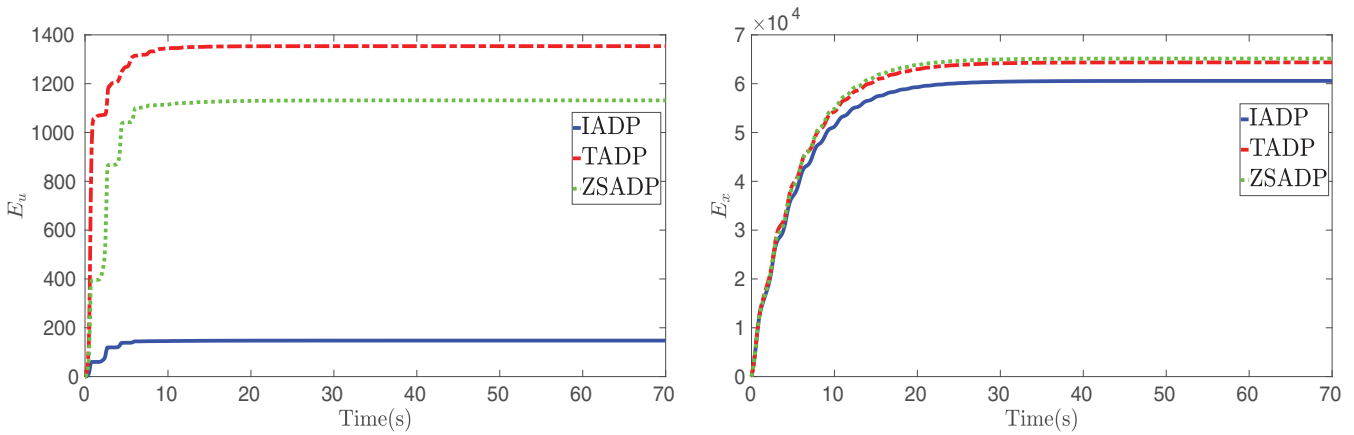


FIGURE 4 The performance comparison between IADP, ZSADP, and TADP under the vanishing disturbance  $d_1(x)$

$\hat{u}$ . The performance analysis shown in Figure 4 also clarifies the conservativeness of ZSADP and TADP. Although the worst-case disturbance related terms (i.e.,  $d_Z^T d_Z$  for ZSADP,  $v_T^T v_T$ ,  $l_M^2$ , and  $d_M^2$  for TADP) incorporated into the cost functions (64)–(65) allow controller designers to address the additive disturbance  $d_1(x)$ , these additionally introduced terms trade off the desired performance indexes of control efforts and state deviations. Thus, a performance compromise problem arises.

Given the aforementioned simulation results, we know that, even though model-based ZSADP and TADP could provide the rigorous robustness guarantee under worst-case disturbances, they perform poorly than our proposed model-free IADP, especially regarding the control energy expenditure.

#### 4.1.2 | Validation under the non-vanishing disturbance, measurement noise, and physical change

To further validate the superiority of our proposed IADP over ZSADP and TADP, this section conducts comparative simulations under the vanishing disturbance used in Section 4.1.1, as well as the newly introduced non-vanishing disturbance, measurement noise, and sudden physical change. It is worth noting that ZSADP<sup>16</sup> and TADP<sup>17</sup> can only deal with state-dependent disturbances in a closed-loop form. Thus, here the amplitudes of the chosen non-vanishing disturbance, measurement noise, and physical change are purposely set to be the level that could be tackled by the inherent robustness of ZSADP and TADP.

