Reinforcement Learning-Based Approximate Optimal Control For Attitude Reorientation Under State Constraints

Hongyang Dong, Xiaowei Zhao, and Haoyang Yang

***Abstract*—This paper addresses the attitude reorientation problems of rigid bodies under multiple state constraints. A novel reinforcement learning (RL)-based approximate optimal control method is proposed to make the trade-off between control cost and performance. The novelty lies in that it guarantees constraint handling abilities on attitude forbidden zones and angular-velocity limits. To achieve this, barrier functions are employed to encode the constraint information into the cost function. Then an RL-based learning strategy is developed to approximate the optimal cost function and control policy. A simplified critic-only neural network (NN) is employed to replace the conventional actor-critic structure once adequate data is collected online. This design guarantees the uniform boundedness of reorientation errors and NN weight estimation errors subject to the satisfaction of a finite excitation condition, which is a relaxation compared with the persistent excitation condition that is typically required for this class of problems. More importantly, all underlying state constraints are strictly obeyed during the online learning process. The effectiveness and advantages of the proposed controller are verified by both numerical simulations and experimental tests based on a comprehensive hardware-in-loop testbed.**

***Index Terms*—Attitude Control; Reinforcement Learning; Adaptive Dynamic Programming; State Constraints; Approximate Optimal Control.**

I. Introduction

Attitude control of rigid bodies is widely investigated in aerospace engineering [1], [2], [3], [4]. Since this type of control action normally consumes very valuable resources (e.g., the fuel and electricity in on-orbit missions), optimal attitude control methods have aroused extensive attention, given their essential abilities to balance the control cost and performance. Theoretically, in order to achieve optimal control, one needs to solve the Hamilton-Jacobi-Bellman (HJB) equation subject to a user-defined cost function. However, due to the high nonlinearity of kinematics and dynamics, analytically solving general optimal attitude control problems is a challenging task. An optimal solution for attitude reorientation problems was proposed in [5]. However, it can only be applied to the rigid bodies whose inertia matrices are diagonal. Krstic and Tsiotras [6] employed an inverse optimal approach to address the optimal attitude reorientation control problem without directly solving the HJB equation. An extension for the tracking cases was studied in Luo et al. [7]. However, the inverse optimal approach only can be applied to special cases with a certain class of cost functions.

In addition to optimizing requirements, the designs of attitude controllers are normally subject to some underlying state constraints. From the standpoint of practical engineering, these constraints are critical to the success of tasks. Unexpected constraint violations may cause severe safety problems and financial loss. Particularly, in attitude reorientation missions (a typical case is the High-Revolution Earth Observation Satellite Systems in China [8]), the control objective is to reorient the payload to the desired direction (e.g., pointing the sensitive payloads to targets). At the same time, underlying state constraints must be obeyed during the control process. Here, mainly two types of state constraints are considered. First, special payloads, such as infrared telescopes, must be kept away from direct exposure to the sunlight or other bright celestial objects [9], [10], which forms a set of attitude forbidden zones. Second, per safety concerns, the angular velocities need to be restricted. For example, NASA's X-ray Timing Explorer (XTE) requires that the angular velocities should always be within the saturation limit of its rate gyros during reorientation actions [10], [11].

Research efforts have been carried out recently to solve these attitude control problems subject to multiple state constraints. An open-loop path planning method was presented in [12] for a single-axis pointing problem under attitude constraints. Gupta et al. [13], [14] proposed model predictive control (MPC) methods for constrained attitude reorientation problems. However, the high computational complexities limit the potential of open-loop path planning and MPC methods. An alternative solution is to design real-time feedback control schemes with state constraint handling abilities. By employing artificial potentials (APs), Lee and Mesbahi [9] designed a feedback attitude reorientation control law subject to multiple attitude constraint zones. Shen et al. [10] extended this result by considering additional angular-velocity limits. Some other constrained feedback controllers for six-degree-of-freedom problems were given in [15], [16]. However, AP-based controllers lack essential optimizing abilities and cannot make the trade-off between control cost and performance.

In this paper, a novel online reinforcement learning (RL)-based controller is proposed to address the optimal attitude reorientation control problems. The RL-based control technique [17], [18], [19], [20], [21], [22], which is commonly referred

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 School of Engineering, University of Warwick, Coventry, CV4 7AL, UK. Emails: {hongyang.dong, xiaowei.zhao}@warwick.ac.uk; H. Yang is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191, China. Email: yanghaoyang8352@buaa.edu.cn.

to as the adaptive dynamic programming (ADP), has aroused great research interests recently. The fundamental principle of this new technique is to improve the control performance by judiciously evaluating system feedback/responses. It can iteratively approximate the optimal control policy and avoid analytically solving the intractable HJB equation. However, the state constraint handling ability of RL-based control is still immature [23]. In this paper, a constrained RL-based control method is proposed, and a critic-only neural network (NN) structure is designed to approximate the optimal cost function and control policy. Lyapunov-based stability analysis guarantees the uniform boundedness (UB) of NN weight estimation errors, barrier functions, and state errors, subject to the satisfaction of a finite excitation condition. The main contribution of the present paper includes:

1) The proposed method brings the essential constraint handling abilities on complex state constraints to the RL-based control framework. Specially designed barrier functions are employed to encode the information of attitude forbidden zones and angular-velocity limits. The boundedness of these barrier functions can ensure that all underlying state constraints are strictly obeyed.
2) The proposed controller has significant advantages over the non-RL feedback control methods for constrained attitude reorientation problems [9], [10]. It has the ability to make a trade-off between control cost and performance. Besides, the only additional computational cost of the proposed method is induced by the updating law of NN. Thus it also has a largely reduced computational complexity when compared with open-loop path planning and MPC methods [13], [14]. These facts greatly enhance the generality and application potentials of the proposed method.
3) A novel online learning strategy with a simplified structure under relaxed excitation conditions is proposed. Inspired by the concurrent learning method [24], [25] and its extensions/applications [26], [27], both real-time data and past measurements are concurrently utilized in the updating law of the NN weights. This design relaxes the persistent excitation (PE) condition that is required by conventional RL-based controllers [20], [22]. An integral-type information matrix is designed to make full use of all incoming data, avoiding the complex selection algorithms employed in [26], [27].

The remainder of this paper is organized as follows. The constrained attitude reorientation problem is formalized in Sec. II. Then the RL-based approximate optimal controller is designed in Sec. III. Simulation and experiment results are given in Sec. IV, and finally, the paper is concluded in Sec. V.

Notations: In this paper, $\|\cdot\|$ denotes the Euclidean norm of vectors and the induced norm of matrices; $\otimes$ is the multiplication operator of quaternions; $\nabla_x(\cdot) = (\partial(\cdot)/\partial x)^{\mathrm{T}}$, where $(\cdot)^{\mathrm{T}}$ is the transpose of the corresponding vector/matrix, and we also denote $\nabla_x^{\mathrm{T}}(\cdot) = (\nabla_x(\cdot))^{\mathrm{T}}$. Besides, $\mathbb{S}^4$ denotes the definition domain of unit quaternions.

II. Preliminaries and Problem Formulation

A. Attitude Kinematics and Dynamics

We employ $\mathcal{F}_r = \{X_r, Y_r, Z_r\}$ and $\mathcal{F}_b = \{X_b, Y_b, Z_b\}$ to denote the inertia frame and the body-fixed frame, respectively. Then the attitude model of $\mathcal{F}_b$ with respect to $\mathcal{F}_r$ is [28]

$$\dot{q}_b = \frac{1}{2}E(q_b)\omega_b^b,\ E(q_b) = \begin{bmatrix} -\xi_b^{\mathrm{T}} \\ \eta_b I_{3\times 3} + S(\xi_b) \end{bmatrix} \tag{1}$$

