自动化学报
ACTA AUTOMATICA SINICA

面向安全与性能动态平衡的自适应流形约束强化学习

刘正堂 吴越 胡春鹤 卢仁智 程骋 郑一力

Adaptive Manifold Constrained Reinforcement Learning for Dynamic Balance Between Safety and Performance

LIU Zheng-Tang, WU Yue, HU Chun-He, LU Ren-Zhi, CHENG Cheng, ZHENG Yi-Li

在线阅读 View online:
http://www.aas.net.cn/article/current/cbe8c48852be0507dc7d6aee536fc5632e7d5b77ed003618b867409708510d13

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

深度强化学习的攻防与安全性分析综述

A Survey of Attack, Defense and Related Security Analysis for Deep Reinforcement Learning

自动化学报. 2022, 48(1): 21−39 https://doi.org/10.16383/j.aas.c200166

性能函数引导的无人机集群深度强化学习控制方法

Performance Function-guided Deep Reinforcement Learning Control for UAV Swarm

自动化学报. 2025, 51(5): 905−916 https://doi.org/10.16383/j.aas.c240519

基于积分强化学习的学习感知型动态事件触发最优控制

Scalable Dynamic Event-triggered Optimal Control for Nonlinear Systems via Integral Reinforcement Learning

自动化学报. 2026, 52(6): 1221−1233 https://doi.org/10.16383/j.aas.c250620

基于自适应噪声的最大熵进化强化学习方法

Adaptive Noise-based Evolutionary Reinforcement Learning With Maximum Entropy

自动化学报. 2023, 49(1): 54−66 https://doi.org/10.16383/j.aas.c220103

基于个性化联邦强化学习的异构多微网能量调度

Energy Scheduling of Heterogeneous Multi-microgrid Based on Personalized Federated Reinforcement Learning

自动化学报. 2025, 51(9): 2072−2084 https://doi.org/10.16383/j.aas.c250130

面向安全与性能动态平衡的自适应流形约束强化学习

刘正堂[1, 2, 3] 吴 越[1, 2, 3] 胡春鹤[1, 2, 3] 卢仁智[4] 程 骋[4] 郑一力[1, 2, 3]

摘 要 机器人走向实际场景时, 如何严格满足安全约束并兼顾学习探索效率, 是机器人实用化的重要挑战. 现有安全学习方法依赖固定控制增益, 难以在训练中动态平衡探索效率与约束满足, 易导致过度保守或约束保持不足. 为此, 提出面向安全–性能动态平衡的状态相关增益调度自适应 ATACOM 方法, 简称 SGA-ATACOM. 在保持原有的约束流形切空间探索结构不变的基础上, 构建约束层与控制器层双层在线调度机制. 其中, 约束层依据约束状态、变化趋势及短时预测, 自适应调节可行性约束参数; 控制器层结合约束残差与速度风险指标, 更新误差修正增益和速度可行域增益, 从而实现在风险指标较高时增强约束回拉与速度保护、在风险指标较低时释放更有效的探索空间的自适应协调. 在 CircularMotion 和 Planar Air Hockey 机械臂任务上, 采用不同强化学习算法开展仿真实验. 结果表明, 该算法在不同任务和强化学习算法下具有通用性, 在维持较低约束违反的同时取得更优的安全–性能平衡. 消融实验验证了约束层与控制器层调度的互补作用. 该算法为受约束强化学习的安全–性能协同优化提供了具有几何可解释性和结构保持性的改进方案.

关键词 安全强化学习; 流形约束; ATACOM; 状态相关增益调度; 安全–性能平衡

引用格式 刘正堂, 吴越, 胡春鹤, 卢仁智, 程骋, 郑一力. 面向安全与性能动态平衡的自适应流形约束强化学习. 自动化学报, xxxx, **xx**(x): x−xx

DOI 10.16383/j.aas.c260122 **CSTR** 32138.14.j.aas.c260122

Adaptive Manifold Constrained Reinforcement Learning for Dynamic Balance Between Safety and Performance

LIU Zheng-Tang[1, 2, 3] WU Yue[1, 2, 3] HU Chun-He[1, 2, 3] LU Ren-Zhi[4] CHENG Cheng[4] ZHENG Yi-Li[1, 2, 3]

Abstract In the deployment of robots to real-world scenarios, balancing strict safety constraint satisfaction and learning exploration efficiency has become a critical challenge restricting the practical application of robotic systems. Existing safe learning methods rely on fixed control-gain designs, making it difficult to dynamically balance exploration efficiency and constraint satisfaction during training, which may lead to overly conservative behavior or insufficient constraint maintenance. To address this issue, this paper proposes a state-dependent gain-scheduled adaptive ATACOM method for dynamic safety-performance balancing, termed SGA-ATACOM. Without altering ATACOM′s original tangent-space exploration structure on the constraint manifold, a two-layer online scheduling mechanism is developed. Specifically, the constraint layer adaptively adjusts viability-constraint parameters according to the current constraint state, its variation trend, and short-term predictions, while the controller layer updates the error-correction gain and velocity-feasibility gain online based on the constraint residual and velocity-risk indicators. Thus, stronger constraint recovery and velocity protection are activated at high risk, whereas more effective exploration space is released at low risk. Simulations are conducted on the CircularMotion task and the Planar Air Hockey manipulator task using different reinforcement learning algorithms. The results show that the method exhibits good generality across different tasks and reinforcement learning algorithms, achieving a better safety-performance trade-off while maintaining low constraint violations. Ablation studies verify the complementary roles of constraint-layer and controller-layer scheduling mechanisms. Overall, the proposed method provides a geometrically interpretable and structure-preserving improvement for coordinated safety-performance optimization in constrained reinforcement learning.

Keywords Safe reinforcement learning; constraint manifold; ATACOM; state-dependent gain scheduling; safety-performance balance

Citation Liu Zheng-Tang, Wu Yue, Hu Chun-He, Lu Ren-Zhi, Cheng Cheng, Zheng Yi-Li. Adaptive manifold constrained reinforcement learning for dynamic balance between safety and performance. *Acta Automatica Sinica*, xxxx, **xx**(x): x−xx

收稿日期 2026-02-10 录用日期 2026-05-07
Manuscript received February 10, 2026; accepted May 7, 2026
国家自然科学基金 (62273053) 资助
Supported by National Natural Science Foundation of China (62273053)
本文责任编委 唐漾
Recommended by Associate Editor TANG Yang

1. 北京林业大学工学院 北京 100083 2. 林木资源高效生产全国重点实验室 北京 100083 3. 林业装备与自动化国家林业和草原局重点实验室 北京 100083 4. 华中科技大学人工智能与自动化学院 武汉 430074

1. School of Technology, Beijing Forestry University, Beijing 100083 2. State Key Laboratory of Efficient Production of Forest Resources, Beijing 100083 3. Key Laboratory of National Forestry and Grassland Administration on Forestry Equipment and Automation, Beijing 100083 4. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074

强化学习 (reinforcement learning, RL) 在解决复杂决策问题方面取得显著成功[1], 然而在机器人等实际应用中, 必须考虑安全性和对物理约束的满足[1−3]. 这些约束可能包括机械限制、安全边界、以及避免损坏环境等要求[2−3]. 传统的强化学习算法通过试错来最大化累积奖励, 往往忽视探索过程中的约束满足问题, 这在实际应用中是不可接受的[1, 4].

为在强化学习中引入约束, 研究者提出多种方法. 如, 在约束马尔科夫决策过程 (constrained Markov decision process, CMDP) 框架下, 强化学习任务可以加入成本约束或状态约束, 并通过优化方法来确保在学习过程中约束得以满足[2]. 近年来, 一些方法尝试将无模型强化学习与约束信息结合, 其中基于控制障碍函数 (control barrier function, CBF) 的方法已成为保障系统安全性的主流手段之一[3−4], 周毅等[4] 对此类方法在未知环境中的应用进行系统综述. 此外, 提出如安全层 (safe layer)、屏蔽 (shielding) 等新技术, 旨在提供更好的安全保障[1, 5−6]. 这种将约束处理与策略优化解耦的设计思路在近期研究中已证明是高效的[5−7], 例如, 董明泽等[7] 提出将神经网络与模型预测控制分离, 利用安全凸空间约束缩小探索范围, 有效提升算法在复杂动态环境中的部署能力. 然而, 这些方法通常依赖于较为复杂的策略设计, 且在多重约束的情况下可能会发生冲突[1, 5].

ATACOM (acting on the tangent space of the constraint manifold) 是由 Liu 等[8] 提出的一种解决机器人约束强化学习问题的方法. 该方法的核心思想是将约束强化学习问题转换为无约束强化学习问题, 通过在约束流形的切空间上进行探索, 从而保证在整个学习过程中满足物理和安全约束.

尽管安全性是机器人部署的先决条件, 但在高速运动控制、工业流水线作业或灾难救援等对时效性要求极高的应用场景中, 任务完成效率 (即累积奖励) 往往具有与安全性同等重要的地位. 如果仅仅为满足硬约束而牺牲过多的探索空间, 会导致机器人动作过于保守、作业效率低下, 甚至因陷入局部最优而无法完成任务. 现有的流形约束方法往往面临着“安全–性能”的零和博弈: 为确保极端情况下的绝对安全, 操作者通常被迫设定较大的误差修正增益与保守的速度限制. 这种“一刀切”的静态参数配置虽然可以保障底线安全, 却严重限制智能体在风险水平较低时减少不必要的保守修正. 因此, 如何在不破坏 ATACOM 原有约束流形几何结构的前提下, 使约束响应强度能够随系统状态和风险水平自适应变化, 从而缓解安全性与任务性能之间的静态折中, 仍是受约束强化学习中值得深入研究的问题.

针对上述问题, 本文以 ATACOM 为基础, 提出一种状态相关增益调度自适应 ATACOM (state-dependent gain-scheduled adaptive ATACOM, SGA-ATACOM) 方法. 在保持原有约束流形切空间探索、解析约束保持与误差修正控制结构不变的基础上, 本文构建约束层与控制器层双层在线调度机制. 其中, 约束层依据当前约束状态、变化趋势及短时预测结果自适应调节可行性约束参数; 控制器层结合约束残差与速度风险在线更新误差修正增益和速度可行域增益, 以增强系统对不同风险状态的响应能力.