Here we follow the time-varying non-vanishing disturbance from Reference 48, which is denoted as  $d_2(t)$  and set as a square wave with amplitude 0.2 and period 5 s. The measurement noise is chosen as a white Gaussian noise with 50 dBW. We add  $d_2(t)$  and the measurement noise into the simulation environment during the time from 20 to 60 s. To simulate a sudden physical change, for example, parameter perturbations due to unknown loads put on the pendulum, at  $t = 20$  s, the dynamics of pendulum (62) is randomly reset as

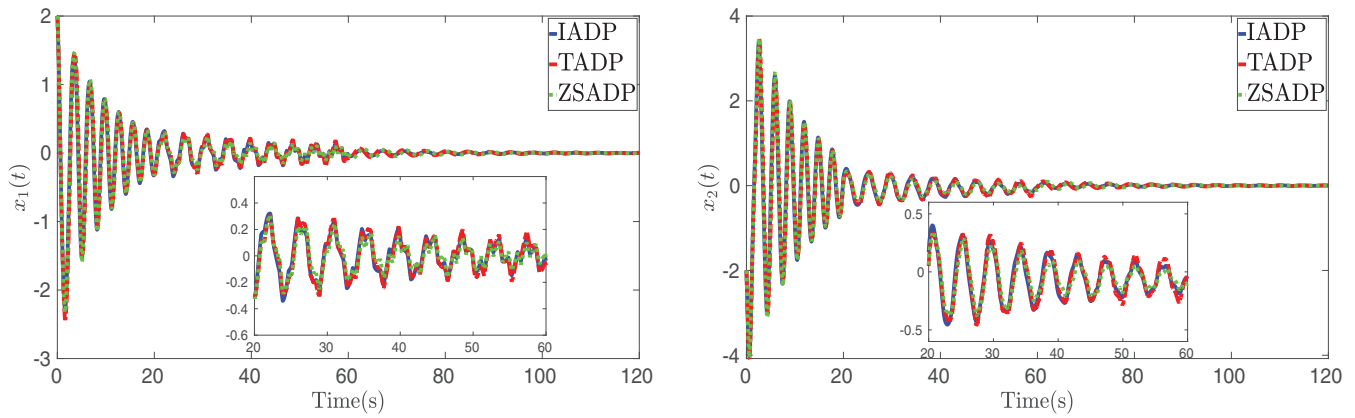


FIGURE 5 The state trajectories of IADP, ZSADP, and TADP under the non-vanishing disturbance  $d_2(t)$

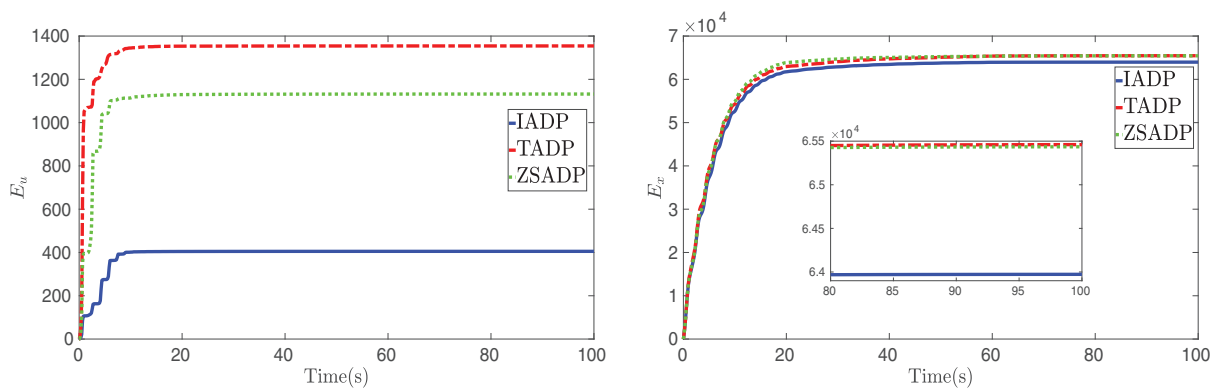


FIGURE 6 The performance comparison between IADP, ZSADP, and TADP under the non-vanishing disturbance  $d_2(t)$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = f(x) \begin{bmatrix} x_2 \\ -2 \sin x_1 - 0.1x_2 \end{bmatrix} + gx \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} u + kx \begin{bmatrix} 1 \\ -0.1 \end{bmatrix} d_1(x). \quad (66)$$

The parameter settings for IADP, ZSADP, and TADP are the same as the settings in Section 4.1.1.

