Learning-Based Attitude Tracking Control with High-Performance Parameter Estimation

Hongyang Dong, Xiaowei Zhao, Qinglei Hu, Haoyang Yang, and Pengyuan Qi

***Abstract*—This paper aims to handle the optimal attitude tracking control tasks for rigid bodies via a reinforcement learning-based control scheme, in which a constrained parameter estimator is designed to compensate system uncertainties accurately. This estimator guarantees the exponential convergence of estimation errors and can strictly keep all instant estimates always within pre-determined bounds. Based on it, a critic-only adaptive dynamic programming (ADP) control strategy is proposed to learn the optimal control policy with respect to a user-defined cost function. The matching condition on reference control signals, which is commonly employed in relevant ADP design, is not required in the proposed control scheme. We prove the uniform ultimate boundedness of the tracking errors and critic weight's estimation errors under finite excitation conditions by Lyapunov-based analysis. Moreover, an easy-to-implement initial control policy is designed to trigger the real-time learning process. The effectiveness and advantages of the proposed method are verified by both numerical simulations and hardware-in-loop experimental tests.**

***Index Terms*—Attitude tracking control; adaptive dynamic programming; parameter estimation; adaptive control.**

I. INTRODUCTION

The attitude control problem has aroused extensive attention [1], [2], due to its essential applications in aerospace and mechanical engineering. Various control techniques, such as passivity-based control [3] and sliding mode control [4], [5], have been successfully employed to solve attitude stabilization, maneuver, and tracking control problems. However, these methods usually lack the ability to balance the closed-loop performance and control cost. Such a trade-off is important for many practical tasks (e.g., fuel and electricity, which are control costs, are the most valuable resources for on-orbit satellites especially in deep-space missions). In this sense, optimal control is a more suitable choice. However, given the high complexity of attitude control systems, analytically solving the corresponding Hamilton-Jacobi-Bellman (HJB) equations is challenging, especially in tracking cases. Ref. [6] presented an inverse optimal approach for attitude regulation, which avoided directly solving the HJB equation. Sharma and Tewari [7] proposed an optimal solution for attitude maneuvers for a specific class of rigid bodies with diagonal inertia matrices. Nevertheless, these elegant results require information on full system dynamics. Due to various issues, including structure changes, payload movement, and fuel consumption, parameter uncertainties are common and inevitable problems that restrict the applications of most existing optimal control methods. For tracking cases, though adaptive control approaches, e.g. [8], [9], can realize attitude tracking objectives subject to uncertainties, these results have no optimizing abilities. They thus may potentially lead to high control costs. Luo et al. [10] extended the result in [6] and proposed an adaptive attitude tracking controller based on the inverse optimal approach. But this elegant result can only be applied to special cost functions. To sum up, the incompatibility between optimizing and adapting abilities has become a severe bottleneck in the development of optimal attitude tracking control systems subject to uncertain parameters.

Adaptive dynamic programming (ADP) [11], [12], [13], [14], which is a class of control strategies based on reinforcement learning, is a promising tool to address the aforementioned technical challenge. ADP has aroused enormous interest and attention recently. Its fundamental principle is to improve control policy by properly evaluating feedbacks from environments, avoiding directly analytically solving the HJB equations. In tracking cases, forgetting factors [15], [16] can be employed in ADP to address the issue induced by the nonautonomous nature of systems (which leads the original cost functions to become ill-defined and time-varying). An ADP-based model-free tracking control method was proposed in [17], in which an approximated preview of the desired reference trajectory was designed. To avoid introducing forgetting factors into the cost functions, Kamalapurkar et al. [18], [19] reformulated the optimal tracking problems to be optimal stationary problems. They employed reference signals as augmented states, addressing the issues induced by the nonautonomous nature. A drawback of the excellent results in [18], [19] is the requirement of explicit matching conditions on reference signals, which are usually strong or subject to the precise knowledge of system models. Moreover, there are severe technical barriers to the design of ADP-based controllers for attitude tracking problems: 1) General attitude tracking control problems require tracking state trajectories instead of reference control signals. Therefore, the reference control signal is not prior information and must be deduced by system dynamics. Thus, no explicit matching conditions can be ensured subject to system uncertainties. 2) Parameter uncertainties are inevitable issues of on-orbit spacecraft. This issue leads to additional difficulties to design ADP-based atti-

This work was funded by the UK Engineering and Physical Sciences Research Council (grant number: EP/S001905/1). H. Dong and X. Zhao (Corresponding Author) are with the Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry CV4 7AL, UK. Emails: hongyang.dong@warwick.ac.uk, xiaowei.zhao@warwick.ac.uk. H. Yang and Q. Hu are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191, China. Emails: yanghaoyang8352@buaa.edu.cn, huql_buaa@buaa.edu.cn. P. Qi is with the Research Institute for Frontier Science, Beihang University, Beijing, 100191, China. Email: qipengyuan@buaa.edu.cn.

tude tracking controllers. Parameter estimators with inadequate performance may result in inaccurate estimations of reference control signals and significantly degrade the overall tracking control performance.

Motivated by these facts and built upon the results in [20], [12], [14], [18], [19], [21], we design a special ADP controller for attitude tracking control tasks in this paper. A novel parameter estimator is proposed to estimate parameters. Based on it, the optimal tracking control task is equivalently transformed to an optimal stationary task. After that, a critic structure is designed to learn control policies. The boundedness of the closed-loop system is guaranteed via Lyapunov-based analysis. We summarize this paper's main contributions as follows.

1) A novel high-performance parameter estimator is designed. Distinct from conventional adaptive estimation methods [22] and the advanced concurrent-learning-based methods [23], [24], [25], our estimator has the ability to keep all instantaneous estimates within pre-determined bounds strictly. In addition, it ensures estimation errors converge to zero exponentially under relaxed excitation conditions. These features are essential for the proposed ADP controller and also the initial control policy.

2) In comparison with the ADP-based tracking control approaches in [18], [19], our controller does not require the explicit matching condition or prior knowledge on the reference control signal. Instead, the reference control signal is estimated online based on the estimator and employed to reformulate the tracking dynamics. This framework enhances the generality and flexibility of the whole control strategy.

3) Per practical implement concerns, we show that a proportional derivative (PD)-like controller, along with the proposed parameter estimator, can be employed to initialize the online learning process. This also indicates that the proposed controller can bring a commonly-employed tracking controller the essential optimizing ability in real-time.

4) The effectiveness of our controller is verified by both numerical simulations and hardware-in-loop experiments.

In the remainder of this paper, we formulate the considered control problem in Sec. II. After that, the parameter estimator and the ADP-based controller are proposed in Sec. III. Simulation and experiment results are provided in Sec. IV, and then we conclude our work in Sec. V.

II. Problem Formulation

A. Problem Formulation

Notations: We employ $\|\cdot\|$ to denote the Euclidean norm. Post-superscript $\cdot^x$ means a vector is expressed in a coordinate system $\mathcal{F}_x$. $S(\cdot)$ indicates the skew-symmetric matrice. In addition, we employ $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to denote a matrix's minimum and maximum eigenvalues, respectively.

Unit quaternions are employed to describe the attitude model. We use $\mathcal{F}_b$, $\mathcal{F}_i$, and $\mathcal{F}_r$ to denote the body-fixed frame, the inertial frame, and the reference frame, respectively. In attitude tracking control problems, $\mathcal{F}_b$ is required to track the motion of $\mathcal{F}_r$, which renders the following error kinematics and dynamics [26]

$$\dot{q}_{br} = \frac{1}{2}E(q_{br})\omega_{br}^b,\ E(q_{br}) = \begin{bmatrix} -\xi_{br}^{\mathrm{T}} \\ \zeta_{br}I_{3\times3} + S(\xi_{br}) \end{bmatrix} \tag{1}$$

$$J\dot{\omega}_{br}^b = -S(\omega_{bi}^b)J\omega_{bi}^b + J[S(\omega_{br}^b)\omega_{ri}^b - C(q_{br})\dot{\omega}_{ri}^r] + u \tag{2}$$

where $q_{br} = [\zeta_{br}, \xi_{br}^{\mathrm{T}}]^{\mathrm{T}}$ denotes a relative unit quaternion of $\mathcal{F}_b$ w.r.t. $\mathcal{F}_r$, and here $\zeta_{br} = \cos(\vartheta_{br}/2)$ and $\xi_{br} = \sin(\vartheta_{br}/2)e_{br}$ are called the quaternion q_{br}'s scalar part and vector part, respectively, with ϑ_{br} and e_{br} are the Euler eigenangle and eigenaxis of q_{br}. Moreover, one has $\zeta_{br}^2 + \xi_{br}^{\mathrm{T}}\xi_{br} = 1$. One can refer to [26] for more detailed introduction and explanation of the quaternion, the Euler eigenagle & eigneaxis, and their relationships. In addition, ω_{br}^b denotes the relative angular velocity, ω_{ri}^r is the reference angular velocity, ω_{ri}^b is the coordinate transformation of ω_{ri}^r in $\mathcal{F}_b$, u denotes the control signal, and $C(\cdot)$ is the transformation matrix defined by $C(q_{br}) = I_{3\times3} - 2\zeta_{br}S(\xi_{br}) + 2S^2(\xi_{br})$. The reference quaternion q_{ri} and angular velocity ω_{ri}^r are both user-defined signals, satisfying $\dot{q}_{ri} = 0.5E(q_{ri})\omega_{ri}^r$ and $\omega_{ri}^r, \dot{\omega}_{ri}^r \in \mathcal{L}_\infty$. Also, J denotes the inertia matrix, which is unknown for controller design.

Following the strategy in [18], [19], the control input u is decomposed to a virtual reference control signal u_r and an optimal part u_o, such that $u = u_r + u_o$. Unlike [18], [19], in which u_r can be directly obtained through the assumption of matching conditions, we construct the virtual reference control signal to be

$$u_r = JC(q_{br})\dot{\omega}_{ri}^r + S(\omega_{ri}^b)J\omega_{ri}^b = Y_r\theta \tag{3}$$

where $\theta = [J_{11}, J_{12}, J_{13}, J_{22}, J_{23}, J_{33}]^{\mathrm{T}}$ is a vector form of J, and Y_r is a regressor matrix. It is noteworthy that since the analytical expression of $JC(q_{br})\dot{\omega}_{ri}^r + S(\omega_{ri}^b)J\omega_{ri}^b$ is available, one can deduce the expression of the regressor matrix Y_r by taking Jacoibian of $JC(q_{br})\dot{\omega}_{ri}^r + S(\omega_{ri}^b)J\omega_{ri}^b$ with respect to θ, while without requiring any specific parameter value of θ.