本文的主要工作可概括为: 1) 针对 ATACOM 固定增益设定带来的参数刚性问题, 提出状态相关增益调度框架, 实现约束响应强度的在线自适应调整; 2) 构建约束层与控制器层相结合的双层调度机制, 在保持几何结构和解析控制优势的同时, 引入面向状态风险的动态调节能力; 3) 在不改变切空间探索与解析约束保持结构的前提下, 给出面向 K_f、K_g、K_c、K_q 的统一状态相关调度实现方式, 增强方法的结构保持性与可扩展性.

1 相关工作

约束强化学习是近年来机器人控制领域的重要研究方向, 旨在解决强化学习过程中的安全约束问题. 张昌昕等[9] 系统综述安全强化学习在机器人领域的应用, 指出各类方法的核心差异在于如何处理探索效率与安全边界的权衡. 现有的约束强化学习方法大致可分为三类: 基于约束马尔科夫决策过程 (CMDP) 的方法、安全探索方法以及基于约束流形的方法.

1.1 基于 CMDP 的约束强化学习

Altman[10] 首次提出 CMDP 框架, 将约束引入强化学习中, 目标是在满足期望累积成本约束的条件下最大化回报. 基于此框架, Achiam 等[11] 提出约束策略优化算法 (constrained policy optimization, CPO), 结合信任域方法确保策略更新过程中的约束满足. Liu 等[12] 开发基于内点法的策略优化算法, 有效处理多重约束问题. 另一主要方向是应用拉格朗日松弛法[13−15], 将约束优化问题转换为无约束优化问题. 此外, Chow 等[16] 提出基于 Lyapunov 函数的安全策略优化方法, 给出约束满足的理论保证. 陈谋等[17] 针对多旋翼无人机吊挂系统, 设计一种预设性能约束下的分布式控制策略, 在保证避碰的同

时严格满足系统的瞬态和稳态性能指标.

然而, 上述方法主要关注学习过程结束时的约束满足, 无法保证学习过程中的安全性, 因此不适合直接应用于实际机器人系统的在线学习任务.

1.2 安全探索方法

安全探索方法致力于在整个学习过程中满足状态依赖约束. Garcia 等[18] 基于风险函数和基线智能体设计安全探索框架. Alshiekh 等[19] 和 Hans 等[20] 分别提出屏蔽 (shielding) 和安全回退策略 (backup policy) 方法, 通过干预不安全动作来保护系统. 近期关于安全探索方法的研究进一步将环境感知信息融入该框架, 例如, 赵静等[21] 提出一种基于深度强化学习的“虚拟管道”避障方法, 通过构建虚拟的安全管道约束无人机的飞行轨迹, 有效实现复杂环境下的自主安全导航. 然而, 这些方法通常需要手动设计复杂的安全策略或依赖精确的障碍物模型, 且在多重约束的情况下可能会发生冲突.

同时, 相关研究开始将模型信息与模型无关的强化学习算法相结合. Dalal 等[22] 提出安全层方法, 分析性地寻找最接近原策略的安全动作. Cheng 等[23] 利用控制障碍函数保证探索安全性. Koller 等[24] 和 Hewing 等[25] 提出基于模型预测控制的安全探索方法. 这些方法通常需要在每一步求解受约束优化问题, 计算复杂度较高.

1.3 基于约束流形的方法

Liu 等[8] 提出的 ATACOM 方法构建一种新的安全强化学习范式, 将约束定义为流形, 并在其切空间上进行探索. 该方法通过将不等式约束转化为等式约束, 构建约束流形, 并在流形的切空间上设计安全控制器. 这种方法的核心优势在于将约束强化学习问题转换为无约束问题, 无需初始可行策略或手动设计安全回退机制, 且可与任何模型无关的强化学习算法结合.

ATACOM 的工作原理是在约束流形的切空间上构建安全动作空间, 通过在这个空间中进行探索, 可自然保证约束的满足. 该方法包含漂移补偿、收缩和切向三个关键组件, 共同作用确保系统状态收敛到约束流形并在流形上移动.

Liu 等[26] 在后续工作中进一步对 ATACOM 进行理论分析和扩展, 证明控制器能够保证系统趋近约束流形, 并分析了在扰动存在时约束违反的有界性. 他们还将 ATACOM 应用于多种机器人控制任务, 包括平面机器人导航、空气曲棍球击打和人机交互场景[27], 展示出该方法在高维复杂任务中的有效性.

尽管 ATACOM 方法为流形约束探索奠定了坚实基础, 但在面对复杂多变的动态环境时, 如何实现控制机制的自适应调节仍是一个亟待解决的难题. 固定的控制增益难以应对学习过程中时变的安全需求, 仅依赖当前状态的评估方式往往难以捕捉历史风险趋势与潜在隐患. 此外, 如何在不同学习阶段自动仲裁安全与性能的权重, 以及缓解离散时间步长带来的数值敏感性, 依然是限制算法鲁棒性与泛化能力的关键挑战. 例如, 张楠杰等[28] 提出一种分层自适应控制框架, 通过实时感知地面特性自适应调整四足机器人的质心高度与步态参数, 显著提升了机器人在复杂地形下的稳定性.

2 基本 ATACOM 算法

本节介绍基本 ATACOM 算法的主要思想与数学形式. ATACOM 的核心目标是在强化学习过程中, 将智能体的探索限制在约束流形的切空间附近, 使动作在尽可能满足状态约束的前提下仍保留足够的探索自由度. 与传统“先采样动作、再做安全修正”的思路不同, ATACOM 直接在满足约束的几何结构上定义可探索方向, 从而将受约束控制问题转化为切空间中的无约束探索问题.

2.1 问题描述

考虑连续状态、连续动作的受约束强化学习问题. 设系统状态为 $s \in \mathcal{S}$, 动作为 $a \in \mathcal{A}$, 即时奖励为 $r(s,\ a)$. 系统满足一组状态约束, 这些约束仅作用于状态中的可直接控制部分. 按照原论文的设定, 将状态分解为

$$s = [q,\ x]^{\mathrm{T}}$$

其中, $q \in \mathbf{R}^Q$ 表示直接可控变量, x 表示其余不可直接控制的状态分量. 约束仅依赖于 q, 写为

$$f(q) = 0,\ g(q) \leq 0$$

其中 $f : \mathbf{R}^Q \to \mathbf{R}^F$ 表示等式约束, $g : \mathbf{R}^Q \to \mathbf{R}^G$ 表示不等式约束, 且通常有 $F < Q$. 因此, 受约束强化学习问题可表示为

$$\begin{aligned} &\max_{\pi} \mathrm{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t,\ a_t)\right] \\ &\text{s.t.} \quad f(q_t) = 0,\ \ g(q_t) \leq 0 \end{aligned}$$

ATACOM 的基本思想是: 不让策略直接在原动作空间中任意探索, 而是先根据约束构造允许运动的切空间, 再让策略只在这一切空间中给出自由探索分量. 由此, 约束满足被编码进控制结构本身, 而不是作为事后修补项.

2.2 约束流形与切空间表示

为统一处理等式约束与不等式约束, ATACOM 为每个不等式约束引入松弛变量 $\mu \in \mathbf{R}^G$, 将原始约束改写为等式形式, 即

$$c(q,\ \mu) = \begin{bmatrix} f(q) \\ g(q) + \dfrac{1}{2}\mu^2 \end{bmatrix} = 0 \tag{1}$$

其中, μ^2 表示逐元素平方. 由此, 原本的可行域边界被嵌入到扩展空间 $(q,\ \mu)$ 中的约束流形上. 对式 (1) 求时间导数, 可得

$$\dot{c}(q,\ \mu,\ \dot{q},\ \dot{\mu}) = J_c(q,\ \mu) \begin{bmatrix} \dot{q} \\ \dot{\mu} \end{bmatrix} = 0 \tag{2}$$

其中, $J_c(q,\ \mu) = \begin{bmatrix} J_f(q) & 0 \\ J_g(q) & \mathrm{diag}\{\mu\} \end{bmatrix}$, $J_f(q)$ 与 $J_g(q)$ 分别为 $f(q)$ 和 $g(q)$ 的雅可比矩阵.

若记 $N_c(q,\ \mu) = Null(J_c(q,\ \mu))$ 为约束雅可比的零空间基, 则切空间中的任意可行速度都可写为

$$\begin{bmatrix} \dot{q} \\ \dot{\mu} \end{bmatrix} = N_c(q,\ \mu)\alpha \tag{3}$$

其中, α 是低维自由坐标. 由于 $J_c N_c = 0$, 代入式 (2) 可得 $\dot{c} = 0$. 这说明, 只要系统速度始终位于约束流形的切空间内, 约束值在连续时间意义下就不会发生一阶偏离. 由此, 原始受约束动作选择问题被转化为对切空间坐标 α 的无约束采样问题.

这一思想揭示出 ATACOM 的几何本质: 策略并不直接输出系统的完整控制量, 而是输出约束流形切空间中的自由探索分量, 系统的实际运动由“约束保持项”和“切空间探索项”共同决定.

2.3 从状态约束到可行性约束

仅在速度层面进行切空间采样并不适用于一般物理系统. 原因在于真实机器人通常要求连续的速度或加速度轨迹; 若策略直接采样 $\dot{q}$, 则难以保证控制输入的连续性. 此外, 对于不等式约束, 仅满足 $g(q) \le 0$ 还不够, 当系统接近边界时, 必须限制其朝违反方向继续演化.

为此, ATACOM 将状态约束进一步转化为可行性约束 (viability constraints), 即

$$f(q) + K_f \dot{f}(q,\ \dot{q}) = f(q) + K_f J_f(q)\dot{q} = 0 \tag{4}$$

$$g(q) + K_g \dot{g}(q,\ \dot{q}) = g(q) + K_g J_g(q)\dot{q} \le 0 \tag{5}$$

其中, K_f 与 K_g 为正定对角增益矩阵. 它们控制约束误差与约束变化率之间的耦合关系. 具体而言, 当约束偏离较大时, 这一形式会自动限制系统沿着恶化方向继续演化, 并推动其以受控速度回到可行区域.

