输入受限下的超紧密航天器编队避撞相对位置强化学习控制

孟亦真[1,2], 黄　静[1,2†], 周绍辉[3], 周　彬[1,2], 朱康武[1,2]

(1. 上海航天控制技术研究所, 上海 201109; 2. 上海市空间智能控制技术重点实验室, 上海 201109;

3. 上海航天空间技术有限公司, 上海 201109)

摘要: 考虑具有外界干扰、避撞约束和固定时间约束的近地轨道超紧密航天器编队的重构控制问题, 本文提出一种多约束条件下的考虑执行机构死区效应的航天器编队鲁棒控制方法. 首先, 建立近地轨道完整的编队航天器相对位置非线性动力学方程和执行机构死区动态响应模型; 其次, 根据状态约束条件设计编队相对位置约束机制, 基于反步法和强化学习评判–动作网络, 提出防避撞约束和固定时间约束的鲁棒控制律, 进一步考虑到执行机构电推力器的死区效应, 基于强化学习的动作网络来近似死区特性, 本文通过最小化评判网络代价函数来解决执行机构死区效应对控制精度带来的影响, 并应用Lyapunov稳定性定理证明其闭环系统的一致有界性; 最后, 在MATLAB/Simulink平台上进行仿真验证, 结果表明所提出方法的有效性.

关键词: 航天器编队; 避撞; 强化学习; 死区效应; 固定时间约束

引用格式: 孟亦真, 黄静, 周绍辉, 等. 输入受限下的超紧密航天器编队避撞相对位置强化学习控制. 控制理论与应用, 2025, 42(4): 659 – 668

DOI: 10.7641/CTA.2023.20792

Reinforcement learning control of collision avoidance for ultra-close formation of spacecraft with input constraints

MENG Yi-zhen[1,2], HUANG Jing[1,2†], ZHOU Shao-hui[3], ZHOU Bin[1,2], ZHU Kang-wu[1,2]

(1. Shanghai Aerospace Control Technology Institute, Shanghai 201109, China;
2. Shanghai Key Laboratory of Aerospace Intelligent Control Technology, Shanghai 201109, China;
3. Shanghai Space Technology Co., Ltd, Shanghai 201109, China)

Abstract: Considering the control problem of reconstructing the ultra-tight formation of near-Earth orbit spacecraft in the presence of external disturbances, collision avoidance constraints, and fixed-time constraints, this study presents a robust control method for spacecraft formation that accounts for the dead-zone effect of the actuator under multiple constraint conditions. Firstly, we establish the nonlinear dynamic equations governing the relative positions of the spacecraft in the complete near-Earth orbit formation, as well as the dynamic response model capturing the dead-zone effect of the actuator. Secondly, we design a constraint mechanism for the relative positions of the formation based on state constraints. Robust control laws, employing a combination of backstepping and a reinforcement learning actor-critic network, are proposed to address collision avoidance constraints and fixed-time constraints. Additionally, we approximate the dead-zone characteristics of the actuator's thrusters by leveraging a reinforcement learning actor network. To mitigate the impact of the dead-zone effect on control accuracy, we minimize the cost function of the actor network. The Lyapunov stability theorem is employed to demonstrate the uniformly boundedness of the closed-loop system. Finally, we conduct simulation verification on the MATLAB/Simulink platform, and the results substantiate the effectiveness of the proposed method.

Key words: spacecraft formation; collision avoidance; reinforcement learning control; dead-zone effect; fixed time constraint

Citation: MENG Yizhen, HUANG Jing, ZHOU Shaohui, et al. Reinforcement learning control of collision avoidance for ultra-close formation of spacecraft with input constraints. *Control Theory & Applications*, 2025, 42(4): 659 – 668

收稿日期: 2022–09–08; 录用日期: 2023–12–07.

†通讯作者. E-mail: huangjing04415@163.com.

本文责任编委: 贾英民.

国家重点研发计划项目(2022YFB3902700, 2022YFB3902702), 空间目标感知全国重点实验室资助.

Supported by the National Key R&D Program of China (2022YFB3902700, 2022YFB3902702) and the National Key Lab of Space Target Awareness.

1 引言

航天器编队飞行因其结构灵活、功能强大、可靠性高、生命周期长及发射风险低等优点, 已成为航天控制领域的研究热点. 与单个航天器任务相比, 航天器编队飞行面临的挑战是在较长时间内维持固定的航天器相对距离, 如深空干涉仪器、合成孔径雷达等. 这些复杂、精密的编队任务需对编队中各个航天器到达编队构型的时间、相对位置、控制精度进行约束, 以满足指定的一致性的任务需求(形成新构型、波束同步等). 由此, 便涉及到航天器系统状态约束、控制时间约束、分布式协同与电推进执行机构的智能控制问题[1].

有别于单个航天器任务, 以主–从结构编队组成的分布式雷达天线为例, 为实现高分辨率的观测, 天线阵面边缘之间的距离必须保持在1米之内. 为实现这种超近距离、高精度的航天器编队飞行任务, 如何快速形成编队构型, 同时避免航天器间的碰撞, 并能满足精度需求, 含有死区约束电推进下的分布式状态约束、避碰控制系统的设计俨然成为超紧密航天器编队重构任务的重中之重.

在航天器超紧密编队任务中, 其超近距离的重构, 涉及到航天器的系统状态约束控制以及受到外界扰动后的避碰问题. 针对于系统状态约束控制问题, 已有众多的研究成果, 如基于模型预测控制方法[2]的不变集方法[3]、障碍李雅普诺夫函数以及积分型障碍李雅普诺夫函数[4]. 文献[5]采用障碍Lyapunov函数结合反步法解决航天器预置性能下的状态约束问题. 而目前基于反步法结合障碍Lyapunov函数的方法在超紧密编队的全状态约束控制设计中, 还需要满足虚拟控制律的可行性条件, 即其必须要满足给定的约束区[6]. 为克服该预先给定约束域带来的应用限制, 本文提出一个仅依赖于系统状态的非线性约束函数, 以此来约束紧密航天器编队的相对位置, 实现控制目标. 同时, 因为超紧密编队, 在系统受到扰动时, 避免航天器间碰撞与系统约束状态的越界成为超紧密编队重构控制中的重要一环. 解决这个问题的常见方法是构造李雅普诺夫势函数, 采用李雅普诺夫函数完成协同控制和障碍势函数进行避撞, 随后, 设计由协调项和避碰项组成的集成控制器, 完成控制目标[7]. 该方法难以实现在避免碰撞的同时, 保持系统的相对位置状态约束. 为解决该问题, 借鉴强化学习能够基于评判网络对当前的控制效果进行评价, 而后通过动作网络的控制调节, 使得系统的收益在未来最大化[6]. 基于此, 本文拟用强化学习的指标函数, 来实现惩罚功能, 当超紧密编队卫星进入碰撞或者发生系统状态越界行为后, 强化学习评判此刻的控制效果差(输出较大的惩罚值), 进而调节动作网络的输出, 从而能够最小化评判网络的输出值, 迫使越界的航天器自动恢复到期望的约束区域, 避免因碰撞或者越界造成的损坏及构型破坏.

另一个值得关注的问题是编队控制的时间约束问题, 尤其是对于快速高精度控制系统, 比如多卫星系统的控制、多导弹飞行控制等, 要求各子系统在有限时间内能够达到一致性. 在已有的有限时间控制问题中, 收敛时间依赖于初值, 从而无法解决初值未知或比较大时, 难以满足系统编队的有限时间控制需求. 为克服该不足, 本文设计基于固定时间的控制策略, 其收敛时间有不依赖于初值的上界, 而收敛过程中的通讯距离由航天器的状态约束来保持[8]. 同时, 为保障航天器编队相对位置控制系统的精度, 电推进因其比冲较化学推进至少高一个量级, 同时兼具精度可调、冲量小且寿命长的优点, 成为航天器编队相对位置控制适宜的执行机构, 其能够提升航天器任务执行能力, 扩展其空间任务范围[9]. 但电推进器的死区效应给控制系统设计带来了负面影响.

针对多约束条件下的超紧密编队的控制问题, 本文的创新点如下:

1) 基于固定时间收敛的非线性控制机制. 在反步法的基础上, 通过引入固定时间收敛的非线性控制项, 解决超紧密编队分布式控制的有限时间约束问题.

2) 基于状态依赖的约束机制. 考虑航天器超紧密编队相对位置的约束, 提出基于状态依赖的约束机制, 以确保编队的相对位置约束.

3) 基于强化学习的评判–动作网络的软约束控制策略. 为解决航天器状态越界导致的碰撞和编队构型破坏问题, 设计基于强化学习的评判–动作网络的软约束控制策略. 通过最小化评判网络输出作为目标, 并优化动作网络的输出, 发挥其状态约束机制, 确保航天器间的相对位置约束.

本文提出的基于强化学习的电推进鲁棒控制算法创新地解决多约束条件下的超紧密编队的控制问题. 通过引入固定时间收敛的非线性控制项、基于状态依赖的约束机制和基于强化学习的评判–动作网络的软约束控制策略, 有效地应对约束条件和执行机构死区效应带来的挑战. 此外, 通过稳定性分析和MATLAB/Simulink平台上的仿真验证, 证明该方法的实时性和有效性.

2 编队航天器相对位置数学模型

本文采用主–从方式建立编队航天器相对位置数学模型, 假设每个航天器都是一个质点, 领航者在以恒定角速度运行在圆形轨道上, 而跟随者在领航者的周围按照所期望的参考轨迹运行. 为获得航天器编队飞行(spacecraft formation flying, SFF)的运动模型, 结合参考文献[10], 引入如图1所示的坐标系:

1) $C_{\rm I}=\{X,Y,Z\}$表示以地心为原点的惯性坐标系. 定义$\boldsymbol{R}$为该系下从地心指向领航者的位置矢量. $\boldsymbol{\rho}\in\mathbb{R}^3$为主航天器与从航天器间的相对位置矢量.

2) $C_{\rm L}=\{x_{\rm l},y_{\rm l},z_{\rm l}\}$为主航天器参考坐标系, $x_{\rm l}$的正方向与瞬时的切向速度相反; $Y_{\rm l}$的正方向为沿着矢径$\boldsymbol{R}$的方向; $Y_{\rm l}$垂直于$X_{\rm l}$和$Z_{\rm l}$组成的平面, $x_{\rm l},y_{\rm l},z_{\rm l}$符合右手准则.

