ADP-Based Spacecraft Attitude Control under Actuator Misalignment and Pointing Constraints

Haoyang Yang, *Graduate Student Member, IEEE,* Qinglei Hu, *Senior Member, IEEE,* Hongyang Dong, and Xiaowei Zhao

***Abstract*—This paper is devoted to real-time optimal attitude reorientation control of rigid spacecraft control. Particularly, two typical practical problems —— actuator misalignment and forbidden pointing constraints are considered. Within the framework of adaptive dynamic programming (ADP), a novel constrained optimal attitude control scheme is proposed. In this design, a special reward function is developed to characterize the environment feedback and deal with the pointing constraints. Notably, a novel argument term is introduced to the reward function for overcoming the inevitable difficulty in actuator misalignment. By virtue of the Lyapunov stability theory, the ultimate boundedness of state error and the optimality of the proposed method can be guaranteed. Finally, the effectiveness and performance of the developed ADP-based controller are evaluated by not only numerical simulations but also experimental tests with a hardware-in-loop platform.**

***Index Terms*—Attitude control; actuator misalignment; pointing constraints; adaptive dynamic programming (ADP); reinforcement learning.**

I. Introduction

Attitude reorientation control is an essential technology for a broad range of space missions, which has drawn wide attention in the past several decades [1]. Various control approaches, such as adaptive control [2] and sliding mode control (SMC) [3], [4], have been designed to address attitude reorientation problems. In realistic applications, because of the increasing complexity of the on-orbit missions, attitude control policy not only requires to complete the basic reorientation, but also should consider other practical cases, such as control cost reduction, actuator mounting, and onboard instruments' safety.

From the viewpoint of the requirement about onboard instruments' safety, some on-board light-sensitive payloads must avoid direct exposure to bright light [5], leading to the spacecraft are prohibited from pointing to some forbidden areas during the attitude reorientation. The problem of attitude control with forbidden pointing constraints is well understood for the case without considering the cost and actuator mounting by employing APF-based methods [5]–[9]. However, it should be pointed out that such APF-based methods usually over-prioritize the constraint handling ability and can lead to unacceptable control consumption. Therefore, the ability to balance control consumption and performance is essential in spacecraft reorientation tasks. In this context, optimal control has emerged as a suitable choice, and some elegant works have given solutions based on the optimization [10], [11]. Theoretically speaking, optimal attitude control problems require solving the Hamilton-Jacobi-Bellman (HJB) equations subject to system dynamics and user-defined cost functions [12]. Due to the high complexity and nonlinearity of attitude dynamics, it is quite hard to obtain analytical solutions or quickly calculate numerical solutions of the HJB equations in practical applications, not to mention achieving other mission requirements.

Another practical issue that has drawn wide attention is actuator misalignment. In practice, actuator misalignment is inevitable due to the finite-manufacturing tolerance or warping of the spacecraft structure [13]. This problem will cause the deviation between the actual control torque and desired torque, resulting in the deterioration of control performance or even the failure of tasks. Previous research efforts have been carried out to deal with actuator misalignment, and most of them employed the SMC-based controller [13], [14]. However, these methods lack the optimizing ability and may lead to high control costs. Recently, Wang [15] proposed a robust optimal controller to address the optimal attitude reorientation control problem with actuator misalignment. However, this method is designed for special cases and cannot handle complex attitude control scenarios like attitude constraints.

It is noteworthy that designing optimal control strategies for attitude reorientation problems under complex motion constraints is still an open problem, and actuator misalignment makes the whole control problem even more challenging. Reinforcement learning (RL) is a promising technology to address such problems. In the control-related field, RL is also referred as the approximate/adaptive dynamic programming (ADP) [16], which is a powerful data-driven method for optimal control problems. The fundamental principle of ADP-based control is that it utilizes previous or online data to iteratively approximate the solution of HJB equation and accordingly improve control performance. ADP-based control has aroused significant research interests recently in spacecraft [17]–[19] and many other motion control systems [20]–[22]. However, most of the existing ADP-based methods neither consider actuator misalignment nor motion constraints.

Motivated by these facts, this work attempts to design a control scheme considering the alignment-error tolerating, pointing constraint, and control performance real-time optimization simultaneously. The contributions of this paper are listed in the following aspects:

This work was funded by the Natural Science Foundation of China under grants 61960206011, 61633003, and 61803012. H. Yang and Q. Hu (Corresponding Author) are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. Emails: {yanghaoyang8352, huql_buaa}@buaa.edu.cn; H. Dong and X. Zhao are with the Intelligent Control and Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry, CV4 7AL, UK. Emails: {hongyang.dong, xiaowei.zhao}@warwick.ac.uk.

1) A novelty ADP-based optimal controller is developed for a typical practical case that spacecraft orientation control under pointing constraint, actuator misalignment and external disturbances. The controller achieves approximately solving the HJB equation by online learning, which presents the superiority in real-time performance. To the authors' best knowledge, this is the first time to provide a real-time optimal policy for the constrained orientation control in the presence of actuator misalignment and external disturbances.

2) Aiming at the actuator misalignment and external disturbances in practical applications, a special augmented term is designed in this paper, which relaxes the traditional precondition that the uncertainties need to converge with the state (e.g., assumed in [15]). Then the reward function embedded the augmented term to characterize environmental feedback.

3) The ultimately uniformly bounded stability of the whole system is derived based on the Lyapunov analysis, which also guarantees the optimality of the control scheme and compliance with constraints. For further evaluating the effectiveness in practical applications of the proposed method, a representative hardware-in-loop experimental validation is presented on a semi-physical experimental platform.

The rest of this article is organized as follows. The attitude dynamics, actuator misalignment, and motion constraints are introduced in Sec. II. Then, the ADP-based control scheme is designed in Sec. III. Numerical simulation and experimental validations are presented in Sec. IV and Sec. V, respectively. Finally, this article ends with some concluding remarks in Sec. VI.

Notation: Throughout the paper, $\mathbb{R}^{n\times m}$ denotes the set of $n \times m$ real matrix. Post superscript $(\cdot)^\times$ denotes the skew-symmetric matrices of three dimensional vectors. The 2-norm and infinity norm of vectors or matrices are presented by $\|\cdot\|$ and $\|\cdot\|_\infty$, respectively. The n-dimensional identity matrix represented as $\boldsymbol{I}_n$. Let $\mathbf{0}_{n\times m}$ be $n\times m$ zero matrix, and $\mathbf{1}_{n\times m}$ denotes $n \times m$ one matrix.

II. Preliminaries and Problem Formulation

In this section, a brief background of the rigid spacecraft dynamics, state constraints, and actuator misalignment will be discussed.

A. Attitude Kinematics and Dynamics

Let $\mathcal{F}_I = \{X_I, Y_I, Z_I\}$ and $\mathcal{F}_B = \{X_B, Y_B, Z_B\}$ denote the inertial frame and the body-fixed frame, respectively. The rotation of the $\mathcal{F}_B$ with respect to the $\mathcal{F}_I$ is represented in the form of MRPs. Then the kinematic equation in the term of MRPs is given as [23]:

$$\dot{\sigma} = H(\sigma)\omega, \text{ with } \quad H(\sigma) \triangleq \frac{1+\sigma^T\sigma}{4}I_3 - \frac{1}{2}\sigma^\times + \frac{1}{2}\sigma^\times\sigma^\times \tag{1}$$

where the MRPs is defined as $\sigma = q_v/(1+q_0) \in \mathbb{R}^3$ with $q = \left[q_0, q_v^T\right]^T \in \mathbb{R}^4$ representing spacecraft quaternions. The angular velocity is denoted by $\omega \in \mathbb{R}^3$. If all the actuators are assembled in the ideal location, the rigid spacecraft dynamics can be given as:

$$J\dot{\omega} = \omega^\times J\omega + \tau \tag{2}$$

where $J \in \mathbb{R}^{3\times 3}$ represents the total inertia matrix of the spacecraft. The control torque is denoted by $\tau \in \mathbb{R}^3$.

