Adaptive Optimal Control of Space Tether System for Payload Capture via Policy Iteration

FENG Yiting[1], *ZHANG Ming*[2], *GUO Wenhao*[2], *WANG Changqing*[1*]

1. School of Automation, Northwestern Polytechnical University, Xi'an 710129, P. R. China;
2. Beijing Institute of Aerospace Systems Engineering, Beijing 100076, P. R. China

(Received 14 May 2021; revised 16 July 2021; accepted 5 August 2021)

Abstract: The libration control problem of space tether system (STS) for post-capture of payload is studied. The process of payload capture will cause tether swing and deviation from the nominal position, resulting in the failure of capture mission. Due to unknown inertial parameters after capturing the payload, an adaptive optimal control based on policy iteration is developed to stabilize the uncertain dynamic system in the post-capture phase. By introducing integral reinforcement learning (IRL) scheme, the algebraic Riccati equation (ARE) can be online solved without known dynamics. To avoid computational burden from iteration equations, the online implementation of policy iteration algorithm is provided by the least-squares solution method. Finally, the effectiveness of the algorithm is validated by numerical simulations.

Key words: space tether system (STS); payload capture; policy iteration; integral reinforcement learning (IRL); state feedback

CLC number: V448.2 **Document code:** A **Article ID:** 1005-1120(2021)04-0560-11

0 Introduction

With the development of aerospace industry, more and more spacecraft have been launched, resulting in a large amount of space debris in low earth orbit (LEO). Due to the complex space disturbance, the orbital altitude of the spacecraft changes, which can cause the collision between different spacecraft, resulting in a large number of space debris. Therefore, the safe and efficient capture of space debris is of great significance for the safe completion of space missions. Space tether system (STS) has been widely studied in debris removal[1-3], orbit transfer[4-5] and artificial gravity generation[6-7] due to its flexible structure and higher reliability of approaching space target than the space manipulator on the floating platform.

An on-orbit capture mission operated by space tether can be mainly divided into three stages: The deployment of tether before capture, rendezvous for capture, post-capture stabilization and retrieval[8]. Due to complex space environment, there are possibly some errors of tether length and pendulum angle in the payload capture process. The capture mechanism installed at the end of the tether is integrated with the target payload in the post-capture period. Uncertain inertial dynamic and undesirable rendezvous position can cause the tether swing and oscillation in the orbital plane, which can lead to the tether winding up with payload if unstable. Therefore, analysis of the the equilibrium position and corresponding stabilization control of STS are essential. Recently, numerous studies regarding to dynamic and control of STS for payload capture have been conducted by scholars from all over the world. Hovell et al.[9] compared the ability of different tether structures for space debris removal by air-floating test platform. Through two-dimensional plane ex-

*Corresponding author, E-mail address: wangcq@nwpu.edu.cn.

How to cite this article: FENG Yiting, ZHANG Ming, GUO Wenhao, et al. Adaptive optimal control of space tether system for payload capture via policy iteration[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2021, 38(4): 560-570.

http://dx.doi.org/10.16356/j.1005-1120.2021.04.003

periment, it was verified that the sub-tethered structure has better performance of towing space debris. In Refs.[10-11], the swing characteristics and stability of STS in the process of in-plane orbit transfer were studied, and some effective orbit maneuver schemes and swing suppression strategy of the tethered system based on tension and continuous constant thrust were proposed. Although the proposed method can suppress the in-plane motion quickly and accurately, the mothod using external input such as electrodynamic force and thrust is not suitable for stability control in station-keeping stage. Tension control is more suitable to stabilize the tether system in the post-capture stage because the control time is not limited in the long orbit period.

In order to ensure STS fulfill the space missions quickly and stably, scholars have designed a variety of control methods for the release and retrieval process of the tether system. Ref.[12] proposed adaptive sliding mode control for deployment of space tether to overcome the disturbance in low-eccentricity orbits. Energy-based control framework was employed into deployment of tethered spacecraft with input saturation[13]. Apart from the above methods, several linear or nonlinear methods were also studied for STS, such as incremental nonlinear dynamic surface control, robust performance control, model predictive control, and so on[14-16]. It should be noted that most existing studies discussed above depend on the accurate model of STS. However, the system parameters will change abruptly in the period of payload capture, so the accurate dynamic of STS cannot be obtained. Due to highly complex dynamic characteristic and unmeasurable dynamic parameters, designing model-free controller for STS is significant. In recent years, with the development of artificial intelligence, a variety of intelligent optimization algorithms have emerged in solving control issues of aerospace system[17-18]. As a representative technology in the field of artificial intelligence, model-free control scheme based on reinforcement learning (RL) has gained wide attention in solving optimization problems with unknown internal dynamics and external disturbances. The optimal controller design of dynamic system can be converted into solving the Hamilton-Jacobi-Bellman (HJB) equations. However, the analytical solutions to the HJB equations are hard to obtain[19]. The key superiority of RL is to approximately solve the HJB equations through an iterative method, including policy iteration (PI) and value iteration[20]. PI is the most widely technique used in RL to approximate the HJB equations. Generally, PI method has two-step iterations: Policy evaluation and policy improvement. For optimal control problem with parametric uncertainties or even unknown dynamics, the online learning algorithms are of great significance, which can be integrated with adaptive control to develop adaptive optimal control algorithms[21]. An online PI algorithm was first presented for optimal control of continuous time system in Ref.[22]. Vrabie et al.[23] proposed an integral reinforcement learning (IRL) algorithm for linear continuous-time systems using only partial knowledge about the system dynamics. Furthermore, Ref.[24] presented online model-free RL algorithm for completely unknown continuous-time linear systems.

Based on the above discussion, an online IRL control scheme based on the policy iteration technique is designed to stabilize STS for payload capture with dynamical uncertainty. A state feedback controller for payload capture is developed by using the online information of the system states and inputs without requiring prior knowledge of the system internal dynamics.

1 Problem Formulation

1.1 Dynamic model

STS is considered as an elastic rod model with mass. Some reasonable assumptions are given to simplify the system dynamics modeling as follows.

(1) The space tug (main satellite) and capture mechanism are connected by an elastic tether, and the centroid of the system is on Kepler's orbit.

(2) The main satellite and sub-satellite in tethered system can be considered as mass points, without regard to the attitude of the satellites.

(3) The tether is regarded as an elastic rod with uniform mass distribution. Only the longitudinal vibration along the tether is considered.

