自动化学报
ACTA AUTOMATICA SINICA

含动力学奖励的航天器编队深度强化学习控制

金伟成 陈提 胡海岩

Deep Reinforcement Learning Control for Spacecraft Formation With Dynamical Reward

JIN Wei-Cheng, CHEN Ti, HU Hai-Yan

在线阅读 View online: https://doi.org/10.16383/j.aas.c250202

您可能感兴趣的其他文章

含动力学奖励的航天器编队深度强化学习控制

金伟成 [1, 2] 陈 提 [1, 2] 胡海岩 [1, 2]

摘 要 提出一种航天器编队的深度强化学习控制方法. 该方法通过引入动力学奖励, 考虑轨迹的动力学可行性并优化燃料消耗量. 在训练环境中, 引入 J_2 摄动相对动力学模型, 基于近端策略优化算法, 将航天器的局部观测信息作为策略网络和评价网络的输入. 策略网络输出航天器的期望位置和速度, 结合动力学模型限制策略任意动作之间的转换控制, 使输出轨迹考虑动力学可行性. 评价网络基于局部观测信息估计由动力学模型限制的优势函数, 从而辅助策略网络更新参数. 进一步地, 以燃料消耗量的负数作为动力学奖励, 结合避撞和任务相关奖励后, 训练得到的策略网络在完成航天器编队任务的同时优化了燃料消耗.

关键词 深度强化学习, 航天器编队, 分布式控制, 动力学

引用格式 金伟成, 陈提, 胡海岩. 含动力学奖励的航天器编队深度强化学习控制. 自动化学报, 2025, **51**(10): 2283−2292

DOI 10.16383/j.aas.c250202 **CSTR** 32138.14.j.aas.c250202

Deep Reinforcement Learning Control for Spacecraft Formation With Dynamical Reward

JIN Wei-Cheng[1, 2] CHEN Ti[1, 2] HU Hai-Yan[1, 2]

Abstract This paper presents a deep reinforcement learning control method for spacecraft formation. The method deals with the dynamical feasibility of the trajectory and optimizes the fuel consumption by introducing dynamical reward. Based on proximal policy optimization algorithm, a dynamic model of relative motion with J_2 perturbation is introduced in the training environment, and the inputs of Actor and Critic networks are the local observed information of the spacecraft. The outputs of the Actor network are the desired position and velocity of the spacecraft. Combining the dynamic model that restricts the control of transitions between two arbitrary actions of the strategy, the Actor network outputs the desired position and velocity, which makes the output trajectory account for the dynamical feasibility. The Critic network estimates the advantage function constrained by the dynamic model based on local observed information, therefore, the Actor network updates the parameters based on the advantage function. Further, the dynamical reward is defined as the negative value of the fuel consumption. As a result, combining collision avoidance and task-related rewards, the obtained Actor network achieves the distributed spacecraft formation task while optimizing the fuel consumption.

Key words Deep reinforcement learning, spacecraft formation, distributed control, dynamics

Citation Jin Wei-Cheng, Chen Ti, Hu Hai-Yan. Deep reinforcement learning control for spacecraft formation with dynamical reward. *Acta Automatica Sinica*, 2025, **51**(10): 2283−2292

与复杂的大型航天器相比, 小型航天器集群自主编队在分布式观测、空间大型结构组装、空间碎片抓捕等领域具有成本低、可靠性高和适应性强等优势[1−3]. 自 1978 年美国国家航空航天局 (National Aeronautics and Space Administration, NASA) 提出模块化航天器概念以来[4], 航天器集群自主编队任务在现代航天中展现出巨大的应用潜力, 相关的在轨任务和理论研究也在不断发展[5−7].

例如, 美国空军研究实验室 (Air Force Research Laboratory, AFRL) 提出 TechSat 21 项目, 用 8 个小型航天器来执行分布式观测任务[8]; 多伦多大学航空航天研究所实现了 CanX-4 与 CanX-5 双航天器编队, 在 50 m 距离下实现厘米级相对位置的高精度编队飞行[9]; 我国则成功将实践九号 A/B 双卫星送入预定轨道, 进行卫星编队技术验证[10].

相关理论研究多集中于复杂场景或复杂动力学下的运动规划和控制器设计. Foust 等[11] 基于模型预测控制[12] 和凸序列优化方法[13] 提出一种用分布式集群航天器组装大型空间结构的算法; Basu 等[14] 提出一种燃料消耗量最优的航天器位置分配算法,

收稿日期 2025-05-07 录用日期 2025-08-28
Manuscript received May 7, 2025; accepted August 28, 2025
国家重点研发计划 (2022YFC2204800), 国家自然科学基金 (12494562, 12472015) 资助
Supported by National Key Research and Development Program of China (2022YFC2204800) and National Natural Science Foundation of China (12494562, 12472015)
本文责任编委 王金枝
Recommended by Associate Editor WANG Jin-Zhi
1. 南京航空航天大学航空学院 南京 210016 2. 航空航天结构力学及控制全国重点实验室 南京 210016
1. College of Aerospace Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016 2. State Key Laboratory of Mechanics and Control for Aerospace Structures, Nanjing 210016

并基于收缩时域模型预测控制算法实现轨迹跟踪. 目前, 有不少学者通过多智能体强化学习算法求解协同包围等场景[15], 并获得了很好效果. 然而, 不同场景的奖励设置、强化学习与复杂动力学模型的结合等问题尚待研究. 面对日益复杂的场景和通用性要求, 可基于多智能体强化学习算法在协同包围等问题中的求解思路, 探索航天器集群自主编队的新途径[16−17].

对多智能体强化学习方法的研究可分为两类: 一是从控制和优化理论出发, 设计并求解优化控制问题; 二是从人工智能角度出发, 学习复杂环境下的任务决策序列. 第一类研究大多针对有领导者的一致性问题, 将系统博弈问题转化为哈密顿−雅可比−埃萨克斯 (Hamilton-Jacobi-Isaacs, HJI) 方程[18]. 例如, Vamvoudakis 等[19] 运用策略迭代方法在线求解双人零和博弈问题的 HJI 方程, 输出的反馈控制策略可保证系统的闭环稳定性. 在第二类研究中, 采用深度神经网络与环境交互来更新网络参数, 可处理复杂环境的决策问题, 但会牺牲可解释性. 比如, Yan 等[20] 基于 Actor-Critic 模型训练一种固定翼无人机编队控制策略, 其算法中的 Actor 和 Critic 网络采用多层感知机模型并建立在连续的状态空间和动作空间中; 他们还采用经验回放机制来提高训练效率. 第二类研究的一个重要区分点是, 策略的训练与执行是集中式还是分布式. 集中式训练通常较为容易, 也能取得更好效果. 例如, Xu 等[21] 基于近端策略优化 (Proximal policy optimization, PPO) 算法和领导者−跟随者框架, 给出固定翼无人机编队的集中决策式避撞编队保持方法, 在单个领导者和五个跟随者的情况下获得较低的通信成本. 但集中式训练往往依赖全局信息, 难以扩展到大规模集群, 而分布式训练可以解决这一问题. 例如, Chen 等[22] 通过深度强化学习, 分布式地学习了无通信多机器人场景下的避撞策略, 其方法的核心是价值网络对给定智能体与其邻居状态下到达目标的时间进行估计. 分布式架构的主要问题是环境非平稳性, 在训练过程中, 智能体不断更新的策略也是环境的一部分, 这会对学习产生负面影响. 除此之外, 多智能体强化学习还可根据其迭代优化的目标分为基于值 (Value-based) 的和基于策略 (Policy-based) 的多智能体强化学习, 以及两者结合的 Actor-Critic 方法[23]. 然而, 在考虑航天器复杂轨道动力学的前提下, 需解决的难题是: 如何仅依据局部观测信息独立决策, 实现不同规模航天器集群的自主编队控制.

本文针对航天器集群自主编队任务, 提出一种含动力学奖励的深度强化学习方法. 该方法的主要创新点如下: 一是得到的策略网络以局部信息作为输入并独立决策, 可适应不同规模的航天器集群和期望构型, 而无需重新训练; 二是策略网络输出航天器的期望位置和速度, 考虑了在 J_2 摄动相对动力学模型下轨迹的动力学可行性, 并优化了燃料消耗量.