B. Actuator Misalignment

The spacecraft is actuated by three orthogonally mounted reaction wheels, which are aligned with the X_B,Y_B, and Z_B of $\mathcal{F}_B$. However, the actuators' alignment error is inevitable due to the warping of spacecraft structure or the imperfection during manufacturing and assembling. As shown in Fig. 1, with alignment errors, the real torque applied on the spacecraft is expressed as [15]:

$$\tau = \Lambda\tau_c + d \tag{3}$$

with

$$\Lambda = \begin{bmatrix} \cos\Delta\alpha_1 & \sin\Delta\alpha_2\cos\Delta\beta_2 & \sin\Delta\alpha_3\cos\Delta\beta_3 \\ \sin\Delta\alpha_1\cos\Delta\beta_1 & \cos\Delta\alpha_2 & \sin\Delta\alpha_3\sin\Delta\beta_3 \\ \sin\Delta\alpha_1\sin\Delta\beta_1 & \sin\Delta\alpha_2\sin\Delta\beta_2 & \cos\Delta\alpha_3 \end{bmatrix} \tag{4}$$

where τ_c is the control torque generated by reaction wheels, $\Delta\alpha_i$ and $\Delta\beta_i (i \in \{1,2,3\})$ are the deviation angles away from their nominal directions. The disturbance torque applied to the spacecraft is denoted by d.

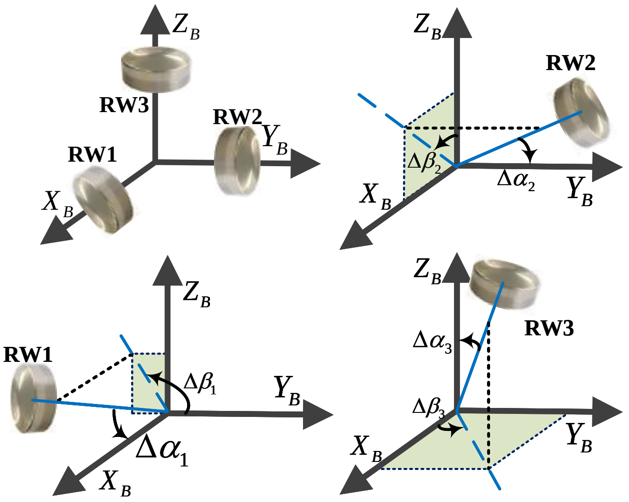


Fig. 1. Actuator misalignment illustration.

Assumption 1. *Disturbance is bounded by a unknown constant $d_M > 0$ such that $\|d\|_\infty \le d_M$.*

Assumption 2. *The deviation angle $\Delta\alpha_i, (i \in \{1,2,3\})$ are limited to an allowable range $\Delta\alpha_i \in [-\alpha_M, \alpha_M]$, which depends on the manufacturing and assembling accuracy. The deviation angle $\Delta\beta_i (i \in \{1,2,3\})$ are represented in $[-\pi,\pi]$.*

C. Pointing Constraints

As mention in Sec. I, the light-sensitive payloads should be kept away from any harmful bright objects during the reorientation. The geometric relationship of them can be illustrated in Fig. 2. In the figure, b_i denotes the normalized sightline vector of the i-th sensitive payload represented in $\mathcal{F}_B$, and n_j is the normalized direction vector toward the j-th bright object represented in $\mathcal{F}_I$. To ensure the safety of

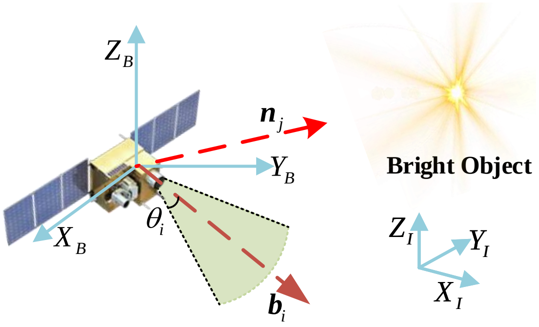


Fig. 2. The geometric relationship between payload and bright objects.

light-sensitive payloads, the vectors n_j, b_i and the half-angle-of-view of the i-th sensitive payload θ_i $(0 \leq \theta_i \leq \pi/2)$ must satisfy that:

$$b_i^T C(\sigma) n_j \leq \cos\theta_i \tag{5}$$

where $C(\sigma)$ is a transformation matrix satisfying that

$$C(\sigma) = I_3 - \frac{4(1-\sigma^T\sigma)}{(1+\sigma^T\sigma)^2}\sigma^\times + \frac{8}{(1+\sigma^T\sigma)^2}\sigma^\times\sigma^\times \tag{6}$$

Then, Eq. (5) can be further organized to be

$$c_{ij}(\sigma) = \cos\theta_i - b_i^T C(\sigma) n_j \geq 0 \tag{7}$$

Thus, once $c_{ij} \geq 0$, the attitude constraints can be guaranteed.

D. Problem Statement

The control objective of the attitude reorientation problem considered in this paper is to design a control scheme τ_c such that the controller can improve the control performance in real-time with the ability of constraints avoidance in the presence of the actuator misalignment and disturbance.

III. ADP-Based Controller Design

A. Reward Function Design

The reward function will be discussed first. The reward function is the mathematical characterization of environmental feedback while the agents are implementing the corresponding action. Using the reward signal to formalize the mission goal is a distinctive feature of reinforcement learning. The basic idea of the reward function design is that feedback a positive reward to desired states and a negative reward to undesired states.

The desired state is set as the target orientation, the corresponding reward function r_t is defined in the form of MRPs and angular velocity given by:

$$r_t = \sigma^T Q_\sigma \sigma + \omega^T Q_\omega \omega \tag{8}$$

where Q_σ and Q_ω denote the weight matrices associated with attitude and angular velocity, respectively. The balance between the reward of attitude and angular velocity can be adjusted by tuning these two weight matrices. It can be seen from (8) that the ”distance” from target states relate to the level of r_t.

Then, the control efficiency also should be considered here. The corresponding reward function is given as the quadratic form:

$$r_u = \tau_c^T R \tau_c \tag{9}$$

where R is the weight matrix associated with the control efficiency. By taking the r_u into account, the control consumption will be considered in the control policy.

Furthermore, according to the analysis in Sec. II-C, the reward function associated with the undesired states designed as:

$$r_c = -\sigma^T Q_\sigma \sigma \sum_{i=1}^{N_i}\sum_{j=1}^{N_j} \mu_{ij} \log\left(\frac{c_{ij}(\sigma)}{1-\cos\theta_i}\right) \tag{10}$$

where N_i and N_j are the numbers of light-sensitive payloads and bright objects. The scale factors μ_{ij} $(i \in \{1,2,...,N_i\}, j \in \{1,2,...,N_j\})$ are used to adjust the level of r_c.

Summing up the above analysis, the reward functions are constructed by (11), considering both desired and undesired states during the spacecraft reorientation mission by characterization the states into the corresponding values.

$$r = \underbrace{r_c}_{\text{undesired states}} + \underbrace{r_t}_{\text{desired states}} + \underbrace{r_u}_{\text{control efficiency}} \tag{11}$$

Remark 1. *Note that, the reward function designed here is different from the APF-based methods (e.g., in [7], [9]). Although both methods characterize the forbidden area by a large value, only the current value relates to the control signal in the APF-based methods, and the whole process's value relates to the control signal in the proposed method.*

B. Nominal Optimal Control Solution Analysis

After designing the reward function, the optimal control solution will be analyzed in this part. The spacecraft attitude model (1),(2) can be reorganized as the compact form:

$$\dot{\chi} = F(\chi) + G\tau_c \tag{12}$$

where $\chi = \left[\sigma^T, (J\omega)^T\right]^T \in \mathbb{R}^6$ is the motion state represented by the compact form, and

$$F(\chi) = \begin{bmatrix} H(\sigma)\omega \\ \omega^\times J\omega \end{bmatrix}, G = \begin{bmatrix} 0_{3\times3} \\ I_3 \end{bmatrix} \tag{13}$$

The cost function of the optimal control is defined as the integral of the reward function:

$$V(\chi) = \int_t^\infty r dt \tag{14}$$

The optimal control policy is denoted by τ_c^*. Thus, the corresponding cost function is represented by $V^*(\chi)$. After taking derivative for both sides of (14) with respect to time, we get the HJB equation as:

$$H(\chi, \tau_c^*, \nabla_\chi V^*(\chi)) \triangleq \nabla_\chi^T V^*(\chi)(F + G\tau_c^*) + r = 0 \tag{15}$$

