自动化学报
ACTA AUTOMATICA SINICA

基于相对阶解耦的多智能体系统安全强化学习

黄 张斯伦

Safe Reinforcement Learning for Multi-agent Systems Based on Relative-degree Decoupling

HUANG Di, ZHANG Si-Lun

在线阅读 View online: https://doi.org/10.16383/j.aas.c260090

您可能感兴趣的其他文章

安全强化学习综述

Safe Reinforcement Learning: A Survey

自动化学报. 2023, 49(9): 1813−1835 https://doi.org/10.16383/j.aas.c220631

基于多智能体强化学习的博弈综述

Multi-agent Reinforcement Learning Based Game: A Survey

自动化学报. 2025, 51(3): 540−558 https://doi.org/10.16383/j.aas.c240478

具有指数图信息通信的大规模情境多智能体强化学习

Large-scale Episodic Multi-agent Reinforcement Learning With Exponential Graph Information Communication

自动化学报. 2026, 52(6): 1304−1318 https://doi.org/10.16383/j.aas.c250633

多智能体强化学习控制与决策研究综述

Survey on Multi-agent Reinforcement Learning for Control and Decision-making

自动化学报. 2025, 51(3): 510−539 https://doi.org/10.16383/j.aas.c240392

DOS攻击下饱和脉冲多智能体系统的安全定制化一致性

Security Customization Consensus of Multi-agent Systems Based on Saturation Impulse Under DOS Attacks

自动化学报. 2024, 50(12): 2499−2512 https://doi.org/10.16383/j.aas.c240203

基于分布式自适应内模的多智能体系统协同最优输出调节

Cooperative Optimal Output Regulation for Multi-agent Systems Based on Distributed Adaptive Internal Model

自动化学报. 2025, 51(3): 678−691 https://doi.org/10.16383/j.aas.c240371

基于相对阶解耦的多智能体系统安全强化学习

黄 崛 [1] 张斯伦 [2]

摘 要 针对存在模型不确定性与外部扰动的多智能体协同控制问题, 提出一种基于模型相对阶解耦的安全强化学习框架. 首先, 利用非线性系统中的相对阶理论, 将高阶状态安全约束映射为输入显式出现的仿射不等式; 在凸多面体参数不确定与有界扰动条件下, 进一步给出由有限个线性不等式交集表征的可认证凸安全动作集, 为在线投影与约束求解提供显式可行域. 其次, 为解决单步屏蔽易产生后继不可行的问题, 设计带终端集的滚动安全滤波器: 在每个时刻维护有限时域备份控制序列, 并通过末端约束与终端控制器保证递归可行性与无限时域安全性. 最后, 利用相对阶标准型构建策略分解架构, 以 LQR 标称控制器提供稳定热启动, 强化学习仅负责对未建模动态进行补偿, 并采用参数共享的独立近端策略优化算法学习补偿量. 卫星编队姿态控制的仿真结果表明: 在执行器效率不确定与扰动条件下, 所提方法能够将约束违规率维持在接近零的水平, 同时实现更快的训练收敛与更低的跟踪误差.

关键词 多智能体系统; 安全强化学习; 相对阶; 控制屏障函数; 安全滤波器

引用格式 黄崛, 张斯伦. 基于相对阶解耦的多智能体系统安全强化学习. 自动化学报, 2026, **52**(10): 2307−2317

DOI 10.16383/j.aas.c260090 **CSTR** 32138.14.j.aas.c260090

Safe Reinforcement Learning for Multi-agent Systems Based on Relative-degree Decoupling

HUANG Di[1] ZHANG Si-Lun[2]

Abstract To address issues in cooperative multi-agent control subject to model uncertainty and external disturbances, this paper proposes a safe reinforcement learning framework based on model relative-degree decoupling. First, utilizing relative-degree theory in nonlinear systems, high-order state safety constraints are mapped into affine inequalities where the control input appears explicitly. Under conditions of convex-polytope parameter uncertainty and bounded disturbances, a certifiable convex safe action set characterized by the intersection of finite linear inequalities is derived to provide an explicit feasible region for online projection and constraint solving. Second, to overcome the issues of subsequent infeasibility often associated with single-step shielding, a receding-horizon safety filter incorporating a terminal set is designed. By maintaining a finite-horizon backup control sequence at each time step, this filter guarantees recursive feasibility and infinite-horizon safety via terminal constraints and a terminal controller. Finally, a policy decomposition architecture is constructed based on the relative-degree normal form. A linear quadratic regulator nominal controller provides a stable warm start, while reinforcement learning is only responsible for compensating for unmodeled dynamics using a parameter-sharing independent proximal policy optimization (IPPO) algorithm to learn the compensation. Simulation results on satellite formation attitude control demonstrate that, under actuator efficiency uncertainty and external disturbances, the proposed method maintains a near-zero constraint violation rate while achieving faster training convergence and reduced tracking error.

Keywords multi-agent systems; safe reinforcement learning; relative degree; control barrier functions; safety filter

Citation Huang Di, Zhang Si-Lun. Safe reinforcement learning for multi-agent systems based on relative-degree decoupling. *Acta Automatica Sinica*, 2026, **52**(10): 2307−2317

收稿日期 2026-02-03 录用日期 2026-04-12
Manuscript received February 3, 2026; accepted April 12, 2026
中国航空科学基金 (202300010S5001), 国家自然科学基金 (61903278) 资助
Supported by the Aeronautical Science Foundation of China (202300010S5001) and National Natural Science Foundation of China (61903278)
本文责任编委 付俊杰
Recommended by Associate Editor FU Jun-Jie
1. 湖北珞珈实验室 武汉 430072 中国 2. 瑞典皇家理工学院数学系 斯德哥尔摩 SE 114 28 瑞典
1. Hubei Luojia Laboratory, Wuhan 430072, China 2. Department of Mathematics, KTH Royal Institute of Technology, Stockholm SE 114 28, Sweden

多智能体系统 (multi-agent systems, MAS) 在机器人协同、无人集群对抗及智能交通管控等领域的应用日益广泛[1−4]. 然而, 在面对模型不确定性及外部扰动时, 如何确保系统严格满足各种安全约束 (如光学载荷视场角限制、机间防碰撞等), 已成为限制其从理论过渡到实际应用的关键问题.

传统基于模型的控制方法, 例如模型预测控制 (model predictive control, MPC)[5]、带约束的非线性编队跟踪控制[6] 以及分布式非线性跟踪控制[7] 等,

均可在已知动力学与通信拓扑的前提下实现多智能体协同, 但这类方法普遍对模型精度与结构化假设依赖较强. 另外, 多智能体强化学习[8−10] 凭借强大的探索与拟合能力, 能够有效应对模型未知与环境时变问题. 但其作为黑盒系统, 缺乏充分的安全性保障, 极易在训练初期或探索阶段引发灾难性故障.

为弥补安全性短板, 安全强化学习应运而生. 早期的研究如奖励约束策略优化[11] 和形式化验证方法[12−14] 多采用拉格朗日乘子法将约束转化为奖励函数中的软惩罚项. 然而, 这类“软约束”方法无法从理论上保证硬约束的零违规, 且惩罚系数的调节极易导致训练振荡. 近年来, 结合控制屏障函数 (control barrier functions, CBF) 的屏蔽器 (shielding) 架构成为一种主流方向. 例如, Cheng 等[15] 提出利用高斯过程回归来学习模型不确定性, 并结合鲁棒 CBF 实现安全交互; Qin 等[16] 通过神经网络学习去中心化的屏障证书 (barrier certificates) 以降低对解析模型的依赖. 但上述方法多针对低相对阶的系统设计, 或在处理高阶非线性动力学时缺乏基于系统结构的理论保障. 针对复杂任务, Ahmad 等[17] 提出 HMARL-CBF 分层架构, 虽然在任务规划层引入协作机制, 但其底层的安全保障本质上仍属于单步屏蔽 (single-step shielding), 未从动力学层面解决高阶非线性带来的问题. 更重要的是, 现有基于 QP (quadratic programming) 的屏蔽方法普遍存在“贪婪近视”的缺陷. 为解决此问题, Hsu 等[18] 提出的 ISAACS 框架引入递推可行性的思想, 通过预测未来的安全状态来避免死锁. 然而其安全性在很大程度上依赖于神经网络的学习效果与泛化能力, 在复杂工况下难以提供确定性的安全承诺.