(4) Some space environment effects are ignored, such as solar light pressure, atmospheric resistance, and the oblateness of the earth.

Fig.1 shows the coordinate frames of the space tether. The inertial frame $OXYZ$ is attached to the center of earth. The OXY plane is the same as orbit plane. The axis of OX points to orbital perigee, and the OZ axis is along the equatorial plane normal toward the celestial north pole. The OY axis represents the third axis of righthanded orthogonal frame. The orbit coordinate frame $CX_oY_oZ_o$ is located at the mass center of STS, with CX_o axis outward from the Earth center along the local vertical. CZ_o axis is directed toward the orbit normal direction, and CY_o axis along the local horizon represents the third axis of right-handed orthogonal frame. The body-fixed frame $CX_tY_tZ_t$ has the same origin as orbit coordinate frame, with CX_t axis along the opposite direction of the tether tension. The direction of CY_t and CZ_t are determined by the in-plane angle θ and out-plane angle φ relative to frame $CX_oY_oZ_o$. The payload target is in the same orbital plane as the space tether. η is the true anomaly of target in the inertial frame.

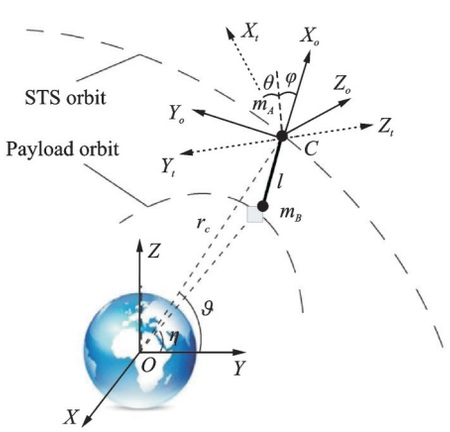

Fig.1 Schematic of capture process by STS

The states of STS can be described by five generalized coordinates: The orbital radius r, true anomaly ϑ, in-plane angle θ, out-plane angle φ, and elastic deformation of tether ε. According to the Euler-Lagrange equation, the differential equations of the system model can be derived as follows

$$\begin{cases} \ddot{\theta}=-2(\dot{\theta}+\dot{\vartheta})\left(-\dot{\varphi}\tan\varphi+\Phi_2\dfrac{\dot{l}_0}{l_0}+\dfrac{\dot{\varepsilon}}{1+\varepsilon}\right)+ \\ \qquad 2\dfrac{\dot{\vartheta}\dot{r}}{r}-\dfrac{3\mu\sin(2\theta)}{2r^3}\left(1+\Phi_1\dfrac{l_0^2\cos^2\varphi}{r^2}\right) \\ \ddot{\varphi}=-2\dot{\varphi}\left(\Phi_2\dfrac{\dot{l}_0}{l_0}+\dfrac{\dot{\varepsilon}}{1+\varepsilon}\right)- \\ \qquad \left((\dot{\theta}+\dot{\vartheta})^2+\dfrac{3\mu}{r^3}\cos^2\theta\right)\sin\varphi\cos\varphi \\ \ddot{\varepsilon}=-\dfrac{2\dot{l}_0\dot{\varepsilon}}{l_0}-(1+\varepsilon)\dfrac{\ddot{l}_0}{l_0}-2\Phi_3\dfrac{\dot{l}_0}{l_0}\left[\dot{\varepsilon}+(1+\varepsilon)\dfrac{\dot{l}_0}{l_0}\right]+ \\ \qquad \dfrac{m_*}{m_A(m_B+m_t)/m}(1+\varepsilon)\left[(\dot{\theta}+\dot{\vartheta})^2\cos^2\varphi+\right. \\ \qquad \left.\dot{\varphi}^2+\dfrac{\mu}{r^3}(3\cos^2\theta\cos^2\varphi-1)\right]- \\ \qquad \dfrac{T}{m_A(m_B+m_t)/m} \end{cases} \tag{1}$$

where m_A denotes the mass of the tug and m_B the mass of the combination of the capture mechanism and target; $m_t=\rho l_0$ represents the mass of tether, with ρ the density of tether and l_0 the original length; the actual length of tether is $l=l_0(1+\varepsilon)$. In addition, T is the tether tension and μ the geocentric gravitation constant. The mass coefficients are defined as

$$\begin{cases} \Phi_1=m_*/m \\ \Phi_2=m_A(m_B+m_t/2)/(mm_*) \\ \Phi_3=(2m_A-m)m_t/[2m_A(m_B+m_t)] \\ \Phi_4=(m_B+m_t/2)/(m_B+m_t) \\ m=m_A+m_B+m_t \\ m_*=(m_A+m_t/2)(m_B+m_t/2)/m-m_t/6 \end{cases} \tag{2}$$

Remark 1 The tether swings around the equilibrium position disturbed by target and the non-nominal libration motion occur in the post-capture stage. The elastic elongation of tether is ignored and the length of tether remains unchanged without control input. The effects from the variation of the tether mass and deformation are ignored. It is assumed that the system moves on the circular orbit, with orbital angular velocity $\dot{\vartheta}=\Omega=\sqrt{\mu/r^3}$. Based on the

above assumptions, the system model can be linearized near the equilibrium position.

Due to little impact on the in-plane stability, the out-plane motion of the system is ignored. For STS in the circular orbit, the equilibrium positions are along the radial direction of the orbit, with in-plane angle $\theta_{1,2}=0,\pi$. Eq.(1) can be linearized around the equilibrium point by ignoring higher-order terms of nonlinear system. By introducing a dimensionless time $\tau=\Omega t$, the dimensionless linear model of the system can be derived as follows

$$\begin{cases}\varepsilon_0''=3\Phi_4(\varepsilon_0+1)+2\Phi_4\theta'-\dfrac{mT}{[m_A\Omega^2(m_B+m_t)]l_c}\\ \theta''=-2\Phi_2\dfrac{\varepsilon_0'}{\varepsilon_0+1}-3\theta\end{cases} \tag{3}$$

where the superscripts "′" "″" mean the first and the second derivative versus τ. The dimensionless length is denoted as $\varepsilon_0=l/l_c-1$, and l_c is the nominal tether length in the post-capture stage.