Further taking partial differential for both sides of (15) with respect to τ_c^*, the closed-form of τ_c^* deduced as:

$$\tau_c^* = -\frac{1}{2}R^{-1}G^T\nabla_\chi V^*(\chi) \tag{16}$$

Then substituting (15) back into (16), the HJB equation can be rewritten as:

$$r_c + r_t + \nabla_\chi V^*(\chi) F - \frac{1}{4}\nabla_\chi^T V^*(\chi) G R^{-1} G \nabla_\chi V^*(\chi) = 0 \tag{17}$$

Note that the model described by (12) is a highly nonlinear system, which increases the intractability of analytically solving the HJB equation (17). Hence, approximation emerges as a way to deal with this problem. According to the Weierstrass approximation theorem [24], a neural network that contains a sufficient set of basis functions can be employed to approximate the optimal cost function (14), given as following:

$$V^*(\chi) = w^T \phi(\chi) + \epsilon(\chi) \tag{18}$$

In which, $\chi \in \mathbb{X}$, and $\mathbb{X} \subset \mathbb{R}^6$ is a compact set. The basis function is denoted by $\phi(\chi) = [\phi_1(\chi), \phi_2(\chi), \dots, \phi_p(\chi)]^T \in \mathbb{R}^p$ (p denotes the number of basis), satisfies that:

$$\begin{array}{l} \phi_i(0_{6\times 1}) = 0 \\ \dot{\phi_i}(0_{6\times 1}) = 0 \end{array}, \qquad i \in \{1, 2, \dots, p\} \tag{19}$$

The optimal weight vector of basis function w is a unknown constant vector, and $\epsilon(\chi) \in \mathbb{R}$ denotes the reconstruction error. Then the closed-form of τ_c^* (16) can be reconstructed as:

$$\tau_c^* = -\frac{1}{2} R^{-1} G^T \left[\psi(\chi) w + \varepsilon(\chi)\right] \tag{20}$$

with

$$\psi(\chi) = \nabla_\chi \phi(\chi), \quad \varepsilon(\chi) = \nabla_\chi \epsilon(\chi) \tag{21}$$

Since w is a unknown vector, the estimation of optimal weight w is denoted by $\hat{w}$. Then, the corresponding approximation of (16) and (18) given as:

$$\tau_c = -\frac{1}{2} R^{-1} G^T \psi(\chi) \hat{w} \tag{22}$$

$$V(\chi) = \hat{w}^T \phi(\chi) \tag{23}$$

Then, a special update law of the $\hat{w}$ will be discussed in the later part.

C. Augmented Term and Learning Law Design

Before proceeding, the augmented term design for actuator misalignment and disturbances will be discussed. Aiming at the alignment error and disturbance introduced in Sec. II-B, a special non-negative augmented term δ_M is designed as:

$$\delta_M = \alpha_1 k_M \lambda_M \hat{w}^T Y^T Y \hat{w} + d_M \|\hat{w}^T Y^T\| + \frac{1}{2} d_M^2 \|Y\|^2 \tag{24}$$

where $Y = G^T \psi(\chi)$ and $\rho_\delta > 0$ is an adjustable coefficient, $k_M = \|R^{-1}\|$. Then recalling Assumption 2, the upper bound of $\|\Lambda - I_3\|$ is exist, which can be denoted by λ_M.

Subsequently, the augmented term will be merged into the nominal optimal form. First, the augmented reward function is defined by the combination of the nominal reward function (11) and the augmented term designed in (24), as following:

$$r_{\text{aug}} = r + \delta_M \tag{25}$$

Accordingly, the cost function (14) is redefined by (26), and HJB equation (17) can be rewritten as (27)

$$V(\chi) = \int_t^\infty r_{\text{aug}} dt \tag{26}$$

$$\begin{aligned} H(\chi, \tau_c^*, \nabla_\chi V^*(\chi)) \triangleq & \nabla_\chi^T V^*(\chi))(F + G\tau_c^*) + r_{\text{aug}} \\ = & -\frac{1}{4}\nabla_\chi^T V^*(\chi) G R^{-1} G \nabla_\chi V^*(\chi) \\ & + r_t + r_c + \delta_M + \nabla_\chi^T V^*(\chi) F = 0 \end{aligned} \tag{27}$$

Then, further considering the following Bellman error:

$$\delta_{\text{HJB}} = \nabla_\chi^T V(\chi)(F + G\tau_c) + r_{\text{aug}} \tag{28}$$

Substituting (22)-(23) into (28), and adding (28) and (27), one has:

$$\begin{aligned} \delta_{\text{HJB}} = & \nabla_\chi^T V(\chi)(F + G\tau_c) + r_{\text{aug}} - H(\chi, \tau_c^*, \nabla_\chi V^*(\chi)) \\ = & \nabla_\chi^T V(\chi)(F + G\tau_c) + \tau_c R \tau_c \\ & - \nabla_\chi^T V^*(\chi)(F + G\tau_c^*) - \tau_c^* R \tau_c^* \\ = & \vartheta^T \tilde{w} + \epsilon_\delta \end{aligned} \tag{29}$$

where $\vartheta = \psi(\chi)(F + G\tau_c)$ is defined for expressing simplicity, $\tilde{w} = \hat{w} - w$ is the weight error, and ϵ_δ denotes the induced reconstruction error defined as the same form with [25].

Assumption 3. *Consider an auxiliary variable defined by $\eta = \vartheta/(1 + \vartheta^T\vartheta) \in \mathbb{R}^p$, and it satisfies that there exist time instants $t_1 < t_2 < \dots < t_l$ and a positive constants c_η such that $\sum_{k=1}^l \eta(t_k)\eta^T(t_k) \geq c_\eta I_{p\times p}$*

It can be noticed from (29) that the bellman error relates to $\tilde{w}$. Therefore, the learning law will be designed by considering the δ_{HJB} as following:

$$\dot{\hat{w}} = -\gamma_1 \frac{\delta_{\text{HJB}} \vartheta}{(1 + \vartheta^T \vartheta)^2} - \gamma_2 \Theta \tag{30}$$

in which, γ_1 and γ_2 are positive constant, and γ_1 is used to adjust the learning rate. The auxiliary variable Θ defined as:

$$\begin{aligned} \Theta = & \sum_{k=1}^l \eta(t_k)\eta^T(t_k)\hat{w} + r\eta(t_k)/[1 + \vartheta^T(t_k)\vartheta(t_k)] \\ = & \Omega \tilde{w} + \epsilon_\Theta \end{aligned} \tag{31}$$

Note that $\Omega = 2\sum_{k=1}^l \eta(t_k)\eta^T(t_k)$, and it contains the data at previous time instance $t_1, t_2, \dots, t_l$. Under the Assumption 3, there will be a constant c_η such that $\Omega \geq 2c_\eta I_{p\times p}$. Besides, $\epsilon_\Theta = \sum_{k=1}^l \eta(t_k)\epsilon_\delta/[1 + \vartheta^T(t_k)\vartheta(t_k)]$ is the residual error.

As mentioned above, the learning law and the augmented term design for actuator misalignment are developed in this part. After that, the relevant analysis will be given in the following part.

D. Convergence and Stability Analysis

While before proposing the convergence and stability analysis, a standard assumption, as given in most of the RL-based control algorithms, are given as follows.

Assumption 4. *For $\chi \in \mathbb{X}$, there exist positive constants b_ϕ, b_ψ, b_ϵ, b_ε, and b_δ, such that $\|\phi\| \leq b_\phi$, $\|\psi\| \leq b_\psi$, $\|\epsilon\| \leq b_\epsilon$, $\|\varepsilon\| \leq b_\varepsilon$, and $\epsilon_\delta \leq b_\delta$.*

To ensure that the state χ is in the compact set $\mathbb{X}$, the following initial control policy will be introduced.