安全性问题在卫星编队飞行 (satellite formation flying) 领域尤为突出. 随着纳卫星技术的发展, 分布式编队与星座任务已成为执行对地观测、空间操作等复杂任务的重要形态[19]. 然而, 此类系统运行于高动态、强非线性的轨道环境中, 不仅面临严苛的构型保持与防碰撞约束, 还需应对日益复杂的空间电磁干扰与链路不确定性[20]. 在这些高阶动力学特性与极端的安全要求下, 现有的“贪婪”屏蔽方法或单纯的数据驱动策略往往难以在保证绝对安全的同时保持编队性能.

针对上述挑战, 本文提出一种基于相对阶[21] 解耦的多智能体安全强化学习方法. 不同于将系统视为黑盒的端到端学习, 该方法充分利用被控对象的物理结构特性. 本文的主要贡献在于:

1) 基于微分几何和非线性系统理论挖掘系统的结构特性, 引入高阶控制屏障函数 (high-order control barrier function, HOCBF)[22]. 通过相对阶变换, 将含有不确定性的非线性状态约束转化为线性不等式, 从而推导出输入空间上的显式安全动作集表示, 为系统的安全控制提供了可认证的可行动作边界.

2) 利用显式的安全动作集设计了递归可行的安全滤波器, 从而确保系统在无限时域内满足安全约束.

3) 利用相对阶构建“标称骨架 + 残差学习”的策略分解架构. 利用解耦后的线性子系统设计稳定控制器作为基准, 让强化学习负责补偿剩余动态, 该机制显著降低了探索空间的维度, 提升了训练收敛速度.

4) 在多星协同姿态编队的仿真环境中验证了所提方法的有效性. 实验结果表明, 相比现有基准方法, 本文方法在实现零安全违规的同时, 显著提升了收敛效率.

1 预备知识与问题描述

1.1 通信拓扑与图论描述

考虑由 N 个智能体组成的多智能体系统, 其通信网络由无向连通图 $G=(\mathcal{V},\ \mathcal{E})$ 描述. 其中 $\mathcal{V}=\{1,\ \cdots,\ N\}$ 为节点集, $\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$ 为边集. 智能体 i 的邻居集定义为

$$\mathcal{N}_i=\{j\in\mathcal{V}\mid(i,\ j)\in\mathcal{E}\} \tag{1}$$

在时刻 t, 智能体 i 可获得其本地状态 $x_i(t)$ 以及邻居状态 $\{x_j(t)\}_{j\in\mathcal{N}_i}$.

1.2 系统相对阶

为处理高阶非线性动力学约束, 本文利用微分几何和非线性系统中的相对阶理论[21] 对系统进行变换. 考虑如下多输入仿射非线性系统动力学方程:

$$\begin{cases}\dot{x}=f(x)+\displaystyle\sum_{j=1}^{m}g_j(x)u_j\\ y=h(x)\end{cases} \tag{2}$$

其中 $x\in\mathbf{R}^n$ 为状态; $u=[u_1,\ \cdots,\ u_m]^{\mathrm{T}}\in\mathbf{R}^m$ 为控制输入; y 为系统输出; $f:\mathbf{R}^n\to\mathbf{R}^n$ 为漂移向量场; $g_j:\mathbf{R}^n\to\mathbf{R}^n$ $(j=1,\ \cdots,\ m)$ 为第 j 个控制输入对应的输入向量场; $h:\mathbf{R}^n\to\mathbf{R}^p$ 为输出映射, p 为输出维数.

给出多输入非线性系统的相对阶定义, 其中 $\mathcal{L}_f$ 和 $\mathcal{L}_{g_j}$ 分别表示沿向量场 f 和 g_j 的李导数算子.

定义 1 (系统相对阶). 对于系统 (2), 若存在整数向量 $r=[r_1,\ \cdots,\ r_p]^{\mathrm{T}}$, 使得在某邻域 $\mathcal{U}^0$ 内满足:

1) $\forall i \in \{1, \cdots, p\}$, $\forall j \in \{1, \cdots, m\}$ 与 $k < r_i - 1$, 均有 $\mathcal{L}_{g_j}\mathcal{L}_f^k h_i(x) = 0$;

2) 解耦矩阵 $A_{ij}(x) = \mathcal{L}_{g_j}\mathcal{L}_f^{r_i-1}h_i(x)$ 非奇异.

则称系统具有相对阶 r.

定义 1 确保所涉系统局部 Byrnes-Isidori 标准型存在. 例如本文关注的飞行器通道具有有限相对阶且零动态具有输入到状态稳定性 (input-to-state stability, ISS) 性质 (详见第 5 节).

1.3　控制屏障函数与高阶控制屏障函数

同样对于系统 (2), 给定连续可微函数 $b: \mathbf{R}^n \to \mathbf{R}$, 其中 $b(x)$ 用于刻画系统的安全约束. 定义安全集

$$\mathcal{C} = \{x \in \mathbf{R}^n \mid b(x) \geq 0\} \tag{3}$$

定义 2 (控制屏障函数). 若存在扩展 $\mathcal{K}_\infty$ 类函数 $\alpha(\cdot)$, 使得对任意 $x \in \mathcal{C}$ 满足

$$\sup_{u \in \mathcal{U}_u} \Big[\mathcal{L}_f b(x) + \sum_{j=1}^{m} \mathcal{L}_{g_j} b(x)\, u_j + \alpha\big(b(x)\big)\Big] \geq 0 \tag{4}$$

则称 $b(x)$ 为控制屏障函数. 其中, $\mathcal{U}_u \subseteq \mathbf{R}^m$ 为允许输入集合 (如幅值/速率约束); sup 表示对允许输入集合 $\mathcal{U}_u$ 中的控制输入 u 取上确界.

当安全函数 $b(x)$ 相对输入 u 的相对阶 $r > 1$ 时, 需引入高阶控制屏障函数[22], 将约束沿导数链提升到输入显式出现的阶次.

定义 3 (高阶控制屏障函数). 对相对阶为 r 的系统, 定义辅助函数序列 $\psi_0(x), \cdots, \psi_r(x)$ 满足

$$\begin{cases} \psi_0(x) = b(x) \\ \psi_k(x) = \dot{\psi}_{k-1}(x) + \alpha_k\big(\psi_{k-1}(x)\big) \end{cases} \tag{5}$$

其中, $k = 1, \cdots, r$ 为递推索引; $\alpha_k(\cdot)$ 为可微扩展 $\mathcal{K}_\infty$ 类函数; $\dot{\psi}_{k-1}$ 表示 ψ_{k-1} 沿系统轨迹的时间导数. 若存在控制输入 u 使得

$$\sup_{u \in \mathcal{U}_u} \Big[\mathcal{L}_f \psi_{r-1}(x) + \sum_{j=1}^{m} \mathcal{L}_{g_j} \psi_{r-1}(x)\, u_j + \alpha_r\big(\psi_{r-1}(x)\big)\Big] \geq 0 \tag{6}$$

则称 $b(x)$ 为系统的高阶控制屏障函数. $\alpha_r(\cdot)$ 为高阶控制屏障函数递推中的第 r 个可微扩展 $\mathcal{K}_\infty$ 类函数.

1.4　动力学建模

考虑包含参数不确定性与外部扰动的智能体动力学

$$\dot{x}_i = f_i(x_i) + g_{i,\,1}(x_i)u_{i,\,1} + g_{i,\,2}(x_i,\, \kappa_i)u_{i,\,2} + D_i(t) \tag{7}$$

其中 $i \in \mathcal{V}$; $u_{i,\,1}$ 与 $u_{i,\,2}$ 分别表示已知增益通道和参数不确定增益通道的控制输入, $g_{i,\,1}(x_i)$ 与 $g_{i,\,2} \times (x_i,\, \kappa_i)$ 为对应的输入增益; 扰动 $D_i(t)$ 满足 $\|D_i(t)\| \leq \overline{D}_i$, $\overline{D}_i$ 表示扰动范数的已知上界; 参数 κ_i 属于凸多面体集.

$$\Lambda_i = \mathrm{conv}\{\vartheta_i^{(1)}, \cdots, \vartheta_i^{(L)}\} \tag{8}$$

其中, $\mathrm{conv}\{\cdot\}$ 为所列顶点的凸包, 即这些顶点的所有凸组合构成的集合; $\vartheta_i^{(1)}, \cdots, \vartheta_i^{(L)}$ 为凸多面体 Λ_i 的顶点; L 为顶点个数.

假设 1 (相对阶有效域). 存在状态空间中的开集 $\mathcal{U}^0$, 使得对于任一点 $x_i \in \mathcal{U}^0$, 非线性系统 (7) 具有相对阶 $r = [r_1, \cdots, r_p]^{\mathrm{T}}$. 在此基础上, 假设存在包含工作点的连通开集 $\Omega \subset \mathcal{U}^0$ (在该集合内解耦矩阵始终非奇异), 使得本文算法实际涉及的闭环轨迹始终满足 $x_i(t) \in \Omega$.