Under the vanishing disturbance  $d_1(x)$ , the non-vanishing square wave disturbance  $d_2(t)$ , the white Gaussian measurement noise, and the sudden physical change from (62) to (66), the simulation results are shown in Figures 5 and 6. The state trajectories displayed in Figure 5 reveal that three methods all successfully stabilize the pendulum without retuning parameters, that is, these three methods possess inherent robustness to the aforementioned uncertainties and disturbances in certain amplitudes.

The performance comparison shown in Figure 6 further validates the significant control energy deduction of our developed IADP. Comparing to Figure 4 in Section 4.1.1, the increased control effort of IADP results from the required additional control energy to deal with the newly introduced uncertainties and disturbances. Besides, Figure 6 also displays that IADP outperforms ZSADP and TADP in terms of the state deviation  $E_x$ .

## 4.2 | Validation of the enhanced robustness of IADP

To highlight the enhanced robustness of our proposed IADP, this section conducts numerical simulations under a more complex simulation environment comparing to Section 4.1.2. The details are as follows: during the time from 20 to 60 s, the added non-vanishing disturbance  $d_3(t)$  is a square wave with amplitude 0.5 and period 1 s, whose amplitude and frequency are both improved comparing to  $d_2(t)$  used in Section 4.1.2; the incorporated measurement noise is set as a

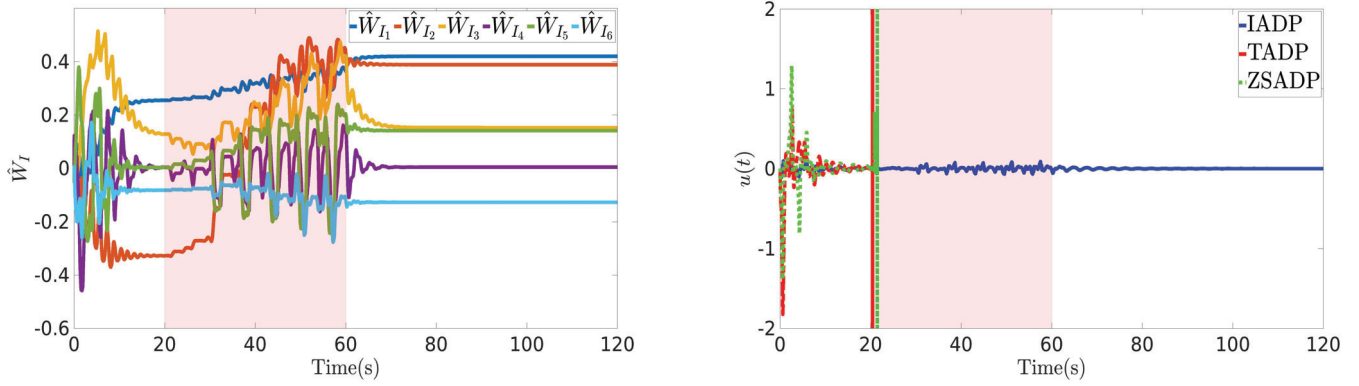


FIGURE 7 The estimated weight trajectory of IADP and the control trajectories of IADP, ZSADP, and TADP