Then the state space equation of the system near the equilibrium point can be compactly expressed as

$$\boldsymbol{X}'=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{B}\boldsymbol{U} \tag{4}$$

where $\boldsymbol{A}$ denotes the state matrix, $\boldsymbol{B}$ the input matrix, $\boldsymbol{X}=[\varepsilon_0 \quad \varepsilon_0' \quad \theta \quad \theta']^{\mathrm{T}}$ the state vector, and $\boldsymbol{U}$ the dimensionless control input.

Remark 2 According to Eq.(3), while adjusting the tether length and velocity, the in-plane angle can be stabilized by the term of $\varepsilon_0'/(\varepsilon_0+1)$ in the differential equation. Adjustment of the tether length and velocity can be realized by releasing/rewinding mechanism and tension control. In this paper, tension control scheme is adopted to stabilize the swing motion of tether in the post-capture stage.

1.2 Preliminaries

The aim of this paper is to design an online adaptive control scheme based on policy iteration to drive the real-time asymptotic stability defined dynamic system. In this section, some concepts and propositions for control design are given.

Definition 1 (Bellman's optimality principle[25]) Bellman optimality principle is a basic foundation of reinforcement learning. According to the Bellman optimal equation, there exists an optimal control strategy to obtain the optimal cost function of any Markov decision process (MDP). For the linear system, its cost function is the quadratic of the state vector and the control input, so the corresponding optimal feedback control can be derived by solving the basic algebraic Riccati equation (ARE).

Consider the linear time-invariant (LTI) dynamical system described by

$$\dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t)+\boldsymbol{B}\boldsymbol{u}(t) \tag{5}$$

where $\boldsymbol{x}(t)\in\mathbf{R}^n$, $\boldsymbol{u}(t)\in\mathbf{R}^m$, and the pair $(\boldsymbol{A},\boldsymbol{B})$ is controllable, subject to the following optimal control problem

$$u^*(t)=\arg\min_{\boldsymbol{u}(t),t\in[t_0,\infty)}V(t_0,x(t_0),\boldsymbol{u}(t)) \tag{6}$$

where the infinite horizon quadratic cost function to be minimized is expressed as

$$V(x(t_0),t_0)=\int_{t_0}^{\infty}(\boldsymbol{x}^{\mathrm{T}}(\tau)\boldsymbol{Q}\boldsymbol{x}(\tau)+\boldsymbol{u}^{\mathrm{T}}(\tau)\boldsymbol{R}\boldsymbol{u}(\tau))\mathrm{d}\tau \tag{7}$$

with $\boldsymbol{Q}\geqslant 0$, $\boldsymbol{R}>0$, and $(\boldsymbol{Q}^{1/2},\boldsymbol{A})$ detectable.

Based on Bellman optimality principle, the solution of this optimal control problem is obtained by $\boldsymbol{u}(t)=-\boldsymbol{K}\boldsymbol{x}(t)$ with

$$\boldsymbol{K}=\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P} \tag{8}$$

where the matrix $\boldsymbol{P}$ is the symmetric positive definite solution of the ARE as Eq.(9). And the unique solution determines the stable close-loop controller.

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}+\boldsymbol{P}\boldsymbol{A}-\boldsymbol{P}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}+\boldsymbol{Q}=0 \tag{9}$$

Lemma 1[26] Consider the linear system expressed as Eq.(5), initialize $\boldsymbol{K}_0\in\mathbf{R}^{m\times n}$ to be any stabilizing feedback gain matrix, and let $\boldsymbol{P}_i$ be the symmetric positive definite solution of the Lyapunov equation

$$(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_i)^{\mathrm{T}}\boldsymbol{P}_i+\boldsymbol{P}_i(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_i)+\boldsymbol{Q}+\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i=0 \tag{10}$$

where $\boldsymbol{K}_i=\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_{i-1}$ with $i=1,2,\cdots,n$. Then the following properties are satisfied: (i) $(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_i)$ is Hurwitz; (ii) $\boldsymbol{P}^*\leqslant\boldsymbol{P}_{i+1}\leqslant\boldsymbol{P}_i$; (iii) $\lim\limits_{k\to\infty}\boldsymbol{K}_i=\boldsymbol{K}^*$, $\lim\limits_{k\to\infty}\boldsymbol{P}_i=\boldsymbol{P}^*$.

Remark 3 It should be noted that model information of the system is needed to solve the above ARE, which means that the system matrix $\boldsymbol{A}$ and control input matrix $\boldsymbol{B}$ are needed to be known. Therefore, designing a model-free controller with-

out using knowledge regarding the system dynamics is particularly important research topic in the optimal control field.

2 Control Design

In this section, an adaptive optimal controller for STS is investigated in the case of unknown internal dynamics. The aim of this paper is to design an online learning control scheme to drive STS asymptotically stable in real time. The proposed IRL policy iteration control diagram is shown as Fig.2, including policy evaluation and policy improvement. Policy evaluation is to calculate the infinite horizon cost associated with the given stability controller, and the purpose of policy improvement is to improve the feedback gain of the system to reduce the cost.

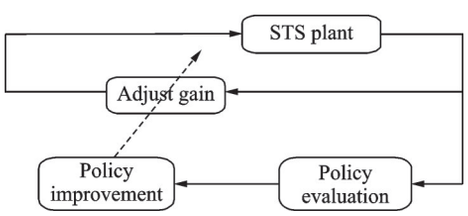

Fig.2 Closed-loop control structure of the system

2.1 IRL policy iteration scheme

Let $\boldsymbol{K}$ be a stabilizing feedback control gain for Eq.(5). Under the assumption that $(\boldsymbol{A}, \boldsymbol{B})$ is controllable, the close-loop system is stable with the control input $\boldsymbol{u}(t)=-\boldsymbol{K}\boldsymbol{x}(t)$. It directly comes from the state feedback control explanation.

