Learning-Based 6-DOF Control for Autonomous Proximity Operations under Motion Constraints

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

***Abstract*—This paper proposes a reinforcement learning (RL)-based six-degree-of-freedom (6-DOF) control scheme for the final-phase proximity operations of spacecraft. The main novelty of the proposed method are from two aspects: 1) the closed-loop performance can be improved in real-time through the RL technique, achieving an online approximate optimal control subject to the full 6-DOF nonlinear dynamics of spacecraft; 2) Nontrivial motion constraints of proximity operations are considered and strictly obeyed during the whole control process. As a stepping stone, the dual-quaternion formalism is employed to characterize the 6-DOF dynamics model and motion constraints. Then, an RL-based control scheme is developed under the dual-quaternion algebraic framework to approximate the optimal control solution subject to a cost function and a Hamilton-Jacobi-Bellman equation. In addition, a specially designed barrier function is embedded in the reward function to avoid motion constraint violations. The Lyapunov-based stability analysis guarantees the ultimate boundedness of state errors and the weight of NN estimation errors. Besides, we also show that a PD-like controller under dual-quaternion formulation can be employed as the initial control policy to trigger the online learning process. The boundedness of it is proved by a special Lyapunov strictification method. Simulation results of prototypical spacecraft missions with proximity operations are provided to illustrate the effectiveness of the proposed method.**

***Index Terms*—Spacecraft Proximity Operations; Constrained 6-DOF Control; Reinforcement Learning; Approximate Optimal Control.**

I. Introduction

AUTONOMOUS spacecraft proximity operations (SPO) is an essential technology for a broad range of space missions, such as docking, servicing, inspection, sample retrieval, active debris removal, and asteroids exploration [1]–[7]. As the primary requirement of these missions, flying safety must be guaranteed during all proximity operations. This requires the spacecraft to obey multiple complex motion constraints. The two most important types of constraints are referred as approach corridors and field-of-view constraints in literature: the approach corridor aims to restrict the translational motion trajectory of the spacecraft, such that the potential collision between the components of spacecraft can be avoided [8]; the field-of-view constraint is arising from the requirement of autonomous rendezvous and capture sensor system (ARCSS) [9], the target must be kept in the detectable zone of ARCSS to provide the essential information for GNC systems.

Due to the practical significance of SPO, guidance and control methods for SPO with the ability to handle complex motion constraints have aroused extensive attention. Various approaches have been investigated, such as artificial potential function (APF) [10]–[15], inverse dynamics in the virtual domain (IDVD) [16], optimization based method [17], [18], and model predictive control (MPC) [16], [19], [20]. These methods can be categorized into two types: optimization-based methods and APF-based methods. The APF-based methods usually establish virtual high-potential areas for obstacles forbidden zones that produce repulsive forces to prevent violations of constraints. Ref. [10] proposed a dual-quaternion-based APF to solve the constrained six-degree-of-freedom (6-DOF) maneuvers, and the local minimum problem was addressed by the selection of control parameters that satisfies a mild condition. Zappulla et al. [14] employed an adaptive APF approach on hardware-in-loop SPO experiments and achieved collision-free SPO. Huang et al. [11] designed a finite-time control law with full-state constraints by incorporating the tan-type barrier Lyapunov function. We note that APF-based methods have shown their capability of handling constraints. However, these results usually lack optimizing abilities. They cannot make the balance between the control performance and control cost, resulting in potentially high control costs that are unacceptable for on-orbit SPO missions. In this context, constrained optimal control (COC) is a promising alternative solution. A second order cone programming (SOCP) based method was presented in [17] for the rendezvous and docking with corridor constraints. However, this method is open-loop which cannot deal with real-time feedback. Although an MPC-based COC approach was proposed in [20] with the capability of feedback control, the receding-horizon characteristic of MPC makes solving a COC for SPO in real-time become a computationally burdensome task, especially for onboard computers which only have very limited computing resources. Thus, it is of significant importance to design a new constrained optimal control scheme for SPO that can efficiently address the 6-DOF optimal tracking control problems while strictly obey underlying motion constraints.

Theoretically speaking, the optimal control of nonlinear systems usually requires to solve the Hamilton-Jacobi-Bellman

This work was supported in part by the National Natural Science Foundation of China under Grant (61960206011, 61633003, 61803012), in part by the Key Laboratory Project of National Defense Science and Technology of China under Grant 6142208190303, in part by the Major Project of Science and Technology Innovation 2030—"New Generation Artificial Intelligence" under Grant 2018AAA0102700, in part by the Beijing Natural Science Foundation under Grant JQ19017, and in part by the Beijing Advanced Discipline Center for Unmanned Aircraft System. (Corresponding author: Qinglei Hu.)

Q. Hu, H. Yang and H. Dong are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. Emails: {huql_buaa, yanghaoyang8352, hdong}@buaa.edu.cn; H. Dong is also with the the School of Engineering, University of Warwick, Coventry, CV4 7AL, UK. X. Zhao is with the School of Engineering, University of Warwick, Coventry, CV4 7AL, UK. Email: xiaowei.zhao@warwick.ac.uk.

(HJB) equation. This is a challenging task and even an accurate numerical solution is hard to be obtained [21]. Besides, the highly nonlinear and coupled model of the 6-DOF dynamics of spacecraft also increase significant difficulties for this nontrivial task. The reinforcement learning (RL) technique is a promising new tool to address this challenge. RL-based control, which is commonly referred as approximate/adaptive dynamic programming (ADP) [22]–[24] in the literature, is a powerful data-driven method to solve the optimal control problems of nonlinear systems. The basic idea of RL-based control or ADP is employing special approximators (such as neural networks) to approximate the cost function as well as the optimal control scheme, and measurement data is implemented in the training process of these approximators. There are many pioneering theoretical works that have emerged based on the ADP framework for the optimal control of various systems [25]–[27]. However, the state constraint handling abilities of these notable results are still immature [28]. They cannot be straightforwardly extended to solve the COC problems for the proximity operations of spacecraft.

Motivated by these facts, a novel RL-based controller with constraint handling abilities is developed in this paper for the autonomous proximity operations of spacecraft. 6-DOF motion constraints are considered during the whole control process, including both the approach corridor and the field-of-view constraint. Compared with other traditional 6-DOF modeling methods of spacecraft motion [2], [11], [29], [30], dual quaternion can accurately represent the 6-DOF dynamics while considering the coupling between the rotational motion and transitional motion. Therefore, the dual quaternion formalism is employed to describe the 6-DOF motion of spacecraft. Then, a special dual quaternion based reward function is designed, which not only represents a trade-off between control performance and control cost but also can encode the constraint information into the controller. Besides, by making full use of the underlying properties of dual quaternion to the 6-DOF coupling motion description, an RL-based online learning algorithm is proposed to approximate the optimal control policy, improving the closed-loop performance while strictly complying with underlying motion constraints. The Lyapunov-based analysis ensures the stability of the closed-loop system. To the best knowledge of the authors, this is the first time an RL-based controller is proposed under the dual-quaternion formalism and applied to the 6-DOF constrained optimal control problem of spacecraft. The advantages of the proposed control scheme include: 1) First, the proposed method is capable of simultaneously dealing with motion constraints, control performance, and computational efficiency. 2) Second, the utilization of the dual quaternion to accurately modeling the 6-DOF dynamics in a compact form, the algorithm design can be more compact. Meanwhile, the feasibility and applicability of our model-based method can be significantly improved. 3) Finally, our method is able to rapidly endow a traditional and easy-to-implement controller with the capabilities of optimization and motion constraints handling by online tuning the weight. We also show that an easy-to-implement controller under dual-quaternion formulation can be employed as the initial control policy to trigger the learning process, and the boundedness of it is proved by a special Lyapunov strictification method.

The rest of this paper is organized as follows. In Sec. II, the concept and operations of the dual quaternion are introduced, and the SPO control problem is introduced based on the dual-quaternion 6-DOF model. Subsequently, Sec. III gives the design of the reward function and the development of the RL-based control scheme and the initial control policy. Numerical simulations and analysis for illustrating the superiority of the proposed method are presented in Sec. IV. Finally, the paper offers some concluding remarks in Sec. V.

Notation: Throughout the paper, $\mathbb{R}^{n\times m}$ denotes the set of $n\times m$ real matrix and $\hat{\mathbb{R}}^{n\times m}$ denotes the corresponding dimensional dual matrix. Post superscript $(\cdot)^{\times}$ denotes the skew-symmetric matrices of three dimensional vectors, and $(\cdot)^{a}$ indicates the corresponding vector expressed in frame $\mathcal{F}_A$. The n-dimensional identity matrix represented as $\boldsymbol{I}_n$, and $\mathbf{0}_{n\times m}$ is $n\times m$ zero matrix, $\mathbf{1}_{n\times m}$ denotes $n\times m$ one matrix.

II. Preliminaries and Problem Formulation

In this section, a brief background of dual quaternion based dynamics and motion constraints will be discussed. More details of dual quaternions can be found in [31]–[33].

A. Definition of Coordinate Frames

As illustrated in Fig. 1, the reference frames employed in this paper include the target coordinate frame $\mathcal{F}_T = \{O_T, X_T, Y_T, Z_T\}$, the inertial coordinate frame $\mathcal{F}_N = \{O_N, X_N, Y_N, Z_N\}$, and the body-fixed frame $\mathcal{F}_B = \{O_B, X_B, Y_B, Z_B\}$. For arbitrary vector $\boldsymbol{a} \in \mathbb{R}^3$ can be described as $\boldsymbol{a}^b$, $\boldsymbol{a}^t$, and $\boldsymbol{a}^n$ in $\mathcal{F}_B$, $\mathcal{F}_T$ and $\mathcal{F}_N$, respectively.

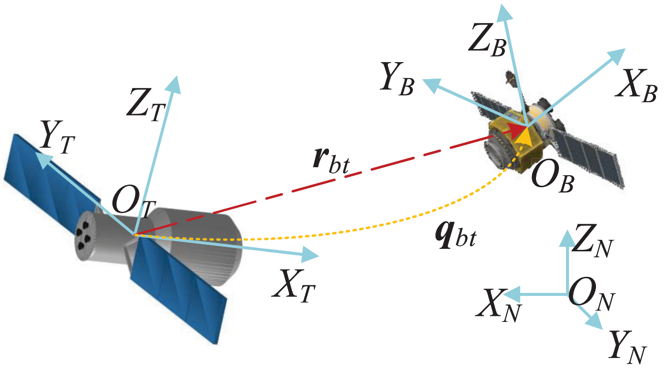


Figure 1. Illustration of coordinate frames.

B. Quaternions and Dual Quaternions

The unit quaternion is widely adopted to represent the relative attitude between two different coordinate frames. A unit quaternion is defined as $\boldsymbol{q} = \left[q_0, \boldsymbol{q}_v^T\right]^T \in \mathbb{Q}$ (the set of unit quaternion, $\mathbb{Q} = \left\{\boldsymbol{q} \in \mathbb{R}^3 \times \mathbb{R} \mid \|\boldsymbol{q}\| = 1\right\}$), which is composed of the vector part $\boldsymbol{q}_v = [q_1, q_2, q_3]^T = \sin\frac{\theta}{2}\boldsymbol{n}$ and the scalar part $q_0 = \cos\frac{\theta}{2}$. The eigen-rotation axis is denoted by $\boldsymbol{n}$ and $\theta \in (-\pi, \pi]$ is the rotation angle around this axis. The relative attitude between $\mathcal{F}_B$ and $\mathcal{F}_T$ denoted by $\boldsymbol{q}_{bt}$.