本文的结构如下: 第 1 节阐述航天器自主编队任务、航天器的 J_2 摄动相对动力学模型和多智能体强化学习的基本知识; 第 2 节介绍本文提出的航天器自主编队深度强化学习框架; 第 3 节给出训练场景和结果, 并对比不同动力学奖励占比对训练后策略的影响; 第 4 节总结全文.

1 问题描述

本节主要介绍航天器自主编队任务场景和控制目标、航天器的 J_2 摄动相对动力学模型以及多智能体强化学习的基本知识.

1.1 任务场景和控制目标

假设 n 个同构的航天器已由运载火箭发射进入任务轨道, 并随机释放在任务轨道附近, 要求各航天器独立决策形成预定的编队. 一般来说, 针对不同的具体任务, 预定的期望编队有两种类型: 第一种给定精确的状态坐标; 第二种给定若干期望区域. 在航天器集群追逃、博弈、对抗、结构组装和重构等空间任务中, 多采用第二种预定期望编队, 尤其是在轨道平面内的期望区域编队. 因此, 本文主要研究平面内期望区域的航天器自主编队任务. 例如, 图 1 给出期望 6 个航天器在“十”字形平面区域内自主形成编队的场景.

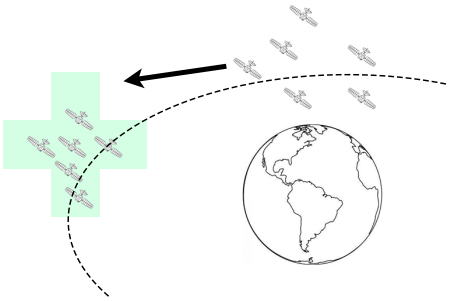


图 1 航天器自主编队示意图

Fig. 1 Schematic diagram of spacecraft autonomous formation

在本文中, 设航天器只有有限的计算和通讯能力, 且其所携带的低功耗传感器仅能探测周围邻域 $r_{\rm obs}$ 内的信息. 此外, 航天器集群的规模和期望区域的形状会根据具体任务需求改变. 在执行任务前, 期望区域信息被预先发送给所有航天器, 并假设期

望区域形状在执行任务期间保持不变. 因此, 本文的控制目标是: 各航天器通过期望区域信息和实时探测到的局部信息, 在不发生碰撞的情况下被控制到期望区域内. 此外, 需要考虑航天器的 J_2 摄动相对动力学模型, 生成的轨迹要符合动力学可行性和较少的燃料消耗, 并要适应不同规模的集群和期望区域.

1.2　J_2 摄动相对动力学模型

空间航天器在进行近距离机动时, 主要受到 J_2 摄动力的影响[14, 24]. 首先, 定义如下两个坐标系: 一是地心惯性系 (Earth central inertial coordinate frame, ECI) $\hat{O}$-$\hat{X}\hat{Y}\hat{Z}$, 其坐标原点 $\hat{O}$ 为地球质心, $\hat{X}$ 轴为地球赤道平面与黄道平面的交线, 指向春分点, $\hat{Z}$ 轴指向北极, $\hat{Y}$ 轴与 $\hat{X}$ 轴和 $\hat{Z}$ 轴构成右手系; 二是主星质心相对运动轨道坐标系 (Local-vertical-local-horizontal coordinate frame, LVLH) O-XYZ, 其坐标原点 O 是在地心惯性系中描述的主星质心或目标轨道上的虚拟主星质心, X 轴由地心指向主星质心, Y 轴在主星的运行轨道平面内且指向主星的速度方向, Z 轴与 X 轴和 Y 轴构成右手系. 集群中第 j 个航天器的相对动力学模型定义在 LVLH 内, 其表达式为

$$\begin{cases} \ddot{x}_j = 2\dot{y}_j\omega_z - x_j\left(n_j^2-\omega_z^2\right) + y_j\alpha_z - \\ \qquad z_j\omega_x\omega_z - (\zeta_j-\zeta)\,\mathrm{s}_{i_o}\mathrm{s}_\theta - r\left(n_j^2-n^2\right) + u_{jx} \\ \ddot{y}_j = -2\dot{x}_j\omega_z + 2\dot{z}_j\omega_x - y_j\left(n_j^2-\omega_z^2-\omega_x^2\right) - \\ \qquad x_j\alpha_z + z_j\alpha_x - (\zeta_j-\zeta)\,\mathrm{s}_{i_o}\mathrm{c}_\theta + u_{jy} \\ \ddot{z}_j = -2\dot{y}_j\omega_x - x_j\omega_x\omega_z - y_j\alpha_x - \\ \qquad z_j\left(n_j^2-\omega_x^2\right) - (\zeta_j-\zeta)\,\mathrm{c}_{i_o} + u_{jz} \end{cases} \tag{1}$$

其中,

$$\omega_z = \dot{\theta} + \dot{\Omega}\mathrm{c}_{i_o} = \frac{h}{r^2}$$

$$n_j^2 = \frac{\mu}{r_j^3} + \frac{k_{J2}}{r_j^5} - \frac{5k_{J2}r_{jZ}^2}{r_j^7}$$

$$r_j = \|\boldsymbol{r}_j\| = \sqrt{(r+x_j)^2 + y_j^2 + z_j^2}$$

$$\alpha_z = \dot{\omega}_z = -\frac{2h\dot{r}}{r^3} - \frac{k_{J2}\mathrm{s}_{i_o}^2\mathrm{s}_{2\theta}}{r^5}$$

$$r_{jZ} = \boldsymbol{r}_j\cdot\hat{\boldsymbol{Z}} = (r+x_j)\,\mathrm{s}_{i_o}\mathrm{s}_\theta + y_j\mathrm{s}_{i_o}\mathrm{c}_\theta + z_j\mathrm{c}_{i_o}$$

$$\omega_x = \dot{i}_o\mathrm{c}_\theta + \dot{\Omega}\mathrm{s}_\theta\mathrm{s}_{i_o} = -\frac{k_{J2}\mathrm{s}_{2i_o}\mathrm{s}_\theta}{hr^3}$$

$$\zeta_j = \frac{2k_{J2}r_{jZ}}{r_j^5},\quad \zeta = \frac{2k_{J2}\mathrm{s}_{i_o}\mathrm{s}_\theta}{r^4}$$

$$n^2 = \frac{\mu}{r^3} + \frac{k_{J2}}{r^5} - \frac{5k_{J2}\mathrm{s}_{i_o}^2\mathrm{s}_\theta^2}{r^5}$$

$$\alpha_x = \dot{\omega}_x = -\frac{k_{J2}\mathrm{s}_{2i_o}\mathrm{c}_\theta}{r^5} + \frac{3\dot{r}k_{J2}\mathrm{s}_{2i_o}\mathrm{s}_\theta}{r^4h} - \frac{8k_{J2}^2\mathrm{s}_{i_o}^3\mathrm{c}_{i_o}\mathrm{s}_\theta^2\mathrm{c}_\theta}{r^6h^2}$$

$$k_{J2} = \frac{3J_2\mu R_e^2}{2} \tag{2}$$

式 (1) 和式 (2) 中, "$\cdot$"表示向量点乘; x_j、y_j 和 z_j, u_{jx}、u_{jy} 和 u_{jz} 分别表示第 j 个航天器在 LVLH 中的位置和施加的控制加速度; c_*, s_* 分别表示 $\cos(*)$ 和 $\sin(*)$; ω_z 表示轨道平面的转动角速度在 Z 轴上的分量; n^2 和 ζ 是包含在 ECI 下描述的目标轨道根数的项; n_j^2 和 ζ_j 是包含目标轨道根数和第 j 个航天器相对位置的项; α_z 表示轨道平面的转动角加速度在 Z 轴上的分量; r_{jZ} 表示 $\boldsymbol{r}_j$ 在 ECI 的 $\hat{Z}$ 轴上的投影; ω_x 表示轨道平面的转动角速度在 X 轴上的分量; α_x 表示轨道平面的转动角加速度在 X 轴上的分量; $R_e = 6\,378\,\mathrm{km}$、$J_2 = 1.082\,62\times10^{-3}$ 和 $\mu = 3.986\times10^5\ \mathrm{km}^3\cdot\mathrm{s}^{-2}$ 分别表示地球的平均半径、扁率常数和标准重力参数. 式 (1) 和式 (2) 中目标轨道的混合轨道根数 $\boldsymbol{r}_s = \begin{bmatrix} r & v_x & h & \Omega & i_o & \theta \end{bmatrix}$ 满足下式