Lemma 1. *[23] Consider the spacecraft attitude dynamics in (1) and (2), design an initial control policy as:*

$$\tau_{init} = -k_\sigma \sigma - k_\omega \omega \tag{32}$$

where k_σ and k_ω are positive coefficient of controller. Then for all initial state $\chi(0)$ is in the admissible domain $\mathbb{D}_\chi \subset \mathbb{R}^6$, exits a compact set $\mathbb{X}$ such that $\chi(t) \in \mathbb{X}$, for all $t \geq 0$.

Remark 2. *The initial controller is used to trigger the online learning mechanism and guarantee asymptotically stability at the beginning, the form of which is not unique. The above controller was employed because of its simplicity and ease of implementation in practice. What's more, the form of (32) can be conveniently reconstructed by the combination of the basis function and weight vector as (38).*

Lemma. 1 guarantees an admissible initial control policy. Then the convergence and stability of entire system is analyzed by following Theorem.

Theorem 1. *Consider the spacecraft attitude dynamics in (1)-(2) with the actuator alignment (3), and the policy described by (22)-(23). Design the learning law as (30). Then, under the Assumption 4, for all $\chi(0) \in \mathbb{D}_\chi$, one has χ and $\tilde{w}$ are ultimately bounded for all $t \geq 0$.*

Proof. Consider the following storage function:

$$L = V^*(\chi) + \frac{\rho_L}{2}\tilde{w}^T\tilde{w} \tag{33}$$

where $\rho_L > 0$ is a constant just for analysis. Then taking the time derivative of (33), and substituting (22) and (27) into it , yield:

$$\begin{aligned}\dot{L} =& \nabla_\chi^T V^*(F + G\Lambda\tau_c + Gd) + \rho_L \tilde{w}^T \dot{\tilde{w}} \\ =& -r_t - r_c - \delta_M - \frac{1}{4}w^T Y^T R^{-1} Y w + \frac{1}{4}\varepsilon^T G R^{-1} G^T \varepsilon \\ & - \frac{1}{2} w Y^T R^{-1} Y \tilde{w} - \frac{1}{2}\varepsilon^T G R^{-1} Y \tilde{w} + \hat{w}^T Y^T d + \tilde{w}^T Y^T d \\ & + \varepsilon^T G d - \frac{1}{2}\hat{w} Y^T (\Lambda - I_3) R^{-1} Y \hat{w} + \rho_L \tilde{w}^T \dot{\tilde{w}} \\ & + \frac{1}{2}\tilde{w} Y^T (\Lambda - I_3) R^{-1} Y \hat{w} - \frac{1}{2}\varepsilon^T G (\Lambda - I_3) R^{-1} Y \hat{w}\end{aligned} \tag{34}$$

Further substituting (24) and (30) into (34), then employing the arithmetic-geometric average inequality, one has

$$\begin{aligned}\dot{L} \leq & -r_t - r_c + \frac{1}{2}\varepsilon^T G R^{-1} G^T \varepsilon + \frac{1}{2}\tilde{w}^T Y^T R^{-1} Y \tilde{w} \\ & - d_M \|\hat{w}^T Y^T\| - \frac{1}{2} d_M^2 \|Y\|^2 - \alpha_1 k_M \lambda_M \hat{w}^T Y^T Y \hat{w} \\ & + \hat{w}^T Y^T d - \tilde{w}^T Y^T d + \varepsilon^T G d + k_M \lambda_M \hat{w} Y^T Y \hat{w} \\ & + \frac{1}{4} k_M \lambda_M \tilde{w} Y^T Y \tilde{w} + \rho_L \tilde{w}^T \dot{\tilde{w}} \\ \leq & -r_t - r_c - (\alpha_1 - 1) k_M \lambda_M \hat{w} Y^T Y \hat{w} + \varepsilon_{L1} + \varepsilon_{L2} \\ & - \frac{1}{2}\tilde{w}^T [\frac{\gamma_1 \rho_L \vartheta \vartheta^T}{(1 + \vartheta^T \vartheta)^2} + (\gamma_2 \rho_L (4c_\eta - 1) - 1) I_p \\ & - \frac{\lambda_M + 2}{2} k_M Y^T Y] \tilde{w}\end{aligned} \tag{35}$$

where $\varepsilon_{L1} = (0.5 + 0.25\lambda_M)\varepsilon^T G R^{-1} G^T \varepsilon + \varepsilon^T G d$, and $\varepsilon_{L2} = 0.5\rho_L \gamma_1 \varepsilon_H^2 / (1 + \vartheta^T \vartheta)^2 + 0.5\gamma_2 \rho_L \epsilon_\Theta^T \epsilon_\Theta$ are bounded by Assumption. 4. Then by setting parameters α_1 and ρ_L such that:

$$\alpha_1 \geq 1, \text{and } \rho_L \geq \frac{1}{\gamma_2(4c_\eta - 1)}(1 + \frac{\lambda_M + 2}{4} k_M b_\psi^2) \tag{36}$$

On this basis, Eq. (35) directly guarantees the ultimate boundedness of χ and $\tilde{w}$. What's more, the result $V(\chi) \in \mathcal{L}_\infty$ lead to $\int_{t_0}^\infty r_c \in \mathcal{L}_\infty$, which ensure the constraints will not be violated. □

Remark 3. *Although the initial policy (32) does not consider the motion constraints, actuator misalignment, and performance optimization, after it triggers the online learning process, the system will have these capabilities by learning from the online states.*

Remark 4. *Notably, according to the processing given in (34) and (35), the precondition of the disturbance does not need to converge relative to states anymore (as assumed in [15]). Thus, as long as meeting the precondition that disturbance is bounded, the system's stability can be ensured in theory. This improvement in the precondition of disturbance is also of great significance in practical applications.*

To assist the readers in understanding the operating mechanism of the proposed ADP-based method, the main framework of the whole system is provided in Fig. 3. Besides, the implementation procedures are also concluded in Fig. 4.

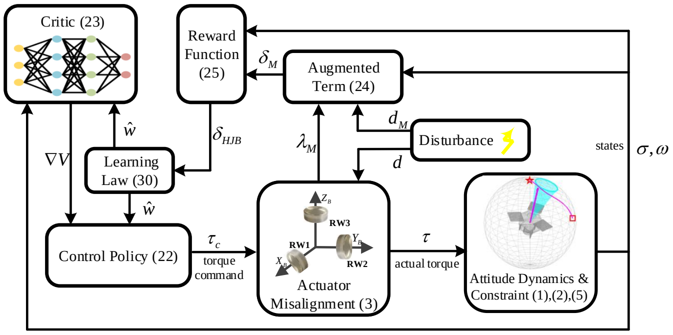


Fig. 3. The main framework of the whole system.

It is worth noting that the learning process is an online manner in practical implementation, as highlighted by the red box in Fig. 4. Particularly, an admissible control policy should be employed as the initial controller at the beginning. Then the learning law acts as a tuner to adjust the control policy according to the online reward (feedback from the environment).

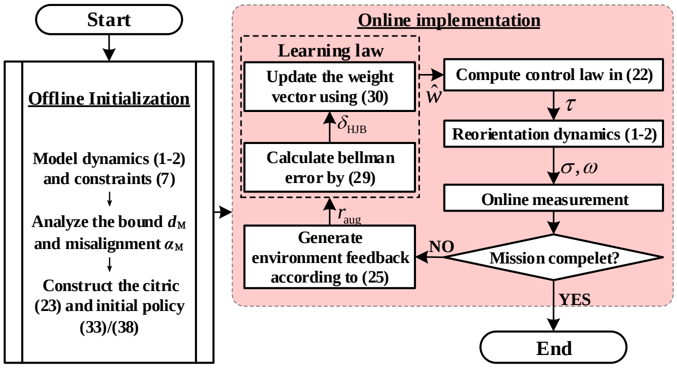

Start
Offline Initialization
Model dynamics (1-2) and constraints (7)
Analyze the bound d_M and misalignment α_M
Construct the citric (23) and initial policy (33)/(38)
Online implementation
Learning law
Update the weight vector using (30)
δ_{HJB}
Calculate bellman error by (29)
r_{aug}
Generate environment feedback according to (25)
$\hat{W}$
Compute control law in (22)
τ
Reorientation dynamics (1-2)
σ, ω
Online measurement
Mission compelet?
NO
YES
End

Fig. 4. The flowchart of implementation manner.

