arXiv:2602.07335v1 [eess.SY] 7 Feb 2026

面向航天器近距离操作中鲁棒且非贪心控制屏障函数的元强化学习

Minduli C. Wijayatunga[1,*], Richard Linares[2], Roberto Armellin[3]

1 伊利诺伊大学厄巴纳-香槟分校，美国伊利诺伊州厄巴纳 2 麻省理工学院，美国马萨诸塞州剑桥 3 奥克兰大学，新西兰奥克兰

*** 通讯作者：minduli@illinois.edu**

摘要

自主航天器巡检和对接任务需要能够在推力约束和不确定性下保证安全的控制器。输入约束控制障碍函数（ICCBF）为有界驱动条件下的安全认证提供了一种框架；然而，传统输入约束控制障碍函数表述可能过于保守，且对不确定性的鲁棒性有限，导致燃料消耗较高、任务可行性降低。本文提出一种框架，将定义 ICCBF 递归的完整 class-$\mathcal{K}$ 函数层级参数化并进行学习，从而能够对安全集进行局部塑形并降低保守性。通过微分代数高效计算控制裕度，使学习得到的连续时间输入约束控制屏障函数能够应用于航天器近距离操作中常见的时间采样动力系统。本文开发了一种元强化学习方案，用于训练策略，使其能够针对一组隐藏的物理参数和不确定性生成 ICCBF 参数，并采用多层感知机（MLP）和循环神经网络（RNN）架构。在巡航控制、航天器巡检和对接场景中的仿真结果表明，与固定的 class-$\mathcal{K}$ ICCBF 相比，所提出的方法能够在保持安全的同时降低燃料消耗并提高可行性；在更复杂的巡检案例中，RNN 尤具优势。

关键词

控制屏障函数、元强化学习、时间采样系统、航天器近距离操作

1 引言

航天器交会与近距离操作（RPO）必须满足严格的安全约束，同时受推力和星载计算限制。它们需要能够在不牺牲燃料效率、时间效率或不确定性鲁棒性的前提下维持安全的控制器，尤其是在自主运行时。在传统控制架构下，实现这种平衡仍然颇具挑战，因为这类架构往往将规划与安全保障分开，并且在实际飞行执行约束下只能提供有限的保证。

控制障碍函数（CBF）通过确保指定安全集的正向不变性，提供了一种基于集合论的安全认证方法；通常借助实时二次规划（CBF–QP）实现，仅对标称指令 [30] 进行最小幅度的调整。这种结构颇适合自主交会与近距离操作，因为它能够提供安全保证且具有计算可处理性。然而，交会与近距离操作的若干特点限制了传统控制障碍函数二次规划安全滤波器的有效性。

(1)相对阶限制。 许多交会与近距离操作约束，例如接近锥、视线（LOS）约束、禁入区（KOZ）和限入区（KIZ），都属于几何约束，且仅取决于相对位置。在这些任务所涉及的有效二阶平移动力学下，对于此类障碍函数，控制不会出现在障碍函数的一阶时间导数中，因此 $L_g h \equiv 0$。因此，标准的一阶 CBF–QP 无法在没有额外构造的情况下直接塑造轨迹演化。

(2)推力约束。 传统的 CBF 公式不会考虑推力约束。因此，在施加推力限制时，由障碍函数定义的标称安全集可能无法达到，因为可能存在一些状态，虽然满足障碍函数约束，却没有任何可容许控制输入能够维持其正向不变性。

(3)时间采样实现。 尽管航天器动力学随时间连续演化，航天器推力指令通常通过采用零阶保持的时间采样控制器执行。由于经典 CBF 条件是在连续时间下推导的，若不考虑采样间效应，直接将其应用于时间采样动力系统可能会丧失安全保证 [6]。

(4) 短视行为。 即使可行，控制障碍函数二次规划安全过滤器通常也是短视的，因为它们瞬时强制执行约束，而不预见未来激活，也不优化安全性之外的目标。这可能导致振荡或抖动控制 [34], 过大的推力消耗，以及局部安全动作与全局高效策略相冲突的次优轨迹。

(5)模型不确定性。 交会与接近操作通常涉及显著的建模不确定性以及部分未知/隐藏参数，尤其是在非合作或碎片交互场景中 [28]。由于标准CBF条件假定已知的控制仿射模型，未建模动力学和参数不确定性可能使障碍函数导数界失效并损害不变性，除非引入鲁棒性机制——而这往往以在严格执行器限制下牺牲可行性为代价。

幸运的是，越来越多的近期工作致力于缓解这些不足。为处理相对阶限制，高阶控制障碍函数构造对障碍条件求导，直至控制显式出现，从而使基于 QP 的过滤器能够影响障碍函数的演化 [26]。输入约束 CBF（ICCBF）同样如此，并且还通过构造输入可容许内安全集，将推力限制直接纳入障碍函数构造 [2]。尽管这些扩展提高了对航天器相关几何约束的适用性，但它们可能通过缩小可容许集并收紧可行性裕度而引入保守性；在接近约束边界或存在扰动时，所得的 QP 优化仍可能变得不可行 [17]。

为解决时间采样问题，已有研究提出离散时间控制屏障函数变体 [1, 13, 24, 33]，以在采样时刻确保安全性，但这类方法通常会导致非仿射控制依赖关系，而高阶约束则会进一步增加计算复杂度。为了在时间采样系统中使用连续时间控制屏障函数公式，同时保持每个制导周期仅求解一次凸二次规划的计算可行性，研究者采用了基于裕度的控制屏障函数实现 [6, 19]，通过利用基于利普希茨论证或可达集论证推导出的显式采样间界限，强化控制屏障函数不等式。然而，对于较长的

更新周期，这些裕度可能会显著缩小可容许集，而且要计算每个时间步的利普希茨常数，通常需要进行计算量很大的数值最小化。

与其他局限相比，探讨 CBF 公式中短视行为的文献相对有限。在 [32,33], 中，通过将 CBF 约束嵌入模型预测控制（MPC）公式来缓解短视问题； [7]则采用预测安全过滤以考虑成本。这些方法能够改善长时域行为，但会引入额外的设计选择（例如时域和权重的选择），且通常比逐步求解控制障碍函数二次规划需要更复杂的在线优化；在与飞行相关的更新频率和星载计算资源受限的情况下，这可能带来挑战。

为应对模型和参数不确定性带来的挑战，研究人员已开发出不确定性感知控制障碍函数变体 [12,16,31] 。此外，研究人员还开发了鲁棒时间采样方法 [8,20] ，将采样间隔裕度纳入考量，从而在零阶保持和有界扰动下维持安全性。航空航天应用已开始采用这些思路，实现不确定性感知的约束轨迹生成 [9]。然而，这些机制可能以更高的保守性为代价换取鲁棒性，从而降低执行器限制严格时的可行性，因此需要采用学习增强设计，调整障碍函数参数，以确保生成燃料与时间最优轨迹。

一种此类方法见 [15], ，该方法将概率学习与基于模型预测控制的验证相结合，针对离散时间输入约束控制障碍函数提出了不确定性感知在线自适应框架，以优化 class-$\mathcal{K}$ 参数。该框架通过显式在线验证提供了强有力的安全保证，但每次控制更新时都依赖多步预测与优化。相比之下，本研究考虑在时间采样零阶保持下执行的连续时间输入约束控制障碍函数，并致力于实现适用于航天器制导频率的低开销在线自适应。具体而言，输入约束控制障碍函数用于编码推力约束，而时间采样执行则通过控制安全裕度来处理；该裕度利用微分代数（DA）高效计算，避免了先前工作中使用的计算密集型基于最小化的裕度估计。

本研究的核心思想是对输入约束控制障碍函数递归中的完整 class-$\mathcal{K}$ 层级进行参数化与学习，从而揭示决定保守性和输入可容许内安全集的自由度。通过涵盖大量隐藏物理与运行参数分布以及状态与推力不确定性的训练，循环元强化学习策略能够在线生成适应任务的 class-$\mathcal{K}$ 参数，同时每个控制步仅需执行一次凸二次规划，从而确保实时可处理性。本研究的主要贡献如下。

1. **可学习输入约束控制障碍函数递归：** 对类-$\mathcal{K}$ 层级进行参数化和学习，以降低保守性并提高可行性。 2. **基于微分代数的时间采样裕度：** 将采样间安全裕度纳入 ICCBF–QP，并利用微分代数高效计算，从而保持实时可处理性。 3. **不确定性下的元强化学习自适应：** 在隐藏参数和扰动分布的大幅变化范围内训练元强化学习策略，以生成稳健、可迁移的输入约束控制障碍函数参数。

为验证所开发的框架，我们针对维度逐渐增加的三个任务开展了蒙特卡罗（MC）研究：一维巡航控制问题、二维对接问题和三维巡检问题。在所有情况下，目标都是在维持安全性并抵达目标的同时节省燃料；巡检案例还旨在最大化巡检指标。元强化学习

训练分别采用标准多层感知机策略和循环长短期记忆（LSTM）策略进行，以作比较。

本文其余部分安排如下。第2节回顾问题设定、时间采样动力学模型，以及有关控制障碍函数、输入约束控制屏障函数和元强化学习的预备知识。第3节介绍所提出的方法论，包括输入约束控制障碍函数递归的参数化、控制裕度计算及相关的二次规划公式。第4节描述元强化学习训练框架，包括循环策略架构、奖励设计和观测结构。第5节报告三个案例研究的结果，第6节总结全文。

2 理论背景

符号说明。 $\mathbb{R}$ 和 $\mathbb{R}_+$ 分别表示实数和非负实数。若函数$\alpha\colon [0,a) \to [0,\infty)$连续、严格递增，且 $\alpha(0)=0$，则称其为 class-$\mathcal{K}$ 函数。李导数记为 $L_{\boldsymbol{f}}h(\boldsymbol{x}) = \frac{\partial h}{\partial \boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})$和 $L_{\boldsymbol{g}}h(\boldsymbol{x}) = \frac{\partial h}{\partial \boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x})$。

2.1 问题设定

考虑控制仿射系统

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{u}, \qquad \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^n,\ \boldsymbol{u} \in \mathcal{U} \subset \mathbb{R}^m, \tag{1}$$

其中 $\boldsymbol{f}$ 和 $\boldsymbol{g}$ 足够光滑。通过连续可微函数 $h\colon \mathcal{X} \to \mathbb{R}$ 指定安全性，其安全集为

$$\mathcal{S} \triangleq \{\boldsymbol{x} \in \mathcal{X} : h(\boldsymbol{x}) \geq 0\}. \tag{2}$$

如果 $\boldsymbol{x}(0) \in \mathcal{S}$ 意味着 $\boldsymbol{x}(t) \in \mathcal{S}$ 对所有 $t \geq 0$均成立，则反馈控制律使 $\mathcal{S}$ 前向不变 。

2.2 时间采样执行与零阶保持

尽管 (1) 在连续时间内演化，但传感和控制更新通常在数字硬件上以离散时刻执行。令 $t_k = kT$ 表示采样时刻，其中 $T > 0$ 为采样时间步长。在采用零阶保持（ ZOH ）的时间采样实现中，仅在 t_k 更新控制输入，并在两次更新之间保持恒定：

$$\boldsymbol{u}(t) = \boldsymbol{u}_k, \qquad t \in [t_k, t_{k+1}), \qquad \boldsymbol{u}_k \in \mathcal{U}, \tag{3}$$

其中， $\boldsymbol{u}_k$ 根据采样状态 $\boldsymbol{x}_k \triangleq \boldsymbol{x}(t_k)$ 计算得出。在 $[t_k, t_{k+1})$ 上得到的轨迹段是方程 (1) 在初始条件 $\boldsymbol{x}(t_k) = \boldsymbol{x}_k$ 和恒定输入 $\boldsymbol{u}_k$ 下的解。状态更新可写成积分形式：