$$\begin{cases} \dot{r} = v_x \\ \dot{v}_x = -\dfrac{\mu}{r^2} + \dfrac{h^2}{r^3} - \dfrac{k_{J2}}{r^4}\left(1-3\mathrm{s}_{i_o}^2\mathrm{s}_\theta^2\right) \\ \dot{h} = -\dfrac{k_{J2}\mathrm{s}_{i_o}^2\mathrm{s}_{2\theta}}{r^3} \\ \dot{\Omega} = -\dfrac{2k_{J2}\mathrm{c}_{i_o}\mathrm{s}_\theta^2}{hr^3} \\ \dot{i}_o = -\dfrac{k_{J2}\mathrm{s}_{2i_o}\mathrm{s}_{2\theta}}{2hr^3} \\ \dot{\theta} = \dfrac{h}{r^2} + \dfrac{2k_{J2}\mathrm{c}_{i_o}^2\mathrm{s}_\theta^2}{hr^3} \end{cases} \tag{3}$$

由于小型航天器的推进器只能提供有限推进力, 所以本文假设式 (1) 中的控制力存在饱和, 即 $\boldsymbol{u}_j = \begin{bmatrix} u_{jx} & u_{jy} & u_{jz} \end{bmatrix}^{\mathrm{T}} = \mathrm{sat}_{\boldsymbol{u}_{\min}}^{\boldsymbol{u}_{\max}}(\boldsymbol{u}_{cj})$. 其中, $\boldsymbol{u}_{cj}$ 是第 j 个航天器计算得到的控制力; $\boldsymbol{u}_{\min} = u_{\min}\boldsymbol{1}$ 和 $\boldsymbol{u}_{\max} = u_{\max}\boldsymbol{1}$ 表示饱和的下限和上限, $\boldsymbol{1}$ 为维度合适的全 1 列向量.

1.3　多智能体强化学习基本知识

利用神经网络的特点, 多智能体强化学习可以处理高维复杂环境下的决策问题. 当环境中存在多个仅可观测局部信息的决策者时, 可以用部分可观马尔科夫博弈 (Partially observable Markov games,

POMG) 进行建模[25].

参考图 2, 将模型表示为 $\langle S, A, P, \{R^i\}, O, \{Z^i\}, N, \gamma\rangle$. 其中, $S = S^1 \times S^2 \times \cdots \times S^N$ 表示联合状态空间; $A = A^1 \times A^2 \times \cdots \times A^N$ 表示联合动作空间; $P = P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a})$ 表示在状态 $\boldsymbol{s}_t$ 下, 所有智能体采取联合动作 $\boldsymbol{a}$ 时, 环境转移到状态 $\boldsymbol{s}_{t+1}$ 的概率, $\boldsymbol{s}_{t+1}$, $\boldsymbol{s}_t \in S$, $\boldsymbol{a} \in A$; R^i 表示智能体 i 的奖励函数, 其取值 r^i 称为智能体 i 获得的即时奖励; $O \subseteq S$ 表示可观测空间; $Z^i = Z^i(\boldsymbol{o}^i|\boldsymbol{s}, \boldsymbol{a}^i)$ 表示在执行动作 $\boldsymbol{a}^i$ 并转移到状态 $\boldsymbol{s}$ 下, 获得观测 $\boldsymbol{o}^i \in O$ 的概率, $\boldsymbol{a}^i \in A^i$; N 表示智能体集合; $\gamma \in [0, 1]$ 表示折扣因子, 其取值决定了当前奖励与未来奖励之间的相对重要性.

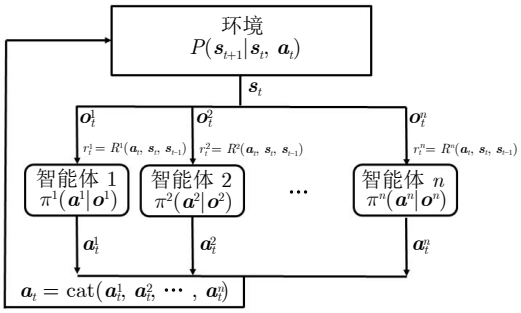


图 2 部分可观马尔科夫博弈

Fig. 2 Partially observable Markov games

智能体 i 的策略 $\pi^i(\boldsymbol{a}^i|\boldsymbol{o}^i)$ 定义了在获得观测 $\boldsymbol{o}^i$ 时采取动作 $\boldsymbol{a}^i$ 的概率. 智能体 i 的目标是最大化自身的未来累积折扣奖励

$$G_t^i = \sum_{k=0}^{\infty} \gamma^k r_{t+k}^i \tag{4}$$

其中, t 是累积折扣奖励开始计算的时间步.

在部分可观马尔科夫博弈模型中, 某个智能体在训练过程中获得的奖励不仅与环境有关, 还与其他智能体的策略相关, 而所有智能体的策略均不断更新, 这无疑增加了训练过程的不稳定性.

从现有文献看, 多智能体强化学习可按训练架构分为三类[26], 分别是分布式架构 (Distributed training with distributed execution, DTDE)、中心式架构 (Centralized training with centralized execution, CTCE) 和中心式训练–分布式执行架构 (Centralized training with distributed execution, CTDE). 在本文的控制目标下, 策略以局部观测信息作为输入, 且需要适应不同规模的集群. 所以, 在训练过程中, 各航天器也应只使用局部观测信息, 而不是所有航天器的全局信息, 以保证策略可以适应不同规模的集群.

2 航天器自主编队深度强化学习框架

针对上节阐述的航天器自主编队任务, 本节提出通过深度强化学习框架训练策略实现控制目标. 通过引入动力学奖励, 考虑航天器近距离机动下的 J_2 摄动相对动力学模型路径可行性.

图 3 是本文所提方法的流程图, 本节其余部分将结合图 3, 从训练环境、训练架构与损失函数、网络结构和奖励函数设置来介绍本文方法的细节.

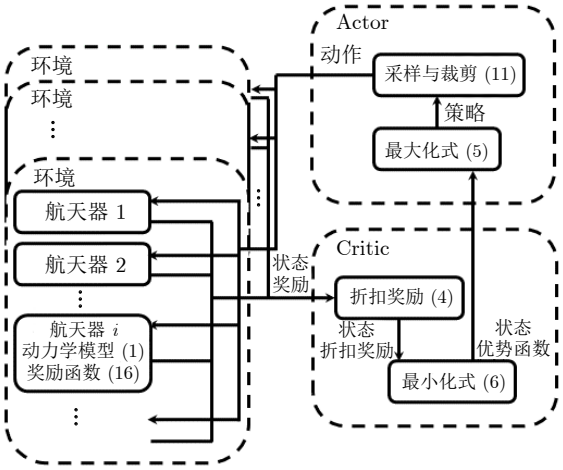


图 3 所提方法流程图

Fig. 3 Flowchart of the proposed method

2.1 训练环境

在多智能体强化学习训练环境中, 大多关心任务目标是否达成. 因此, 认为智能体状态可根据神经网络输出动作而随时更改. 但事实上, 航天器状态受到其动力学方程制约, 因此, 本文建立考虑动力学方程 (1) 的航天器集群自主编队训练环境. 将航天器执行任务可能到达的空间 R 划分为 m 个互不相交的子区域 R_1, R_2, $\cdots$, R_m, 且 $R = \cup_{i=0}^{m} R_i$. 记向量 $\boldsymbol{d} \in \mathbf{R}^m$ 的维数与子区域个数相等, 用来标记对应子区域是否为期望子区域, 其第 j 个元素为 1 表示子区域 R_j 为期望区域, 为 0 则表示 R_j 不是期望区域. 在本文中, 设最终期望区域可由某多边形表示, 即期望区域 $R_d = \cup_{k,\, d_k=1} R_k$ 为多边形, 如图 4 中的绿色区域. 该训练环境可包含任意 n 个航天器, 其状态更新遵循式 (1). 在图 4 中, 红色圆点表示航天器, 圆点的半径即为避撞半径; 与红色圆点同心的半透明黄色圆为该航天器可探测信息的范围, 其半径为低功耗雷达的最大探测半径. 单个航天器规划终止有两种: 第一种是航天器与其他航天器或环境边界发生碰撞, 此时仅发生碰撞的航天器被重新随机分配位置; 第二种是所有航天器都保持

在期望区域内若干步 s_f 后, 此时所有航天器被重新随机分配位置并开始下一轮仿真.