Substituting u_r back into (2) renders

$$J\dot{\omega}_{br}^b = -S(\omega_{bi}^b)(J\omega_{bi}^b) + JS(\omega_{br}^b)\omega_{ri}^b + S(\omega_{ri}^b)J\omega_{ri}^b + u_o \tag{4}$$

One can readily verify that $q_{br} \to q_I$, $\omega_{br}^b \to 0_{3\times1}$ and $u_o \to 0_{3\times1}$ are the closed-loop system's equilibrium, and here $q_I = [1, 0, 0, 0]^{\mathrm{T}}$. Thus, by designing u_r, the original attitude tracking control problem is reformulated by (1) and (4). We aim to minimize the following performance metric by designing u_o.

$$U = \int_0^\infty [r(\tau) + u_o^{\mathrm{T}}(\tau)Ru_o(\tau)]\mathrm{d}\tau \tag{5}$$

here $r = (q_{br} - q_I)^{\mathrm{T}}Q_q(q_{br} - q_I)^{\mathrm{T}} + (\omega_{br}^b)^{\mathrm{T}}Q_\omega\omega_{br}^b$, and $Q_q \in \mathbb{R}^{4\times4}$, $Q_\omega \in \mathbb{R}^{3\times3}$ and $R \in \mathbb{R}^{3\times3}$ are positive-definite.

However, since θ is unknown, only its estimation can be employed in the controller design. Therefore, u_r also needs to be estimated, formalized by $\hat{u}_r = Y_r\hat{\theta}$. Here $\hat{u}_r$ and $\hat{\theta}$ denote the estimates of u_r and θ, respectively. A novel parameter estimator will be designed in Sec.III.A to address this issue.

B. Optimal Solution Analysis

To formulate the optimal attitude tracking control task, the system model is re-organized to the following governing form

$$\dot{\eta} = F(\eta) + Gu_o \tag{6}$$

where $\eta = [(q_{br} - q_I)^{\mathrm{T}}, (\omega_{br}^b)^{\mathrm{T}}, (q_{ri} - q_I)^{\mathrm{T}}, (\omega_{ri}^r)^{\mathrm{T}}, (\dot{\omega}_{ri}^r)^{\mathrm{T}}]^{\mathrm{T}} = [\eta_1^{\mathrm{T}}, \eta_2^{\mathrm{T}}, \eta_3^{\mathrm{T}}, \eta_4^{\mathrm{T}}, \eta_5^{\mathrm{T}}]^{\mathrm{T}}$, and

$$F(\eta) = \begin{bmatrix} 0.5E(\eta_1 + q_I)\eta_2 \\ (Y + Y_r)\theta \\ 0.5E(\eta_3 + q_I)\eta_4 \\ \eta_5 \end{bmatrix}, \; G = \begin{bmatrix} 0_{4\times 3} \\ I_{3\times 3} \\ 0_{4\times 3} \\ 0_{3\times 3} \end{bmatrix}.$$

Besides, Y is a regressor matrix satisfying $Y\theta = -S(\omega_{bi}^b)(J\omega_{bi}^b) + J[S(\omega_{br}^b)\omega_{ri}^b - C(q_{br})\dot{\omega}_{ri}^r]$.

Based on the cost index U, we define a cost function (or called performance metric) in the following equation.

$$V = \int_t^{\infty} [r(\tau) + u_o^{\mathrm{T}}(\tau) R u_o(\tau)] \mathrm{d}\tau \tag{7}$$

We denote the optimal control policy and cost function by u_o^* and V^*, respectively. As discussed in [18], [19], V^* is a time-invariant function of η, and it satisfies the so-called Hamiltonian function [27], [28]:

$$H(\eta, u_o, \nabla_\eta V) = \nabla_\eta^{\mathrm{T}} V[F + Gu_o] + r + u_o^{\mathrm{T}} R u_o = 0 \tag{8}$$

By taking the partial differential for Eq. (8), one can get the closed-form optimal control policy: $u_o^* = -0.5R^{-1}G^{\mathrm{T}}\nabla_\eta V^*$. Then the HJB equation in terms of $\nabla_\eta V^*$ can be obtained by introducing u_o^* back into Eq. (8):

$$r + \nabla_\eta^{\mathrm{T}} V^* F - \frac{1}{4}\nabla_\eta^{\mathrm{T}} V^* G R^{-1} G^{\mathrm{T}} \nabla_\eta V^* = 0 \tag{9}$$

Conventional optimal control methods aim to directly solve the analytical solutions for u_o^* and V^*. However, as discussed in Introduction, such a task is challenging for the attitude tracking systems subject to uncertainties. We address this non-trivial task by adaptive dynamic programming (ADP) [27]. ADP is a state-of-the-art control method that combines the ideas of reinforcement learning and dynamic programming (DP), which learns through environment feedback instead of directly solving optimal control problems. It addresses the "curse of dimensionality" problem in DP. In this paper, a special parameter-estimator-based ADP approach is developed to learn the optimal solutions of Eq. (9), in which a novel parameter estimator is designed for system uncertainty compensation and a critic structure is introduced to estimate $\nabla_\eta V^*$ and u_o^*.

III. Parameter-Estimator-Based Critic-Only ADP

A. Parameter Estimator Design

The parameter estimator plays an essential role in the attitude tracking control problem considered here. As shown in the last section, system uncertainties influence the design of both u_r and u_o. Besides, for many practical systems such as on-orbit satellites, though the inertia matrix J is usually unknown, it is trivial to have prior knowledge about its lower and upper bounds based on the structure and component information. To be specific, there exist $\theta_{k,\min}, \theta_{k,\max} \in \mathbb{R}$, such that $\theta_k \in (\theta_{k,\min}, \theta_{k,\max})$, and here θ_k denotes the kth entry, $k = 1, 2, ..., 6$. The estimates out of these bounds make no sense and potentially degrade the performance of the whole estimation process. Moreover, conventional online estimators/identifiers usually require system states or control signals to satisfy the persistence explication (PE) condition [23] to guarantee precise estimation. However, such conditions are quite strong and unattainable for attitude control systems. To address these issues, we propose a parameter estimator that achieves exponential convergence under relaxed excitation conditions. Our design also ensures instant estimates are always within pre-determined bounds.

Recalling the definition of the regressor matrix Y, the original attitude tracking dynamics (2) can be re-organized to a compact form: $J\dot{\omega}_{br} = Y\theta + u$. Then we construct the following filtered variables

$$\begin{aligned} \dot{Y}_f(t) &= -\alpha Y_f(t) + Y(t),\; Y_f(0) = 0_{3\times 6} \\ \dot{u}_f(t) &= -\alpha u_f(t) + u(t),\; u_f(0) = 0_{3\times 1} \\ \dot{\omega}_f(t) &= -\alpha \omega_f(t) + \omega_{br}^b(t),\; \omega_f(0) = (1/\alpha)\omega_{br}^b(0) \end{aligned} \tag{10}$$

where α is a user-defined positive constant. Based on Eq. (10), one has

$$\frac{\mathrm{d}}{\mathrm{d}t}(J\dot{\omega}_f - Y_f\theta - u_f) = -\alpha(J\dot{\omega}_f - Y_f\theta - u_f) \tag{11}$$

Therefore,

$$J\dot{\omega}_f = Y_f\theta + u_f + \gamma \tag{12}$$

and here the term $\gamma(t) = \gamma(0)\mathrm{e}^{-\alpha t}$ is vanishing exponentially. Based on Eqs. (10)-(12), the initial condition of γ satisfies

$$\gamma(0) = J[-\alpha\omega_f(0) + \omega_{br}^b(0)] - Y_f(0) - u_f(0)$$

Here $Y_f(0)$, $u_f(0)$ and $\omega_f(0)$ are expected to be properly chosen such that $\gamma(0) = 0$, which can further ensure $\forall t \geq 0, \gamma(t) \equiv 0$. One can readily verify that the settings given in Eq. (10) satisfy such a requirement. Therefore, Eq. (12) shows

$$u_f = J\dot{\omega}_f - Y_f\theta \tag{13}$$

The above equation indicates that the filtered signal u_f contains the information of unknown parameters. Another important fact is that, $\dot{\omega}_f$ is available for the parameter estimator design (since it only relies on the information of ω_{br}^b). These facts are our key motivations of employing the filtering structure in Eq. (10).

In order to keep instantaneous estimates always within user-defined bounds, we define a projection law in the following equation.

$$\theta_k = (\theta_{k,\max} - \theta_{k,\min})\mathrm{sig}(\psi_k) + \theta_{k,\min} \tag{14}$$

with $\mathrm{sig}(\cdot) : \mathbb{R} \to (0, 1)$ is the sigmoid function. We employ the sigmoid function because it is a smooth function that can project $x \in (-\infty, +\infty)$ to $\mathrm{sig}(x) \in (0, 1)$. Other alternatives can also be considered, such as the tanh function, but then the projection law in Eq. (14) also needs to be re-designed. Based on Eq. (14), one can see that θ_k on $(\theta_{k,\min}, \theta_{k,\max})$ is projected to ψ_k on $\mathbb{R}$, for $k = 1, 2, ..., 6$.

Theorem 1: Given the original attitude tracking model in (2) and the filtered dynamics in (12), designing the following parameter update law

$$\begin{aligned} \dot{\hat{\psi}}(t) = &- \mu_1(Y_\theta^{\mathrm{T}}(t)Y_\theta(t)\hat{\theta}(t) - Y_\theta^{\mathrm{T}}(t)u_f(t)) \\ &- \mu_2\sum_{i=1}^{l}(Y_\theta^{\mathrm{T}}(t_i)Y_\theta(t_i)\hat{\theta}(t) - Y_\theta^{\mathrm{T}}(t_i)u_f(t_i)) \end{aligned} \tag{15}$$

here μ_1 and μ_2 are positive constants, and $\hat{\theta}$ and $\hat{\psi}$ are respectively the orginal & projected estimates of θ. Y_θ is an auxiliary regressor matrix satisfying $Y_\theta \theta = J\dot{\omega}_f - Y_f\theta$, and t_i denotes time indices with $0 \leq t_i \leq t$, $i = 1, 2, ...l$. Taking

$$\hat{\theta}_k = (\theta_{k,\max} - \theta_{k,\min})\text{sig}(\hat{\psi}_k) + \theta_{k,\min} \tag{16}$$

with $\hat{\theta}_k$ and $\hat{\psi}_k$ are the k^{th} entries of $\hat{\theta}$ and $\hat{\psi}$, respectively. Then the estimation error is bounded, and all instantaneous estimates are within the pre-determined bounds: $\forall t \geq 0$, $\hat{\theta}_k(t) \in (\theta_{k,\min}, \theta_{k,\max})$. Moreover, if $\sum_{i=1}^{l} Y_\theta^{\mathrm{T}}(t_i)Y_\theta(t_i)$ is full-rank, one has $\tilde{\theta} = \hat{\theta} - \theta$ exponentially converges to zero.