$$J\dot{\omega}_b^b = -S(\omega_b^b)(J\omega_b^b) + u \tag{2}$$

where $q_b = [\eta_b, \xi_b^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{S}^4$ is a unit quaternion, which is employed to describe the attitude kinematics between the two frames. Here $\eta_b = \cos(\vartheta_b/2)$ and $\xi_b = \sin(\vartheta_b/2)e_b$ are the scalar part and vector part of q_b, respectively, with ϑ_b and e_b are the eigenangle and eigenaxis associated with q_b. Under this definition, one can easily verify that $q_b^{\mathrm{T}}q_b = \eta_b^2 + \xi_b^{\mathrm{T}}\xi_b = 1$. Besides, J is the inertia matrix of the rigid body; ω_b^b is the angular velocity of $\mathcal{F}_b$ with respect to $\mathcal{F}_r$; u is the control input to be designed; $I_{3\times 3}$ denotes the identity matrix; and $S(\cdot)$ denotes the skew-symmetric matrices of three dimensional vectors. The post-superscript $\cdot^x$ indicates the corresponding vector is expressed in the a frame $\mathcal{F}_x$. We employ $q_d = [\eta_d, (\xi_d)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{S}^4$ to denote the desired attitude, where η_d and ξ_d are respectively the scalar and vector parts of q_d. Then the error quaternion can be defined by $q_e = [\eta_e, \xi_e^{\mathrm{T}}]^{\mathrm{T}} = q_d^* \otimes q_b$, where $q_d^* = [\eta_d, -(\xi_d)^{\mathrm{T}}]^{\mathrm{T}}$ is the conjugate of q_d, and the operator "$\otimes$" is defined by $q_1 \otimes q_2 = [\eta_1\eta_2 - \xi_1^{\mathrm{T}}\xi_2, (\eta_1\xi_2 + \eta_2\xi_1 + \xi_1 \times \xi_2)^{\mathrm{T}}]^{\mathrm{T}}$, for any $q_1, q_2 \in \mathbb{S}^4$. Morevoer, the error kinematics satisfies [28]

$$\dot{q}_e = \frac{1}{2}E(q_e)\omega_b^b,\ E(q_e) = \begin{bmatrix} -\xi_e^{\mathrm{T}} \\ \eta_e I_{3\times 3} + S(\xi_e) \end{bmatrix}. \tag{3}$$

In conventional attitude reorientation problems, the control objective is to maneuver the attitude of the rigid body from $q_b(0)$ to q_d, formalized as $\lim_{t\to\infty} q_e(t) = q_I$, $\lim_{t\to\infty} \omega_b^b(t) = 0_{3\times 1}$, and here $q_I = [1, 0, 0, 0]^{\mathrm{T}}$ denotes the identity unit quaternion. However, as mentioned in the introduction, underlying state constraints are often required to be obeyed during this process. This is analyzed in the following subsections.

B. Attitude Constraints

As discussed in the introduction, in many important applications, attitude reorientation operations should keep sensitive payloads away from any direct exposure to harmful bright objects. This is illustrated in Fig. 1. In the figure, α_i denotes the normalized boresight vector of a sensitive payload i, and β_j is the normalized direction vector towards an object j that needs to be avoided. To avoid unexpected exposure, the bright object should be kept off from a cone-like field-of-view of the payload, and the corresponding half-cone angle is denoted by θ_{ij} with $0 \le \theta_{ij} \le \pi/2$.

This requirement can be described by the following inequality, based on the geometric relation illustrated in Fig. 1.

$$\alpha_i^b \cdot \beta_j^b - \cos\theta_{ij} < 0. \tag{4}$$

In (4), both α_i and β_j are expressed in the frame $\mathcal{F}_b$ (α_i^b and β_j^b, respectively). Note that α_i^b is a constant vector under

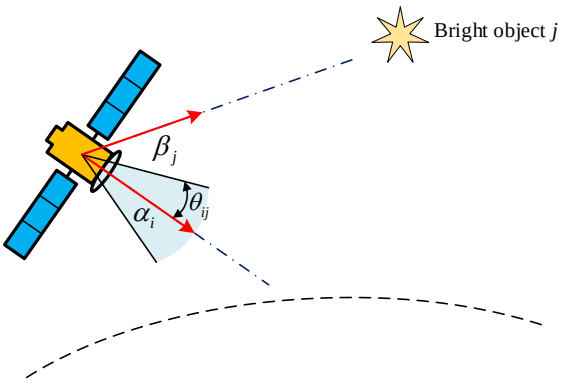


Figure 1: Attitude constraint illustration

the reasonable assumption that sensitive payloads are fixed in the rigid body. Moreover, given the coordinate transformation law, we have $\beta_j^b = C(q_b)\beta_j^r$, where $C(q_b)$ is a transformation matrix satisfying $C(q_b) = I_{3\times3} - 2\eta_b S(\xi_b) + 2S(\xi_b)S(\xi_b)$, and β_j^r is the expression of β_j in the frame $\mathcal{F}_r$. Thus (4) is equivalent to

$$\alpha_i^b \cdot [C(q_b)\beta_j^r] - \cos\theta_{ij} < 0 \tag{5}$$

and (5) can be further organized to be

$$q_e^{\mathrm{T}}\Xi^{\mathrm{T}}(q_d)M_{ij}\Xi(q_d)q_e < 0 \tag{6}$$

with $M_{ij} = \begin{bmatrix} (\alpha_i^b)^{\mathrm{T}}\beta_j^r, & (\alpha_i^b \times \beta_j^r)^{\mathrm{T}} \\ \alpha_i^b \times \beta_j^r, & 2\alpha_i^b(\beta_j^r)^{\mathrm{T}} - ((\alpha_i^b)^{\mathrm{T}}\beta_j^r)I_{3\times3} \end{bmatrix} - (\cos\theta_{ij})I_{4\times4}$ and $\Xi = \begin{bmatrix} \eta_d, & -\xi_d^{\mathrm{T}} \\ \xi_d, & \eta_d I_{3\times3} + S(\xi_d) \end{bmatrix}$. Note that the relationship $q_b = q_d \otimes q_e$ is employed in (6).

To encode the information of all attitude constraints, consider the following barrier function:

$$V_a = -\sum_{i}^{k}\sum_{j}^{l}\gamma_{ij}\|q_e - q_I\|^2 \ln[-\Omega_{ij}(q_e)/2] \tag{7}$$

where $\Omega_{ij}(q_e) = q_e^{\mathrm{T}}\Xi^{\mathrm{T}}(q_d)M_{ij}\Xi(q_d)q_e$, and γ_{ij} are non-negative constant gains, $i = 1, 2, ..., k$, $j = 1, 2, ..., l$. Here k and l denote the total number of sensitive loads and corresponding constraint zones, respectively. Then we define $\mathfrak{D} = \{q_e \in \mathbb{S}^4 \mid \Omega_{ij}(q_e) < 0, i = 1, 2, .., k, j = 1, 2, ..., l\}$ to be the attitude admissible domain, and one can readily verify that $V_a \geq 0$ for all $q_e \in \mathfrak{D}$. Furthermore, $V_a = 0$ when $q_e = q_I$ (the desired attitude should always be in the admissible domain), and $V_a \to +\infty$ when any $\Omega_{ij}(q_e)$ goes to zero.

Remark 1: It is noteworthy that, M_{ij} in (6) a constant matrix. This is distinct from the relevant results in [9] and [10]; and it is achieved through fully utilizing the invariance property of α_i^b and β_j^r. Moreover, this property can greatly benefit the design of the subsequent approximate optimal controller, since it renders V_a to be a function with the unique argument q_e.

C. Angular Velocity Constraints

Per safety concerns, the angular velocities of on-orbit satellites are usually restricted, formalized by

$$\|\omega_{bi}^b\| \leq \omega_{\max,i} \tag{8}$$

where ω_{bi}^b is the i^{th} entry of ω_b^b, and $\omega_{\max,i} > 0$ is the maximum acceptable angular velocity, $i = 1, 2, 3$. A barrier function as follows is designed to encode this constraint.

$$V_\omega = -\gamma_\omega \sum_{i=1}^{3} \|\omega_{bi}^b\|^2 \ln\left(\frac{\omega_{\max,i}^2 - \omega_{bi}^2}{\omega_{\max,i}^2}\right) \tag{9}$$

where $\gamma_\omega > 0$ is the weight constant. We define $\mathfrak{F} = \{\omega_b^b \in \mathbb{R}^3 \mid \omega_{bi}^b < \omega_{\max\, i}, i = 1, 2, 3\}$ to be the admissible set of ω_b^b. Then it can be readily verified that for all $\omega_b^b \in \mathfrak{F}$, one has $V_\omega \geq 0$ (the equation holds when $\omega_b^b = 0_{3\times1}$); and $V_\omega \to +\infty$ when $\omega_{bi}^b \to \omega_{\max,i}$, $i = 1, 2, 3$.