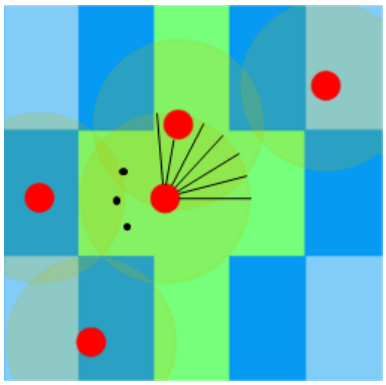


图 4　训练环境渲染图

Fig. 4　Rendering graph of the training environment

每个航天器的局部观测信息由环境信息和实时探测信息组成. 环境信息由期望区域多边形 R_d 的顶点位置坐标、环境的边界组成, 每步的实时探测信息则为自身当前位置、速度以及相邻航天器和环境边界与自身的距离, 如图 4 中的黑色线段所示, 可由低功耗雷达扫描得到.

2.2　训练架构与损失函数

本节在图 4 所示的环境上建立深度强化学习训练架构. 该架构基于近端策略优化算法[27−28], 在训练过程中各航天器共享 Actor 和 Critic 网络参数. 具体地, Actor 网络 $\pi_{\boldsymbol{\theta}}$ 是以 $\boldsymbol{\theta}$ 为参数的策略网络, 且 $\pi_{\boldsymbol{\theta}}=\pi^i$, $i=1,\ 2,\ \cdots,\ n$, 目标为学习最终的控制策略. 它收集并行训练环境中 n 个航天器的局部观测, 给出这 n 个航天器应采取的下一步动作, 然后将这些动作分发给各航天器. 在每一步中, Actor 网络根据 Critic 网络对优势函数的估计来更新自身参数 $\boldsymbol{\theta}$, 谋求下述目标函数的最大化

$$\begin{aligned} L\left(\boldsymbol{\theta}\right)=&\frac{1}{Bn}\sum_{i=1}^{B}\sum_{k=1}^{n}\min(r_{\boldsymbol{\theta},\,i}^{k}A_i^k,\\ &\mathrm{clip}\left(r_{\boldsymbol{\theta},\,i}^{k},\ 1-\varepsilon,\ 1+\varepsilon\right)A_i^k)+\\ &\sigma\frac{1}{Bn}\sum_{i=1}^{B}\sum_{k=1}^{n}S\left[\pi_{\boldsymbol{\theta}}\left(\boldsymbol{a}_i^k\left|\boldsymbol{o}_i^k\right.\right)\right] \end{aligned}\tag{5}$$

其中, $r_{\boldsymbol{\theta},\,i}^{k}=\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_i^k|\boldsymbol{o}_i^k)/\pi_{\boldsymbol{\theta}_{\mathrm{last}}}(\boldsymbol{a}_i^k|\boldsymbol{o}_i^k)$ 是更新前后策略网络的概率比, 用以表征更新前后策略的差异, $\pi_{\boldsymbol{\theta}_{\mathrm{last}}}(\boldsymbol{a}_i^k|\boldsymbol{o}_i^k)$ 表示更新前的策略网络; B 是批大小, 即每次梯度迭代的样本数量; n 是航天器的个数; A_i^k 是优势函数, 由 Critic 网络估计; ε 是策略网络裁剪超参数; $\mathrm{clip}(r_{\boldsymbol{\theta},\,i}^{k},\ 1-\varepsilon,\ 1+\varepsilon)$ 是裁剪函数, 将 $r_{\boldsymbol{\theta},\,i}^{k}$ 裁剪在 $[1-\varepsilon,\ 1+\varepsilon]$ 区间; $S[\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_i^k|\boldsymbol{o}_i^k)]$ 是策略 $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_i^k|\boldsymbol{o}_i^k)$ 的信息熵; σ 是与策略信息熵相关的超参数, 保持策略在训练过程中对环境的探索, 避免提前收敛至局部最优解.

与此同时, Critic 网络 $V_{\boldsymbol{\phi}}$ 是以 $\boldsymbol{\phi}$ 为参数的评价网络, 且 $V_{\boldsymbol{\phi}}=V^i$, $i=1,\ 2,\ \cdots,\ n$, 用于估计优势函数. 它收集并行训练环境中的 n 个航天器的状态和环境返回的即时奖励, 并拟合累积折扣奖励值. Critic 网络谋求下述损失函数的最小化

$$\begin{aligned} L(\boldsymbol{\phi})=\frac{1}{Bn}\sum_{i=1}^{B}\sum_{k=1}^{n}\max\Big(&\left(V_{\boldsymbol{\phi}}\left(\boldsymbol{s}_i^k\right)-G^i\right)^2,\\ &\left(\mathrm{clip}\left(V_{\boldsymbol{\phi}}\left(\boldsymbol{s}_i^k\right),\ V_{\boldsymbol{\phi}_{\mathrm{last}}}\left(\boldsymbol{s}_i^k\right)-\delta,\right.\right.\\ &\left.\left.V_{\boldsymbol{\phi}_{\mathrm{last}}}\left(\boldsymbol{s}_i^k\right)+\delta\right)-G^i\right)^2\Big) \end{aligned}\tag{6}$$

其中, G^i 是累积折扣奖励, 通过 Actor 网络 $\pi_{\boldsymbol{\theta}}$ 与环境交互得到; $V_{\boldsymbol{\phi}_{\mathrm{last}}}(\boldsymbol{s}_i^k)$ 表示更新前的 Critic 网络; δ 是 Critic 网络裁剪超参数.

在训练阶段, 如图 5 所示的 Actor 和 Critic 网络均以局部观测信息作为输入, 交替更新各自参数改进策略; 在执行阶段, 各航天器仅依赖 Actor 网络以局部观测信息来独立做出决策.

2.3　网络结构

本节介绍 Actor 和 Critic 网络的具体结构和输入输出的具体形式. 在本文中, 这两个网络除输出之外具有一致的结构. 网络结构及其输入输出对各航天器都是一致的. 因此, 在不引起歧义的情况下, 本节剩下部分省略表示航天器编号的上标或下标.