Proof: See Appendix A.

Remark 1: The idea of introducing both real-time measurements $[Y_\theta^{\mathrm{T}}(t)Y_\theta(t)]$ and past measurements $[\sum_{i=1}^{l} Y_\theta^{\mathrm{T}}(t_i)Y_\theta(t_i)]$ into the estimation law is inspired by the concurrent learning (CL) method proposed in [23], [24]. This design ensures the convergence of $\tilde{\theta}$ under the requirement that $\sum_{i=1}^{l} Y_\theta^{\mathrm{T}}(t_i)Y_\theta(t_i)$ is full-rank. This requirement is much weaker than conventional online estimation/identification methods, which require Y_θ to satisfy the persistent excitation condition.

Remark 2: Significantly distinct from the CL method and its recent advances [23], [24], [25], [19], our parameter estimator has the ability to keep all instantaneous estimates within user-defined bounds. Moreover, we strictly prove that $\tilde{\theta}$ can converge exponentially under the projected framework. In addition, the original CL method requires the information of immeasurable state (which is $\dot{\omega}_{br}^b$ in our study). This issue is addressed by the filtering structure in our design (as shown in Eqs. (10)-(13)). All these important results show the proposed estimator's advantages and greatly enhance our ADP-based control strategy's performance and applicability.

B. Parameter-Estimator-Based ADP Controller Design

In this section, we design a parameter-estimator-based ADP controller to approxiamte the optimal control policy. According to the Weierstrass approximation theorem, for $\eta \in \mathcal{X}$, with $\mathcal{X}$ denotes a compact set, a set of basis functions $\sigma(\eta) = [\sigma_1(\eta), \sigma_2(\eta), ..., \sigma_p(\eta)]^{\mathrm{T}} \in \mathbb{R}^p$ can be employed to approximate V^*, formalized by

$$V^* = W^{\mathrm{T}}\sigma(\eta) + \epsilon(\eta) \tag{17}$$

where $\sigma_i(\eta)$ satisfies $\sigma_i(0) = 0$ and $(\mathrm{d}\sigma_i/\mathrm{d}\eta)|_{\eta=0} = 0$, $i = 1, 2, ..., p$. Moreover, the vector $W \in \mathbb{R}^p$ is the weight of basis functions, and the term $\epsilon(\eta) \in \mathbb{R}$ denotes the error induced by reconstruction. Based on that, we have

$$u_o^* = -\frac{1}{2}R^{-1}G^{\mathrm{T}}(\nabla_\eta \sigma W + \nabla_\eta \epsilon) \tag{18}$$

Given the reconstruction in Eqs. (17) and (18), the critic aims to evaluate the unknown weight vector W such that the following function can approximate the optimal cost function V^* as in (17),

$$V = \hat{W}^{\mathrm{T}}\sigma \tag{19}$$

where $\hat{W}$ denotes the estimate of W. If $\hat{W} \to W$, the critic can provide a good estimate for the Hamiltonian as follows.

$$\hat{H}(\eta, u_o^*, \hat{W}^{\mathrm{T}}\nabla_\eta^{\mathrm{T}}\sigma) = \hat{W}^{\mathrm{T}}\nabla_\eta^{\mathrm{T}}\sigma(\hat{F} + Gu_o^*) + r + (u_o^*)^{\mathrm{T}}Ru_o^* \tag{20}$$

Here $\hat{F}$ is the guess of F in which θ is estimated by $\hat{\theta}$. Moreover, we define δ_b in the following equation as the approximation error of the Hamiltonian (or referred to as the Bellman error).

$$\delta_b = \hat{H}(\eta, u_o^*, \hat{W}^{\mathrm{T}}\nabla_\eta^{\mathrm{T}}\sigma) - H(\eta, u_o^*, \nabla_\eta V^*) \tag{21}$$

Recall the fact that $H(\eta, u_o^*, \nabla_\eta V^*) = 0$, we have

$$\delta_b = \hat{H}(\eta, u_o^*, \hat{W}^{\mathrm{T}}\nabla_\eta^{\mathrm{T}}\sigma)$$

A commonly-used strategy [20], [11], [12] for critic training is updating $\hat{W}$ in order to minimize the squared Bellman error $E = 0.5\delta_b^2$, and an approximate control law: $\hat{u} = -0.5R^{-1}G^{\mathrm{T}}\nabla_\eta \sigma \hat{W}$ is employed during the learning process. Based on (20) and the normalized gradient descent algorithm, one can minimize E by designing the following weight update law [11], [12]

$$\begin{aligned}\dot{\hat{W}} &= -c_1 \frac{1}{(\varpi^{\mathrm{T}}(t)\varpi(t) + 1)^2}\frac{\partial E}{\partial \hat{W}} \\ &= -c_1 \frac{\varpi}{(\varpi^{\mathrm{T}}(t)\varpi(t) + 1)^2}(\varpi^{\mathrm{T}}\hat{W} + r + u_o^{\mathrm{T}}Ru_o)\end{aligned} \tag{22}$$

where the positive constant c_1 is user-defined, and $\varpi = \nabla_\eta^{\mathrm{T}}\sigma(\hat{F} + Gu_o)$. Recall the property of Hamiltonian, one can show that Eq. (22) is equivalent to

$$\dot{\hat{W}} = -c_1 \frac{\varpi}{(\varpi^{\mathrm{T}}(t)\varpi(t) + 1)^2}(\varpi^{\mathrm{T}}\tilde{W} + W\nabla_\eta^{\mathrm{T}}\sigma\tilde{F} - \epsilon_H) \tag{23}$$

where $\tilde{W} = \hat{W} - W$, $\tilde{F} = \hat{F} - F$, and $\epsilon_H = -\nabla_\eta^{\mathrm{T}}\epsilon(F + Gu_o^*)$ denotes the residual error [11], [12].

In this paper, we employ both past & real-time measurements to update $\dot{\hat{W}}$. This is built upon the concurrent learning (CL) approach [19], [23], [12], aiming to improve the critic updating performance. Before providing the design details of our $\dot{\hat{W}}$, we make several assumptions as follows.

Assumption 1: For $\eta \in \mathcal{X}$, where $\mathcal{X}$ can be any compact set, one has ϵ, $\nabla_\eta \epsilon$ and ϵ_H are bounded. Moreover, these reconstruction and residual errors go to zero if sufficient basis functions are selected.

Assumption 2: The variable $D = \nabla_\eta^{\mathrm{T}}\sigma GRG^{\mathrm{T}}\nabla_\eta \sigma$ is bounded. Therefore, there exists a constant $b_D > 0$ so that $\|D\| \leq b_D$ for η.

Assumption 3: The variable $\phi = \varpi/(\varpi^{\mathrm{T}}\varpi + 1)$ follows an finite excitation (FE) condition [23]. To be specific, there exist t_{w1}, t_{w2}, c_w with $0 \leq t_{w1} \leq t_{w2} \leq t$ and $c_w > 0$ such that $\int_{t_{w1}}^{t_{w2}} \phi(\tau)\phi^{\mathrm{T}}(\tau)\mathrm{d}\tau \geq c_w I_{p\times p}$.

Assumption 4: The initial control policy u_o is admissible on the whole state definition domain and can stabilize the system (6) from any initial condition.

We note that the assumptions 1 & 2 are standard [20], [29], [11], [12], [30]. The assumption 3 is mild for attitude tracking dynamics. The last assumption is given the fact that the attitude tracking system is controllable on the whole state definition domain. More importantly, the assumption 4 ensures that the initial controller can keep system states inside a compact set $\mathcal{X}$, which lays the foundation for the application of Weierstrass approximation.

For ease of notations, we define the following auxiliary variables $\varsigma = \varpi^{\mathrm{T}}\varpi + 1$, $\varphi_1 = \phi\phi^{\mathrm{T}}$, and $\varphi_2 = \phi(r + u_o^{\mathrm{T}} R u_o)/\varsigma$. Then we design the critic's update law as follows,

$$\begin{aligned}\dot{\hat{W}}(t) = &- c_1 \frac{\varpi}{(\varpi^{\mathrm{T}}(t)\varpi(t)+1)^2}(\varpi^{\mathrm{T}}\hat{W} + r + u_o^{\mathrm{T}} R u_o) \\ &- c_2 \Xi(t, t_{w2}, t_{w1})\end{aligned} \tag{24}$$

with $c_1, c_2 > 0$. In addition, $\Xi(t, t_{w2}, t_{w1}) = \xi_1(t_{w2}, t_{w1})\hat{W}(t) + \xi_2(t_{w2}, t_{w1})$ with

$$\dot{\xi}_1(t, t_{w1}) = -\kappa\xi_1(t, t_{w1}) + \varphi_1(t), \quad \xi_1(t_{w1}) = 0_{p\times p} \tag{25}$$

$$\dot{\xi}_2(t, t_{w1}) = -\kappa\xi_2(t, t_{w1}) + \varphi_2(t), \quad \xi_2(t_{w1}) = 0_{p\times 1} \tag{26}$$

Based on the definition of Ξ, we have

$$\begin{aligned}\Xi(t, t_{w2}, t_{w1}) &= \int_{t_{w1}}^{t_{w2}} \mathrm{e}^{-\kappa(t_{w2}-\tau)}(\varphi_1(\tau)\hat{W}(t) + \varphi_2(\tau))\mathrm{d}\tau \\ &= \xi_1(t_{w2}, t_{w1})\tilde{W}(t) + \Omega(t_{w2}, t_{w1})\end{aligned} \tag{27}$$

where $\xi_1(t_{w2}, t_{w1}) = \int_{t_{w1}}^{t_{w2}} \mathrm{e}^{-\kappa(t_{w2}-\tau)}\phi(\tau)\phi^{\mathrm{T}}(\tau)\mathrm{d}\tau$ functions as an information matrix which "stores" the information of ϖ throughout the time interval $[t_{w1}, t_{w2}]$, and $\Omega(t_{w2}, t_{w1}) = \int_{t_{w1}}^{t_{w2}} \mathrm{e}^{-\kappa(t_{w2}-\tau)}[W\nabla_\eta\sigma(\tau)\tilde{F}(\tau) + \epsilon_H(\tau)]\phi(\tau)/\varsigma(\tau)\mathrm{d}\tau$ denotes a constant error vector. With the assumption 3, we have $\xi_1(t_{w2}, t_{w1}) \geq c_\Phi$, and here $c_\Phi = e^{-\kappa(t_{w2}-t_{w1})}c_w$. Based on these preliminaries, we are ready to propose our ADP controller.