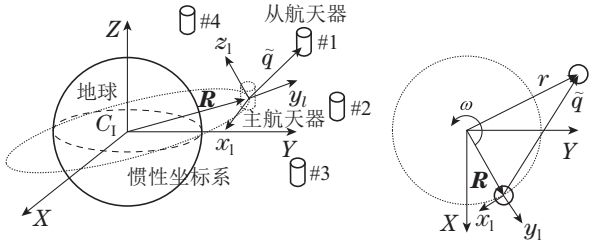

图 1 航天器编队运行示意图

Fig. 1 Schematic representation of the SFF system

在$C_{\rm I}$坐标系下, 主航天器和从航天器之间相对位置运动的动力学方程如下[1]:

$$m_{\rm f}\ddot{\boldsymbol{\rho}}+m_{\rm f}M_{\rm G}{\rm G}(\frac{\boldsymbol{R}+\boldsymbol{\rho}}{\|\boldsymbol{R}+\boldsymbol{\rho}\|^3}-\frac{\boldsymbol{R}}{\|\boldsymbol{R}\|^3})+\boldsymbol{F}_{\rm d}=\boldsymbol{u}_t, \tag{1}$$

其中: $m_{\rm l}$和$m_{\rm f}$分别表示主航天器和从航天器的质量; G为万有引力常数; $M_{\rm G}$表示地球的质量; $\boldsymbol{F}_{\rm d}\triangleq\boldsymbol{F}_{\rm df}-(\dfrac{m_{\rm f}}{m_{\rm l}})\boldsymbol{F}_{\rm dl}$为编队系统受到的扰动量, $\boldsymbol{F}_{\rm df},\boldsymbol{F}_{\rm dl}$分别表示作用在主航天器和从航天器上的外力干扰; $\boldsymbol{u}_{\rm t}=\boldsymbol{u}_{\rm f}-(\dfrac{m_{\rm f}}{m_{\rm l}})\boldsymbol{u}_{\rm l}$为系统的控制量, $\boldsymbol{u}_{\rm f},\boldsymbol{u}_{\rm l}$分别表示作用在主航天器和从航天器上的控制力. 进一步, 将位置矢量$\boldsymbol{\rho}$映射到动坐标系$C_{\rm L}$中, 可得

$$\boldsymbol{\rho}=x\vec{\boldsymbol{i}}_{\rm l}+y\vec{\boldsymbol{j}}_{\rm l}+z\vec{\boldsymbol{k}}_{\rm l}, \tag{2}$$

其中$\boldsymbol{i}_{\rm l},\boldsymbol{j}_{\rm l},\boldsymbol{k}_{\rm l}$表示$C_{\rm L}$三坐标轴的单位矢量, 那么它的二阶导数方程可以改写为

$$\begin{aligned}\ddot{\boldsymbol{\rho}}=&(\ddot{x}-2\omega_0\dot{y}-\omega_0^2x)\vec{\boldsymbol{i}}_{\rm l}+\\&(\ddot{y}+2\omega_0\dot{x}-\omega_0^2y)\vec{\boldsymbol{j}}_{\rm l}+\ddot{z}\vec{\boldsymbol{k}}_{\rm l},\end{aligned} \tag{3}$$

进一步, 将式(3)代入式(1)中, 可得跟随者相对于领航者的非线性相对位置平动动力学方程为

$$m_{\rm f}\ddot{\boldsymbol{\rho}}+m_{\rm f}\boldsymbol{C}(\boldsymbol{\omega})\dot{\boldsymbol{\rho}}+m_{\rm f}\boldsymbol{N}(\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{R},\boldsymbol{u}_{\rm l})=\boldsymbol{u}_{\rm f}, \tag{4}$$

其中: $\boldsymbol{C}(\omega_0)=2\omega_0\begin{bmatrix}0&-1&0\\1&0&0\\0&0&0\end{bmatrix}\in\mathbb{R}^{3\times3}$, $\boldsymbol{N}(\cdot)$为

$$\boldsymbol{N}(\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{R},\boldsymbol{u}_{\rm l})=M_{\rm G}{\rm G}[\frac{x}{\|\boldsymbol{R}+\boldsymbol{\rho}\|^3}-\frac{\omega_0^2yz}{\|\boldsymbol{R}+\boldsymbol{\rho}\|^3}-\omega_0^2x(\frac{y+\|\boldsymbol{R}\|}{\|\boldsymbol{R}+\boldsymbol{\rho}\|^3}-\frac{1}{\|\boldsymbol{R}\|^2})]^{\rm T}.$$

为了便于后续的推导, 基于式(1)(4), 可得SFF的模型为

$$\begin{cases}\dot{\boldsymbol{x}}_1=\boldsymbol{f}_1(\bar{\boldsymbol{x}}_1,\boldsymbol{p}_1)+\boldsymbol{g}_1(\bar{\boldsymbol{x}}_1)\boldsymbol{x}_2,\\\dot{\boldsymbol{x}}_2=\boldsymbol{f}_2(\bar{\boldsymbol{x}}_2,\boldsymbol{p}_2)+\boldsymbol{g}_2(\bar{\boldsymbol{x}}_2)\boldsymbol{N}(\boldsymbol{v}(t))+\boldsymbol{D},\end{cases} \tag{5}$$

其中: $\boldsymbol{x}_1=\boldsymbol{\rho},\boldsymbol{x}_2=\dot{\boldsymbol{\rho}},\boldsymbol{f}_1(\bar{\boldsymbol{x}}_1,\boldsymbol{p}_1)=0,\boldsymbol{g}_1(\bar{\boldsymbol{x}}_1)=\boldsymbol{I}_{3\times3}$, $\boldsymbol{f}_2(\bar{\boldsymbol{x}}_2,\boldsymbol{p}_2)=-\boldsymbol{M}_{\rm f}^{-1}\boldsymbol{C}(\boldsymbol{x}_2)\boldsymbol{x}_2-\boldsymbol{M}_{\rm f}^{-1}\boldsymbol{N}(\boldsymbol{x}_1)$, $\boldsymbol{g}_2(\bar{\boldsymbol{x}}_2)=\boldsymbol{M}_{\rm f}^{-1}$.

航天器受到的干扰$\boldsymbol{D}=-\boldsymbol{M}_{\rm f}^{-1}\boldsymbol{F}_{\rm d}$. $\boldsymbol{N}(\boldsymbol{v}(t))$为系统的控制输入量. 由于受限于与小卫星的重量、体积等因素, 具有重量小、比冲高的电推进器成为各国争相研究的重点. 在本文中, 采用电推力器作为航天器编队控制的执行机构, 重点考虑执行机构的死区特性, 定义为系统带有死区约束的输入控制量, 为执行机构提供的控制量, $\boldsymbol{N}(\boldsymbol{v}(t))=[N_1(v_1(t))\ \ N_2(v_2(t))\ \ N_3(v_3(t))]^{\rm T}$具体表达如下[11–12]:

$$N_i(v_i(t))=\begin{cases}h_{\rm r}(v_i(t)-o_{\rm r}),\ v_i(t)>o_{\rm r},\\0,\qquad\qquad\quad |v_i(t)|\leqslant o_{\rm r},\\h_{\rm l}(v_i(t)+o_{\rm l}),\ v_i(t)<o_{\rm l},\end{cases} \tag{6}$$

其中: $i=1,2,3$; $h_{\rm l},h_{\rm r}$分别为死区的左、右斜率; $o_{\rm l}$, $o_{\rm r}$分别为执行器的非线性断点. 经分析, $\boldsymbol{N}(\boldsymbol{v}(t))=\boldsymbol{v}(t)-\Delta\boldsymbol{v}(t),\Delta\boldsymbol{v}(t)$为死区效应带来的执行误差.

3 固定时间强化学习优化控制

3.1 航天器状态约束

考虑到编队航天器之间的避撞需求, 本文将引入系统的非对称状态约束, 设计如下的非线性状态依赖函数:

$$\zeta_i(t)=\frac{x_{1,i}(t)}{(F_{i1}(t)+x_{1,i}(t))(F_{i2}(t)-x_{1,i}(t))}, \tag{7}$$

其中: $i=1,2,3$; $\boldsymbol{\zeta}(t)=[\zeta_1(t)\ \ \zeta_2(t)\ \ \zeta_3(t)]^{\rm T}$, $\boldsymbol{x}_1(t)=[x_{11}(t)\ \ x_{12}(t)\ \ x_{13}(t)]^{\rm T}$; $F_{i1}(t)$和$F_{i2}(t)$为系统状态约束函数, 并且系统初始值$x_{1i}(0)$满足$-F_{i1}(0)<x_{1i}(0)<F_{i2}(0)$. 同时, 式(7)的引入可以直接处理系统的全状态约束, 并且避免了对虚拟控制律可行性条件的判别. 为便于后续的推导分析, 本文基于反步法引入如下的坐标变换:

$$\begin{cases}\boldsymbol{z}_1=\boldsymbol{\zeta}_1-\boldsymbol{\alpha}_0,\ \boldsymbol{z}_2=\boldsymbol{x}_2-\boldsymbol{\alpha}_{\rm 2f},\\\alpha_{0,i}=\dfrac{y_{{\rm d},i}}{(F_{i,1}(t)+y_{{\rm d},i})(F_{i,2}(t)-y_{{\rm d},i})},\end{cases} \tag{8}$$

其中: $i=1,2,3$; 状态约束下的跟踪误差$\boldsymbol{z}_1$定义为$\boldsymbol{z}_1=[z_{1,1}\ \ z_{1,2}\ \ z_{1,3}]^{\rm T}$; $\boldsymbol{\alpha}_0=[\alpha_{0,1}\ \ \alpha_{0,2}\ \ \alpha_{0,3}]^{\rm T}$为虚拟控制量; $\boldsymbol{\alpha}_{\rm 2f}=[\alpha_{\rm 2f,1}\ \ \alpha_{\rm 2f,2}\ \ \alpha_{\rm 2f,3}]^{\rm T}$为避免维数爆炸, 引入如下的一阶滤波器状态:

$$\varepsilon_1\dot{\boldsymbol{\alpha}}_{{\rm 2f},i}+\boldsymbol{\alpha}_{{\rm 2f},i}=\boldsymbol{\alpha}_{1,i}, \tag{9}$$

其中: $i=1,2,3$; $\varepsilon_1>0$; 虚拟控制量$\boldsymbol{\alpha_1}=[\alpha_{1,1}\ \ \alpha_{1,2}\ \ \alpha_{1,3}]^{\rm T}$作为一阶惯性环节的输入. 进一步, 定义如下因一阶惯性环节引入引起的误差量$\boldsymbol{y}_1$, 具体为

$$\boldsymbol{y}_1=\boldsymbol{\alpha}_{\rm 2f}-\boldsymbol{\alpha}_1, \tag{10}$$

其中$\boldsymbol{y}_1=[y_{11}\ \ y_{12}\ \ y_{13}]^{\mathrm{T}}$.