记网络的输入数据为 $\boldsymbol{o}=[\hat{\boldsymbol{r}}_s^{\mathrm{T}},\ \tilde{\boldsymbol{r}}_d^{\mathrm{T}},\ \hat{\boldsymbol{r}}_d^{\mathrm{T}},\ \hat{\boldsymbol{r}}^{\mathrm{T}},\ \boldsymbol{v}^{\mathrm{T}},\ \hat{\boldsymbol{l}}^{\mathrm{T}}]^{\mathrm{T}}$. 具体地, $\hat{\boldsymbol{r}}_s$ 为归一化的轨道根数, $\boldsymbol{v}$ 表示航天器的当前速度, 令 $\boldsymbol{r}$ 表示航天器的当前位置, 则

$$\hat{\boldsymbol{r}}=(\boldsymbol{r}-\boldsymbol{r}_{\mathrm{c}})\,./\,(\boldsymbol{r}_{\mathrm{box}}-2\boldsymbol{r}_{\mathrm{c}})\tag{7}$$

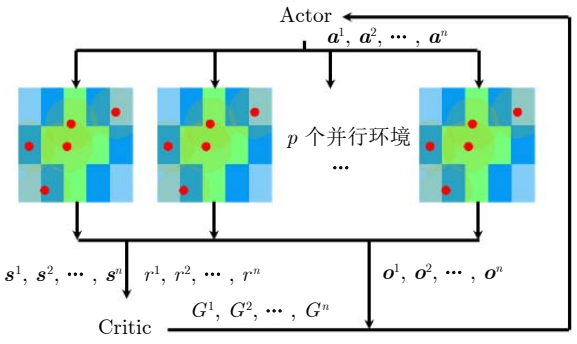


图 5　训练架构示意图

Fig. 5　Schematic diagram of the training framework

其中, $\hat{\boldsymbol{r}}$ 为位置的归一化值; $\boldsymbol{r}_{\rm c}=r_{\rm c}\boldsymbol{1}$, $r_{\rm c}$ 为避撞半径; $\boldsymbol{r}_{\rm box}$ 表示环境的尺寸; "./"表示对两向量的各分量逐一做除法. 由于当航天器与环境边界之间的距离小于避撞半径时就认为发生了碰撞, 所以式 (7) 相当于归一化的航天器当前位置. 假设 $\boldsymbol{r}_d$ 为航天器当前位置指向其与期望区域边界最近处的向量, 即 $\boldsymbol{r}_d=\arg\min\limits_{\boldsymbol{p}\in\partial R_d}(\|\boldsymbol{r}-\boldsymbol{p}\|)-\boldsymbol{r}$, 其中 ∂R_d 为期望区域的边界, $\boldsymbol{p}$ 是期望区域边界上的点, 则 $\tilde{\boldsymbol{r}}_d$ 与 $\hat{\boldsymbol{r}}_d$ 定义为

$$\tilde{\boldsymbol{r}}_d=\begin{cases}|\boldsymbol{r}_d|, & \boldsymbol{r}\in R_d\\ -|\boldsymbol{r}_d|, & \text{否则}\end{cases}\tag{8}$$

$$\hat{\boldsymbol{r}}_d=\frac{\boldsymbol{r}_d}{\|\boldsymbol{r}_d\|}\tag{9}$$

式 (8) 中的 $|\boldsymbol{r}_d|$ 代表对 $\boldsymbol{r}_d$ 的每个分量取绝对值. 虽然式 (8) 的编码在理论上存在冗余, 但是为尽量减少不连续因素、提高网络的拟合效果, 本文采用式 (8) 的形式而非 One-Hot 编码, 以区分航天器是否在期望区域内. 令 $\boldsymbol{l}$ 表示低功耗雷达探测得到的原始数据, 如图 6 左侧所示, 将其展开、归一化后得到图 6 右侧的数据

$$\hat{\boldsymbol{l}}=(\boldsymbol{r}_{\rm obs}-\boldsymbol{l})./(\boldsymbol{r}_{\rm obs}-\boldsymbol{r}_{\rm c})\tag{10}$$

其中, $\boldsymbol{r}_{\rm obs}=r_{\rm obs}\boldsymbol{1}$, $r_{\rm obs}$ 是观测半径. 式 (10) 在数值上对 $\boldsymbol{l}$ 进行归一化, 从直观上看, $\hat{\boldsymbol{l}}$ 中各元素之和的相对大小表示航天器与其他航天器或边界发生碰撞的概率.

Actor 网络的输出是下一步航天器期望位置增量 $\Delta\boldsymbol{r}_d$ 和期望速度 $\boldsymbol{v}_d$ 的概率分布. 在本文中, 假设这两个分布为正态分布. 控制策略的下一步动作由这两个分布采样得到, 并进行裁剪操作, 以防止极端值的影响. 其中裁剪区间 $[\boldsymbol{a}_{\min},\ \boldsymbol{a}_{\max}]$ 是与航天器机动能力有关的参数, $\boldsymbol{a}_{\min}$ 和 $\boldsymbol{a}_{\max}$ 分别表示最小和最大控制加速度向量, 定义为

$$\boldsymbol{a}^i=\left[\Delta\boldsymbol{r}_d^{\rm T},\ \boldsymbol{v}_d^{\rm T}\right]^{\rm T}={\rm clip}\left({\rm sample}\left(\pi_{\boldsymbol{\theta}}\right),\ \boldsymbol{a}_{\min},\ \boldsymbol{a}_{\max}\right)\tag{11}$$

其中, sample(·) 表示对输出的分布进行采样. Critic 网络的输出为标量, 且在输出层不设计激活函数.

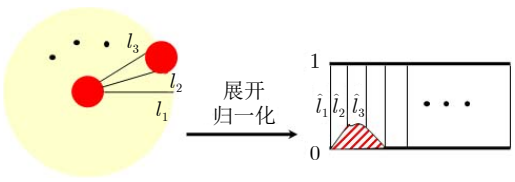


图 6 局部观测信息

Fig. 6 Local observed information

Actor 网络与 Critic 网络均采用多层感知机 (Multilayer perceptron, MLP) 模型, 激活函数为 LeakyReLU.

2.4 奖励函数设置

本文的奖励函数由三部分组成, 分别是避撞奖励 $r_{\rm col}$、期望奖励 $r_{\rm d}$ 和动力学奖励 $r_{\rm dyn}$.

避撞奖励 $r_{\rm col}$ 在每步开始时被赋值为 0, 当航天器与其他航天器或边界发生碰撞时, $r_{\rm col}$ 被加上相应的数值, 即

$$r_{\rm col}=\begin{cases}r_{\rm col}+r_{\rm b}, & d_{\rm b}<r_{\rm c}\\ r_{\rm col}+r_{\rm inter}, & d_{\rm inter}<r_{\rm c}\\ r_{\rm col}+0, & \text{否则}\end{cases}\tag{12}$$

其中, $r_{\rm b}$, $r_{\rm inter}<0$ 为常数, $r_{\rm b}$ 表示航天器与边界发生碰撞的奖励, $r_{\rm inter}$ 表示与其他航天器发生碰撞的奖励; $d_{\rm b}$ 表示航天器与边界的最近距离; $d_{\rm inter}$ 表示航天器与最近航天器之间的距离. 需注意的是, 式 (12) 中的条件并不互斥, 在每步中都需要逐条检查.

期望奖励 $r_{\rm d}$ 用于在训练过程中引导航天器进入期望区域, 定义为

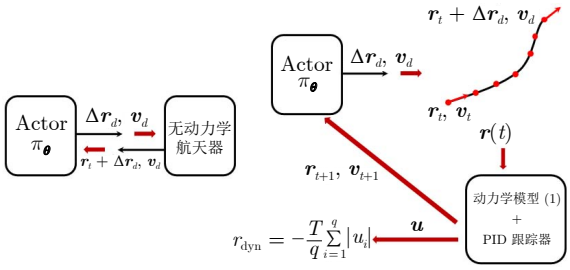
$$r_{\rm d}=\begin{cases}r_{\rm base}+r_{\rm inc}\left(\|\boldsymbol{r}_{d,\,t}\|-\|\boldsymbol{r}_{d,\,t-1}\|\right), & \boldsymbol{r}\notin R_d\\ 0, & \text{否则}\end{cases}\tag{13}$$

其中, $r_{\rm base}$, $r_{\rm inc}<0$ 为常数, $r_{\rm base}$ 表示不在期望区域内获得的基础奖励, $r_{\rm inc}$ 是折扣因子; $\|\boldsymbol{r}_{d,\,t}\|$ 和 $\|\boldsymbol{r}_{d,\,t-1}\|$ 分别表示当前步和前一步航天器与期望区域边界最近处的距离.