通过相对阶坐标变换 $\Phi_i : \mathcal{U}^0 \to \mathbf{R}^n$, 将状态 x_i 映射为

$$z_i = \Phi_i(x_i) = \begin{bmatrix} \xi_i \\ \eta_i \end{bmatrix} \tag{9}$$

其中, z_i 为经相对阶坐标变换后的状态向量, ξ_i 为由输出及其导数组成的“外部 (输出) 坐标”, η_i 为补足维度的零动态坐标. 具体对应关系为

$$\xi_i := \big[y_i^{\mathrm{T}},\ \dot{y}_i^{\mathrm{T}},\ \cdots,\ (y_i^{(r_i-1)})^{\mathrm{T}}\big]^{\mathrm{T}} \tag{10}$$

在该坐标下, 控制输入在 y_i 的第 r_i 阶导数中首次显式出现, 从而得到输入仿射表达:

$$y_i^{(r_i)} = \mathcal{A}_i(x_i) + \mathcal{B}_{i,\,1}(x_i)u_{i,\,1} + \mathcal{B}_{i,\,2}(x_i,\, \kappa_i)u_{i,\,2} + \Delta_i \tag{11}$$

其中 $\mathcal{A}_i(x_i)$ 为由 f_i 与输出函数 h_i 的高阶李导数决定的等效漂移项; $\mathcal{B}_{i,\,1}(x_i)$、$\mathcal{B}_{i,\,2}(x_i,\, \kappa_i)$ 为输入增益项, 分别对应于已知增益通道与含参数不确定增益通道在第 r_i 阶输出导数中的系数; Δ_i 为扰动在输出空间的投影.

注 1. 对于本文关注的飞行器姿态动力学, 解耦矩阵奇异或相对阶变化通常对应少数几类退化姿态/执行器几何退化点, 在工程可达域内往往呈现为测度很小的集合, 因此假设 1 在实际任务中具有合理性.

假设 2 (参数仿射不确定性). 假设不确定通道 $g_{i,\,2}(x_i,\, \kappa_i)$ 关于参数 κ_i 具有线性参数化结构, 即存在已知基函数 $\{\hat{g}_{i,\,2,\,k}(x_i)\}_{k=1}^{p}$, 使得

$$g_{i,\,2}(x_i,\,\kappa_i)=\sum_{k=1}^{p}\kappa_{i,\,k}\,\hat{g}_{i,\,2,\,k}(x_i) \tag{12}$$

其中 $\kappa_i=[\kappa_{i,\,1},\,\cdots,\,\kappa_{i,\,p}]^{\mathrm{T}}\in\Lambda_i$.

注 2. 该线性参数化覆盖常见乘性效率不确定性 (如推力/力矩系数).

1.5 基于相对阶的双重解耦

针对非线性系统 (7) 的多输入仿射结构, 本文引入基于相对阶的双重解耦机制, 以统一进行性能优化与安全约束. 分别以性能输出 $y_i=h_{\mathrm{perf}}(x_i)$ 与安全屏障函数 $b_\ell(x_i)\geq 0$ 为对象, 其中 $h_{\mathrm{perf}}(\cdot)$ 表示性能输出映射, 利用李导数构造两条输入–输出映射链路.

记 r_{sys} 为控制输入 $u_i=[u_{i,\,1}^{\mathrm{T}},\,u_{i,\,2}^{\mathrm{T}}]^{\mathrm{T}}$ 首次显式进入 y_i 导数的阶数 (通常情况下即第 1.2 节定义的 r), 则可据此建立局部微分同胚变换并得到积分链标准型, 从而支撑标称模型下的反馈线性化与稳定控制器设计.

同理, 对第 ℓ 个安全约束定义其相对阶 $r_{\mathrm{cbf},\,\ell}$, 并将高阶安全条件等价映射为输入空间的不等式约束. 由于安全约束通常定义在状态空间或其子空间上, 一般满足 $r_{\mathrm{cbf},\,\ell}\leq r_{\mathrm{sys}}$. 利用高阶李导数运算, 可将不确定增益 $g_{i,\,2}(x_i,\,\kappa_i)$ 中的参数 κ_i 显式分离, 得到输入仿射的线性形式:

$$b_\ell^{(r_{\mathrm{cbf},\,\ell})}=\Psi_{\ell,\,0}(x_i)+\Psi_{\ell,\,1}(x_i)u_{i,\,1}+\sum_{q=1}^{p}\kappa_{i,\,q}\,\Psi_{\ell,\,2,\,q}(x_i)\,u_{i,\,2} \tag{13}$$

其中, 各项系数均为当前状态 x_i 的函数, 具体由李导数给出:

$$\begin{cases}\Psi_{\ell,\,0}(x_i)=\mathcal{L}_{f_i}^{r_{\mathrm{cbf},\,\ell}}b_\ell(x_i)+\mathfrak{R}(x_i)\\ \Psi_{\ell,\,1}(x_i)=\mathcal{L}_{g_{i,\,1}}\mathcal{L}_{f_i}^{r_{\mathrm{cbf},\,\ell}-1}b_\ell(x_i)\\ \Psi_{\ell,\,2,\,q}(x_i)=\mathcal{L}_{\hat{g}_{i,\,2,\,q}}\mathcal{L}_{f_i}^{r_{\mathrm{cbf},\,\ell}-1}b_\ell(x_i)\end{cases} \tag{14}$$

式中 $\mathfrak{R}(x_i)$ 代表 HOCBF 递推过程中产生的低阶 α 函数项之和.

该表达式为后续在参数不确定条件下构造安全约束提供了直接基础.

2 分布式安全保证

多智能体系统中的安全约束既可来自个体状态限制, 也可来自交互约束. 以下以单体屏障函数 $b(x_i)$ 为统一形式, 推导其在非线性系统 (7) 下对应的输入空间显式约束, 并给出可认证的凸安全动作集表示.

2.1 HOCBF 约束构造

考虑第 i 个智能体的状态安全集

$$\mathcal{C}_i:=\{x_i\in\mathbf{R}^n\mid b(x_i)\geq 0\} \tag{15}$$

安全控制目标是在 $x_i(0)\in\mathcal{C}_i$ 时, 使得 $\forall t\geq 0$ 有 $x_i(t)\in\mathcal{C}_i$.

取可微扩展 $\mathcal{K}_\infty$ 类函数 $\{\alpha_\nu(\cdot)\}_{\nu=1}^{r_{\mathrm{cbf}}}$, 按定义 3 构造 $\psi_0(x_i)=b(x_i)$, $\psi_\nu(x_i)=\dot{\psi}_{\nu-1}(x_i)+\alpha_\nu(\psi_{\nu-1}(x_i))$. 为保证系统前向不变性, 控制输入 u_i 需满足如下 HOCBF 约束:

$$\mathcal{L}_{f_i}\psi_{r_{\mathrm{cbf}}-1}(x_i)+\sum_{j=1}^{m}\mathcal{L}_{g_{i,\,j}}\psi_{r_{\mathrm{cbf}}-1}(x_i)\;u_{i,\,j}+\alpha_{r_{\mathrm{cbf}}}(\psi_{r_{\mathrm{cbf}}-1}(x_i))\geq 0 \tag{16}$$

结合前文的非线性系统 (7) 与假设 2, 式 (16) 等价于输入的仿射不等式

$$\Psi_0+\Psi_1u_{i,\,1}+\sum_{q=1}^{p}\kappa_{i,\,q}\Psi_{2,\,q}\,u_{i,\,2}+\Psi_d(x_i,\,t)\geq 0 \tag{17}$$

其中 Ψ_0, Ψ_1, $\Psi_{2,\,q}$ 为李导数展开得到的状态相关系数; $\Psi_d(x_i,\,t)$ 表示扰动 $D_i(t)$ 在约束导数链中的有界投影项.

因此, 高阶约束通过相对阶导数链将输入显式分离, 这为后续构造可行域提供起点.

2.2 参数不确定性与扰动约束下的安全动作集

为保证约束在 $\kappa_i\in\Lambda_i$ 及有界扰动下均成立, 对 (17) 取最坏情形. 设扰动投影项满足下界

$$\Psi_d(x_i,\,t)\geq-\Delta_{\max}(x_i) \tag{18}$$

其中, $\Delta_{\max}(x_i)$ 可通过扰动上界与导数链相关函数的梯度范数进行保守估计. 具体地, 若扰动满足 $\|D_i(t)\|\leq\overline{D}_i$, 则由 Cauchy-Schwarz 不等式可取

$$\Delta_{\max}(x_i)=\overline{D}_i\;\|\nabla\psi_{r_{\mathrm{cbf}}-1}(x_i)\|$$

从而得到式 (18). 更一般地, 也可利用屏障函数及其高阶导数链对状态的 Lipschitz 常数构造保守界. 这种利用扰动上界与梯度范数 (或 Lipschitz 常数) 构造鲁棒安全余量的处理思路可参考具有扰动鲁棒性的 CBF[23].