Then some important operations of unit quaternions are given as follows:

$$\mathbf{q}\otimes\mathbf{p}=\left[\begin{array}{c} q_0p_0-\boldsymbol{q}_v^T\boldsymbol{p}_v \\ q_0\boldsymbol{p}_v+p_0\boldsymbol{q}_v+\boldsymbol{q}_v\times\boldsymbol{p}_v \end{array}\right]\triangleq[\mathbf{q}]_{\otimes}\,\mathbf{p} \tag{1}$$

where

$$[\mathbf{q}]_{\otimes}=\left[\begin{array}{cc} q_0 & -\boldsymbol{q}_v^T \\ \boldsymbol{q}_v & \boldsymbol{q}_v^{\times}+q_0\mathbf{I}_3 \end{array}\right],\boldsymbol{q}_v^{\times}=\left[\begin{array}{ccc} 0 & -q_3 & q_2 \\ q_3 & 0 & -q_1 \\ -q_2 & q_1 & 0 \end{array}\right]$$

$$\boldsymbol{q}^*=\left[q_0,-\boldsymbol{q}_v^T\right]^T,\ \text{and}\ \boldsymbol{q}^*\otimes\boldsymbol{q}\otimes\boldsymbol{p}=\boldsymbol{p} \tag{2}$$

for all $\boldsymbol{q},\boldsymbol{p}\in\mathbb{Q}$. Also, for $\boldsymbol{a}\in\mathbb{R}^3$ and $\boldsymbol{q}\in\mathbb{Q}$, define the multiplication:

$$\mathbf{q}\otimes\mathbf{a}\triangleq\mathbf{q}\otimes\left[0,\mathbf{a}^T\right]^T \tag{3}$$

The coordinate transformation of any vector $\boldsymbol{a}\in\mathbb{R}^3$ from frame $\mathcal{F}_T$ to frame $\mathcal{F}_B$, can be represented by the following equation:

$$\boldsymbol{a}^b=\boldsymbol{q}_{bt}^*\otimes\boldsymbol{a}^t\otimes\boldsymbol{q}_{bt} \tag{4}$$

The concept of dual quaternions was developed by Clifford and Study [34], [35]. Before introducing the dual quaternions, the definitions of dual numbers and dual vectors should be given first.

A dual vector (or number) defined as $\hat{\boldsymbol{a}}=\boldsymbol{a}_r+\varepsilon\boldsymbol{a}_d$, where $\boldsymbol{a}_r,\ \boldsymbol{a}_d\in\hat{\mathbb{R}}^m$ ($m=1$ in number case) denote the real part and the dual part, respectively. In addition, ε is called the dual unit, satisfying $\varepsilon\neq0$ but $\varepsilon^2=0$. The swap operation of a dual vector defined as: $\hat{\boldsymbol{a}}^s=\boldsymbol{a}_d+\varepsilon\boldsymbol{a}_r$.

Dual quaternions can be regarded as the combination of dual numbers and regular quaternions comprised of two quaternions: real part and dual part. The dual quaternion $\hat{\boldsymbol{q}}_{bt}\in\hat{\mathbb{Q}}$ (set of dual quaternions) of frame $\mathcal{F}_T$ with respect to frame $\mathcal{F}_B$ is defined as:

$$\hat{\boldsymbol{q}}_{bt}=\boldsymbol{q}_{bt}+\varepsilon\frac{1}{2}\boldsymbol{q}_{bt}\otimes\boldsymbol{r}_{bt}^b=\boldsymbol{q}_{bt}+\varepsilon\frac{1}{2}\boldsymbol{r}_{bt}^t\otimes\boldsymbol{q}_{bt} \tag{5}$$

where $\boldsymbol{r}_{bt}^b$ and $\boldsymbol{r}_{bt}^t$ are the relative position vector of $\mathcal{F}_B$ with respect to $\mathcal{F}_T$ and expressed in frames $\mathcal{F}_B$ and $\mathcal{F}_T$ respectively. A dual quaternion can also be defined as $\hat{\boldsymbol{q}}=\left[\hat{q}_0,\hat{\boldsymbol{q}}_v^T\right]^T\in\hat{\mathbb{Q}}$. Similar to quaternion, $\hat{q}_0\in\hat{\mathbb{R}}$ and $\hat{\boldsymbol{q}}_v\in\hat{\mathbb{R}}^3$ called the scalar part and the vector part of $\hat{\boldsymbol{q}}$, respectively.

Corresponding to the operation of quaternion, some necessary operations of dual quaternion (or matrix) are introduced as follows:

$$\hat{\mathbf{q}}\otimes\hat{\mathbf{p}}=\left[\begin{array}{c} \hat{q}_0\hat{p}_0-\hat{\boldsymbol{q}}_v^T\hat{\boldsymbol{p}}_v \\ \hat{q}_0\hat{\boldsymbol{p}}_v+\hat{p}_0\hat{\boldsymbol{q}}_v+\hat{\boldsymbol{q}}_v\times\hat{\boldsymbol{p}}_v \end{array}\right]\triangleq[\hat{\mathbf{q}}]_{\otimes}\,\hat{\mathbf{p}} \tag{6}$$

where $\hat{\mathbf{q}}=\left[\hat{q}_0,\hat{\boldsymbol{q}}_v^T\right]^T,\hat{\mathbf{p}}=\left[\hat{p}_0,\hat{\boldsymbol{p}}_v^T\right]^T\in\hat{\mathbb{Q}}$, and

$$[\hat{\mathbf{q}}]_{\otimes}=\left[\begin{array}{cc} \hat{q}_0 & -\hat{\boldsymbol{q}}_v^T \\ \hat{\boldsymbol{q}}_v & \hat{\boldsymbol{q}}_v^{\times}+\hat{q}_0\mathbf{I}_3 \end{array}\right],\hat{\boldsymbol{q}}_v^{\times}=\left[\begin{array}{ccc} 0 & -\hat{q}_3 & \hat{q}_2 \\ \hat{q}_3 & 0 & -\hat{q}_1 \\ -\hat{q}_2 & \hat{q}_1 & 0 \end{array}\right]$$

$$\hat{\boldsymbol{q}}^*=\left[\hat{q}_0,-\hat{\boldsymbol{q}}_v^T\right]^T,\ \text{and}\ \hat{\boldsymbol{q}}^*\otimes\hat{\boldsymbol{q}}\otimes\hat{\boldsymbol{p}}=\hat{\boldsymbol{p}} \tag{7}$$

$$\operatorname{vec}(\hat{\boldsymbol{q}})=\hat{\boldsymbol{q}}_v \tag{8}$$

$$\hat{\boldsymbol{q}}\circ\hat{\boldsymbol{p}}=\boldsymbol{q}_r^T\boldsymbol{p}_r+\boldsymbol{q}_d^T\boldsymbol{p}_d \tag{9}$$

$$\hat{a}\odot\hat{\boldsymbol{x}}=a_r\boldsymbol{x}_r+\varepsilon a_d\boldsymbol{x}_d \tag{10}$$

$$\hat{\nabla}_{\hat{\boldsymbol{x}}}V=\frac{\partial}{\partial\boldsymbol{x}_r}V+\epsilon\frac{\partial}{\partial\boldsymbol{x}_d}V \tag{11}$$

where V $\in\mathbb{R}$ and $\hat{\boldsymbol{x}}\in\hat{\mathbb{R}}^n$

C. Kinematics and Dynamics

The target can be regarded as a stationary target during the SPO. Then, according to the dual quaternion formulation, the 6-DOF motion kinematics and dynamics of frame $\mathcal{F}_B$ with respect to frame $\mathcal{F}_T$ are given as follows:

$$\dot{\hat{\boldsymbol{q}}}_{bt}=\frac{1}{2}\hat{\boldsymbol{q}}_{bt}\otimes\hat{\boldsymbol{\omega}}_{bt}^b \tag{12}$$

$$\hat{\boldsymbol{J}}_b\dot{\hat{\boldsymbol{\omega}}}_{bt}^b=-\hat{\boldsymbol{\omega}}_{bt}^b\times(\hat{\boldsymbol{J}}_b\hat{\boldsymbol{\omega}}_{bt}^b)+\hat{\boldsymbol{u}} \tag{13}$$

where $\hat{\boldsymbol{\omega}}_{bt}^b=\boldsymbol{\omega}_{bt}^b+\varepsilon\boldsymbol{v}_{bt}^b$ is the relative dual angular velocity between the target and chaser, then $\boldsymbol{\omega}_{bt}^b,\boldsymbol{v}_{bt}^b$ denote angular velocity and translational velocity between target and chaser represented in $\mathcal{F}_B$, respectively. Furthermore, $\hat{\boldsymbol{u}}=\boldsymbol{f}^b+\varepsilon\boldsymbol{\tau}^b$ is called the total dual control input applied to the spacecraft, and here $\boldsymbol{f}^b,\boldsymbol{\tau}^b\in\mathbb{R}^3$ are the force and the torque applied to the chaser spacecraft, respectively. $\hat{J}_b$ is the dual inertia of the spacecraft, with the definition:

$$\hat{\boldsymbol{J}}_b=m_b\boldsymbol{I}_3\frac{d}{d\varepsilon}+\varepsilon\boldsymbol{J}_b \tag{14}$$

where $m_b\in\hat{\mathbb{R}}$ is the mass of spacecraft and $\boldsymbol{J}_b\in\hat{\mathbb{R}}^{3\times3}$ represent the inertial tensor of spacecraft. Then the inverse of $\hat{\boldsymbol{J}}_b$ is defined as follows:

$$\hat{\boldsymbol{J}}_b^{-1}=\boldsymbol{J}_b^{-1}\frac{d}{d\varepsilon}+\varepsilon\frac{1}{m_b}\boldsymbol{I}_3 \tag{15}$$

D. Motion Constraints

During the approaching stage, the chaser spacecraft should comply with both the approach corridor and field-of-view constraints. In this part, the aforementioned constraints are discussed in detail.

1) Field-of-View Constraint: The field-of-view constraint is caused by the limit field-of-view of the optical instruments. To ensure the target can be captured by the chaser spacecraft during the mission, the angle of line-of-sight should be restricted [36]. Then the field-of-view constraint can be defined as a cone around the line-of-sight in the body frame, as shown in Fig. 2. In the illustration, unit vector $\boldsymbol{c}_{sight}$ denotes the central line-of-sight of the vision sensor system in the body frame, and α_{sight} represents the maximum allowable line-of-sight angle. To satisfy this constraint, the angle between $\boldsymbol{c}_{sight}$ and $-\boldsymbol{r}_{bt}$ should never greater than α, formulating as follows:

$$\frac{-(\boldsymbol{r}_{bt}^b)^T}{\|\boldsymbol{r}_{bt}^b\|}\boldsymbol{c}_{sight}\geq\cos\alpha_{sight} \tag{16}$$

Aided by the property of quaternions, one has:

$$(\boldsymbol{r}_{bt}^b)^T\boldsymbol{c}_{sight}=(\boldsymbol{r}_{bt}^b\otimes\boldsymbol{q}_{bt})^T(\boldsymbol{c}_{sight}\otimes\boldsymbol{q}_{bt}) \tag{17}$$

then Eq.(16) can be further reformulated as:

$$c_1\triangleq-\frac{\hat{\boldsymbol{q}}_{bt}\circ(\hat{\boldsymbol{\Theta}}_{sight}\hat{\boldsymbol{q}}_{bt})}{\|2\varepsilon\circ\hat{\boldsymbol{q}}_{bt}^s\|}-\cos\alpha_{sight}\geq0 \tag{18}$$

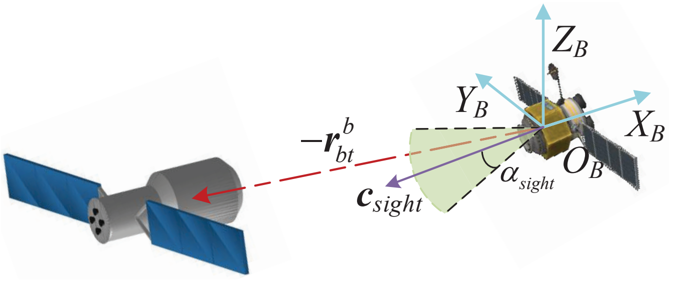


Figure 2. Illustration of field-of-view constraint

where $\boldsymbol{\Theta}_{sight} = \begin{bmatrix} 0 & \boldsymbol{c}_{sight} \\ \boldsymbol{c}_{sight} & -(\boldsymbol{c}_{sight})^{\times} \end{bmatrix}$, and $\hat{\boldsymbol{\Theta}}_{sight} = \boldsymbol{\Theta}^{T}_{sight}\frac{\mathrm{d}}{\mathrm{d}\varepsilon} + \varepsilon\boldsymbol{\Theta}_{sight}$. Thus, when $c_1 \geq 0$ the field-of-view constraint can be guaranteed.