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \int_0^T \Big(\boldsymbol{f}\big(\boldsymbol{x}(t_k+\tau)\big) + \boldsymbol{g}\big(\boldsymbol{x}(t_k+\tau)\big)\boldsymbol{u}_k\Big)\, d\tau. \tag{4}$$

请注意，在此设置下，仅在采样时刻强制满足连续时间控制障碍函数，并不能保证在 $(t_k,\ t_{k+1})$ 上的安全性，除非明确界定采样间效应，或将其纳入条件 [9]。

2.3 控制障碍函数

标准的（相对阶为一）控制障碍函数条件通过要求存在一个 class-$\mathcal{K}$ 函数 α 来保证正向不变性，使得

$$\sup_{\boldsymbol{u} \in \mathcal{U}} \left[L_{\boldsymbol{f}}h(\boldsymbol{x}) + L_{\boldsymbol{g}}h(\boldsymbol{x})\boldsymbol{u}\right] \geq -\alpha(h(\boldsymbol{x}))\,, \qquad \forall \boldsymbol{x} \in \mathcal{X}. \tag{5}$$

等价地，任何满足逐点不等式的控制

$$L_{\boldsymbol{f}} h(\boldsymbol{x}) + L_{\boldsymbol{g}} h(\boldsymbol{x})\boldsymbol{u} + \alpha(h(\boldsymbol{x})) \geq 0 \tag{6}$$

在标准正则性假设下，使 $\mathcal{S}$ 前向不变（例如，参见 [2,3]）。

2.4 控制李雅普诺夫函数

对于本文介绍的某些任务（巡航控制和对接），通过在二次规划中与控制屏障函数约束并用的控制李雅普诺夫函数（CLF）不等式来编码目标到达。设 $V\colon\ \mathcal{X} \to \mathbb{R}_+$ 连续可微。标准 CLF 条件要求存在一个 class-$\mathcal{K}$ 函数 α_V，使得

$$\inf_{\boldsymbol{u}\in\mathcal{U}} \left[L_{\boldsymbol{f}} V(\boldsymbol{x}) + L_{\boldsymbol{g}} V(\boldsymbol{x})\boldsymbol{u}\right] \leq -\alpha_V(V(\boldsymbol{x}))\,, \qquad \forall \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

在本文中，α_V 被限制为单调 class-$\mathcal{K}$ 函数族，使得$\alpha_V(s) = c_V s$，其中标量增益为 $c_V > 0$。

2.5 输入约束控制障碍函数

在交会与接近操作（RPO）中，必须明确考虑推力饱和，而且许多安全条件由位置定义，因此在二阶平移动力学下通常会产生更高的相对阶（即，对于仅位置障碍函数，$L_{\boldsymbol{g}} h = 0$）。输入约束控制屏障函数（ICCBF）可同时解决这两个问题。

在有界驱动条件下，标称安全集 $\mathcal{S} = \{h \geq 0\}$ 可能不是前向不变的。输入约束控制屏障函数通过递归地将 h 与李导数及类-$\mathcal{K}$ 函数 [2]组合，构造输入可容许内安全集 $\mathcal{C}^\star \subseteq \mathcal{S}$。ICCBF 的表述如下。定义：

$$b_0 \triangleq h(\boldsymbol{x}), \tag{8}$$

$$b_i \triangleq \inf_{\boldsymbol{u}\in\mathcal{U}} \left[L_{\boldsymbol{f}} b_{i-1}(\boldsymbol{x}) + L_{\boldsymbol{g}} b_{i-1}(\boldsymbol{x})\boldsymbol{u} + \alpha_{i-1}(b_{i-1}(\boldsymbol{x}))\right], \tag{9}$$

对于 $i = 1, \ldots\ldots, N$，其中 $\{\alpha_i\}_{i=0}^{N-1}$ 为类-$\mathcal{K}$ 函数。令 $\mathcal{C}_i \triangleq \{\boldsymbol{x}\colon\ b_i(\boldsymbol{x}) \geq 0\}$，并且

$$\mathcal{C}^\star \triangleq \bigcap_{i=0}^{N} \mathcal{C}_i. \tag{10}$$

如果存在类-$\mathcal{K}$ 函数 α_N，使得

$$\sup_{\boldsymbol{u}\in\mathcal{U}} \left[L_{\boldsymbol{f}} b_N(\boldsymbol{x}) + L_{\boldsymbol{g}} b_N(\boldsymbol{x})\boldsymbol{u} + \alpha_N(b_N(\boldsymbol{x}))\right] \geq 0, \qquad \forall \boldsymbol{x} \in \mathcal{C}^\star, \tag{11}$$

那么，任何满足以下条件的局部利普希茨反馈

$$L_{\boldsymbol{f}} b_N(\boldsymbol{x}) + L_{\boldsymbol{g}} b_N(\boldsymbol{x})\boldsymbol{u} + \alpha_N(b_N(\boldsymbol{x})) \geq 0 \tag{12}$$

使 $\mathcal{C}^\star$ 前向不变 [2]。

2.6 元强化学习

强化学习（RL）旨在解决序贯决策问题：智能体与环境交互，并通过反复试错来改进自身行为 [25]。元强化学习（Meta-RL）将强化学习从针对单一任务学习单一策略，扩展为在一组相关任务的分布中学习适应机制 $\mathcal{M} \sim p(\mathcal{M})$ [5]。

在这种设定下，各个任务因质量、推力限制或扰动水平等参数的变化而有所不同；因此，最优行为取决于一些无法通过瞬时观测完全确定的属性。

策略 π_θ 可以表示为前馈映射，也可以表示为循环动力系统。在前馈情形下，多层感知机（MLP）策略根据当前观测 $\boldsymbol{o}_k$选择动作 $\boldsymbol{a}_k = \pi_\theta(\boldsymbol{o}_k)$，从而形成无记忆决策规则。在循环情形下，可以使用 LSTM 为映射增加内部隐藏状态，即 $\boldsymbol{a}_k = \pi_\theta(\boldsymbol{o}_k,\ \boldsymbol{s}_k)$，其中 $\boldsymbol{s}_{k+1} = \phi_\theta(\boldsymbol{s}_k,\ \boldsymbol{o}_k,\ \boldsymbol{a}_k)$，$\boldsymbol{s}_k$为当前隐藏状态，$\phi_\theta$ 为隐藏网络的状态更新函数，使策略能够随时间整合信息。这种内部记忆能够在单个回合内推断隐藏的任务参数并据此进行适应，因此非常适用于元强化学习问题 [14]。

3 方法论

给定式（1）中的控制仿射系统，其有界输入集为 $\mathcal{U}$，安全规范由 $h(\boldsymbol{x}) \geq 0$ 编码，本工作的目标是在推力饱和、零阶保持（ZOH）下的时间采样执行，以及显著的参数不确定性和噪声条件下执行安全控制，同时减少燃料消耗及基于控制障碍函数的过滤带来的其他短视效应。

本文的方法结合了三个要素：(1) 在有界驱动下可实施的基于输入约束控制障碍函数的内层安全集 $\mathcal{C}^\star$，(2) 为促进零阶保持下的正向不变性行为而设计的时间采样裕度，以及 (3) 在输入约束控制障碍函数递归中对完整 class-$\mathcal{K}$ 层级的学习得到的参数化，并通过元强化学习进行训练，以提升在不同隐藏参数和不确定性下的可行性并减少燃料消耗。在运行时，每个时间步仅通过一个凸二次规划计算控制输入。

3.1 ICCBF递归与Class-$\mathcal{K}$ 参数化

第2.2.5节中的 ICCBF 递归定义了序列 $\{b_i(x)\}_{i=0}^{N}$ 以及式 (9)–(10) 中的内层安全集 $\mathcal{C}^\star$。集合 $\mathcal{C}^\star$ 在很大程度上取决于 class-$\mathcal{K}$ 函数$\{\alpha_i\}_{i=0}^{N}$的选择。在本文中，每个 α_i 均参数化为具有正增益的单调 class-$\mathcal{K}$ 函数，

$$\alpha_i(s_k) = \theta_{i,k}\, s_k, \qquad \theta_{i,k} > 0, \qquad i = 0, \ldots, N, \tag{13}$$

其中，增益是状态自适应的，并在每个采样时刻 k更新。将这些增益汇集起来可得

$$\boldsymbol{\theta}_k \triangleq [\theta_{0,k}, \theta_{1,k}, \ldots, \theta_{N,k}]^\top \in \mathbb{R}^{N+1}_{>0}. \tag{14}$$

一个通过强化学习训练的策略 π_ψ 将当前状态 x_k 映射为 ICCBF参数，即$\boldsymbol{\theta}_k = \pi_\psi(x_k)$，从而在线调整内层安全集的保守性与可行性。当存在控制李雅普诺夫函数时，该策略还会输出其类别-$\mathcal{K}$ 参数，即 $[\boldsymbol{\theta}_k, c_{V,k}] = \pi_\psi(x_k)$。

3.2 控制裕度计算

时间采样控制障碍函数必须保证对于所有 $t \geq 0$，$\boldsymbol{x}(t) \in \mathcal{C}^\star$ 均成立，而不仅仅是在采样时刻成立。为此，在采样状态的不变性条件中加入零阶保持输入约束控制障碍函数裕度，使得式 (12) 变为

$$L_f b_N(\boldsymbol{x}) + L_g b_N(\boldsymbol{x}) u + \alpha_N(b_N(\boldsymbol{x})) \geq \nu(T, \boldsymbol{x}) \tag{15}$$

其中， $\nu(T,\ \boldsymbol{x})$ 是所添加的裕度。在本研究中， $\nu(T,\ \boldsymbol{x})$ 使用定理 1 计算，其中 [6],

$$\nu(T,\boldsymbol{x}) = l_1(\boldsymbol{x})T\Delta(\boldsymbol{x}) \tag{16}$$

$$\Delta(\boldsymbol{x}) = \sup_{\boldsymbol{z}\in\mathcal{X}(\boldsymbol{x},T),u\in\mathcal{U}} \|f(\boldsymbol{z}) + g(\boldsymbol{z})u\| \tag{17}$$

$$l_1(\boldsymbol{x}) = l_{L_f b_N}(\boldsymbol{x}) + l_{L_g b_N}(\boldsymbol{x})u_{\max} + l_{\alpha(b_N)}(\boldsymbol{x}) \tag{18}$$

其中 $l_{L_f b_N}$、 $l_{L_g b_N}$ 和 $l_{\alpha(b_N)}$ 是集合 $\mathcal{R}(\boldsymbol{x},T)$ 的局部 Lipschitz 常数。在该裕度下正向不变性的证明可见于 [6]。在 [6],中，式 (17) 和式 (18) 通过由网格采样近似的数值最大化计算。相比之下，本文使用微分代数计算这些量，从而以更低的计算负担得到上界。

在每个采样时间，定义一个保守的局部超集为 $\mathcal{R}(\mathbf{x}_k,\zeta)\ \supseteq\ \mathcal{X}(\mathbf{x}_k,T)$，将其选为以 $\mathbf{x}_k$ 为中心、半宽度为 $\zeta\in\mathbb{R}^n_{>0}$ 的超矩形：

$$\mathcal{R}(\mathbf{x}_k,\zeta) = \{\boldsymbol{x}\in\mathbb{R}^n : |x_i - (\mathbf{x}_k)_i| \le \zeta_i,\ i = 1,\dots,n\}\,. \tag{19}$$

随后，针对域 $\boldsymbol{x} = \mathbf{x}_k + \delta\mathbf{x}$，在 $\delta x_i \in [-\zeta_i,\zeta_i]$ 条件下构造裕度计算中所有相关量的泰勒近似。