继续引入松弛变量后, 可将可行性约束统一写为

$$c(q,\ \dot{q},\ \mu) = \begin{bmatrix} f(q) + K_f J_f(q)\dot{q} \\ g(q) + K_g J_g(q)\dot{q} + \dfrac{1}{2}\mu^2 \end{bmatrix} = 0 \tag{6}$$

对式 (6) 再求时间导数, 得到加速度层面的线性约束, 即

$$J_c(q,\ \mu) \begin{bmatrix} \ddot{q} \\ \dot{\mu} \end{bmatrix} + \psi(q,\ \dot{q}) = 0 \tag{7}$$

其中

$$J_c(q,\ \mu) = \begin{bmatrix} K_f J_f(q) & 0 \\ K_g J_g(q) & \mathrm{diag}(\mu) \end{bmatrix}$$

$$\psi(q,\ \dot{q}) = \begin{bmatrix} J_f(q)\dot{q} + K_f b_f(q,\ \dot{q}) \\ J_g(q)\dot{q} + K_g b_g(q,\ \dot{q}) \end{bmatrix}$$

而 $b_f(q,\ \dot{q}) = \dot{q}^{\mathrm{T}} H_f(q)\dot{q}$, $b_g(q,\ \dot{q}) = \dot{q}^{\mathrm{T}} H_g(q)\dot{q}$ 分别由约束函数的 Hessian 项给出. 于是, 该线性方程的一般解为

$$\begin{bmatrix} \ddot{q} \\ \dot{\mu} \end{bmatrix} = -J_c^{\dagger}(q,\ \mu) \cdot \psi(q,\ \dot{q}) + N_c(q,\ \mu) \cdot \alpha \tag{8}$$

其中, $J_c^{\dagger}$ 为伪逆矩阵, $N_c(q,\ \mu)$ 为零空间基. 式 (8) 是 ATACOM 的核心控制结构: 第 1 项负责补偿约束曲率并维持系统沿流形演化; 第 2 项则是策略可以自由选择的切空间探索分量.

2.4 速度可行域与加速度约束

除位置或几何约束外, 机械系统通常还受到速度上限与加速度上限的限制. 为防止系统在接近速度边界时发生过冲, ATACOM 进一步在加速度层面引入可行性边界. 设速度上下界分别为 $v_{\min}$,$v_{\max}$, 加速度上下界分别为 $a_{\min}$,$a_{\max}$, 则可行加速度区间定义为

$$a_u = \max\big(\min(a_{\max},\ -K_q(\dot{q} - v_{\max})),\ a_{\min}\big) \tag{9}$$

$$a_l = \min\big(\max(a_{\min},\ -K_q(\dot{q} - v_{\min})),\ a_{\max}\big) \tag{10}$$

其中, $K_q > 0$ 为固定增益. 该设计的含义是: 当速度远离边界时, 可用加速度区间较宽, 系统可以更自由地运动; 当速度逐渐接近上限或下限时, 可用加速度区间会收缩, 从而抑制继续朝危险方向加速. 由此, 速度约束被转化为随状态变化而定义的可行动力学范围.

2.5 离散采样下的误差修正

上述推导建立在连续时间框架下, 而强化学习训练与仿真控制本质上是在离散时间步长下执行的. 即使理论上系统始终沿切空间运动, 数值积分误差、控制离散化误差以及模型误差仍会导致系统逐步偏离约束流形. 在此情况下, 在控制律中加入基于约束残差的比例误差修正项, 即

$$\begin{bmatrix}\ddot{q}_E\\ \dot{\mu}_E\end{bmatrix} = -J_c^{\dagger}K_c \cdot c(q,\ \dot{q},\ \mu) \tag{11}$$

其中, K_c 为正定对角增益矩阵. 该项的作用是: 一旦离散执行导致约束残差 $c(q,\ \dot{q},\ \mu)$ 不为零, 系统便沿着约束误差减小的方向施加附加修正, 使状态重新回到约束流形附近.

将误差修正项与式 (8) 结合, 可得 ATACOM 的最终控制律, 即

$$\begin{aligned}\begin{bmatrix}\ddot{q}\\ \dot{\mu}\end{bmatrix} =& -J_c^{\dagger}(q,\ \mu)\big[K_c c(q,\ \dot{q},\ \mu) +\\ & \psi(q,\ \dot{q})\big] + N_c(q,\ \mu)\alpha\end{aligned} \tag{12}$$

也即 ATACOM 的控制策略由三部分组成: 1) 约束保持项, 用于补偿约束流形的几何曲率; 2) 误差修正项, 用于抵消离散化带来的约束偏离; 3) 切空间探索项, 用于保留策略学习所需的自由度.

最终, 系统再根据 $\ddot{q}$ 通过逆动力学模型或底层跟踪控制器转换为真实动作 a.

3 状态相关增益调度自适应 ATACOM

本文提出的状态相关增益调度自适应 ATACOM (SGA-ATACOM) 是在 ATACOM 基础上的一种结构保持型扩展.

ATACOM 通过切空间参数化、可行性约束构造、速度可行域约束以及误差修正项, 将受约束强化学习问题转化为流形切空间上的探索问题. 该框架具有清晰的几何解释, 也能够与常见的 model-free 强化学习算法直接结合. 然而, 原始方法中的若干关键增益均为固定常数, 包括可行性约束中的 K_f 和 K_g、误差修正增益 K_c, 以及速度可行域增益 K_q. 固定参数的优点是实现简单, 但其隐含假设是同一组参数能够在所有状态下同时兼顾约束保持与任务探索, 这在实际中往往过于理想化.

当系统远离约束边界且运动平稳时, 过于“激进”的固定参数设定 (如较小的 K_f、K_g 或较大的 K_c、K_q) 会使控制器长期保持较强修正, 从而压缩切空间中的有效探索范围; 而当系统逼近约束边界或速度上限时, 固定增益又可能无法提供足够快的风险响应. 为此, 本文引入一种状态相关增益调度自适应机制, 将原始固定增益扩展为由当前状态风险驱动的在线调节量. 该机制包含两个层次: 1) 约束层增益调度, 用于在线调整可行性约束中的 K_f 和 K_g; 2) 控制器层增益调度, 用于在线调整误差修正增益 K_c 以及速度可行域增益 K_q.

3.1 整体框架

本文在约束流形的切空间内, 由策略输出自由探索变量 α_t, 并通过约束保持项与误差修正项共同生成实际控制量, 改变固定增益为时变量, 从而得到

$$\begin{aligned}\begin{bmatrix}\ddot{q}\\ \dot{\mu}\end{bmatrix} =& -J_c^{\dagger}(q,\ \mu)\big(K_c(t)\cdot c(q,\ \dot{q},\ \mu) +\\ & \psi(q,\ \dot{q}; K_f(t),\ K_g(t))\big) + N_c(q,\ \mu)\alpha_t\end{aligned} \tag{13}$$

其中, $K_f(t)$、$K_g(t)$、$K_c(t)$ 和 $K_q(t)$ 均由当前状态在线更新. 这里 $K_q(t)$ 用于定义速度可行域对应的加速度截断边界.

本文提出的状态反馈增益调度自适应机制策略在切空间中探索, 约束满足主要由解析结构保证, 约束响应强度和速度保护强度根据当前风险水平在每一步进行调度, 分为约束层和控制器层两层, 以此使得智能体在严格满足安全约束的同时兼顾学习探索效率, 整体结构如图 1 所示.

3.2 约束层的状态相关增益调度

由于固定的 K_f 和 K_g 无法区分“系统远离约束边界”和“系统正在逼近约束边界”这两类状态, 因此需将其扩展为状态相关量.

对任一约束分量记为 $h(q)$, 首先定义三类局部信息:

$$r_t = h(q_t) \tag{14}$$

$$\dot{r}_t = J_h(q_t)\dot{q}_t \tag{15}$$

$$\hat{r}_{t+1} = r_t + \dot{r}_t\Delta t \tag{16}$$

其中, r_t 表示当前约束值, $\dot{r}_t$ 表示约束沿当前速度方向的变化趋势, $\hat{r}_{t+1}$ 为前向预测一步的结果. 三者共同刻画当前时刻的约束风险: r_t 反映当前离边界的距离, $\dot{r}_t$ 反映系统是否正朝着更危险的方向运动, 而 $\hat{r}_{t+1}$ 则给出短时预测下的潜在越界趋势.

对于等式约束, 风险主要来自偏离约束流形的幅值、偏离变化率以及短时预测偏离; 对于不等式约束, 风险则主要来自是否接近边界、是否正在朝违反方向运动, 以及一步预测后是否进一步逼近甚

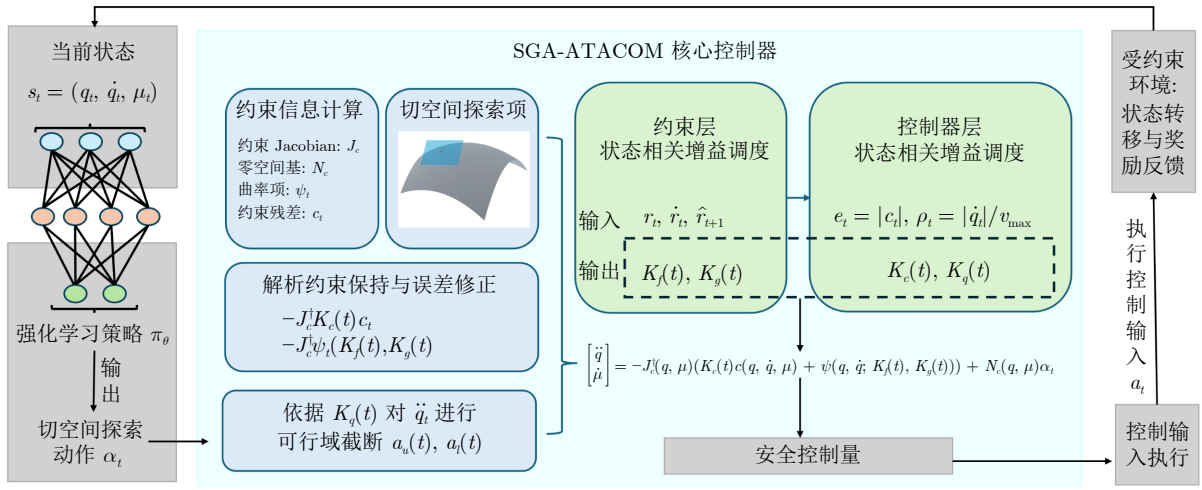


图 1 SGA-ATACOM 的整体框架

Fig. 1 The overall framework of SGA-ATACOM

至跨越边界. 基于这些量, 本文分别构造危险度指标 d_t 和松弛度指标 l_t. 其中, d_t 用于表征当前状态需要增强约束纠偏的程度, 而 l_t 用于表征系统处于安全、平稳状态并可以逐步回归基线参数的程度.