动力学奖励 $r_{\rm dyn}$ 的计算依赖于控制策略给出的动作、航天器的动力学方程和底层跟踪器. 为便于理解, 参考图 7(b), 由 Actor 网络 $\pi_{\boldsymbol{\theta}}$ 得到下一步的期望位置增量 $\Delta\boldsymbol{r}_d$ 和期望速度 $\boldsymbol{v}_d$, 航天器结合当前时刻的位置 $\boldsymbol{r}_t$ 和速度 $\boldsymbol{v}_t$, 求解满足如下条件的期望轨迹 $\boldsymbol{r}(t)$

$$\begin{cases}\boldsymbol{r}(0)=\boldsymbol{r}_t\\ \boldsymbol{r}(T)=\boldsymbol{r}_t+\Delta\boldsymbol{r}_d\\ \dot{\boldsymbol{r}}(0)=\boldsymbol{v}_t\\ \dot{\boldsymbol{r}}(T)=\boldsymbol{v}_d\end{cases}\tag{14}$$

其中, T 为常数, 表示期望轨迹经历的时间, 属于超参数. 将时间离散为 q 段后, 采用底层的 PID 跟踪器跟踪期望轨迹 $\boldsymbol{r}(t)$, 得到最终的实际位置 $\boldsymbol{r}_{t+1}$ 和速度 $\boldsymbol{v}_{t+1}$ 作为环境的返回. 将 PID 跟踪器计算的控制力施加到 J_2 摄动相对动力学方程 (1) 上, 如图 7(b) 所示, 并基于式 (1) 来更新航天器下一刻的



(a) 无动力学　(b) 有动力学
(a) Without dynamic　(b) With dynamic

图 7　有无动力学对比图

Fig. 7　Comparison graph with and without dynamic

状态. 随后, PID 跟踪器依据新状态计算控制力, 如此循环. 在控制过程中, 产生的控制力向量 $\boldsymbol{u} \in \mathbf{R}^q$ 用于计算动力学奖励

$$r_{\text{dyn}} = -\frac{T}{q}\sum_{i=1}^{q}|u_i| \tag{15}$$

其中, u_i 表示 $\boldsymbol{u}$ 的第 i 个分量. 若假设燃料消耗量与输出控制力的大小呈线性关系, 则该奖励可理解为跟踪期望轨迹所需燃料的负数. 在图 7(b) 中, 最终的实际位置 $\boldsymbol{r}_{t+1}$ 和速度 $\boldsymbol{v}_{t+1}$ 通常并不等于期望位置 $\boldsymbol{r}_t + \Delta \boldsymbol{r}_d$ 和期望速度 $\boldsymbol{v}_d$, 其值与当前位置和速度、跟踪器参数、航天器动力学模型、期望轨迹经历的时间等有关. 但对于图 7(a) 所示的无动力学情况, 最终的实际位置 $\boldsymbol{r}_{t+1}$ 和速度 $\boldsymbol{v}_{t+1}$ 等于期望位置 $\boldsymbol{r}_t + \Delta \boldsymbol{r}_d$ 和期望速度 $\boldsymbol{v}_d$.

本文中, 最终奖励是上述三部分奖励的加权和

$$r_{\text{total}} = r_{\text{col}} + \alpha_1 r_{\text{d}} + \alpha_2 r_{\text{dyn}} \tag{16}$$

其中, α_1 和 α_2 为调整期望奖励和动力学奖励在总奖励中占比的超参数. Actor 网络期望最大化总奖励 (16) 的未来累积折扣奖励, 这包括了避撞、任务和动力学相关三部分. 将式 (16) 代入式 (4), 则 Critic 网络的目标函数 (6) 中计入了动力学效应. 随后 Actor 网络对目标函数 (5) 进行优化, 使训练后的策略计入动力学效应.

由于强化学习得到的最优策略对奖励经过仿射变换具有不变性[29], 可以结合 $\boldsymbol{a}_{\min}$, $\boldsymbol{a}_{\max}$ 的值调整 r_{base}, r_{inc}, 使得 $r_{\text{d}} < 0$, 这样理论上能得到的最高奖励是 0, 以方便判断训练结果的收敛性.

3　结果与分析

基于第 2 节建立的航天器自主编队深度强化学习框架, 本节给出具体任务场景下的训练结果, 以及与无动力学、动力学奖励为零等场景的对比. 此外, 详细对比动力学奖励占比不同对训练结果的定量影响.

3.1　训练结果

在本节中, 设置并行训练场景的参数一致, 并以不同的随机数种子驱动运行. 在场景中, 航天器可能到达的空间为 4 m × 4 m 的正方形, 被平均划分为 3 × 5 的子区域. 如图 4 所示, 期望区域被设置为“十”字形. 该任务针对 8 个航天器, 其避撞半径均为 0.15 m, 观测半径均为 0.5 m. 只要将所有航天器在不发生碰撞情况下控制在期望区域内持续 100 步, 即被认为完成任务. Actor 网络和 Critic 网络拥有相同的两层隐藏层, 神经元个数分别是 64 和 32, 部分超参数取值如表 1 所示. 表中取值为 $\{*, *\}$ 的参数是指在训练过程中该参数随着训练轮数从前者线性更新至后者.

Actor 网络输出的裁剪区间为 $\boldsymbol{a}_{\max} = -\boldsymbol{a}_{\min} = \begin{bmatrix}0.02 & 0.02 & 0.01 & 0.01\end{bmatrix}^{\mathrm{T}}$, 期望轨迹经历的时间为 T =10 s. 根据表 1 给出的控制加速度上下限 $\pm$ 0.001 m/s^2 可以看出, 跟踪器的速度无法在相邻步内由 − 0.01 m/s 增加到 0.01 m/s. 因此, 策略网络需要考虑输出轨迹的动力学可行性.

在本节的每轮训练中, 共有 128 个并行环境, 分别以随机数种子 1 234, 1 235, ⋯, 1 361 驱动生成航天器在环境中无碰撞的初始位置. 每个环境中包含 8 个航天器, 每个航天器在每轮探索环境 75 步,

表 1　部分超参数取值

Table 1　The values of some hyperparameters

参数符号	含义	取值
p	并行训练环境数	128
σ	策略信息熵加权系数	$\{3\times10^{-3},\ 3\times10^{-5}\}$
ε, δ	裁剪超参数	0.2
γ	折扣因子	0.99
$u_{\min}$	控制加速度下限	−0.001 m/s^2
$u_{\max}$	控制加速度上限	0.001 m/s^2
$\dim(\boldsymbol{l})$	低功耗雷达检测数据数	90
r_{b}	与边界碰撞的奖励	−500
r_{inter}	相互碰撞的奖励	−10
r_{base}, r_{inc}	期望奖励相关超参数	1, 35
α_1	期望奖励加权系数	0.6
l_r	学习率	$\{2\times10^{-4},\ 0\}$
a	半长轴	7 100 km
ω	近地点幅角	−20°
f	真近点角	20°
Ω	升交点经度	0°
e	离心率	0.05
i_o	轨道倾角	15°
k_p, k_i, k_d	PID 跟踪器参数	1.0, 0.01, 2.0

每轮共探索环境 76 800 步. 在将 76 800 步数据分为 30 个 batch 后, 每轮通过随机选择 batch 对 Actor 网络进行 450 次参数更新, 整个 500 轮训练过程总计探索环境 3840 万步.

平均奖励随训练轮数的变化如图 8 所示. 其中, 无动力学曲线表示的训练环境认为航天器可在相邻步实现动作空间中任意两个动作之间的转移, 即总能到达期望动作, 因此该曲线快速收敛; 而 $\alpha_2 = 1.0$ 的曲线表示训练环境中存在动力学方程约束, 航天器不能在相邻步实现动作空间中任意两个动作之间的转移, 故该曲线的收敛速度明显低于无动力学曲线. 图 9(a) 和图 9(b) 分别表示无动力学和 $\alpha_2 = 1.0$ 的环境训练得到的航天器轨迹, 对比可见, 图 9(b) 的轨迹更具动力学可行性.