于是, 针对参数与扰动不确定性的安全条件为

$$\Psi_0+\Psi_1u_{i,\,1}+\min_{\kappa_i\in\Lambda_i}\sum_{q=1}^{p}\kappa_{i,\,q}\Psi_{2,\,q}u_{i,\,2}-\Delta_{\max}(x_i)\geq 0 \tag{19}$$

由于 Λ_i 为凸多面体, 线性函数在该集合上的最小值在其顶点集处取得, 因此可认证的凸安全动作

集可表示为有限个线性不等式所确定的半空间的交集:

$$\mathcal{K}^i_{\text{safe}}(x_i) := \Big\{ u_i \in \mathcal{U}_u \;\Big|\; \Psi_0(x_i) + \Psi_1(x_i)\, u_{i,\,1} + \sum_{q=1}^{p} \vartheta^{(s)}_{i,\,q}\, \Psi_{2,\,q}(x_i)\, u_{i,\,2} - \Delta_{\max}(x_i) \geq 0,\ \ \forall s \in \{1,\,\cdots,\,L\} \Big\} \tag{20}$$

该集合为凸多面体, 将用于后续的 QP 投影/安全滤波器求解.

由此, 将状态域安全约束 $b(x_i) \geq 0$ 映射为有关输入的仿射不等式, 并在不确定参数与有界扰动下得到可认证的凸安全动作集 $\mathcal{K}^i_{\text{safe}}(x_i)$, 便于后续安全滤波与策略输出投影的实现.

对于多智能体避撞等依赖邻居状态的耦合约束 $b_{ij}(x_i,\, x_j) \geq 0$, 其高阶导数通常同时显含 u_i 与 u_j. 为实现分布式求解, 可以引入互惠系数 $\omega_{ij} \in (0,\, 1)$ 对安全余量进行责任分配[24], 并满足 $\omega_{ij} + \omega_{ji} = 1$. 其中, ω_{ij} 表示智能体 i 在成对安全约束 b_{ij} 中承担的安全责任份额, ω_{ij} 越大, 表示智能体 i 需留出更大的局部安全裕度. 当局部责任分配满足互补关系 $\omega_{ij} + \omega_{ji} = 1$ 时, 由两侧局部约束叠加得到的总安全余量与原耦合约束保持一致, 满足全局安全要求. 在无额外先验信息时, 可取对称分配 $\omega_{ij} = \omega_{ji} = 0.5$; 若已知不同智能体的机动能力或任务优先级等存在差异, 则可按要求进行责任份额分配.

3　基于 HOCBF 的在线安全滤波器

为将安全保证推广至无限时域, 本文设计一种基于 HOCBF 的在线安全滤波器. 为了克服 QP 求解可能出现的不可行 (infeasibility) 问题, 本文借鉴了文献 [18] 中安全滤波器的递归思想. 该滤波器在每个时刻维护有限时域的备份控制序列, 并采用滚动移位机制确保递归可行性.

3.1　终端控制器

假设在安全约束集 $\mathcal{C}$ 内存在一个非空的最大控制不变集 $\mathcal{S}_\infty \subseteq \mathcal{C}$, 满足对任意 $x_i \in \mathcal{S}_\infty$, 存在可行输入 $u \in \mathcal{U}_u$, 使得闭环状态始终保持于 $\mathcal{S}_\infty$ 内.

注 3. 在实际实现中, $\mathcal{S}_\infty$ 可由标称稳定控制器 (如 LQR (linear quadratic regulator)) 闭环系统的局部不变域近似表征, 例如由相应李雅普诺夫函数的某个水平集给出.

为补偿连续时刻在实际中离散化实现引起的近似误差, 并增强对数值计算偏差的鲁棒性, 引入安全裕度 $\varepsilon_{\mathcal{T}} > 0$, 定义内缩屏障函数

$$b_{\mathcal{T}}(x_i) := b(x_i) - \varepsilon_{\mathcal{T}} \tag{21}$$

记 $b_{\mathcal{T}}$ 对输入 u_i 的相对阶为 r_{cbf}, 取可微扩展 $\mathcal{K}_\infty$ 类函数 $\{\alpha_\nu(\cdot)\}_{\nu=1}^{r_{\text{cbf}}}$, 并按式 (5) 对 $b_{\mathcal{T}}$ 进行 HOCBF 递推构造, 从而得到 $\psi^{\mathcal{T}}_\nu$. 据此定义终端集为

$$\mathcal{T}_i := \mathcal{S}_\infty \;\cap\; \bigcap_{\nu=0}^{r_{\text{cbf}}-1} \Big\{ x_i \in \mathbf{R}^n \;\big|\; \psi^{\mathcal{T}}_\nu(x_i) \geq 0 \Big\} \tag{22}$$

其中, $\bigcap_{\nu=0}^{r_{\text{cbf}}-1}$ 表示各个 ν 对应的集合取交集. 由 $b_{\mathcal{T}}(x_i) \geq 0 \Rightarrow b(x_i) \geq \varepsilon_{\mathcal{T}}$ 可知, $\mathcal{T}_i$ 相较于原安全集, 引入了裕度内缩.

为保证系统进入终端集 $\mathcal{T}_i$ 后能够长期驻留, 本文定义一个能够保持该集合前向不变的终端控制器, 记为 $\pi^{\mathcal{T}}_i(x_i)$. 其物理意义为有限时域备份序列末端的"安全锚点": 一旦系统进入终端集, 该控制器即接管控制权, 通过维持集合的前向不变性, 将有限时域的安全规划扩展为无限时域的安全保证.

定义 4 (终端控制器). 对任意 $x_i \in \mathcal{T}_i$, 终端控制器通过求解如下二次规划问题 (HOCBF-QP) 得出:

$$\begin{aligned} \pi^{\mathcal{T}}_i(x_i) = \arg\min_{u \in \mathcal{U}_u} \ & \|u - u^{\text{nom}}_i(x_i)\|^2 \\ \text{s.t.} \quad & u \in \mathcal{K}^i_{\text{safe},\,\mathcal{T}}(x_i) \end{aligned} \tag{23}$$

其中 $\mathcal{K}^i_{\text{safe},\,\mathcal{T}}(x_i)$ 表示将第 2 节安全动作集构造式中的 b 替换为内缩后的 $b_{\mathcal{T}}$, 从而得到的 HOCBF 可行输入集合; $u^{\text{nom}}_i(x_i)$ 为第 i 个智能体在状态 x_i 下的标称控制输入, 由基于反馈线性化的 LQR 标称控制器给出.

将式 (21) 定义的内缩屏障函数 $b_{\mathcal{T}}$ 代替原屏障函数 b, 按照第 2 节的方法重新构造 HOCBF 约束及安全动作集, 可得到终端安全动作集 $\mathcal{K}^i_{\text{safe},\,\mathcal{T}}(x_i)$ 对应的系数 $\Psi^{\mathcal{T}}_0(x_i)$、$\Psi^{\mathcal{T}}_1(x_i)$、$\Psi^{\mathcal{T}}_{2,\,q}(x_i)$ 以及扰动保守界 $\Delta^{\mathcal{T}}_{\max}(x_i)$. 对每个顶点 $s \in \{1,\,\cdots,\,L\}$, 定义

$$S^{\mathcal{T}}_s(x_i) := \max_{u \in \mathcal{U}_u} \Big(\Psi^{\mathcal{T}}_1(x_i) u_{i,\,1} + \sum_{q=1}^{p} \vartheta^{(s)}_{i,\,q} \Psi^{\mathcal{T}}_{2,\,q}(x_i) u_{i,\,2} \Big) \tag{24}$$

由于目标是线性的, 约束集合 $\mathcal{U}_u$ 为盒约束, 故 $S^{\mathcal{T}}_s(x_i)$ 可显式计算.

进一步说明, $\pi^{\mathcal{T}}_i(x_i)$ 由一个严格凸二次规划给出, 且约束函数关于状态变量连续可微. 因此, 在终端 QP 可行并满足线性无关约束资格条件时, 其最优解映射关于状态 x_i 是局部 Lipschitz 连续的. 类似地, CBF-QP 控制器正则性讨论可参考文献 [25].