D. Problem Formulation and Analysis

To formalize the optimal control problem considered in this paper, first we re-organize the system model to be the following compact form based on (1), (2) and (3):

$$\dot{x} = f(x) + gu \tag{10}$$

where $x = [(q_e - q_I)^{\mathrm{T}}, (J\omega_b^b)^{\mathrm{T}}]$, and

$$f(x) = \begin{bmatrix} 0.5E(q_e)\omega_b^b \\ -S(\omega_b^b)(J\omega_b^b) \end{bmatrix},\ g = \begin{bmatrix} 0_{4\times3} \\ I_{3\times3} \end{bmatrix}. \tag{11}$$

Then, based on the barrier functions of state constraints, the control objective is to design a control policy u to render $\lim_{t\to\infty} x(t) = 0_{7\times1}$ (i.e., $\lim_{t\to\infty} q_e(t) = q_I$, and $\lim_{t\to\infty} \omega_b^b(t) = 0_{3\times1}$), while minimizing the following cost function:

$$V(x(t)) = \int_t^\infty h(x(\tau))\mathrm{d}\tau,\ h = r + V_a + V_\omega + u^{\mathrm{T}}Ru \tag{12}$$

where $r = (q_e - q_I)^{\mathrm{T}}Q_q(q_e - q_I) + (\omega_b^b)^{\mathrm{T}}Q_\omega\omega_b^b$, 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 diagonal constant matrices.

Assuming the optimal control policy $u^*(x)$ exists, and the corresponding optimal cost function $V^*(x)$ is $\mathcal{C}^1$. Then by taking time derivative for both sides of (12), one has

$$\begin{aligned} H(x, u^*, \nabla_x V^*) =& \nabla_x^{\mathrm{T}}V^*[f + gu^*] \\ &+ r + V_a + V_\omega + (u^*)^{\mathrm{T}}Ru^* \\ =& 0. \end{aligned} \tag{13}$$

It should be emphasized that (13) can be established for any admissible controllers and corresponding cost functions (not only for the optimal ones).

Then the closed-form of u^* can be deduced by taking partial differential for both sides of (13) with respect to u^*:

$$u^* = -\frac{1}{2}R^{-1}g^{\mathrm{T}}\nabla_x V^*. \tag{14}$$

Further introducing u^* back into Eq. (13) leads to the following HJB equation:

$$r + V_a + V_\omega + \nabla_x^{\mathrm{T}}V^* f - \frac{1}{4}\nabla_x^{\mathrm{T}}V^* g R^{-1} g^{\mathrm{T}}\nabla_x V^* = 0. \tag{15}$$

However, due to the high nonlinearity of the system model, it is intractable to analytically solve (15). In the following section, an RL-based online controller will be designed to approximate the optimal solution u^*.

III. Online RL-based Control Algorithm

Based on the Weierstrass approximation theorem [19], [20], a NN that contains a sufficient set of basis functions can reconstruct the optimal cost function V^*(x) for $x \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^7$ is a compact set. This can be formalized by

$$V^*(x) = W^{\mathrm{T}}\sigma(x) + \epsilon(x) \tag{16}$$

and here $\sigma(x) = [\sigma_1(x), \sigma_2(x), ..., \sigma_p(x)]^{\mathrm{T}} \in \mathbb{R}^{p\times 1}$ is the basis function vector, with $\sigma_i(x)$ satisfying $\sigma_i(0) = 0$ and $(\mathrm{d}\sigma_i(0))/(\mathrm{d}t) = 0$, $i = 1, 2, ..., p$. $W \in \mathbb{R}^{p\times 1}$ denotes the unknown constant weigh vector of basis functions, and $\epsilon(x) \in \mathbb{R}$ is the reconstruction error. Then the optimal control policy can be transformed to

$$u^*(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}[\nabla_x\sigma(x)W + \nabla_x\epsilon(x)]. \tag{17}$$

In RL-based control, usually two sets of estimates for W (i.e. critic and actor) are employed to respectively approximate the optimal cost function and control policy, formalized by

$$V(x, \hat{W}_c) = \hat{W}_c^{\mathrm{T}}\sigma(x) \tag{18}$$

$$u(x, \hat{W}_a) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}\nabla_x\sigma(x)\hat{W}_a \tag{19}$$

where $\hat{W}_c$ and $\hat{W}_a$ denote the weights of critic and actor, respectively.

Then considering the following Bellman error:

$$\delta_b = \nabla_x^{\mathrm{T}}V[f + gu] + h \tag{20}$$

and substituting (18) and (19) into (20) yields

$$\begin{aligned}\delta_b =&\delta_b - H(x, u^*, \nabla_x V^*)\\ =&\nabla_x^{\mathrm{T}}V[f + gu] + u^{\mathrm{T}}Ru - \nabla_x^{\mathrm{T}}V[f + gu^*] - (u^*)^{\mathrm{T}}Ru^*\\ =&\varpi^{\mathrm{T}}\tilde{W}_c + \epsilon_H\end{aligned} \tag{21}$$

where $\varpi = \nabla_x^{\mathrm{T}}\sigma(f + gu)$ and $D = \nabla_x^{\mathrm{T}}\sigma gR^{-1}g^{\mathrm{T}}\nabla_x\sigma$ are employed for ease of notation, and we denote $\tilde{W}_c = \hat{W}_c - W$ and $\tilde{W}_a = \hat{W}_a - W$. Besides, ϵ_H is a residual error defined same with [26], [29], [30]. Since δ_b has the information of the weight estimation errors, it has been commonly employed to design the update law of $\hat{W}$. In this paper, not only the real-time information of δ_b but also its past measurements are employed. Before presenting the specific design of our RL-based controller, some necessary definitions and assumptions are given as follows.

Definition 1 (Finite Excitation, FE) [24]: A bounded signal $y(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times m}$ is said to be finite exciting over an interval $[t, t + T]$, where $t \geq 0$ is a finite time index, if there exist constants $T > 0$ and $c > 0$ such that $\int_t^{t+T} y^{\mathrm{T}}(\tau)y(\tau)\mathrm{d}\tau \geq cI_{m\times m}$.

Definition 2 (Persistent Excitation, PE) [31]: A bounded signal $y(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times m}$ is said to be persistently exciting if there exist positive constants c and T such that for arbitrary $t \geq 0$, one has $\int_t^{t+T} y^{\mathrm{T}}(\tau)y(\tau)\mathrm{d}\tau \geq cI_{m\times m}$.

Assumption 1: Consider an auxiliary variable defined by $\phi = \varpi/(\varpi^{\mathrm{T}}\varpi + 1)$, it satisfies a FE condition, i.e., 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 2: For $x \in \mathcal{X}$, there exist positive constants b_σ, b_{∇_σ}, b_ϵ, b_{∇_ϵ} and b_{ϵ_H}, such that $\|\sigma\| \leq b_\sigma$, $\|\nabla\sigma\| \leq b_{\nabla_\sigma}$, $\|\epsilon\| \leq b_\epsilon$, $\|\nabla_x\epsilon\| \leq b_{\nabla_\epsilon}$ and $\|\epsilon_H/(\varpi^{\mathrm{T}}\varpi + 1)\| \leq b_{\epsilon_H}$.

Note that Assumption 1 is mild and it is much weaker than the conventional PE conditions that are required in many RL-based controllers such as [19], [20], [22]. Moreover, Assumption 2 is a standard assumption in relevant research.