Theorem 2: Considering the governing model in (6), the estimator proposed in Theorem 1 with assumption of $\sum_{i=1}^{l}\Phi(t_i)$ is full-rank, and the critic-only ADP strategy

$$u_o = -\frac{1}{2}R^{-1}G^{\mathrm{T}}\nabla_\eta\sigma\hat{W}. \tag{28}$$

Under Assumptions 1-4, designing the update law of critic weights as (24). Then $\tilde{W}$, ξ_{br} and ω_{br}^b are uniformly ultimately bounded (UUB).

Proof: See Appendix B.

Remark 3: Theorem 2 shows that tracking errors and critic weight's estimation errors converge to a residual set bounded by $\sqrt{b/b_\lambda}$. It should be emphasized that this residual set can be significantly reduced if sufficient basis functions are selected such that $\epsilon, \nabla_\eta\epsilon, \epsilon_H \to 0$. It also shrinks quickly with the convergence of x.

Remark 4: By employing the past measurements to update $\dot{\hat{W}}$, the closed-loop system's stability is ensured under an FE condition. This design also allows us to replace the commonly-used actor-critic structure in ADP with a simplified critic-only structure. Besides, different from relevant results [12], [18], [19], [21], our integral-form data collection approach makes use of all incoming data, and it is also arguably easier to implement.

Remark 5: It should be emphasized that the parameter estimator designed in Sec. III.A plays an essential role in the whole control strategy's stability and performance. On the one hand, as shown in Sec. III.A and Eq. (44), our estimator's non-certainty-equivalence (non-CE) feature introduces the negative quadratic term of estimation error (i.e. $-\|\tilde{\theta}\|^2$) in the stability analysis, which is indispensable to the closed-loop system's UUB. On the other hand, the exponential convergence feature of $\|\tilde{\theta}\|^2$ ensures the accurate estimation for the reference control signal u_r, which is the key to achieving high-performance tracking control.

Remark 6: This remark explains the difference between our work and the studies in [11], [12]. Though these three studies all employ the squared Bellman error to update the critic's weights, they have significant and different task-oriented designs far beyond this backbone. Particularly, Ref. [11] focuses on achieving H_∞-based ADP control, and it also employs the event-triggering strategy to reduce the potential communication burden in specific applications. Ref. [12] aims to address the input saturation issue for an ADP-based controller and shows the closed-loop stability subject to input saturation. In contrast, our paper shows how to manage the optimal tracking control problem subject to system uncertainties under the ADP framework. The methods in [11], [12] cannot handle this problem because they lack the ability to estimate uncertain parameters and reference control inputs, while such information is crucial to ensure closed-loop stability. We propose a novel estimator to estimate this information accurately and achieve high-performance tracking. As explained in Remark 5, our estimator's non-CE and exponential convergence features are essential for the closed-loop stability and the whole control task. As an application-oriented study, we further explain in Subsection III.C how to choose the initial control policy for the optimal attitude tracking control problem considered here. Moreover, hardware-in-loop experiments are conducted in Section IV to test the performance of our estimator-based ADP control method.

C. Construction of the Initial Control Policy

Same with many other online ADP control methods [12], [18], [19], [21], an admissible initial control policy is needed to trigger real-time learning. In this subsection, we show that a PD-like controller

$$u = -k_p\xi_{br} - k_d\omega_{br}^b + Y_r\hat{\theta} \tag{29}$$

is an admissible initial control policy, where k_p and k_d are positive constants, and $\tilde{\theta}$ is updated by our estimator. To show the system stability under (29), we employ a candidate Lyapunov function in the equation below.

$$V_{pd} = k_p(1-\zeta_{br})^2 + k_p\xi_{br}^{\mathrm{T}}\xi_{br} + \frac{1}{2}(\omega_{br}^b)^{\mathrm{T}}(J\omega_{br}^b) + \varrho V_I \tag{30}$$

where ϱ is employed for analysis purposes and it satisfies $\varrho > \max_{t\geq 0}\{\|Y_r(t)\|^2\}/(2\mu_2\mu_\theta)$. Since Y_r is bounded, one can always find such a proper ϱ. Recalling Eqs. (4), (39) and (29), the time derivative of V_{pd} satisfies

$$\begin{aligned}\dot{V}_{pd} = &- k_d\|\omega_{br}^b\|^2 - (\omega_{br}^b)^{\mathrm{T}}(Y_r\tilde{\theta}) + \varrho\dot{V}_I \\ \leq &- 0.5k_d\|\omega_{br}^b\|^2 - 0.5\varrho\mu_2\mu_\theta\|\tilde{\theta}\|^2\end{aligned} \tag{31}$$

So one has $\omega_{br}^b \in \mathcal{L}_\infty \cap \mathcal{L}_2$. Then by analyzing $\dot{\omega}_{br}^b$ and employing the Barbalat lemma, we can conclude that $\lim_{t\to\infty}\{\xi_{br}, \omega_{br}^b\} = 0$.

This design indicates that one can choose the initial policy for our parameter-estimator-based ADP controller as follows

$$u_o = -k_p\xi_{br} - k_d\omega_{br}^b \tag{32}$$

Besides, this initial controller can be easily described by a set of polynomial basis functions: $\sigma_{pd} = [\xi_{br1}\omega^b_{br1}, \xi_{br2}\omega^b_{br2}, \xi_{br3}\omega^b_{br3}, \frac{1}{2}(\omega^b_{br1})^2, \frac{1}{2}(\omega^b_{br2})^2, \frac{1}{2}(\omega^b_{br3})^2]^{\mathrm{T}}$ with the corresponding weights to be $\hat{W}_{pd} = [2r_{11}k_p, 2r_{22}k_p, 2r_{33}k_p, 2r_{11}k_d, 2r_{22}k_d, 2r_{33}k_d]^{\mathrm{T}}$, and here ξ_{bri}, ω^b_{bri} and r_{ii} are the i^{th} entries of ξ_{br}, ω^b_{br} and R, respectively. One can readily verify that $u_o = -0.5R^{-1}G^{\mathrm{T}}\nabla_\eta\sigma_{pd}\hat{W}_{pd}$ is consistent with (32).

A remaining issue is that the assumption 1 requires the boundedness of $G^{\mathrm{T}}\nabla_\eta\sigma$ for any η, thus σ_{pd} cannot be directly employed to construct σ. Recalling the fact that $\|\xi_{br}\| < 1$, this boundedness requirement can be satisfied by modifying σ_{pd} to be $\sigma_{pd} = [\xi_{br1}\omega^b_{br1}, \xi_{br2}\omega^b_{br2}, \xi_{br3}\omega^b_{br3}, \int_0^{\omega^b_{br1}} s(\omega)\mathrm{d}\omega, \int_0^{\omega^b_{br2}} s(\omega)\mathrm{d}\omega, \int_0^{\omega^b_{br3}} s(\omega)\mathrm{d}\omega]^{\mathrm{T}}$, where $s(\cdot) : \mathbb{R} \to \mathbb{R}$ is defined by

$$s(x) = \begin{cases} x & \text{if abs}(x) \leq k_s \\ k_s\text{sign}(x) & \text{if abs}(x) > k_s \end{cases} \tag{33}$$

and here k_s is a user-defined positive constant. Then we have

$$\begin{aligned} u_o =& -0.5R^{-1}G^{\mathrm{T}}\nabla_\eta\sigma_{pd}\hat{W}_{pd} \\ =& -k_p\xi_{br} - k_d s(\omega^b_{br}) \end{aligned} \tag{34}$$

with $s(\omega^b_{br}) = [s(\omega^b_{br1}), s(\omega^b_{br2}), s(\omega^b_{br3})]^{\mathrm{T}}$. Eq. (34) ensures that $G^{\mathrm{T}}\nabla_\eta\sigma_{pd}$ is always bounded and u_o keeps the form in (32) when $|\omega^b_{bri}| \leq k_s$, $i = 1, 2, 3$. Actually, the controller in (34) can also stabilize the system since $\tilde{\theta}$ converge to zero exponentially.

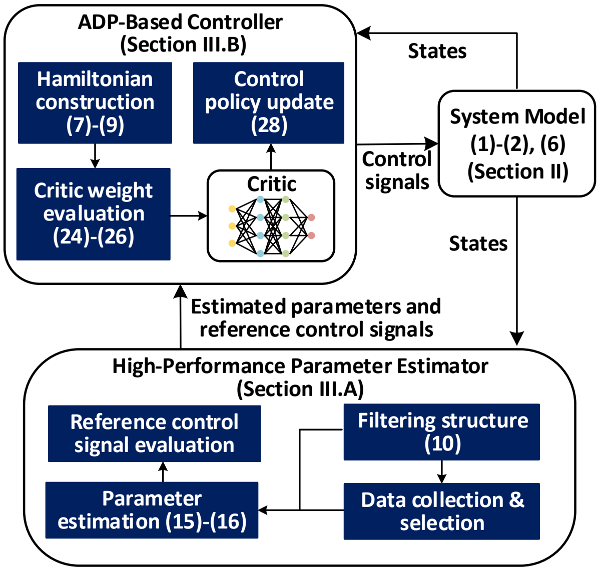
ADP-Based Controller
(Section III.B)
Hamiltonian construction (7)-(9)
Control policy update (28)
Critic weight evaluation (24)-(26)
Critic
States
Control signals
System Model (1)-(2), (6) (Section II)
States
Estimated parameters and reference control signals
High-Performance Parameter Estimator
(Section III.A)
Reference control signal evaluation
Filtering structure (10)
Parameter estimation (15)-(16)
Data collection & selection

Figure 1: The main framework of our parameter-estimator-based ADP control method.

All these results show that our control method can be easily implemented by employing the parameter-estimator-based PD-like controller for initialization. It also indicates that our method can bring the essential optimizing ability to a conventional control method. These facts indicate the flexibility and versatility of our method.