对于连续可微标量函数 $p(x)$，域 $\mathcal{R}(\mathbf{x}_k,\zeta)$ 上的局部 Lipschitz 常数定义为 $l_p(\mathbf{x}_k) \coloneqq \sup_{x\in\mathcal{R}(\mathbf{x}_k,\zeta)} \|\nabla p(x)\|_2$，并满足

$$l_p(\mathbf{x}_k) \le \left(\sum_{i=1}^{n}\left(\sup_{x\in\mathcal{R}(\mathbf{x}_k,\zeta)}\left|\frac{\partial p}{\partial x_i}(x)\right|\right)^2\right)^{1/2}. \tag{20}$$

利用微分代数，可以通过对微分代数多项式求导来计算 $\partial p/\partial x_i$，并可利用微分代数界定，在 $\mathcal{R}(\mathbf{x}_k,\ \zeta)$上获得各偏导数泰勒近似的区间包络。

令 $\overline{d}_i$ 表示所得上界 $d_i \ge \sup_{x\in\mathcal{R}(\mathbf{x}_k,\zeta)} \left|\partial p/\partial x_i(x)\right|$。于是， $l_p(\mathbf{x}_k)$的保守界可定义为

$$\widehat{l}_p(\mathbf{x}_k)\ =\ \left(\sum_{i=1}^{n}\overline{d}_i^{\,2}\right)^{1/2}, \tag{21}$$

$\widehat{\ }$使得 $l_p(\mathbf{x}_k) \ge l_p(\mathbf{x}_k)$。这可用于计算$l_{L_f b_N}$、 $l_{L_g b_N}$和 $l_{L_{\alpha(b_N)}}$的保守替代量。

接下来，为了界定 $\Delta(\mathbf{x}_k)$，在$\mathcal{R}(\mathbf{x}_k,\ \zeta)$上计算系统动力学的微分代数包络。对于式（1）中具有可容许输入$\mathcal{U} = \{u : \|u\|_2 \le u_{\max}\}$的控制仿射动力学，根据三角不等式可得

$$\begin{aligned}\sup_{\substack{x\in\mathcal{R}(\mathbf{x}_k,\zeta)\\ u\in\mathcal{U}}} \|f(x) + g(x)u\|_2\ \le\ &\sup_{x\in\mathcal{R}(\mathbf{x}_k,\zeta)} \|f(x)\|_2 \\ &+\ u_{\max}\sup_{x\in\mathcal{R}(\mathbf{x}_k,\zeta)} \|g(x)\|_2.\end{aligned} \tag{22}$$

其中， $\|\cdot\|_2$ 表示 $g(x)$的诱导矩阵 2-范数。通过微分代数区间包络（对 f采用逐分量界，对 g采用范数界），可得到式（22）中上确界的保守值，从而得到 $\Delta(\mathbf{x}_k)$的保守值。

最后可将所得裕度汇总为

$$\widehat{l}_1(\mathbf{x}_k) = \widehat{l}_{L_f b_N}(\mathbf{x}_k) + \widehat{l}_{L_g b_N}(\mathbf{x}_k)\,u_{\max} + \widehat{l}_{b_N}(\mathbf{x}_k), \tag{23}$$

$$\widehat{\nu}(T,\mathbf{x}_k) = \widehat{l}_1(\mathbf{x}_k)\,T\,\widehat{\Delta}(\mathbf{x}_k). \tag{24}$$

由于 $\mathcal{X}(\mathbf{x}_k, T) \subseteq \mathcal{R}(\mathbf{x}_k, \zeta)$，且对于泰勒近似，每个 DA 界在 $\mathcal{R}(\mathbf{x}_k, \zeta)$ 上都是保守的，因此 $\hat{\nu}_1^{\ell}$ 可充分估计所需裕度，无需通过任何网格搜索方法计算利普希茨常数。

然而，需要注意的是，DA包络是针对截断后的泰勒近似计算的，而非针对精确函数计算的。对于非多项式表达式，截断会引入建模误差，原则上可能损害保守性。在本工作中，我们通过选择足够高的DA展开阶数并限制局部域来减轻截断效应，使真实函数值落在所得界限内。所得界限通过大量蒙特卡罗测试进行了验证，如配套的Jupyter笔记本所示。所有测试用例均未观察到界限违反。虽然泰勒模型 [10] 可用于为非多项式表达式获得具有形式化保证的界限，但这会大幅增加计算复杂度，因此鉴于DA界限已展现出很强的经验有效性，这超出了本工作的范围。

3.3 二次规划公式

在每个时间步 k，都会求解一个二次规划，以在满足时间采样ICCBF安全条件的同时计算控制输入。基线逐步二次规划公式为

$$\begin{aligned} \mathbf{u}_k^{*} = & \arg\min_{\mathbf{u}_k \in \mathbb{R}^m} \|\mathbf{u}_k\|^2 \\ \text{s.t. } & L_f b_N(\boldsymbol{x}_k) + L_g b_N(\boldsymbol{x}_k)\,\mathbf{u}_k \\ & \quad + \theta_{N,k} b_N(\boldsymbol{x}_k) \geq \widehat{\nu}_k\big(T, \boldsymbol{x}_k\big), \quad \mathbf{u}_k \in \mathcal{U}. \end{aligned} \tag{25}$$

其中，b_N 是终端ICCBF函数，$\mathcal{K}_N > 0$ 是终端类-$\mathcal{K}$ 增益，$\widehat{\nu}_k$ 是时间采样裕度，用于考虑零阶保持下的采样间演化。

对于还要求满足收敛目标的场景，我们会加入带松弛变量的CLF约束，以在不损害安全性的前提下保证可行性。由此得到的CLF–ICCBF–QP为

$$\begin{aligned} \mathbf{u}_k^{*} = & \arg\min_{\mathbf{u}_k \in \mathbb{R}^m,\ \varepsilon_k \geq 0} \tfrac{1}{2}\|\mathbf{u}_k\|^2 + p\,\varepsilon_k^2 \\ \text{s.t. } & L_f b_N(\boldsymbol{x}_k) + L_g b_N(\boldsymbol{x}_k)\,\mathbf{u}_k \\ & \quad + \theta_{N,k}\, b_N(\boldsymbol{x}_k) \geq \widehat{\nu}_k\big(T, \boldsymbol{x}_k\big), \\ & L_f V(\boldsymbol{x}_k) + L_g V(\boldsymbol{x}_k)\,\mathbf{u}_k \leq -c_{V,k}\, V(\boldsymbol{x}_k) + \varepsilon_k, \\ & \mathbf{u}_k \in \mathcal{U}. \end{aligned} \tag{26}$$

其中，V是控制李雅普诺夫函数，$\alpha_V(\cdot)$是类-$\mathcal{K}$ 函数，$p > 0$ 用于对CLF松弛进行加权。松弛变量 ε 用于放宽CLF约束，使满足CLF的要求不会以牺牲CBF为代价，也就是说，不会损害安全性。由于ICCBF终端条件关于$\mathbf{u}_k$是仿射的，因此这些优化问题都是凸二次规划。在本研究的所有实验中，这些二次规划均使用MOSEK [18]求解。

4 元强化学习公式

该学习问题被建模为一个基于时间采样动力学的情节式离散时间控制任务。在每个时间步 t_k，智能体选择 ICCBF整形参数 $\boldsymbol{\theta}_k$，并由第3节.3.3中的二次规划公式据此计算实际执行的控制输入。下文将详细介绍所用的网络架构、隐藏参数分布以及强化学习环境。

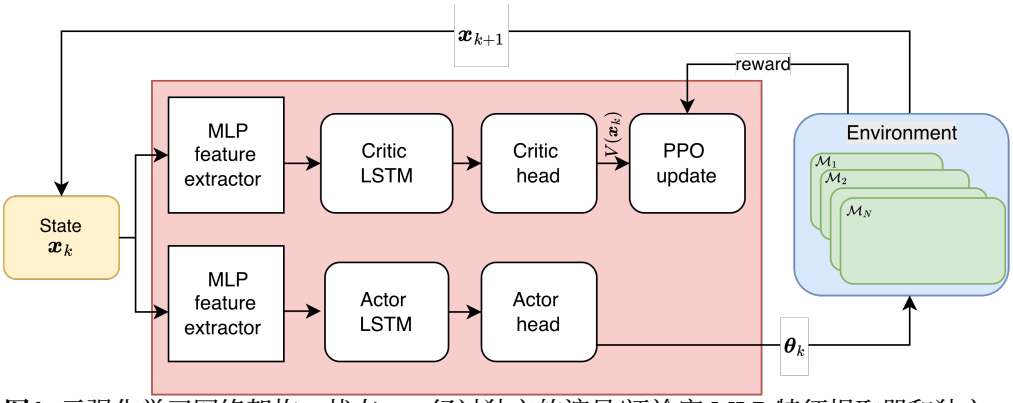

图1. 元强化学习网络架构。状态 x_k 经过独立的演员/评论家 MLP 特征提取器和独立的长短期记忆网络映射，由演员头输出 ICCBF 参数 $\boldsymbol{\theta}_k$，由评论家头输出价值估计 $V(x_k)$。近端策略优化通过跨任务 $\mathcal{M}_i \sim p(\mathcal{M})$ 的轨迹采样进行更新。

4.1 网络架构

训练采用在 Stable-Baselines3 [22] 中实现的循环执行器—评论家策略。网络架构如图 1 所示。观测为完整的系统状态 x_k。执行器和评论家使用彼此独立的特征提取器及循环模块。因此，执行器通过专用的多层感知机特征提取器处理 x_k，并生成 ICCBF整形参数 $\boldsymbol{\theta}_k$，这些参数经过缩放以满足$\boldsymbol{\theta}_k \in [\boldsymbol{\theta}_{\min}, \boldsymbol{\theta}_{\max}]$。评论家则使用单独的多层感知机特征提取器，后接 LSTM，以整合时间信息并推断与任务变化相关的隐藏参数知识，从而进行价值估计；评论家输出价值函数 $V(x_k)$，用于优势计算，并在策略更新中加以利用。策略参数采用近端策略优化（PPO） [23, 29]进行优化。

4.2 初始状态和隐藏参数分布

每个回合都通过采样一个初始条件和若干可变环境参数进行初始化，例如质量、最大推力 $u_{\max}$、KIZ/KOZ 的大小等。这种回合级随机化使元强化学习智能体接触到各种任务实例的分布，从而提升鲁棒性并减少对单一标称场景的过拟合。

在回合开始时，环境从有界可容许集 $\mathcal{X}_0 \in \mathcal{S}$ 中抽取一个初始状态 x_0。每个回合还会从以标称值 $\bar{\mathbf{p}}$为中心的有界超矩形中采样一个隐藏参数向量$\mathbf{p} \in \mathbb{R}^{n_p}$。每个参数 p_i 均独立抽取，具体如下：

$$p_i \sim \mathrm{Uniform}\big(\big[(1-\delta_i)\,\bar{p}_i,\ (1+\delta_i)\,\bar{p}_i\big]\big)\,, \tag{27}$$

对于 $i = 1, \ldots, n_p$，其中 $\bar{p}_i$ 表示第 i 个参数的标称值，$\delta_i > 0$指定其相对变化水平。结果中各测试用例采用的百分比变化量 δ_i 汇总于表2。在后文给出的蒙特卡罗结果中，每个数据集内各控制器使用固定的参数样本，以确保性能差异完全源于所提出的 ICCBF 调节机制。

4.3 强化学习环境

如前所述，本研究使用近端策略优化训练一个神经网络，使其在时间步 k 输出式（26）中的 ICCBF 类-$\mathcal{K}$ 增益 $\boldsymbol{\theta}_k$，并可选择输出 CLF增益 $c_{V,\,k}$。