Then we define the following auxiliary variable:

$$\Phi(t, t_{w2}, t_{w1}) = \phi_1(t_{w2}, t_{w1})\hat{W}_c(t) + \phi_2(t_{w2}, t_{w1}) \tag{22}$$

with

$$\dot{\phi}_1(t, t_{w1}) = -\kappa\phi_1(t, t_{w1}) + \varphi_1(t), \quad \phi_1(t_{w1}) = 0_{p\times p} \tag{23}$$

$$\dot{\phi}_2(t, t_{w1}) = -\kappa\phi_2(t, t_{w1}) + \varphi_2(t), \quad \phi_2(t_{w1}) = 0_{p\times 1} \tag{24}$$

and here $\varphi_1 = \phi\phi^{\mathrm{T}}$ and $\varphi_2 = h\phi/(\varpi^{\mathrm{T}}\varpi + 1)$ are employed for ease of notation, and κ is a positive constant. Substituting the solutions of (23) and (24) into the definition of Φ yields

$$\begin{aligned}\Phi(t, t_{w2}, t_{w1}) =& \int_{t_{w1}}^{t_{w2}} \mathrm{e}^{\kappa(\tau - t_{w2})}(\varphi_1(\tau)\hat{W}_c(t) + \varphi_2(\tau))\mathrm{d}\tau\\ =& \phi_1(t_{w2}, t_{w1})\tilde{W}_c(t) + \epsilon_\Phi.\end{aligned} \tag{25}$$

Notice $\phi_1(t_{w2}, t_{w1}) = \int_{t_{w1}}^{t_{w2}} \mathrm{e}^{-\kappa(t_{w2}-\tau)}\phi(\tau)\phi^{\mathrm{T}}(\tau)\mathrm{d}\tau$, it "stores" the information of ϕ throughout the time interval $[t_{w1}, t_{w2}]$. Under the assumption that ϕ satisfies an FE condition, we have $\phi_1(t_{w2}, t_{w1}) \geq c_\Phi I_{m\times m}$, and here $c_\Phi = e^{-\kappa(t_{w2}-t_{w1})}c_w$. Besides, $\epsilon_\Phi = \int_{t_{w1}}^{t_{w2}}(\mathrm{e}^{-\kappa(t_{w2}-\tau)}\phi\epsilon_H/(\varpi^{\mathrm{T}}\varpi + 1))\mathrm{d}\tau$ is the residual error vector, and it is a constant for a fixed time interval $[t_{w1}, t_{w2}]$.

The employment of Φ, ϕ_1, and ϕ_2 can greatly benefit the design of the update laws of $\hat{W}_c$ and $\hat{W}_a$. However, since ϕ_1 is positive-definite only after adequate data is collected online (i.e. the FE condition of ϕ must be satisfied), how to guarantee the boundedness and handle the state constraints during this data collection process should be addressed first. To this end and motivated by the results given in [20], [22], [32], a solution is given in following lemma.

Lemma 1: Consider the attitude model in (10), and the actor-critic architecture described in (18) and (19). Design the weight update laws (with satisfying Assumption 2) to be

$$\dot{\hat{W}}_c = -c\frac{\varpi(t)\delta_b(t)}{(\varpi^{\mathrm{T}}(t)\varpi(t) + 1)^2} \tag{26}$$

$$\dot{\hat{W}}_a = -(a_1\hat{W}_a - a_2\phi\phi^{\mathrm{T}}\hat{W}_c) \tag{27}$$

where c, a_1, and a_2 are positive constants. Then for all $q_e(0) \in \mathfrak{D}$ and $\omega_b^b(0) \in \mathfrak{F}$, one has q_e, ω_b^b, $\tilde{W}_c$ and $\tilde{W}_a$ are ultimately bounded. Besides, the barrier functions V_a and V_ω are also bounded.

Proof: See Appendix A.

Lemma 1 guarantees an admissible online data collection process. Then, once adequate data is collected (i.e. ϕ satisfies an FE condition), the following critic-only structure can replace the actor-critic controller given in Lemma 1.

Theorem 1: Consider the attitude model in (10) and the RL-based control framework described in (18) and (19). Under Assumptions 1 and 2, design the weight update laws to be

$$\dot{\hat{W}}_c(t) = -c_1 \frac{\varpi(t)\delta_b(t)}{(\varpi^{\mathrm{T}}(t)\varpi(t)+1)^2} - c_2\Phi(t,t_{w2},t_{w1}) \quad (28)$$

$$\hat{W}_a(t) = \hat{W}_c(t) \quad (29)$$

where c_1 and c_2 are positive constants. Then, for all $q_e(0) \in \mathfrak{D}$ and $\omega_b^b(0) \in \mathfrak{F}$, one has q_e, ω_b^b, and $\tilde{W}$ are UB, and $V_a, V_\omega \in \mathcal{L}_\infty$.

Proof: Consider the following storage function for the critic-only architecture proposed in Theorem 1:

$$L = V^* + \frac{\rho_c}{2}\tilde{W}_c^{\mathrm{T}}\tilde{W}_c \quad (30)$$

where the positive constant ρ_c is employed just for analysis purpose. Then substituting (13), (19), and (28) into the time derivative of L yields

$$\begin{aligned}
\dot{L} =& \nabla_x^{\mathrm{T}}V^*(f+gu) + \rho_c\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c \\
=& \nabla_x^{\mathrm{T}}V^*(f+gu^*) + \rho_c\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c + \nabla_x^{\mathrm{T}}V^*(gu-gu^*) \\
=& \nabla_x^{\mathrm{T}}V^*(f+gu^*) + \rho_c\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c \\
& - \frac{1}{2}(W^{\mathrm{T}}\nabla_x^{\mathrm{T}}\sigma + \nabla_x^{\mathrm{T}}\epsilon)gR^{-1}g^{\mathrm{T}}(\nabla_x\sigma\tilde{W}_c - \nabla_x\epsilon) \quad (31) \\
\le& -r - V_a - V_\omega + \frac{1}{2}\tilde{W}_c^{\mathrm{T}}D\tilde{W}_c + \rho_c\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c + \epsilon_{L1} \\
\le& -(q_e-q_I)^{\mathrm{T}}Q_q(q_e-q_I) - (\omega_b^b)^{\mathrm{T}}Q_\omega\omega_b^b - V_a \\
& - V_\omega - \tilde{W}_c^{\mathrm{T}}C\tilde{W}_c + \epsilon_{L1} + \epsilon_{L2}
\end{aligned}$$

where $C = 0.5[\rho_c c_1\phi\phi^{\mathrm{T}} + \rho_c c_2 c_\Phi I_{p\times p} - D]$, $\epsilon_{L1} = 0.5\nabla_x^{\mathrm{T}}\epsilon g R^{-1}g^{\mathrm{T}}\nabla_x\epsilon$, and $\epsilon_{L2} = 0.5\rho_c c_1\|\epsilon_H\|^2/(\varpi^{\mathrm{T}}\varpi+1)^2 + 0.5\rho_c c_2\|\epsilon_\Phi\|^2/c_\Phi$. By Assumption 2, one has e_{L1}, e_{L2}, and $\|D\|$ are bounded. Thus one has $C > 0$ by setting $\rho_c > b_D/(c_2 c_\Phi)$, where b_D is the upper bound of $\|D\|$. On this basis, the result in (31) indicates $q_e - q_I$, ω_b^b, and $\tilde{W}_c$ are uniformly bounded, and $V_a, V_\omega \in \mathcal{L}_\infty$. The proof is complete.

Remark 2: Conventional RL-based controllers, especially online actor-critic architectures (e.g., [20], [32], [22]), usually require PE conditions to ensure the convergence of the estimation errors of NN weights. This is also indicated by the result given in Lemma 1. The term $-\tilde{W}_c^{\mathrm{T}}\phi\phi^{\mathrm{T}}\tilde{W}_c$ in the proof of Lemma 1 implies that $\tilde{W}_c$ could converge to zero only if ϕ satisfies the PE condition. However, the PE condition is typically strong, and it is impractical for the attitude reorientation problems considered in this paper, given the fact that $\phi \to 0_{p\times 1}$ when $x \to 0_{7\times 1}$. Therefore, though the actor-critic architecture in Lemma 1 can meet the boundedness requirement, it may lead to severe performance degradation. To address this issue, a novel critic-only learning architecture is proposed in Theorem 1. By employing a specially designed information matrix Φ, both real-time data and past measurements are employed in the controller. This design not only simplifies the whole control structure, but also relaxes the excitation condition (for convergence) from PE to FE.

Remark 3: It should be emphasized that the idea of employing past measurements to relax the PE condition is inspired by the concurrent learning method [24], [25] and its extensions/applications [26], [27] in the RL community. But these elegant results lack the constraint handling abilities. This essential problem is addressed in our paper by designing special barrier functions. Besides, in these elegant results, discrete historical data stacks are employed to collect past measurements: $\sum_{i=1}^{l}\phi(i)\delta_b(i)/(\varpi^{\mathrm{T}}(i)\varpi(i)+1)$, where $\delta_b(i)$ denotes the Bellman error given in (21) at a past (discrete) time point i while replacing $\hat{W}(i)$ with its real-time counterpart. However, this design needs to employ complicated algorithms for data selection purposes, and the complexity of such algorithms grows significantly with the increase of the total number of basis functions. In contrast, the integral-form data collecting structure in (22) to (24) can take full use of the incoming data, which is arguably easier to implement and computationally much cheaper.