下面基于反步法的思想, 分步骤详细阐述控制器算法的设计过程:

步骤 1 对$\boldsymbol{\zeta}_1$关于时间求导, 可得

$$\dot{\boldsymbol{\zeta}}_1=\boldsymbol{\eta}_1\dot{\boldsymbol{x}}_1, \tag{11}$$

因此, 结合式(11), 可得$\boldsymbol{z}_1=\boldsymbol{\zeta}_1-\boldsymbol{\alpha}_0$关于时间$t$的导数

$$\dot{\boldsymbol{z}}_1=\boldsymbol{\eta}_1(\boldsymbol{g}_1(\boldsymbol{x}_1)\boldsymbol{x}_2+\boldsymbol{f}_1)-\boldsymbol{\eta}_{\mathrm{d}1}\dot{\boldsymbol{y}}_{\mathrm{d}}, \tag{12}$$

其中:

$$\boldsymbol{\eta}_{\mathrm{d}1}=[\eta_{\mathrm{d}1,1}\ \ \eta_{\mathrm{d}1,2}\ \ \eta_{\mathrm{d}1,3}]^{\mathrm{T}},$$
$$\eta_{\mathrm{d}1,i}=\frac{F_{i,1}F_{i,2}+y_{\mathrm{d},i}^2}{(F_{i,1}+y_{\mathrm{d},i})^2(F_{i,2}-y_{\mathrm{d},i})^2},\ i=1,2,3.$$

结合$\boldsymbol{z}_2$与$\boldsymbol{y}_1$的定义, 可得$\boldsymbol{x}_2=\boldsymbol{z}_2+\boldsymbol{y}_1+\boldsymbol{\alpha}_1$. 进一步, 对$\frac{1}{2}\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1$求关于时间的$t$的导数, 可得

$$\boldsymbol{z}_1^{\mathrm{T}}\dot{\boldsymbol{z}}_1=\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_1\boldsymbol{g}_1\boldsymbol{\alpha}_1+\mathcal{P}_1, \tag{13}$$

其中$\mathcal{P}_1=\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_1\boldsymbol{f}_1+\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_1\boldsymbol{g}_1(\boldsymbol{z}_2+\boldsymbol{y}_1)-\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_{\mathrm{d}1}\dot{\boldsymbol{y}}_{\mathrm{d}}$. 根据如下的杨氏不等式:

$\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_1\boldsymbol{f}_1\leqslant\underline{g}_1\|\boldsymbol{z}_1\|^2\|\boldsymbol{\eta}_1\|^2\|\boldsymbol{f}_1\|^2+\frac{1}{4\underline{g}_1}$, $\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_1\boldsymbol{g}_1\boldsymbol{z}_2\leqslant\underline{g}_2\|\boldsymbol{z}_1\|^2\|\boldsymbol{\eta}_1\|^2\|\boldsymbol{z}_2\|^2+\frac{\bar{g}_1^2}{4\underline{g}_2}$, $\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_1\boldsymbol{g}_1\boldsymbol{y}_1\leqslant\underline{g}_1\frac{\bar{g}_1^2}{\underline{g}_1}\|\boldsymbol{z}_1\|^2\times\|\boldsymbol{\eta}_1\|^2+\frac{1}{4}\|\boldsymbol{y}_1\|^2$, $-\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_{\mathrm{d}1}\dot{\boldsymbol{y}}_{\mathrm{d}}\leqslant\underline{g}_1\|\boldsymbol{z}_1\|^2\|\boldsymbol{\eta}_{\mathrm{d}1}\|^2\|\dot{\boldsymbol{y}}_{\mathrm{d}}\|^2+\frac{1}{4\underline{g}_1}$, 因此, $\mathcal{P}_1$可化简为

$$\mathcal{P}_1\leqslant\underline{g}_1b_1\|\boldsymbol{z}_1\|^2\Pi_1+\Delta_1+\underline{g}_2\|\boldsymbol{z}_1\|^2\|\boldsymbol{\eta}_1\|^2\|\boldsymbol{z}_2\|^2+\frac{\|\boldsymbol{y}_1\|^2}{4}, \tag{14}$$

其中: $b_1=\max\{1,\|\boldsymbol{f}_1\|^2,\ \frac{\bar{g}_1^2}{\underline{g}_1}\}$; $\underline{g}_1,\underline{g}_2$为正定参数; $\Pi_1=2\|\boldsymbol{\eta}_1\|^2+\|\boldsymbol{\eta}_{\mathrm{d}1}\|^2\|\dot{\boldsymbol{y}}_{\mathrm{d}}\|^2$已知, 同时为可计算的标量函数; $\Delta_1=\frac{1}{2\underline{g}_1}+\frac{\bar{g}_1^2}{4\underline{g}_2}$为未知的正常数; $\underline{g}_1,\underline{g}_2>0$代表控制分配矩阵范数的下界, 同理$\bar{g}_1,\bar{g}_2>0$指代的是控制分配矩阵的上界.

进一步, 可将式(13)化简为

$$\boldsymbol{z}_1^{\mathrm{T}}\dot{\boldsymbol{z}}_1\leqslant\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\eta}_1\boldsymbol{g}_1\boldsymbol{\alpha}_1+\underline{g}_1b_1\|\boldsymbol{z}_1\|^2\Pi_1+\Delta_1+\underline{g}_2\|\boldsymbol{z}_1\|^2\|\boldsymbol{\eta}_1\|^2\|\boldsymbol{z}_2\|^2+\frac{1}{4}\|\boldsymbol{y}_1\|^2, \tag{15}$$

从而, 设计如下的虚拟控制律$\boldsymbol{\alpha}_1$:

$$\begin{cases}\boldsymbol{\alpha}_1=-\mathrm{diag}^{-1}\{\boldsymbol{\eta}_1\}(\boldsymbol{k}_1\boldsymbol{z}_1+\boldsymbol{z}_1\hat{b}_1\Pi_1+\\ \qquad l_1\mathrm{sig}^{\frac{m_1}{n_1}}(\boldsymbol{z}_1)+l_2\mathrm{sig}^{\frac{p_1}{r_1}}(\boldsymbol{z}_1)),\\ \dot{\hat{b}}_1=\ell_1(\|\boldsymbol{z}_1\|^2\Pi_1-\ell_{\hat{b}_1}\hat{b}_1),\end{cases} \tag{16}$$

其中: $\boldsymbol{k}_1=\mathrm{diag}\{k_{11},k_{22},k_{33}\}$, $k_{11},k_{22},k_{33},\ell_1,\ell_{\hat{b}1}$为待设计的正常数; $\tilde{b}_1=b_1-\hat{b}_1$为估计误差, $\hat{b}_1$为b_1的估计值. 函数$\mathrm{sig}^{\frac{p_1}{r_1}}(\cdot)=\mathrm{sgn}(\cdot)|\cdot|^{\frac{p_1}{r_1}}$.

在进行对虚拟控制量$\boldsymbol{\alpha}_1$的稳定性分析之前, 结合$\boldsymbol{y}_1$的定义(10), 可得$\dot{\boldsymbol{y}}_1=-\frac{\boldsymbol{y}_1}{\varepsilon_1}+\boldsymbol{\hbar}_1$, $\boldsymbol{\hbar}_1=-\dot{\boldsymbol{\alpha}}_1=-\frac{\partial\boldsymbol{\alpha}_1}{\partial\boldsymbol{\eta}_1}\dot{\boldsymbol{\eta}}_1-\frac{\partial\boldsymbol{\alpha}_1}{\partial\boldsymbol{z}_1}\dot{\boldsymbol{z}}_1-\frac{\partial\boldsymbol{\alpha}_1}{\partial\hat{b}_1}\dot{\hat{b}}_1-\frac{\partial\boldsymbol{\alpha}_1}{\partial\Pi_1}\dot{\Pi}_1$. 因此可得

$$\boldsymbol{y}_1^{\mathrm{T}}\dot{\boldsymbol{y}}_1=-\frac{\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1}{\varepsilon_1}+\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{\hbar}_1\leqslant(\frac{1}{4}-\frac{1}{\varepsilon_1})\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1+\boldsymbol{\hbar}_1^{\mathrm{T}}\boldsymbol{\hbar}_1.$$

而后, 通过选取如下的Lyapunov函数: $V_1=\frac{1}{2}\boldsymbol{z}_1^{\mathrm{T}}\times\boldsymbol{z}_1+\frac{g_1}{2\ell_1}\tilde{b}_1^{\mathrm{T}}\tilde{b}_1+\frac{1}{2}\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1$, 并对其求关于时间$t$的导数

$$\begin{aligned}\dot{V}_1\leqslant&-\underline{g}_1\lambda_{\min}(\boldsymbol{k}_1)\|\boldsymbol{z}_1\|^2-\frac{\underline{g}_1\ell_{\hat{b}1}}{2}\|\tilde{b}_1\|^2-\bar{\varepsilon}_1\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1+\\&\boldsymbol{\hbar}_1^{\mathrm{T}}\boldsymbol{\hbar}_1+\underline{g}_2\|\boldsymbol{z}_1\|^2\|\boldsymbol{\eta}_1\|^2\|\boldsymbol{z}_2\|^2+\Gamma_1-\\&\boldsymbol{z}_1^{\mathrm{T}}\underline{g}_1(l_1\mathrm{sig}^{\frac{m_1}{n_1}}(\boldsymbol{z}_1)+l_2\mathrm{sig}^{\frac{p_1}{r_1}}(\boldsymbol{z}_1)),\end{aligned} \tag{17}$$

其中: $0<\underline{g}_1\leqslant g_1$, $\underline{g}_1\ell_{\hat{b}1}\tilde{b}_1^{\mathrm{T}}\hat{b}_1\leqslant-\frac{\underline{g}_1\ell_{\hat{b}1}}{2}\|\tilde{b}_1\|^2+\frac{\underline{g}_1\ell_{\hat{b}1}}{2}\times\|b_1\|^2$, $\Gamma_1=\Delta_1+\frac{\underline{g}_1\ell_{\hat{b}1}}{2}\|b_1\|^2$. $\underline{g}_2\|\boldsymbol{z}_1\|^2\|\boldsymbol{\eta}_1\|^2\|\boldsymbol{z}_2\|^2$将会在后续的控制器设计中消除, 此外, $\boldsymbol{\hbar}_1$将会在反步控制算法的最后一项中处理.

3.2 强化学习评判网络设计

结合控制框图(如图2所示), 为满足系统的状态约束, 提高系统的安全性, 本节在第3.1节状态约束的基础上, 设计如下的效用函数为软约束:

$$\boldsymbol{J}(t)=\int_t^{\infty}\gamma^{\frac{-\xi+t}{T}}\boldsymbol{p}(\boldsymbol{s}(\xi))\mathrm{d}\xi, \tag{18}$$

其中: ξ为关于时间变量t的积分变量; $T>0$为强化学习效用函数的解算区间; $\gamma\in(0,1)$为折扣因子, 用来减小当前权重对未来系统收益的影响, 更符合客观事实规律. 如果$\boldsymbol{s}$足够小, 则达到系统的优化目标.

1) 将受到干扰的系统状态重新约束到硬约束范围内;

2) 优化电推进器的控制输出, 使得系统的能量消耗越小越好, 即强化学习的效用函数不会再持续增长, 如果$\boldsymbol{s}$比较大, 则自适应调节控制器输出使得状态约束下的跟踪误差$\boldsymbol{z}_1$, $\dot{\boldsymbol{z}}_1$和$\boldsymbol{J}(t)$减小.