Algorithm 1 Parameter-estimator-based ADP for attitude tracking control.

```
Formalize the attitude tracking control problem via the quaternion description, as shown in Eqs. (1), (2) and (6).
Based on task requirements, construct the performance metric as shown in Eq. (5).
Initialize user-defined parameters, including $Q_q$, $Q_\omega$, $R$, $\alpha$, $\mu_1$, $\mu_2$, $c_1$, $c_2$, $\kappa$, $k_p$, $k_d$, $k_s$.
Set $Y_f(0)$, $u_f(0)$, $\omega_f(0)$, $\hat{\theta}(0)$, $\hat{W}(0)$, $\xi_1(0)$, and $\xi_2(0)$.
Initialize the terminal time $T_s$.
while $t < T_s$ do
    if $t = 0$ then
        Set $t_{w1} = 0$ and apply the initial control policy $u_o(0)$ in (32) to the system.
    end if
    Update the filtered variables in the estimator, i.e. $Y_f(t)$, $u_f(t)$ and $\omega_f(t)$, by (10).
    Based on the measurements at time $t$, employ the selection algorithm to maximize the eigenvalue of the information matrix $\sum_{i=1}^{l} Y_\theta^{\mathrm{T}}(t_i)Y_\theta(t_i)$.
    Update the projected parameter estimate $\hat{\psi}(t)$ by (15).
    Calculate the original parameter estimate $\hat{\theta}(t)$ by (16).
    if $\xi_1(t_{w2}, t_{w1})$ is not full-rank then
        Update $t_{w2}$ by setting $t_{w2} = t$.
    else
        Set $t_{w2} = t_{w2}$.
    end if
    Evaluate the weight vector $\hat{W}(t)$ by (24), (25) and (26).
    Calculate the control input $u_o(t)$ via (28) and apply it to the system.
end while
```

The main framework of our parameter-estimator-based ADP controller is provided in Fig. 1, and the specific implementation steps are summarized in Algorithm 1.

Remark 7: We analyze the computational complexity of our parameter-estimator-based ADP control method in this remark. Given the estimation laws in Eqs. (10), (15), (16) and the ADP-based control law in Eqs. (24)-(26) and (28), one can see that a limited number of additional integral and addition/subtraction operations are required in our method compared with conventional control methods, such as the PD-like controller. Particularly, the most time-costly operations in our parameter estimator are from Eq. (10), which require the users to calculate the filtered states via integral. As for our ADP-based attitude tracking controller, its complexity is directly related to the critic network's size. Based on the Weierstrass approximation theorem, the critic network's universal approximation and information processing abilities usually increase as more basis functions are embedded in it. But from (24)-(26), the size of the critic network also quadratically increases the computational complexity, and a trade-off is required in practical applications. The case studies in Section IV show that a critic network with six quadratic-form basis functions can already make a good approximation to the potential optimal tracking control solution and render superior performance to popular attitude control methods.

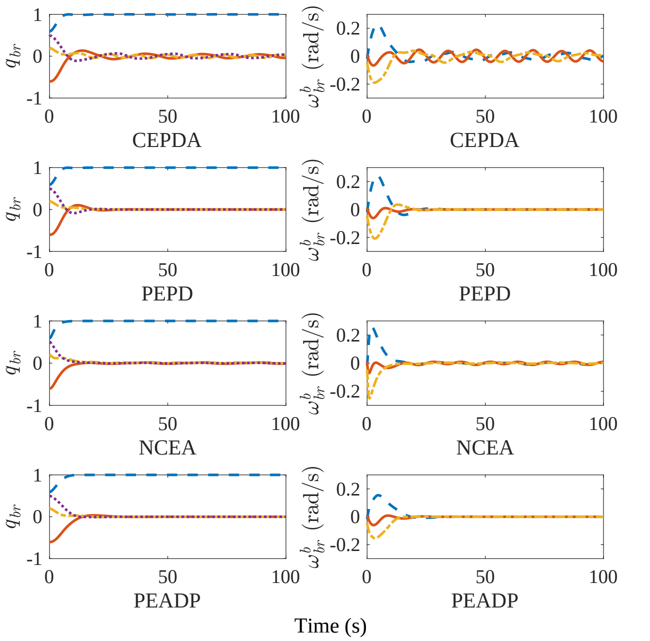

Figure 2: Results of q_{br} and ω^b_{br} under different controllers.

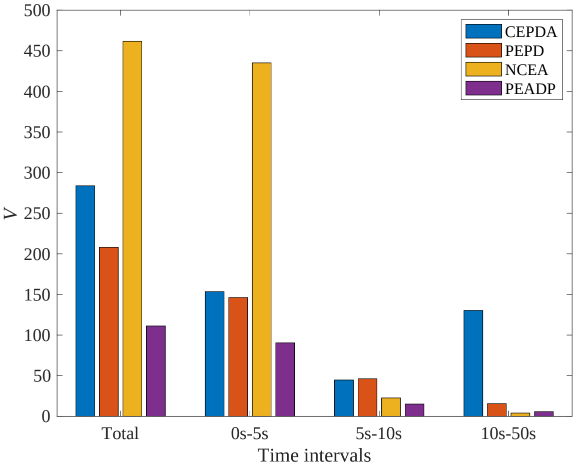

Figure 3: Simulation results of V.

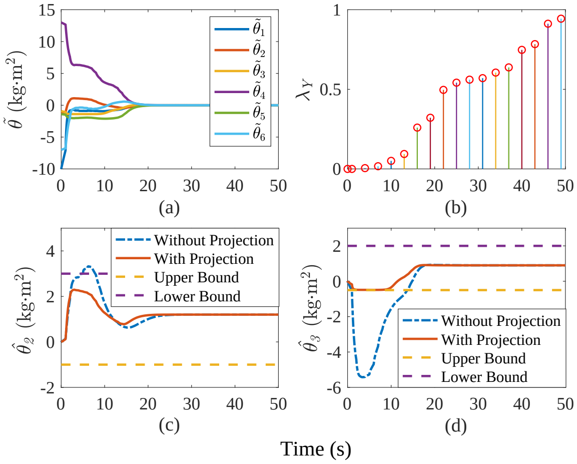

Figure 4: Illustrations of parameter estimation.

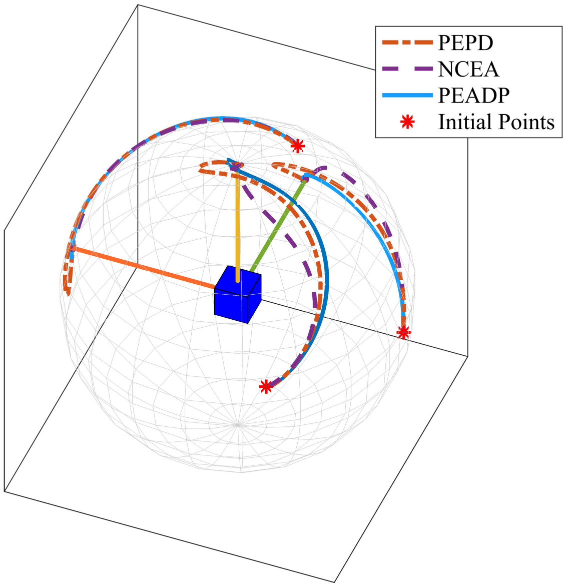

Figure 5: 3D illustration of numerical simulations.

IV. Simulation Results

A. Numerical Simulations

Numerical simulation results are presented in this subsection to show the performance of our control method. We set $\theta = [20, 1.2, 0.9, 17, 1.4, 15]^{\mathrm{T}}$kg·m^2, and the reference signals satisfy $q_{ri}(0) = q_I, \omega^r_{ri}(t) = [0.1\sin(\pi t/12), 0.05\cos(\pi t/6), -0.1\sin(\pi t/12)]^{\mathrm{T}}$rad/s. We also set $q_{br}(0) = [0.5916, -0.6, 0.2, 0.5]^{\mathrm{T}}$, $\omega^b_{br}(0) = [0, 0, 0]^{\mathrm{T}}$rad/s, and $\hat{\theta}(0) = [10, 0, 0, 30, 0, 8]^{\mathrm{T}}$kg·m^2. The upper and lower bounds of $\hat{\theta}$ are $\theta_{\max} = [25, 3, 2, 35, 3, 20]^{\mathrm{T}}$ kg·m^2 and $\theta_{\max} = [5, -1, -0.5, 12, -1, 5]^{\mathrm{T}}$kg·m^2, respectively. The cost function follows $Q_q = 10I_{4\times4}$, $Q_w = 20I_{3\times3}$ and $R = 10I_{3\times3}$. Control parameters for the proposed method are chosen to be $\alpha = 0.05$, $\mu_1 = 5$, $\mu_2 = 20$, $c_1 = 5$, $c_2 = 2$, $\kappa = 0.1$. Following the design and analysis in Sec.III.C, a PD-like controller with $k_p = 4$ and $k_d = 6$ is employed to trigger online learning, and we set $\sigma = \sigma_{pd}$ with $k_s = 0.3$. Thus we have $\hat{W}(0) = [80, 80, 80, 120, 120, 120]^{\mathrm{T}}$. Besides, by monitoring the rank of ξ_1, we set $t_{w1} = 0s$ and $t_{w2} = 5s$. After system states are stabilized ($\|\xi_{br}\| < 0.01$ and $\|\omega^b_{br}\| < 0.002$), we release the data stored in $\Xi(t, t_{w2}, t_{w1})$ to reduce the residual error.

For comparison purpose, not only the proposed parameter-estimator-based ADP method (denote as "PEADP"), but also three other controllers are considered:

1) *The Parameter-Estimator-based PD-like Controller* (denote as "PEPD"). This is the initial control policy as discussed in Sec. III.C. Comparing PEDP with PEADP can clearly show the optimizing ability of the proposed method.

2) *A Certainty Equivalence (CE)-based PD-Like Adaptive Controller* (denote as "CEPDA"). This controller also follows the form $u = -k_p\xi_{br} - k_d\omega^b_{br} + Y_r\hat{\theta}$, while $\hat{\theta}$ is updated by

the widely-employed CE-based adaptive law $\dot{\hat{\theta}} = -k_{ce} Y_r^{\mathrm{T}} \omega_{br}^b$, where we set $k_{ce} = 20$ in simulations.