算法 1 状态转移与奖励计算

要求： T, t, t_f, 状态 x_k, $\sigma_{\|\mathbf{u}\|}$, $\sigma_{\mathbf{u}_\theta}$, $\sigma_{\mathbf{x}}$, θ_k, （可选） α'1：更新时间 $t \leftarrow t + T$。2：设置 done $=$ false 和 failure $=$ false。3：使用式（23）和（24）计算 $\widehat{\nu}(T, \mathrm{x}_k)$。4：按式（26）构建 ICCBF–QP。5：求解（26）以获得标称控制 $\mathrm{u}_k^\star$。6：如果二次规划不可行或发生求解器故障，则 7：设置 done $\leftarrow$ true；failure $\leftarrow$ true；$\mathrm{u}_k^\star \leftarrow \mathbf{0}$。8：结束 if。9：按式（29）施加控制噪声，得到 u_k。10：使用式（1），在 T 上通过 u_k 推进状态，得到 x_{k+1}。11：计算原始安全函数值 h_{k+1}，以及（如有）控制李雅普诺夫函数值 V_{k+1}。12：使用式（28）添加噪声以形成观测 $\boldsymbol{S}_k$，并使用式（32）进行缩放，得到 $\boldsymbol{S}_{k+1}^*$。13：如果 $t \geq t_f$，则 14：done $\leftarrow$ true。15：结束 if。16：如果 $h_{k+1} < 0$，则 17：done $\leftarrow$ true，failure $\leftarrow$true。18：结束 if。19：使用式（33）计算奖励 R_k。20：返回（$\boldsymbol{S}_{k+1}, R_k$, done）。

执行器提供这些量后，即可使用式（25）或（26）中的 ICCBF-QP公式计算控制量 $\mathbf{u}_k$，以确保安全性维持至下一个时间步，同时还可（选择性地）满足 CLF 条件。随后，使用式（1）在零阶保持下将状态传播一个采样周期 T，得到 $\mathbf{x}_{k+1}$。算法 1 总结了这一过程。

状态误差 本文考虑状态误差，并将其建模为真实状态上的加性零均值高斯噪声。在每个时间步 k，学习智能体可获取的状态计算如下：

$$\boldsymbol{x}_k^E = \boldsymbol{x}_k + \boldsymbol{\epsilon}_{\boldsymbol{x}_k}, \qquad \boldsymbol{\epsilon}_{\boldsymbol{x}_k} \sim \mathcal{N}\big(\mathbf{0},\, \boldsymbol{\sigma}_x^2 \mathbf{I}\big)\,, \tag{28}$$

其中， $\boldsymbol{\sigma}_x = [\boldsymbol{\sigma}_r, \boldsymbol{\sigma}_v]^T$ 表示位置和速度误差的标准差。

推力误差为对推力执行误差进行建模，传播前先对$\mathbf{u}_k$ 的大小和方向进行扰动，以此得到实际施加的控制$\mathbf{u}_k^\star$。令$\mathbf{u}_k^\star = [u_x, u_y, u_z]^\top$ 表示在动力学所用的同一控制坐标系下的表达形式。定义大小 $u_k = \|\mathbf{u}_k^\star\|_2$ 以及面外角和面内角（β, γ），随后分别施加大小扰动和角度扰动，并将执行指令重构为

$$u_k = \|\mathbf{u}_k^\star\|_2, \qquad \beta = \sin^{-1}\!\left(\frac{u_z}{u_k}\right), \qquad \gamma = \tan^{-1}\!\left(\frac{u_x}{u_y}\right), \tag{29}$$

$$u_k^E = u_k\,(1 + \delta_u)\,, \qquad \beta^E = \beta + \delta_\beta, \qquad \gamma^E = \gamma + \delta_\gamma, \tag{30}$$

$$\mathbf{u}_k = u_k^E \begin{bmatrix} \cos\beta^E \sin\gamma^E \\ \cos\beta^E \cos\gamma^E \\ \sin\beta^E \end{bmatrix}, \qquad \mathbf{u}_k \leftarrow \frac{u_{\max}}{\max(u_{\max}, \|\mathbf{u}_k\|_2)}\,\mathbf{u}_k. \tag{31}$$

where $\delta_u \sim \mathcal{N}(0, \sigma_u^2)$, $\delta_\beta \sim \mathcal{N}(0, \sigma_\beta^2)$, and $\delta_\gamma \sim \mathcal{N}(0, \sigma_\gamma^2)$.

观测在每个时间步，强化学习环境还必须提供观测 $\boldsymbol{S}_k$ ，其中包含足够的信息，以便估计每一步动作的潜在奖励 k。为使输入归一化，从而确保稳定性和特征缩放的一致性，采用简单的最小-最大归一化，使得

$$\boldsymbol{S}_k^* = 2\left(\frac{\boldsymbol{S}_k - \boldsymbol{S}_{\min}}{\boldsymbol{S}_{max} - \boldsymbol{S}_{\min}}\right) - 1. \tag{32}$$

其中 $\boldsymbol{S}_{\min}$ 和 $\boldsymbol{S}_{\max}$ 分别是 $\boldsymbol{S}$中各分量的最小值和最大值向量。

对于所考虑的巡航控制和对接情形， $\boldsymbol{S}_k = \boldsymbol{x}_k^E$。对于检查情形， $\boldsymbol{S}_k = [\boldsymbol{x}_k^E, \theta_S, n_{insp}, \hat{\mathbf{d}}]^T$ 其中 θ_S 为太阳角， n_{insp} 为已巡检点的数量，而 d 是指向剩余未巡检点最大聚类的方向向量，该向量使用 K-means 按照 [27]所述方法计算得出。

奖励每步奖励会惩罚控制量消耗和失效事件，并在回合时域末端额外施加终端控制李雅普诺夫函数惩罚。具体而言，

$$R_k = \begin{cases} r_k, & t < t_f, \\ r_k - w_V P_V, & t \geq t_f, \end{cases} \tag{33}$$

其中

$$r_k = -w_u \|\mathbf{u}_k^\star\|_2 - w_{\text{fail}} P_{\text{fail}} - w_h \max\big(0, -h_{k+1}\big). \tag{34}$$

这里， w_u、 w_{fail}、 w_h和 w_V 是用户选定的权重，而 P_{fail} 用于标记二次规划失败：

$$P_{\text{fail}} = \begin{cases} 1, & \text{QP infeasible or solver failure}, \\ 0, & \text{otherwise}. \end{cases} \tag{35}$$

使用控制李雅普诺夫函数时，若其在整个回合内未能降至预设阈值 ρ_V以下，则施加终端惩罚。令 $N_f = \lceil t_f/T \rceil$ 和$\mathbf{V} = \{V(\mathbf{x}_j)\}_{j=0}^{N_f}$。那么

$$P_V = \begin{cases} \min(\mathbf{V}), & \min(\mathbf{V}) > \rho_V, \\ 0, & \text{otherwise}. \end{cases} \tag{36}$$

5 结果

本节考虑了三个维度逐渐增加的测试用例：(i) 一维巡航控制，(ii) 带旋转目标的二维航天器对接，这两者均改编自 [2], ；以及(iii) 改编自 [27]的三维巡检场景。如果在整个时域内或直至达成任务目标时，所有安全约束始终得到满足，则将该回合标记为安全。燃料用量通过随时间累积的总燃料消耗 $\sum_k \|\mathbf{u}_k\| T$ 或总量$\Delta v \,(/ m \sum_k \|\mathbf{u}_k\| T)$报告。燃料消耗结果（以及巡检问题的巡检指标）以均值 $\mu \pm$ 标准差 σ表示，并同时给出四分位数 $[Q_1, Q_2, Q_3, Q_4]$ 和第99百分位数 P_{99}。

每个案例均对比了三种控制器：（i）使用恒定 $[\boldsymbol{\theta}_k, c_{V,k}]$ 值的未调优 ICCBF 基线；（ii）由前馈式近端策略优化策略输出 $[\boldsymbol{\theta}_k, c_{V,k}]$ 的*MLP* 调优 ICCBF；以及（iii）由采用长短期记忆行动者-评论家的循环式近端策略优化策略输出 $[\boldsymbol{\theta}_k, c_{V,k}]$ 的*RNN* 调优 ICCBF。本节详细介绍评估方案、训练所用的超参数，以及三个测试案例的详情和结果。

表 1. 三个测试案例的主要超参数。

超参数	巡航控制	对接	检查
环境/轨迹展开			
时间步 T	0.1	0.5	10
总时间步数	10^5	10^6	10^6
批量大小	64	64	256
每次更新的轮数	10	10	10
优化/*PPO*设置			
学习率	10^{-4}	10^{-4}	2×10^{-4}
折扣因子	0.99	0.995	0.99
GAE	0.95	0.95	0.95
裁剪范围	0.10	0.10	0.20
熵系数	0.01	0.01	0.01
网络架构（除非另有说明，在演员/评论家之间共享）			
激活函数	`Tanh`	`Tanh`	`Tanh`
隐藏层	3	4	4
每层节点数	32	64	128
std	0.2	0.2	0.2
LSTM（仅循环神经网络）			
LSTM隐藏层大小	64	64	128
# LSTM层	1	1	1

表2. 每回合的隐藏参数变化（标称值及相对偏差δ_i）

类别	巡航控制	对接	检查
从航天器质量 m	1650 千克（±20%）	1000 千克（±10%）	12 kg（±10%）
最大推力 $u_{\max}$	0.25 N（±20%）	0.25 kN（±10%）	1.0 N（±10%）
最大速度 $v_{\max}$	24.0 m s^{-1} (±20%)	–	–
前车速度 v_0	13。89 米/秒$^{-1}$ （±10%）	–	–
副航天器半径 R_D	–	–	5 m (±10%)
主航天器半径 R_C	–	2.4米（±10%）	10米（±10%）
主航天器自旋速率 ω	–	0.6°秒$^{-1}$ （±10%）	–
主航天器轨道半径 r	–	6771千米（±10%）	6771 千米（±10%）
接近锥半角 γ	–	10° （±10%）	–
KIZ 半径 $R_{\max}$	–	–	800 米（±10%）

训练和网络超参数汇总于表 1。我们对超参数进行了迭代调优，以确保各任务中的训练稳定，并在每个测试案例中对所有基线采用相同的近端策略优化设置。

为确保对学习策略进行公平比较，每个测试用例的评估都在固定的、包含 N_{MC} 回合的蒙特卡洛数据集上进行，$\mathcal{D}_{\mathrm{MC}} \triangleq \{ (\mathbf{x}_0^{(i)}, \mathbf{p}^{(i)}) \}_{i=1}^{N_{\mathrm{MC}}}$ 其中初始条件和回合参数均经过预采样，并在所有方法中保持固定。

在讨论的所有测试案例中，ICCBF 的阶数 N 设为 2，这是在所考虑的二阶平移动力学下，位置定义约束所需的最小 ICCBF 阶数。对于巡航控制任务，由于安全约束的相对阶为 1，$N = 2$ 并非必需，但为保持一致并允许对内层安全集进行额外整形，仍予以保留。

表 3. 所施加的推力和状态不确定性的标准差

噪声源	巡航控制	对接	检查
σ_u	0.1	0.05	0.05
σ_γ	–	0.1°	0.1°
σ_β	–	–	0.1°
σ_r （各轴）	2 m	0.1 米	0.1 米
σ_v （每个轴）	0.5 米秒$^{-1}$	2 毫米秒$^{-1}$	2 毫米秒$^{-1}$