在此基础上, 约束层目标增益定义为

$$K_t^{\star}=K_0\frac{1+\lambda_l l_t}{1+\lambda_d d_t} \tag{17}$$

其中, K_0 为对应约束的基线增益, $\lambda_l>0$ 与 $\lambda_d>0$ 分别控制松弛项和危险项的影响强度. 式 (17) 表明, 当系统处于高风险状态时, 目标增益会减小; 当系统远离危险区域时, 目标增益则会逐步恢复到基线附近. 需要指出的是, 在可行性约束形式中, 较小的 K_f 或 K_g 对应更强的约束纠偏速度, 因此危险状态下增益减小并不表示约束作用减弱, 而是意味着更积极的约束回拉.

为降低离散噪声引起的高频抖动, 本文对 d_t 和 l_t 引入指数滑动平均, 分别记为 d_t 和 l_t. 进一步地, 采用不对称更新策略, 即

$$K_{t+1}=K_t+\eta_t(K_t^{\star}-K_t) \tag{18}$$

其中, 更新率 η_t 取决于目标方向. 当 $K_t^{\star}\le K_t$ 时, 采用较大的更新率以快速增强约束响应; 当 $K_t^{\star}>K_t$ 时, 采用较小的恢复率, 使增益缓慢回落, 避免系统在安全边界附近频繁切换. 通过这一机制, 约束层增益可以在“快速响应风险”与“抑制参数振荡”之间取得平衡.

3.3 控制器层的状态相关增益调度

除约束层的 K_f 和 K_g 外, ATACOM 还使用固定的误差修正增益 K_c 以及速度可行域增益 K_q. 其中, K_c 决定离散时间下约束残差的回拉强度, K_q 决定速度接近边界时加速度可行域的收缩速度. 本文进一步将其扩展为状态相关量, 以增强控制器对即时风险的响应能力.

对于误差修正增益 K_c, 采用 ATACOM 内部实际参与控制的约束残差作为风险输入, 即

$$e_t=|c(q_t,\ \dot{q}_t,\ \mu_t)| \tag{19}$$

这一选择使增益调度直接作用于控制器实际需要修正的量, 从而与式 (13) 中的误差修正项保持一致.

同时, 为刻画速度风险, 定义归一化速度指标

$$\rho_t=\frac{|\dot{q}_t|}{v_{\max}} \tag{20}$$

其分量值越接近 1, 表示系统越接近速度上限. 当 ρ_t 超过给定触发阈值时, 说明系统已进入高速度风险区域, 应增强速度保护.

基于上述两个量, 本文分别构造控制器层的两类目标增益.

对 K_q 而言, 其调度主要由速度风险驱动, 记为

$$K_q^{\star}(t)=K_{q,\,0}\ (1+\gamma_q\ \phi_q(\rho_t)) \tag{21}$$

其中, $K_{q,\,0}$ 为基线速度可行域增益, $\phi_q(\cdot)$ 为由触发阈值归一化得到的风险映射函数, γ_q 为放大系数. 速度越接近边界, K_q 越大, 对应的加速度可行域越收紧.

对 K_c 而言, 其调度同时考虑约束残差与速度风险. 记残差风险映射为 $\phi_c(e_t)$, 则

$$K_c^{\star}(t)=K_{c,\,0}\big(1+\gamma_c\max(\phi_c(e_t),\ \phi_q(\rho_t))\big) \tag{22}$$

其中, $K_{c,\,0}$ 为基线误差修正增益. 式 (22) 表明, 当约束残差显著增大时, K_c 会增强; 当速度已逼近极限, 即使当前几何残差不大, 也会通过速度风险通

道提高误差修正强度, 从而抑制潜在的离散越界.

与约束层类似, 控制器层同样采用指数滑动平均与不对称更新方式. 风险上升时增益快速提高, 风险消退时增益缓慢回落. 这一设计一方面避免由瞬时噪声引起的频繁切换, 另一方面则可保留对突发风险的快速响应能力.

3.4　加速度可行域与闭环执行

在得到更新的 $K_q(t)$ 后, 系统按照 ATACOM 的思想对加速度进行可行域截断, 即

$$a_u(t) = \max\big(\min(a_{\max},\ -K_q(t)(\dot{q}_t - v_{\max})),\ -a_{\max}\big) \tag{23}$$

$$a_l(t) = \min\big(\max(-a_{\max},\ -K_q(t)(\dot{q}_t - v_{\min})),\ a_{\max}\big) \tag{24}$$

由式 (23) 和式 (24) 可见, $K_q(t)$ 越大, 允许的加速度区间越窄. 因此, 当系统逼近速度边界时, 控制器会自动更强地抑制继续朝危险方向加速的趋势. 随后, 截断后的 $\ddot{q}_t$ 被送入底层控制接口, 转换为环境实际执行的控制输入.

3.5　算法流程

基于上述设计, SGA-ATACOM 的在线执行过程如算法 1 所示. 该方法在每个控制时刻分别对约束层增益 K_f、K_g 和控制器层增益 K_c、K_q 进行在线自适应, 并在此基础上构造控制律与速度可行域截断.

算法 1. SGA-ATACOM 算法

输入. 约束函数 f 和 g, 当前策略 $\pi_{\boldsymbol{\theta}}$, 基线增益 K_f^{base}, K_g^{base}, K_c^{base}, K_q^{base}, 时间步长 Δt, 状态相关增益调度参数 Θ.

输出. 满足约束的策略 $\pi_{\boldsymbol{\theta}}$.

1) 初始化约束层增益和控制器层增益:
$K_f \leftarrow K_f^{\text{base}}$, $K_g \leftarrow K_g^{\text{base}}$, $K_c \leftarrow K_c^{\text{base}}$, $K_q \leftarrow K_q^{\text{base}}$;
2) 初始化约束层与控制器层的平滑状态;
3) **while** 训练未结束 **do**
4)　初始化环境状态 s_0;
5)　**for** 每个时间步 $t = 0, \cdots, T-1$ **do**
6)　　根据当前策略采样切空间动作 $\boldsymbol{\alpha}_t \sim \pi_{\boldsymbol{\theta}}(\cdot \mid s_t)$;
7)　　计算当前约束风险指标、约束变化趋势及一步前向预测量:
$$e_t = \begin{bmatrix} f(q_t) \\ g(q_t,\ \dot{q}_t) \end{bmatrix},\ \dot{e}_t = \frac{e_t - e_{t-1}}{\Delta t}$$
$$\hat{e}_{t+1} = e_t + \Delta t\ \dot{e}_t;$$
8)　　根据 e_t、$\dot{e}_t$ 和 $\hat{e}_{t+1}$ 更新约束层增益 K_f 和 K_g;
9)　　构造当前约束残差 $\boldsymbol{c}_t$、约束 Jacobian 矩阵 J_c、切空间基 N_c 和曲率补偿项 ψ_t;
10)　　根据约束残差和速度风险更新控制器层增益 K_c 与 K_q;
11)　　计算 ATACOM 结构下的控制输入:
$$\begin{bmatrix} \ddot{q}_t \\ \dot{\mu}_t \end{bmatrix} = -J_c^{\dagger}\left(K_c c_t + \psi_t\right) + N_c \alpha_t;$$
12)　　根据速度调度增益 K_q 对 $\ddot{q}_t$ 进行速度可行域截断:
$$\ddot{q}_t \leftarrow \text{clip}_{\mathcal{V}(K_q)}\left(\ddot{q}_t\right);$$
13)　　执行控制输入, 获得下一时刻状态 s_{t+1} 与即时奖励 r_t, 并存储轨迹数据;
14)　**end for**
15)　利用采样轨迹更新策略参数 θ;
16) **end while**
17) **return** $\pi_{\boldsymbol{\theta}}$.

4　仿真实验

为验证 SGA-ATACOM 方法的有效性, 本文在两个具有代表性的受约束强化学习任务上开展仿真实验, 分别为二维圆周运动任务和平面空气曲棍球任务. 前者具有清晰的几何约束结构, 便于分析方法在约束流形附近的行为特性; 后者具有更高的状态维数和更复杂的机器人约束, 可用于检验方法在复杂控制场景中的适用性与泛化能力. 本文重点关注以下三个问题: 1) 所提出方法能否在不显著破坏约束的前提下提升任务性能; 2) 该方法在不同强化学习算法下是否具有一致的收益; 3) 约束层增益调度与控制器层增益调度分别起到何种作用.

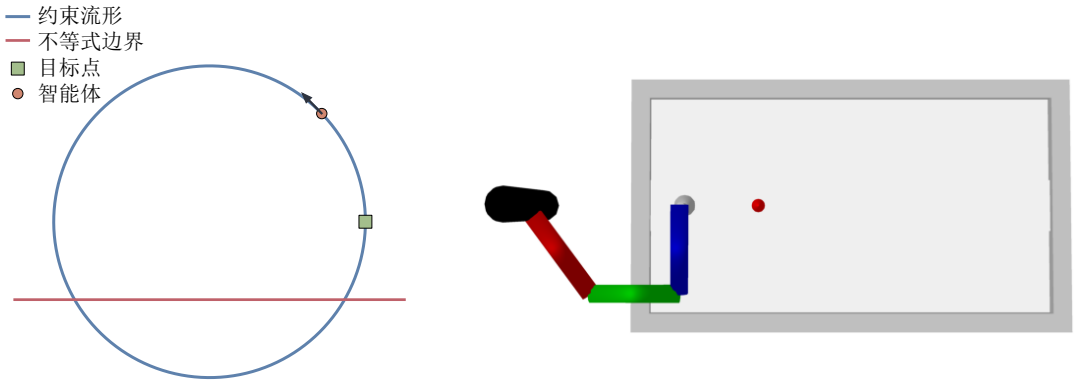
4.1　实验设置

4.1.1　实验环境

本文采用两个仿真环境进行评估, 其示意如图 2 所示. 第 1 个环境为 CircularMotion 任务[8], 如图 2(a) 所示. 系统在二维平面内运动, 目标是在满足约束的前提下沿单位圆附近完成控制任务. 该环境同时包含等式约束与不等式约束: 等式约束要求状态保持在单位圆流形附近; 不等式约束限制系统不能越过下方边界. 此外, 系统还受到速度上限约束. 因此, 该环境能够较为完整地体现可行性约束、误差修正以及速度可行域控制三者的耦合关系, 适合作为分析所提方法作用机理的基础平台.