IV. Numerical Simulation

A. Comparison Simulation

To demonstrate the effectiveness and performance of the developed method, a set of numerical simulations are conducted in the presence of attitude constraints and alignment error. The simulation results are obtained by MATLAB/Simulink on a 3.40 GHz Intel-i7 desktop computer with 16 GB of RAM.

Consider the inertia matrix of rigid spacecraft as $J = [20, 1.2, 0.9; 1.2, 17, 1.4; 0.9, 1.4, 15]\,\text{kg}\cdot\text{m}^2$ and the initial states are set to be $\sigma(0) = [-0.2735, -0.2099, -0.0844]^T$, $\omega(0) = [0, 0, 0]^T$rad/s. The light-sensitive instrument is assumed to be aligned with the Z_B axis of the frame $\mathcal{F}_B$, thus, $b_1 = [0, 0, 1]^T$, and the half-angle-of-view is set to be 15 deg. During the reorientation process, the vectors of bright objects need to be avoided are given as: $n_1 = [-0.2310, 0.4077, 0.8834]^T$, $n_2 = [-0.2750, 0.0050, 0.3250]^T$, $n_3 = [0.0864, 0.7564, 0.6484]^T$. The simulated disturbances are assumed to occur at the beginning of simulation, and its form defined in (37).

$$\begin{aligned} d =& 5\times 10^{-4} \\ &\times \begin{bmatrix} 3\cos(10\|\omega\|t) + 4\sin(3\|\omega\|t) + 5\text{rand}(1) \\ -1.5\cos(2\|\omega\|t) + 3\sin(5\|\omega\|t) - 7.5\text{rand}(1) \\ 3\cos(10\|\omega\|t) - 8\sin(4\|\omega\|t) - 2.5\text{rand}(1) \end{bmatrix} \text{Nm} \end{aligned} \tag{37}$$

Besides, the misalignment angles are assumed to be: $\Delta\alpha_1 = 14.3\text{deg}$, $\Delta\alpha_2 = 15.0\text{deg}$, $\Delta\alpha_3 = -14.5\text{deg}$, $\Delta\beta_1 = 36.0\text{deg}$, $\Delta\beta_2 = -20.0\text{deg}$, $\Delta\beta_3 = -15.4\text{deg}$.

To straightforwardly demonstrate the advantages of the proposed method, the initial controller and other two controllers in related literature are compared in the numerical simulation, and their parameters are also given as following:

a) Initial controller: A PD-like controller is employed as the initial control policy for the proposed controller here. The basis-weight form of (32) are reconstructed as follow

$$\tau_{\text{init}} = -\frac{1}{2}R^{-1}G^T\psi\hat{w}_0 \tag{38}$$

where $\hat{w}_0 = [0.25\times\mathbf{1}_{1\times3}, 5\times\mathbf{1}_{1\times3}, \mathbf{0}_{1\times9}]^T$ denotes the initial value of $\hat{w}$, and $R = I_3$. The basis functions are given by:

$$\begin{aligned} \phi = [&\sigma_1\omega_1, \sigma_2\omega_2, \sigma_3\omega_3, \omega_1^2, \omega_2^2, \omega_3^2, \sigma_1^2\omega_1^2, \sigma_2^2\omega_2^2, \sigma_3^2\omega_3^2, \\ &..., \sigma_1^2\omega_2^2, \sigma_1^2\omega_3^2, \sigma_2^2\omega_1^2, \sigma_2^2\omega_3^2, \sigma_3^2\omega_1^2, \sigma_3^2\omega_2^2]^T \end{aligned} \tag{39}$$

b) Proposed controller: The proposed controller is improved from the initial controller, and its parameters are set to be: $Q_\sigma = I_3$, $Q_\omega = I_3$, $R = I_3$, $\gamma_1 = \gamma_2 = 2$, $u_{11} = u_{12} = u_{13} = 1.5$, $\alpha_2 = 2$.

c) Controller in [17]: A reinforcement learning controller without the abilities of handling misalignment and disturbances (abbreviated as ADPC). It keeps the same parameters as the proposed controller, but no augmented term is designed for misalignment and disturbances.

d) Controller in [15]: An inverse optimal controller with the abilities of handling misalignment and disturbances (abbreviated as IOC). The parameters of it are set to be: $K = 0.25I_3$, $R = I_3$, $Q = I_6$, $\sigma_t(t) = e^{-0.2t}$.

For fair comparison, we employed the same cost function, given in (40), to evaluate the performance of these controllers.

$$r_{\text{compare}} = \sigma^T Q_\sigma \sigma + \omega^T Q_\omega \omega + \tau_c^T R \tau_c \tag{40}$$

in which, $Q_\sigma = Q_\omega = R = I_3$.

The simulation results are given in the Figs. 5-8. The time responses of the attitude, angular velocity and control signal under the above mention controllers are depicted in Fig. 5, which shows that all these controllers successfully reorientate to the desire states. Fig. 6 shows the trajectories of sightline of the light-sensitive payload b_1 in $\mathcal{F}_\mathcal{I}$ on unit sphere and corresponding 2D projection under different controllers. It is obvious that the initial controller and IOC controller fail to avoid the undesired states. Another noteworthy feature is that a short section of ADPC's trajectory does not bypass the prohibit area but the proposed controller do it, as shown in the local magnification of Figs. 6 (b) and (c). The major reason is that the misalignment of the actuators leads to the control torque cannot be accurately applied to the ideal direction. To further demonstrate the performance of the proposed controller, the overall cost of different controllers during the orientation is given in Fig. 7. It can be seen that the proposed method improves the performance from the initial controller and it shows effective optimizing abilities compared with the ADPC method under the actuator misalignment. Note that, the IOC controller presents the lowest cost because it's trajectory crossed the prohibited area without handling the constraints. Fig. 8 shows the learning process of the proposed method. The bellman error δ_{HJB} converges to near zero after 100s and estimation of weight vector $\hat{w}$ also tends to be stable, which indicates that the proposed method achieves (approximate) optimal control.

Then, the influence of uncertain actuator misalignment is studied in Fig. 9. Based on the parameters set in the

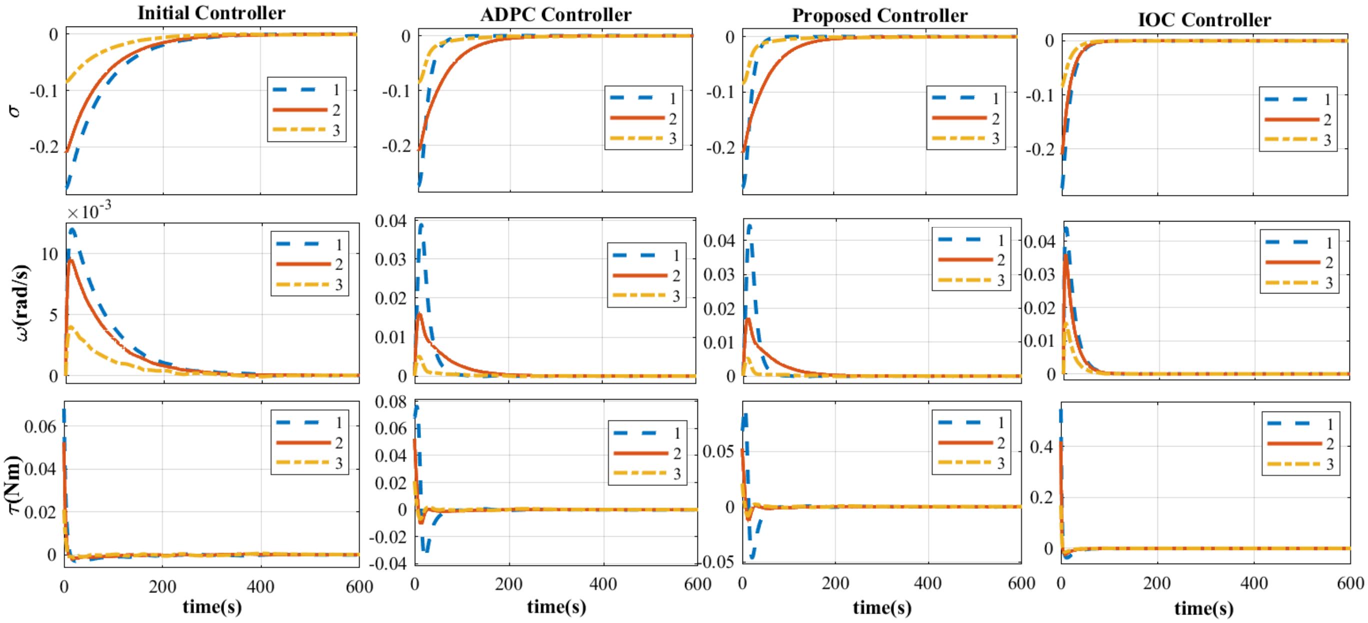
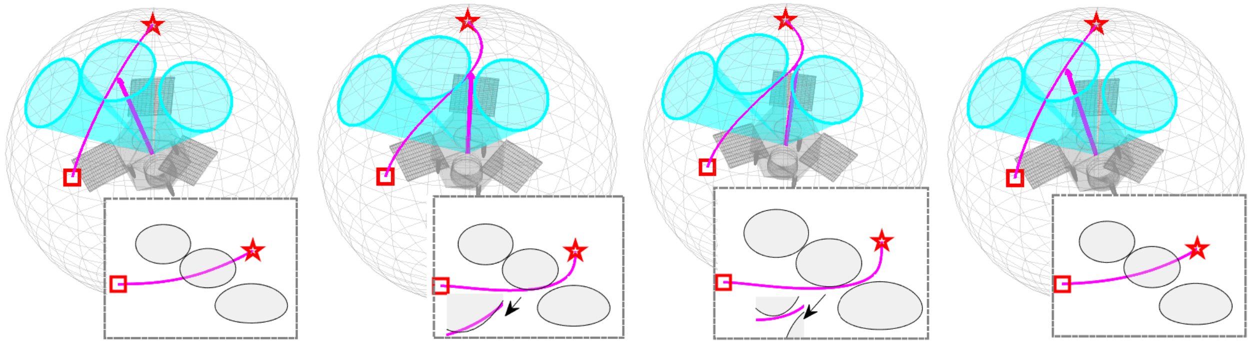