Substituting the expression of control input into Eq.(7), the corresponding infinite horizon quadratic cost becomes

$$V(\boldsymbol{x}(t))=\int_t^{\infty}\boldsymbol{x}^{\mathrm{T}}(\tau)(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K})\boldsymbol{x}(\tau)\mathrm{d}\tau=\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{P}\boldsymbol{x}(t) \tag{11}$$

where $\boldsymbol{P}$ is the real symmetric positive definite solution of the Lyapunov matrix equation

$$(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K})^{\mathrm{T}}\boldsymbol{P}+\boldsymbol{P}(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K})=-(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}+\boldsymbol{Q}) \tag{12}$$

As a Lyapunov function candidate for controlled plant, the cost function $V(\boldsymbol{x}(t))$ can be written as

$$V(\boldsymbol{x}(t))=\int_t^{t+T}\boldsymbol{x}^{\mathrm{T}}(\tau)(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K})\boldsymbol{x}(\tau)\mathrm{d}\tau+\int_{t+T}^{\infty}\boldsymbol{x}^{\mathrm{T}}(\tau)(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K})\boldsymbol{x}(\tau)\mathrm{d}\tau=\int_t^{t+T}\boldsymbol{x}^{\mathrm{T}}(\tau)(\boldsymbol{Q}+\boldsymbol{K}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K})\boldsymbol{x}(\tau)\mathrm{d}\tau+V(\boldsymbol{x}(t+T)) \tag{13}$$

Using the conventionalized expression $\boldsymbol{x}(t)=\boldsymbol{x}_t$, the cost function can be rewritten as $V(\boldsymbol{x}_t)=\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x}_t$, and the initial stable control gain is defined as $\boldsymbol{K}_1$. So we can get the following online policy iteration scheme

$$\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_t=\int_t^{t+T}(\boldsymbol{x}_\tau^{\mathrm{T}}(\boldsymbol{Q}+\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i)\boldsymbol{x}_\tau)\mathrm{d}\tau+\boldsymbol{x}_{t+T}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_{t+T} \tag{14}$$

$$\boldsymbol{K}_{i+1}=\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_i \tag{15}$$

Note that Eqs.(14) and (15) form a new policy iteration algorithm without involving the plant matrix $\boldsymbol{A}$. The whole design procedure of the online control scheme can be summarized as Algorithm 1. The algorithm design of state feedback control based on online IRL can be summarized as the following theorem.

Algorithm 1 Continuous-time IRL policy iteration algorithm

Input: Initial condition of the system x_0, initial stabilizing control gain $\boldsymbol{K}_1$, initial $\boldsymbol{P}_0=0$, initial iteration number $i=1$, a positive error constant ε, positive definite symmetric matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$.

while $\|V_{t+\delta T}-V_t\|\geqslant\varepsilon$ **do**

(Policy evaluation)

Calculate $\boldsymbol{P}_i$ from

$$\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_t-\boldsymbol{x}_{t+\delta T}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_{t+\delta T}=\int_t^{t+\delta T}(\boldsymbol{x}_\tau^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}_\tau+\boldsymbol{u}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{u})\mathrm{d}\tau=\int_t^{t+\delta T}(\boldsymbol{x}_\tau^{\mathrm{T}}(\boldsymbol{Q}+\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i)\boldsymbol{x}_\tau)\mathrm{d}\tau$$

(Policy improvement)

Update the feedback gain using $\boldsymbol{K}_{i+1}=\boldsymbol{R}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_i$

Update the control policy $u_t=-\boldsymbol{K}_{i+1}\boldsymbol{x}_t$

$i\leftarrow i+1$

end while

return $\boldsymbol{u}_t, \boldsymbol{K}_{i+1}, \boldsymbol{P}_i$

Theorem 1 For the system model described in Eq.(4), if $\boldsymbol{K}_0$ is initialized to guarantee $\boldsymbol{A}_0=\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_0$ stable, and $\boldsymbol{P}_i$ and $\boldsymbol{K}_i$ are updated as the policy iteration scheme with proper positive definite symmetric matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, the closed-loop system is always stable during the iteration period.

Proof Since the positive definite cost function $V_i(\boldsymbol{x}_t)=\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_t$ is defined as the Lyapunov function candidate and

$$\frac{\mathrm{d}(\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_t)}{\mathrm{d}t}=\boldsymbol{x}_t^{\mathrm{T}}(\boldsymbol{A}_i^{\mathrm{T}}\boldsymbol{P}_i+\boldsymbol{P}_i\boldsymbol{A}_i)\boldsymbol{x}_t=-\boldsymbol{x}_t^{\mathrm{T}}(\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i+\boldsymbol{Q})\boldsymbol{x}_t \tag{16}$$

then for any $\delta T>0$, the unique solution of the Lyapunov equation satisfies

$$\int_t^{t+\delta T}(\boldsymbol{x}_\tau^{\mathrm{T}}(\boldsymbol{Q}+\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i)\boldsymbol{x}_\tau)\mathrm{d}\tau=-\int_t^{t+\delta T}\frac{\mathrm{d}(\boldsymbol{x}_\tau^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_\tau)}{\mathrm{d}\tau}\mathrm{d}\tau=\boldsymbol{x}_t^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_t-\boldsymbol{x}_{t+T}^{\mathrm{T}}\boldsymbol{P}_i\boldsymbol{x}_{t+T} \tag{17}$$

Taking the derivative of $V_i(\boldsymbol{x}_t)$ along the state trajectories generated by the control policy u_t, one obtains

$$\begin{aligned}\dot{V}_i(\boldsymbol{x}_t)=&\boldsymbol{x}_t^{\mathrm{T}}[\boldsymbol{P}_i(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_{i+1})+(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_{i+1})^{\mathrm{T}}\boldsymbol{P}_i]\boldsymbol{x}_t=\\&\boldsymbol{x}_t^{\mathrm{T}}[\boldsymbol{P}_i(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_i)+(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_i)^{\mathrm{T}}\boldsymbol{P}_i]\boldsymbol{x}_t+\\&\boldsymbol{x}_t^{\mathrm{T}}[\boldsymbol{P}_i\boldsymbol{B}(\boldsymbol{K}_i-\boldsymbol{K}_{i+1})+(\boldsymbol{K}_i-\boldsymbol{K}_{i+1})^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_i]\boldsymbol{x}_t\end{aligned} \tag{18}$$

According to Eq.(12), the first term in Eq.(18) can be written as $-\boldsymbol{x}_t^{\mathrm{T}}[\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i+\boldsymbol{Q}]\boldsymbol{x}_t$, and the second term can be rewritten using Eq.(15). Then the transformed form of Eq.(18) can be obtained as

$$\begin{aligned}\dot{V}_i(\boldsymbol{x}_t)=&-\boldsymbol{x}_t^{\mathrm{T}}[(\boldsymbol{K}_i-\boldsymbol{K}_{i+1})^{\mathrm{T}}\boldsymbol{R}(\boldsymbol{K}_i-\boldsymbol{K}_{i+1})]\boldsymbol{x}_t-\\&\boldsymbol{x}_t^{\mathrm{T}}[\boldsymbol{Q}+\boldsymbol{K}_{i+1}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_{i+1}]\boldsymbol{x}_t\end{aligned} \tag{19}$$

With the consideration of $\boldsymbol{Q}>0$ and $\boldsymbol{R}>0$, $\dot{V}_i(\boldsymbol{x}_t)<0$ is guaranteed when the state vector is nonzero, which proves the updated control policy in Algorithm 1 is stable. The proof of Theorem 1 is completed.