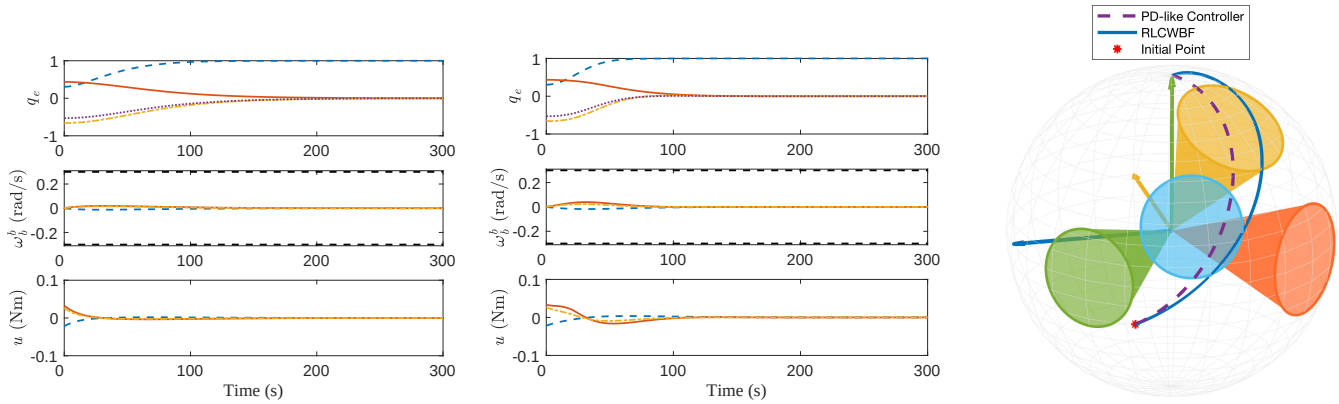
Remark 4: In Theorem 1, we show that system states converge to a residual set whose size is related to ϵ_{L1} and ϵ_{L2}. We want to emphasize that this residual set shrink significantly with the convergence of x. Besides, to potentially reduce the residual error further, the information matrix $\Phi(t,t_{w2},t_{w1})$ can be updated to $\Phi(t,\bar{t}_{w2},\bar{t}_{w1})$ if ϕ also satisfies the FE condition on a new time interval $[\bar{t}_{w1},\bar{t}_{w2}]$. This is because the convergence of x and $\tilde{W}_c$ can potentially render a smaller $\|\epsilon_\Phi\|$ on the new interval $[\bar{t}_{w1},\bar{t}_{w2}]$.

IV. Numerical Simulations and Hardware-in-Loop Experiments

A. Numerical Simulation Results

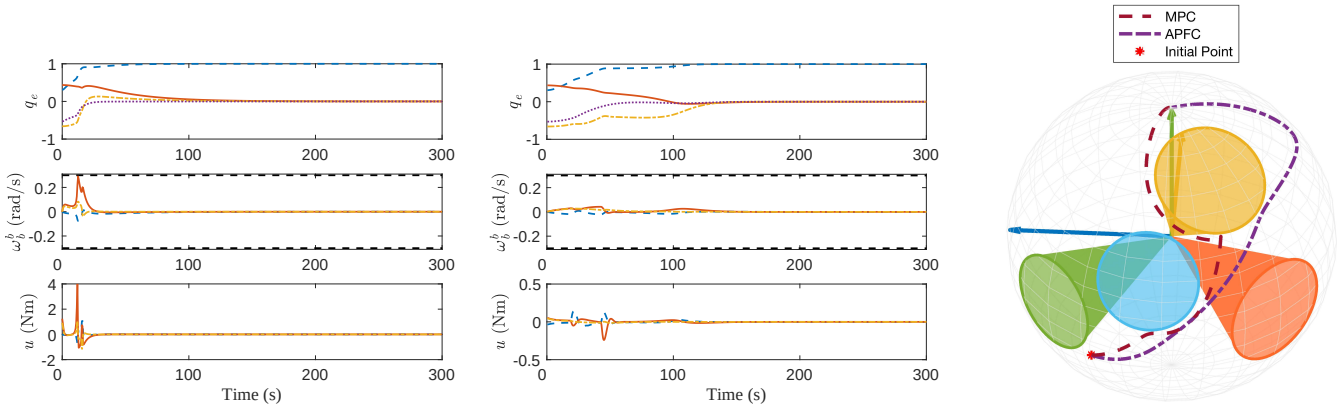
Numerical simulation results are given in this section to show the effectiveness and advantages of the proposed method. Consider a rigid body with $J = [20,0,0;0,17,0;0,0,15]$ kg·m^2. Its initial attitude is $q_b(0) = [0.3062, 0.4356, -0.6597, -0.5303]^{\mathrm{T}}$ with $\omega_b^b(0) = 0_{3\times 1}$ rad/s. Assuming a sensitive payload is fixed on the rigid body, and its boresight vector coincides with the z-axis of the frame $\mathcal{F}_b$, thus $\alpha_1^b = [0,0,1]^{\mathrm{T}}$. The control objective is to reorient the frame $\mathcal{F}_b$ to be consistent with $\mathcal{F}_r$, so we have $q_d = q_I$ and $q_e = q_b$.

During the reorientation process, there are four attitude constraint zones that are required to be avoided: 1) $\beta_{11} = [-0.9245, 0.0925, 0.3698]^{\mathrm{T}}$, $\theta_{11} = 18^\circ$; 2) $\beta_{12} = [-0.4602, -0.2761, 0.8438]^{\mathrm{T}}$, $\theta_{12} = 20^\circ$, 3) $\beta_{13} = [-0.7071, -0.7071, 0]^{\mathrm{T}}$, $\theta_{13} = 20^\circ$, and 4) $\beta_{14} = [-0.7071, 0.7071, 0]^{\mathrm{T}}$, $\theta_{14} = 18^\circ$. Please note β_{11} to β_{14} are expressed in $\mathcal{F}_r$. Besides, the maximum angular velocity is set to be $\omega_{\max,i} = 0.3$ rad/s, $i = 1,2,3$. The cost function follows: $\gamma_{11} = 0.4$, $\gamma_{12} = 0.6$, $\gamma_{13} = 0.2$, $\gamma_{14} = 0.2$, $\gamma_\omega = 10$, $Q_q = I_{4\times 4}$, $Q_\omega = 10I_{3\times 3}$, and $R = 20I_{3\times 3}$. A PD-like controller with $k_p = 0.05$ and $k_d = 1.5$ is employed as the initial control policy for the RL-based controller (denoted by RLC) proposed in this paper. To straightforwardly show the learning ability of RLC, we choose $\sigma = [\xi_{e1}\omega_{b1}^b, \xi_{e2}\omega_{b2}^b, \xi_{e3}\omega_{b3}^b, (\omega_{b1}^b)^2, (\omega_{b2}^b)^2, (\omega_{b3}^b)^2]^{\mathrm{T}}$. Thus the proposed method can keep the same structure with the PD-like control method, helping to illustrate whether it can bring a conventional controller the essential optimizing and constraint handling abilities in an online pattern. Accordingly,



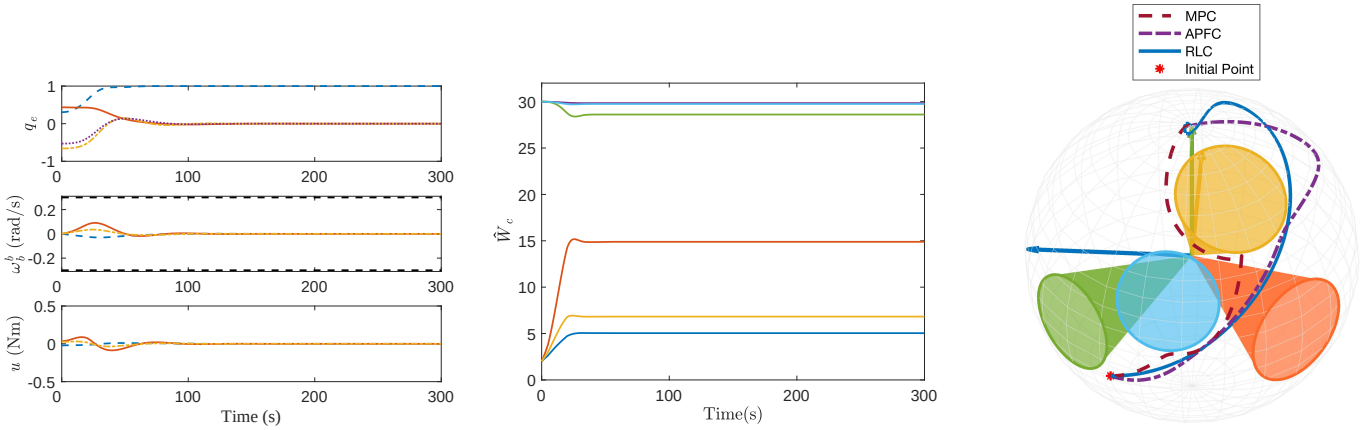
(a) Simulation results of the PD-like controller (b) Simulation results of the RLCWBF method (c) 3D illustration 1