本文的控制目标主要是使状态约束下的跟踪误差$\boldsymbol{z}_1$, 强化学习的效用函数$\boldsymbol{p}(\boldsymbol{s})$, 状态量$\boldsymbol{s}$足够小的同时保障其有界. 进而, 容易获得强化学习指标函数的期望值$\boldsymbol{J}_{\mathrm{d}}=[0\ \ 0\ \ 0]^{\mathrm{T}}$. 根据参考文献[13]的设计如下:

$$p_i(s_i(\xi))=\begin{cases}0,\ F_{i1}(t)<x_{1i}<F_{i2}(t),\ s_i(v(t))\leqslant\varepsilon\\ \varepsilon,\ F_{i1}(t)<x_{1i}<F_{i2}(t),\ s_i(v(t))>\varepsilon\\ \wp,\ x_{1i}\geqslant F_{i2}(t),\ x_{1i}\leqslant F_{i1}(t),\end{cases} \tag{19}$$

其中$\boldsymbol{s}(v(t))=\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{R}_{\mathrm{s}}\boldsymbol{z}_1+\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Q}_{\mathrm{s}}\boldsymbol{v}$. 为实现上述两项

优化目标, 强化学习目标函数$p_i(s_i(\xi))$与优化目标的关系分析如下:

1) 当系统相对距离约束在界限F_{i1}内, 同时, 系统的能量消耗小于设定阈值ε时, 判定系统的控制效果较好, 使得$p_i(s_i(\xi))=0$, 表明此时有效补偿电推进器死区特性的影响;

2) 当$p_i(s_i(\xi))=\varepsilon$时, 表明此时的死区特性控制补偿量并未达到较好的效果, 需通过调节控制输出, 减小效用函数的输出, 以此来优化死区补偿控制量的输出. 在最小化效用函数的过程中, 有效解决电推进器的死区效应, 完成目标2;

3) 当$p_i(s_i(\xi))=\wp$时, $\wp>0$为一个足够大的常数, 当系统状态溢出约束边界后, 起到惩罚作用, 使得系统状态重新返回约束边界内, 实现优化目标1.

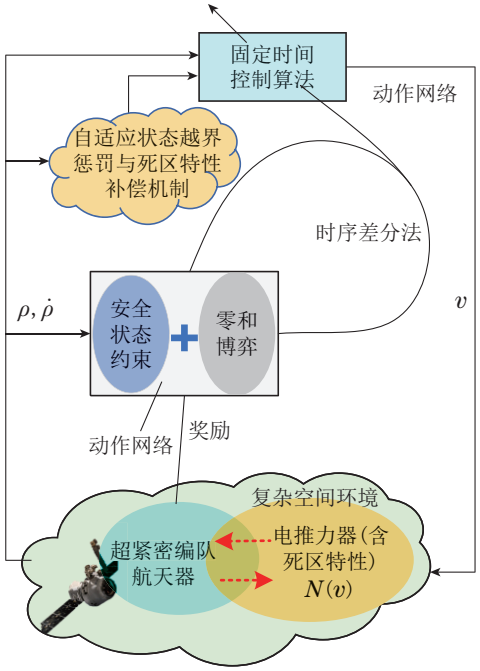

图 2　基于强化学习的超紧密航天器编队防避撞相对位置控制算法框图

Fig. 2　Diagram of relative position control for ultra-tight spacecraft formation collision avoidance based on reinforcement learning

依据参考文献[13], 强化学习的目标是使效用函数$\boldsymbol{J}(t)$越小越好. 结合式(18), 由于$\boldsymbol{J}(t)$的积分是到达无穷的, 为解决该优化指标无穷时域内的求解问题, 本文建如下的Bellman误差方程, 反映$\boldsymbol{J}(t-T)$与$\boldsymbol{J}(t)$的递推迭代关系:

$$\boldsymbol{J}(t-T)=\int_{t-T}^{\infty}\gamma^{\frac{-\xi+t-T}{T}}\boldsymbol{p}(\boldsymbol{s}(\xi))\mathrm{d}\xi=\gamma^{-1}(\boldsymbol{J}(t)+\boldsymbol{p}_{\mathrm{c}}),\tag{20}$$

其中$\boldsymbol{p}_{\mathrm{c}}=\int_{t-T}^{t}\gamma^{[\frac{(-\xi+1)}{T}]}\boldsymbol{p}(\boldsymbol{s}(\xi))\mathrm{d}\xi$为价值函数在积分区间$[t-T,t)$的约束值.

进一步, $\boldsymbol{p}_{\mathrm{c}}=[p_{\mathrm{c}1}\ \ p_{\mathrm{c}2}\ \ p_{\mathrm{c}3}]^{\mathrm{T}}$可得

$$p_{\mathrm{c}i}=\int_{t-T}^{t}\gamma^{\frac{-\xi+1}{T}}p_i(s_i(\xi))\mathrm{d}\xi=\begin{cases}0, & F_{i1}(t)<x_{1,i}<F_{i2}(t),\ s(\boldsymbol{v}(t))\leqslant\varepsilon,\\ \dfrac{\varepsilon(\gamma-1)}{\ln\gamma}T, & F_{i1}(t)<x_{1,i}<F_{i2}(t),\ s(\boldsymbol{v}(t))>\varepsilon,\\ \dfrac{\wp(\gamma-1)}{\ln\gamma}T, & x_{1,i}\geqslant F_{i2}(t),\ x_{1,i}\leqslant F_{i1}(t),\end{cases}\tag{21}$$

其中: $i=1,2,3$; $\|\boldsymbol{p}_{\mathrm{c}}\|\leqslant b_{p_{\mathrm{c}}}$, $b_{p_{\mathrm{c}}}>0$为一正常数.

基于参考文献[7], 采用径向基神经网络来逼近强化学习评判指标函数$\boldsymbol{J}(t)$, 具体为

$$\boldsymbol{J}(t)=\boldsymbol{W}_{\mathrm{c}}^{*\mathrm{T}}\boldsymbol{\phi}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}}(t))+\boldsymbol{\varepsilon}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}}(t)),\tag{22}$$

其中: $\boldsymbol{W}_{\mathrm{c}}^{*}$为理想的权重, 存在正常数$b_{W_{\mathrm{c}}}$使得$\|\boldsymbol{W}_{\mathrm{c}}^{*}\|_{\mathrm{F}}\leqslant b_{W_{\mathrm{c}}}$; $\boldsymbol{x}_{\mathrm{c}}=[\boldsymbol{x}_1^{\mathrm{T}}(t)\ \ \boldsymbol{z}_1^{\mathrm{T}}(t)\ \ \boldsymbol{x}_{\mathrm{d}}^{\mathrm{T}}(t)]^{\mathrm{T}}$为评判神经网络的输入量. 基于高斯基函数, 径向基神经网络的$\boldsymbol{\phi}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}})$的表达式为

$$\boldsymbol{\phi}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}})=\exp[\frac{-\|\boldsymbol{x}_{\mathrm{c}}-\boldsymbol{c}_{\mathrm{r}}\|}{b_{\mathrm{r}}^2}],\tag{23}$$

其中: $\boldsymbol{c}_{\mathrm{r}}$为径向基神经网络的中心值, 同时, b_{r}为径向基神经网络的宽度. 进一步, 采用时间差分算法, 可得

$$\begin{aligned}&\boldsymbol{e}_{\mathrm{c}}=\\&\hat{\boldsymbol{J}}(t)-\gamma\hat{\boldsymbol{J}}(t-T)+\boldsymbol{p}_{\mathrm{c}}=\hat{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{c}}(t)+\boldsymbol{p}_{\mathrm{c}}=\\&\hat{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{c}}(t)+\hat{\boldsymbol{p}}_{\mathrm{c}}+\boldsymbol{W}_{\mathrm{c}}^{*\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{c}}(t),\end{aligned}\tag{24}$$

其中$\Delta\boldsymbol{\phi}_{\mathrm{c}}(t)=[\boldsymbol{\phi}_{\mathrm{c}}(x_{\mathrm{c}}(t))-\gamma\boldsymbol{\phi}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}}(t-T))]$. 进一步, 可得$\|\Delta\boldsymbol{\phi}_{\mathrm{c}}(t)\|\leqslant(1+\gamma)b_{\phi_{\mathrm{c}}}$. 最终, 评判网络的权重$\hat{\boldsymbol{W}}_{\mathrm{c}}$更新律, 设计为

$$\dot{\hat{\boldsymbol{W}}}_{\mathrm{c}}=-\boldsymbol{\Gamma}_{\mathrm{c}}\Delta\boldsymbol{\phi}_{\mathrm{c}}(t)[\hat{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{c}}(t)+\boldsymbol{p}_{\mathrm{c}}]^{\mathrm{T}}-\sigma_{\mathrm{c}}\boldsymbol{\Gamma}_{\mathrm{c}}\hat{\boldsymbol{W}}_{\mathrm{c}},\tag{25}$$

其中: $\boldsymbol{\Gamma}_{\mathrm{c}}=\mathrm{diag}\{\Gamma_{\mathrm{c}_1},\Gamma_{\mathrm{c}_2},\Gamma_{\mathrm{c}_3}\}>0$为神经网络的学习率; $\sigma_{\mathrm{c}}\boldsymbol{\Gamma}_{\mathrm{c}}\hat{\boldsymbol{W}}_{\mathrm{c}}$为神经网络权重更新率的修正项, 用于放宽神经网络的权重参数学习的持续激励约束. 在式(25)中, 第1项用来减小估计误差$\boldsymbol{E}_{\mathrm{c}}=\dfrac{1}{2}\boldsymbol{e}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{e}_{\mathrm{c}}$, 同时, 第2项用来增强神经网络权重参数学习过程的鲁棒性, 并解决持续激励缺失引起的评价网络权重更新梯度消失问题. 至此, 强化学习的评判网络设计完成.

3.3　融合强化学习执行网络的固定时间控制算法设计

将跟踪误差$\boldsymbol{z}_2=[z_{21}\ \ z_{22}\ \ z_{23}]^{\mathrm{T}}$关于时间求导数, 可得

$$\dot{\boldsymbol{z}}_2=\boldsymbol{f}_2+\boldsymbol{g}_2\boldsymbol{N}(\boldsymbol{v}(t))+\boldsymbol{D}-\dot{\boldsymbol{\alpha}}_{2\mathrm{f}},\tag{26}$$

进一步, 可得

$$\boldsymbol{z}_2^{\mathrm{T}}\dot{\boldsymbol{z}}_2=\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{g}_2\boldsymbol{N}(\boldsymbol{v}(t))+\mathcal{P}_2,\tag{27}$$

其中$\mathcal{P}_2 = \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{f}_2 - \boldsymbol{z}_2^{\mathrm{T}}\dot{\boldsymbol{\alpha}}_{2\mathrm{f}} + \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{D}$.