Remark 4 In terms of online implement of IRL algorithm, that is, the on-policy method, the target policy and behavior policy are unified into one policy. Thus, the gain matrix calculated in each iteration is immediately applied to the system, which enhances the running speed of the algorithm. Compared with off-policy method, there is no need to introduce exploration noise into the controller in the proposed algorithm.

2.2 Online implementation of IRL algorithm

In this section, an online adaptive learning algorithm is presented to implement the IRL policy iteration scheme in real time. The algorithm performs the online IRL iterations by measuring the present state $\boldsymbol{x}_t$ and the next state $\boldsymbol{x}_{t+T}$ with fixed sampling time T. The information of the system matrix $\boldsymbol{A}$ is involved in the measured states, which leads to the policy update without knowing the internal dynamic of the system.

The symmetric matrix $\boldsymbol{P}_i$ in the value function $V_i(\boldsymbol{x}(t))$ can be calculated at each iteration i by measuring the states along the system trajectory. For the convenience of computing, the value function is written as

$$\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{P}_i\boldsymbol{x}(t)=\bar{\boldsymbol{p}}_i^{\mathrm{T}}\bar{\boldsymbol{x}}(t) \tag{20}$$

where $\bar{\boldsymbol{x}}_t$ denotes the Kronecker product quadratic polynomial basis vector with the elements $\{x_i(t)x_j(t)\}_{i=1,m;j=1,n}$. The parameter vector $\bar{\boldsymbol{p}}_i$ contains the elements of the matrix $\boldsymbol{P}_i$ ordered by columns with the redundant elements removed. Then Eq.(14) can be rewritten as

$$\bar{\boldsymbol{p}}_i^{\mathrm{T}}(\bar{\boldsymbol{x}}(t)-\bar{\boldsymbol{x}}(t+T))=\int_t^{t+T}(\boldsymbol{x}^{\mathrm{T}}(\tau)(\boldsymbol{Q}+\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i)\boldsymbol{x}(\tau))\mathrm{d}\tau \tag{21}$$

where $\bar{\boldsymbol{p}}_i$ is the vector of unknown parameters and $(\bar{\boldsymbol{x}}(t)-\bar{\boldsymbol{x}}(t+T))$ acts as a regression vector. The right hand side is the integral reinforcement on the time interval $[t,t+T]$, which can be denoted as

$$d(\bar{\boldsymbol{x}}(t),\boldsymbol{K}_i)=\int_t^{t+T}(\boldsymbol{x}^{\mathrm{T}}(\tau)(\boldsymbol{Q}+\boldsymbol{K}_i^{\mathrm{T}}\boldsymbol{R}\boldsymbol{K}_i)\boldsymbol{x}(\tau))\mathrm{d}\tau \tag{22}$$

where $d(\bar{\boldsymbol{x}}(t),\boldsymbol{K}_i)$ represents a desired value or target function and the estimate of the parameter $\bar{\boldsymbol{p}}_i$ is to be found such that the parameter satisfies the equation as closely as possible. To compute it efficiently, define a new controller state $V(t)$ as the augmented state of the system, and add the state equation $\dot{V}(t)=\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}\boldsymbol{x}(t)+\boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{R}\boldsymbol{u}(t)$ to the controller dynamics. The value of $d(\bar{\boldsymbol{x}}(t),\boldsymbol{K}_i)$ can be computed by using $d(\bar{\boldsymbol{x}}(t),\boldsymbol{K}_i)=V(t+T)-$

$V(t)$.

The unknown parameter vector $\bar{\boldsymbol{p}}_i$ of the value function is involved in the scalar equation Eq.(21), which can be solved by a batch solution method in the least-squares sense. Firstly, some relevant vectors of parameters are defined as

$$\boldsymbol{X}=[\bar{\boldsymbol{x}}_{\Delta}^{1} \quad \bar{\boldsymbol{x}}_{\Delta}^{2} \quad \cdots \quad \bar{\boldsymbol{x}}_{\Delta}^{N}] \tag{23}$$

$$\bar{\boldsymbol{x}}_{\Delta}^{i}=\bar{\boldsymbol{x}}^{i}(t)-\bar{\boldsymbol{x}}^{i}(t+T) \tag{24}$$

$$\boldsymbol{Y}=[\,d(\bar{\boldsymbol{x}}^{1},\boldsymbol{K}_i) \quad d(\bar{\boldsymbol{x}}^{2},\boldsymbol{K}_i) \quad \cdots \quad d(\bar{\boldsymbol{x}}^{N},\boldsymbol{K}_i)\,]^{\mathrm{T}} \tag{25}$$

Assume the square loss function as $J(\bar{\boldsymbol{p}}_i)=\sum_{n=1}^{N}(\boldsymbol{y}_n-\bar{\boldsymbol{p}}_i^{\mathrm{T}}\bar{\boldsymbol{x}}_{\Delta}^{n})^2$. To minimize the square loss, let the derivative of $J(\bar{\boldsymbol{p}}_i)$ with respect to $\bar{\boldsymbol{p}}_i$ equals to zero, one can obtain

$$\nabla_{\bar{p}_i}J(\bar{\boldsymbol{p}}_i)=\nabla_{\bar{p}_i}\left(\sum_{n=1}^{N}(\boldsymbol{y}_n-\bar{\boldsymbol{p}}_i\bar{\boldsymbol{x}}_{\Delta}^{n})^2\right)=$$

$$-\sum_{n=1}^{N}2(\boldsymbol{y}_n-\bar{\boldsymbol{p}}_i\bar{\boldsymbol{x}}_{\Delta}^{n})\bar{\boldsymbol{x}}_{\Delta}^{n}=0 \tag{26}$$

$$\sum_{n=1}^{N}\bar{\boldsymbol{p}}_i^{\mathrm{T}}\bar{\boldsymbol{x}}_{\Delta}^{n}\bar{\boldsymbol{x}}_{\Delta}^{n}=\sum_{n=1}^{N}\boldsymbol{y}_n\bar{\boldsymbol{x}}_{\Delta}^{n} \tag{27}$$