表 4. 各任务的性能总结

Task	Case	$\boldsymbol{\mu \pm \sigma}$	$[\boldsymbol{Q_1}, \boldsymbol{Q_2}, \boldsymbol{Q_3}, \boldsymbol{P_{99}}]$	Safe
		巡航控制的燃料消耗 [Ns，其他任务的单位为 m/s]		$N_{\text{safe}}/N_{\text{MC}}$ (%)
巡航控制	未调优	4.75 ± 0.74	[4.26, 4.77, 5.33, 5.96]	99.82%
	MLP 调优	3.86 ± 0.51	[3.53, 3.86, 4.21, 4.91]	99.82%
	RNN 调优	2.77 ± 0.62	[2.27, 2.74, 3.24, 4.08]	99.84%
对接	未调优	8.20 ± 1.94	[7.35, 8.05, 9.30, 12.00]	96.88%
	MLP 调优	5.70 ± 1.79	[4.45, 5.77, 6.79, 11.09]	97.06%
	RNN 调优	6.19 ± 1.51	[5.51, 6.15, 7.00, 10.09]	97.28%
检查	未调优	289.32 ± 194.89	[109.54, 297.13, 429.18, 866.62]	100%
	MLP 调优	27.39 ± 74.98	[7.83, 8.50, 9.55, 392.12]	88.4%
	RNN 调优	3.87 ± 5.53	[0.84, 1.10, 2.00, 15.27]	100%
		检查得分 (%)		
	未调优	99.99 ± 0.14	[100.00, 100.00, 100.00, 100.00]	
	MLP 调优	31.59 ± 14.51	[24.00, 26.00, 33.00, 88.50]	
	RNN 调优	85.41 ± 27.25	[100.00, 100.00, 100.00, 100.00]	

5.1 巡航控制问题

在 [4]和 [2], 中讨论的该问题，考虑一个沿直线运动的车辆（主航天器）质点模型。从航天器以距离 d跟随前车，前车以已知的恒定速度 v_0行驶。目标是设计一个控制器，使从航天器达到限速 v_{max} ，同时确保碰撞规避。碰撞规避安全约束定义为 $d \geq 1.8v$, 使跟随车辆达到限速的 CLF约束定义为$V(x) = (v - v_{\max})^2$。将状态向量定义为 $x = [d,\, v]^T$后，动力学模型为

$$\begin{bmatrix} \dot{d} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v_0 - v \\ -\frac{F(v)}{m} \end{bmatrix} + \begin{bmatrix} 0 \\ g_0 \end{bmatrix} u, \qquad \mathcal{U} = \{u : |u| \leq u_{\max}\}, \tag{37}$$

where $u$is the control input. The resistive force $F(v)$ is modeled
其中$F(v) = 0.1 + 5v + 0.25v^2$， m 为从航天器质量， g_0 为重力加速度。安全集 $\mathcal{S}$ 随后定义为

$$\mathcal{S} = \{x \in \mathcal{X} \mid h_0(x) = x_1 - 1.8x_2 \geq 0\}. \tag{38}$$

初始状态的抽取满足$\mathbf{x}_0 \in \mathcal{S}$。隐参数向量包括车辆质量 m、前车速度 v_0、目标速度 $v_{\max}$以及最大可用推力 $u_{\max}$。在每个回合开始时，隐藏参数按照式(27)进行采样，标称值与偏差见表2。对于所分析的5000个蒙特卡洛样本，抽取的隐藏参数如图2所示。所加入的位置与速度不确定性的标准差见表3。

结果 图 3 和表 4 总结了巡航控制任务的推力消耗统计数据。未调参ICCBF（ $\boldsymbol{\theta} = [4, 7, 2]$ 和 $\alpha_V = 10$）的推力需求相对较高，且上尾较重。对 ICCBF 进行调优可显著降低推力消耗。MLP调参ICCBF 将

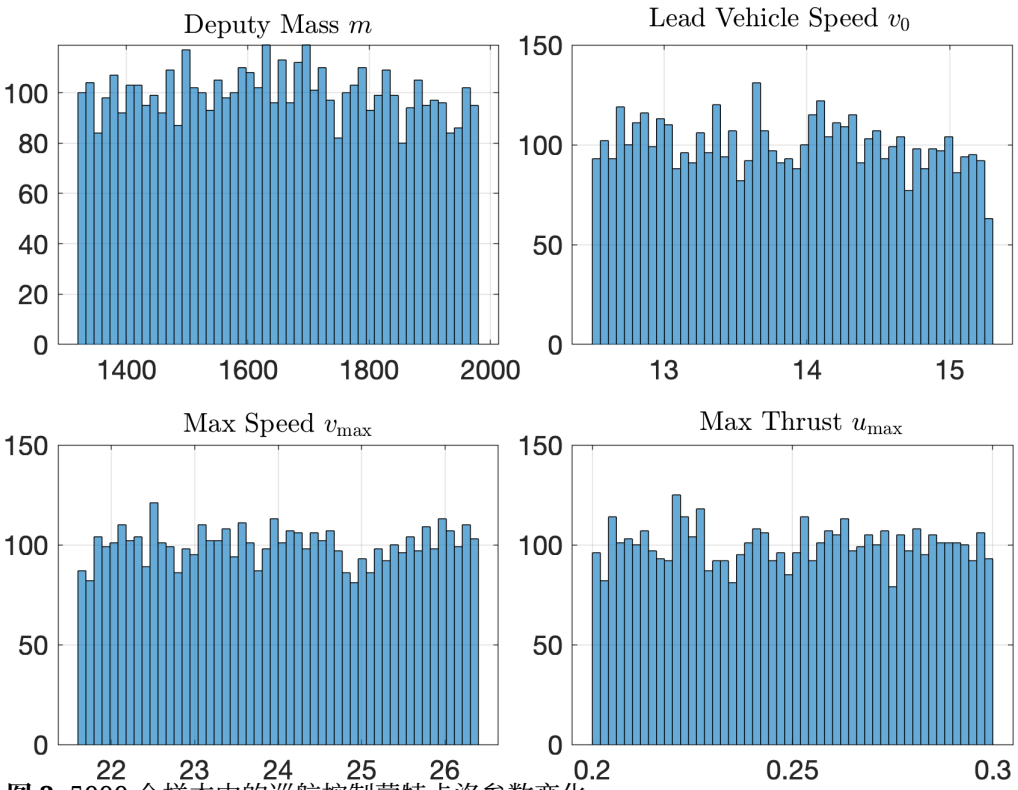

图 2. 5000 个样本中的巡航控制蒙特卡洛参数变化

平均推力降低了 18.7%，相较于未调参ICCBF。经 RNN 调优的 ICCBF 实现了最大的燃油节省：相较于未调参ICCBF降低了 41.7%，相较于 MLP调参ICCBF降低了 28.2%。中位数和四分位数统计数据也体现了这一降幅，其中，经 RNN 调优的 ICCBF 的上四分位数 Q_3 降低了 39.2%。上尾也显著收窄， P_{99} 降低了 31.5%。

图 4 展示了部分生成的蒙特卡洛轨迹，以及控制李雅普诺夫函数和控制障碍函数条件的变化情况和观察到的失败案例数量。可以看到，三种方法的失败次数都较低且相近（约为$\sim 0.2\%$），其中经 RNN 调优的 ICCBF 略微提高了经验成功率。

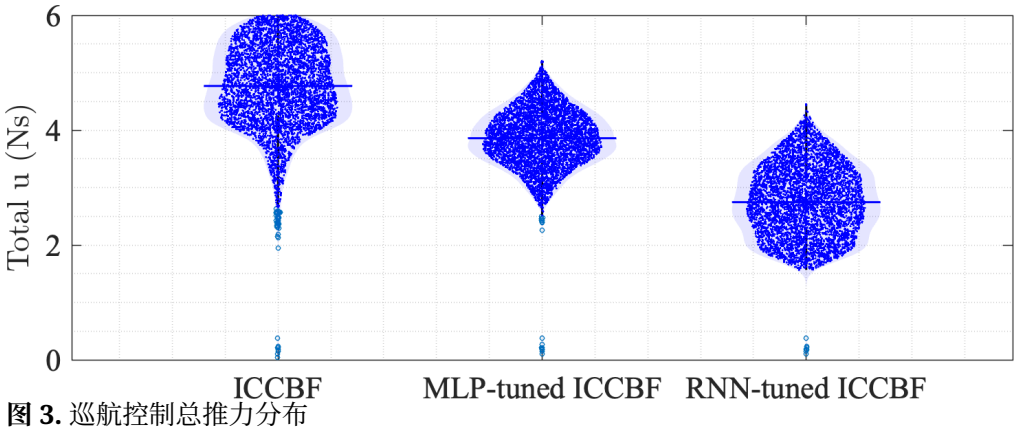

图 3. 巡航控制总推力分布

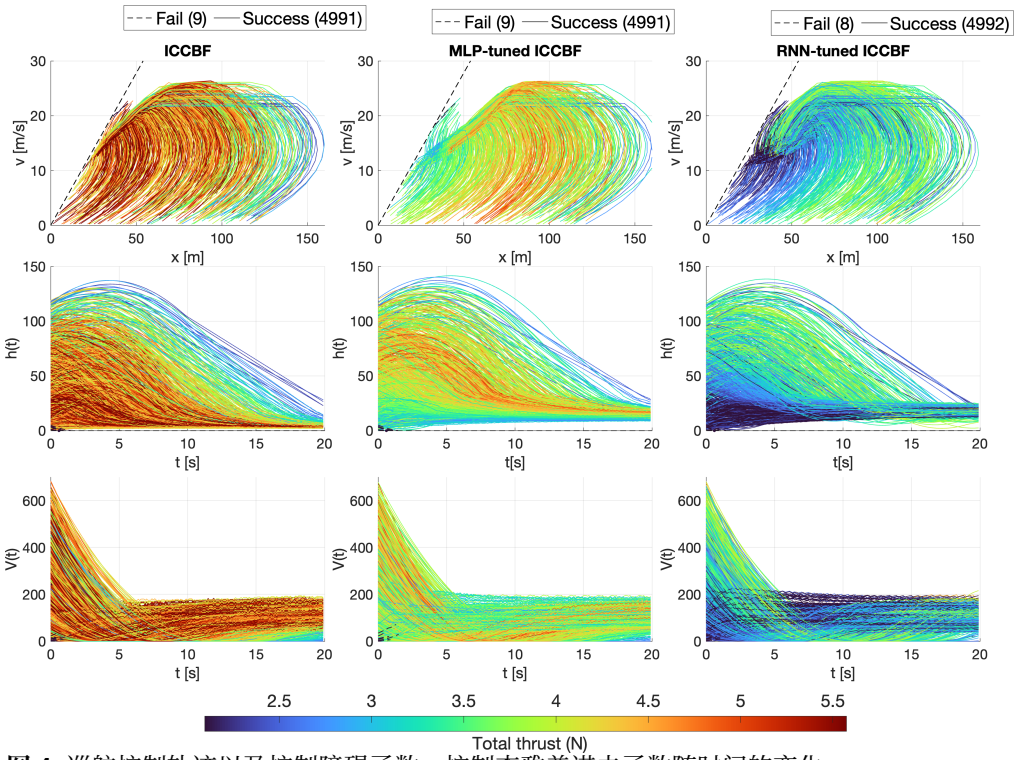

图 4. 巡航控制轨迹以及控制障碍函数、控制李雅普诺夫函数随时间的变化

5.2 航天器与旋转目标交会

在 [2] 和 [21], 中讨论的这个测试用例是追踪航天器与目标天体之间的自主交会场景，二者均建模为质点。目标被表示为圆盘上的一点，圆盘半径为 R_C，并以恒定角速度ω相对于局部垂直局部水平（LVLH）坐标系旋转。该任务旨在确定所需推力，使追踪航天器从 100 m 的初始距离移动至目标 3 m 范围内。该测试用例如图 6 所示。