设计如下的Lyapunov函数:

$$\begin{cases} V_2 = V_{21}+V_{22},\ V_{22} = \dfrac{\underline{g}_2 l_{\mathrm{v}}}{2}\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{c}}^{-1}\tilde{\boldsymbol{W}}_{\mathrm{c}}),\\ V_{21} = V_1 + \dfrac{1}{2}\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2 + \dfrac{\underline{g}_2}{2\ell_2}\tilde{b}_2^{\mathrm{T}}\tilde{b}_2 + \\ \qquad \dfrac{\underline{g}_2}{2}\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{a}}^{-1}\tilde{\boldsymbol{W}}_{\mathrm{a}}), \end{cases} \tag{28}$$

其中: $\tilde{b}_2 = b_2 - \hat{b}_2$为参数估计误差, $\hat{b}_2$为b_2的估计值; $\ell_2 > 0$为设计参数. 对V_{21}求关于时间的导数, 可得

$$\begin{aligned} &\dot{V}_{21} \leqslant \\ &\dot{V}_1 + \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{g}_2\boldsymbol{N}(\boldsymbol{v}(t)) + \mathcal{P}_2 - \frac{\underline{g}_2}{\ell_2}\tilde{b}_2^{\mathrm{T}}\dot{\hat{b}}_2 + \\ &\underline{g}_2\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Gamma}_{\mathrm{a}}^{-1}\dot{\tilde{\boldsymbol{W}}}_{\mathrm{a}}), \end{aligned} \tag{29}$$

进一步, 再次结合杨氏不等式, 可得

$$\begin{cases} \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{f}_2 \leqslant \underline{g}_2\|\boldsymbol{z}_2\|^2\|\boldsymbol{f}_2\|^2 + \dfrac{1}{4\underline{g}_2},\\ -\boldsymbol{z}_2^{\mathrm{T}}\dot{\boldsymbol{\alpha}}_{2\mathrm{f}} \leqslant \underline{g}_2\|\boldsymbol{z}_2\|^2\|\dot{\boldsymbol{\alpha}}_{2\mathrm{f}}\|^2 + \dfrac{1}{4\underline{g}_2}, \end{cases} \tag{30}$$

进一步, 可得

$$\mathcal{P}_2 \leqslant \underline{g}_2 b_2\|\boldsymbol{z}_2\|^2\Pi_2 + \Delta_2 + \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{D}, \tag{31}$$

其中: $b_2 = \max\{1, \|\boldsymbol{f}_2\|^2\}$, $\Pi_2 = 1 + \|\boldsymbol{\eta}_1\|^2\|\boldsymbol{z}_1\|^2 + \|\dot{\boldsymbol{\alpha}}_{2\mathrm{f}}\|^2$. $\Delta_2 = \dfrac{1}{2\underline{g}_2} > 0$. 同时, 考虑到电推进器的死区特性, $\Delta\boldsymbol{v}(t)$为未知量, 以及系统受到的外部有界干扰$\boldsymbol{D}$, 本文引入如下的复合未知项, 具体为[12, 14]

$$\boldsymbol{\Xi} = \boldsymbol{D} + \boldsymbol{g}_2\Delta\boldsymbol{v}(t). \tag{32}$$

本文借助于强化学习的Actor网络对上述未知量进行估计, 在消除未知量影响的同时, 优化系统的控制效果, 具体为

$$\boldsymbol{\Xi} = \boldsymbol{W}_{\mathrm{a}}^{*}\boldsymbol{\phi}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}}) + \boldsymbol{\varepsilon}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}}), \tag{33}$$

其中: $\boldsymbol{W}_{\mathrm{a}}^{*} \in \mathbb{R}^{N_{\mathrm{a}}\times n}$为理想权重, 满足$\|\boldsymbol{W}_{\mathrm{a}}^{*}\|_{\mathrm{F}} \leqslant b_{W_{\mathrm{a}}}$, $b_{W_{\mathrm{a}}}$为大于零的正常数; N_{a}为动作网络的隐层; $\boldsymbol{x}_{\mathrm{a}}$为强化学习动作网络的输入. $\boldsymbol{\phi}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}}) \in \mathbb{R}^{N_{\mathrm{a}}}$为径向基神经网络的激活函数, 类似于$\boldsymbol{\phi}_{\mathrm{c}}$, 见式(23). 同理, 其满足$\|\boldsymbol{\phi}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}})\| \leqslant b_{\phi_{\mathrm{a}}}$, 并且$b_{\phi_{\mathrm{a}}}$为大于零的正整数. 由于$\boldsymbol{W}_{\mathrm{a}}^{*}$为理想权重, 无法直接获得, 在本文中, 采用其估计, 具体如下:

$$\hat{\boldsymbol{\Xi}} = \hat{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\phi}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}}). \tag{34}$$

强化学习动作网络的目标有两个:

1) 使得越界的系统状态能够重新恢复到安全约束界限内, 起到系统状态软约束的目的;

2) 优化系统的控制输出, 使其输出能量越小越好, 即使得评判网络的估计输出$\hat{\boldsymbol{J}}(t)$最小化来逼近去最优值$\boldsymbol{J}_{\mathrm{d}} = \boldsymbol{0}$.

从而定义强化学习动作网络的估计误差为

$$\boldsymbol{e}_{\mathrm{a}} = \boldsymbol{z}_2 + \hat{\boldsymbol{J}}(t) - \boldsymbol{J}_{\mathrm{d}} = \boldsymbol{z}_2 + \hat{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{\phi}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}}), \tag{35}$$

由此, 设计$\hat{\boldsymbol{W}}_{\mathrm{a}}$的更新律如下:

$$\dot{\hat{\boldsymbol{W}}}_{\mathrm{a}} = \boldsymbol{\Gamma}_{\mathrm{a}}\boldsymbol{\phi}_{\mathrm{a}}(t)[\boldsymbol{z}_2 + \hat{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}})]^{\mathrm{T}} - \ell_{\mathrm{a}}\boldsymbol{\Gamma}_{\mathrm{a}}\hat{\boldsymbol{W}}_{\mathrm{a}}, \tag{36}$$

其中: $\boldsymbol{\Gamma}_{\mathrm{a}} = \mathrm{diag}\{\Gamma_{\mathrm{a}_1}, \Gamma_{\mathrm{a}_2}, \Gamma_{\mathrm{a}_3}\} > 0$为神经网络的学习率; $\ell_{\mathrm{a}}\boldsymbol{\Gamma}_{\mathrm{a}}\hat{\boldsymbol{W}}_{\mathrm{a}}$为神经网络权重更新率的修正项, 用于放宽神经网络的权重参数学习的持续激励约束. 在式(36)中, 第1项用于第2项的功能与评判网络权重$\hat{\boldsymbol{W}}_{\mathrm{c}}$更新律相同.

进一步, 设计如下的系统的控制器:

$$\begin{cases} \boldsymbol{v} = -(\boldsymbol{k_2}\boldsymbol{z}_2 + \hat{\boldsymbol{b}}_2\boldsymbol{z}_2\Pi_2 + l_3\mathrm{sig}^{\frac{m_2}{n_2}}(\boldsymbol{z}_2) + \\ \qquad \hat{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\phi}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}}) + l_4\mathrm{sig}^{\frac{p_2}{r_2}}(\boldsymbol{z}_2)),\\ \dot{\hat{b}}_2 = \ell_2(\|\boldsymbol{z}_2\|^2\Pi_2 - \ell_{\hat{b}_2}\hat{b}_2),\ \hat{b}_2(0) \geqslant 0, \end{cases} \tag{37}$$

其中:

$$\begin{aligned} &\mathrm{sig}^{\frac{m_2}{n_2}}(\boldsymbol{z}_2) = \\ &[\mathrm{sig}^{\frac{m_2}{n_2}}(z_{2,1})\ \ \mathrm{sig}^{\frac{m_2}{n_2}}(z_{2,2})\ \ \mathrm{sig}^{\frac{m_2}{n_2}}(z_{2,3})]^{\mathrm{T}},\\ &\mathrm{sig}^{\frac{p_2}{r_2}}(\boldsymbol{z}_2) = \\ &[\mathrm{sig}^{\frac{p_2}{r_2}}(z_{2,1})\ \ \mathrm{sig}^{\frac{p_2}{r_2}}(z_{2,2})\ \ \mathrm{sig}^{\frac{p_2}{r_2}}(z_{2,3})]^{\mathrm{T}}. \end{aligned}$$

4 稳定性证明

定理 1 考虑带有死区特性电推进器的航天器系统(4)–(6), 针对其状态安全约束控制(7), 采用固定时间控制器及其自适应律(16)(37), 结合强化学习评判网络(18)及其更新律(25), 以及用于克服死区约束的控制输出与越界状态的强化学习动作网络(33)及其更新律(36), 可得以下结论:

1) 所有闭环系统的信号都是一致最终有界的. 其收敛域为

$$\begin{aligned} &\|\tilde{\boldsymbol{W}}_{\mathrm{a}}\|_{\mathrm{F}} \leqslant \sqrt{\frac{2(\Upsilon + V_2(0))}{\underline{g}_2\lambda_{\min}(\boldsymbol{\Gamma}_{\mathrm{a}}^{-1})}},\\ &\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}} \leqslant \sqrt{\frac{2(\Upsilon + V_2(0))}{l_{\mathrm{v}}\lambda_{\min}(\boldsymbol{\Gamma}_{\mathrm{c}}^{-1})}},\\ &\|\tilde{b}_i\|_{\mathrm{F}} \leqslant \sqrt{\frac{2\gamma_i(\Upsilon + V_2(0))}{\underline{g}_i}},\ i = 1, 2,\\ &\|\boldsymbol{y}_2\|_{\mathrm{F}} \leqslant \sqrt{2(\Upsilon + V_2(0))}. \end{aligned}$$

2) 系统状态可在固定时间收敛至以下约束区域:

$$\begin{aligned} &\|z_{1,j}\| \leqslant \min\{(\frac{M}{(1-\theta)a})^{\frac{n}{n+m}}, (\frac{M}{(1-\theta)b})^{\frac{r}{r+p}}\},\\ &t_{\mathrm{s1}} \leqslant \frac{1}{a}\frac{2n}{m-n} + \frac{1}{b}\frac{2r}{r-p}, \end{aligned}$$

可通过适当参数调节来调节系统状态的收敛时间及收敛域.

3) 系统的状态将会保持在式(7)定义的安全约束范围内运行, 一旦受扰, 偏离安全约束范围, 强化学习惩罚机制(18)将使其重新恢复到约束域中.