Noting that $\bar{\boldsymbol{p}}_i^{\mathrm{T}}\bar{\boldsymbol{x}}_{\Delta}^{n}$ is a scalar, the left hand side of Eq.(27) has the following form after rearrangement

$$\sum_{n=1}^{N}(\bar{\boldsymbol{p}}_i^{\mathrm{T}}\bar{\boldsymbol{x}}_{\Delta}^{n})\bar{\boldsymbol{x}}_{\Delta}^{n}=\sum_{n=1}^{N}\bar{\boldsymbol{x}}_{\Delta}^{n}(\bar{\boldsymbol{p}}_i^{\mathrm{T}}\bar{\boldsymbol{x}}_{\Delta}^{n})=\sum_{n=1}^{N}\bar{\boldsymbol{x}}_{\Delta}^{n}\bar{\boldsymbol{x}}_{\Delta}^{n\,\mathrm{T}}\bar{\boldsymbol{p}}_i \tag{28}$$

Using the above transformation equation, Eq.(23) can be rewritten as

$$\sum_{n=1}^{N}\bar{\boldsymbol{x}}_{\Delta}^{n}\bar{\boldsymbol{x}}_{\Delta}^{n\,\mathrm{T}}\bar{\boldsymbol{p}}_i=\sum_{n=1}^{N}\boldsymbol{y}_n\bar{\boldsymbol{x}}_{\Delta}^{n} \tag{29}$$

Then the batch least-squares solution of $\bar{\boldsymbol{p}}_i$ is obtained in the matrix form

$$\bar{\boldsymbol{p}}_i=(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{Y} \tag{30}$$

Until now, the least-squares problem can be solved online with a sufficient number of data collected along the state trajectory. According to Lemma 2, the convergence of the online adaptive IRL algorithm can be guaranteed in finite iteration steps.

Remark 5 The developed adaptive optimal control is a type of data-driven method, where the system matrix is not needed. In fact, the algorithm can be also employed into time-varying system. If matrix $\boldsymbol{A}$ of the system changes suddenly, as long as the current controller of the new matrix is stable, the algorithm converges to the corresponding solution of the new ARE.

3 Numerical Simulation

In this section, numerical simulations are conducted to validate the performance of the proposed online adaptive IRL control scheme for stabilizing the swing motion of STS after capturing payload. Some parameters of the system are provided in Table 1. The desired dimensionless state of STS is $(\varepsilon_0 \;\; \dot{\varepsilon}_0 \;\; \theta \;\; \dot{\theta})=(0 \quad 0 \quad 0 \quad 0)$. Due to the impact of payload, the tether swings up to certain libration angle with varying length, the initial state of STS is assumed as $(\varepsilon_0 \;\; \dot{\varepsilon}_0 \;\; \theta \;\; \dot{\theta})=(-0.003\,3 \;\; 0 \;\; 0.174\,6 \;\; 0)$.

Table 1 Specific parameters of the tethered system

Parameter	Value
Orbital radius r/km	7 371
Mass of space tug m_A/kg	1 600
Mass of capture mechanism m_1/kg	50
Mass of payload m_p/kg	500
Nominal length of tether l_c/m	1 000
Density of the tether ρ/(kg•km^{-1})	0.198

In view of the unknown mass parameter of payload in general cases, the the initial parameters of the controller is deduced based on the linearized model of STS before capturing payload, which can guarantee the initial stability of controller. Then the feedback gain will be updated to satisfy the optimal control of STS in the post-capture period until convergence. It should be noted that the designed controller is applied into the original nonlinear plant. Since the onboard computational source limited, the sampling interval is set as $\tau=0.05$.

In addition to parameters of mission scenario, the controller parameters are well selected to make sure the convergence of algorithm and control performance. It is reasonable that the better control performance of the in-plane libration angle can be obtained by selecting larger corresponding weights in cost function. Therefore, the symmetric weight matrices are chosen as $\boldsymbol{Q}=\mathrm{diag}(10\ 2\ 1\ 1)$ and $\boldsymbol{R}=1$. The parameter N_N represents that an iterative up-

date is performed after N_N sampling time steps. Here N_N is set to 20, which means that the system iterates once every $N_N \cdot \tau = 20 \times 0.05 = 1$(rad) to update the control gain $\boldsymbol{K}$ and critic parameter matrix $\boldsymbol{P}$. The parameter $\varepsilon = 10^{-3}$ denotes the error threshold, representing that the iteration process will stop when the cost function V_t satisfies the requirement.

The simulation results are shown in Figs.3—8. Fig.3 shows the evolution of the parameters of matrix $\boldsymbol{P}$ in the Riccati equation. The matrix $\boldsymbol{P}$ converges to a constant optimal value after four iterations, which means that the online learning process is completed and the final control gain is determined after around 4 rad. The error norms of critic matrix and gain matrix in the policy iteration are presented in Fig.4. It can be seen that the matrix $\boldsymbol{K}$ will be close to the ideal optimal control gain after several policy