Fig. 5. The time responses of the attitude, angular velocity and control signal under different controllers.


(a)

(b)

(c)

(d)

Fig. 6. 3D trajectories on unit sphere and corresponding 2D projection on cylindrical projection under the different controllers. (a) Initial controller. (b) ADPC controller. (c) Proposed controller. (d) IOC controller.

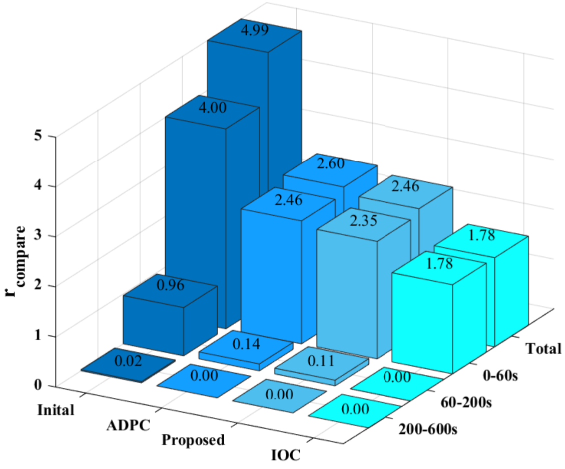


Fig. 7. Comparison of performance cost under different controllers.

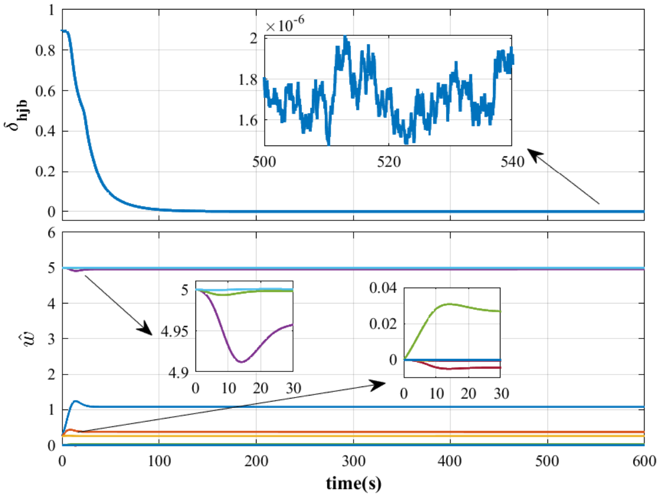


Fig. 8. Learning result of the proposed controller.

above case, three comparison results ($\alpha_M = 12\text{deg}$, $\alpha_M = 8\text{deg}$, $\alpha_M = 4\text{deg}$) are given in Fig. 9. It can be seen from the two local magnifications that the dynamic response and steady accuracy under these deviation angles have little difference. These results also present the characteristics that the smaller deviation angle leads to faster response and higher stable accuracy.

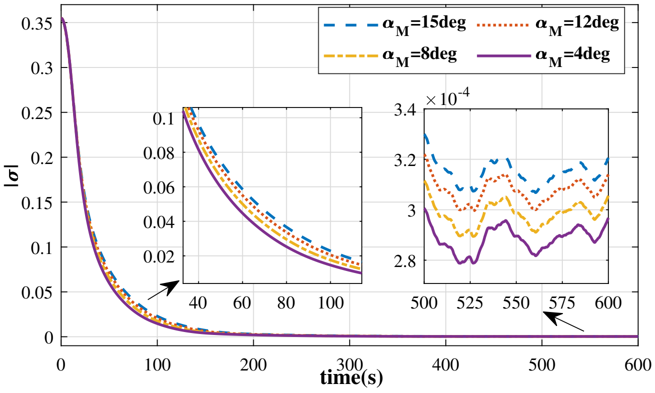


Fig. 9. Dynamic response and steady state accuracy under different deviation angle.

B. Mentor-Carlo Simulation

In order to verify the comprehensive insight into the proposed method's performance, a 500-run Mentor-Carlo simulations are given in this part. The initial attitude and parameters of the system are randomly chosen in the ranges listed in Table I.

TABLE I
RANGES OF INITIAL ATTITUDE AND SYSTEM'S PARAMETERS

Parameter	Values/Ranges
$\sigma(0)$	$(-0.272, 0.274) \times (-0.20, 0.22)$ $\times(-0.075, -0.0095)$
$\alpha_i(i=1,2,3)$, deg	$(-15, 15)$
$\beta_i(i=1,2,3)$, deg	$(-180, 180)$
b_1 (single constraint)	in a 15 deg cone with $[-0.23, 0.41, 0.88]^T$ as the central axis
J, $\text{kg} \cdot \text{m}^2$	$[20, 1.2, 0.9; 1.2, 17, 1.4; 0.9, 1.4, 15]$ $+[j_1, j_4, j_5; j_4, j_2, j_6; j_5, j_6, j_3]$
$j_i(i=1,\cdots,6)$, $\text{kg} \cdot \text{m}^2$	$(-0.1, 0.1)$

The results of the Mentor-Carlo simulation are summarized in Fig. 10. It can be noted from the first subfigure that the minimum angle between b_1 and n_1 of every single run is larger than the constraint angle. The second and third subfigures show the terminal attitude error and convergence time of each run in the Mentor-Carlo simulation, respectively. The convergence time t_{conv} is defined as the time duration from initial state to a given allowable state error ($\|[\sigma^T, \omega^T]^T\| \le 1.0 \times 10^{-3}$). These Mentor-Carlo simulation results indicate that the proposed method can excellently achieve control objectives under variable initial states and parameter uncertainties without any constraint violations.

To sum up, the simulation results show the performance of the proposed RL-based control scheme, in terms of forbidden pointing avoidance, actuator misalignment handling, and anti-disturbances ability.