证　结合设计的Lyapunov函数, 并代入控制律(37)及其自适应律, 可得

$$\begin{aligned}\dot{V}_{21} \leqslant\ & -\bar{\varepsilon}_1\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1-\sum_{i=1}^{2}\underline{g}_i\lambda_{\min}(\boldsymbol{k}_i)\|\boldsymbol{z}_i\|^2-\\ &\sum_{i=1}^{2}\frac{\underline{g}_i\ell_{\hat{b}2}}{2}\|\tilde{b}_i\|^2+\sum_{i=1}^{2}\varGamma_i+\hbar_1^{\mathrm{T}}\hbar_1-\\ &\underline{g}_2\boldsymbol{z}_2^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\phi}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}})+\boldsymbol{z}_2^{\mathrm{T}}\underline{g}_2\varepsilon_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}})-\\ &\ell_{\mathrm{a}}\underline{g}_2\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\hat{\boldsymbol{W}}_{\mathrm{a}})+\underline{g}_2\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{a}}(t)\boldsymbol{z}_2^{\mathrm{T}})+\\ &\underline{g}_2\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{a}}(t)[\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{c}}(\boldsymbol{x}_{\mathrm{c}})]^{\mathrm{T}}),\end{aligned}\tag{38}$$

其中$\varGamma_2=\varDelta_2+\dfrac{\ell_{\hat{b}_2 2}}{2}\|b_2\|^2$. 进一步, 结合不等式

$$\begin{aligned}&\ell_{\mathrm{a}}\underline{g}_2\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\hat{\boldsymbol{W}}_{\mathrm{a}})\leqslant-\frac{\ell_{\mathrm{a}}\underline{g}_2}{2}\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{a}})+\frac{\ell_{\mathrm{a}}\underline{g}_2 b_{W_{\mathrm{a}}}^2}{2},\\ &-\boldsymbol{z}_2^{\mathrm{T}}\underline{g}_2 k_{21}\boldsymbol{z}_2+\boldsymbol{z}_2\underline{g}_2\boldsymbol{\varepsilon}_{\mathrm{a}}(\boldsymbol{x}_{\mathrm{a}})\leqslant\\ &-\underline{g}_2\lambda_{\min}(\boldsymbol{k}_{21})\|\boldsymbol{z}_2\|^2+\underline{g}_2 b_{\varepsilon_{\mathrm{a}}}\|\boldsymbol{z}_2\|\leqslant\frac{\underline{g}_2 b_{\varepsilon_{\mathrm{a}}}^2}{4\lambda_{\min}(\boldsymbol{k}_{21})},\\ &-l_{\mathrm{a}}\|\tilde{\boldsymbol{W}}_{\mathrm{a}}\|_{\mathrm{F}}^2+b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}b_{W_{\mathrm{c}}}\|\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\|_{\mathrm{F}}\leqslant\frac{b_{\phi_{\mathrm{a}}}^2 b_{\phi_{\mathrm{c}}}^2 b_{W_{\mathrm{c}}}^2}{4l_{\mathrm{a}}},\\ &b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}\|\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\|_{\mathrm{F}}\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}\leqslant\\ &\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}\|\tilde{\boldsymbol{W}}_{\mathrm{a}}\|_{\mathrm{F}}^2+\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}^2\|\hat{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}\leqslant\\ &\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}+b_{W_{\mathrm{c}}},\end{aligned}\tag{39}$$

进一步, 可得

$$\begin{aligned}&\dot{V}_{21}\leqslant\\ &-\underline{g}_1\lambda_{\min}(\boldsymbol{k}_1)\|\boldsymbol{z}_1\|^2-\underline{g}_2\lambda_{\min}(\boldsymbol{k}_{22})\|\boldsymbol{z}_2\|^2-\\ &\underline{g}_2(\frac{\ell_{\mathrm{a}}}{2}-\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}-l_{\mathrm{a}})\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{a}})-\bar{\varepsilon}_1\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1+\\ &\underline{g}_2\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{c}})-\sum_{i=1}^{2}\frac{\underline{g}_i\ell_{\hat{b}2}}{2}\|\tilde{b}_i\|^2+b_{V_1},\end{aligned}\tag{40}$$

其中$b_{V_1}=\dfrac{\ell_{\mathrm{a}}\underline{g}_2 b_{W_{\mathrm{a}}}^2}{2}+\dfrac{\underline{g}_2 b_{\phi_{\mathrm{a}}}^2 b_{\phi_{\mathrm{c}}}^2 b_{W_{\mathrm{c}}}^2}{4l_{\mathrm{a}}}+\sum_{k=1}^{2}\varGamma_k+\hbar^{\mathrm{T}}\hbar$.

进一步, 将V_{22}关于时间求导数, 可得

$$\dot{V}_{22}\leqslant-\underline{g}_2 l_{\mathrm{v}}\sigma_{\mathrm{c}}\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{c}})+b_{V_2}(\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}),\tag{41}$$

其中$b_{V_2}=\underline{g}_2(1+\gamma)l_V b_{\phi_{\mathrm{c}}}b_{p_{\mathrm{c}}}+\underline{g}_2(1+\gamma)^2 l_V b_{\phi_{\mathrm{c}}}^2 b_{W_{\mathrm{c}}}$.

进一步, 可得

$$\begin{aligned}&\dot{V}_2=\dot{V}_{21}+\dot{V}_{22}\leqslant\\ &-\underline{g}_1\lambda_{\min}(\boldsymbol{k}_1)\|\boldsymbol{z}_1\|^2-\underline{g}_2\lambda_{\min}(\boldsymbol{k}_{22})\|\boldsymbol{z}_2\|^2-\\ &\underline{g}_2(\frac{\ell_{\mathrm{a}}}{2}-\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}-l_{\mathrm{a}})\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{a}})-\bar{\varepsilon}_1\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1+b_{V_1}-\\ &\underline{g}_2(l_{\mathrm{v}}\sigma_{\mathrm{c}}-\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}-l_{\mathrm{c}})\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{c}})+b_{V_2}\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}-\\ &l_{\mathrm{c}}\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{c}})-\sum_{i=1}^{2}\frac{\underline{g}_i\ell_{\hat{b}i}}{2}\|\tilde{b}_i\|^2,\end{aligned}\tag{42}$$

结合不等式$-l_{\mathrm{c}}\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}^2+b_{V_2}\|\tilde{\boldsymbol{W}}_{\mathrm{c}}\|_{\mathrm{F}}\leqslant\dfrac{b_{V_2}^2}{2l_{\mathrm{c}}}$, 进一步, 可得

$$\dot{V}_2\leqslant-\Re V_2+\varTheta,\tag{43}$$

其中:

$$\begin{aligned}\Re=\min\{&\underline{g}_1\lambda_{\min}(\boldsymbol{k}_1),\ \underline{g}_2\lambda_{\min}(\boldsymbol{k}_{22}),\\ &\frac{\sigma_{\mathrm{a}}-b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}-2l_{\mathrm{a}}}{\lambda_{\min}(\boldsymbol{\varGamma}_{\mathrm{a}}^{-1})},\ell_{\hat{b}_1},\ell_{\hat{b}_2},\bar{\varepsilon}_2,\\ &\frac{2l_{\mathrm{v}}\sigma_{\mathrm{c}}-b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}-2l_{\mathrm{c}}}{l_{\mathrm{v}}\boldsymbol{\lambda}_{\min}(\boldsymbol{\varGamma}_{\mathrm{c}}^{-1})}\},\\ \varTheta=b_{V_1}+&\frac{b_{V_2}^2}{2l_{\mathrm{c}}},\end{aligned}$$

同时对不等式(43)两侧乘以$\mathrm{e}^{\Re t}$, 并定义$\varUpsilon=\varTheta/\Re$, 在区间$[0,t]$上积分式(43)可得

$$V_2(t)\leqslant\varUpsilon+V_2(0)\mathrm{e}^{-\Re t},\tag{44}$$

其中$V_2(0)=\dfrac{1}{2}\sum_{i=1}^{2}[\boldsymbol{z}_i^{\mathrm{T}}(0)\boldsymbol{z}_i(0)+\dfrac{\underline{g}_i}{2\gamma_i}\tilde{b}_i^2(0)]+\dfrac{1}{2}\times\boldsymbol{y}_2^{\mathrm{T}}(0)\boldsymbol{y}_2(0)+\dfrac{\underline{g}_2}{2}\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}(0)\boldsymbol{\varGamma}_{\mathrm{a}}^{-1}\tilde{\boldsymbol{W}}_{\mathrm{a}}(0))+\dfrac{\underline{g}_2 l_{\mathrm{v}}}{2}\times\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}(0)\boldsymbol{\varGamma}_{\mathrm{c}}^{-1}\tilde{\boldsymbol{W}}_{\mathrm{c}}(0))$. 此时, 参数$\ell_{\mathrm{a}},l_{\mathrm{v}}\sigma_{\mathrm{c}}$的选取满足$\dfrac{\ell_{\mathrm{a}}}{2}-\dfrac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}-l_{\mathrm{a}}>0, l_{\mathrm{v}}\sigma_{\mathrm{c}}-\dfrac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}-l_{\mathrm{c}}>0$, 可得$\boldsymbol{z}_1,\boldsymbol{z}_2,\tilde{b}_1,\tilde{b}_2,\boldsymbol{y}_1,\tilde{\boldsymbol{W}}_{\mathrm{a}},\tilde{\boldsymbol{W}}_{\mathrm{c}}$都是有界的, 完成定理1中第1部分的证明.

此时, 注意到在紧集$\mathcal{D}_{\mathrm{d}}=\{\boldsymbol{y}_{\mathrm{d}}\in\mathbb{R}^3:-\boldsymbol{A}_i 0\leqslant\boldsymbol{y}_{i\mathrm{d}}\leqslant\boldsymbol{A}_i 0, i=1,2,3\}$中, $\alpha_0\in L_\infty$, 结合$\boldsymbol{z}_1=\boldsymbol{\zeta}_1-\boldsymbol{\alpha}_0$, 可知$\boldsymbol{\zeta}_1$为有界的. 由于有界的

$$\frac{(F_{i1}F_{i2}+x_{1,i}^2)}{[(F_{i1}+x_{1,i})(F_{i2}-x_{1,i})]^2},$$

可得$\varPi_1\in L_\infty$, 进一步结合式(16)可得$\boldsymbol{\alpha}_1\in L_\infty,\dot{\hat{b}}_1\in L_\infty$. 随后, 结合式(9), 可得$\boldsymbol{\alpha}_{2\mathrm{f}}$为有界量. 同理, 可分析$\boldsymbol{\zeta}_2,\boldsymbol{\alpha}_2,\dot{\hat{b}}_2$为有界量, 以及系统的控制输入$\boldsymbol{v}$也是有界量. 同时, 根据$b_1,b_2$的有界定义, 可得$\hat{b}_1,\hat{b}_2$也是有界量. 再次基于$\boldsymbol{z}_1=\boldsymbol{\zeta}_1-\boldsymbol{\alpha}(0)$, 结合式(7), 可得

$$\boldsymbol{\delta}_1\boldsymbol{z}_1=\boldsymbol{\delta}_2\boldsymbol{e},\tag{45}$$