2) Approach Corridor Constraint: In the actual missions, the chaser spacecraft should keep in the designated zone to avoid obscuring the sightline of observing docking port, as well as collisions with the components of the target. For avoiding this problem, the chaser spacecraft should approach the docking port from a direction. The approach corridor constraint is defined as a cone around the central axis (denoted as $\boldsymbol{c}_{path}$) of docking port that lies in the frame $\mathcal{F}_T$, as shown in Fig. 3. The half-angle of cone represented by α_{path}. To satisfy this constraint, one has:

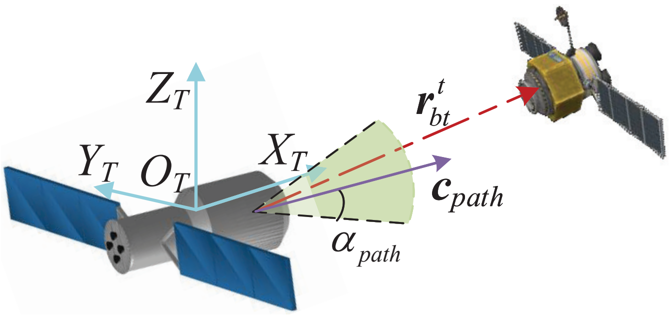


Figure 3. Illustration of approach corridor constraint

$$\frac{(\boldsymbol{r}_{bt}^{t})^{T}}{\|\boldsymbol{r}_{bt}^{t}\|}\boldsymbol{c}_{path} \geq \cos\alpha_{path} \tag{19}$$

Similar to the process in Sec. II-D1, Eq.(20) can be guaranteed by the following inequation:

$$c_2 \triangleq \frac{\hat{\boldsymbol{q}}_{bt} \circ (\hat{\boldsymbol{\Theta}}_{path}\hat{\boldsymbol{q}}_{bt})}{\|2\varepsilon \circ \hat{\boldsymbol{q}}_{bt}^{s}\|} - \cos\alpha_{path} \geq 0 \tag{20}$$

where $\boldsymbol{\Theta}_{path} = \begin{bmatrix} 0 & \boldsymbol{c}_{path} \\ \boldsymbol{c}_{path} & -(\boldsymbol{c}_{path})^{\times} \end{bmatrix}$, and $\hat{\boldsymbol{\Theta}}_{path} = \boldsymbol{\Theta}^{T}_{path}\frac{\mathrm{d}}{\mathrm{d}\varepsilon} + \varepsilon\boldsymbol{\Theta}_{path}$

E. Control Objective

The objective is to develop an online learning control scheme to achieve the control law evolution for performance optimization in the SPO mission, in the presence of approach corridor and orientation constraints.

III. Controller Design

A. State Reward Function Design

Before proceeding, we first discuss the reward function. The reward function is the feedback of the environment while agents are implementing the corresponding action. The use of a reward signal to formalize the idea of a goal is one of the most distinctive features of reinforcement learning [22]. The basic idea of the reward function design is that giving a high reward (present as a small value herein) to desired states and a low reward to undesired states (a large value).

Further, according to the analysis in Sec. II-D, the reward function associated with the undesired state is designed as:

$$\Upsilon_{sight} = -\beta_1(\hat{\boldsymbol{q}}_{bt}-\hat{\boldsymbol{q}}_I)\circ(\hat{\boldsymbol{Q}}_q(\hat{\boldsymbol{q}}_{bt}-\hat{\boldsymbol{q}}_I))\log\left(\frac{c_1}{1-\cos\alpha_{sight}}\right) \tag{21}$$

$$\Upsilon_{path} = -\beta_2(\hat{\boldsymbol{q}}_{bt}-\hat{\boldsymbol{q}}_I)\circ(\hat{\boldsymbol{Q}}_q(\hat{\boldsymbol{q}}_{bt}-\hat{\boldsymbol{q}}_I))\log\left(\frac{c_2}{1-\cos\alpha_{path}}\right) \tag{22}$$

where $\hat{\boldsymbol{q}}_I = \boldsymbol{q}_I + \epsilon\boldsymbol{0}_{4\times1}$, $\boldsymbol{q}_I = [1,0,0,0]^T$ is identity quaternion, and β_1, β_2 are the scale factors interpreted as the 'level' of reward, $\hat{\boldsymbol{Q}}_q = \boldsymbol{Q}_q\frac{\mathrm{d}}{\mathrm{d}\varepsilon} + \varepsilon\boldsymbol{Q}_q$ is the dual weight matrix. Note that, when the highest reward is meeting the target is at the center of the spacecraft's field view. Contrarily, the reward will rapid decline when the target close to the edge of the spacecraft's field view. Similarly, the center of the approach corridor corresponds to the high reward and the edge corresponds to the low reward.

The desired states are set as the target's states, the relevant reward function defined by the form of error dual quaternion and dual angular velocity given by:

$$\Upsilon_{state} = (\hat{\boldsymbol{q}}_{bt}-\hat{\boldsymbol{q}}_I)\circ(\hat{\boldsymbol{Q}}_q(\hat{\boldsymbol{q}}_{bt}-\hat{\boldsymbol{q}}_I)) + (\hat{\boldsymbol{\omega}}_{bt}^{b})\circ(\hat{\boldsymbol{Q}}_\omega\hat{\boldsymbol{\omega}}_{bt}^{b}) \tag{23}$$

where $\hat{\boldsymbol{Q}}_\omega = \boldsymbol{Q}_\omega\frac{\mathrm{d}}{\mathrm{d}\varepsilon} + \varepsilon\boldsymbol{Q}_\omega$ is also dual weight matrix. The balance between the reward of dual quaternion and dual angular velocity can be adjusted by tuning the dual weight matrix $\hat{\boldsymbol{Q}}_q, \hat{\boldsymbol{Q}}_\omega$. Evidently, according to (23) the distance from target's states relate to the level of this reward.

Remark 1. *It is noteworthy that, although the design ideas of reward functions is given a high "penalty" in the prohibited area, it is distinct from the APF-based method (e.g., [10], [14]), the control signal not only related to the current "penalty", but also to the throughout the whole process's "penalty", which will be reflected in the next part. Furthermore, the factors* $(-1+\cos\alpha_{sight})$ *and* $(-1+\cos\alpha_{path})$ *are introduced in (21) and (22), respectively, for adjusting the logarithm operation maps into* $[0,+\infty)$. *Thus, there is no penalty at the most desired state (that is* $\Upsilon_{sight\backslash path} = 0$*).*

Summing up the above analysis, the reward functions are constructed by (24), considering both desired and undesired states during the SPO mission by mapping the states into the corresponding value.

$$\Upsilon = \underbrace{\Upsilon_{state}}_{\text{desired states}} + \underbrace{\Upsilon_{path} + \Upsilon_{sight}}_{\text{undesired states}} \tag{24}$$

B. Optimal Control Solution Analysis

After designing the reward functions, the optimal control solution analysis will be discussed in this part. To formalize the optimal control problem, the model of spacecraft Eqs.(12),(13) are rewritten as the compact form:

$$\dot{\hat{\boldsymbol{x}}} = \hat{\boldsymbol{F}} + \boldsymbol{G}\hat{\boldsymbol{u}} \tag{25}$$

where $\hat{\boldsymbol{x}} = \left[\hat{e}^T, (\hat{\boldsymbol{J}}_b\hat{\boldsymbol{\omega}}_{bt})^T\right]^T$ is motion state, with $\hat{e} = Vec(\hat{\boldsymbol{q}}_{bt}^s - \hat{\boldsymbol{q}}_I^s)$, and

$$\hat{\boldsymbol{F}} = \begin{bmatrix} \frac{1}{2}Vec(\hat{\boldsymbol{q}}_{bt} \otimes \hat{\boldsymbol{\omega}}_{bt}^b)^s \\ -\hat{\boldsymbol{\omega}}_{bt}^b \times (\hat{\boldsymbol{J}}_b\hat{\boldsymbol{\omega}}_{bt}^b) \end{bmatrix}, \boldsymbol{G} = \begin{bmatrix} \boldsymbol{0}_{3\times3} \\ \boldsymbol{I}_3 \end{bmatrix}$$

In space missions, control cost is also a considerable factor due to the high cost of the energy. The control cost and state error should be both considered in policy design. Therefore, the cost function of the optimal control $V(\hat{\boldsymbol{x}})$ is defined as the integral of the non-negative reward function $r(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}) = \Upsilon(\hat{\boldsymbol{x}}) + \hat{\boldsymbol{u}} \circ \hat{\boldsymbol{u}}$ will be discussed in the following part.

$$V(\hat{\boldsymbol{x}}) = \int_t^\infty r(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}})dt \tag{26}$$

The optimal control policy is $\hat{\boldsymbol{u}}^*$ (if exist), thus the corresponding cost function is denoted by $V^*(\hat{\boldsymbol{x}})$. Then $\hat{\boldsymbol{u}}^*$ satisfies

$$H\left(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}^*, \hat{\nabla}_{\hat{\boldsymbol{x}}}V^*\right) = 0 \tag{27}$$

where the Hamilton equation is defined by

$$H\left(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}^*, \hat{\nabla}_{\hat{\boldsymbol{x}}}V^*\right) = \hat{\nabla}_{\hat{\boldsymbol{x}}}V^* \circ (\hat{\boldsymbol{F}} + \boldsymbol{G}\hat{\boldsymbol{u}}) + r(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}^*) \tag{28}$$

Further taking partial differential for (28), the closed-form of $\hat{\boldsymbol{u}}^*$ can be deduced as:

$$\hat{\boldsymbol{u}}^* = -\frac{1}{2}\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}V^* \tag{29}$$

Then substituting (29) back into (28) leads to the following HJB equation:

$$\hat{\nabla}_{\hat{\boldsymbol{x}}}V^* \circ \hat{\boldsymbol{F}} + \Upsilon - \frac{1}{4}(\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}V^*) \circ (\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}V^*) = 0 \tag{30}$$

Note that, the high nonlinearity of the system model (25) increases the intractability of analytically solving the HJB equation (30). Hence, an RL-based online controller will be designed to approximate the optimal solution $\boldsymbol{u}^*$ in the following part.