Figure 2: Simulation results of the PD-like controller and the RLCWBF method



(a) Simulation results of the APFC (b) Simulation results of the MPC (c) 3D illustration 2

Figure 3: Simulation results of the APFC and the MPC



(a) Time response of system states under the RLC (b) Simulation results of $\hat{W}_c$ (c) 3D illustration 3

Figure 4: Simulation results of RLC

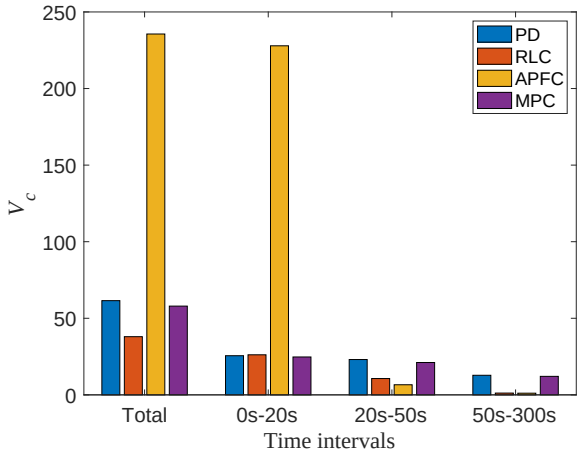


Figure 5: Simulation results of V_c under different controllers

one has $\hat{W}_c(0) = \hat{W}_a(0) = [2, 2, 2, 30, 30, 30]^{\mathrm{T}}$, and other control gains are set to be: $\kappa = 0.1$, $a_1 = 0.05$, $a_2 = 0.1$, $c_1 = 3$, and $c_2 = 0.3$. Besides, under the actor-critic structure in Lemma 1, ϕ_1 becomes full-rank after 3.1s. Based on that, the data in the first 5s ($t_{w1} = 0$ and $t_{w2} = 5$) is employed to build Φ, and this matrix is released (set Φ to be a zero matrix) later to reduce the residual error.

For comparison purposes, some other attitude reorientation controllers are also employed in simulations: 1) The AP-based feedback controller in [10] (denoted by APFC), which also considers both attitude constraints and angular-velocity limits. Its parameters are set to be: $k = 0.05$, $\alpha = 1/30$, $k_1 = 2\Upsilon^{-1}$, and $k_2 = 0.004$. 2) An RL-based controller without barrier functions (denoted by RLCWBF). This controller keeps the same structures and parameters as the proposed method, while no barrier functions are introduced into the cost function. It is employed to show the effectiveness of barrier functions and illustrate the constraint handling abilities of RLC. 3) The PD-like controller: $u = -k_p\xi_e - k_d\omega_b^b$ with $k_p = 0.05$ and $k_d = 1.5$. This is the initial control policy of RLC. It is employed to illustrate the learning ability of the RLC method proposed in this paper. 4) The nonlinear MPC controller in [14], which also has optimizing and constraint handling abilites. Please refer to [14] for the specific design of this controller. The time step of MPC is set to be 0.4s, and other parameters include: $P_1 = Q_1 = 4I_{3\times3}$, $P_2 = Q_2 = 0.25I_{3\times3}$, and $Q_3 = I_{3\times3}$.

Based on all these settings, time responses of system states under the PD-like controller and RLCWBF are illustrated in Figs. 2a and 2b, respectively. Due to the online learning ability of RLCWBF, it has an improved closed-loop performance and faster convergence process compared with the PD-like controller. Besides, a three-dimensional (3D) illustration is provided in Fig. 2c to show the reorientation trajectories of the boresight vector $\alpha_1^b = [0, 0, 1]^{\mathrm{T}}$ in the frame $\mathcal{F}_r$. In this figure, the axes of $\mathcal{F}_r$ are denoted by mutually perpendicular lines, and the attitude forbidden zones are illustrated by cones. One can see that though the PD-like controller and RLCWBF can achieve the reorientation objective, they both violate the attitude constraints. Distinct with PD and RLCWBF, the APFC in [10] and the MPC in [14] avoid the violation of all underlying state constraints during the whole reorientation process, as illustrated in Figs. 3a to 3c. Please note that the bounds of ω_b^b are indicated by dashed lines. Then, simulation results of RLC are given in Fig. 4. Specifically, the time responses of q_e, ω_b^b, and u are in Fig. 4a; the time response of $\hat{W}_c$ is in Fig. 4b; and a 3D illustration which contains the trajectories of not only RLC but also APFC and MPC is provided in Fig. 4c. From Fig. 4c, one can see that RLC, APFC, and MPC can achieve control objectives while obeying all underlying constraints. Besides, RLC has smoother trajectories and less fluctuations than MPC and APFC.

To quantitatively compare the performance of different controllers, we define a new cost function: $V_c = (q_e - q_I)^{\mathrm{T}} Q_q (q_e - q_I) + (\omega_b^b)^{\mathrm{T}} Q_\omega \omega_b^b + u^{\mathrm{T}} R u$, which can show the overall cost of different controllers without considering state constraints. The simulation results of V_c under PD, RLC, APFC, and MPC are given in Fig. 5. One can see that the proposed RLC method shows effective optimizing abilities. It has a significant lower cost when compared with APFC (reduced by 84%), PD (reduced by 38%), and also MPC (reduced by 34%).

The computational complexity of different controllers is also worth to be analyzed. For the simulation scenario here, RLC only needs 4.096466 seconds to complete a simulation of 300 seconds (under a computer with Intel Core i7-8565U CUP @ 1.80GHz, 16GB RAM). In contrast, MPC needs to numerically solve a sub-optimal control problem at every time step (every 0.4s), which renders a much higher computational complexity. It requires 279.1197s to finish the 300-second simulation under the same computation conditions with RLC. Besides, the simulation time of PD and APFC is 0.972631s and 2.862144s, respectively. These results indicate that though RLC renders a higher computational complexity than conventional feedback controllers (PD and APFC), the additional cost is acceptable and much less than some other sub-optimal controllers (MPC).

B. Hardware-in-Loop Experimental Results

To test the performance of the proposed RL-based controller under disturbances and noisy measurements, hardware-in-loop (HL) experiment is conducted in this subsection. The HL testbed is demonstrated in Fig. 6, which contains: 1) A triaxial turntable. It can simulate the attitude motion of spacecraft in real time; a raster and a gyroscope are equipped with the turntable to respectively measure the Euler angles and angular velocities, providing state measurements (q_b and ω_b^b, where Euler angles are transferred to unit quaternions) to the controller. 2) A high-reliability real-time simulation computer (HRSC). It collects the sensor outputs and calculates the control command signal u. Then a control allocation algorithm is carried out based on the the configuration of reaction wheels; the resulting control command ($\hat{\tau}_{cmd}$) is transferred to the underlying controller. 3) An ARM-based underlying control PCB. It receives the control command from HRSC, then implements these signals (τ_{cmd}) to reaction wheels via RS-422. 4) Four reaction wheels, serving as actuator simulators. They provide the final control signal τ_c (i.e., u in Eq. (2)) to the simulated attitude dynamic in the simulation computer. Moreover, a weighted pseudo-inverse algorithm [33] is employed for control allocation.