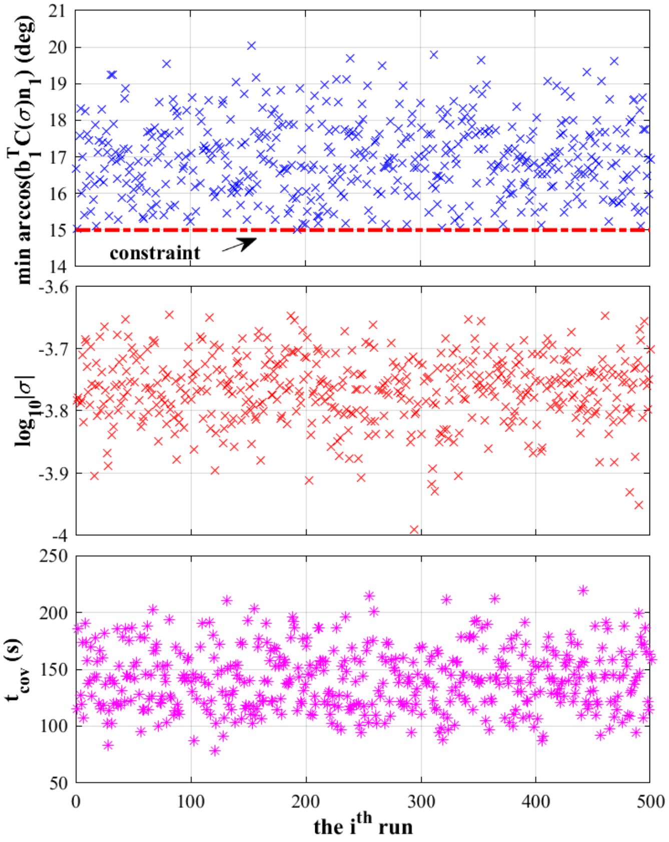


Fig. 10. Results of the Mentor-Carlo simulation.

V. Experimental Verification

A. Experimental Platform

To further validate the practical applicability of the proposed method from the perspective of implementation, experimental studies are conducted in this section by using a hardware-in-loop (HIL) experimental platform, as shown in Fig. 11. The mainly consists of this HIL experiment platform are listed as following: 1) A high-performance real-time simulation computer (HPRTSC), which runs the dynamics model on the VxWorks system. 2) A three-axis turntable for rotational motion simulation. 3) An angular velocity measurement unit comprising four fibre-optic gyroscopes (FOGs). 4) A reaction wheel assembly (RWA), which executes the control commands from HPRTSC and feedbacks its output to HPRTSC. 5) Other hardware and software including power modules and monitor interface. Other detailed information can be found in [26]. The proposed controller of the spacecraft attitude reorientation can be tested and illustrated by this HIL platform.

The experimental procedures are arranged as shown in Fig. 12. First, the attitude angle can be obtained from the turntable, and the angular velocity is measured by the FOGs. Second, the motion information is sent to the HPRTSC as state feedback, prior to this, the attitude angle is converted into the form of MRPs to meet the requirement of the controller.

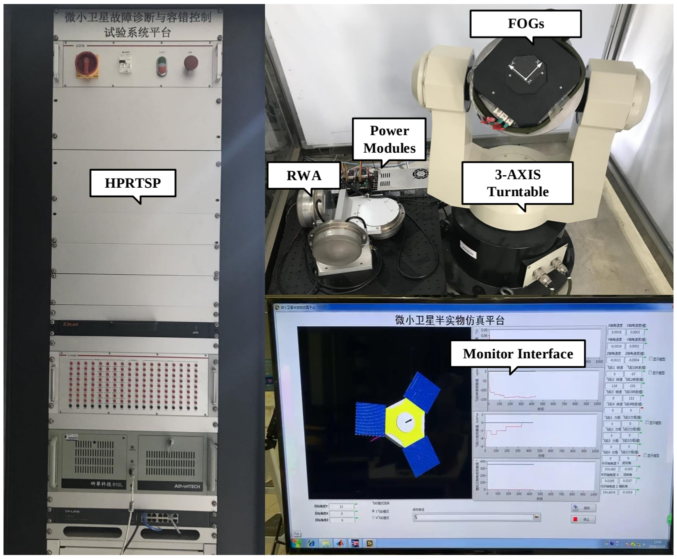


Fig. 11. Hardware-in-loop experiment platform.

Third, the control signal is calculated from the controller and implemented on the RWA. Note that, we add a saturation function to restrict the control signal within the maximum output torque of each reaction wheel (0.1 Nm). Besides, the output torque of reaction wheels can be measured by its inner sensor, then the misalignment simulator receives these signals and calculates the real torque applied to the spacecraft. Fourth, the attitude dynamics run in the HPRTSC, which received the torque and sent the attitude angle to the turntable for motion simulation. In addition, the monitor interface is used to display some key states and parameters during the experiment process.

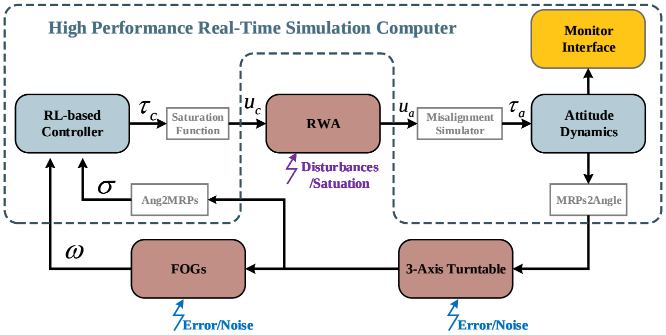


Fig. 12. Block diagram of hardware-in-loop experiment system.

B. Experimental Result

In the experiment scenario, due to physical limitations (e.g., real-time performance of the sensor and actuators), the control period is set to be 50ms. Besides, practical measurement noise of states and disturbances of control signals is inevitably introduced into the closed-loop system. Therefore the numerical noise (37) will not be added anymore. The other parameters keep the same as in the numerical simulation scenario.

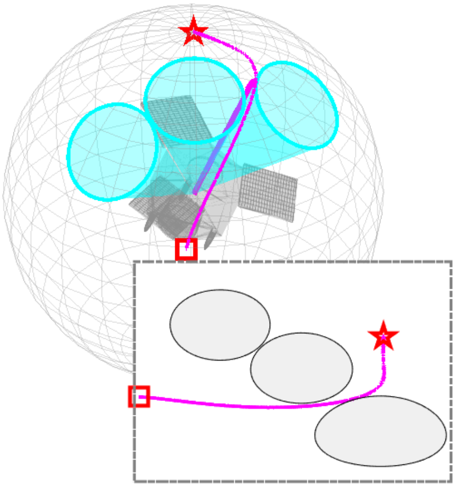


Fig. 13. Motion illustration of the experiment.

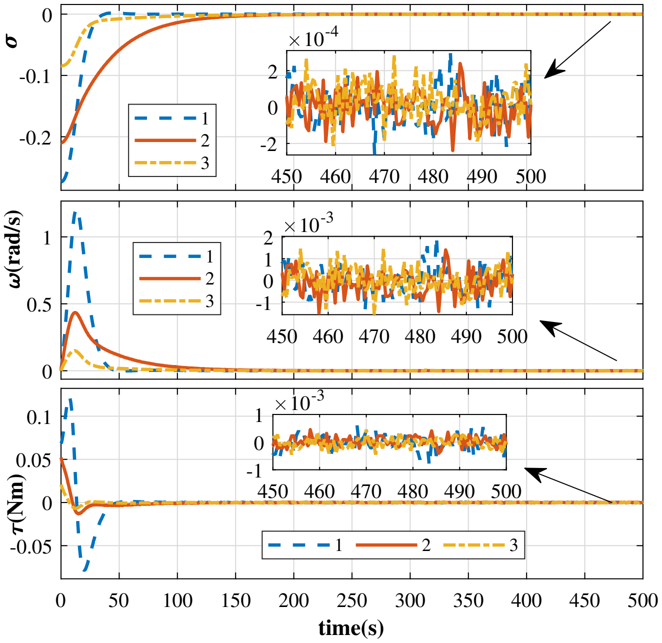


Fig. 14. Experimental results of motion state and control torque.

From Figs. 13-14, we can see that the spacecraft ultimately maneuvers to the desired pointing without crossing any forbidden areas. Fig. 15 depicts the learning process in HIL experiment. It can be seen that the HJB error δ_{HJB} converges to zero and the weighed vector $\hat{w}$ trend to stable within 80s. Although the bellman error δ_{HJB} is larger than the numerical simulation, this result can be still be regarded as achieving optimal control, which meets the requirement of the mission. As the similar in numerical simulation, dynamic response and steady state accuracy under different deviation angles are given in Fig. 16. The dynamic response still follows the law that the smaller the angle leads to the faster the response. But there is almost no difference in stability accuracy due to the measurement and actuator accuracy, as shown in the local magnifications of Fig. 16.