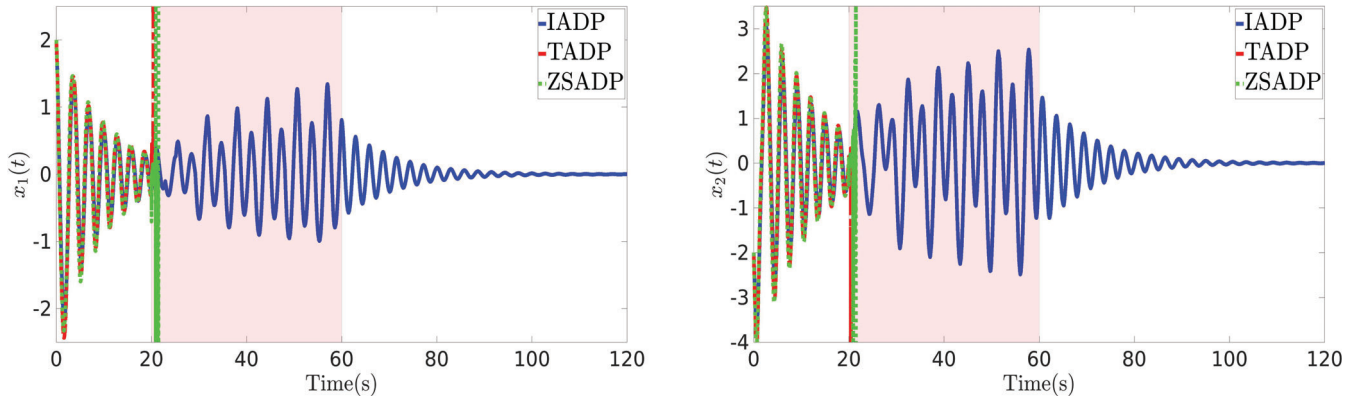


FIGURE 8 The state trajectories of IADP, ZSADP, and TADP under a complex simulation environment

white Gaussian noise with 10 dBW, whose magnitude is 5 times larger than the one chosen in Section 4.1.2. Besides, to model a significant physical change, at  $t = 20$  s, the pendulum (62) is reset as

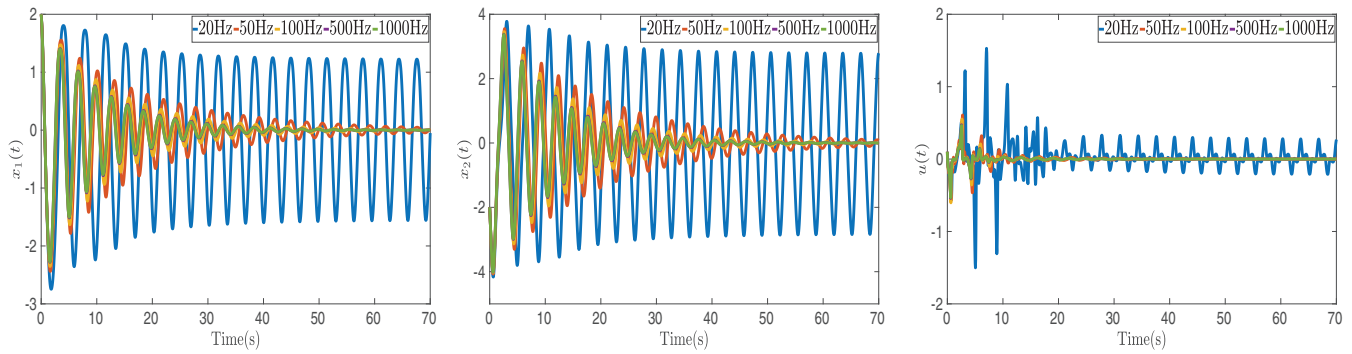
$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = f(x) \begin{bmatrix} -x_2 \\ 4.9 \sin x_1 - 0.2x_2 \end{bmatrix} + g(x) \begin{bmatrix} 0 \\ -0.25 \end{bmatrix} u + kx \begin{bmatrix} 1 \\ -0.2 \end{bmatrix} d_1(x). \quad (67)$$

Comparing to Section 4.1.2, the simulated physical change here is more aggressive by inverting the sign of model parameters. The parameter settings for IADP, ZSADP, and TADP follow the settings in Section 4.1.1. The sampling rate is chosen as 1000 Hz.