3) *A Non-Certainty-Equivalence (Non-CE) Adaptive Controller in [8]* (denote as "NCEA"). This advanced adaptive control method is deduced by the immersion and invariance (I&I) philosophy [31], which has been demonstrated to have improved closed-loop performance than conventional attitude tracking control methods. One can refer to Ref. [8] for the design details of NCEA, and the control parameters in the simulation are chosen as $k_p = 0.5$, $k_v = 2$, $\alpha = 2.5$, $\gamma = 20$.

Under all these settings, the time responses of q_{br} and ω_{br}^b are given in Fig. 2. One can observe that all the four controllers ensure the boundedness of system tracking errors, while due to the existence of uncertain parameters, CEPDA leads to large residual errors. In contrast, PEPD guarantees the precise convergence of q_{br} and ω_{br}^b, which demonstrates the effectiveness of the proposed parameter estimator. NCEA and PEADP further ensure improved performance with smoother trajectories and fewer fluctuations, and one can see that PEADP permits higher convergence precision than NCEA. Simulation results of V are illustrated in Fig. 3. It indicates that the good performance of NCEA comes at the expense of a high cost. In contrast, PEADP significantly reduces the cost when compared with all the other controllers (reduced by 60.8%, 46.5% and 75.9% with respect to CEPDA, PEPD and NCEA, respectively), which ensures its optimizing ability.

Parameter estimation performance is given in Fig. 4. Fig. 4.a indicates that $\tilde{\theta}$ quickly converge to zero, which is benefited by introducing the information matrix $\sum_{i=1}^{l} Y_\theta^{\mathrm{T}}(t_i) Y_\theta(t_i)$ into the update law. Denoting $\lambda_Y = \lambda_{\min}\{\mu_2 \sum_{i=1}^{l} Y_\theta^{\mathrm{T}}(t_i) Y_\theta(t_i)\}$, Fig. 4.b further demonstrates the change of λ_Y.

Figs. 4.c and 4.d show the function of the parameter projection law, in which the convergence trajectories of $\tilde{\theta}_2$ and $\tilde{\theta}_3$ with & without projection are employed as examples. One can see that though the precise parameter estimation can be ensured independent of the projection, the estimates escape the prescribed bounds without the projection during the estimation process. While employing the parameter projection mechanism can always restrict the estimates to the pre-determined bounds, and accordingly improve the convergence process.

A three-dimensional illustration on the frame $\mathcal{F}_r$ is given in Fig. 5 to show the tracking process of $\mathcal{F}_b$ with respect to $\mathcal{F}_r$ (illustrated by the trajectories of axes). In Fig. 5, the central cube is used to show the origin of coordinate systems, and the mutually perpendicular lines pointing from it denote the axes of $\mathcal{F}_r$. The axes moving trajectories from $\mathcal{F}_b$ to $\mathcal{F}_r$ under PEPD, NCEA and PEADP are shown in Fig. 5 (CEPDA is excluded in this illustration because it cannot guarantee the precise convergence within 100s in the employed simulation case). One can see that the PEADP approach proposed in this paper renders smoother trajectories with fewer fluctuations when compared with PEPD and NCEA.

For the case study considered in this subsection, the time costs for the 100-second simulation runs (on a computer with 2.9GHz Intel Core i7 and 16GB RAM) under different controllers are illustrated in Fig. 6, which verify the analysis in Remark 7. Particularly, the time costs of CEPDA and PEPD are 0.7036s and 1.1639s, respectively, showing that the computational burden caused by our parameter estimator is quite small. Moreover, the time cost of our PEADP is 1.8399s, which is even smaller than that of the NCEA method (2.0881s). In addition, the following subsection also verifies that our method works perfectly fine in real-time hardware-in-loop experiments. All these results indicate that the additional computational burden induced by our PEADP (with respect to other popular attitude controllers such as CEPDA and NCEA) is mild and will have limited influence on its potential onboard implementation.

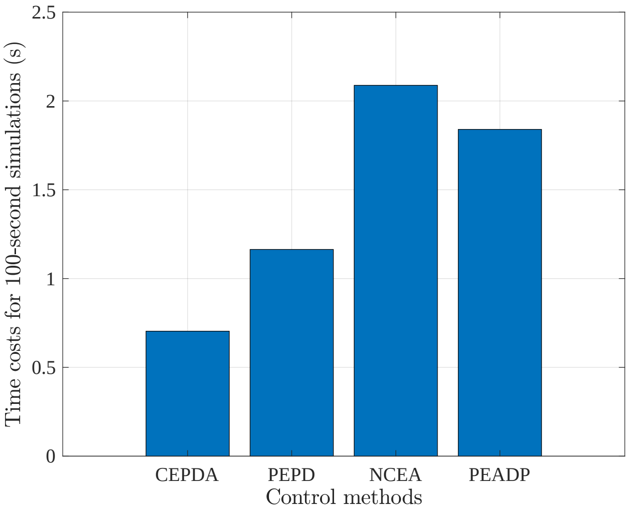

Figure 6: Time costs for 100-second simulations under different controllers.

B. Hardware-in-Loop Experiments

Hardware-in-loop (HL) experimental results are given in this subsection to further demonstrate the features and effectiveness of the proposed control method. The HL testbed is illustrated in Fig. 7. Its main components include 1) A triaxial turntable to simulate the attitude motion of spacecraft. 2) A reliable real-time simulation computer. Particularly, the weighted pseudo-inverse algorithm designed in [32] is employed for control allocation. 3) An ARM-based underlying control PCB. 4) Four reaction wheels serve as actuator simulators. With all the hardware, the overall performance of control methods can be comprehensively tested under practical measurement noises and control signal disturbances.

The time step in experiments is 0.05s. Based on the physical constraints of the testbed (the maximum reaction wheel output is 0.1Nm, with the maximum slope to be 0.01Nm/s), the reference angular velocity is reset to be $\boldsymbol{\omega}_{ri}^r(t) = [0.01\sin(0.1t), 0.02\sin(0.05t), 0.015\sin(0.08t)]^{\mathrm{T}}$ rad/s. Correspondingly, control parameters are modified as follows. For PEADP, we set $c_1 = 1$ and $c_2 = 0.1$. For PEPD, we set $k_p = 0.1$ and $k_d = 3$. The NCEA method follows $k_p = 0.05$, $k_d = 0.2$, $\alpha = 0.25$ and $\gamma = 5$. All the other settings are same as the ones in numerical simulations.

Experimental results of q_{br} and ω_{br} under different controllers are illustrated in Fig. 8. One can see that PEADP still has superior performance to all the other methods, permitting

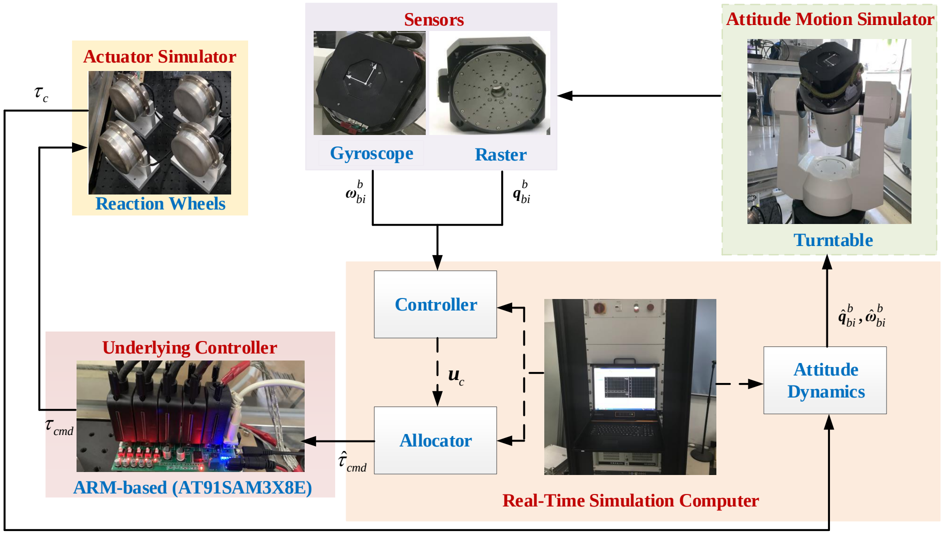

Figure 7: The hardware-in-loop testbed for experiments.

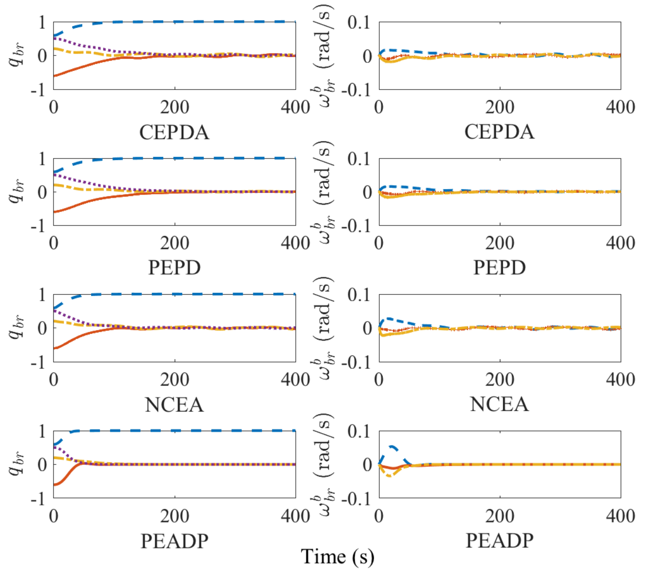

Figure 8: Experimental results of q_{br} and ω_{br}^{b} under different controllers.

a faster convergence process and better precision. It also renders less overall cost as shown in Fig. 9. The 3D trajectory illustration is given in Fig. 10. All these experiment results further indicate the effectiveness and advantages of our control method.

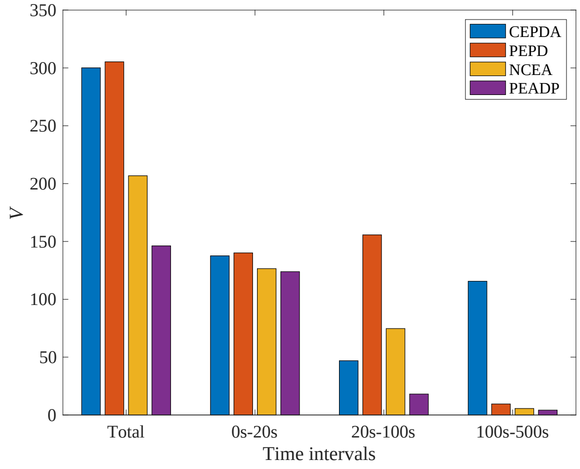

Figure 9: Experimental results of V.