在 CircularMotion 环境中, 可直接控制状态取值为 $q = [x,\ y]^{\mathrm{T}}$, 其中, x, y 表示质点在二维平面中的位置, 对应不等式约束引入的 μ 为松弛变量. 该环境中策略每个离散控制步输出一次 α, 随后经式 (12) 计算得到 $\ddot{q}$, 并将其作为环境中的二维加速度控制量 $a = [\ddot{x},\ \ddot{y}]^{\mathrm{T}}$ 执行. 离散时间步长设为 $\Delta t = 0.01$ s.

第 2 个环境为 Planar Air Hockey 任务[8], 如



(a) 圆周运动环境
(a) Circular motion environment

(b) 平面空气曲棍球环境
(b) Planar Air Hockey environment

图 2 仿真实验环境示意图
Fig. 2 Schematic diagram of the simulation environments

图 2(b) 所示. 该环境采用一个 3 关节平面机器人执行 air hockey 击球任务. 机器人在将球击向目标区域的同时, 需要始终满足多类不等式约束, 包括末端执行器保持在球台工作区域内、关节位置不超过各自限制以及关节速度不超过各自限制. 该环境的控制输入为基于逆动力学模型得到的关节力矩. 与 CircularMotion 中的二维质点加速度控制任务相比, Planar air hockey 同时引入机器人动力学、关节空间约束和末端执行器工作空间约束, 因此更接近实际机械臂受约束控制问题, 可用于评估所提方法在复杂机器人场景中的适用性与泛化能力. 按照原始环境设定, 实验中假设动力学模型已知, 以减少由跟踪误差或模型失配引起的额外约束违反.

4.1.2 对比方法

为全面评估所提方法, 本文在不同环境中设置相应的对比方法. 其中, 在 CircularMotion 环境中, 选取 ATACOM、SGA-ATACOM、ErrorCorrection 和 Terminated 四类方法作为比较对象. 在对比方法中, ATACOM 使用固定的 K_f、K_g、K_c 和 K_q; 本文方法将上述增益扩展为状态相关的在线调度量; ErrorCorrection 方法仅保留误差修正思想而不包含完整的切空间约束结构; Terminated 方法则通过违反约束时的提前终止实现约束处理.

在 Planar Air Hockey 环境中, 选取 ATACOM、SGA-ATACOM 和 Unconstrained 三类方法进行比较. 在对比方法中, Unconstrained 方法不显式引入 ATA-COM 的约束控制结构, 可用于观察不施加解析约束控制时的性能与违反约束行为.

4.1.3 强化学习算法与评价指标

在 CircularMotion 环境中, 分别采用 DDPG、PPO 和 TRPO 三种强化学习算法; 在 Planar Air Hockey 环境中, 采用 PPO 和 TRPO 两种算法. 所有结果均基于 5 个随机种子重复实验获得, 以降低训练随机性带来的偶然影响.

本文采用四项指标对方法进行评估: 平均回报 R 用于衡量任务性能; 平均约束违反程度 c_{avg} 用于表征训练过程中整体约束保持水平; 最大约束违反程度 $c_{\max}$ 用于刻画最坏情况下的约束偏离; 最大速度约束违反程度 $c_{\dot{q},\max}$ 用于衡量速度可行域控制效果. 其中, R 越大表示任务性能越优, 而 c_{avg}、$c_{\max}$ 和 $c_{\dot{q},\max}$ 越小表示约束满足能力越强. 本文的定量分析对 R 和 c_{avg} 采用多随机种子最终评估结果的均值进行统计, 而 $c_{\max}$ 和 $c_{dq,\max}$ 用于描述最坏情况下的约束偏离. 因此本文报告其在多随机种子中的峰值, 以更准确地反映方法的安全边界表现.

除环境参数与强化学习训练参数外, SGA-ATACOM 还包含约束层与控制器层两类状态相关调度超参数. 考虑到这些参数直接决定增益随风险状态的调节方式, 本文将其关键取值汇总于表 1.

4.2 CircularMotion 环境实验结果

4.2.1 不同约束处理方法的比较

图 3 给出了 CircularMotion 环境下 DDPG、PPO 和 TRPO 三种强化学习算法在不同约束处理方法上的训练曲线对比, 表 2 汇总了对应的最终统计结果. 可以看到, 所提出的 SGA-ATACOM 在三种算法下均同时提升了任务回报与约束满足能力, 且整体优于 ATACOM、ErrorCorrection 和 Terminated.

具体而言, 对于 DDPG, SGA-ATACOM 的平均回报由 ATACOM 的 235.7015 提升至 247.1787, 同时 $c_{\max}$ 由 0.031 223 降至 0.012 514, 降幅约为

59.92%, c_{avg} 由 0.006 402 降至 0.003 576, 如图 3(a) ~ 3 (d) 所示. 对于 PPO, 平均回报由 278.928 2 提升至 300.638 1, $c_{\max}$ 由 0.031 222 降至 0.012 518, 降幅约为 59.91%, c_{avg} 由 0.006 523 降至 0.002 561, 如图 3(e) ~ 3(h) 所示. 对于 TRPO, 平均回报由 264.6153 提升至 293.580 4, $c_{\max}$ 由 0.031 204 降至 0.012 515, 降幅约为 59.89%, c_{avg} 由 0.006 437 降至 0.003 045, 如图 3(i) ~ 3(l) 所示.

与 ErrorCorrection 相比, 所提方法在性能和约束控制两方面均具有明显优势. 例如, 由表 2 可知, 在 PPO 下, ErrorCorrection 的平均回报仅为 192.177 3, 且 $c_{\max}$ 达到 0.139 112; 相比之下, 所提方法的平均回报为 300.638 1, $c_{\max}$ 仅为 0.012 518. 与 Terminated 相比, 所提方法的优势更加明显. Terminated 虽然通过终止机制抑制部分严重违反约束的行为, 但训练过程更不稳定, 且性能显著偏低. 例如, 由表 2 可知, 在 DDPG 下, Terminated 的平均回报为−95.739 3, 同时 $c_{\dot{q}\max}$ 达到 0.498 038. 这说明, 在连续控制任务中, 单纯依赖违反约束后的终止机制难以替代将约束信息显式嵌入控制律的解析方法.

表 1　SGA-ATACOM 的关键增益调度参数

Table 1　Key gain-scheduling parameters of SGA-ATACOM

实验环境	参数类别	约束层		控制器层	
		K_f	K_g	K_c	K_q
CircularMotion	基线增益	0.1	2.0	100	20
	相对增益范围	[0.35, 1.0]	[0.40, 1.0]	[1.0, 2.0]	[1.0, 2.0]
	更新速率	0.25	0.30	0.15	0.20
Planar Air Hockey	基线增益	0.5	1.0	240	$2a_{\max}/v_{\max}$
	相对增益范围	[0.45, 1.0]	[0.50, 1.0]	[1.0, 1.8]	[1.0, 1.8]
	更新速率	0.18	0.16	0.12	0.16

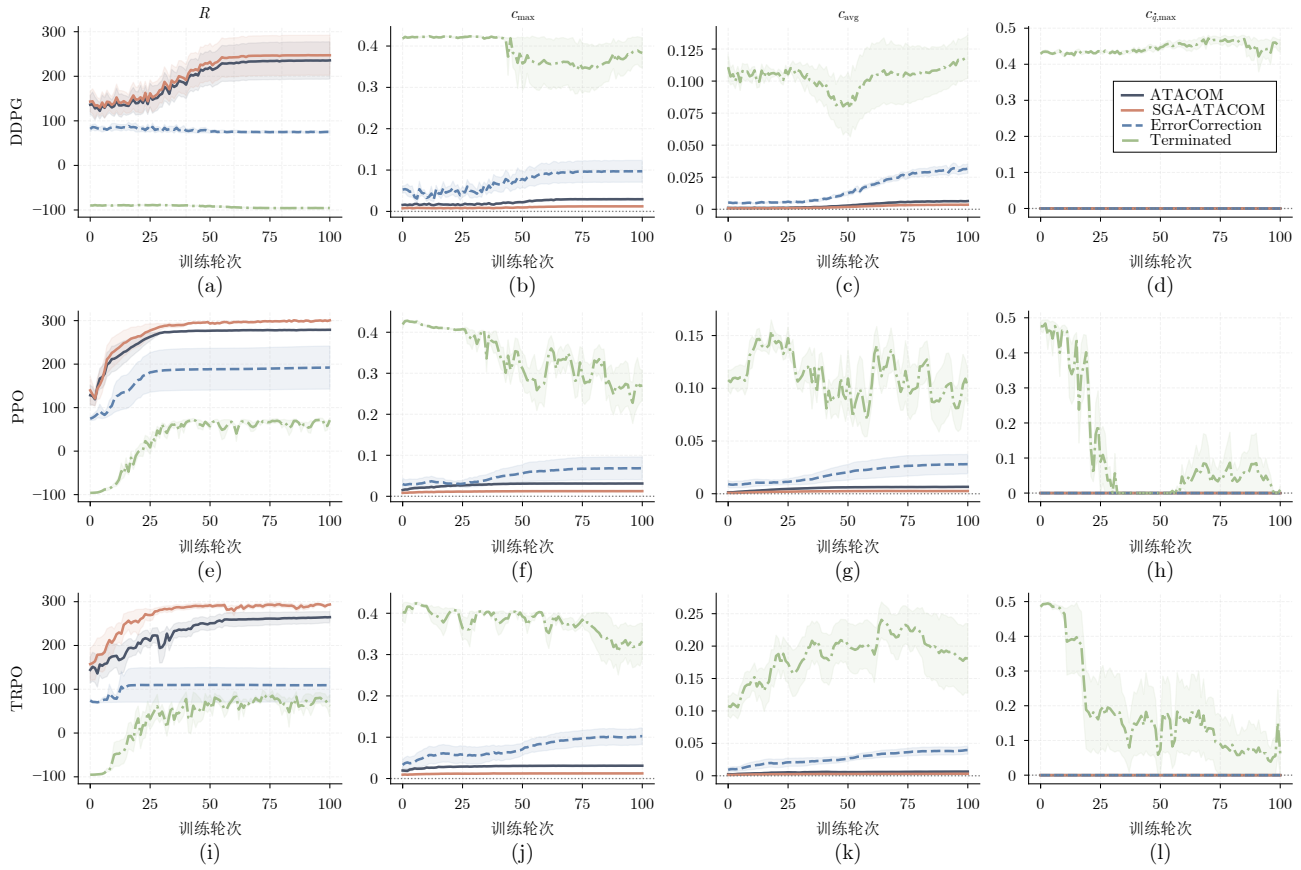


图 3　CircularMotion 环境下不同方法的性能与约束对比

Fig. 3　Comparison of performance and constraints of different methods in the CircularMotion environment