To this end, the above results show that the proposed method not only works well in the numerical case but also in the HIL experiment.

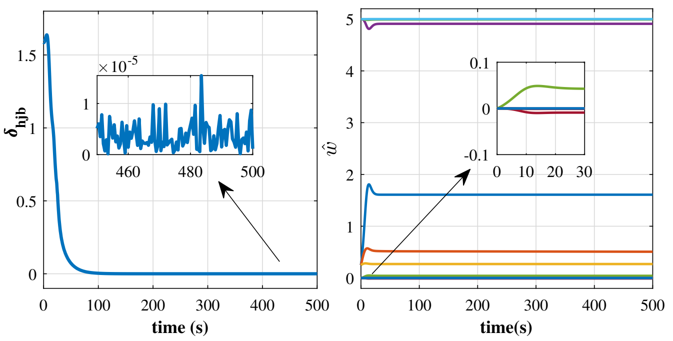


Fig. 15. Experimental results of learning.

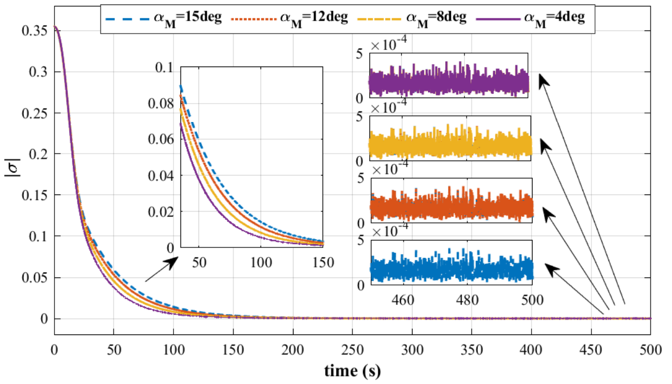


Fig. 16. Experimental results of dynamic response and steady state accuracy under different deviation angles.

VI. Conclusion

In this paper, a novel ADP-based controller was proposed for attitude reorientation in the presence of actuator misalignment, motion constraints and disturbances. Barrier functions were employed to encode motion constraints and a specially designed augmented term was introduced into the ADP framework to mitigate the influence of actuator misalignment and disturbances. The ultimate boundedness of state errors was guaranteed by the Lyapunov-based analysis. Numerical simulations verified the effectiveness and advantages of the proposed method. Notably, the HIL experiment results showed the potential for practical implementation of our method.

References

[1] S. Yin, B. Xiao, S. X. Ding, and D. Zhou, “A Review on Recent Development of Spacecraft Attitude Fault Tolerant Control System,” *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3311–3320, 2016.

[2] C. Wang, L. Guo, C. Wen, Q. Hu, and J. Qiao, “Event-Triggered Adaptive Attitude Tracking Control for Spacecraft with Unknown Actuator Faults,” *IEEE Trans. Ind. Electron.*, vol. 67, no. 3, pp. 2241–2250, 2020.

[3] D. Li, H. Yu, K. P. Tee, Y. Wu, S. S. Ge, and T. H. Lee, “On time-synchronized stability and control,” *IEEE IEEE Trans. Syst. Man Cybern. Syst.*, 10.1109/TSMC.2021.3050183.

[4] Y. Xia, Z. Zhu, M. Fu, and S. Wang, “Attitude tracking of rigid spacecraft with bounded disturbances,” *IEEE Trans. Ind. Electron.*, vol. 58, no. 2, pp. 647–659, 2011.

[5] U. Lee and M. Mesbahi, “Feedback control for spacecraft reorientation under attitude constraints via convex potentials,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 4, pp. 2578–2592, 2014.

[6] M. D. Ramos and H. Schaub, “Kinematic steering law for conically constrained torque-limited spacecraft attitude control,” *J. Guid. Control. Dyn.*, vol. 41, no. 9, pp. 1990–2001, 2018.

[7] Q. Hu, B. Chi, and M. R. Akella, “Anti-unwinding attitude control of spacecraft with forbidden pointing constraints,” *J. Guid. Control. Dyn.*, vol. 42, no. 4, pp. 822–835, 2019.

[8] X. Shao, Q. Hu, Y. Shi, and B. Yi, “Data-driven immersion and invariance adaptive attitude control for rigid bodies with double-level state constraints,” *IEEE Trans. Control Syst. Technol.*, 10.1109/TCST.2021.3076439.

[9] S. Kulumani and T. Lee, “Constrained geometric attitude control on so (3),” *Int. J. Control Autom. Syst.*, vol. 15, no. 6, pp. 2796–2809, 2017.

[10] 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,” *J. Guid. Control. Dyn.*, vol. 40, no. 2, pp. 309–319, 2017.

[11] H. C. Kjellberg and E. G. Lightsey, “Discretized quaternion constrained attitude pathfinding,” *J. Guid. Control. Dyn.*, vol. 39, no. 3, pp. 710–715, 2016.

[12] B. Luo, D. Liu, H. N. Wu, D. Wang, and F. L. Lewis, “Policy Gradient Adaptive Dynamic Programming for Data-Based Optimal Control,” *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3341–3354, 2017.

[13] B. Xiao, Q. Hu, D. Wang, and E. K. Poh, “Attitude tracking control of rigid spacecraft with actuator misalignment and fault,” *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 6, pp. 2360–2366, 2013.

[14] F. Zhang and G. R. Duan, “Robust adaptive integrated translation and rotation finite-time control of a rigid spacecraft with actuator misalignment and unknown mass property,” *Int. J. Syst. Sci.*, vol. 45, no. 5, pp. 1007–1034, 2014.

[15] Z. Wang and Y. Li, “Rigid spacecraft robust optimal attitude stabilization under actuator misalignments,” *Aerosp. Sci. Technol.*, vol. 105, p. 105990, 2020.

[16] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming with Applications in Optimal Control*, ser. Advances in Industrial Control. Cham: Springer International Publishing, 2017.

[17] H. Dong, X. Zhao, and H. Yang, “Reinforcement Learning-Based Approximate Optimal Control for Attitude Reorientation Under State Constraints,” *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 4, pp. 1664–1673, 2021.

[18] Q. Hu, H. Yang, H. Dong, and X. Zhao, “Learning-based 6-dof control for autonomous proximity operations under motion constraints,” *IEEE Trans. Aerosp. Electron. Syst.*, 10.1109/TAES.2021.3094628.

[19] C. Wei, J. Luo, H. Dai, and G. Duan, “Learning-based adaptive attitude control of spacecraft formation with guaranteed prescribed performance,” *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 4004–4016, 2019.

[20] C. Mu, Z. Ni, C. Sun, and H. He, “Air-Breathing Hypersonic Vehicle Tracking Control Based on Adaptive Dynamic Programming,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 3, pp. 584–598, 2017.

[21] C. Wei, J. Luo, H. Dai, Z. Bian, and J. Yuan, “Learning-based adaptive prescribed performance control of postcapture space robot-target combination without inertia identifications,” *Acta Astronautica*, vol. 146, pp. 228–242, 2018.

[22] B. Luo, H. N. Wu, and T. Huang, “Optimal Output Regulation for Model-Free Quanser Helicopter with Multistep Q-Learning,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 4953–4961, 2018.

[23] P. Tsiotras, “Further passivity results for the attitude control problem,” *IEEE Trans. Automat. Contr.*, vol. 43, no. 11, pp. 1597–1600, 1998.

[24] 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.

[25] S. Xue, B. Luo, and D. Liu, “Event-Triggered Adaptive Dynamic Programming for Zero-Sum Game of Partially Unknown Continuous-Time Nonlinear Systems,” *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 50, no. 9, pp. 3189–3199, 2020.

[26] H. Yang and Q. Hu, “Research and experiment on dynamic weight pseudo-inverse control allocation for spacecraft attitude control system,” in *Proc. Chinese Control Conf.*, Guangzhou, China, Jul. 2019, pp. 8200–8205.