V. Conclusions

This paper developed a parameter-estimator-based ADP control scheme for attitude tracking control tasks. The proposed parameter estimator can ensure exponential convergence and keep all instant estimates within user-defined bounds. Based on it, a critic-only structure was proposed to learn the optimal control policy w.r.t. the cost function. Both past & real-time measurements were employed to improve the performance of both parameter estimation and reference tracking. The features and effectiveness of our method were illustrated by not only numerical simulations but also hardware-in-loop experiments, under comparisons with conventional and advanced adaptive control methods. Deep RL methods, such as deep deterministic policy gradient, will be investigated in the future to realize data-driven, model-free tracking control for more complicated on-orbit tasks. Considering measurement models and optimal state estimation subject to measurement errors is also an interesting topic that is worth investigating in the future.

Appendix

A. Proof of Theorem 1

We define $\Phi(t) = Y_\theta^{\mathrm{T}}(t)Y_\theta(t)$ for ease of notation. Recalling the fact that $u_f = Y_\theta\theta$, one has (15) satisfies

$$\dot{\tilde{\psi}}(t) = -\mu_1\Phi(t)\tilde{\theta} - \mu_2[\sum_{i=1}^{l}\Phi(t_i)]\tilde{\theta}(t) \tag{35}$$

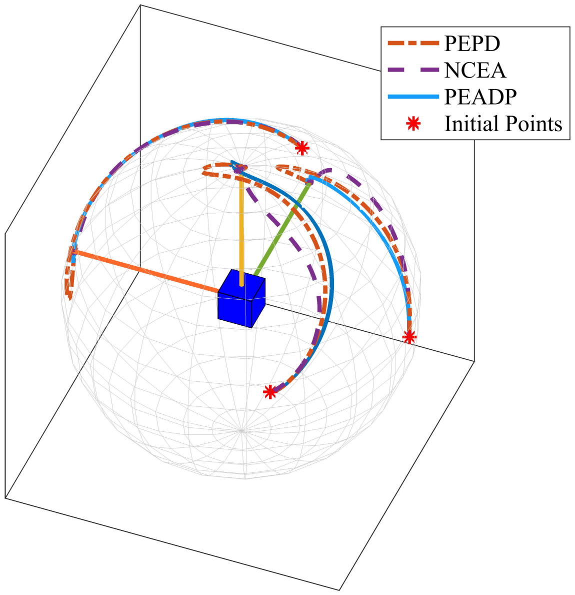

Figure 10: 3D illustration of experimental results.

where $\tilde{\theta} = \hat{\theta} - \theta$. Based on it, we employ a storage function in the equation below.

$$\begin{aligned} V_I =& (\theta_{k,\max} - \theta_{k,\min}) \sum_{k=1}^{6} [\tilde{\psi}_k + \ln(1 + \mathrm{e}^{-\tilde{\psi}_k - \psi_k}) \\ & - \tilde{\psi}_k \mathrm{sig}(\psi_k) - \ln(1 + \mathrm{e}^{-\psi_k})] \end{aligned} \tag{36}$$

and here $\tilde{\psi}_k = \hat{\psi}_k - \psi_k$. One can readily prove that $V_I \to \infty$ when $\|\tilde{\psi}\| \to \infty$. Besides, we have

$$\frac{\partial V_I}{\partial \tilde{\psi}_k} = (\theta_{k,\max} - \theta_{k,\min})(\mathrm{sig}(\hat{\psi}_k) - \mathrm{sig}(\psi_k)) = \tilde{\theta}_k \tag{37}$$

This indicate $\tilde{\psi}_k \le 0$ if $\partial V_I / \partial \tilde{\psi}_k \le 0$ and vice versa. Thus V_I is a valid Lyapunov candidate function of $\tilde{\psi}$. Therefore, $\dot{V}$ satisfies,

$$\dot{V}_I(t) = -\mu_1 \|Y_\theta(t)\tilde{\theta}(t)\|^2 - \mu_2 \tilde{\theta}^{\mathrm{T}}(t) [\sum_{i=1}^{l} \Phi(t_i)] \tilde{\theta}(t) \tag{38}$$

One has $V_I(t) \ge 0$ and $\dot{V}_I(t) \le 0$, so $V_I \in \mathcal{L}_\infty$ and $\tilde{\psi} \in \mathcal{L}_\infty$. Recalling (16), we have $\forall t \ge 0$, $\hat{\theta}_k(t) \in (\theta_{k,\min}, \theta_{k,\max})$.

Moreover, when $\sum_{i=1}^{l} \Phi(t_i)$ is full-rank, we have $\mu_\theta = \lambda_{\min}[\sum_{i=1}^{l} \Phi(t_i)] > 0$. Under this condition, one can further obtain

$$\dot{V}_I(t) = -\mu_1 \|Y_\theta(t)\tilde{\theta}(t)\|^2 - \mu_2 \mu_\theta \|\tilde{\theta}(t)\|^2 \tag{39}$$

We state that V_I has an important property as described in the following equation.

$$c_{\min} \|\tilde{\theta}\|^2 \le V_I \le c_{\max} \|\tilde{\theta}\|^2 \tag{40}$$

where $c_{\min} = \inf_{t \ge 0, k=1,2,\ldots,6} \gamma_k(t)$, $c_{\max} = \sup_{t \ge 0, k=1,2,\ldots,6} \gamma_k(t)$, and here $\gamma_k(t) = (1 + \mathrm{e}^{-\tilde{\psi}_k(t) - \psi_k})^2 / [(\theta_{k,\max} - \theta_{k,\min}) \mathrm{e}^{-\tilde{\psi}_k(t) - \psi_k}]$. It is noteworthy that, since $\tilde{\psi} \in \mathcal{L}_\infty$, $c_{\min}$ and $c_{\max}$ are bounded. To prove (40), first we consider the right-hand side of it. Define an auxiliary variable $M(\tilde{\psi}_k) = c_{\max} \tilde{\theta}_k^2 - (\theta^k_{\max} - \theta^k_{\min}) [\tilde{\psi}_k + \ln(1 + \mathrm{e}^{-(\tilde{\psi}_k + \psi_k)}) - \tilde{\psi}_k \mathrm{sig}(\psi_k) - \ln(1 + \mathrm{e}^{-\psi_k})]$. Then it can be verified that

$$\frac{\partial M}{\partial \tilde{\psi}_k} = \tilde{\theta}_k [\frac{2c_{\max}}{\gamma_k} - 1] \tag{41}$$

Thus if $\tilde{\psi}_k \le 0$, one has $\partial M / \partial \tilde{\psi}_k \le 0$, and vice versa. These facts indicate $M(\tilde{\psi}_k) \ge M(0) = 0$ for all $\tilde{\psi}_k \in \mathbb{R}$. Therefore, by summing up $M(\tilde{\psi})$ for all k, the right-hand side of (40) is ensured, and similar analysis can be employed to prove the other part. Combining the result of Eqs. (39) and (40), one has

$$\|\tilde{\theta}(t)\|^2 \le \frac{V_I(0) \mathrm{e}^{-\mu_2 \mu_\theta / c_{\max}}}{c_{\min}} \tag{42}$$

Thus $\tilde{\theta}$ converges exponentially. The proof is complete.

B. Proof of Theorem 2

Considering the storage function in the equation below,

$$L = V^* + 0.5 \rho_1 \tilde{W}^{\mathrm{T}} \tilde{W} + \rho_2 V_I \tag{43}$$

where ρ_1 and ρ_2 are positive constants.

$\dot{L}$ is analyzed in (44), in which $D = \nabla_\eta^{\mathrm{T}} \sigma G R G^{\mathrm{T}} \nabla_\eta \sigma$ and $b = 0.5 \nabla_\eta^{\mathrm{T}} \epsilon G R^{-1} G^{\mathrm{T}} \nabla_\eta \epsilon + \rho_1 c_1 \epsilon_H^2 / (\varpi^{\mathrm{T}} \varpi + 1)^2 + 0.5 \|(W^{\mathrm{T}} \nabla_\eta^{\mathrm{T}} \sigma + \nabla_\eta^{\mathrm{T}} \epsilon) G Y_r\|^2 + 0.5 \rho_1 c_2 \|\Omega(t_{w1}, t_{w2})\|^2 / c_\Phi$. Notice that the arithmetic-geometric mean inequality is employed in (44). Recalling the assumption 1, one has $D \in \mathcal{L}_\infty$. Furthermore, for $\eta \in \mathcal{X}$, there exists positive constant b_F so that $\|W \nabla_\eta \sigma \tilde{F} / (1 + \varpi^{\mathrm{T}} \varpi + 1)\| \le b_F \|\tilde{\theta}\|$. Therefore, by setting $\rho_1 = 2b_D / (c_2 c_\Phi)$ and $\rho_2 = (2b_F^2 + 1)/(\mu_2 \mu_\theta)$, $\dot{L}$ satisfies

$$\begin{aligned} \dot{L} \le & - (q_{br} - q_I)^{\mathrm{T}} Q_q (q_{br} - q_I)^{\mathrm{T}} - (\omega^b_{br})^{\mathrm{T}} Q_\omega \omega^b_{br} \\ & - \frac{1}{2} \rho_1 \tilde{W}^{\mathrm{T}} (c_1 \phi \phi^{\mathrm{T}} + c_2 c_\Phi I) \tilde{W} - \frac{1}{2} \rho_2 \mu_2 \mu_\theta \|\tilde{\theta}\|^2 + b \end{aligned} \tag{45}$$

We denote $z = [(q_{br} - q_I)^{\mathrm{T}}, (\omega^b_{br})^{\mathrm{T}}, \tilde{W}^{\mathrm{T}}]^{\mathrm{T}}$ and $b_\lambda = \lambda_{\min}\{Q_q, Q_\omega, 0.5\rho_1(c_1 \phi \phi^{\mathrm{T}} + c_2 c_\Phi I)\}$. Then Eq. (45) indicates that $\dot{L} \le 0$ if $\|z\| \ge \sqrt{b/b_\lambda}$.

Based on these results, the final step of this proof is to guarantee that the hypersphere with a radius $\sqrt{b/b_\lambda}$ (in the state definition domain) lies inside a compact $\mathcal{X}$ that is required by the assumption 2. We denote this hypersphere by $\mathcal{S}$. Recalling the assumptions 1 & 2, one has the following two facts: 1) $\sqrt{b/b_\lambda}$ is irrelevant to the size of $\mathcal{X}$ and 2) there is no restriction on the upper bound of the size of $\mathcal{X}$. Thus there always exists a compact set $\mathcal{X}$ so that $\mathcal{S} \subset \mathcal{X}$. Moreover, if $\eta(0) \in \mathcal{X}$, then $\forall t \ge 0$, $\eta(t) \in \mathcal{S} \subset \mathcal{X}$. This ensures the UUB of q_{br}, ω^b_{br} and $\tilde{W}$. The proof is complete.