在此测试用例中施加了视线安全约束，要求追踪航天器的相对位置始终保持在一个半角为 γ 、与对接轴对齐的圆锥内。系统状态定义为 $x \in \mathbb{R}^5$，包括相对位置（p_x, p_y）、相对速度（v_x, v_y）以及对接口角度 ψ。系统动力学为非线性相对运动方程，见式 [1]。控制输入 $\boldsymbol{u} = [u_x, u_y]^T$ 表示推进力，并受到约束 $\| \boldsymbol{u} \| \leq u_{\max}$。视线约束表示为 $h_0(x) \geq 0$，其中

$$h_0(x) = \cos\theta - \cos\gamma = \frac{\vec{r}_{c-p} \cdot \hat{e}}{\|\vec{r}_{c-p}\|} - \cos(\gamma) \tag{39}$$

这里，$\vec{r}_{c-p} = \left(p_x - R_C \cos\psi,\ p_y - R_C \sin\psi\right)^T$ 是追踪航天器相对于对接口的位置向量，而$\hat{e} = (\cos\psi, \sin\psi)^T$ 是对接轴单位向量。请注意，此安全约束 $h_0(x)$ 存在相对阶方面的挑战，因为 $L_g h_0(x) = 0$。为引导追踪航天器驶向对接口，采用以下控制李雅普诺夫函数

$$V(x) = \left(v_x + \frac{p_x - R_C\cos\psi}{10}\right)^2 + \left(v_y + \frac{p_y - R_C\sin\psi}{10}\right)^2. \tag{40}$$

在此情况下，初始状态仍按$\mathbf{x}_0 \in \mathcal{S}$的方式抽取。回合参数向量包括从航天器质量 m、追踪器半径 R_C、目标的旋转速率

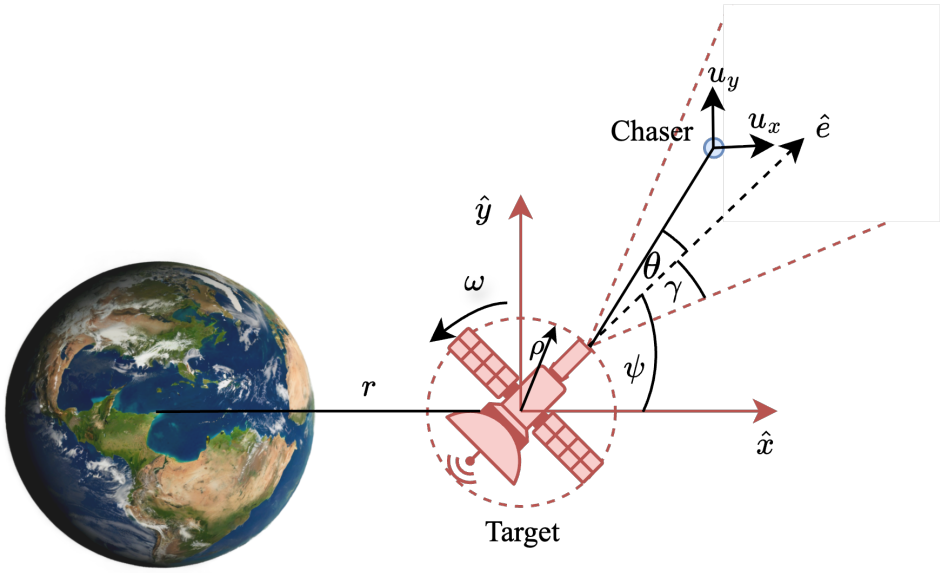

图 5. 航天器对接问题

ω，以及最大可用推力 $u_{\max}$。每个回合开始时，参数均根据式 (27) 抽样，标称值和偏差见表2。所加入的状态与推力不确定性的标准差见表3。

结果

图 7、图 8 和表 4 总结了对接任务的 Δv 消耗和安全性能。与未调优的 ICCBF 基线（$\boldsymbol{\theta} = [0.25, 0.85, 0.05]$ 和$\alpha_V = 0.1$）相比，两种学习调优方法都降低了控制力度并减少了失败案例。MLP调参ICCBF的平均总 Δv最低，将均值从 8.20 m s^{-1} 降至 5.70 m s^{-1}，较未调参ICCBF降低了 30.3%。经 RNN 调优的 ICCBF 也显著降低了推力消耗，均值为 6.19 m s^{-1}（降低了 24.5%）。中位数和四分位数统计也一致地体现了这些改进。尽管经 RNN 调优的 ICCBF 并非平均 Δv最低，但其上尾较短，且 P_{99} 值低于MLP调参ICCBF。这表明它能更好地抵御高推力异常值，鲁棒性有所提升。

所有三个控制器在蒙特卡洛集合中的故障率都较低。未调参ICCBF的故障率为1.86%，而经 MLP 和 RNN 调优的 ICCBF 故障率分别为 1.80% 和 1.74%。经 RNN 调优的 ICCBF 故障率较低，这表明递归可能通过利用时间上下文，增强对潜在参数变化和累积扰动的鲁棒性，这与巡航控制案例类似。

5.3 三维航天器检查

最后，ICCBF调参框架应用于三维航天器检查场景，其中从航天器必须安全地绕球形主航天器目标机动，同时逐步检查其表面，如图 9 所示。主航天器被建模为半径为 R_C 的球体，从航天器的有效碰撞半径设为 R_D。主航天器表面离散为 N_p 个点，检查进度则通过满足允许的观测和照明条件的点数来量化。每检查一个新点，即可获得奖励。