定理 1 (终端 QP 可行的充分条件). 对任意 $x_i \in \mathcal{T}_i$, 若

$$\min_{s\in\{1,\cdots,L\}}\left(S_s^{\mathcal{T}}(x_i)+\Psi_0^{\mathcal{T}}(x_i)-\Delta_{\max}^{\mathcal{T}}(x_i)\right)\geq 0 \quad (25)$$

则 $\mathcal{K}^i_{\text{safe},\,\mathcal{T}}(x_i)\neq\emptyset$, 从而终端 QP (23) 可行.

引理 1 (HOCBF 前向不变性[22]). 考虑多输入仿射系统 $\dot{x}=f(x)+\sum_{j=1}^m g_j(x)u_j$. 设 $b(x)$ 相对输入 u 的相对阶为 r, 并按式 (5) 构造 $\{\psi_\nu(x)\}_{\nu=0}^{r-1}$, 定义 $\mathcal{C}:=\bigcap_{\nu=0}^{r-1}\{x\mid\psi_\nu(x)\geq 0\}$. 在动力学局部 Lipschitz 且控制律 $u(t)$ 局部 Lipschitz 连续的前提下, 若闭环输入始终满足 HOCBF 给出的可行输入约束 (见式 (16)), 则集合 $\mathcal{C}$ 前向不变.

引理 2 (终端集不变性). 若系统向量场 $f_i(\cdot)$ 与 $g_i(\cdot)$ 均为局部 Lipschitz 连续, 且内缩屏障函数 $b_{\mathcal{T}}$ 相对 u_i 具有相对阶 r_{cbf}, 同时闭环控制律 $u_i(t)=\pi_i^{\mathcal{T}}(x_i(t))$ 满足局部 Lipschitz 条件, 并始终满足 $u_i(t)\in\mathcal{K}^i_{\text{safe},\,\mathcal{T}}(x_i(t))$ 以及 $x_i(t)\in\mathcal{S}_\infty$, 则终端集 $\mathcal{T}_i$ 前向不变.

证明. 由式 (22) 可得

$$\mathcal{T}_i=\mathcal{S}_\infty\cap\mathcal{C}_{\mathcal{T}},\qquad \mathcal{C}_{\mathcal{T}}:=\bigcap_{\nu=0}^{r_{\text{cbf}}-1}\{x_i\mid\psi_\nu^{\mathcal{T}}(x_i)\geq 0\}$$

若 $x_i(t_0)\in\mathcal{T}_i$, 则有 $x_i(t_0)\in\mathcal{C}_{\mathcal{T}}$ 且 $x_i(t_0)\in\mathcal{S}_\infty$. 又由于终端控制器满足 $u_i(t)\in\mathcal{K}^i_{\text{safe},\,\mathcal{T}}(x_i(t))$, 而 $\mathcal{K}^i_{\text{safe},\,\mathcal{T}}(x_i)$ 是由终端屏障函数 $b_{\mathcal{T}}$ 构造的 HOCBF 可行输入集. 故由引理 1 可知, 集合 $\mathcal{C}_{\mathcal{T}}$ 前向不变, 即对任意 $\tau\geq t_0$, 有 $x_i(\tau)\in\mathcal{C}_{\mathcal{T}}$. 另外, 由 $x_i(t)\in\mathcal{S}_\infty$ 的定义可知, 在终端控制器作用下闭环轨迹保持于 $\mathcal{S}_\infty$ 内. 故对任意 $\tau\geq t_0$, 均有

$$x_i(\tau)\in\mathcal{S}_\infty\cap\mathcal{C}_{\mathcal{T}}=\mathcal{T}_i$$

因此终端集 $\mathcal{T}_i$ 前向不变. □

3.2 备份策略

为实现递归安全, 本文提出一种基于“尝试–回退”机制的在线安全滤波器. 该方法的核心在于维护一条能够安全回退至终端集 $\mathcal{T}_i$ 的轨迹, 即备份策略.

定义 5 (备份策略). 给定步长 H, 若控制序列 $\Pi_i^B(t)=\{u^B_{i,\,t|t},\,u^B_{i,\,t+1|t},\,\cdots,\,u^B_{i,\,t+H-1|t}\}$ 及其产生的预测轨迹同时满足以下约束:

$$\begin{cases}u^B_{i,\,t+k|t}\in\mathcal{U}_u, & \forall k\in\{0,\,\cdots,\,H-1\}\\ u^B_{i,\,t+k|t}\in\mathcal{K}^i_{\text{safe}}(x_{i,\,t+k|t}), & \forall k\in\{0,\,\cdots,\,H-1\}\\ x_{i,\,t+H|t}\in\mathcal{T}_i\end{cases} \quad (26)$$

则称 $\Pi_i^B(t)$ 为一个备份策略.

注 4. 预测步长 H 应根据系统最大制动能力选取, 以确保系统能从工作区边界回归至终端集.

基于上述定义, 在线安全滤波器的核心任务是求解备份可行性问题 $\mathcal{P}_B(t;u_i^{\text{task}})$, 即验证以当前任务动作 u_i^{task} 为首元素的备份策略是否存在. 问题 $\mathcal{P}_B$ 可行当且仅当满足如下集合非空条件:

$$\left\{\text{备份策略 }\Pi_i^B(t)\;\middle|\;u^B_{i,\,t|t}=u_i^{\text{task}}\right\}\neq\emptyset \quad (27)$$

若式 (27) 成立, 说明任务动作 u_i^{task} 安全, 可直接执行; 否则, 滤波器将触发回退机制 (见算法 1), 利用上一时刻的备份策略保证系统安全.

3.3 在线滤波流程

具体而言, 滤波器采用滚动时域机制执行: 尝试求解 $\mathcal{P}_B(t;u_i^{\text{task}})$. 若问题有解, 说明任务动作安全, 直接执行并更新备份; 若问题无解, 则触发回退机制, 利用上一时刻备份序列的移位 (shift) 结果作为保底策略. 详细流程如算法 1 所示.

算法 1. 在线安全滤波器 $\phi(x_{i,\,t},\,u_i^{\text{task}})$

输入. 当前状态 $x_{i,\,t}$, 任务动作 u_i^{task}, 上一时刻备份序列 $\Pi_i^B(t-1)$, 预测时域 H.

输出. 安全动作 $u_i^*(t)$, 更新备份序列 $\Pi_i^B(t)$.

1: 求解 $\mathcal{P}_B(t;u_i^{\text{task}})$
2: **if** 可行 **then**
3: 更新备份序列 $\Pi_i^B(t)\leftarrow\Pi_i^B(t)\big|_{\mathcal{P}_B(t;u_i^{\text{task}})}$
4: $u_i^*(t)\leftarrow u_i^{\text{task}}$
5: **else**
6: $\Pi_i^B(t)\leftarrow\{u^B_{i,\,t|t-1},\,\cdots,\,u^B_{i,\,t+H-2|t-1},\,\pi_i^{\mathcal{T}}(x_{i,\,t+H-1|t-1})\}$
7: $u_i^*(t)\leftarrow u^B_{i,\,t|t-1}$
8: **end if**

由算法 1 可得, 实际执行输入 $u_i^*(t)$ 始终满足 HOCBF 可行输入约束; 结合引理 1, 可得安全集在无限时域内前向不变.

同时, 备份序列在滚动移位末端接入 $\pi_i^{\mathcal{T}}$, 并通过末端约束 $x_{i,\,t+H|t}\in\mathcal{T}_i$ 保证轨迹可到达终端集; 由引理 2 可知, 一旦到达 $\mathcal{T}_i$ 即可持续驻留, 从而避免后续时刻安全滤波问题因不存在满足约束的控制序列而不可行, 并提供了可用的安全回退通道.

4 基于相对阶坐标变换的多智能体安全强化学习

为解决高维非线性系统下的多智能体协同控制问题, 本文提出一种安全强化学习框架. 该框架以独立近端策略优化 (independent proximal policy optimization, IPPO) 为基础, 通过策略分解降低探

索难度, 并结合在线安全滤波器保证训练的安全性.

4.1　参数共享的多智能体学习架构

本文采用参数共享的 IPPO 算法. 所有智能体共用一套 Actor 网络参数 θ 与 Critic 网络参数 ω, 显著减少了训练时的参数总量. 在时刻 t, 智能体 i 基于局部观测 $o_{i,\,t}$, 利用共享策略网络采样动作补偿量, 并利用共享价值网络评估当前状态价值:

$$\Delta u_{i,\,t} \sim \pi_\theta(\cdot \mid o_{i,\,t}), \qquad V_{i,\,t} = V_\omega(o_{i,\,t}) \tag{28}$$

其中, $\Delta u_{i,\,t}$ 为智能体 i 在时刻 t 采样的动作补偿量; $\pi_\theta(\cdot \mid o_{i,\,t})$ 为共享 Actor 网络在局部观测 $o_{i,\,t}$ 条件下给出的补偿动作概率分布; $V_{i,\,t}$ 为对应的价值估计; $V_\omega(o_{i,\,t})$ 为共享 Critic 网络输出的价值估计. 符号“$\sim$”表示按该概率分布采样.