C. Online Learning Control Algorithm

As aforementioned, such a high nonlinearity of the cost function (26) makes the HJB equation (30) hard to solve. Approximation emerges as a way to deal with it. According to the Weierstrass Approximation theorem [25], [37] a NN that contains a sufficient set of basis functions can be employed to approximate the optimal cost function (26), given by

$$V^*(\hat{\boldsymbol{x}}) = \boldsymbol{w}^T\boldsymbol{\sigma}(\hat{\boldsymbol{x}}) + \epsilon(\hat{\boldsymbol{x}}) \tag{31}$$

for $\hat{\boldsymbol{x}} \in \mathcal{X}$, where $\mathcal{X} \subset \hat{\mathbb{R}}^6$ is a compact set. The basis function vector, denoted by $\boldsymbol{\sigma}(\hat{\boldsymbol{x}}) = [\sigma_1(\hat{\boldsymbol{x}}), \sigma_2(\hat{\boldsymbol{x}}), \ldots, \sigma_p(\hat{\boldsymbol{x}})]^T \in \mathbb{R}^p$, satisfies that:

$$\begin{array}{l} \sigma_i(\hat{\boldsymbol{0}}_{6\times1}) = 0 \\ \dot{\sigma}_i(\hat{\boldsymbol{0}}_{6\times1}) = 0 \end{array}, \qquad i = 1, 2, \ldots, p$$

The weight vector of basis function $\boldsymbol{w}$ is a unknown constant vector, and $\epsilon(\hat{\boldsymbol{x}}) \in \mathbb{R}$ is the reconstruction error. Then Eq.(29) can be equivalently rewritten by

$$\hat{\boldsymbol{u}}^* = -\frac{1}{2}\boldsymbol{G}^T(\hat{\nabla}_{\hat{\boldsymbol{x}}}\boldsymbol{\sigma}(\hat{\boldsymbol{x}})\boldsymbol{w} + \hat{\nabla}_{\hat{\boldsymbol{x}}}\epsilon(\hat{\boldsymbol{x}})) \tag{32}$$

Based on the RL technique, the function of actor-critic is to online approximate the weigh vector $\boldsymbol{w} \in \mathbb{R}^p$. Then a weight estimation vector $\boldsymbol{w}_{est}$ was employed to construct the estimate the cost function and control policy:

$$V(\hat{\boldsymbol{x}}, \boldsymbol{w}_{est}) = \boldsymbol{w}_{est}^T\boldsymbol{\sigma}(\hat{\boldsymbol{x}}) \tag{33}$$

$$\hat{\boldsymbol{u}} = -\frac{1}{2}\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\boldsymbol{\sigma}\boldsymbol{w}_{est} \tag{34}$$

Subsequently, further consider the following Bellman error:

$$\delta_b = \hat{\nabla}_{\hat{\boldsymbol{x}}}V \circ (\hat{\boldsymbol{F}} + \boldsymbol{G}\hat{\boldsymbol{u}}) + r(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}) \tag{35}$$

Recalling (30), the Bellman error can be rewritten as follows:

$$\begin{aligned} \delta_b =& \delta_b - H\left(\hat{\boldsymbol{x}}, \hat{\boldsymbol{u}}^*, \hat{\nabla}_{\hat{\boldsymbol{x}}}V^*\right) \\ =& \tilde{\boldsymbol{w}}^T\boldsymbol{\vartheta} + \epsilon_\delta \end{aligned} \tag{36}$$

where $\boldsymbol{\vartheta} = \hat{\nabla}_{\hat{\boldsymbol{x}}}\boldsymbol{\sigma} \circ (\hat{\boldsymbol{F}} + \boldsymbol{G}\hat{\boldsymbol{u}})$ is defined for expressing simplicity; $\tilde{\boldsymbol{w}} = \boldsymbol{w}_{est} - \boldsymbol{w}$ is the weight error, and ϵ_δ denotes the induced reconstruction error.

It can be noticed that Eq.(36) contains the information of $\tilde{\boldsymbol{w}}$, it has been commonly employed to design the learning law of the estimated weight $\boldsymbol{w}_{est}$. Specially, not only the real-time information of δ_b but also the past measurements are utilized in this paper. Before proceeding further, we make the following assumptions.

Assumption 1. *For $x \in \mathcal{X}$, there exist positive constants $b_\sigma, b_{\nabla_\sigma}$ and b_{ϵ_δ}, such that, $\|\sigma\| \leq b_\sigma$, $\|\nabla_x\sigma\| \leq b_{\nabla\sigma}$, and $\epsilon_\delta \leq b_{\epsilon_\delta}$.*

Assumption 2. *Introduce an auxiliary variable defined by $\boldsymbol{\eta} = \boldsymbol{\vartheta}/(\boldsymbol{\vartheta}^T\boldsymbol{\vartheta}+1)$, it satisfies a finite excitation (FE) condition [38], i.e., there exist t_{k1}, t_{k2}, γ_w with $0 \leq t_{k1} \leq t_{k2} \leq t$ and γ_w such that $\int_{t_{k1}}^{t_{k2}} \boldsymbol{\eta}(\tau)\boldsymbol{\eta}^T(\tau)\mathrm{d}\tau \geq \gamma_w I_p$.*

Note that, Assumption 1 is a standard assumption. Assumption 2 is much weaker than the conventional persistent excitation PE assumptions in online RL-based controllers in [39].

Afterward, we introduced an auxiliary variable $\boldsymbol{\Psi}$ here to utilize the online data designed as follows:

$$\boldsymbol{\Psi}(t, t_{k2}, t_{k1}) = \boldsymbol{\psi}_1(t_{k2}, t_{k1})\hat{\boldsymbol{w}}_{est} + \boldsymbol{\psi}_2(t_{k2}, t_{k1}) \tag{37}$$

with

$$\dot{\boldsymbol{\psi}}_1(t, t_{k1}) = -\kappa\boldsymbol{\psi}_1(t, t_{k1}) + \boldsymbol{\varphi}_1(t) \tag{38}$$

$$\dot{\boldsymbol{\psi}}_2(t, t_{k1}) = -\kappa\boldsymbol{\psi}_2(t, t_{k1}) + \boldsymbol{\varphi}_2(t) \tag{39}$$

where $\boldsymbol{\psi}_1(t_{k1}) = \mathbf{0}_{p\times p}$, $\boldsymbol{\psi}_2(t_{k1}) = \mathbf{0}_{p\times 1}$, $\boldsymbol{\varphi}_1 = \boldsymbol{\eta}\boldsymbol{\eta}^T$, $\boldsymbol{\varphi}_3 = \boldsymbol{\eta}/(\boldsymbol{\vartheta}^T\boldsymbol{\vartheta}+1)$, $\boldsymbol{\varphi}_2 = r\boldsymbol{\varphi}_3$, and κ is a positive constant. According to (37)-(39), one has

$$\begin{aligned}\boldsymbol{\Psi}(t, t_{k2}, t_{w1}) &= \int_{t_{k1}}^{t_{k2}} \mathrm{e}^{\kappa(\tau - t_{k2})}(\boldsymbol{\varphi}_1(\tau)\boldsymbol{w}_{est} + \boldsymbol{\varphi}_2(\tau))\mathrm{d}\tau \\ &= \boldsymbol{Y}(t_{k2}, t_{k1})\tilde{\boldsymbol{w}} + \boldsymbol{\epsilon}_\Psi\end{aligned} \tag{40}$$

where $\boldsymbol{Y}(t_{k2}, t_{k1}) = \int_{t_{k1}}^{t_{k2}} \mathrm{e}^{\kappa(\tau-t_{k2})}\boldsymbol{\varphi}_1 \mathrm{d}\tau$ is an information matrix, which "stores" the information of $\boldsymbol{\eta}$ throughout the time interval $[t_{k1}, t_{k2}]$ and the residual error vector is denoted by $\boldsymbol{\epsilon}_\Psi = \int_{t_{k1}}^{t_{k2}} \mathrm{e}^{\kappa(\tau-t_{k2})}\epsilon_\delta\boldsymbol{\vartheta}/(\boldsymbol{\vartheta}^T\boldsymbol{\vartheta}+1)^2\mathrm{d}\tau$. Furthermore, under Assumption 2, one has $\boldsymbol{Y}(t_{k2}, t_{k1}) \geq \mathrm{e}^{-\kappa(t_{k2}-t_{k1})}\gamma_w\boldsymbol{I}_m \doteq \gamma_\Psi\boldsymbol{I}_m$

Introducing the above auxiliary variable (40) into the learning law of $\boldsymbol{w}_{est}$ is significantly beneficial to improve learning efficiency. Nevertheless, considering $\boldsymbol{Y}(t_{k2}, t_{k1})$ is positive-define only one has sufficient online data is collected, it is necessary to ensure the boundedness of state first. With regard to this, the following theorem is given as a solution.

Theorem 1. *Consider the system defined in (25), and the policy described in (34). With Assumption 2, design the learning law of $\boldsymbol{w}_{est}$ as:*

$$\dot{\boldsymbol{w}}_{est} = -\gamma_1\delta_b\boldsymbol{\varphi}_3 - \gamma_2\boldsymbol{\Psi}(t, t_{k2}, t_{k1}) \tag{41}$$

where γ_1 and γ_2 are positive constants. Then, the estimated weight $\tilde{\boldsymbol{w}}$, the state $\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I$ and $\hat{\boldsymbol{\omega}}_{bt}$ are ultimately bounded, if the condition (44) is satisfied.

Proof. Consider the following storage function:

$$\mathcal{L} = V^* + \frac{a_1}{2}\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}} \tag{42}$$

where $a_1 > 0$ is a constant. Then taking the time derivative of (42) along (26) and (41) yield:

$$\begin{aligned}\dot{\mathcal{L}} &= \hat{\nabla}_{\hat{\boldsymbol{x}}}V \circ (\hat{\boldsymbol{F}} + \boldsymbol{G}\hat{\boldsymbol{u}}) + a_1\tilde{\boldsymbol{w}}^T\dot{\tilde{\boldsymbol{w}}} \\ &= -r - \frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Gamma}\tilde{\boldsymbol{w}} - \frac{1}{4}\boldsymbol{w}^T\boldsymbol{\Gamma}\boldsymbol{w} + a_1\tilde{\boldsymbol{w}}^T\dot{\tilde{\boldsymbol{w}}} + \epsilon_1 \\ &\leq -r + \frac{1}{2}\tilde{\boldsymbol{w}}^T\boldsymbol{\Gamma}\tilde{\boldsymbol{w}} + a_1\tilde{\boldsymbol{w}}^T\dot{\tilde{\boldsymbol{w}}} + \epsilon_2 \\ &\leq -r - \tilde{\boldsymbol{w}}^T\boldsymbol{M}\tilde{\boldsymbol{w}} + \epsilon_3\end{aligned} \tag{43}$$

where $\epsilon_1 = -0.5(\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\epsilon) \circ (\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\boldsymbol{\sigma}\tilde{\boldsymbol{w}}) + 0.25(\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\epsilon) \circ (\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\epsilon)$, $\epsilon_2 = 0.5(\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\epsilon) \circ (\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\epsilon)$, $\epsilon_3 = \epsilon_2 + 0.5a_1\gamma_2\gamma_\Phi\epsilon_\delta^2 + 0.5a_1\gamma_1\epsilon_\delta^2/(\boldsymbol{\vartheta}^T\boldsymbol{\vartheta}+1)$ is a bounded value, $\boldsymbol{\Gamma} = (\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\boldsymbol{\sigma}) \circ (\boldsymbol{G}^T\hat{\nabla}_{\hat{\boldsymbol{x}}}\boldsymbol{\sigma})$, $\boldsymbol{M} = 0.5(-\boldsymbol{\Gamma} + a_1\gamma_1\boldsymbol{\eta}\boldsymbol{\eta}^T + a_1\gamma_2\gamma_\Psi\boldsymbol{I}_m)$. Recall the Assumption 1, it can be deduced that $\|\boldsymbol{\eta}\| < \frac{1}{2}$ and $\|\boldsymbol{\Gamma}\| < b_\Gamma$, where b_Γ is the a positive constant. Thus, by adjusting γ_1, γ_2 and a_1 to satisfy

$$a_1 > \frac{b_\Gamma}{\gamma_1 + 4\gamma_2\gamma_\Psi} \tag{44}$$

one has $\boldsymbol{M} > 0$. Then Eq.(43) indicates the $\tilde{\boldsymbol{w}}$ is ultimately bounded, as well as $\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I$ and $\hat{\boldsymbol{\omega}}_{bt}$. □

Remark 2. *Note that, constant a_1 is a coefficient of $\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}$ which is employed just for convergence analysis purpose. Therefore, we do not need to set a value to it. As long as there is an a_1 that satisfies the condition (44), a function then can be constructed to guarantee the convergence of the entire system. Hence, in practical applications, parameters γ_1 and γ_2 can be chosen by an empirical way according to the actual situation.*

D. Initial Control Policy Design

An initial control policy is required to ensure the system states to a compact set $\mathcal{X}$, and it must be represented by the basis functions $\boldsymbol{\sigma}(\hat{\boldsymbol{x}})$. For the spacecraft SPO problem considered herein, the initial policy designed in (45) has the capability of meeting the requirement.

$$\hat{\boldsymbol{u}}_{init} = \hat{k}_p \odot \hat{\boldsymbol{e}} - \hat{k}_d \odot (\hat{\boldsymbol{\omega}}_{bt}^b)^s \tag{45}$$

Coefficients of PD controller are positive constants denoted by $\hat{k}_p = k_{pr} + \varepsilon k_{pd}$, $\hat{k}_d = k_{dr} + \varepsilon k_{dd}$. This PD-like controller can guarantee the asymptotic convergence of system states (though it lacks optimizing and constraint handling abilities). What's more, it can be reconstructed by the following subset:

$$\boldsymbol{\sigma}_{pd}(\hat{\boldsymbol{x}}) = [\boldsymbol{e}_r^T, \boldsymbol{e}_d^T, (\boldsymbol{v}_{bt}^b)^T, (\boldsymbol{\omega}_{bt}^b)^T]^T \tag{46}$$

Thus the corresponding weights are set to be $\boldsymbol{w}_{pd} = [k_{pr}\mathbf{1}_{1\times3}, k_{pd}\mathbf{1}_{1\times3}, k_{dr}\mathbf{1}_{1\times3}, k_{dd}\mathbf{1}_{1\times3}]^T$, with $\boldsymbol{e}_r$ and $\boldsymbol{e}_d$ represent the real part and dual part of $\hat{e}$, respectively.