其中: 当$-F_{i1}<x_{1i}<F_{i2}$时, 可得$\boldsymbol{\delta}_1=[\delta_{1,1}\ \delta_{1,2}\ \delta_{1,3}]^{\mathrm{T}}$, 具体为$\delta_{1i}=(F_{i1}-x_{1i})(F_{i2}+x_{1i})(F_{i1}-y_{\mathrm{d}i})(F_{i2}+y_{\mathrm{d}i})>0$, 同时, $\boldsymbol{\delta}_2=[\delta_{2,1}\ \delta_{2,2}\ \delta_{2,3}]^{\mathrm{T}}$, 具体是$\delta_{2i}=F_{i1}F_{i2}+x_{1i}y_{\mathrm{d}i}>0$, 此时, 跟踪误差满足$\boldsymbol{e}=\boldsymbol{x}_1-\boldsymbol{y}_{\mathrm{d}}$. 进一步

$$\boldsymbol{e}=\boldsymbol{\delta}\boldsymbol{z}_1,\tag{46}$$

其中$\boldsymbol{\delta}=\mathrm{diag}\{\frac{\delta_{11}}{\delta_{21}},\frac{\delta_{12}}{\delta_{22}},\frac{\delta_{13}}{\delta_{23}}\}$, 由于$\boldsymbol{z}_1$是有界的, 则$\boldsymbol{e}\in L_\infty$. 至此, 所有的闭环系统的信号都是有界的. 进一步, 基于上述有界性的分析, 可知, 存在正常数M使得$-\sum\limits_{i=1}^{2}\frac{\underline{g}_i\ell_{\hat{b}2}}{2}\|\tilde{b}_i\|^2-\underline{g}_2(\frac{\ell_{\mathrm{a}}}{2}-\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}-l_{\mathrm{a}})\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\times\tilde{\boldsymbol{W}}_{\mathrm{a}})-\underline{g}_2(l_{\mathrm{v}}\sigma_{\mathrm{c}}-\frac{b_{\phi_{\mathrm{a}}}b_{\phi_{\mathrm{c}}}}{2}-l_{\mathrm{c}})\mathrm{tr}(\tilde{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\mathrm{c}})-\bar{\varepsilon}_1\boldsymbol{y}_1^{\mathrm{T}}\boldsymbol{y}_1+\Theta\leqslant M$. 在此基础上, 考虑Lyapunov函数$V_3=\frac{\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1}{2}+\frac{\boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2}{2}$, 并对其求关于时间的导数, 可得

$$\dot{V}_3\leqslant -aV_3^{\frac{n+m}{2n}}-bV_3^{\frac{r+p}{2r}}+M,\tag{47}$$

其中: $a=2^{1-\frac{n+m}{2n}}2^{\frac{n+m}{2n}}a_1, a_1=\min\{\underline{g}_1l_1,\underline{g}_2l_3\}$, $\frac{m}{n}=\min\{\frac{m_1}{n_1},\frac{m_2}{n_2}\}$, $b=2^{\frac{r+p}{2r}}b_1$, $b_1=\min\{\underline{g}_1l_2,\underline{g}_2l_5\}$, $\frac{p}{r}=\max\{\frac{p_1}{r_1},\frac{p_2}{r_2}\}$. 同时, $\frac{m+n}{2n}>1$, $0<\frac{p+r}{2r}<1$.

结合文献[15]中的引理2, 系统固定时间的收敛域Σ可表示为$\Sigma=\{\lim\limits_{t\to t_{\mathrm{s}1}}z_1|V_3\leqslant\min\{(\frac{M}{(1-\theta)a})^{\frac{2n}{n+m}},(\frac{M}{(1-\theta)b})^{\frac{2r}{r+p}}\}\}$. 系统收敛的固定时间上界为$t_{\mathrm{s}1}\leqslant\frac{1}{a}\frac{1}{\frac{m+n}{2n}-1}+\frac{1}{b}\frac{1}{1-\frac{p+r}{2r}}\leqslant\frac{1}{a}\frac{2n}{m-n}+\frac{1}{b}\frac{2r}{r-p}$. 此时, 一旦系统的状态$\boldsymbol{z}_1$收敛至$\Sigma$内, 在固定时间上界$t_{\mathrm{s}1}$, 系统状态的收敛域可为

$$|z_{1,j}|\leqslant\min\{(\frac{M}{(1-\theta)a})^{\frac{n}{n+m}},(\frac{M}{(1-\theta)b})^{\frac{r}{r+p}}\},$$

至此, 系统的稳定性证明结束. 证毕.

5 结果与讨论

5.1 仿真初始条件及控制器参数

本节在MATLAB/Simulink环境下进行数值仿真验证, 控制目标是考虑外部干扰和执行机构死区特性的条件下, 使航天器编队在100 s内达到期望构型, 并满足0.01 m的相对位置保持精度和防避撞要求(两颗星相对距离不小于1.5 m), 在系统的初始参数如表1所示. 本文的仿真环境是Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz, 8 G内存.

表 1 系统仿真参数

Table 1 System parameters for simulation

序号	系统参数	值
1	m_{l}	59 kg
2	m_{f}	59 kg
3	R_{l}	7578.17 km
4	ω_0	9.57×10^{-4} rad/s
5	$\boldsymbol{R}_0(0)$	$[6\ \ 7\ \ 5]^{\mathrm{T}}$ m
6	$\boldsymbol{v}_0(0)$	$[0\ \ 0\ \ 0]^{\mathrm{T}}$ m/s

在数学仿真中, 假设两颗星的期望相对位置为$[3\ \ 0\ \ 0]^{\mathrm{T}}$ m, 期望相对速度为$[0\ \ 0\ \ 0]^{\mathrm{T}}$(m/s). 在实际应用中, 本文考虑超紧密航天器在近地轨道附近受到的气动阻力、太阳光压等干扰力幅值不超过10^{-3} N, 故而, 本文假设系统干扰力上限为1.5×10^{-3} N. 电推进器死区特性的参数为$h_{\mathrm{r}}=1, h_{\mathrm{l}}=1, o_{\mathrm{r}}=0.3, o_{\mathrm{l}}=-0.3$.

考虑到系统的状态约束问题, 本文设计的状态约束函数的参数为$F_{11}=2, F_{21}=F_{31}=-4, F_{12}=F_{22}=F_{32}=15$. 另外, 采用的电推进器死区特性的参数为$h_{\mathrm{r}}=1, h_{\mathrm{l}}=1, o_{\mathrm{r}}=0.3, o_{\mathrm{l}}=-0.3$.

根据系统的控制需求, 选取位置控制回路的参数为$\boldsymbol{k}_1=[0.05\ \ 0.05\ \ 0.05]^{\mathrm{T}}, l_1=l_2=0.01, m_1=9, n_1=5, p_1=5, r_1=7, \ell_{\hat{b}1}=\ell_{\hat{b}2}=1$. 速度控制回路参数为$\boldsymbol{k}_2=[15\ \ 15\ \ 15]^{\mathrm{T}}, l_3=0.1, l_4=0.01, m_2=11, n_2=9, p_2=5, r_2=7$.

强化学习评判网络的初始参数为

$$\Gamma_{\mathrm{c}}=0.1,\ \sigma_{\mathrm{c}}=0.2,$$

$$c_{\mathrm{critic}}=\begin{bmatrix}-1.5&-1&-0.5&0&0.5&1&1.5\\-1.5&-1&-0.5&0&0.5&1&1.5\\-1.5&-1&-0.5&0&0.5&1&1.5\end{bmatrix},$$

$$b_{\mathrm{critic}}=2,$$

此外, 强化学习动作网络的初始参数为

$$\Gamma_{\mathrm{a}}=0.01,\ \sigma_{\mathrm{a}}=0.1, c_{\mathrm{actor}}=c_{\mathrm{critic}},\ b_{\mathrm{actor}}=2.$$

其余参数为$\boldsymbol{R}_{\mathrm{s}}=\boldsymbol{I}_3, \boldsymbol{Q}_{\mathrm{s}}=\boldsymbol{I}_7$. $\boldsymbol{I}_n$为n维的单位矩阵. $\varepsilon=0.1, \wp=10$. 仿真时间为400 s, 仿真结果如图3–10所示.

图3与图4分别为超紧密编队航天器的三维相对位置曲线和相对位置随时间变化曲线, 图5为编队航天器的相对速度随时间变化的曲线. 从图3–5可以看出, 两颗星在90 s达到期望的相对位置, 满足固定时间的控制要求, 相对位置保持精度达到0.031 m.

图6为强化学习奖励函数变化曲线, 图7和图8分别为强化学习评判网络中的函数$\hat{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{c}}$和强化学习动作网络的补偿量$\hat{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\Delta\boldsymbol{\phi}_{\mathrm{a}}$变化曲线. 从图中可以看出在航天器编队相对位置重构的过程中, 在20 s左右, 超紧密编队行为出现的状态越界的行为, 此时, 提出的基于强化学习的软约束机制开始工作: 由图6可以看出, 此时出现相对位置轴方向的越界行为, 致使轴方向的强化学习奖励函数值急剧增大, 与目标奖励函数越小越好($p_{\mathrm{c}1}$越小越好)的相背离. 进而随着$p_{\mathrm{c}1}$的增大, 强化学习评判网络中的$(\hat{\boldsymbol{W}}_{\mathrm{c}}^{\mathrm{T}}\Delta\phi_{\mathrm{c}})_1$(见于图7)也迅速增大, 进一步引发图6中强化学习动作网络的在x轴方向的补偿量$(\hat{\boldsymbol{W}}_{\mathrm{a}}^{\mathrm{T}}\Delta\phi_{\mathrm{a}})_1$也迅速增大, 形成负反馈调节机制, 调节超紧密编队航天器x轴方向的运动位置ρ_1(见图4), 重新回归到安全的避碰约束范围(1.5 m)外以满足超紧密编队的防避撞的控制需求.

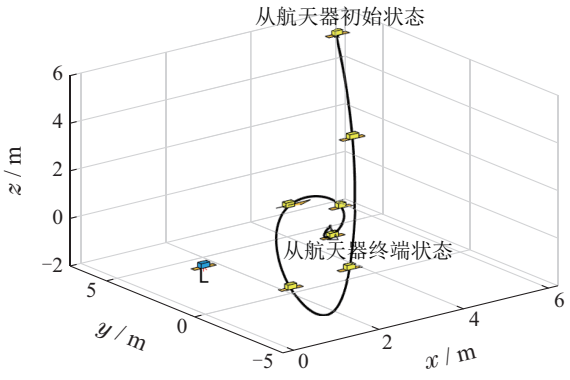

图 3 航天器超紧密编队三维运行轨迹

Fig. 3 Moving trajectories of the spacecraft in ultra close formation in view of 3-*D*

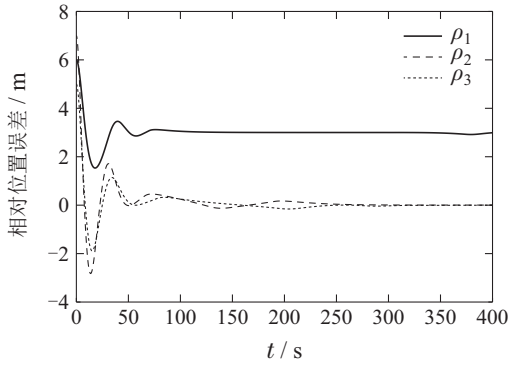

图 4 航天器超紧密编队的相对位置误差

Fig. 4 Trajectories of relative position error

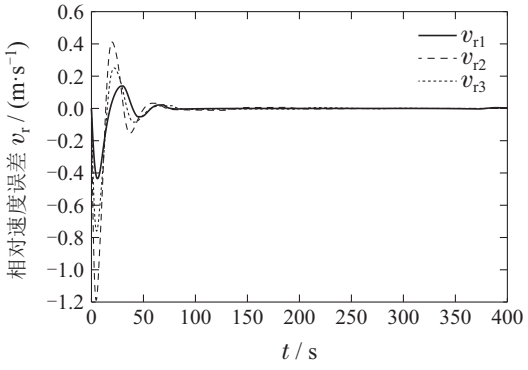

图 5 航天器超紧密编队相对速度误差

Fig. 5 Trajectories of relative velocity error

为进一步说明所提出的算法能减小执行机构存在的死区带来的精度影响, 本节在控制器设计中, 不考虑基于强化学习的死区补偿机制, 在同样的仿真参数下, 仿真结果如图9–10所示. 图9与图10分别为超紧密编队航天器的三维相对位置和相对距离随时间变化的曲线. 由于减小外部扰动的影响, 推力器在构型保持阶段一直在工作, 以保持超紧密编队的高精度控制行为. 由于执行机构死区效应的影响, 致使执行机构无法实时响应超紧密编队高精度的控制需求, 如图10所示, 如不进行基于强化学习的实施补偿, 航天器相对位置不能满足保持0.01 m的相对位置精度的控制需求, 最终只能达到0.3 m左右的位置保持精度, 对比图4, 可以看出, 本文所提出的基于强化学习的死区补偿机制能够处理死区效应带来的控制精度影响, 可以使航天器编队的相对位置保持精度达到0.01 m. 仿真结果表明本文所提出控制方法的有效性.