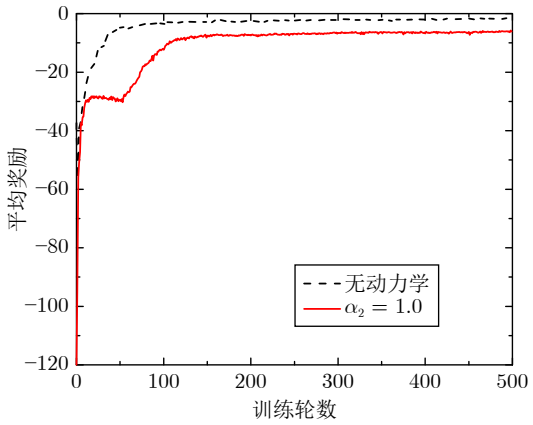
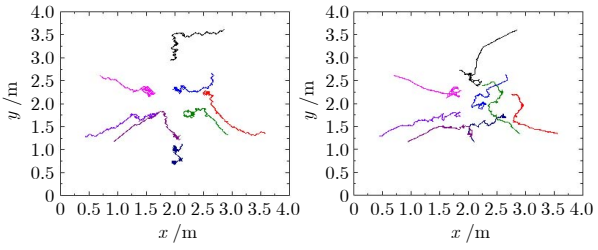


图 8 平均奖励随训练轮数变化图

Fig. 8 Graph of average reward over training epoch



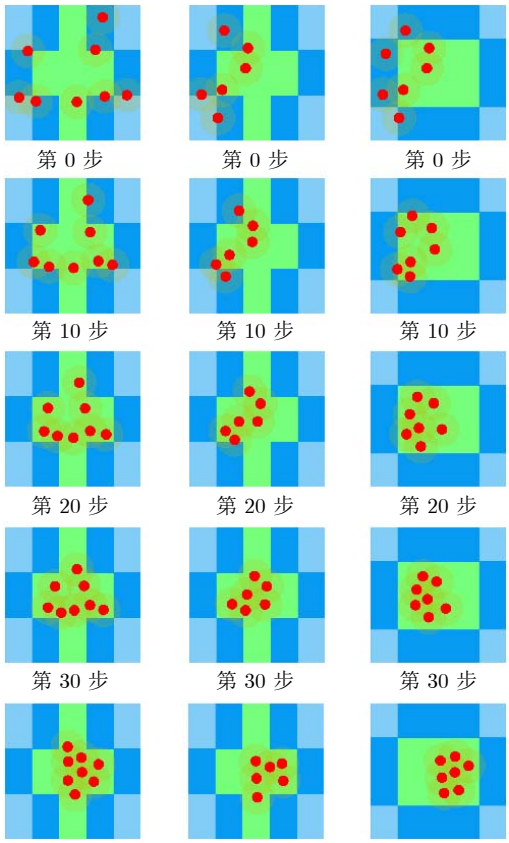
(a) 无动力学的轨迹 (b) $\alpha_2 = 1.0$ 的轨迹

(a) Trajectories without dynamic (b) Trajectories of $\alpha_2 = 1.0$

图 9 有无动力学奖励的轨迹对比

Fig. 9 Comparison of trajectories with or without dynamical reward

图 10(a) ~ 10(c) 分别给出了三个不同环境下某些时刻的快照图. 图 10(a) 与训练环境一致, 图 10(b) 含有不同个数的航天器, 图 10(c) 含有不同个数的航天器且期望区域也不同, 表明所给出的策略可以适应多种规模的集群和不同的期望构型.



(a) 场景 1 (b) 场景 2 (c) 场景 3

(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

图 10 策略在不同场景下的渲染图

Fig. 10 Rendering graphs of strategy in different scenarios

3.2 不同动力学奖励占比对训练结果的影响

为更好地对比分析不同动力学奖励占比对训练结果的影响, 本节采用 200 个随机生成的环境来评估训练后的策略. 这 200 个环境除航天器初始位置不同外, 其他参数与训练环境一致. 现定义如下三种评估指标: 一是定义完成率为完成任务的环境数与总的测试环境数的比值; 二是定义平均完成步数为所有完成任务的环境所需的平均步数, 任务的完成步数不得超过 500 步; 三是定义燃料消耗率为平均燃料消耗量与最大燃料消耗的比值, 其中平均燃料消耗量是控制力曲线下面积的平均值, 最大燃料消耗是驱动器全功率输出时控制力的曲线下面积. 假设燃料消耗量与输出控制力的大小呈线性关系, 燃料消耗率的物理含义是在平均完成时间下需要作动器全功率运行的时间占比, 而燃料消耗的绝对值为平均完成步数与燃料消耗率的乘积.

表 2 给出了动力学奖励在总奖励中占比不同时上述评估指标的值, “-”表示不计动力学效应. 可以

表 2　500 轮后不同动力学奖励占比的结果
Table 2　Results under different percentages of dynamical reward after 500 epochs

α_2	完成率 (%)	平均完成步数	燃料消耗率 (%)	燃料消耗绝对值
-	91.5	142.95		
0	84.5	123.95	61.21	75.88
0.5	69.5	125.42	58.17	72.96
1.0	64.0	127.79	46.83	59.84
1.5	0			

看到, 对于本文的网络结构、参数设置、训练架构和训练轮数, 当不考虑动力学效应时, 任务的完成率达到 91.5%, 平均完成步数为 142.95; 在考虑了动力学效应后, 即 α_2 分别为 0、0.5、1.0 时, 完成率下降到了 84.5%、69.5% 和 64.0%, 但平均完成步数缩短到 123.95、125.42 和 127.79, 这意味着任务的完成时间至少缩短了 10.6%. 在综合燃料消耗率后, $\alpha_2 = 1.0$ 时所得到的策略比 $\alpha_2 = 0$ 时的策略节省了 21.14% 的燃料. 值得指出的是, 动力学奖励在总奖励中的占比过高会使得机动成本较高, 当机动代价高于不前往期望区域的代价时, 策略会倾向于停在原地不动. 为避免策略无法完成任务, 可按以下步骤确定参数: 1) 确定避撞相关的奖励 $r_{\rm col}$; 2) 根据避撞奖励的值, 确定期望奖励及其超参数 α_1, 此时可设 $\alpha_2 = 0$, 保证策略可在有效避撞情况下前往期望区域; 3) 确定动力学奖励超参数 α_2.

4　结束语

本文提出一种航天器集群自主编队任务场景下的深度强化学习控制方法, 引入燃料消耗量为动力学奖励. 训练采用近端策略优化算法, 得到的策略执行过程是完全分布式的, 即每个航天器仅根据局部观测信息独立做出决策. 在训练场景中, 引入航天器的 J_2 摄动相对动力学模型, 隐式地限制了策略网络输出动作之间的任意转换. 进一步地, 将控制器给出的燃料消耗量负数作为动力学奖励, 引导策略优化控制所需的燃料量. 在仿真环境中训练得到的策略, 可适应各种规模的航天器集群和不同的期望区域. 仿真结果表明, 动力学奖励可显著提高策略生成轨迹的动力学可行性, 并降低编队控制过程所需的燃料消耗量, 验证了所提出的深度强化学习方法和引入的动力学奖励的有效性. 本文提出的框架会受限于具体航天任务对特定指标的强约束, 如要求任务成功率极高且燃料携带量有限. 更改网络结构和训练超参数可得到统计意义上满足某些强约束的策略, 但理论上满足强约束的策略训练框架尚需进一步研究.

References

1　Xue Z H, Liu J G, Wu C C, Tong Y C. Review of in-space assembly technologies. *Chinese Journal of Aeronautics*, 2021, **34**(11): 21−47

2　Ma Ya-Jie, Jiang Bin, Ren Hao. Adaptive direct fault-tolerant control design for spacecraft integrated attitude and orbit system. *Acta Automatica Sinica*, 2023, **49**(3): 678−686
(马亚杰, 姜斌, 任好. 航天器位姿运动一体化直接自适应容错控制研究. 自动化学报, 2023, **49**(3): 678−686)

3　Lymer J, Hanson M, Tadros A, Boccio J, Hollenstein B, Emerick K, et al. Commercial application of in-space assembly. In: Proceedings of the AIAA SPACE. Long Beach, USA: AIAA, 2016. 5236−5253

4　Bartlett R O. NASA standard multimission modular spacecraft for future space exploration. In: Proceedings of the 16th Goddard Memorial Symposium. Washington, USA: American Astronautical Society and Deutsche Gesellschaft für Luft- und Raumfahrt, 1978.