Remark 3. *The initial controller is designed based on the dual-quaternion framework. It is distinct to the PD-like control scheme proposed in [10], [40], in which the dual-quaternion error term is denoted by $\mathrm{vec}(\hat{\boldsymbol{q}}_{bt}^* \otimes (\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I)^s)$. Thus, this term is not suit for the initial policy in this framework (this point will be mention in the Remark 5). To deal with that, we redesigned the PD-like initial controller and give the proof by employing a special Lyapunov strictification method.*

Theorem 2. *Given the system defined in (12) and (13), consider the initial policy designed in (45). Then it can be guaranteed that $\lim_{t\to\infty}\hat{\boldsymbol{q}}_{bt}(t) = \hat{\boldsymbol{q}}_I$ and $\lim_{t\to\infty}\hat{\boldsymbol{\omega}}_{bt}^b(t) = \hat{\mathbf{0}}_{3\times1}$.*

Proof. Before proving above theorem, some properties about dual quaternions are listed as follows:

$$\hat{\boldsymbol{a}}^s \circ (\hat{\boldsymbol{a}} \times \hat{\boldsymbol{b}}) = \hat{\mathbf{0}}_3, \quad \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}} \in \hat{\mathbb{R}}^3 \tag{47}$$

$$\hat{\boldsymbol{q}}_1 \circ (\hat{\boldsymbol{q}}_2 \otimes \hat{\boldsymbol{q}}_3) = \hat{\boldsymbol{q}}_3^s \circ [\hat{\boldsymbol{q}}_2^* \otimes (\hat{\boldsymbol{q}}_1^s)], \quad \hat{\boldsymbol{q}}_1, \hat{\boldsymbol{q}}_2, \hat{\boldsymbol{q}}_3 \in \hat{\mathbb{Q}} \tag{48}$$

The detailed proofs of these properties are given in [41]. To analyze the stability of the closed-loop system, consider the following Lyapunov-like function candidate:

$$\mathcal{V}_I = \varrho\hat{k}_p^s \odot (\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I) \circ (\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I) + \frac{\varrho}{2}(\hat{\boldsymbol{\omega}}_{bt}^b)^s \circ (\hat{\boldsymbol{J}}_b\hat{\boldsymbol{\omega}}_{bt}^b) + \mathcal{N}_I \tag{49}$$

where $\mathcal{N}_I = [2\varepsilon\,\mathrm{vec}(\hat{\boldsymbol{q}}_{bt}^* \otimes (\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I)^s)] \circ [\varepsilon(\hat{\boldsymbol{J}}_b\hat{\boldsymbol{\omega}}_{bt}^b)]$ is a cross term for prove purpose. By applying Binet–Cauchy identity of cross product along with Cauchy–Schwarz inequality, one has:

$$\begin{aligned}\mathcal{V}_I \geq & 2\varrho k_{pd}(1 - q_0) + \frac{1}{2}\varrho k_{dd}(\boldsymbol{\omega}_{bt}^b)^T\boldsymbol{J}_b\boldsymbol{\omega}_{bt}^b \\ & + (\varrho\frac{k_{pr}}{4} - \frac{m}{2})\|\boldsymbol{r}_{bt}^b\|^2 + \frac{m}{2}(\varrho k_{dr} - 1)\|\boldsymbol{v}_{bt}^b\|^2\end{aligned} \tag{50}$$

Setting $\varrho > \max\{2m/k_{pr}, 1\}$, Eq.(50) guarantees $\mathcal{V}_I \geq 0$, and $\mathcal{V}_I = 0$ only when $\hat{\boldsymbol{q}}_{bt} = \hat{\boldsymbol{q}}_I$ and $\hat{\boldsymbol{\omega}}_{bt}^b = \hat{\mathbf{0}}_{3\times1}$. Therefore, $\mathcal{V}_I$ can be regarded as a valid Lyapunov-like function candidate. Then employing the above properties (47) and (48) and substituting (12) and (13) into (45), the time derivative of $\mathcal{V}$ can be written as:

$$\begin{aligned}\dot{\mathcal{V}}_I =&\varrho(\hat{\boldsymbol{\omega}}_{bt}^b)^s \circ [-\hat{k}_d^s(\hat{\boldsymbol{\omega}}_{bt}^b)^s + ((1-q_0)\boldsymbol{I}_3 + \boldsymbol{q}^\times)\boldsymbol{r}_{bt}^b + \varepsilon\mathbf{0}_{3\times1}]\\ &- k_{pr}q_0\|\boldsymbol{r}_{bt}^b\|^2 - k_{dr}(\boldsymbol{r}_{bt}^b)^T\boldsymbol{v}_{bt}^b + m\|\boldsymbol{v}_{bt}^b\|^2\\ =&-(\varrho k_{dr} - 1)m\|\boldsymbol{v}_{bt}^b\|^2 - \varrho k_{dd}\|\boldsymbol{\omega}_{bt}^b\|^2 - k_{pr}q_0\|\boldsymbol{r}_{bt}^b\|^2\\ &+ (\boldsymbol{r}_{bt}^b)^T[(\varrho - \varrho q_0 - k_{dr})\boldsymbol{I}_3 + \varrho\boldsymbol{q}_v^\times]\boldsymbol{v}_{bt}^b\\ \leq&-\varrho k_{dd}\|\boldsymbol{\omega}_{bt}^b\|^2 - (\varrho m k_{dr} - m - \frac{\mu_M}{2})\|\boldsymbol{v}_{bt}^b\|^2\\ &-(k_{pr}q_0 - \frac{\mu_M}{2})\|\boldsymbol{r}_{bt}^b\|^2\end{aligned} \tag{51}$$

with $\mu_M = \|(\varrho - \varrho q_0 - k_{dr})\boldsymbol{I}_3 + \varrho\boldsymbol{q}_v^\times\|$. Thus, by adjusting ϱ, $\hat{k}_p$ and $\hat{k}_d$ to satisfy:$k_{pr}q_0 - \frac{\mu_M}{2} \geq 0$ and $\varrho m k_{dr} - m - \frac{\mu_M}{2} \geq 0$, one has $\dot{\mathcal{V}}_I \leq 0$, and $\mathcal{V}_I = 0$ only when $\hat{\boldsymbol{q}}_{bt} = \hat{\boldsymbol{q}}_I$, $\hat{\boldsymbol{\omega}}_{bt}^b = \hat{\mathbf{0}}_{3\times1}$. Based on Barbalat's lemma [42], it can be guaranteed that $\lim_{t\to\infty}\hat{\boldsymbol{q}}_{bt}(t) = \hat{\boldsymbol{q}}_I$ and $\lim_{t\to\infty}\hat{\boldsymbol{\omega}}_{bt}^b(t) = \hat{\mathbf{0}}_{3\times1}$. □

Remark 4. *The initial policy is given as a PD-like controller for its simplicity and effectiveness, moreover, it's easy to be reconstructed by a set of simple basis functions. But the initial policy is not limited to the PD-like controller, and the basis is not limited to the simple polynomial-type basis functions. As long as an appropriate set of basis function is designed for reconstructing a given controller, it can evolve into the (sub)optimal controller during the control processing.*

Remark 5. *Reconstructing the initial control policy is tricky work, the basis functions are chosen according to the initial controller design, because it needs states are bounded at the initial stage. So the elements of the basis function should contain the element of the initial controller (such as the terms of PD). The initial controller (45) allows us to select the basis functions more convenient. We can employ the terms of the initial controller as a part of basis functions. Then, some other basis functions can be appropriately added to improve the performance of learning. Note that, it is not recommended to use the terms independent of $\hat{\boldsymbol{\omega}}_{bt}^b$ as basis functions here, which will vanish after multiply by $\boldsymbol{G}$.*

According to the above analyses, the proposed method in this paper can be intuitively summarized by diagram as shown in Fig. 4.

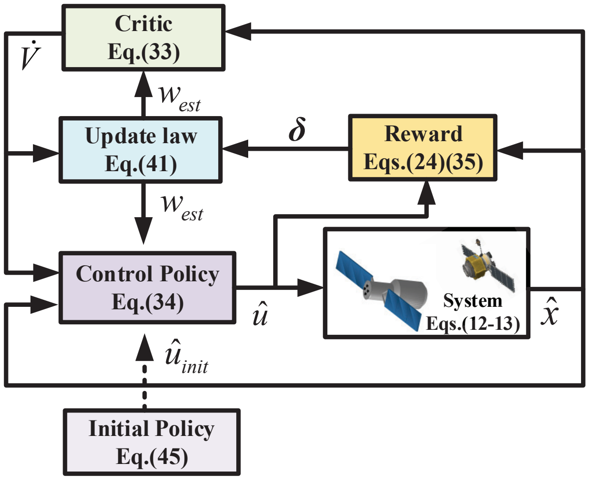


Figure 4. Structure of the system.

IV. Simulation Study

In this section, numerical simulation examples are demonstrated to illustrate the efficacy and superiority of the proposed method. The control objective is to drive the chaser spacecraft to the desired pose concerning the target spacecraft. In the simulation scenarios, the mass and inertia of the chaser spacecraft are $m = 15$ kg, $\boldsymbol{J} = diag\,[20.8, 21.1, 32.6]\ \mathrm{kg}\cdot\mathrm{m}^2$, respectively. The structure of the NN is employed as (46).

A. Case1: Point to Point Maneuver without Constraints

The first case assumes that the chaser spacecraft need to maneuver to the desired position and attitude. In this study, the initial relative quaternion and position are $\boldsymbol{q}_{bt/0} = [0.8772, 0.3426, -0.2764, 0.1918]^T$ and $\boldsymbol{r}_{bt/0}^t = [500, -185, 163]^T$ m. The desired final relative attitude and position are $\boldsymbol{q}_{bt/des} = [1, 0, 0, 0]^T$ and $\boldsymbol{r}_{bt/des}^t = [0, 0, 0]^T$ m, which renders $\hat{\boldsymbol{q}}_{bt/des} = [1, 0, 0, 0]^T + \varepsilon\,[0, 0, 0, 0]^T$. Furthermore, the initial relative dual angular velocity of the chaser is assumed to $\hat{\boldsymbol{\omega}}_{bt/0} = [0, 0, 0]^T + \varepsilon\,[0, 0, 0]^T$,and the target is considered to be stationary. The cost function chosen as: $r(\hat{\boldsymbol{q}}_{bt}, \hat{\boldsymbol{\omega}}_{bt}, \hat{\boldsymbol{u}}_b) = (\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I) \circ (\hat{\boldsymbol{Q}}_q(\hat{\boldsymbol{q}}_{bt} - \hat{\boldsymbol{q}}_I)) + (\hat{\boldsymbol{\omega}}_{bt}^b) \circ (\hat{\boldsymbol{Q}}_\omega\hat{\boldsymbol{\omega}}_{bt}^b) + \hat{\boldsymbol{u}} \circ \hat{\boldsymbol{u}}$, where $\hat{\boldsymbol{Q}}_q = \boldsymbol{I}\frac{\mathrm{d}}{\mathrm{d}\varepsilon} + \varepsilon 2\boldsymbol{I}$ and $\hat{\boldsymbol{Q}}_\omega = 5\boldsymbol{I}\frac{\mathrm{d}}{\mathrm{d}\varepsilon} + \varepsilon 10\boldsymbol{I}$.