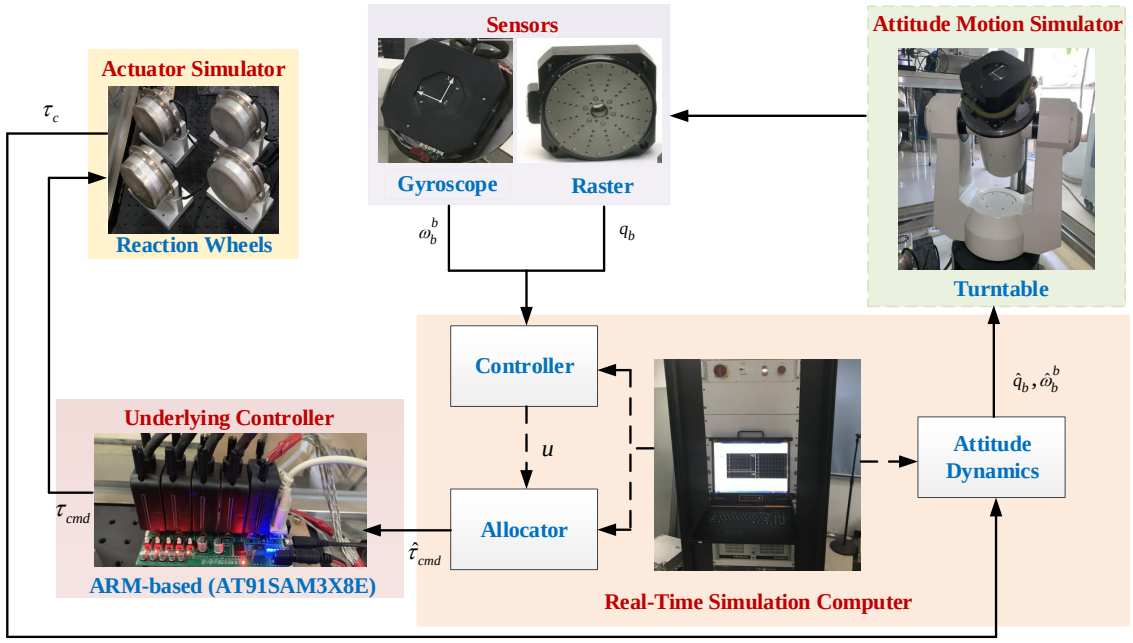
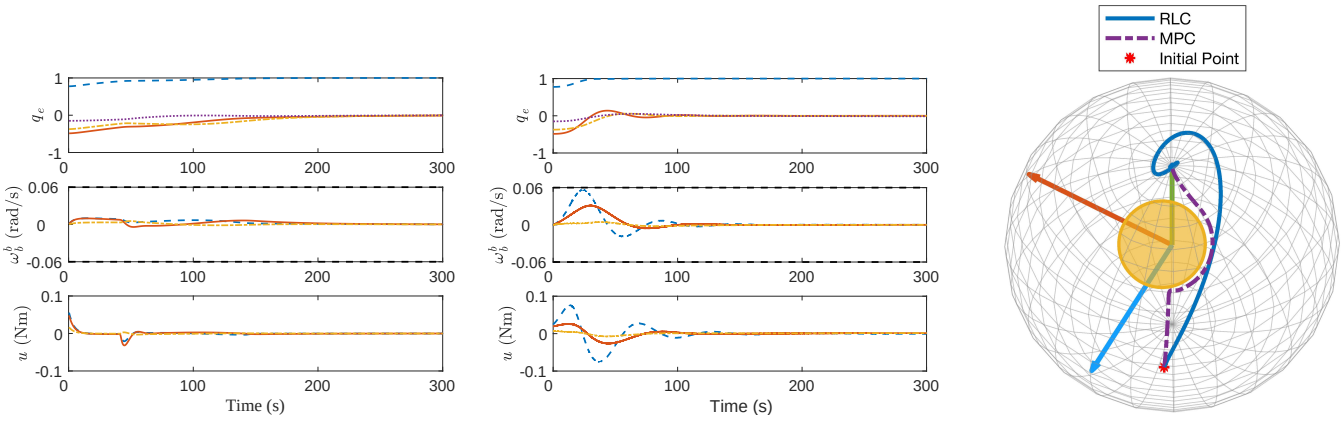


Figure 6: Framework of the hardware-in-loop testbed



(a) Experimental results of MPC (b) Experimental results of RLC (c) 3D illustration 4

Figure 7: Experimental results of MPC and RLC

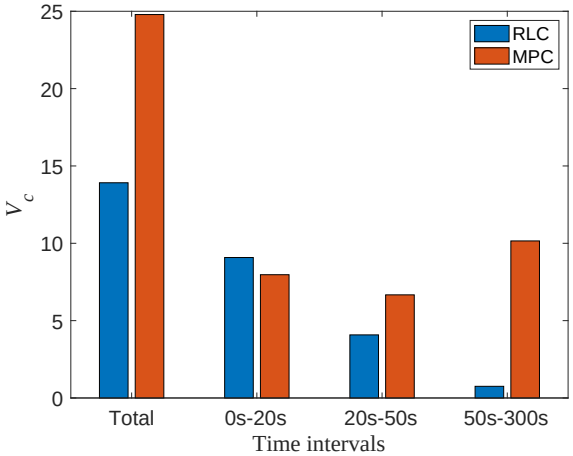


Figure 8: Experiment results of V_c under RLC and MPC

The time step of the testbed is set to be 0.05s. By employing this HL testbed, practical measurement noise of states and disturbances of control signals are inevitably introduced into the closed-loop system. Besides, the turntable has two main physical restrictions: 1) the maximum output of the reaction wheels is 0.1Nm with slopes no larger than 0.01Nm/s; 2) the rotational degree-of-freedom of the Y-axis is $[-90°, 90°]$ (it has singularities at $\pm 90°$ on this direction). Given these restrictions, the simulation scenario in Sec. IV-A is modified as follows: 1) For safety concerns, the allowable maximum angular velocity is reduced to be $\omega_{\max,i} = 0.06$rad/s, $i = 1, 2, 3$; 2) To reduce the computational complexity of MPC, only a single attitude forbidden zone is considered, with $\beta_{11} = [-0.1761, 0.4402, 0.8805]^{\mathrm{T}}$, $\theta_{11} = 15°$, and $\gamma_{11} = 6$; 3) The initial attitude is changed to be $q_b(0) = [0.7762, -0.4858, -0.3728, -0.15]^{\mathrm{T}}$. All the other settings are

kept same with Sec. IV-A.

The two sub-optimal controllers that have constraint handling abilities, i.e. MPC and RLC, are employed to carry out experiments; the results are given in Figs. 7a to 7c. One can see that, even in the presence of disturbances and noisy measurements, RLC and MPC can still achieve the reorientation goal with high precision. For this experimental case, MPC renders a shorter convergence trajectory (see Fig. 7c) but a longer convergence time (see Figs. 7a and 7b). Moreover, the experiment results of V_c under RLC and MPC are given in Fig. 8. One can see that RLC still has a better performance and renders a lower overall cost.

In summary, simulation and experiment results show that the RLC method proposed in this paper has the ability to make a trade-off between performance and control cost and to handle underlying state constraints. It can address the drawbacks of MPC and APFC while keeping their advantages. Table I summaries the features of APFC, MPC, and RLC, based on both theoretical analysis and simulation & experiment results. But a limitation of the RLC in the present paper is the lack of control constraint handling abilities. This issue will be explored in the future research.

Table I: Features of Different Controllers

	APFC in [10]	MPC in [14]	RLC in this paper
State Constraint	$\surd$	$\surd$	$\surd$
Online Complexity	$\surd$ (Low)	$\times$ (High)	$\surd$ (Low)
Optimizing Ability	$\times$	$\surd$	$\surd$
Control Constraint	$\times$	$\surd$	$\times$

V. Conclusions

A reinforcement-learning based controller for attitude reorientation control problems under multiple state constraints was proposed in this paper. Specially designed barrier functions were introduced into the cost function to encode the information of attitude forbidden zones and angular-velocity limits. Then an online RL-based control scheme was developed to ensure the convergence of NN weights and reorientation errors, with guaranteed constraint handling abilities during the whole online learning process. Besides, a simplified critic-only neural network was designed to replace the conventional actor-critic structure once adequate data was collected online. Both numerical simulations and experimental tests verified the effectiveness and advantages of the proposed method. Future work in this direction will consider system uncertainties and control constraints.