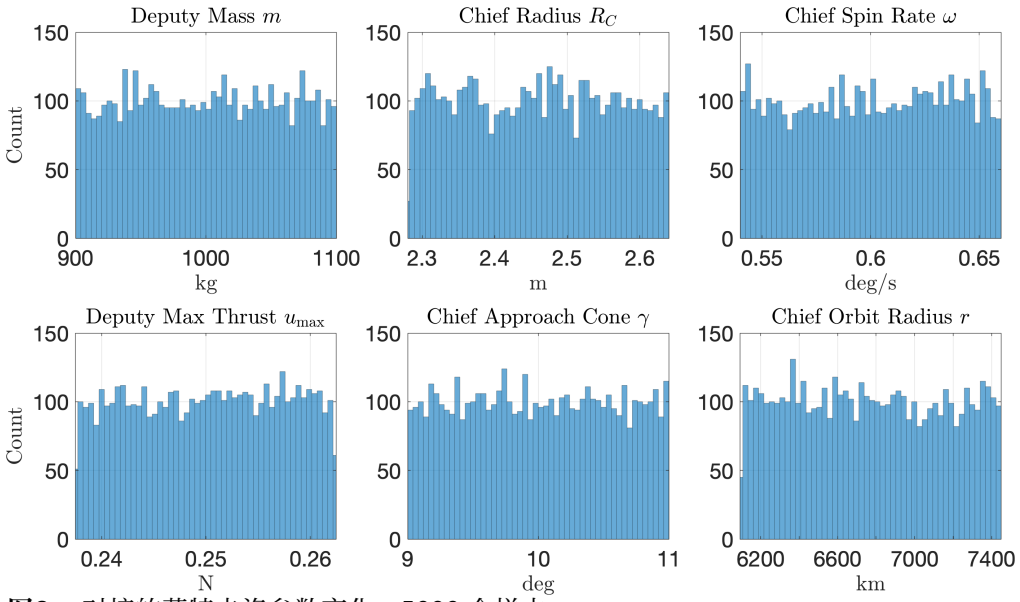

图6。 对接的蒙特卡洛参数变化，5000 个样本

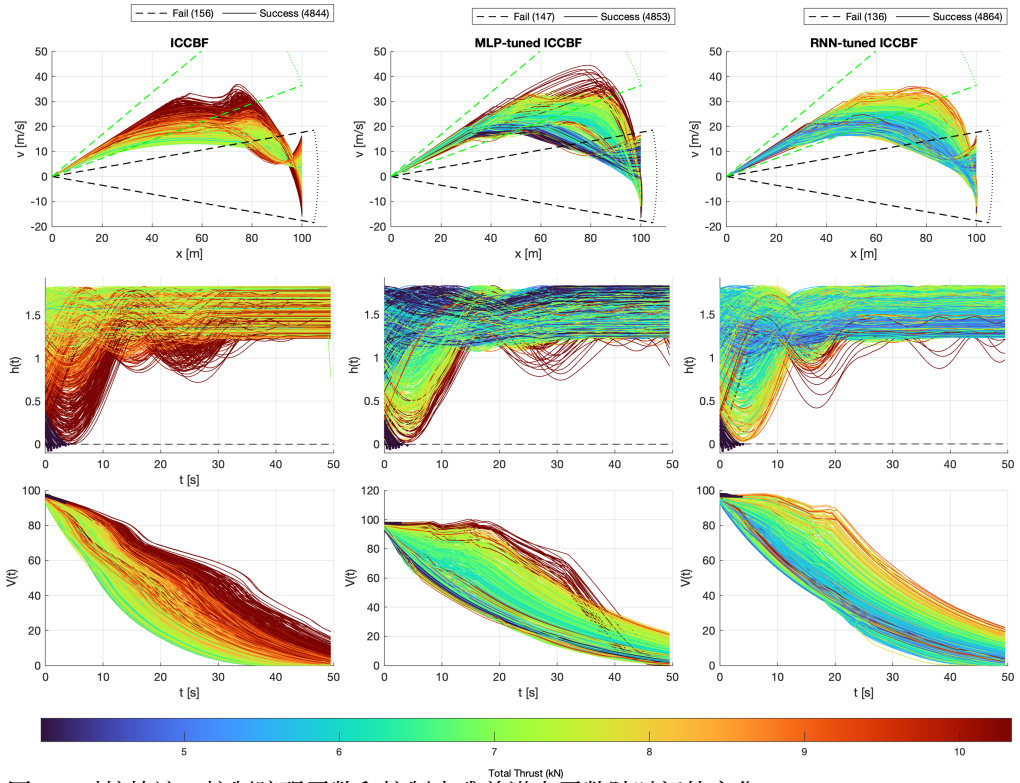

图7。 对接轨迹、控制障碍函数和控制李雅普诺夫函数随时间的变化

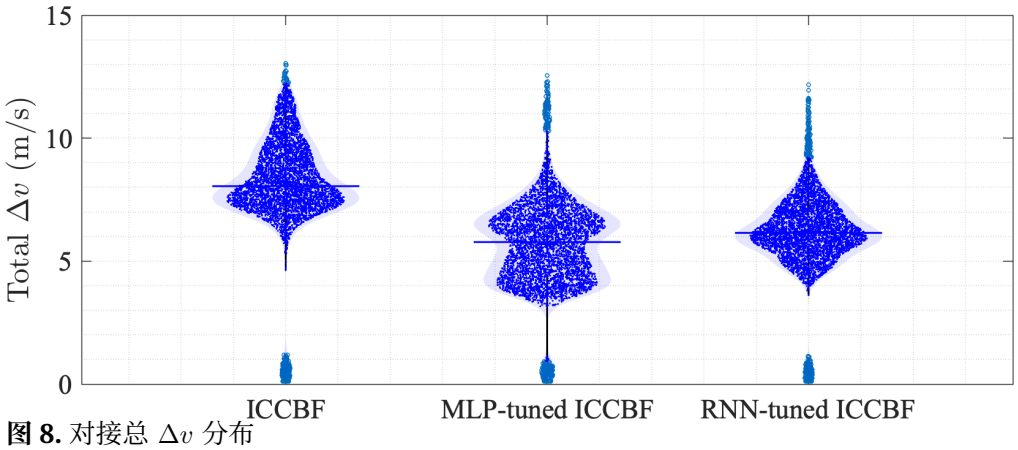

图 8. 对接总 Δv 分布

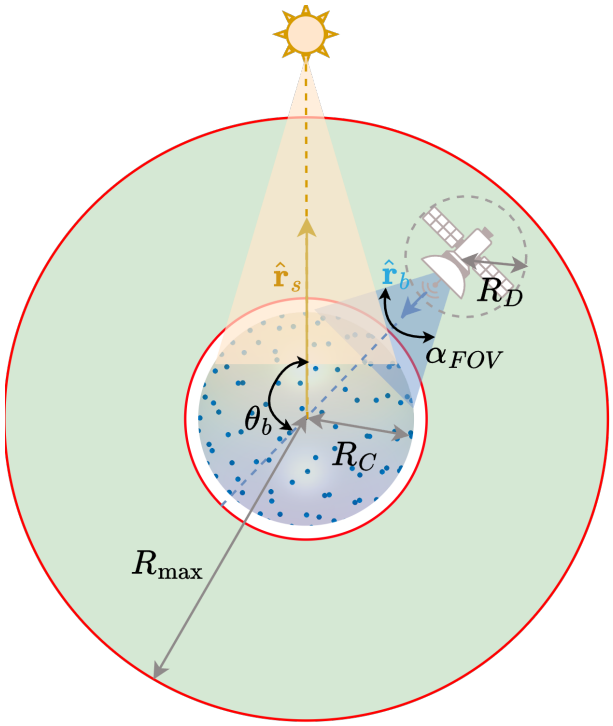

图 9. 航天器检查问题。 α_{FOV} 表示从航天器传感器的视场，而 θ_b 是传感器视轴与太阳之间的夹角。

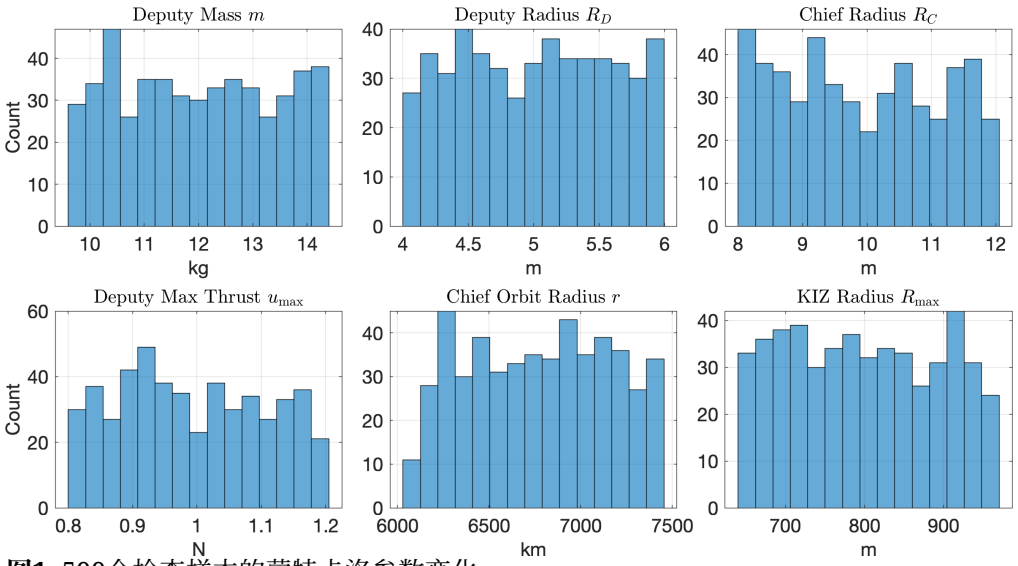

图1 500个检查样本的蒙特卡洛参数变化

请注意，该问题改编自 [27], ，其采用八个一阶控制障碍函数，并结合速度和加速度限制，以缓解安全条件仅取决于位置和/或姿态时可能出现的退化问题。相比之下，本文仅考虑KOZ、KIZ和传感器避日约束，并使用输入约束控制屏障函数来施加推力限制。

问题动力学与采样时间 从航天器状态记为$\mathbf{x} = [\mathbf{r}^\top,\ \mathbf{v}^\top]^\top \in \mathbb{R}^6$，其中$\mathbf{r} = [x\ y\ z]^\top$ 和 $\mathbf{v} = [\dot{x}\ \dot{y}\ \dot{z}]^\top$ 分别是在以主航天器为中心的局部垂直局部水平坐标系中表示的相对位置和速度。相对运动采用Clohessy–Wiltshire（Hill）方程建模，该方程具有闭式解及相应的状态转移表示 [11]。控制量为三轴推力输入$\mathbf{u} \in \mathbb{R}^3$，采用零阶保持，并受逐轴饱和约束 $\|\mathbf{u}\|_\infty \le u_{\max}$。

采样周期为 $T = 10$ s，最大时域长度 $N_{\max}$ 设为 1224 步，以匹配 [27] 中的条件。因此，最长任务时间为 $T_{\mathrm{TOF}} = N_{\max}\,\Delta t = 3.40$h；不过，如果所有点都已在适当条件下完成检查，则允许回合提前结束。

初始化 每个回合通过对初始距离 $\|\mathbf{r}_0\| \in [50, 100]$ m 进行均匀采样来初始化，并随机选取方位角和仰角，同时设置 $\mathbf{v}_0 = \mathbf{0}$ m/s。采样太阳角参数 $\theta_S \in [0, 2\pi]$，并据此定义 Hill x–y 平面内从主航天器指向太阳的太阳方向单位向量 $\hat{\mathbf{r}}^S(\theta_S)$。每个回合中变化的隐藏参数包括从航天器质量 m、从航天器半径 R_D、主航天器半径 R_C、 $u_{\max}$、 r以及 KIZ 半径 $R_{\max}$. 。这些参数的标称值和相对偏差再次列于表2，并在图10中展示。添加的推力不确定性的标准差见表3。

观测模型 如前所述，奖励根据被检查的点的数量提供。若某一点位于从航天器的视线内、处于视场角 α_{FOV} 之内、该点不处于日食中，且该点得到充分照明（既不过亮也不过暗），则认为该点可见。照明使用 Blinn–Phong 反射模型以及一个要求

表面法线与太阳方向的点积为正的阴影门控进行评估。该观测模型在 [27] 中有详细讨论。

ICCBF 安全滤波器 本研究将三项安全条件编码为 ICCBF，它们分别是：

1. 用于碰撞规避的禁入区（KOZ）约束，其中$h_1(\mathbf{x}) = \|\mathbf{r}\| - r_{coll} \geq 0$ 且 $r_{\text{coll}} = R_C + R_D$。 2. 用于强制执行 $h_2(\mathbf{x}) = R_{\max} - \|\mathbf{r}\| \geq 0$的保持区（KIZ）。
3. 一项角度条件，确保从航天器不会将其传感器对准太阳$h_3(\mathbf{x}) = \theta_b - \alpha_{FOV}/2 \geq 0$。

这些条件通过 ICCBF 递归得到强制执行，从而产生关于控制输入的仿射终端阶段约束：

$$L_f b_{i,N}(x) + L_g b_{i,N}(x)\, u + \theta_{i,N} b_{i,N}(x) \geq \nu_i(T, x), \tag{41}$$

对于 $i = 1,2,3$，其中 $\nu_i(T, x)$ 是第3.3.2节中定义的裕度。

请注意，与巡航控制和对接案例不同，在此测试用例中，检查任务的完成度无法表示为控制李雅普诺夫函数。因此，所有方法都需要学习得到的标称控制策略 $\boldsymbol{u}_{RL}$，包括未调优的 ICCBF 基线。在未调优的情况下，ICCBF 充当监督式安全滤波器，通过修改标称控制指令来确保安全性。

在每个时间步 k，安全滤波器被构造为二次规划

$$\mathbf{u}_k^\star = \arg\min_{\mathbf{u}_k \in \mathbb{R}^m} \ \|\mathbf{u}_k - \mathbf{u}_{k,\text{rl}}\|_2^2 \tag{42}$$

$$\text{s.t.} \quad \text{Eq. (41)} \tag{43}$$

$$\|\mathbf{u}_k\|_\infty \leq u_{\max}. \tag{44}$$

除了 $\boldsymbol{\theta}_k = [\theta_{1,0}, ..., \theta_{1,N}, \theta_{2,0}, ..., \theta_{2,N}, \theta_{3,0}, ..., \theta_{3,N}]^T$，强化学习 actor 在此情况下还会提供一个名义推力强化学习指令 $\mathbf{u}_{\text{k,rl}}$。

强化学习状态、奖励和终止标准 强化学习状态／观测向量与 [27] 中保持一致。它包括带有不确定性 $\boldsymbol{x}_k^E$ 的从航天器相对状态、ˆ 太阳角 θ_S、已巡检点的数量 n_{insp}，以及一个方向向量 $\mathbf{d}$，该向量指向通过 K-均值聚类启发式计算得出的、包含最多剩余未巡检点的簇。在实验中使用 $N_p = 100$ 个网格和 $K = 4$ 个簇。

在式 (34) 给出的奖励中加入一项基于已巡检点数量的奖励，使得

$$\begin{aligned} r_k = & -w_u \frac{T}{m} \|\mathbf{u}_k^\star\|_2 - w_{\text{fail}} P_{\text{fail}} - w_{h_1} \mathbb{I}\{h_{1,k+1} < 0\} \\ & - w_{h_2} \mathbb{I}\{h_{2,k+1} < 0\} - w_{h_3} \mathbb{I}\{h_{3,k+1} < 0\} \end{aligned} \tag{45}$$

$$+ w_P \sum \text{Points inspected}. \tag{46}$$

其中 $w_u = 0.1$、$w_{\text{fail}} = w_{h_1} = w_{h_2} = w_{h_3} = 1.0$ 和 $w_P = 0.1$。当 $P_{\text{fail}} = 1$ 时，回合终止；这可能是由于二次规划求解器故障或违反任何控制屏障函数约束所致，也可能是因为所有检查点均已成功检查。

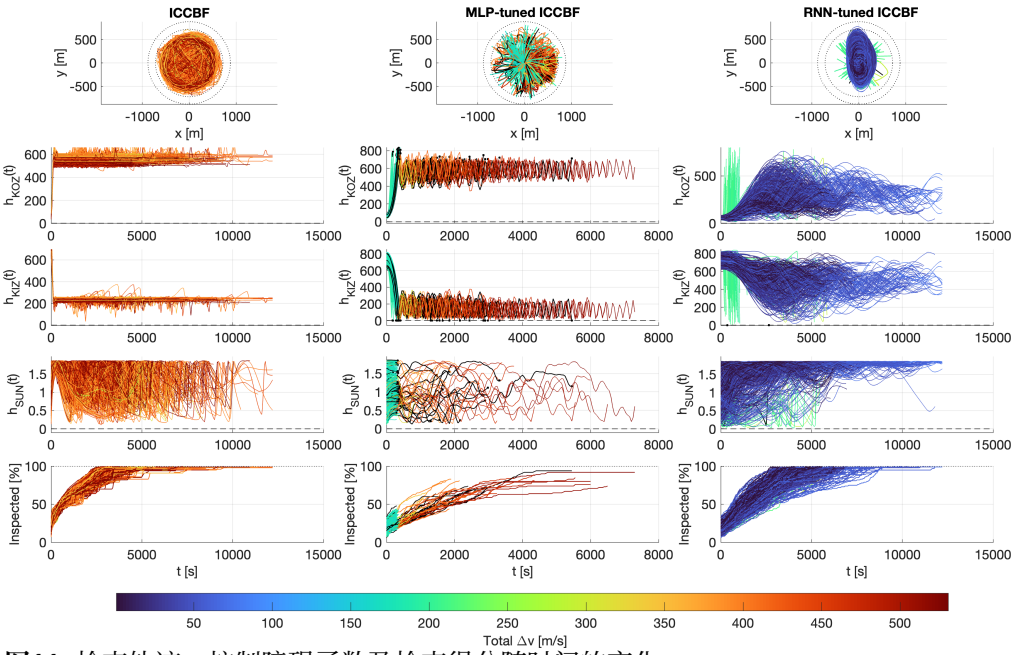

图11. 检查轨迹、控制障碍函数及检查得分随时间的变化

结果 检查结果见图 11、图 12 和表 4，清楚地展现了在调优控制屏障函数时引入递归的优势。经 RNN 调优的 ICCBF 整体表现最为出色，在消耗燃料最少的同时，几乎完成了全部检查覆盖。这一表现表明，递归策略能够利用时间上下文，并推断检查几何结构中无法在单个时间步直接观测到的潜在结构。