表 2 CircularMotion 环境下不同方法的定量结果

Table 2 Quantitative results of different methods in the CircularMotion environment

算法	方法	R	c_{avg}	$c_{\max}$	$c_{\dot{q},\max}$
DDPG	ATACOM	235.701 5	0.006 402	0.031 223	0
	SGA-ATACOM	247.178 7	0.003 576	0.012 514	0
	ErrorCorrection	74.954 0	0.031 649	0.140 175	0
	Terminated	−95.739 3	0.118 951	0.428 035	0.498 038
PPO	ATACOM	278.928 2	0.006 523	0.031 222	0
	SGA-ATACOM	300.638 1	0.002 561	0.012 518	0
	ErrorCorrection	192.177 3	0.028 069	0.139 112	0
	Terminated	71.208 9	0.101 304	0.401 823	0.046 248
TRPO	ATACOM	264.615 3	0.006 437	0.031 204	0
	SGA-ATACOM	293.580 4	0.003 045	0.012 515	0
	ErrorCorrection	109.174 2	0.039 597	0.135 171	0
	Terminated	62.178 4	0.180 728	0.405 737	0.313 938

综合图 3 与表 2 可以看出, 在 CircularMotion 环境中, SGA-ATACOM 在 DDPG、PPO 和 TRPO 下均取得更高的平均回报, 且同时降低了平均约束违反程度和最大约束违反程度. 相较于 Terminated 和 ErrorCorrection, 所提方法在连续控制任务中表现出更稳定的约束保持能力. 在具有清晰几何约束结构的任务中, 状态相关增益调度自适应能够更有效地调节约束响应强度与探索强度之间的关系, 从而减小约束偏离并提升任务表现.

4.2.2 强化学习算法与评价指标

图 4 展示了 SGA-ATACOM 在 CircularMotion 环境下结合 DDPG、PPO 和 TRPO 时的训练曲线. 由于该任务具有清晰的几何约束结构, 不同算法下的训练过程能够更直接地反映所提方法对约束流形附近运动的稳定调节能力. 可以看出, 尽管三种算法在收敛速度和训练波动性方面存在差异, 但所提方法在三者下均能够维持较低的约束违反, 并持续提升任务回报. 该结果表明, SGA-ATACOM 在约束结构明确的低维任务中具有稳定的跨算法适配能力.

4.3 Planar Air Hockey 环境实验结果

4.3.1 与不同方法的比较

在 Planar Air Hockey 环境中, PPO 和 TRPO 两种算法在不同方法上的训练结果对比如图 5 所示, 表 3 汇总了对应的最终统计结果. 具体而言, 在 PPO 下, SGA-ATACOM 的平均回报达到 525.2286, 明显高于 ATACOM 的 440.1124 和 Unconstrained 的 280.5240. 同时, ATACOM 与所提方法在 c_{avg}、$c_{\max}$ 和 $c_{\dot{q}\max}$ 上均保持为 0, 而 Unconstrained 方法则表现出显著的速度违反, $c_{\dot{q}\max}$ 达到 1.178 097. 该结果表明, 在该任务下, 所提方法在保持零约束违反的同时进一步提升了任务性能.

从表 3 和图 5 可以看出, 在 Planar Air Hockey 环境中, SGA-ATACOM 在 PPO 和 TRPO 下均取得高于基线方法的平均回报. 需要指出的是, 在个

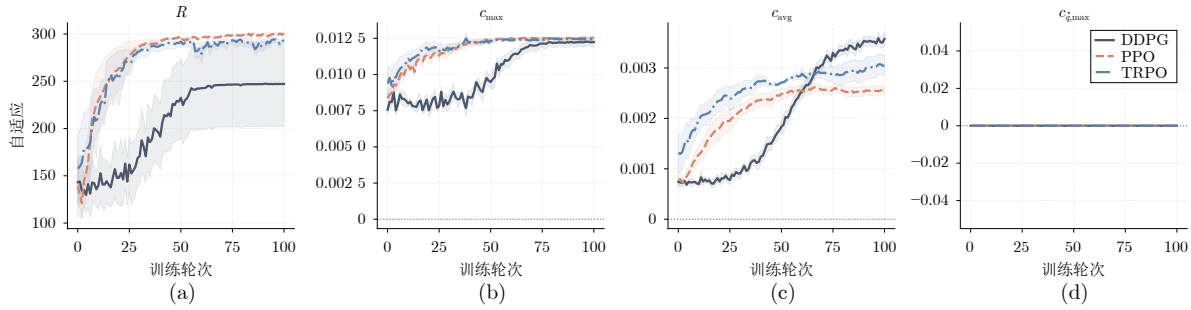


图 4 CircularMotion 环境下 SGA-ATACOM 在不同强化学习算法上的训练曲线对比

Fig. 4 Training curves of SGA-ATACOM under different reinforcement learning algorithms in the CircularMotion environment

别实验设置下, SGA-ATACOM 的 $c_{dq,\max}$ 峰值可能高于原始 ATACOM. 考虑到 $c_{\dot{q}\max}$ 反映的是最坏情况下的瞬时速度违反, 该现象主要对应少数高风险时刻的局部过冲, 而非整体约束控制能力的退化. 结合 R 与 c_{avg} 的结果可以看出, SGA-ATACOM 仍在整体上得到了更高的任务回报和相同水平的平均约束违反, 这说明该方法的收益主要体现在整体控制质量的提升, 而最坏情况下的瞬时速度边界行为仍存在进一步改进空间.

在 TRPO 下, SGA-ATACOM 在保持极低速度约束违反的条件下取得明显更高的平均回报, 说明状态相关增益调度自适应在复杂动力学任务中能够更有效地调节约束响应强度与探索强度之间的关系. 这表明, 在更高维、更复杂的机器人约束场景中, 所提方法能够在维持约束满足的基础上进一步提升任务性能.

4.3.2　强化学习算法与评价指标

图 6 给出了 SGA-ATACOM 在 Planar Air Hockey 环境下结合 PPO 和 TRPO 时的训练曲线. 与 CircularMotion 环境相比, 该任务具有更高的状态维数和更复杂的约束耦合关系, 因此更适合用于考察所提方法在复杂场景中的泛化能力. 可以看出, 尽管两种算法具有不同的更新机制, SGA-ATACOM 在二者下均表现出较稳定的训练趋势和良好的约束控制能力. 该结果表明, SGA-ATACOM 在更复杂的高维机器人任务中仍具有较好的方法泛化能力.

4.4　消融实验

为分析所提方法中不同组成部分的作用, 本文在 CircularMotion 环境中进一步进行消融实验, 分别比较 ATACOM、仅启用控制器层增益调度、仅启用约束层增益调度以及同时启用两层调度的完整方法. 实验基于 TRPO 开展, 所有结果均基于 5 个随机种子统计得到, 结果如图 7 所示, 定量结果见表 4.

从表 4 可以看出, ATACOM 的平均回报为 264.615 3, 同时对应 $c_{\max}=0.031\,204$ 和 $c_{\text{avg}}=0.006\,44$.

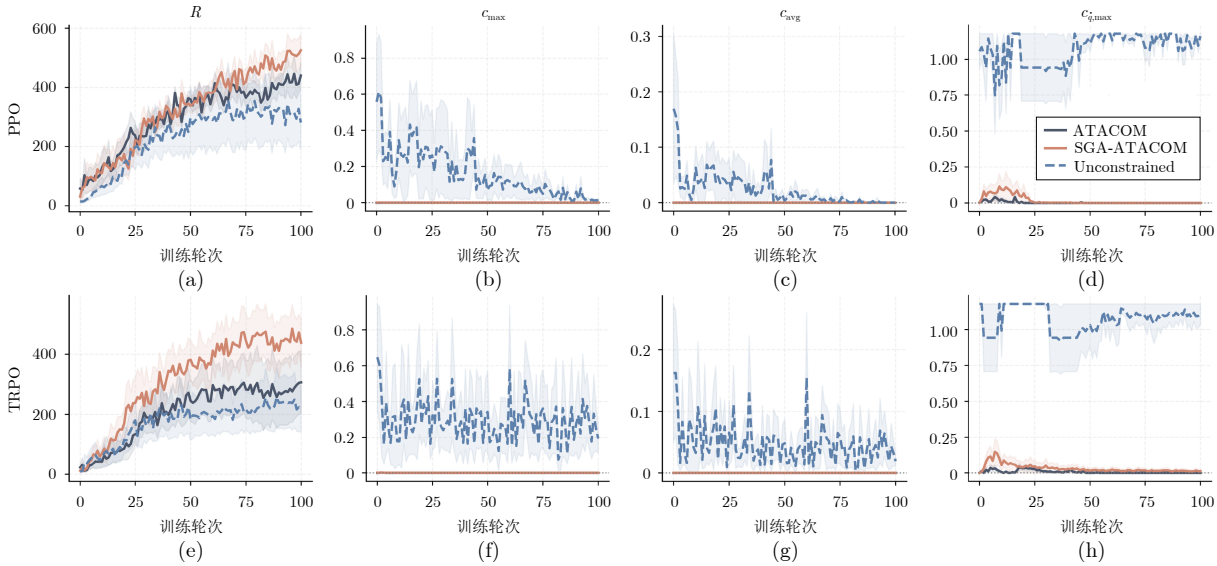


图 5　Planar Air Hockey 环境下不同方法的性能与约束对比

Fig. 5　Performance and constraint comparison of different methods in the Planar Air Hockey environment