Appendix

A. Proof of Lemma 1

Consider the following storage function:

$$L_{ac} = V^* + \frac{\rho_1}{2}\tilde{W}_c^{\mathrm{T}}\tilde{W}_c + \frac{\rho_2}{2}\tilde{W}_a^{\mathrm{T}}\tilde{W}_a \tag{32}$$

where $\rho_1, \rho_2 > 0$ are constants which are employed just for analysis purpose. Substituting (13) into the time derivative of L_{ac} and employing the arithmetic-geometric average inequality (AGAI), one has

$$\begin{aligned}\dot{L}_{ac} =&\nabla_x^{\mathrm{T}} V^*(f+gu) + \rho_1\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c + \rho_2\tilde{W}_a^{\mathrm{T}}\dot{\tilde{W}}_a\\ =&\nabla_x^{\mathrm{T}} V^*(f+gu^*) + \rho_1\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c + \rho_2\tilde{W}_a^{\mathrm{T}}\dot{\tilde{W}}_a\\ &-\frac{1}{2}(W^{\mathrm{T}}\nabla_x^{\mathrm{T}}\sigma + \nabla_x^{\mathrm{T}}\epsilon)gR^{-1}g^{\mathrm{T}}(\nabla_x\sigma\tilde{W}_a - \nabla_x\epsilon)\\ \leq& -r - V_a - V_\omega + \frac{1}{2}\tilde{W}_a^{\mathrm{T}}D\tilde{W}_a + \epsilon_{L1}\\ &+\rho_1\tilde{W}_c^{\mathrm{T}}\dot{\tilde{W}}_c + \rho_2\tilde{W}_a^{\mathrm{T}}\dot{\tilde{W}}_a\end{aligned} \tag{33}$$

where $\epsilon_{L1} = 0.5\nabla_x^{\mathrm{T}}\epsilon gR^{-1}g^{\mathrm{T}}\nabla_x\epsilon$. Substituting (21), (26) and (27) into (33) and employing AGAI again yields

$$\begin{aligned}\dot{L}_{ac} \leq& -r - V_a - V_\omega + \frac{1}{2}\tilde{W}_a^{\mathrm{T}}D\tilde{W}_a + \epsilon_{L1}\\ &-\frac{\rho_1 c}{2}\tilde{W}_c^{\mathrm{T}}\phi\phi^{\mathrm{T}}\tilde{W}_c + \frac{\rho_1 c\|\epsilon_H\|^2}{2(\varpi^{\mathrm{T}}\varpi+1)^2}\\ &-\rho_2\tilde{W}_a^{\mathrm{T}}[a_1(\tilde{W}_a+W) - a_2\phi\phi^{\mathrm{T}}(\tilde{W}_c+W)]\\ \leq& -(q_e-q_I)^{\mathrm{T}}Q_q(q_e-q_I) - (\omega_b^b)^{\mathrm{T}}Q_\omega\omega_b^b - V_a\\ &-V_\omega - \tilde{W}_c^{\mathrm{T}}\phi\Lambda_c\phi^{\mathrm{T}}\tilde{W}_c - \tilde{W}_a^{\mathrm{T}}\Lambda_a\tilde{W}_a + \epsilon_a\end{aligned} \tag{34}$$

where $\Lambda_c = (1/2)\rho_1 c - 2\rho_2 a_2$, $\Lambda_a = [(1/4)\rho_2 a_1 - (1/2)D]$, and $\epsilon_a = \rho_1 c\|\epsilon_H\|^2/(2(\varpi^{\mathrm{T}}\varpi+1)^2) + (1/2)\rho_2 a_1 W^{\mathrm{T}}W + 2\rho_2 a_2 W^{\mathrm{T}}W + \epsilon_{L1}$. Please note the fact $\|\phi\| < 1$ is employed in (34). Thus, one has $\Lambda_a, \Lambda_c > 0$ by setting

$$\rho_1 > 4\rho_2 a_2/c,\ \rho_2 > 2b_D/a_1$$

and here $b_D = \max_{t\geq 0}\{D(t)\}$, which is bounded under Assumption 2. On this basis, Eq. (34) indicates that $q_e - q_I, \omega_b^b, \tilde{W}_c, \tilde{W}_a, V_a, V_\omega \in \mathcal{L}_\infty$. This completes the proof.

References

[1] M. Lovera and A. Astolfi, "Spacecraft attitude control using magnetic actuators," *Automatica*, vol. 40, no. 8, pp. 1405–1414, 2004.

[2] Q. Shen, D. Wang, S. Zhu, and K. Poh, "Finite-time fault-tolerant attitude stabilization for spacecraft with actuator saturation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 3, pp. 2390–2405, 2015.

[3] H. Cai and J. Huang, "Leader-following attitude consensus of multiple rigid body systems by attitude feedback control," *Automatica*, vol. 69, pp. 87–92, 2016.

[4] H. Gui and G. Vukovich, "Finite-time output-feedback position and attitude tracking of a rigid body," *Automatica*, vol. 74, pp. 270–278, 2016.

[5] 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.

[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] 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.

[8] X. Gu and X. Tong, "Overview of China earth observation satellite programs [space agencies]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 3, no. 3, pp. 113–129, 2015.

[9] U. Lee and M. Mesbahi, "Feedback control for spacecraft reorientation under attitude constraints via convex potentials," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 4, pp. 2578–2592, 2014.

[10] Q. Shen, C. Yue, C. H. Goh, B. Wu, and D. Wang, "Rigid-body attitude stabilization with attitude and angular rate constraints," *Automatica*, vol. 90, pp. 157–163, 2018.

[11] B. Wie and J. Lu, "Feedback control logic for spacecraft eigenaxis rotations under slew rate and control constraints," *Journal of Guidance, Control, and Dynamics*, vol. 18, no. 6, pp. 1372–1379, 1995.
[12] E. L. de Angelis, F. Giulietti, and G. Avanzini, "Single-axis pointing of underactuated spacecraft in the presence of path constraints," *Journal of Guidance, Control, and Dynamics*, vol. 38, no. 1, pp. 143–147, 2014.
[13] R. Gupta, U. V. Kalabić, S. Di Cairano, A. M. Bloch, and I. V. Kolmanovsky, "Constrained spacecraft attitude control on $SO(3)$ using fast nonlinear model predictive control," in *2015 American Control Conference*. IEEE, 2015, pp. 2980–2986.
[14] D. Y. Lee, R. Gupta, U. V. Kalabić, S. Di Cairano, A. M. Bloch, J. W. Cutler, and I. V. Kolmanovsky, "Geometric mechanics based nonlinear model predictive spacecraft attitude control with reaction wheels," *Journal of Guidance, Control, and Dynamics*, vol. 40, no. 2, pp. 309–319, 2017.
[15] H. Dong, Q. Hu, and M. R. Akella, "Dual-quaternion-based spacecraft autonomous rendezvous and docking under six-degree-of-freedom motion constraints," *Journal of Guidance, Control, and Dynamics*, vol. 41, no. 5, pp. 1150–1162, 2017.
[16] H. Dong, Q. Hu, Y. Liu, and M. R. Akella, "Adaptive pose tracking control for spacecraft proximity operations under motion constraints," *Journal of Guidance, Control, and Dynamics*, Early Access, 2019.
[17] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming with Applications in Optimal Control*. Cham, Switzerland: Springer, 2017.
[18] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. New York, NY, USA: Wiley, 2017.
[19] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.
[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] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, 2014.
[22] 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.
[23] D. Görges, "Relations between model predictive control and reinforcement learning," in *2017 IFAC World Congress*. IFAC, 2017, pp. 4920–4928.
[24] G. Chowdhary, M. Mühlegg, and E. Johnson, "Exponential parameter and tracking error convergence guarantees for adaptive controllers without persistency of excitation," *International Journal of Control*, vol. 87, no. 8, pp. 1583–1603, 2014.
[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] 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.
[27] 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.
[28] H. Schaub and J. L. Junkins, *Analytical Mechanics of Space Systems*. Reston, USA: AIAA, 2018.
[29] 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.
[30] S. Xue, B. Luo, and D. Liu, "Event-triggered adaptive dynamic programming for zero-sum game of partially unknown continuous-time nonlinear systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018.
[31] P. Ioannou and J. Sun, *Robust Adaptive Control*. Upper Saddle River, USA: Prentice Hall, 1996.
[32] 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.
[33] 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. From 2015-2017, he was a joint Ph.D. student at the Cockrell School of Engineering, University of Texas at Austin, Texas, USA. His current research interests include reinforcement learning, deep learning, intelligent control, and adaptive control.

Xiaowei Zhao is Professor of Control Engineering and an EPSRC Fellow at the School of Engineering, University of Warwick, Coventry, UK. He obtained his 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.

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