基线未调优 ICCBF采用固定的屏障权重（$\boldsymbol{\theta} = 0.05\,\mathbf{1}_9$）；然而，在该情形下，所施加的控制输入$u_{k,\mathrm{RL}}$ 仍通过学习获得。尽管该方法在大多数情况下都能取得较高的检查完成度，但这是以最大的总 Δv为代价的，表明鲁棒性与可观测性是通过保守控制力度实现的。相比之下，MLP调参ICCBF更激进地最小化 Δv，但这种降低是以牺牲检查性能为代价的。相当一部分回合在覆盖率不完整时终止，表明在所选奖励结构下，前馈策略倾向于优先考虑燃料效率而非任务完成度。经 RNN 调优的 ICCBF提供了最为有利的权衡。检查得分分布集中于100%，所有四分位数均达到完全覆盖，这意味着至少75%的回合实现了完整检查。与此同时，总燃料消耗保持在较低水平， P_{99} 仅为 15.27 m s^{-1}，凸显了较强的尾部性能以及各试验间的一致性。未调优情形与RNN 调优情形在所有测试情形中均保持了安全性。

最后需要指出的是，这些结果本质上受到检查进度与 Δv 最小化在奖励函数中相对权重的影响。此处报告的趋势对应于检查实验中采用的恒定权重配置；采用其他权重策略将改变任务完成度与燃料最优性之间的平衡。

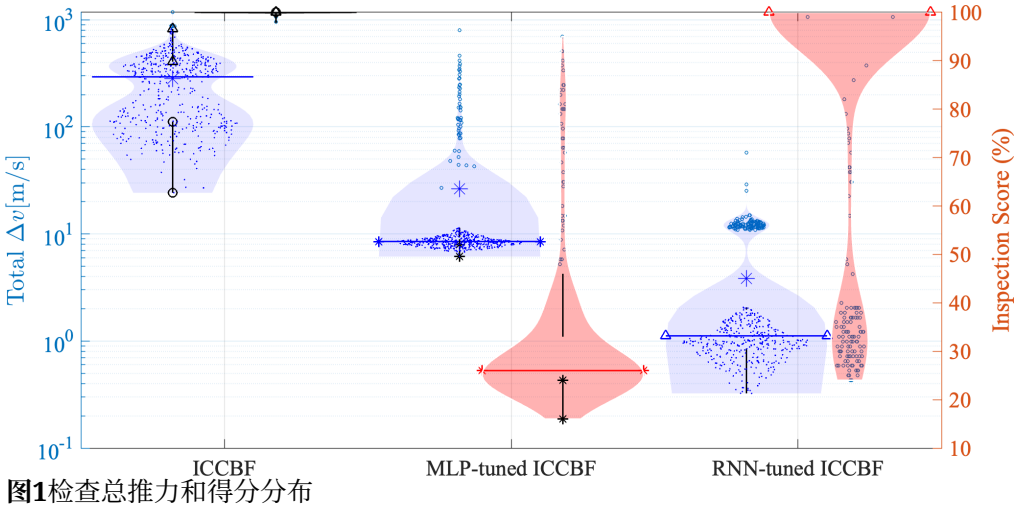

图1检查总推力和得分分布

6 结论

本文提出了一种元强化学习框架，用于调节 ICCBF，使其表现出非短视行为，并增强对状态不确定性、推力限制和隐藏参数变化的鲁棒性。对隐藏参数的鲁棒性在行星机器人作业（PRO）中尤为关键，因为目标物体的属性通常存在一定程度的不确定性。所提出的方法学习用于推导连续时间输入约束控制屏障函数的局部类-$\mathcal{K}$ 函数层级，使控制器能够在保持经验安全保证的同时，平衡安全性、燃料最优性和可行性。学习得到的调节机制在时间采样动力系统上实现，并利用通过微分代数（DA）高效计算的控制裕度，从而保持机载计算的可处理性。

在具有代表性的巡航控制、对接和检查场景中，结果表明，LSTM 等循环架构在调优 ICCBF 方面尤其有效。经 RNN 调优的 ICCBF 在保持更多可行且成功的轨迹的同时，始终能降低燃料消耗，尤其是在隐藏参数较多的情况下。特别是在巡检问题中，基于多层感知机的方法往往会以牺牲检查完整性为代价来减少 Δv；而 RNN 则能成功学习非贪心检查轨迹，在满足约束的同时降低燃料消耗并提高检查得分。这些发现凸显了时间记忆对于在结构化但部分可观测的环境中学习安全关键控制律的重要性。

参考文献

1. A. Agrawal 和 K. Sreenath。用于离散系统安全关键控制的离散控制屏障函数及其在双足机器人导航中的应用。载于机器人学：科学与系统（*RSS*），2017年。2. D. R. Agrawal 和 D. Panagou。通过输入约束控制屏障函数进行安全控制综合。载于*2021*年第*60*届*IEEE*决策与控制会议（*CDC*），第6113–6118页。IEEE，2021年。3. A. D. Ames、S. Coogan、M. Egerstedt、G. Notomista、K. Sreenath 和 P. Tabuada。控制屏障函数：理论与应用。载于*2019*年第*18*届欧洲控制会议（*ECC*），第3420–3431页，2019年。

4. A. D. Ames、J. W. Grizzle 和 P. Tabuada。基于控制屏障函数的二次规划及其在自适应巡航控制中的应用。载于第*53*届 *IEEE* 决策与控制会议，第6271–6278页，2014年。5. J. Beck、R. Vuorio、E. Liu、Z. Xiong、L. Zintgraf、C. Finn 和 S. Whiteson。元强化学习教程。人工智能基础与趋势丛书。Now Publishers，2025年。6. J. Breeden、K. Garg 和 D. Panagou。采样数据系统中的控制屏障函数。*IEEE* 控制系统快报，6:367–372，2022年。7. J. Breeden 和 D. Panagou。用于在线安全关键控制的预测控制屏障函数。载于*2022* 年 *IEEE* 第*61*届决策与控制会议（*CDC*），第924–931页。IEEE，2022年。8. S. Breeden 等。具有输入约束的采样数据系统高相对阶鲁棒控制屏障函数。*IEEE* 控制系统快报，2022年。9. S. Breeden 和 D. Panagou。存在扰动和不确定性的卫星轨迹鲁棒控制屏障函数。*Automatica*，2023年。10. B. Chachuat 和 M. Villanueva。参数化常微分方程解的界：当泰勒模型遇上微分不等式。载于计算机辅助化学工程，第30卷，第1307–1311页。爱思唯尔，2012年。11. W. H. Clohessy 和 R. S. Wiltshire。卫星交会末端制导系统。航空航天科学杂志，27(9):653–658，1960年。12. S. Dean 等。基于控制屏障函数的基于感知的安全。arXiv 预印本 arXiv:2010.16001，2020年。13. K. Dunlap、K. Bennett、D. van Wijk、N. Hamilton 和 K. Hobbs。用于六自由度航天器检查的运行时保障强化学习。美国航空航天学会，2024年。14. B. Gaudet、R. Linares 和 R. Furfaro。结合强化元学习的自适应制导与综合导航。《宇航学报》，169:180–190，2020年。15. T. Kim、R. I. Kee 和 D. Panagou。通过不确定性感知在线参数自适应学习改进输入约束控制屏障函数。载于*2025* 年 *IEEE* 国际机器人与自动化会议（*ICRA*），第3868–3875页，2025年。16. S. Kolathaya 和 A. D. Ames。基于控制屏障函数的输入到状态安全。*IEEE* 控制系统快报，2019年。17. S. Liu、W. Xiao 和 C. Belta。具有可行性保证的安全关键控制及其在异构车队中的应用。载于*2024* 年 *IEEE* 第*63*届决策与控制会议（*CDC*），第8066–8073页，2024年。18. MOSEK ApS。*MOSEK* 优化套件。MOSEK ApS，丹麦哥本哈根，2023年。第10版。文档和软件可从 https://www.mosek.com/ 获取。19. P. S. Oruganti、P. Naghizadeh 和 Q. Ahmed。采样数据系统的鲁棒控制屏障函数。*IEEE* 控制系统快报，8:103–108，2024年。

20. V. K. Oruganti 等。针对具有输入约束的采样数据系统的鲁棒控制屏障函数。*IEEE* 控制系统快报，2024年。21. H. Park、S. Di Cairano 和 I. Kolmanovsky。针对航天器交会与对接、旋转/翻滚平台以及碎片规避的模型预测控制。载于*2011*年美国控制会议论文集，第 1922–1927 页，2011年。22. A. Raffin、A. Hill、A. Gleave、A. Kanervisto、M. Ernestus 和 N. Dormann。Stable-baselines3：可靠的强化学习实现。机器学习研究杂志，22(268):1–8，2021年。23. J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017年。24. A. Singletary、Y. Chen 和 A. D. Ames。具有输入时延的采样数据系统控制屏障函数。载于*2020*年第*59*届*IEEE*决策与控制会议（*CDC*），第 804–809 页，2020年。25. R. S. Sutton 和 A. G. Barto。《强化学习：导论》。麻省理工学院出版社，第二版，2018年。26. X. Tan、W. S. Cortez 和 D. V. Dimarogonas。高阶屏障函数：鲁棒性、安全性与性能关键控制。*IEEE* 自动控制汇刊，67(6):3021–3028，2022年。27. D. van Wijk、K. Dunlap、M. Majji 和 K. Hobbs。通过深度强化学习与离散控制屏障函数实现安全的航天器巡检。航空航天信息系统杂志，21(12):996–1013，2024年。28. R. Volpe、C. Circi、M. Sabatini 和 G. B. Palmerini。利用被动单目相机对非合作翻滚目标进行最优交会的制导、导航与控制架构。《宇航学报》，196:380–393，2022年7月。29. M. C. Wijayatunga、R. Armellin 和 H. Holt。考虑安全性与可观测性的强化学习远距离交会鲁棒轨迹设计与制导。《航空航天科学与技术》，159:109996，2025年

.

30. W. Xiao、C. G. Cassandras 和 C. Belta。《基于控制屏障函数的安全自主性》。计算机科学综合讲座。施普林格，第1版，2023年。31. X. Xu、P. Tabuada 和 J. W. Grizzle。控制屏障函数的鲁棒性。载于*IFAC*混合系统分析与设计会议（*ADHS*），2015年。32. J. Zeng、Z. Li 和 K. Sreenath。利用离散时间控制屏障函数增强非线性模型预测控制的可行性与安全性。载于*2021*年第*60*届*IEEE*决策与控制会议（*CDC*），第6137–6144页。IEEE，2021年。33. J. Zeng、B. Zhang 和 K. Sreenath。采用离散时间控制屏障函数的安全关键模型预测控制。载于*2021*年美国控制会议（*ACC*），第3882–3889页，2021年。34. Z. Zhang 等。非线性电压-电流轨迹约束下并网变流器平滑瞬态运行的高级安全滤波器。*IEEE*电力电子汇刊，2023年。参见相关讨论，其中指出安全过滤器可能引入抖振。