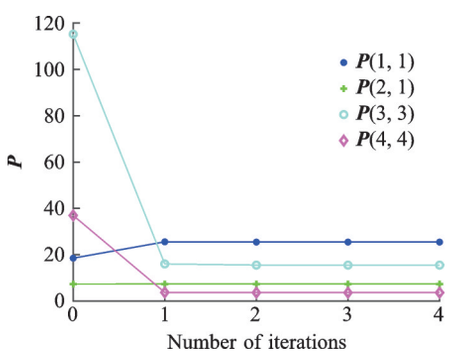

Fig.3 Critic matrix $\boldsymbol{P}$ of controller

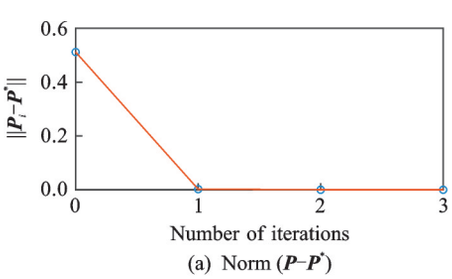

Fig.4 Error norms of matrices $\boldsymbol{P}$ and $\boldsymbol{K}$

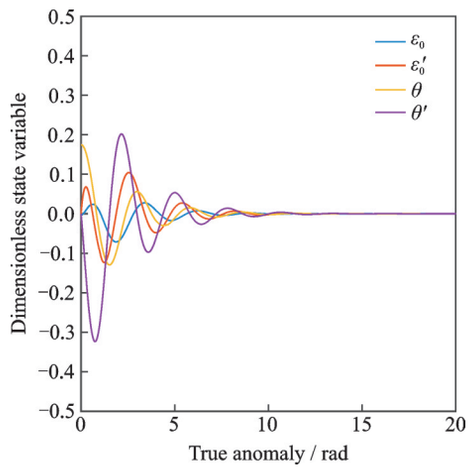

Fig.5 Dimensionless state variables under IRL control

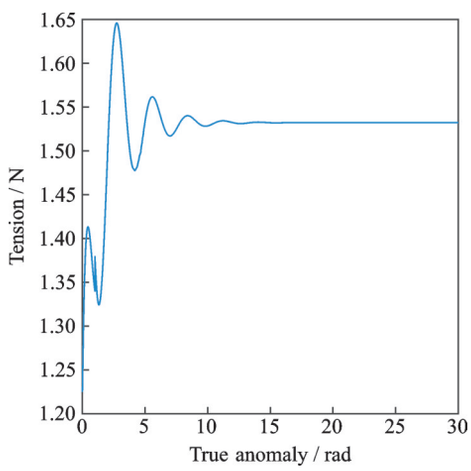

Fig.6 Variation curve of tether tension in the control process

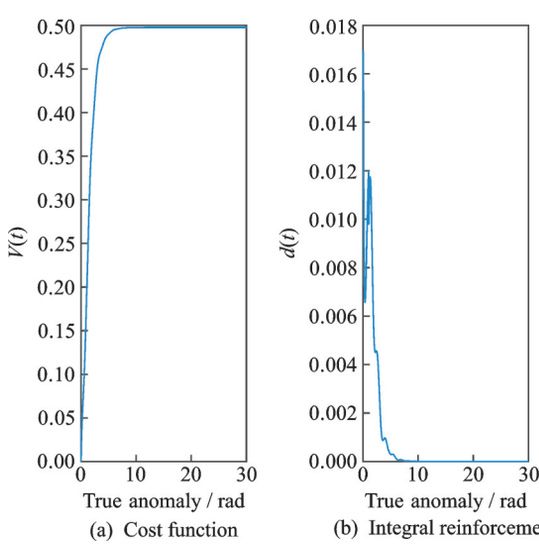

Fig.7 Cost function and integral reinforcement of the controller

updates. From Fig.5, the dimensionless states of STS converge from the initial position in the post-

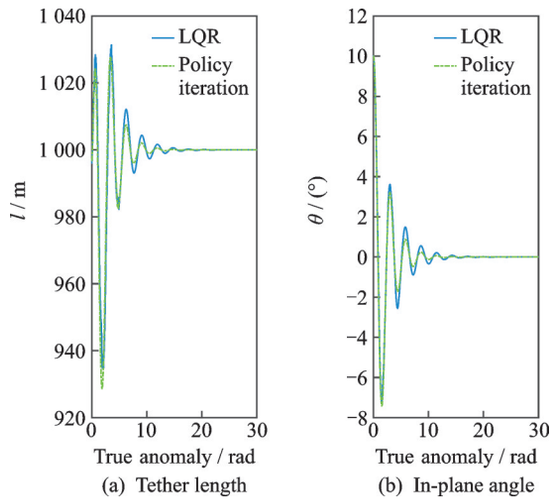

Fig.8 Comparison between policy iteration and LQR

capture phase to zeros within 15 rad. The maximum libration angle is no more than the initial value, and the amplitude of tether oscillation is less than its nominal length, avoiding the risk of collision between the tether and the satellite.

The variation curve of tether tension in Fig.6 indicates that the tether always keeps tense in the control stage, and magnitude of the tension meets the physical characteristics of the tether. Fig.7 shows the evolution of the total cost function (augmented state) and the integral reinforcement signal (one time-step cost) in the optimal control. It can be seen that the cost function $V(t)$ increases gradually over time and ultimately converges to a positive constant. On the contrary, the integral reinforcement has the trend of decreasing gradually, even though transient rise occurs in the online learning stage due to large deviation between initial control gain matrix and desired one.

In order to illustrate the robustness performance of the proposed algorithm, we compared it with the classic LQR controller by simulations under the presence of stochastic disturbance. For our simulations, the Gaussian noise with mean value of 0 and variance of 1 is adopted as stochastic disturbance, and the control parameters $\boldsymbol{Q}$ and $\boldsymbol{R}$ are chosen to be the same for both controllers. The gain matrix of LQR controller is deduced by the known dynamic model of STS before payload capture. Simulation results are shown in Fig.8. Both methods can ensure that the tether length and libration angle converge to desired values, while the method based on policy iteration has better convergence speed and control performance.

4 Conclusions

An adaptive optimal controller based on IRL policy iteration is conducted to address stabilization control of the tether libration after capturing the payload by STS. Due to lack of accurate dynamic model of the system in the post-capture stage, the classic model-based control methods will result in poor control effect. The proposed algorithm can achieve continuous time optimal control without accurately understanding the internal dynamics of the system, thus effectively solving the libration control problem of STS. Firstly, the basic dynamic model of STS is derived considering tether elasticity. Then the policy iteration based IRL algorithm is designed and the batch solution method is proposed for online implementing the algorithm. Finally, the effectiveness of the proposed control scheme is validated by the numerical simulation. The drawback is that the proposed method relies on linear dynamic system, causing poor control performance or even instability for the case of large libration amplitude. Our future work will focus on developing model-free adaptive optimal control scheme that can be directly applied into nonlinear dynamic system.

References

[1] ASLANOV V S, LEDKOV A S. Tether-assisted reentry capsule deorbiting from an elliptical orbit[J]. Acta Astronautica, 2016. DOI: 10.1016/j. actaastro.2016.10.028.

[2] YAMAIGIWA Y, HIRAGI E, KISHIMOTO T. Dynamic behavior of electrodynamic tether deorbit system on elliptical orbit and its control by Lorentz force [J]. Aerospace Science and Technology, 2005, 9 (4): 366-373.

[3] LU H, WANG C, LI A, et al. Low orbit debris mitigation using momentum exchange tether system[J] Journal of Beijing Institute of Technology, 2016, 4 (2): 106-120.

[4] LU H, WANG C, YURIY Z, et al. Optimal control

of payload tossing using space tethered system[J]. IFAC-PapersOnLine, 2016, 49(17): 272-277.