The parameters of initial PD controller are set as: $\hat{k}_p = 0.1 + \varepsilon 0.1$, $\hat{k}_d = 5 + \varepsilon 5$. For comparison purpose, the PD-like controller is employed in this case. The initial control scheme is the same as the PD-like controller for fair comparison and demonstrating the learning result. Time responses of the relative translation and rotation state under initial controller and proposed controller are depicted in Figs. 7-8. These figures show that both control methods successfully achieve control objectives. It can be seen that under the proposed controller, the relative attitude and position converge faster than the initial controller. In particular, the performance cost of the proposed method is improved by online learning. This fact is also confirmed by the comparison of control cost under two controllers in Fig. 5.

Then, the learning process is analyzed in Fig. 6. It can been seen that the weight estimation vector $\hat{\boldsymbol{w}}_{est}$ changes quickly in the first 30s, and stabilizes after 60s. This result can be explained in the time response of Bellman error δ_b, which tends to 0 at about 60s, that means the controller tends to an optimal controller.

B. Case2: Docking to Target with Constraints

The second case assumes that the target spacecraft is relatively stationary with $\mathcal{F}_N$ during the mission, and the chaser spacecraft should be at the same motion state when close to the target with satisfying the approach corridor and field of view constraints. In this case, in order to trigger the constraints, the initial states of chaser set to

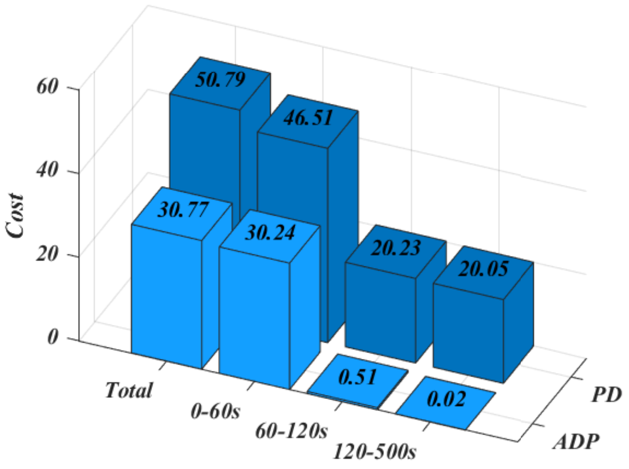


Figure 5. Comparison of performance cost between initial controller (PD) and proposed controller (ADP).

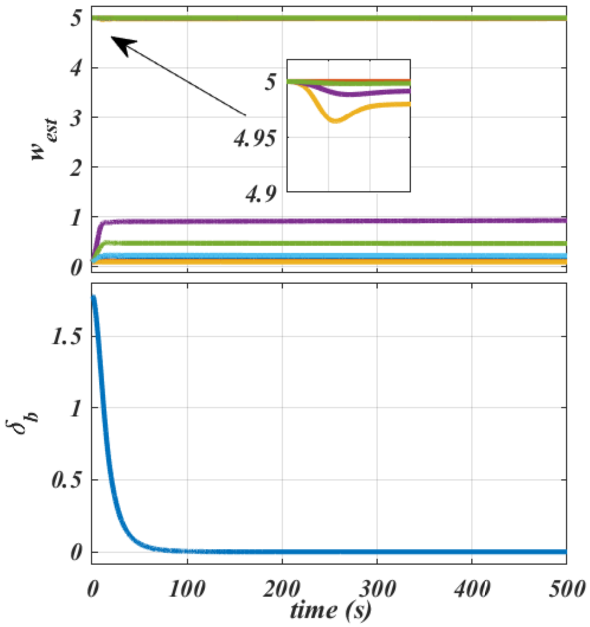


Figure 6. Simulation results of $\hat{\boldsymbol{w}}_{est}$ and δ_b.

be $\boldsymbol{q}_{b/0} = [0.8999, -0.1544, 0.0234, -0.4071]^T$, $\boldsymbol{r}^t_{b/0} = [500, -140, 250]^T$ m, $\boldsymbol{v}^b_{b/0} = [-0.5, -2.0, -1.0]^T$ m/s and $\boldsymbol{\omega}^b_{b/0} = [0, 0, 0]^T$ rad/s. The states of target set to be $\boldsymbol{q}_t = [1, 0, 0, 0]^T$, $\boldsymbol{r}^t_t = [0, 0, 0]^T$ m, $\boldsymbol{\omega}^t_t = [0, 0, 0]^T$ rad/s, and $\boldsymbol{v}^t_t = [0, 0, 0]^T$ m/s. The motion constraints parameters are set as: $c_{sight} = [-1, 0, 0]^T$, $c_{path} = [1, 0, 0]^T$, $\alpha_{path} = \alpha_{sight} = \pi/6$rad.

The parameters of initial controller set to be $\hat{k}_p = 0.2 + \varepsilon 1$, $\hat{k}_d = 5 + \varepsilon 2$. The simulation result of states, chaser's control input and learning process under the proposed method are shown in Fig. 9. It's shown that the proposed method achieves the control object under the motion constraints. To clearly show the relative motion between the target and chaser, the 3D figure drawn upon the $\mathcal{F}_T$ is given in Fig. 10. The instantaneous positions and attitude of the chaser at different simulation times are intuitively provided by the craft models, and the cone of sight and approach corridor are also drawn here.

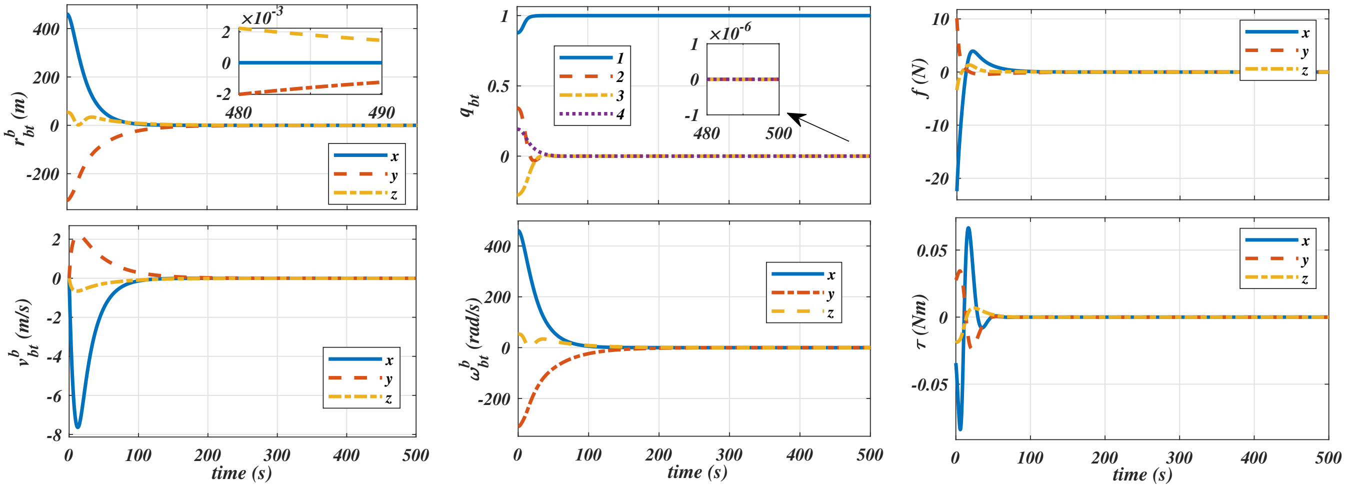
To further demonstrate the performance of the proposed method, we added the proposed method without considering constraints, APF-based method in [10], and initial control method (abbreviated as ADPF, APF, and PD, respectively, and proposed method abbreviated as ADPC in figures) into comparison. It should be emphasized that Ref. [10], to some degree, addresses a very similar problem as in this paper. It employed APF to deal with the motion constraints under the dual quaternion mechanism. However, this method shows its ability of avoiding constraints but it can not deal with performance optimization. Therefore, our method presents better control performance and task completion ability. Related comparisons results between our method and the APF-based method in Ref. [10] are shown as follows. Fig. 11 intuitively presents the translational trajectories of chaser spacecraft during the proximity operations, and the corridor approach presented by the green cone. In Figs. 12-13, the time respond of (18) and (20) are shown under 4 different methods. It can be seen from Figs. 11-13, under the PD method and ADPF method, the chaser spacecraft flies out of the cone and loses the target in its field of view. In contrast, the proposed method and APF method can guarantee that the target is always in the field of view and satisfy its translational trajectories constraints.

Both the proposed method and APF method are shown their ability of constraints handling according to above result. Then the following control effort comparison, given in Fig. 14, shows the superiority in control performance of the proposed method. Define energy consume function of force and torque as $H_f = \int \|\boldsymbol{f}^b\|^2 dt$ and $H_\tau = \int \|\boldsymbol{\tau}^b\|^2 dt$, respectively. The results show that the APF method pay more control effort for avoiding constraints violation at the first stage. The proposed method saves more of energy consumption compared to APF method.

C. Monte-Carlo Simulations

To further verify the comprehensive insight into the performance of the proposed method, a 500-run Monte-Carlo simulations are presented in this part. To this end, the initial states are randomly selected in the ranges listed in Table I. In the simulation, the PWPF modulators [43] are employed for pulse modulation. The parameters of interest are the prefilter coefficients are chosen as $K_m = 0.8$, $T_m = 0.1$, the Schmidt Trigger parameters set as $\delta_{\text{on}} = 0.45, \delta_{\text{off}} = 0.15$, and the thrust magnitude $u_{max} = 20$N.