算法采用 PPO-Clip 机制限制策略更新幅度, 以保证训练过程的稳定性.

4.2　策略分解

为了降低纯数据驱动学习在复杂动力学下的探索难度, 本文利用式 (11) 中的输入结构特性, 在学习侧引入相对阶策略分解机制, 在相对阶坐标下输出导数链满足式 (11).

该结构的意义在于: 在已知的可建模通道上 (对应 $\mathcal{B}_{i,\,1}$), 系统可在局部等价为积分链标准型, 从而允许基于局部线性模型构造稳定的标称控制骨架 (如反馈线性化). 同时, 将难以精确建模的部分、未建模动态与不确定通道的影响统一留给数据驱动补偿学习. 由此得到策略分解

$$u_i = \begin{bmatrix} u_{i,\,1} \\ u_{i,\,2} \end{bmatrix} = \begin{bmatrix} u_{i,\,1}^{\mathrm{nom}}(\xi_i) + u_{i,\,1}^{\mathrm{res}}(o_i;\theta) \\ u_{i,\,2}^{\mathrm{RL}}(o_i;\theta) \end{bmatrix} \tag{29}$$

其中 $u_{i,\,1}^{\mathrm{nom}}$ 由基于反馈线性化的 LQR 标称控制器给出; 残差项 $u_{i,\,1}^{\mathrm{res}}$ 用于补偿局部线性化误差等; 不确定通道 $u_{i,\,2}^{\mathrm{RL}}$ 由策略直接生成, 以降低参数失配导致的风险.

4.3　安全训练机制

训练时 Actor 仅输出补偿量而非完整控制:

$$\Delta u_{i,\,t} = \begin{bmatrix} u_{i,\,1}^{\mathrm{res}}(o_{i,\,t};\theta) \\ u_{i,\,2}^{\mathrm{RL}}(o_{i,\,t};\theta) \end{bmatrix}, \ \Delta u_{i,\,t} \sim \pi_\theta(\cdot \mid o_{i,\,t}) \tag{30}$$

与标称骨架叠加得到任务动作

$$u_{i,\,t}^{\mathrm{task}} = \begin{bmatrix} u_{i,\,1}^{\mathrm{nom}}(x_{i,\,t}) \\ 0 \end{bmatrix} + \Delta u_{i,\,t} \tag{31}$$

再经在线安全滤波器投影/回退为实际执行动作

$$u_{i,\,t}^{*} = \phi(x_{i,\,t},\ u_{i,\,t}^{\mathrm{task}}) \in \mathcal{K}_{\mathrm{safe}}^{i}(x_{i,\,t}) \tag{32}$$

PPO 的动作样本为 $\Delta u_{i,\,t}$ (用于 $\ln \pi_\theta(\Delta u_{i,\,t} | o_{i,\,t})$), 环境执行为 $u_{i,\,t}^{*}$.

由此形成“标称骨架保证稳定、学习补偿性能细节、安全滤波器保证可行性”的闭环学习机制. 在提高学习效率的同时, 基于本文所述条件, 该机制保障闭环系统在训练与执行过程中满足所设定的安全约束.

5　仿真实验与分析

5.1　仿真环境建模

为验证算法的有效性, 仿真考虑由 $N = 4$ 颗微型卫星组成的分布式空间干涉测量任务场景. 如图 1 所示, 在该场景中, 卫星编队需在保持链式领航–跟随的同时, 通过姿态协同控制来维持光学干涉基线的稳定性并且满足光学载荷避光约束.

定义智能体 i 的状态为

$$x_i = \begin{bmatrix} \Theta_i^{\mathrm{T}} & (\omega_i^{\mathrm{body}})^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \tag{33}$$

其中 $\Theta_i = [\phi_i^{\mathrm{att}},\ \theta_i^{\mathrm{att}},\ \psi_i^{\mathrm{att}}]^{\mathrm{T}}$ 为欧拉角, $\omega_i^{\mathrm{body}} = [p_i^{\mathrm{body}},\ q_i^{\mathrm{body}},\ r_i^{\mathrm{body}}]^{\mathrm{T}}$ 为机体角速度. ϕ_i^{att}, θ_i^{att}, ψ_i^{att} 分别为滚转角, 俯仰角和偏航角; p_i^{body}, q_i^{body}, r_i^{body} 分别为绕机体 x, y, z 轴的角速度分量.

姿态动力学为

$$\dot{\Theta}_i = R(\Theta_i)\omega_i^{\mathrm{body}} \tag{34}$$

$$J_i \dot{\omega}_i^{\mathrm{body}} = -\omega_i^{\mathrm{body}} \times (J_i \omega_i^{\mathrm{body}}) + \tau_i + d_i(t) \tag{35}$$

其中, $R(\Theta_i)$ 为欧拉角 Θ_i 决定的姿态运动学变换矩阵; $J_i = \mathrm{diag}\{4.0,\ 3.5,\ 2.5\}\ \mathrm{kg \cdot m^2}$; $d_i(t)$ 为有界外部扰动, 用于模拟重力梯度力矩与残余气动干扰.

为与式 (7) 中不确定增益形式一致, 将控制力矩写为含参数不确定的执行器效率映射

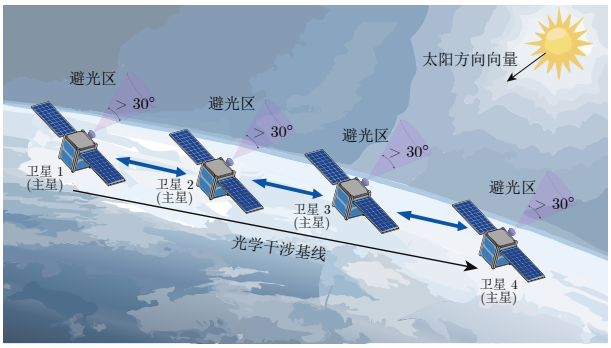


图 1　卫星拓扑图

Fig. 1　Satellite topology diagram

$$\tau_i = \begin{bmatrix} u_{i,\,x} \\ u_{i,\,y} \\ \kappa_i\, u_{i,\,z} \end{bmatrix} \tag{36}$$

其中仅偏航通道存在乘性不确定性, 且 $u_{i,\,2} = u_{i,\,z} \in \mathbf{R}$ 为标量输入通道的特例. 仿真取

$$\kappa_i \sim \mathrm{U}[0.65,\ 0.75] \tag{37}$$

其中, U[0.65, 0.75] 表示区间 [0.65, 0.75] 上的均匀分布. 扰动满足无穷范数界

$$\|d_i(t)\|_\infty \le 2 \times 10^{-3}\ \mathrm{N \cdot m} \tag{38}$$

选取性能输出为姿态误差 $y_i = \Theta_i - \Theta_{\mathrm{ref}}$, 三个姿态输出分量相对于控制输入的相对阶均为 2, 因此系统的向量相对阶为 $r = [2,\ 2,\ 2]^{\mathrm{T}}$, 从而可在相对阶标准型下构造 LQR 标称控制并进行策略分解. 其中 Θ_{ref} 为固定参考姿态.

5.2 安全约束和训练设置

实验考虑光学载荷避光约束. 设太阳方向向量在惯性系下为 v_{sun}, 则其在机体系下为 $v_{\mathrm{sun}}^B = R^{\mathrm{T}}(\Theta_i) v_{\mathrm{sun}}$. 要求太阳方向向量与机体 z 轴单位向量 e_3 的夹角不小于 $\gamma_{\min} = 30°$, 定义安全函数

$$b_{\mathrm{sun}}(x_i) = \cos(\gamma_{\min}) - e_3^{\mathrm{T}} v_{\mathrm{sun}}^B(x_i) \ge 0 \tag{39}$$

该约束为相对阶 $r = 2$ 的高阶状态约束, 通过安全滤波器保证可行性.

仿真训练任务为: 领航卫星 (Sat 1) 追踪固定参考姿态 Θ_{ref} 并执行凝视任务; 编队同步卫星 (Sat i, $i = 2,\ 3,\ 4$) 仅观测并追踪前一序位卫星姿态, 实现全队同步与基线对齐. 仿真与训练相关的关键参数汇总于表 1.