表 3　Planar air hockey 环境下不同方法的定量结果

Table 3　Quantitative results of different methods in the Planar Air Hockey environment

算法	方法	R	c_{avg}	$c_{\max}$	$c_{\dot{q},\max}$
PPO	ATACOM	440.112 4	0	0	0
	SGA-ATACOM	525.228 6	0	0	0
	Unconstrained	280.524 0	0.000 182	0.033 132	1.178 097
TRPO	ATACOM	306.453 7	0	0	0
	SGA-ATACOM	438.229 5	0	0	0.055 845
	Unconstrained	215.184 6	0.019 395	0.374 530	1.178 097

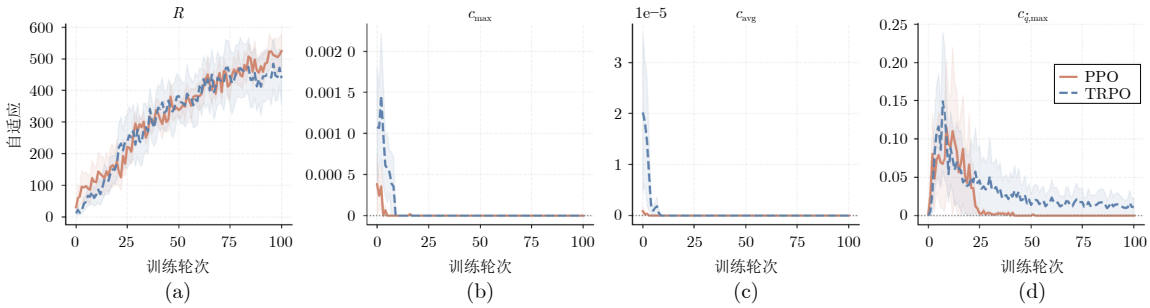


图 6 Planar Air Hockey 环境下 SGA-ATACOM 在不同强化学习算法上的训练曲线对比

Fig. 6 Comparison of training curves of SGA-ATACOM on different reinforcement learning algorithms in the Planar Air Hockey environment

仅启用控制器层增益调度时, $c_{\max}$ 降至 0.024 616, c_{avg} 降至 0.004 486, 说明控制器层调度能够直接增强误差修正与速度保护能力, 但平均回报下降至 254.084 3. 该现象表明, 仅增强 K_c 和 K_q 虽然能够抑制约束偏离, 但却会使控制器在离散执行过程中表现得更加保守. 由于约束层的 K_f 和 K_g 仍保持固定, 可行性约束的动态响应并未同步调整, 导致约束层与控制器层之间出现一定程度的动态失配, 从而削弱了切空间中的有效探索能力.

仅启用约束层增益调度时, $c_{\max}$ 进一步降至 0.015 066, c_{avg} 降至 0.004 151, 表明约束层调度对约束边界附近的主动纠偏作用更为明显, 但其平均回报为 235.940 3, 且方差相对较大. 这说明仅对 K_f 和 K_g 进行状态相关调节, 虽然可以使系统在接近约束边界时更积极地回拉到可行区域, 但由于误差修正增益 K_c 和速度可行域增益 K_q 仍保持固定, 约束层的时变响应无法在控制器层得到同步支撑. 特别是在离散时间执行条件下, 这种不完全自适应会降低闭环响应的一致性, 使系统更容易出现保守恢复或任务推进不足的现象, 因此平均回报反而低于原始 ATACOM.

相比之下, 本文方法的平均回报达到 293.580 4, 为四种配置中的最高回报, 同时 $c_{\max} = 0.012\,515$, $c_{\text{avg}} = 0.003\,045$, 在性能和约束指标上均取得最佳结果. 该结果表明, 控制器层调度与约束层调度并非两个彼此独立、可简单叠加的增强项, 而是共同作用于同一闭环控制系统的耦合机制. 前者主要增强误差修正与速度保护能力; 后者主要改善可行性约束本身的动态响应特性. 若仅启用其中一层调度, 虽然能够在某一方面降低约束违反, 但会破坏原始 ATACOM 中约束响应、误差修正与动作可行域之间的平衡关系, 因此难以同时获得更高回报和更低约束偏离. 只有两层调度共同作用时, 系统才能在高风险状态下快速增强保护, 并在安全区域保持较高的有效探索能力, 从而实现最佳的综合表现.

4.5 增益调度可视化分析

为进一步从时序层面展示状态相关增益调度自适应机制的动态作用方式, 本节进行增益调度可视化分析. 考虑到图 7 和表 4 已基于 CircularMotion 环境下 TRPO 算法的 5 个随机种子统计结果给出不同配置间的定量比较, 本节进一步选取相同实验设置下一次单次训练运行的结果, 对训练过程中的增益演化趋势以及单个回合内状态指标与增益参数之间的对应关系进行展示.

图 8 展示了训练过程中的增益演化曲线. 可以

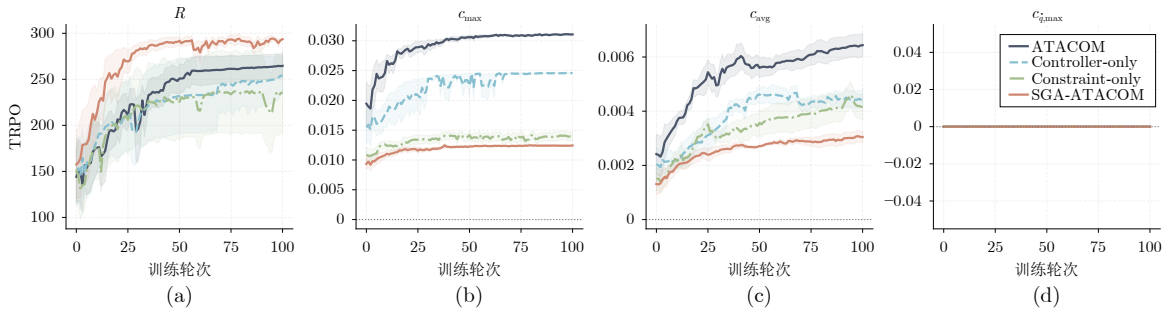


图 7 CircularMotion 环境下 SGA-ATACOM 的消融实验结果

Fig. 7 Ablation results of SGA-ATACOM in the CircularMotion environment

表 4　CircularMotion 环境下消融实验的定量结果
Table 4　Quantitative results of ablation experiments in CircularMotion environment

方法	R	c_{avg}	$c_{\max}$	$c_{\dot{q},\max}$
ATACOM	264.615 3	0.006 437	0.031 204	0
仅控制器层调度	254.084 3	0.004 486	0.024 616	0
仅约束层调度	235.940 3	0.004 151	0.015 066	0
SGA-ATACOM	293.580 4	0.003 045	0.012 515	0

看出, 所引入的各类增益参数并非固定常数, 也并非沿训练轮次作简单的单调递增或递减变化. 整体而言, 控制器层相关增益在训练前中期表现出较明显的阶段性调整, 随后逐步进入相对稳定区间; 约束层相关增益则总体保持在有界范围内缓慢变化. 该结果表明, 本文所采用的参数更新方式在训练过程中随系统状态与局部风险水平的变化发生阶段性重配置, 从而使固定增益 ATACOM 中的静态参数设定扩展为状态相关的动态调度过程.

图 9 展示了单个回合内的约束层状态指标与约束层增益之间的对应关系. 可以看出, 约束层会随约束状态、变化趋势及短时预测信息所表征的局部风险水平发生在线调整. 当系统表现出更明显的边界逼近趋势或更高的潜在违反风险时, 对应的约束层增益会发生相应变化, 以增强可行性约束的主动纠偏能力; 当系统逐渐远离危险边界、约束状态趋于平稳后, 相关增益又逐步回落. 该结果表明, 约束层调度能够在状态层面对潜在风险进行前馈式响应.

图 10 展示了单个回合内控制器层状态指标与增益参数之间的对应关系. 可以看到, 当速度风险 ρ_t 处于较高水平时, 速度可行域增益 $K_q(t)$ 相应提

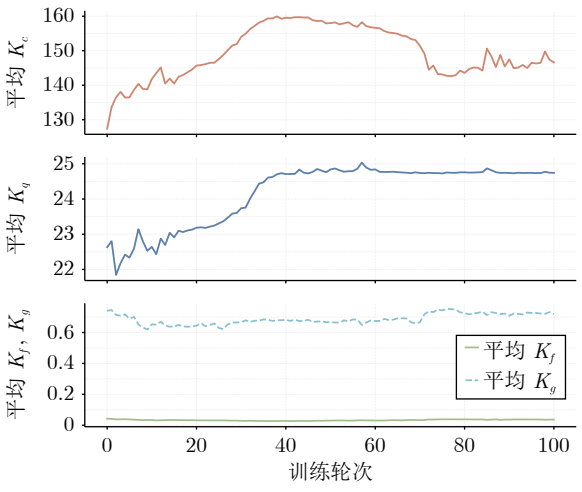


图 8　训练过程中的增益演化曲线
Fig. 8　Gain evolution curve during training

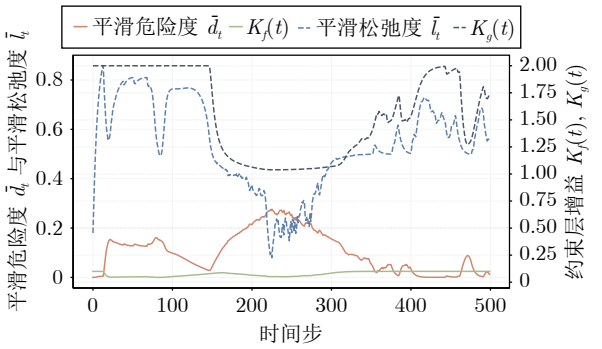


图 9　约束层状态–参数对应曲线
Fig. 9　Constrained layer state-parameter correspondence curves

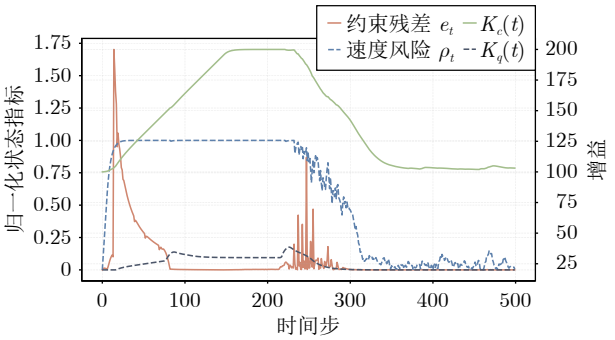


图 10　控制器层状态–参数对应曲线
Fig. 10　Controller layer state-parameter correspondence curves

高, 从而增强对速度边界附近可行域的收缩作用; 当约束残差 e_t 增大时, 误差修正增益 $K_c(t)$ 保持较高水平, 以增强对约束偏离的回拉能力. 随着风险水平逐步下降, $K_c(t)$ 与 $K_q(t)$ 也相应回落. 该结果说明, 控制器层调度能够依据即时状态变化进行在线反馈式响应, 从而体现出“风险指标较高增强保护、风险指标较低释放探索”的动态特征, 与图 9 所示控制器层的残差–风险反馈调节共同构成双层协同的动态保护机制.