[5] SUN L, ZHAO G W, HUANG H. Effect of mass variation on dynamics of tethered system in orbital maneuvering[J]. Acta Astronautica, 2018, 146: 14-23.

[6] MARTIN K, LANDAU D, LONGUSKI J. Method to maintain artificial gravity during transfer maneuvers for tethered spacecraft[J]. Acta Astronautica, 2016, 120: 138-153.

[7] GOU X W, LI A J, TIAN H C, et al. Overload control of artificial gravity facility using spinning tether system for high eccentricity transfer orbits[J]. Acta Astronautica, 2018, 147(6): 383-392.

[8] YU B S, WEN H, JIN D P. Review of deployment technology for tethered satellite systems[J]. Acta Mechanica Sinica, 2018, 34(4): 754-768.

[9] HOVELL K, ULRICH S. Postcapture dynamics and experimental validation of subtethered space debris[J]. Journal of Guidance Control and Dynamics, 2017, 41(2): 1-7.

[10] SUN X, ZHONG R. Libration control for the low-thrust space tug system using electrodynamic force[J]. Journal of Guidance Control and Dynamics, 2018, 41(7): 1-8.

[11] ZHONG R, ZHU Z H. Attitude stabilization of tug-towed space target by thrust regulation in orbital transfer[J]. IEEE/ASME Transactions on Mechatronics, 2019, 24(1): 373-383.

[12] WANG C Q, WANG P B, LI A J, et al. Deployment of tethered satellites in low-eccentricity orbits using adaptive sliding mode control[J]. Journal of Aerospace Engineering, 2017, 30(6): 04017077.

[13] KANG J, ZHU Z H. A unified energy-based control framework for tethered spacecraft deployment[J]. Nonlinear Dynamics, 2018, 95: 1117-1131.

[14] LU Y, HUANG P, MENG Z. Adaptive neural network dynamic surface control of the post-capture tethered spacecraft[J]. IEEE Transactions on Aerospace and Electronic Systems, 2019, 56(2): 1406-1419.

[15] REZA M. Robust performance control of space tether deployment using fractional order tension law[J]. Journal of Guidance Control and Dynamics, 2019, 43(2): 1-7.

[16] WANG B, MENG Z, JIA C, et al. Anti-tangle control of tethered space robots using linear motion of tether offset[J]. Aerospace Science and Technology, 2019, 89: 163-174.

[17] LIU C, DONG C, ZHOU Z, et al. Barrier Lyapunov function-based reinforcement learning control for air-breathing hypersonic vehicle with variable geometry inlet[J]. Aerospace Science and Technology, 2020, 96: 105537.

[18] RAZMI H, AFSHINFAR S. Neural network-based adaptive sliding mode control design for position and attitude control of a quadrotor UAV[J]. Aerospace Science and Technology, 2019, 91: 12-27.

[19] VAMVOUDAKIS K G, LEWIS F L, HUDAS G R. Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality[J]. Automatica, 2012, 48(8): 1598-1611.

[20] LEWIS F L, VRABIE D, VAMVOUDAKIS K G. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers[J]. IEEE Control Systems Magazine, 2012, 32(6): 76-105.

[21] MODARES H, LEWIS F L. A policy iteration approach to online optimal control of continuous-time constrained-input systems[J]. ISA Transactions, 2013, 52(5): 611-621.

[22] DOYA K. Reinforcement learning in continuous time and space[J]. Neural Computation, 2000, 12(1): 219-245.

[23] VRABIE D, PASTRAVANU O, ABU-KHALAF M, et al. Adaptive optimal control for continuous-time linear systems based on policy iteration[J]. Automatica, 2009, 45(2): 477-484.

[24] JIANG Y, JIANG Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics[J]. Automatica, 2012, 48(10): 2699-2704.

[25] LEWIS F L, SYRMOS V L. Optimal control[M]. 2nd ed. New York: John Wiley & Sons, 1995.

[26] KLEINMAN D L. On an iterative technique for Riccati equation computations[J]. IEEE Transactions on Automatic Control, 1968, 13(1): 114-115.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No.62111530051), the Fundamental Research Funds for the Central Universities (No.3102017JC06002) and the Shaanxi Science and Technology Program, China (No.2017KW-ZD-04).

Authors Mr. **FENG Yiting** received the M.S. degree in control science and engineering from Northwestern Polytechnical University, China, in 2020. He is currently pursuing the Ph.D. degree in control science and engineering at Northwestern Polytechnical University, China. His current research interests include nonlinear control, adaptive control, intelligent control and space tether system dynamic analysis and control.

Prof. **WANG Changqing** is currently a professor with School of Automation, Northwestern Polytechnical University. He received the B.S. degree in mechanical design and manufacturing and the M.S. degree in navigation guidance and control from Northwestern Polytechnical University, Xi'an, China, in 1996 and 2001. He received the Ph.D. degree in system analysis, control and information processing from National Research University Moscow Power Engineering Institute, Russia, in 2006. His current research interests include adaptive control, and space tether system dynamic analysis and control.

Author contributions Mr. **FENG Yiting** designed the control algorithm, contributed to the simulation and the analysis of the study and wrote the manuscript. Mr. **ZHANG Ming** contributed to the simulation and the analyzation of the study. Dr. **GUO Wenhao** contributed to the data for the simulation and validation of model. Prof. **WANG Changqing** contributed to the research approach, background of the study and interpreted the results. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: XU Chengting)

基于策略迭代的空间系绳载荷捕获自适应最优控制

冯毅庭[1], 张 鸣[2], 郭闻昊[2], 王长青[1]

(1.西北工业大学自动化学院,西安 710129,中国; 2.北京宇航系统工程研究所,北京 100076,中国)

摘要:研究了基于空间系绳系统载荷捕获后的摆振控制问题。载荷捕获会造成系绳的摆振并导致系绳偏离标称位置。由于捕获后系统存在未知的动力学参数,提出了基于策略迭代的自适应最优控制算法,应用于载荷捕获后系绳系统摆动的稳定控制。通过引入积分强化学习方法,在系统动力学未知情况下在线求解代数黎卡提方程。为了避免迭代方程求解的计算负担,采用最小二乘方法在线实施策略迭代算法。最后,通过数值仿真验证了算法的有效性。

关键词:空间系绳系统;载荷捕获;策略迭代;积分强化学习;状态反馈