5　Wei Zheng-Tao. Autonomous Assembly Control and Ground Experiment of Modular Spacecraft [Ph.D. dissertation], Nanjing University of Aeronautics and Astronautics, China, 2023.
(韦正涛. 模块化航天器自主组装控制及地面实验研究 [博士学位论文], 南京航空航天大学, 中国, 2023.)

6　Dennison K, Stacey N, D′Amico S. Autonomous asteroid characterization through nanosatellite swarming. *IEEE Transactions on Aerospace and Electronic Systems*, 2023, **59**(4): 4604−4624

7　Zheng S K, Li T J, Zhao J, Ma X F, Zhu J L, Huang Z R, et al. Deployment impact experiment and dynamic analysis of modular truss antenna. *International Journal of Aerospace Engineering*, **2022**: Article No. 2038932

8　Burns R, McLaughlin C A, Leitner J, Martin M. TechSat 21: Formation design, control, and simulation. In: Proceedings of the IEEE Aerospace Conference. Proceedings (Cat. No.00TH 8484). Big Sky, USA: IEEE, 2000. 19−25

9　Bonin G, Roth N, Armitage S, Newman J, Risi B, Zee R E. CanX-4 and CanX-5 precision formation flight: Mission accomplished! In: Proceedings of the 29th Annual AIAA/USU Conference on Small Satellites. 2015.

10　Zheng Zhong, Li Peng, Qian Mo-Shu. Spacecraft attitude coordination control with angular velocity and input constraints. *Acta Automatica Sinica*, 2021, **47**(6): 1444−1452
(郑重, 李鹏, 钱默抒. 具有角速度和输入约束的航天器姿态协同控制. 自动化学报, 2021, **47**(6): 1444−1452)

11　Foust R C, Lupu E S, Nakka Y K, Chung S J, Hadaegh F Y. Autonomous in-orbit satellite assembly from a modular heterogeneous swarm. *Acta Astronautica*, 2020, **169**: 191−205

12　Camacho E F, Bordons C. *Model Predictive Control*. London: Springer, 2007.

13　Ortolano N, Geller D K, Avery A. Autonomous optimal trajectory planning for orbital rendezvous, satellite inspection, and final approach based on convex optimization. *The Journal of the Astronautical Sciences*, 2021, **68**(2): 444−479

14　Basu H, Pedari Y, Almassalkhi M, Ossareh H R. Computationally efficient collision-free trajectory planning of satellite swarms under unmodeled orbital perturbations. *Journal of Guidance, Control, and Dynamics*, 2023, **46**(8): 1548−1563

15　Yu Jie. Research on Multi-agent Cooperative Encirclement Algorithm Based on Deep Reinforcement Learning [Master thesis], Harbin University of Science and Technology, China, 2024.
(于杰. 基于深度强化学习的多智能体协同包围算法研究 [硕士学位论文], 哈尔滨理工大学, 中国, 2024.)

16　Wang Long, Huang Feng. An interdisciplinary survey of multi-agent games, learning, and control. *Acta Automatica Sinica*,

2023, **49**(3): 580−613
(王龙, 黄锋. 多智能体博弈、学习与控制. 自动化学报, 2023, **49**(3): 580−613)

17 Zhao Chun-Yu, Lai Jun. Survey on meta reinforcement learning. *Application Research of Computers*, 2023, **40**(1): 1−10
(赵春宇, 赖俊. 元强化学习综述. 计算机应用研究, 2023, **40**(1): 1−10)

18 Luo B, Wu H N, Huang T W. Off-policy reinforcement learning for H_∞ control design. *IEEE Transactions on Cybernetics*, 2015, **45**(1): 65−76

19 Vamvoudakis K G, Lewis F L. Online solution of nonlinear two-player zero-sum games using synchronous policy iteration. *International Journal of Robust and Nonlinear Control*, 2012, **22**(13): 1460−1483

20 Yan C, Xiang X J, Wang C. Fixed-wing UAVs flocking in continuous spaces: A deep reinforcement learning approach. *Robotics and Autonomous Systems*, 2020, **131**: Article No. 103594

21 Xu D, Guo Y X, Yu Z Y, Wang Z F, Lan R Z, Zhao R H, et al. PPO-exp: Keeping fixed-wing UAV formation with deep reinforcement learning. *Drones*, 2023, **7**(1): Article No. 28

22 Chen Y F, Liu M, Everett M, How J P. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 285−292

23 Hu Cheng. Research on Key Technologies of Multi-agent Collaboration Based on Deep Reinforcement Learning [Ph.D. dissertation], Beijing University of Posts and Telecommunications, China, 2025.
(胡成. 基于深度强化学习的多智能体协作关键技术研究 [博士学位论文], 北京邮电大学, 中国, 2025.)

24 Wang D W, Wu B L, Poh E K. *Satellite Formation Flying*. Singapore: Springer, 2017.

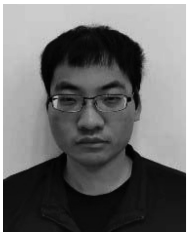
25 Hansen E A, Bernstein D S, Zilberstein S. Dynamic programming for partially observable stochastic games. In: Proceedings of the 19th National Conference on Artifical Intelligence. San Jose, USA: AAAI, 2004. 709−715

26 Wen Guang-Hui, Yang Tao, Zhou Jia-Ling, Fu Jun-Jie, Xu Lei. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems. *Control and Decision*, 2023, **38**(5): 1200−1230
(温广辉, 杨涛, 周佳玲, 付俊杰, 徐磊. 强化学习与自适应动态规划: 从基础理论到多智能体系统中的应用进展综述. 控制与决策, 2023, **38**(5): 1200−1230)

27 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347, 2017.

28 Yu C, Velu A, Vinitsky E, Gao J X, Wang Y, Bayen A, et al. The surprising effectiveness of PPO in cooperative multi-agent games. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1787

29 Zhao S Y. *Mathematical Foundations of Reinforcement Learning*. Singapore: Springer, 2025.



金伟成 南京航空航天大学航空学院博士研究生. 2021 年获得南京航空航天大学工程力学专业学士学位. 主要研究方向为航天器集群的导航、动力学与控制, 马尔科夫过程, 分布式控制系统.
E-mail: jinweich@nuaa.edu.cn
(**JIN Wei-Cheng** Ph.D. candidate at the College of Aerospace Engineering, Nanjing University of Aeronautics and Astronautics. He received his bachelor degree in engineering mechanics from Nanjing University of Aeronautics and Astronautics in 2021. His research interest covers guidance, dynamics and control for spacecraft swarm, Markov process, and distributed control system.)

陈 提 南京航空航天大学教授. 2017 年获得南京航空航天大学动力学与控制专业博士学位. 主要研究方向为在轨自主组装, 绳系卫星, 复杂结构的动力学与控制.
E-mail: chenti@nuaa.edu.cn
(**CHEN Ti** Professor at Nanjing University of Aeronautics and Astronautics. He received his Ph.D. degree in dynamics and control from Nanjing University of Aeronautics and Astronautics in 2017. His research interest covers in-orbit autonomous assembly, tethered satellite, and complex structure dynamics and control.)

胡海岩 南京航空航天大学教授. 1988 年获得南京航空航天大学固体力学专业博士学位. 主要研究方向为柔性结构的时滞控制, 飞行器结构颤振主动抑制, 空间结构展开动力学. 本文通信作者.
E-mail: hhyae@nuaa.edu.cn
(**HU Hai-Yan** Professor at Nanjing University of Aeronautics and Astronautics. He received his Ph.D. degree in solid mechanics from Nanjing University of Aeronautics and Astronautics in 1988. His research interest covers the delayed control of flexible structures, the active flutter suppression of aircraft structures, and the deployment dynamics of space structures. Corresponding author of this paper.)