5.3 仿真结果与分析

为验证所提方法的有效性, 本文从任务性能、安全性能、计算时效性与鲁棒性四个维度开展仿真分析. 选取以下方法作为对比基线: 纯 PPO、单步 CBF-RL (control barrier function-based reinforcement learning) (采用文献 [17] 中的单步投影式安全机制)、LQR + 单步投影安全滤波器 (作为基于模型的参考基线, 简写为 LQR + 安全滤波器).

图 2 显示本文方法在训练初期奖励值更高且收敛更快. 这得益于相对阶解耦下的标称控制骨架提供了稳定“热启动”, 降低了随机探索的发散风险.

进一步地, 图 3 给出评估阶段的平均绝对误差的变化趋势, 表明引入 RL 残差项能有效补偿未建模动态与参数失配引起的偏差. 相比之下, 纯 PPO 方法收敛缓慢, 而 LQR + 安全滤波器方法因缺乏在线补偿能力, 跟踪精度受限. 结果表明, 策略分解机制在保障姿态跟踪精度的同时显著提升了学习效率.

图 4 给出了各方法的安全约束违规率对比结果. 可以看出, 在无安全滤波器时, 违规率呈现频繁的尖峰并长期维持在较高水平, 说明策略在探索阶

表 1 训练超参数设置

Table 1 Training hyperparameter settings

参数	符号	取值
智能体数量	N	4
网络结构	—	2 层隐藏层 (每层 256)
激活函数	—	tanh
姿态追踪权重	w_{track}	8.0
角速度惩罚	w_{rate}	0.05
优化器	—	Adam
Actor 学习率	—	3×10^{-5}
Critic 学习率	—	5×10^{-4}
相对熵早停阈值	—	0.015
预测步长	H	6

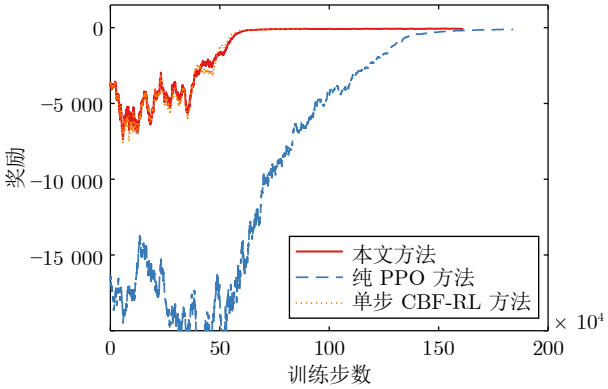


图 2 不同方法的训练奖励曲线

Fig. 2 Training reward curves of different methods

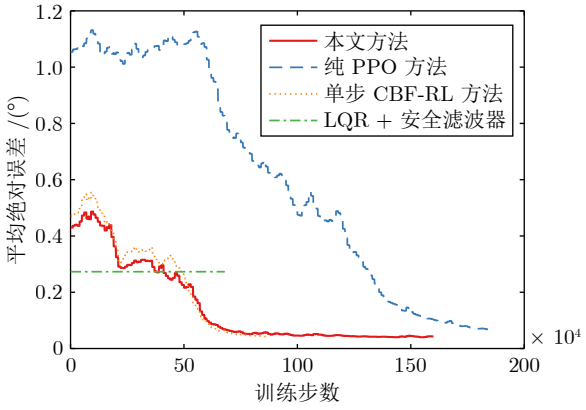


图 3 不同方法的评估误差曲线

Fig. 3 Evaluation error curves of different methods

段容易触发硬约束违规. 而本文方法的违规率始终接近于零, 表明在线安全滤波器能够有效抑制训练与执行过程中的不安全动作.

为进一步增强安全评估的全面性, 图 5 给出了各方法最小安全距离 $h_{\min}$ 的统计结果. 可以看出, 本文方法的 $h_{\min}$ 始终保持为正, 且整体波动较小; 其余方法均出现不同程度的 $h_{\min}<0$ 情形, 说明其在任务过程中仍存在约束侵入现象. 该结果表明, 所提方法在维持安全性上具有更稳定的表现.

图 6 进一步给出了各方法最大侵入时间的统计结果. 可以看到, 本文方法在整个训练过程中最大侵入时间始终保持为零, 而其他方法均存在不同程度的持续违规时段. 尤其是单步 CBF-RL, 虽然

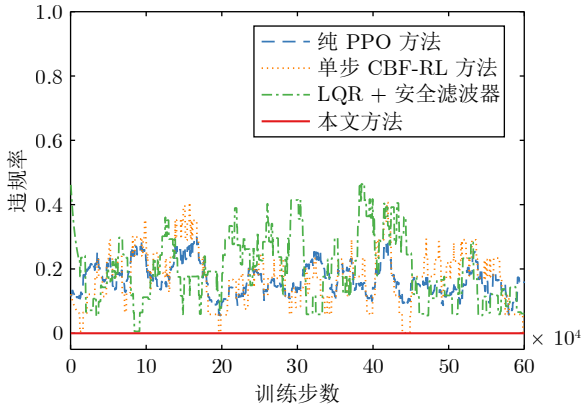
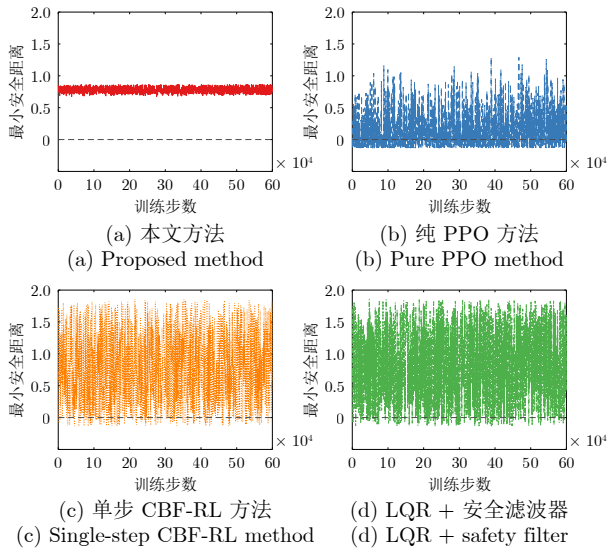


图 4　不同方法的安全约束违规率对比

Fig. 4　Comparison of safety-constraint violation rate for different methods



(a) 本文方法　(b) 纯 PPO 方法

(a) Proposed method　(b) Pure PPO method

(c) 单步 CBF-RL 方法　(d) LQR + 安全滤波器

(c) Single-step CBF-RL method　(d) LQR + safety filter

图 5　不同方法的最小安全距离对比

Fig. 5　Comparison of minimum safety distance for different methods

能够降低部分瞬时违规, 但仍无法完全避免连续侵入现象. 该结果表明, 本文方法不仅能够有效降低违规率, 而且在更严格的安全性指标上同样表现出明显优势, 从而验证了递推安全滤波机制在长期安全维持方面的有效性.

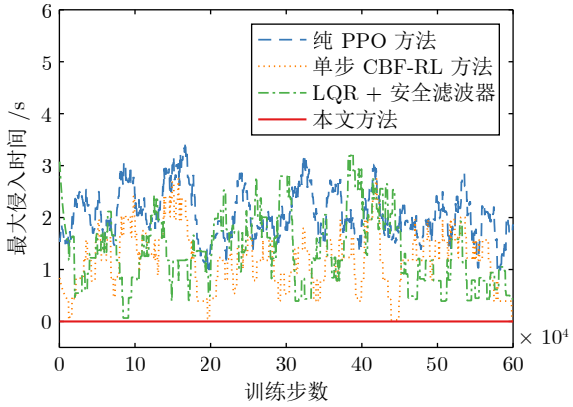


图 6　不同方法的最大侵入时间对比

Fig. 6　Comparison of maximum intrusion time for different methods

在线安全滤波器需要在每个控制周期内维护滚动备份序列并求解安全投影问题, 因此其实时计算代价是方法可实施性的重要指标. 仿真实验均在一台配备 Intel Core i7-13700H CPU、16 GB 内存的计算机上进行. 统计结果显示, 本文方法平均单步耗时约 3.0 ms, 虽略高于单步 CBF-RL 方法的 2.2 ms 和纯 PPO 方法的 1.7 ms, 但远小于 50.0 ms 的控制周期, 满足实时性要求.

此外, 在扰动强度放大至标称值的 20 倍、执行器效率不确定性范围扩大至标称条件下的 6 倍时, 所提方法的平均绝对误差分别仅增加约 1.0% 和 2.5%, 且平均绝对误差始终保持在 2.3° 以内, 展现出较强的鲁棒性.