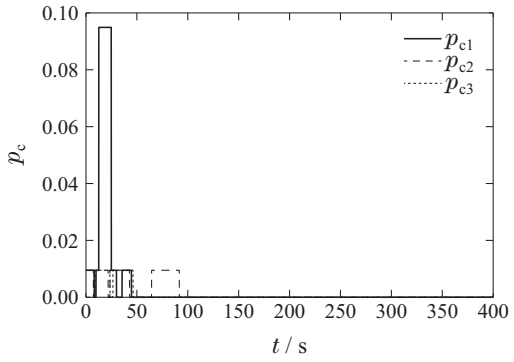

图 6 强化学习奖励函数曲线

Fig. 6 Reinforcement learning reward function curves

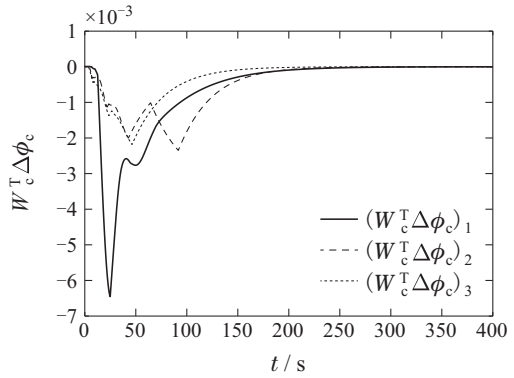

图 7 强化学习评判网络输出曲线

Fig. 7 Reinforcement learning critic network function

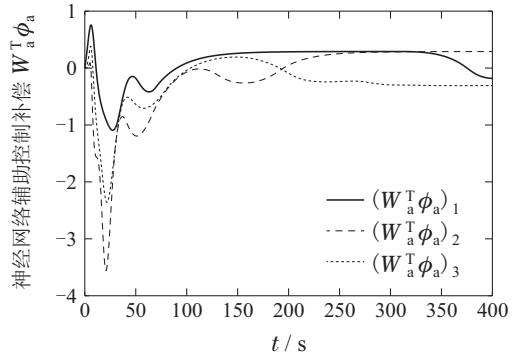

图 8 强化学习动作网络对死区效应的补偿量

Fig. 8 Dead-zone compensation output via action network of reinforcement learning

6 结论

本文对近地轨道超紧密航天器的编队重构问题进行研究, 考虑到电推进器的死区特性、系统受到的外界环境干扰以及避免碰撞的安全环约束需求, 设计了一种基于强化学习的电推力固定时间超紧密编队控制器设计方法: 1) 将超紧密编队构型保持转化为依赖

于系统状态的状态约束控制, 同时, 考虑到紧密编队建立构型的时间约束, 引入固定时间的控制设计, 保障系统在外部扰动作用下, 能够在有限时间内到达满足编队构型约束的状态约束控制目标; 2) 考虑到电推进器死区效应的影响, 本文引入强化学习评判–动作机制, 通过最小化评判网络输出的方式, 辅助以动作网络的辅助补偿, 来解决死区特性带来的影响. 与此同时, 在该评判–动作控制架构上设计状态越界软约束拉回机制, 解决因外部扰动引起的航天器碰撞与紧密编队构型破坏问题, 提高航天器紧密编队的鲁棒性; 3) 在后续的研究中, 将在此基础上, 研究多航天器超紧密编队的分布式协同优化控制问题, 设计满足全局优化的协同编队控制算法.

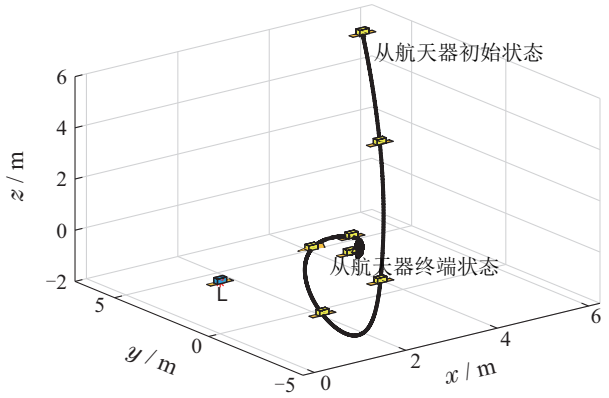

图 9 忽略无死区效应的航天器超紧密编队三维运行轨迹

Fig. 9 Moving trajectories of the spacecraft formation without dead-zone compensation in view of 3-D

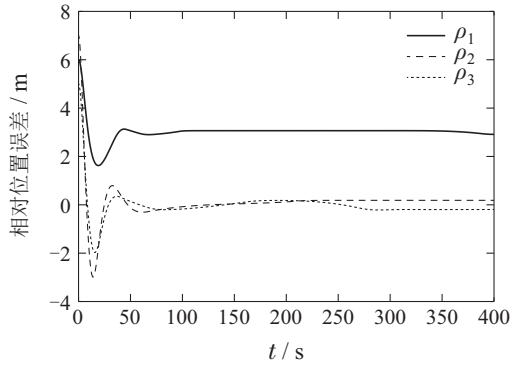

图 10 忽略无死区效应的航天器超紧密编队相对位置误差

Fig. 10 Trajectories of relative position without function of dead-zone compensation

参考文献:

[1] HUANG Jing, SUN Lujun, SUN Jun, et al. Cooperative control for collision avoidance of ultra-close spacecraft formation based on prescribed performance control. *Flight Control and Detection*, 2019, 2(3): 52 – 60.
(黄静, 孙禄君, 孙俊, 等. 基于预设性能控制的超紧密航天器编队防避撞协同控制. 飞控与探测, 2019, 2(3): 52 – 60.)

[2] MAYNE D Q, RAWLINGS J B, RAO C V, et al. Constrained model predictive control: Stability and optimality. *Automatica*, 2000, 36(6): 789 – 814.

[3] BURGER M, GUAY M. Robust constraint satisfaction for continuous time nonlinear systems in strict feedback form. *IEEE Transactions on Automatic Control*, 2010, 55(11): 2597 – 2601.

[4] NGO K B, MAHONY R, JIANG Z P. Integrator backstepping using barrier functions for systems with multiple state constraints. *Proceedings of the 44th IEEE Conference on Decision and Control*. Seville, Spain: IEEE, 2005: 8306 – 8312.

[5] HU Q, SHAO X, GUO L. Adaptive fault-tolerant attitude tracking control of spacecraft with prescribed performance. *IEEE/ASME Transactions on Mechatronics*, 2018, 23(1): 331 – 341.

[6] ZHAO K, SONG Y D. Removing the feasibility conditions imposed on tracking control designs for state-constrained strict-feedback systems. *IEEE Transactions on Automatic Control*, 2019, 64(3): 1265 – 1272.

[7] AN L, YANG G H. Collisions-free distributed optimal coordination for multiple Euler-Lagrangian systems. *IEEE Transactions on Automatic Control*, 2021, 67(1): 460 – 467.

[8] HONG Huifen. *Research on finite-time consensus of multi-agent systems with limited communication and its related problems.* Nanjing: Southeast University, 2019.
(洪会粉. 通信受限的多智能体系统有限时间一致性及其相关问题研究. 南京: 东南大学, 2019.)

[9] HAN Luofeng, ZHU Kangwu, HUANG Wenbin, et al. Development and on orbit verification of M2 microwave ion propulsion system. *Vacuum and Cryogenics*, 2022, 28(1): 98 – 105.
(韩罗峰, 朱康武, 黄文斌, 等. M2型微波离子推进系统研制及在轨验证. 真空与低温, 2022, 28(1): 98 – 105.)

[10] LV Y, HU Q, MA G, et al. 6 DOF synchronized control for spacecraft formation flying with input constraint and parameter uncertainties. *ISA Transactions*, 2011, 50(4): 573 – 580.

[11] ZHANG D, KONG L, ZHANG S, et al. Neural networks-based fixed-time control for a robot with uncertainties and input deadzone. *Neurocomputing*, 2020, 390: 139 – 147.

[12] HE W, HUANG B, DONG Y T, et al. Adaptive neural network control for robotic manipulators with unknown deadzone. *IEEE Transactions on Cybernetics*, 2018, 48(9): 2670 – 2682.

[13] GUO X, YAN W, CUI R. Integral reinforcement learning-based adaptive NN control for continuous-time nonlinear MIMO systems with unknown control directions. *IEEE Transactions on Systems, Man, and Cybernetics*, 2020, 50(11): 4068 – 4077.

[14] LI H, WU Y, CHEN M. Adaptive fault-tolerant tracking control for discrete-time multiagent systems via reinforcement learning algorithm. *IEEE Transactions on Cybernetics*, 2021, 51(3): 1163 – 1174.

[15] CHEN Q, XIE S, SUN M, et al. Adaptive nonsingular fixed-time attitude stabilization of uncertain spacecraft. *IEEE Transactions on Aerospace and Electronic Systems*, 2018, 54(6): 2937 – 2950.

作者简介:

孟亦真 中级工程师, 目前研究方向为航天器集群智能避障与追逃博弈控制, E-mail: 15151830168@163.com;

黄 静 高级工程师, 目前研究方向为航天器振动抑制与容错控制, E-mail: huangjing04415@163.com;

周绍辉 高级工程师, 目前研究方向为航天器电离子推进器控制, E-mail: zsh@sastspace.com;

周 彬 中级工程师, 目前研究方向为航天器姿态最优控制, E-mail: 732200801@qq.com;

朱康武 副研究员, 目前研究方向为航天器电离子推进器结构设计与智能控制, E-mail: zjuzkw@zju.edu.cn.