The results of the Monte-Carlo simulations can be summarized in Figs. 15-16. Fig. 15 presents the overall of the 500-run simulations, from which we can visually see that the trajectories are converged. It can been seen from the subfigure of Fig. 15 that the results terminal error $\|r^b_{bt}(500)\|$ of each single run is lower than $10^{-0.5}$m, which is admissible in practice. The distributions of the maximum field-of-view

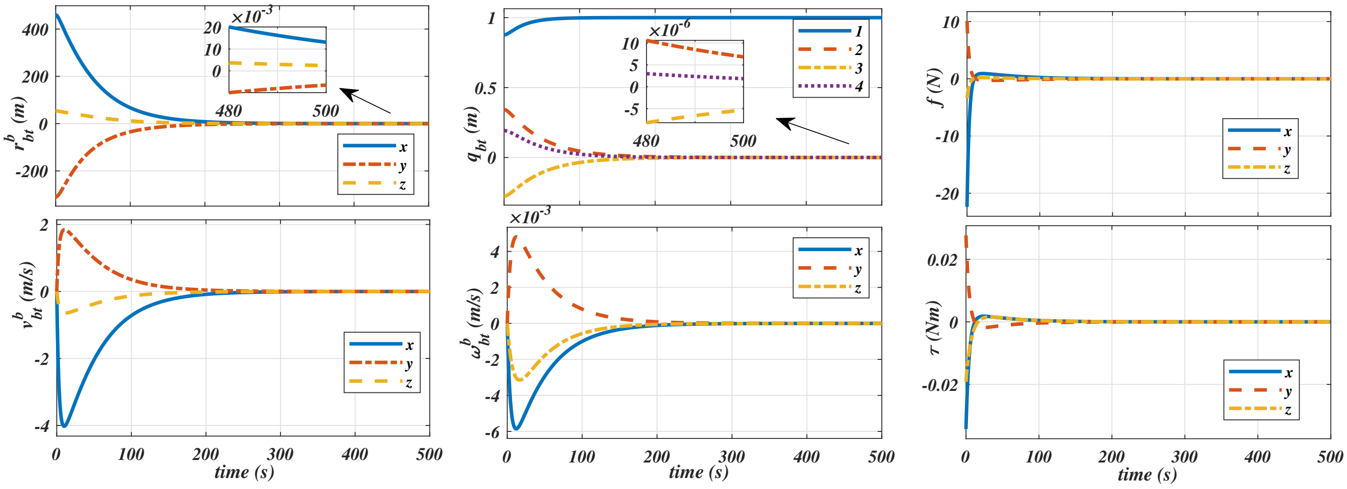


(a)

(b)

(c)

Figure 7. Simulation result under the proposed controller. (a) Simulation result of relative translation. (b) Simulation result of relative rotation. (c) Simulation result of chaser's control inputs

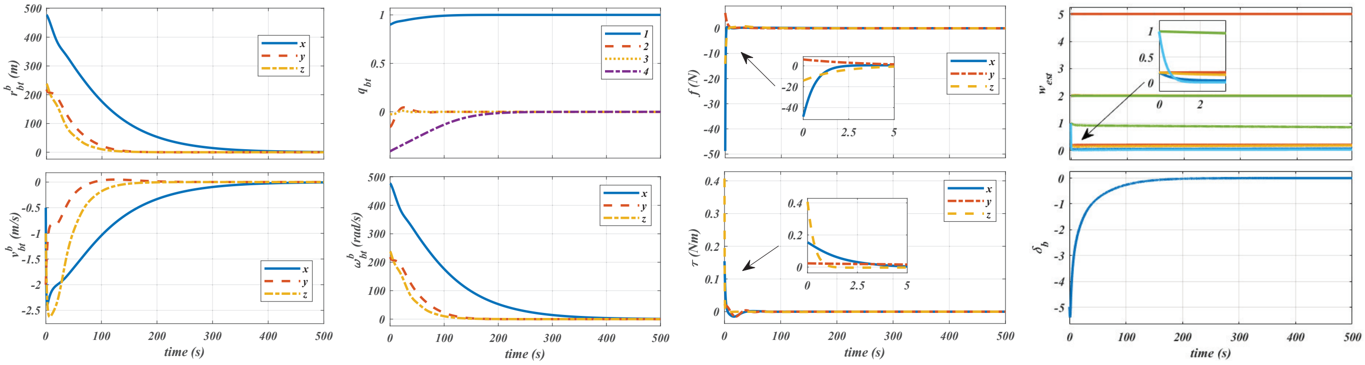


(a)

(b)

(c)

Figure 8. Simulation result under the initial controller (a) Simulation result of relative translation. (b) Simulation result of relative rotation. (c) Simulation result of chaser's control inputs



(a)

(b)

(c)

(d)

Figure 9. Simulation result under the proposed controller. (a) Simulation result of relative translation. (b) Simulation result of relative rotation. (c) Simulation result of chaser's control inputs (d) Simulation results of $\hat{\boldsymbol{w}}_{est}$ and δ_b.

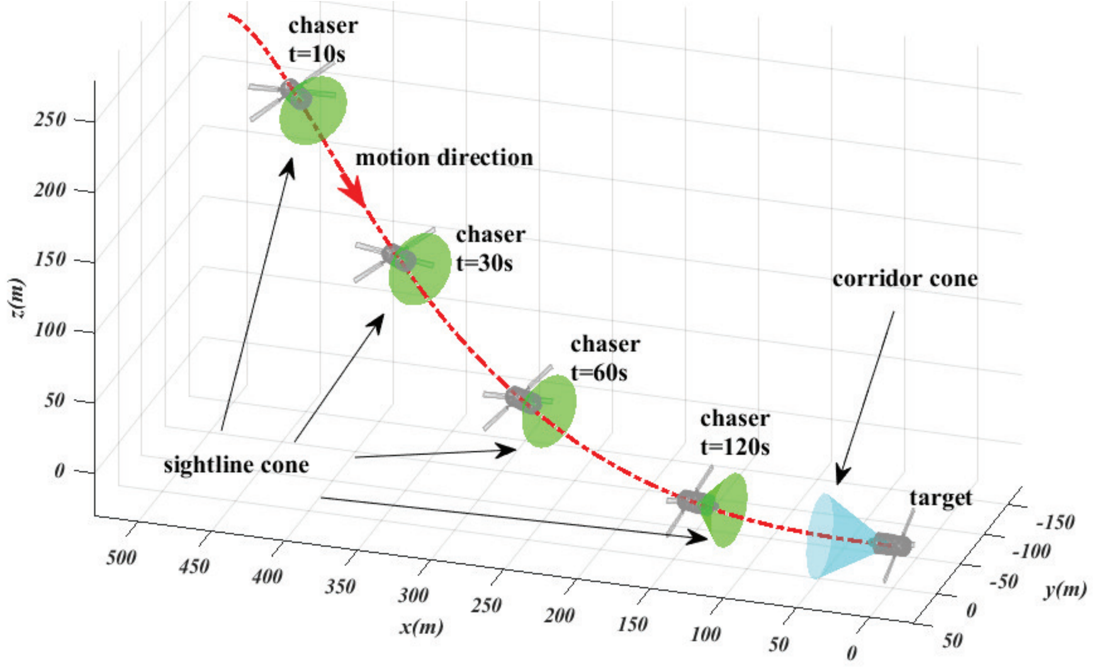
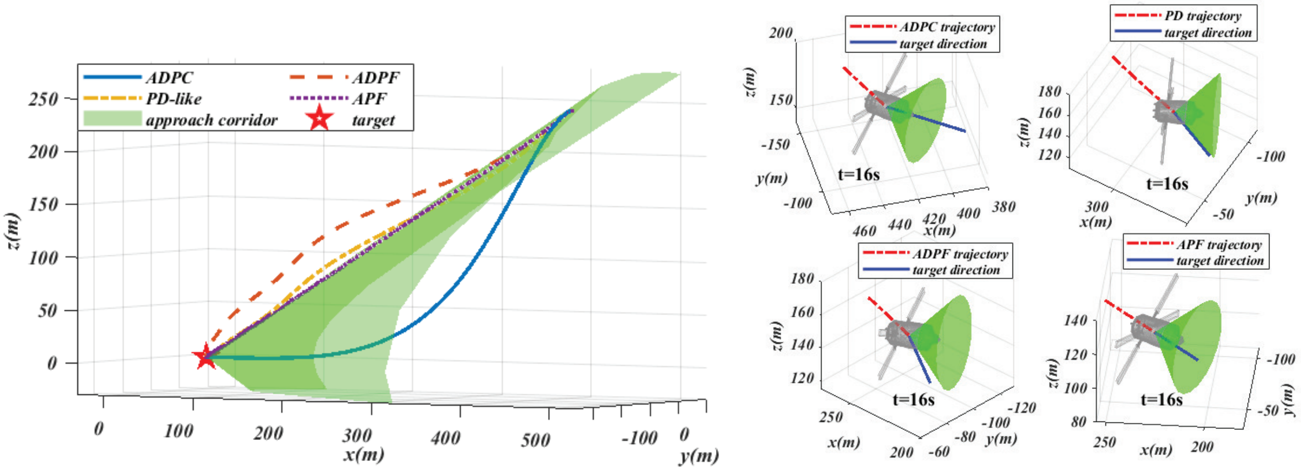


Figure 10. 3D illustration of approach process.



(a)

(b)

Figure 11. Comprehensive 3-D illustration under 4 different controllers.(a) 3D illustration of approach corridor constraint. (b) 3D illustration of field-of-view constraint at 16s.

angle $\arccos\left(\frac{-(\boldsymbol{r}_{bt}^{b})^{T}}{\|\boldsymbol{r}_{bt}^{b}\|}\boldsymbol{c}_{sight}\right)$ and approaching corridor angle $\arccos\left(\frac{(\boldsymbol{r}_{bt}^{t})^{T}}{\|\boldsymbol{r}_{bt}^{t}\|}\boldsymbol{c}_{path}\right)$ (drawn on the corresponding y–z planes) of every single run are given in Fig. 16. These figures indicate that the proposed method still completes the mission under variable initial states without any constraint violations. As it is well known that the characteristics of the actuator, such as the minimum impulse bit and saturation inevitably lead to degradation in performance and precision. But the result of the Monte-Carlo simulations is acceptable.

To sum up, simulation results demonstrate the effectiveness of the proposed method. Compared with conventional methods, the proposed ADP control scheme not only achieves the precise convergence of motion errors with better performance, but also has the ability to handle the constraints.

V. Conclusion

A reinforcement learning (RL)-based six-degree-of-freedom (6-DOF) control method was developed in this paper for the spacecraft proximity operations. In conjunction with the dual-quaternion algebraic framework, a specially designed barrier function was embedded in the reward function to cope with the nontrivial motion constraints of proximity operations. Subsequently, an RL-based control scheme was presented, achieving online approximate the optimal control subject to the 6-DOF nonlinear dynamics and motion constraints. The

Table I
RANGES OF INITIAL STATES.

Parameter	Values/Ranges
r_{mc}, m	$(0, 300)$
θ_{mc}, rad	$(-\pi, \pi)$
$r_{bt}^b(0)$, m	$[500, r_{mc}\cos\theta_{mc}, r_{mc}\sin\theta_{mc}]^T$
$v_{bt}^b(0)$, m/s	$(-1, 1) \times (-1, 1) \times (-1, 1)$
$\omega_{bt}^b(0)$, rad/s	$(-0.01, 0.01) \times (-0.01, 0.01) \times (-0.01, 0.01)$
$q_{bt}(0)$	Euler angle(y-x-z):$(-\pi/6, \pi/6) \times (-\pi/6, \pi/6) \times (-\pi/6, \pi/6)$

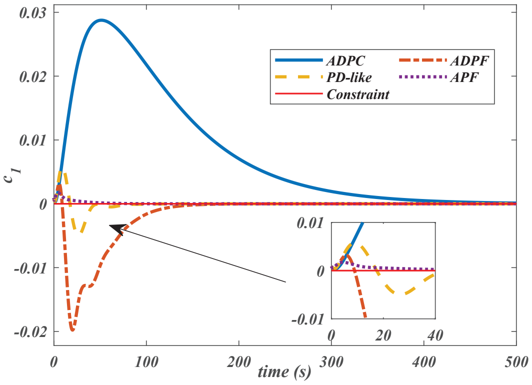


Figure 12. Comparison of field-of-view constraint.

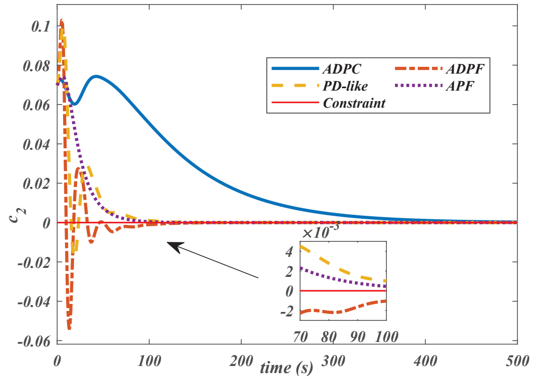


Figure 13. Comparison of approach corridor constraint.

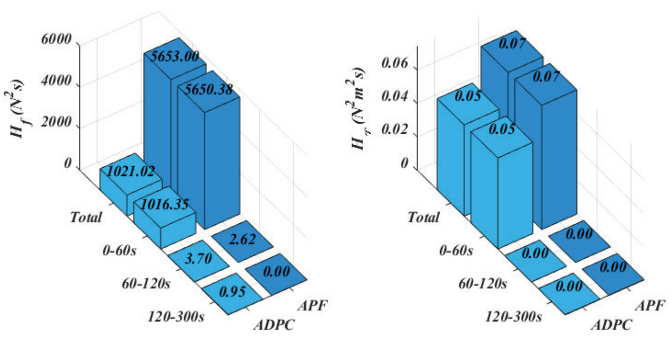


Figure 14. Control effort comparison with APF method.