The estimated weight trajectory of IADP shown in Figure 7 illustrates that under multiple sources of uncertainties and disturbances, our proposed off-policy weight update law (35) enables us to collect real-time data in time and finally achieve weight convergence. The control trajectories shown in Figure 7, and the state trajectories displayed in Figure 8 clarify the enhanced robustness of IADP. Specifically, IADP successfully stabilizes the pendulum under multiple sources of uncertainties and disturbances, however, the robustness of ZSADP and TADP are not enough to tackle such a complex environment. Thus, the control inputs and states of ZSADP and TADP diverge far away immediately when the simulation environment significantly changes at  $t = 20$ s.

### 4.3 | Validation of the performance of IADP under different sampling rates

This subsection conducts multiple comparative numerical simulations to investigate the influence of different sampling rates on IADP's performance. Since we directly chooses the delay time  $L$  as the sampling period, this subsection also



**FIGURE 9** The state and control trajectories of IADP under different sampling rates

investigates the effect of different delay time  $L$  on the controller performance. Note that except for different values of the sampling rate, the conducted simulations in this subsection follow the same simulating environment and parameter settings as Section 4.1.1.

The evolution trajectories of states  $x_1$ ,  $x_2$  and control input  $u$  under different sampling rates are displayed in Figure 9. It is shown that a higher sampling rate leads to better performance. Specifically, for the considered robust optimal regulation control task, the sampling frequency of 50 Hz is enough to achieve satisfying performance. However, a system working under a higher sampling rate is more sensitive to measurement noises, requires faster converters and more storage, and consumes more computing resources. Thus, in practical applications, practitioners need to be aware of the trade-offs mentioned above and choose a suitable sampling rate accordingly.

## 5 | CONCLUSION

This article presents an efficient and low-cost model-free control strategy for robust optimal stabilization of continuous-time nonlinear control-affine systems. To reduce dependence on accurate mathematical models, the TDE technique permits us to obtain a measured input-state data based incremental dynamics, which is an equivalent of the original dynamics, without requiring explicit model knowledge or tedious identification procedures. Then, the HJB equation, which is constructed based on the incremental dynamics, is approximately solved through a single critic structure. The resulting approximate optimal incremental control strategy stabilizes the controlled plant incrementally. Besides, by transforming the critic NN weight learning as a parameter identification process and further using the collected experience data, we develop an efficient weight update law with guaranteed weight convergence. The following properties of our proposed IADP are promising for practical applications: the simultaneous consideration of stability, optimality and robustness, the utilized simplified single critic structure, and the easily implemented off-policy weight update law. Multiple conducted numerical simulations have shown that IADP outperforms common ADP methods in terms of reduced control efforts and enhanced robustness. The proposed IADP builds on the assumption that the full internal states and their derivatives are available, which restricts IADP's generality and practicality. Thus, future works attempt to combine state observer and state derivative estimation techniques with IADP to address the scenario when internal states and their derivatives are not measurable. In addition, since the efficacy of IADP depends on accurate sensor measurements, we will investigate and address the influence of sensor biases or delays on our developed IADP.

### CONFLICT OF INTEREST


Authors have no conflict of interest.

### DATA AVAILABILITY STATEMENT

Research data are not shared.