综上所述, 实验结果表明: 策略分解机制显著提升了训练初期的稳定性与收敛速度; 在线安全滤波器有效减少了安全违规的现象. 两者结合使得系统在存在模型不确定性与外部扰动的条件下, 仍能够同时兼顾高效学习、严格安全保证与满足实时需求的在线控制.

6　结束语

本文面向包含不确定扰动与高阶硬约束的多智能体系统, 给出一套“学习–执行”闭环方案: 执行端采用在线安全滤波器, 通过备份序列、末端约束及滚动移位确保递归可行性; 学习端采用基于相对阶坐标变换的策略分解, Actor 仅输出补偿量, 并由滤波器生成实际执行动作.

仿真结果显示: 策略分解提升了训练初期稳定性与收敛速度, 安全滤波器将探索阶段的约束违规率降至接近零. 两者的结合使执行器效率不确定与高阶安全约束并存时仍能保持较低误差与稳定性能.

需要指出的是, 本文方法依赖于相对阶可定义以及终端集可构造等前提, 因此在更强耦合动力学和更高维场景下存在一定的保守性与计算负担. 后续工作将聚焦两方面: 其一, 系统地分析备份时域与终端裕度选取中性能与保守性之间的权衡; 其二, 面向更强耦合动力学与通信受限的场景, 进一步研究滤波器与分布式学习在异步信息下的鲁棒性与可扩展性.

参考文献

1 Jiang Bi-Tao, Wen Guang-Hui, Zhou Jia-Ling, Zheng De-Zhi. Cross-domain cooperative technology of intelligent unmanned swarm systems: Current status and prospects. *Strategic Study of CAE*, 2024, **26**(1): 117−126
(江碧涛, 温广辉, 周佳玲, 郑德智. 智能无人集群系统跨域协同技术研究现状与展望. 中国工程科学, 2024, **26**(1): 117−126)

2 Zhang J, Yang D S, Li W H, Zhang H G, Li G D, Gu P. Resilient output control of multiagent systems with DoS attacks and actuator faults: Fully distributed event-triggered approach. *IEEE Transactions on Cybernetics*, 2024, **54**(12): 7681−7690

3 Zhou P P, Chen B M. Distributed optimal solutions for multiagent pursuit-evasion games for capture and formation control. *IEEE Transactions on Industrial Electronics*, 2024, **71**(5): 5224−5234

4 Zhao R, Li Y, Gao F, Gao Z H, Zhang T Y. Multi-agent constrained policy optimization for conflict-free management of connected autonomous vehicles at unsignalized intersections. *IEEE Transactions on Intelligent Transportation Systems*, 2024, **25**(6): 5374−5388

5 Guo R, Feng J W, Zhao Y, Huang T W, Liu X Z, Wang J Y. Robust model predictive control for perturbed nonlinear multiagent systems via dynamic event-triggered scheme. *Science China Information Sciences*, 2026, **69**(4): Article No. 142205

6 Wen G H, Fu J J, Wang B W, Wang X D, Shen H. Robust formation tracking of ASVs with nonconservative collision avoidance: A control barrier function-based approach. *IEEE Transactions on Industrial Electronics*, 2026, **73**(6): 9174−9186

7 Wen G H, Yu W W, Xia Y Q, Yu X H, Hu J Q. Distributed tracking of nonlinear multiagent systems under directed switching topology: An observer-based protocol. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017, **47**(5): 869−881

8 Li Y H, Xie G M, Lu Z Q. Revisiting cooperative off-policy multi-agent reinforcement learning. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. Article No. 1439

9 Mai Y X, Yin Q Y, Ni W C, Xu P, Huang K Q. Constructive conflict-driven multi-agent reinforcement learning for strategic diversity. In: Proceedings of the 34th International Joint Conference on Artificial Intelligence. Montreal, Canada: ijcai.org, 2025. Article No. 24

10 Li D P, Lou N, Xu Z W, Zhang B, Fan G L. Efficient communication in multi-agent reinforcement learning with implicit consensus generation. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 23240−23248

11 Tessler C, Mankowitz D J, Mannor S. Reward constrained policy optimization. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net, 2019.

12 Fulton N, Platzer A. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press, 2018. 4485−4491

13 Fisac J F, Lugovoy N F, Rubies-Royo V, Ghosh S, Tomlin C J. Bridging Hamilton-Jacobi safety analysis and reinforcement learning. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE, 2019. 8550−8556

14 Polymenakos K, Abate A, Roberts S. Safe policy search using Gaussian process models. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. Montreal, Canada: International Foundation for Autonomous Agents and Multiagent Systems, 2019. 1565−1573

15 Cheng R, Khojasteh M J, Ames A D, Burdick J W. Safe multi-agent interaction through robust control barrier functions with learned uncertainties. In: Proceedings of the 59th IEEE Conference on Decision and Control (CDC). Jeju, Korea (South): IEEE, 2020. 777−783

16 Qin Z Y, Zhang K Q, Chen Y X, Chen J K, Fan C C. Learning safe multi-agent control with decentralized neural barrier certificates. In: Proceedings of the 9th International Conference on Learning Representations. Austria: OpenReview.net, 2021.

17 Ahmad H M S, Sabouni E, Wasilkoff A, Budhraja P, Guo Z J, Zhang S Y, et al. HMARL-CBF-hierarchical multi-agent reinforcement learning with control barrier functions for safety-critical autonomous systems. In: Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego, USA: 2025.

18 Hsu K C, Nguyen D P, Fisac J F. ISAACS: Iterative soft adversarial actor-critic for safety. In: Proceedings of the 5th Learning for Dynamics and Control Conference. Philadelphia, USA: PMLR, 2023. 90−103

19 Chung S J, Bandyopadhyay S, Foust R, Subramanian G P, Hadaegh F Y. Review of formation flying and constellation missions using nanosatellites. *Journal of Spacecraft and Rockets*, 2016, **53**(3): 567−578

20 Boumeftah A, Ben Yahia O, Frigon J F, Falco G, Kurt G K. Adaptive detection of on-orbit jamming for securing GEO satellite links. In: Proceedings of the 12th Advanced Satellite Multimedia Systems Conference and the 18th Signal Processing for Space Communications Workshop (ASMS/SPSC). Sitges, Spain: IEEE, 2025. 1−8

21 Isidori A. Exact linearization methods. *Nonlinear Control Systems: An Introduction*. London: Springer, 1995. 178−253

22 Xiao W, Belta C. High-order control barrier functions. *IEEE Transactions on Automatic Control*, 2022, **67**(7): 3655−3662

23 Ames A D, Xu X R, Grizzle J W, Tabuada P. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 2017, **62**(8): 3861−3876

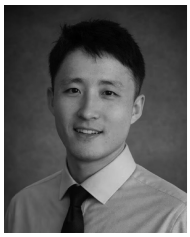
24 van den Berg J, Lin M, Manocha D. Reciprocal velocity obstacles for real-time multi-agent navigation. In: Proceedings of the IEEE International Conference on Robotics and Automa-

tion. Pasadena, USA: IEEE, 2008. 1928−1935

25 van Wijk D E J, Coogan S, Molnar T G, Majji M, Hobbs K L. Disturbance-robust backup control barrier functions: Safety under uncertain dynamics. *IEEE Control Systems Letters*, 2024, **8**: 2817−2822



黄 峁 湖北珞珈实验室副教授. 2017 年获得北京大学力学系统与控制专业博士学位. 主要研究方向为航天器智能姿态与轨道控制, 航天器编队控制, 航天器故障诊断及容错控制. 本文通信作者.
E-mail: dhuang@whu.edu.cn
(**HUANG Di** Associate professor at the Hubei Luojia Laboratory. He received his Ph.D. degree in mechanical systems and control from Peking University in 2017. His research interests include spacecraft intelligent attitude and orbit control, spacecraft formation control, and fault diagnosis and fault-tolerant control of spacecraft. Corresponding author of this paper.)



张斯伦 瑞典皇家理工学院数学系助理教授. 2019 年获得瑞典皇家理工学院优化与系统理论专业博士学位. 主要研究方向为非线性控制, 大规模复杂系统, 激励驱动算法和风险规避/安全强化学习. E-mail: silunz@kth.se
(**ZHANG Si-Lun** Assistant professor in the Department of Mathematics, KTH Royal Institute of Technology. He received his Ph.D. degree in optimization and systems theory from KTH Royal Institute of Technology in 2019. His research interests include nonlinear control, large-scale complex systems, incentive-driven algorithms, and risk-averse/safe reinforcement learning.)