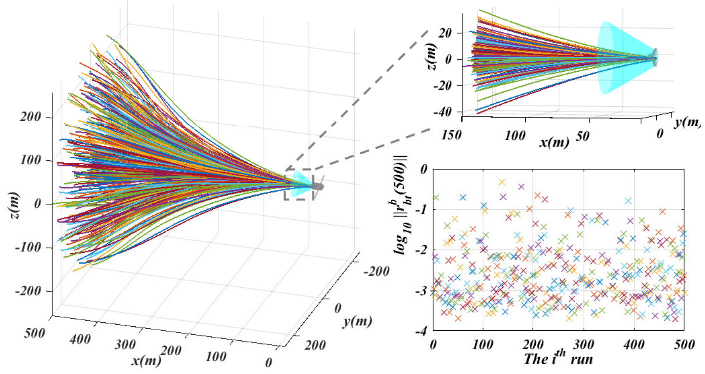
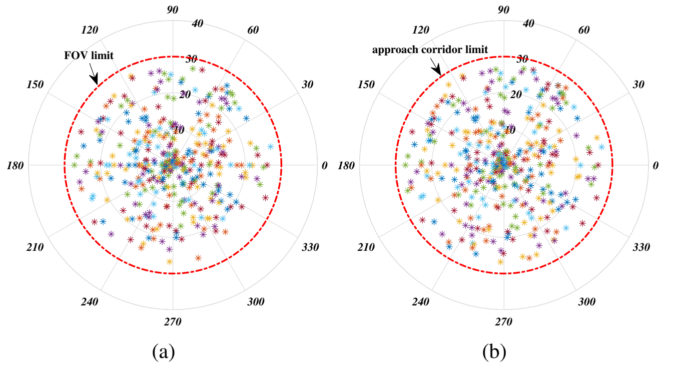


Figure 15. Trajectories and state error of Monte-Carlo simulations



(a)

(b)

Figure 16. (a) Maximum field-of-view angle of Monte-Carlo simulations. (b) Maximum approach corridor angle of Monte-Carlo simulations

ultimate boundedness of state errors and the weight of NN estimation errors in the closed-loop system was guaranteed by the Lyapunov-based method. The efficacy and effectiveness of the proposed method were carefully evaluated through a set of numerical simulations. Further work will aim at RL-based control with the uncertainties of the model.

REFERENCES

[1] W. Fehse, *Automated rendezvous and docking of spacecraft*. Cambridge university press, 2003, vol. 16.

[2] L. Sun and W. Huo, “6-dof integrated adaptive backstepping control for spacecraft proximity operations,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 3, pp. 2433–2443, 2015.

[3] M. B. Quadrelli, L. J. Wood, J. E. Riedel, M. C. McHenry, M. Aung, L. A. Cangahuala, R. A. Volpe, P. M. Beauchamp, and J. A. Cutts, “Guidance, navigation, and control technology assessment for future

planetary science missions," *Journal of Guidance, Control, and Dynamics*, vol. 38, no. 7, pp. 1165–1186, 2015.

[4] C. Pirat, M. Richard-Noca, C. Paccolat, F. Belloni, R. Wiesendanger, D. Courtney, R. Walker, and V. Gass, "Mission design and gnc for in-orbit demonstration of active debris removal technologies with cubesats," *Acta Astronautica*, vol. 130, pp. 114–127, 2017.

[5] H. Hinkel, J. J. Zipay, M. Strube, and S. Cryan, "Technology development of automated rendezvous and docking/capture sensors and docking mechanism for the asteroid redirect crewed mission," in *2016 IEEE Aerospace Conference*. IEEE, 2016, pp. 1–8.

[6] W.-J. Li, D.-Y. Cheng, X.-G. Liu, Y.-B. Wang, W.-H. Shi, Z.-X. Tang, F. Gao, F.-M. Zeng, H.-Y. Chai, W.-B. Luo *et al.*, "On-orbit service (oos) of spacecraft: A review of engineering developments," *Progress in Aerospace Sciences*, 2019.

[7] G. Vukovich and H. Gui, "Robust adaptive tracking of rigid-body motion with applications to asteroid proximity operations," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 1, pp. 419–430, 2017.

[8] L. S. Breger and J. P. How, "Safe trajectories for autonomous rendezvous of spacecraft," *Journal of Guidance, Control, and Dynamics*, vol. 31, no. 5, pp. 1478–1489, 2008.

[9] T. Weismuller and M. Leinz, "Gnc technology demonstrated by the orbital express autonomous rendezvous and capture sensor system," in *29th annual AAS guidance and control conference*. American Astronautical Society, 2006, pp. 06–016.

[10] 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.

[11] Y. Huang and Y. Jia, "Adaptive finite-time 6-dof tracking control for spacecraft fly around with input saturation and state constraints," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 3259–3272, 2019.

[12] 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 Transactions on Control Systems Technology*, 2021, doi:10.1109/TCST.2021.3076439.

[13] X. Shao and Q. Hu, "Immersion and invariance adaptive pose control for spacecraft proximity operations under kinematic and dynamic constraints," *IEEE Transactions on Aerospace and Electronic Systems*, 2021, doi:10.1109/TAES.2021.3053134.

[14] R. Zappulla, H. Park, J. Virgili-Llop, and M. Romano, "Real-time autonomous spacecraft proximity maneuvers and docking using an adaptive artificial potential field approach," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 6, pp. 2598–2605, 2019.

[15] R. Bevilacqua, T. Lehmann, and M. Romano, "Development and experimentation of lqr/apf guidance and control for autonomous proximity maneuvers of multiple spacecraft," *Acta Astronautica*, vol. 68, no. 7-8, pp. 1260–1275, 2011.

[16] J. Virgili-Llop, C. Zagaris, H. Park, R. Zappulla, and M. Romano, "Experimental evaluation of model predictive control and inverse dynamics control for spacecraft proximity and docking maneuvers," *CEAS Space Journal*, vol. 10, no. 1, pp. 37–49, 2018.

[17] P. Lu and X. Liu, "Autonomous trajectory planning for rendezvous and proximity operations by conic optimization," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 2, pp. 375–389, 2013.

[18] M. Xin and H. Pan, "Indirect robust control of spacecraft via optimal control solution," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 2, pp. 1798–1809, 2012.

[19] C. Jewison, R. S. Erwin, and A. Saenz-Otero, "Model predictive control with ellipsoid obstacle constraints for spacecraft rendezvous," *IFAC-PapersOnLine*, vol. 48, no. 9, pp. 257–262, 2015.

[20] U. Lee and M. Mesbahi, "Dual quaternion based spacecraft rendezvous with rotational and translational field of view constraints," in *AIAA/AAS Astrodynamics Specialist Conference*, 2014, p. 4362.

[21] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 2015.

[22] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[23] Y. Jiang and Z.-P. Jiang, *Robust adaptive dynamic programming*. John Wiley & Sons, 2017.

[24] D. Wang, H. He, and D. Liu, "Adaptive critic nonlinear robust control: A survey," *IEEE Transactions on Cybernetics*, vol. 47, no. 10, pp. 3429–3451, 2017.

[25] 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.

[26] C. Wei, J. Luo, H. Dai, and G. Duan, "Learning-based adaptive attitude control of spacecraft formation with guaranteed prescribed performance," *IEEE Transactions on Cybernetics*, vol. 49, no. 11, pp. 4004–4016, 2019.

[27] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and Autonomous Control Using Reinforcement Learning: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042–2062, 2018.

[28] D. Görges, "Relations between model predictive control and reinforcement learning," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 4920–4928, 2017.

[29] Q. Hu, X. Shao, and W. Chen, "Robust fault-tolerant tracking control for spacecraft proximity operations using time-varying sliding mode," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 1, pp. 2–17, 2018.

[30] X. Wang, P. Shi, C. Wen, and Y. Zhao, "Design of Parameter-Self-Tuning Controller Based on Reinforcement Learning for Tracking Noncooperative Targets in Space," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 6, pp. 4192–4208, 2020.

[31] A. T. Yang, "Application of quaternion algebra and dual numbers to the analysis of spatial mechanisms," Ph.D. dissertation, Columbia University Morningside Heights, New York, 1963.

[32] N. Filipe, A. Valverde, and P. Tsiotras, "Pose tracking without linearand angular-velocity feedback using dual quaternions," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 1, pp. 411–422, feb 2016.

[33] J.-Y. Wang, H.-Z. Liang, Z.-W. Sun, S.-N. Wu, and S.-J. Zhang, "Relative motion coupled control based on dual quaternion," *Aerospace Science and Technology*, vol. 25, no. 1, pp. 102–113, 2013.

[34] W. K. Clifford, "Preliminary sketch of biquaternions," *Proceedings of the London Mathematical Society*, vol. s1-4, no. 1, pp. 381–395, 1873.

[35] E. Study, "Von den bewegungen und umlegungen," *Mathematische Annalen*, vol. 39, no. 4, pp. 441–565, 1891.

[36] Y. Cheng, J. L. Crassidis, and F. L. Markley, "Attitude estimation for large field-of-view sensors," *The Journal of the Astronautical Sciences*, vol. 54, no. 3-4, pp. 433–448, 2006.

[37] 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.

[38] 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.

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

[40] N. Filipe and P. Tsiotras, "Adaptive position and attitude-tracking controller for satellite proximity operations using dual quaternions," *Journal of Guidance, Control, and Dynamics*, vol. 38, no. 4, pp. 566–577, 2015.

[41] H. Dong, Q. Hu, M. R. Akella, and F. Mazenc, "Partial lyapunov strictification: Dual-quaternion-based observer for 6-dof tracking control," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 6, pp. 2453–2469, 2019.

[42] I. K. M. Krstic and P. V. Kokotovic, *Nonlinear and Adaptive Control Design*. New York: Wiley, 1995.

[43] G. Song, N. V. Buck, and B. N. Agrawal, "Spacecraft vibration reduction using pulse-width pulse-frequency modulated input shaper," *Journal of Guidance, Control, and Dynamics*, vol. 22, no. 3, pp. 433–440, may 1999.

Qinglei Hu (Senior Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Zhengzhou University, Zhengzhou, China, in 2001, and the Ph.D. degree in control science and engineering with the specialization in guidance and control from the Harbin Institute of Technology, Harbin, China, in 2006.

From 2003 to 2014, he was with the Department of Control Science and Engineering, Harbin Institute of Technology. He joined Beihang University, Beijing, China, in 2014, as a Full Professor. His current research interests include variable structure control and applications, and fault-tolerant control and applications. In these areas, he has authored or coauthored more than 80 technical articles. Prof. Hu serves as an Associate Editor for Aerospace Science and Technology.

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 with Beihang University, Beijing, China.

His current research interests include reinforcement learning-based control, intelligent control, and attitude and 6-DOF motion control. He is also working on the hardware-in-loop experiments for various nonlinear control systems.

Hongyang Dong received the Ph.D. degree in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2018.

From 2015 to 2017, he was a joint Ph.D. Student with the Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, USA. He is currently a Research Fellow in machine learning and intelligent control with the School of Engineering, University of Warwick, Coventry, U.K. His current research interests include reinforcement learning, deep learning, intelligent control, and adaptive control.

Xiaowei Zhao received the Ph.D. degree in control theory from Imperial College London, London, U.K., in 2010.

Before joining the University of Warwick in 2013, he worked as a Post-Doctoral Researcher with the University of Oxford, Oxford, U.K., for three years. He is currently a Professor of control engineering and an EPSRC Fellow with the School of Engineering, University of Warwick, Coventry, U.K. His main research areas are control theory with applications on offshore renewable energy systems, local smart energy systems, and autonomous systems.