### ORCID

Cong Li  <https://orcid.org/0000-0002-1103-4818>

Fangzhou Liu  <https://orcid.org/0000-0002-1275-4809>

Qingchen Liu  <https://orcid.org/0000-0002-5892-3591>



## REFERENCES

1. Kormushev P, Calinon S, Caldwell DG. Reinforcement learning in robotics: applications and real-world challenges. *Robotics*. 2013;2(3):122-148.
2. Koch W, Mancuso R, West R, Bestavros A. Reinforcement learning for UAV attitude control. *ACM Trans Cyber-Phys Syst*. 2019;3(2):1-21.
3. Kuutti S, Bowden R, Jin Y, Barber P, Fallah S. A survey of deep learning applications to autonomous vehicle control. *IEEE Trans Intell Transp Syst*. 2020;22(2):712-733.
4. Recht B. A tour of reinforcement learning: the view from continuous control. *Annu Rev Control Robot Auton Syst*. 2019;2:253-279.
5. Buşoniu L, de Bruin T, Tolić D, Kober J, Palunko I. Reinforcement learning for control: performance, stability, and deep approximators. *Annu Rev Control*. 2018;46:8-28.
6. Khalil HK, Grizzle JW. *Nonlinear Systems*. Vol 3. Prentice Hall; 2002.
7. Vamvoudakis KG, Lewis FL. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*. 2010;46(5):878-888.
8. Kamalapurkar R, Dinh H, Bhasin S, Dixon WE. Approximate optimal trajectory tracking for continuous-time nonlinear systems. *Automatica*. 2015;51:40-48.
9. Vamvoudakis KG, Vrabie D, Lewis FL. Online adaptive algorithm for optimal control with integral reinforcement learning. *Int J Robust Nonlinear Control*. 2014;24(17):2686-2710.
10. Bhasin S, Kamalapurkar R, Johnson M, Vamvoudakis KG, Lewis FL, Dixon WE. A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*. 2013;49(1):82-92.
11. Li Y, Liu Y, Tong S. Observer-based neuro-adaptive optimized control of strict-feedback nonlinear systems with state constraints. *IEEE Trans Neural Netw Learn Syst*. 2021.
12. Zhang H, Cui L, Zhang X, Luo Y. Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Trans Neural Netw*. 2011;22(12):2226-2236.
13. Tong S, Sun K, Sui S. Observer-based adaptive fuzzy decentralized optimal control design for strict-feedback nonlinear large-scale systems. *IEEE Trans Fuzzy Syst*. 2017;26(2):569-584.
14. Boedecker J, Springenberg JT, Wülfing J, Riedmiller M. Approximate real-time optimal control based on sparse Gaussian process models. Proceedings of the 2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL); 2014:1-8.
15. Sun J, Liu C. Disturbance observer-based robust missile autopilot design with full-state constraints via adaptive dynamic programming. *J Franklin Inst*. 2018;355(5):2344-2368.
16. Vamvoudakis KG, Lewis FL. Online solution of nonlinear two-player zero-sum games using synchronous policy iteration. *Int J Robust Nonlinear Control*. 2012;22(13):1460-1483.
17. Liu D, Yang X, Wang D, Wei Q. Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints. *IEEE Trans Cybern*. 2015;45(7):1372-1385.
18. Vrabie D, Lewis F. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Netw*. 2009;22(3):237-246.
19. Liu D, Wei Q. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Trans Neural Netw Learn Syst*. 2013;25(3):621-634.
20. Sokolov Y, Kozma R, Werbos LD, Werbos PJ. Complete stability analysis of a heuristic approximate dynamic programming control design. *Automatica*. 2015;59:9-18.
21. Al-Dabooni S, Wunsch DC. Online model-free N-step HDP with stability analysis. *IEEE Trans Neural Netw Learn Syst*. 2019;31(4):1255-1269.
22. Hsia TC, Gao L. Robot manipulator control using decentralized linear time-invariant time-delayed joint controllers. Proceedings of the IEEE International Conference on Robotics and Automation; 1990:2070-2075.
23. Youcef-Toumi K, Wu ST. Input/output linearization using time delay control. *J Dyn Syst Meas Control*. 1992;114(1):10-19.
24. Zhou Y, Van Kampen EJ, Chu QP. Nonlinear adaptive flight control using incremental approximate dynamic programming and output feedback. *J Guid Control Dyn*. 2016;40(2):493-496.
25. Zhou Y, Van Kampen EJ, Chu QP. Incremental model based online dual heuristic programming for nonlinear adaptive control. *Control Eng Pract*. 2018;73:13-25.
26. Zhou Y, Van Kampen EJ, Chu QP. Incremental model based online heuristic dynamic programming for nonlinear adaptive tracking control with partial observability. *Aerosp Sci Technol*. 2020;105:106013.
27. Kiumarsi B, Lewis FL. Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems. *IEEE Trans Neural Netw Learn Syst*. 2014;26(1):140-151.
28. Li C, Liu F, Wang Y, Buss M. Concurrent learning-based adaptive control of an uncertain robot manipulator with guaranteed safety and performance. *IEEE Trans Syst Man Cybern Syst*. 2021.
29. Formentin S, Garatti S, Rallo G, Savaresi SM. Robust direct data-driven controller tuning with an application to vehicle stability control. *Int J Robust Nonlinear Control*. 2018;28(12):3752-3765.
30. Sastry S. *Nonlinear Systems: Analysis, Stability, and Control*. Vol 10. Springer Science & Business Media; 2013.
31. Jin M, Lee J, Chang PH, Choi C. Practical nonsingular terminal sliding-mode control of robot manipulators for high-accuracy tracking control. *IEEE Trans Ind Electron*. 2009;56(9):3593-3601.
32. Franklin GF, Powell JD, Workman ML. *Digital Control of Dynamic Systems*. Vol 3. Addison-Wesley; 1998.