References

[1] Q. Hu and B. Jiang, "Continuous finite-time attitude control for rigid spacecraft based on angular velocity observer," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 3, pp. 1082–1092, 2017.

[2] J. Qiao, H. Wu, and X. Yu, "High-precision attitude tracking control of space manipulator system under multiple disturbances," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, Early Access.

[3] Y. Igarashi, T. Hatanaka, M. Fujita, and M. W. Spong, "Passivity-based attitude synchronization in $se(3)$," *IEEE Transactions on Control Systems Technology*, vol. 17, no. 5, pp. 1119–1134, 2009.

$$
\begin{aligned}
\dot{L} =&\nabla_\eta^{\mathrm{T}} V^*(F+Gu_o-Gu_o^*+Gu_o^*+G\hat{u}_r-Gu_r)+\rho_1\tilde{W}^{\mathrm{T}}\dot{\hat{W}}+\rho_2\dot{V}_I\\
=&\nabla_\eta^{\mathrm{T}} V^*(F+Gu_o^*)+\nabla_\eta^{\mathrm{T}} V^*(Gu_o-Gu_o^*+GY_r\tilde{\theta})+\rho_1\tilde{W}^{\mathrm{T}}\dot{\hat{W}}+\rho_2\dot{V}_I\\
=&\nabla_\eta^{\mathrm{T}} V^*(F+Gu_o^*)+(W^{\mathrm{T}}\nabla_\eta^{\mathrm{T}}\sigma+\nabla_\eta^{\mathrm{T}}\epsilon)(-\frac{1}{2}GR^{-1}G^{\mathrm{T}}\nabla_\eta\sigma\tilde{W}+\frac{1}{2}GR^{-1}G^{\mathrm{T}}\nabla_\eta\epsilon+GY_r\tilde{\theta})+\rho_1\tilde{W}^{\mathrm{T}}\dot{\hat{W}}+\rho_2\dot{V}_I\\
=&-r-\frac{1}{4}W^{\mathrm{T}}DW-\frac{1}{2}W^{\mathrm{T}}D\tilde{W}-\frac{1}{2}\nabla_\eta^{\mathrm{T}}\epsilon GR^{-1}G^{\mathrm{T}}\nabla_\eta\sigma\tilde{W}+\frac{1}{4}\nabla_\eta^{\mathrm{T}}\epsilon GR^{-1}G^{\mathrm{T}}\nabla_\eta\epsilon+(W^{\mathrm{T}}\nabla_\eta^{\mathrm{T}}\sigma+\nabla_\eta^{\mathrm{T}}\epsilon)GY_r\tilde{\theta}\\
&-\rho_1\tilde{W}^{\mathrm{T}}[c_1\varpi(\varpi^{\mathrm{T}}\hat{W}+r+u_o^{\mathrm{T}}Ru_o)/(\varpi^{\mathrm{T}}\varpi+1)^2+c_2\Xi(t,t_{w2},t_{w1})]-\rho_2[\mu_1\|Y_\theta(t)\tilde{\theta}(t)\|^2+\mu_2\mu_\theta\|\tilde{\theta}(t)\|^2]\\
\leq&-r-\frac{1}{2}\tilde{W}^{\mathrm{T}}[\rho_1c_1\phi\phi^{\mathrm{T}}+\rho_1c_2c_\Phi I_{p\times p}-D]\tilde{W}-(\rho_2\mu_2\mu_\theta-\frac{1}{2})\|\tilde{\theta}\|^2+\|W\nabla_\eta\sigma\tilde{F}/(1+\varpi^{\mathrm{T}}\varpi+1)\|^2+b
\end{aligned}
\tag{44}
$$

[4] A.-M. Zou, K. D. Kumar, Z.-G. Hou, and X. Liu, "Finite-time attitude tracking control for spacecraft using terminal sliding mode and chebyshev neural network," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 4, pp. 950–963, 2011.

[5] S. S.-D. Xu, C.-C. Chen, and Z.-L. Wu, "Study of nonsingular fast terminal sliding-mode fault-tolerant control," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 6, pp. 3906–3913, 2015.

[6] M. Krstic and P. Tsiotras, "Inverse optimal stabilization of a rigid spacecraft," *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 1042–1049, 1999.

[7] R. Sharma and A. Tewari, "Optimal nonlinear tracking of spacecraft attitude maneuvers," *IEEE Transactions on Control Systems Technology*, vol. 12, no. 5, pp. 677–682, 2004.

[8] D. Seo and M. R. Akella, "High-performance spacecraft adaptive attitude-tracking control through attracting-manifold design," *Journal of guidance, control, and dynamics*, vol. 31, no. 4, pp. 884–891, 2008.

[9] Q. Liu, M. Liu, and J. Yu, "Adaptive fault-tolerant control for attitude tracking of flexible spacecraft with limited data transmission," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, Early Access.

[10] W. Luo, Y.-C. Chu, and K.-V. Ling, "Inverse optimal adaptive control for attitude tracking of spacecraft," *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1639–1654, 2005.

[11] D. Wang, C. Mu, D. Liu, and H. Ma, "On mixed data and event driven design for adaptive-critic-based nonlinear H_∞ control," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 993–1005, 2017.

[12] K. G. Vamvoudakis, M. F. Miranda, and J. P. Hespanha, "Asymptotically stable adaptive–optimal control algorithm with saturating actuators and relaxed persistence of excitation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2386–2398, 2016.

[13] B. Luo, D. Liu, T. Huang, and J. Liu, "Output tracking control based on adaptive dynamic programming with multistep policy evaluation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 10, pp. 2155–2165, 2017.

[14] H. Dong, X. Zhao, and H. Yang, "Reinforcement learning-based approximate optimal control for attitude reorientation under state constraints," *IEEE Transactions on Control Systems Technology*, 2020, Early Access.

[15] B. Luo, D. Liu, T. Huang, and D. Wang, "Model-free optimal tracking control via critic-only q-learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 2134–2144, 2016.

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

[17] F. Köpf, S. Ramsteiner, L. Puccetti, M. Flad, and S. Hohmann, "Adaptive dynamic programming for model-free tracking of trajectories with time-varying parameters," *International Journal of Adaptive Control and Signal Processing*, 2020, published online.

[18] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, 2015.

[19] R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 753–758, 2017.

[20] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[21] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, 2016.

[22] P. A. Ioannou and J. Sun, *Robust adaptive control*. Courier Corporation, 2012.

[23] G. Chowdhary, M. Muhlegg, and E. Johnson, "Exponential parameter and tracking erorr convergence guarantees for adaptive controllers without persistency of excitation," *International Journal of Control*, vol. 87, no. 8, pp. 1583–1603, 2014.

[24] G. Chowdhary and E. Johnson, "Theory and flight-test validation of a concurrent-learning adaptive controller," *Journal of Guidance, Control, and Dynamics*, vol. 34, no. 2, pp. 592–607, 2012.

[25] R. Kamalapurkar, B. Reish, G. Chowdhary, and W. E. Dixon, "Concurrent learning for parameter estimation using dynamic state-derivative estimators," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3594–3601, 2017.

[26] H. Schaub and J. L. Junkins, *Analytical Mechanics of Space Systems*. Reston, VA: AIAA Education Series, 2003.

[27] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming with Applications in Optimal Control*, Springer, 2017.

[28] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. John Wiley & Sons, 2017.

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

[30] C. Mu, Y. Zhang, Z. Gao, and C. Sun, "Adp-based robust tracking control for a class of nonlinear systems with unmatched uncertainties," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, Early Access.

[31] A. Astolfi and R. Ortega, "Immersion and invariance: A new tool for stabilization and adaptive control of nonlinear systems," *IEEE Transactions on Automatic control*, vol. 48, no. 4, pp. 590–606, 2003.

[32] H. Yang and Q. Hu, "Research and experiment on dynamic weight pseudo-inverse control allocation for spacecraft attitude control system," in 38^{th} *Chinese Control Conference*. IEEE, 2019, Guangzhou, China, pp. 8200–8205.

Hongyang Dong is currently a Research Fellow in Machine Learning and Intelligent Control at the School of Engineering, University of Warwick, Coventry, UK. He obtained his Ph.D. degree in Control Science and Engineering from Harbin Institute of Technology, Harbin, China, in 2018. His current research interests include reinforcement learning, deep learning, intelligent control, adaptive control, and their applications.

Xiaowei Zhao is Professor of Control Engineering and an EPSRC Fellow at the School of Engineering, University of Warwick, Coventry, UK. He obtained the PhD degree in Control Theory from Imperial College London in 2010. After that he worked as a postdoctoral researcher at the University of Oxford for three years before joining Warwick in 2013. His main research areas are control theory with applications on offshore renewable energy systems, local smart energy systems, and autonomous systems.

Qinglei Hu received the B.Eng. degree in electrical and electronic engineering from Zhengzhou University, Zhengzhou, China, in 2001, and the Ph.D. degree in control science and engineering with the specialization in guidance and control from the Harbin Institute of Technology, Harbin, China, in 2006. From 2003 to 2014, he was with the Department of Control Science and Engineering, Harbin Institute of Technology. He joined Beihang University, Beijing, China, in 2014, as a Full Professor. His current research interests include variable structure control and applications, and fault-tolerant control and applications. In these areas, he has authored or coauthored more than 80 technical articles. Dr. Hu serves as an Associate Editor for Aerospace Science and Technology.

Haoyang Yang received the B.Eng. degree in measurement and control technique and instruments from Harbin Institute of Technology, Harbin, China, in 2017. He is currently pursuing the Ph.D. degree in navigation, guidance, and control with Beihang University, Beijing, China. His current research interests include reinforcement learning-based control, intelligent control, and attitude and 6-DOF motion control. He is also working on the hardware-in-loop experiments for various nonlinear control systems.

Pengyuan Qi received the Ph.D. degree at the University of Warwick, Coventry, U.K, in 2020. He is currently a Lecturer in control engineering with Beihang University, Beijing, China. His research interests include modeling & control of very flexible aircraft and control of electro-hydraulic servo systems.