33. Bhasin S, Kamalapurkar R, Dinh HT, Dixon WE. Robust identification-based state derivative estimation for nonlinear systems. *IEEE Trans Automat Contr*. 2012;58(1):187-192.
34. Levant A. Robust exact differentiation via sliding mode technique. *Automatica*. 1998;34(3):379-384.
35. Chang PH, Jung JH. A systematic method for gain selection of robust PID control for nonlinear plants of second-order controller canonical form. *IEEE Trans Control Syst Technol*. 2008;17(2):473-483.
36. Wang W, Gao Z. A comparison study of advanced state observer design techniques. Proceedings of the 2003 American Control Conference; Vol. 6, 2003:4754-4759.
37. Acquatella P, Van Kampen EJ, Chu QP. Incremental backstepping for robust nonlinear flight control. Proceedings of the EuroGNC 2013; 2013.
38. Semplicio P, Pavel M, Van Kampen EJ, Chu QP. An acceleration measurements-based approach for helicopter nonlinear flight control using incremental nonlinear dynamic inversion. *Control Eng Pract*. 2013;21(8):1065-1077.
39. Abu-Khalaf M, Lewis FL. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*. 2005;41(5):779-791.
40. Chen WH, Yang J, Guo L, Li S. Disturbance-observer-based control and related methods. an overview. *IEEE Trans Ind Electron*. 2015;63(2):1083-1095.
41. Shtessel Y, Edwards C, Fridman L, Levant A. *Sliding Mode Control and Observation*. Springer; 2014.
42. Flanders H. Differentiation under the integral sign. *Am Math Mon*. 1973;80(6):615-627.
43. Tao G. *Adaptive Control Design and Analysis*. Vol 37. John Wiley & Sons; 2003.
44. Finlayson BA. *The Method of Weighted Residuals and Variational Principles*. Vol 73. SIAM; 2013.
45. Yang X, Liu D, Ma H, Xu Y. Online approximate solution of HJI equation for unknown constrained-input nonlinear continuous-time systems. *Inf Sci*. 2016;328:435-454.
46. Si J, Wang YT. Online learning control by association and reinforcement. *IEEE Trans Neural Netw*. 2001;12(2):264-276.
47. Wang D, Liu D, Li H. Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems. *IEEE Trans Autom Sci Eng*. 2014;11(2):627-632.
48. Jankovic M. Robust control barrier functions for constrained stabilization of nonlinear systems. *Automatica*. 2018;96:359-367.

**How to cite this article:** Li C, Wang Y, Liu F, Liu Q, Buss M. Model-free incremental adaptive dynamic programming based approximate robust optimal regulation. *Int J Robust Nonlinear Control*. 2022;32(5):2662-2682. doi: 10.1002/rnc.5964