5　结束语

针对固定增益设定难以灵活适应不同风险状态的问题, 本文提出一种状态相关增益调度自适应 ATACOM (SGA-ATACOM) 方法. 该方法在保持原有约束流形几何结构和解析控制框架不变的基础上, 引入约束层与控制器层的双层在线调度机制, 实现对约束回拉强度、误差修正强度以及速度保护强度的自适应调整.

仿真实验表明, 所提方法能够有效改善固定增益设定下的参数刚性问题, 并在典型受约束强化学

习任务中取得更好的综合表现. 进一步的消融分析说明, 约束层调度与控制器层调度分别作用于可行性约束动态响应和闭环误差修正过程, 二者共同构成方法性能提升的关键机制. 需要指出的是, 本文当前主要从结构设计与实验验证角度分析状态相关双层调度机制, 尚未给出时变增益调度下的严格稳定性证明, 相关理论分析将作为后续研究的重要方向.

综上, SGA-ATACOM 为受约束强化学习中的安全–性能协调提供了一种具有几何可解释性和结构保持性的改进方案. 未来可进一步面向更复杂的机器人任务开展理论分析与硬件验证[29], 以推动该方法在真实场景中的应用.

参考文献

1 Gu S D, Yang L, Du Y L, Chen G, Walter F, Wang J, et al. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 11216−11235

2 Lee J, Schroth L, Klemm V, Bjelonic M, Reske A, Hutter M. Exploring constrained reinforcement learning algorithms for quadrupedal locomotion. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 11132−11138

3 Keyumarsi S, Atman M W S, Gusrialdi A. LiDAR-based online control barrier function synthesis for safe navigation in unknown environments. *IEEE Robotics and Automation Letters*, 2024, **9**(2): 1043−1050

4 Zhou Yi, Zhang Hao, Shi Meng-Jie, Lin Bo-Xian, Shi Lei. A review of research on robot safety control based on control barrier functions in unknown environments. *Journal of University of Electronic Science and Technology of China*, 2025, **54**(1): 29−38
(周毅, 张浩, 施孟佶, 林伯先, 石磊. 未知环境中基于控制障碍函数的机器人安全控制研究综述. 电子科技大学学报, 2025, **54**(1): 29−38)

5 Könighofer B, Bloem R, Jansen N, Junges S, Pranger S. Shields for safe reinforcement learning. *Communications of the ACM*, 2025, **68**(11): 80−90

6 Dawood M, Shokry A, Bennewitz M. A dynamic safety shield for safe and efficient reinforcement learning of navigation tasks. In: Proceedings of the 7th Annual Learning for Dynamics & Control Conference. Ann Arbor, USA: PMLR, 2025. 686−697

7 Dong Ming-Ze, Wen Zhuang-Lei, Chen Xi-Ai, Yang Jiong-Kun, Zeng Tao. Research on robot navigation method integrating safe convex space and deep reinforcement learning. *Acta Armamentarii*, 2024, **45**(12): 4372−4382
(董明泽, 温庄磊, 陈锡爱, 杨炅坤, 曾涛. 安全凸空间与深度强化学习结合的机器人导航方法. 兵工学报, 2024, **45**(12): 4372−4382)

8 Liu P Z, Tateo D, Ammar H B, Peters J. Robot reinforcement learning on the constraint manifold. In: Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR, 2022. 1357−1366

9 Zhang Chang-Xin, Zhang Xing-Long, Xu Xin, Lu Yang. Safe reinforcement learning and its applications in robotics: A survey. *Control Theory & Applications*, 2023, **40**(12): 2090−2103
(张昌昕, 张兴龙, 徐昕, 陆阳. 安全强化学习及其在机器人系统中的应用综述. 控制理论与应用, 2023, **40**(12): 2090−2103)

10 Altman E. *Constrained Markov Decision Processes*. New York: Routledge, 2021. 11−15

11 Achiam J, Held D, Tamar A, Abbeel P. Constrained policy optimization. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org, 2017. 22−31

12 Liu Y S, Ding J X, Liu X. IPO: Interior-point policy optimization under constraints. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 4940−4947

13 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.

14 Stooke A, Achiam J, Abbeel P. Responsive safety in reinforcement learning by PID lagrangian methods. In: Proceedings of the 37th International Conference on Machine Learning. JMLR.org, 2020. Article No. 847

15 Ding D S, Wei X H, Yang Z R, Wang Z R, Jovanović M R. Provably efficient safe exploration via primal-dual policy optimization. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. San Diego, USA: PMLR, 2021. 3304−3312

16 Chow Y, Nachum O, Duenez-Guzman E, Ghavamzadeh M. A lyapunov-based approach to safe reinforcement learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montrél, Canada: Curran Associates Inc., 2018. 8103-8112

17 Chen Mou, Liu Wei, Zhang Peng. Distributed collision avoidance tracking control for quadrotor cooperative suspension system under performance constraints. *Acta Automatica Sinica*, 2024, **50**(12): 2392−2406
(陈谋, 刘伟, 张鹏. 性能约束下的四旋翼无人机协同吊挂系统分布式避碰跟踪控制. 自动化学报, 2024, **50**(12): 2392−2406)

18 Garcia J, Fernández F. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 2012, **45**(1): 515−564

19 Alshiekh M, Bloem R, Ehlers R, Könighofer B, Niekum S, Topcu U. Safe reinforcement learning via shielding. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press, 2018.

20 Hans A, Schneega\beta D, Schäfer A M, Udluft S. Safe exploration for reinforcement learning. In: Proceedings of the 16th European Symposium on Artificial Neural Networks. Bruges, Belgium, 2008. 143−148

21 Zhao Jing, Pei Zi-Nan, Jiang Bin, Lu Ning-Yun, Zhao Fei, Chen Shu-Feng. Virtual tube visual obstacle avoidance for UAV based on deep reinforcement learning. *Acta Automatica Sinica*, 2024, **50**(11): 2245−2258
(赵静, 裴子楠, 姜斌, 陆宁云, 赵斐, 陈树峰. 基于深度强化学习的无人机虚拟管道视觉避障. 自动化学报, 2024, **50**(11): 2245−2258)

22 Dalal G, Dvijotham K, Vecerik M, Hester T, Paduraru C, Tassa Y. Safe exploration in continuous action spaces. arXiv preprint arXiv: 1801.08757, 2018

23 Cheng R, Orosz G, Murray R M, Burdick J W. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press, 2019. 3387−3395

24 Koller T, Berkenkamp F, Turchetta M, Krause A. Learning-based model predictive control for safe exploration. In: Proceedings of the IEEE Conference on Decision and Control (CDC). Miami, USA: IEEE, 2018. 6059−6066

25 Hewing L, Wabersich K P, Menner M, Zeilinger M N. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2020, **3**: 269−296

26 Liu P Z, Bou-Ammar H, Peters J, Tateo D. Safe reinforcement learning on the constraint manifold: Theory and applications. *IEEE Transactions on Robotics*, 2025, **41**: 3442−3461

27 Liu P Z, Zhang K, Tateo D, Jauhri S, Hu Z Y, Peters J, et al.

Safe reinforcement learning of dynamic high-dimensional robotic tasks: Navigation, manipulation, interaction. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE, 2023. 9449−9456

28 Zhang Nan-Jie, Chen Yu-Quan, Ji Mao-Qin, Sun Yun-Kang, Wang Bing. Adaptive control method for quadruped robot facing floors of different roughness. *Acta Automatica Sinica*, 2025, **51**(7): 1585−1598
(张楠杰, 陈玉全, 季茂沁, 孙运康, 王冰. 面向不同粗糙程度地面的四足机器人自适应控制方法. 自动化学报, 2025, **51**(7): 1585−1598)

29 Miki T, Lee J, Hwangbo J, Wellhausen L, Koltun V, Hutter M. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science robotics*, 2022, **7**(62): eabk2822



刘正堂 北京林业大学工学院硕士研究生. 主要研究方向为安全强化学习和多机器人协同控制.
E-mail: liuzhengtang@bjfu.edu.cn
(**LIU Zheng-Tang** Master student at the School of Technology, Beijing Forestry University. Her research interests include safety-reinforcement learning and multi-robot cooperative control.)



吴 越 北京林业大学工学院副教授. 主要研究方向为多机器人协同控制与强化学习. 本文通信作者.
E-mail: wuyue_a@bjfu.edu.cn
(**WU Yue** Associate professor at the School of Technology, Beijing Forestry University. His research interests include multi-robot cooperative control and reinforcement learning. Corresponding author of this paper.)



胡春鹤 北京林业大学工学院副教授. 主要研究方向为机器人安全强化学习, 机器人自主控制.
E-mail: huchunhe@bjfu.edu.cn
(**HU Chun-He** Associate professor at the School of Technology, Beijing Forestry University. His research interests include safe reinforcement learning for robots and autonomous robot control.)



卢仁智 华中科技大学人工智能与自动化学院副教授. 主要研究方向为强化学习与无人系统.
E-mail: rzlu@hust.edu.cn
(**LU Ren-Zhi** Associate professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests include reinforcement learning and unmanned systems.)



程 骋 华中科技大学人工智能与自动化学院副教授. 主要研究方向为系统辨识与智能混线装配.
E-mail: c_cheng@hust.edu.cn
(**CHENG Cheng** Associate professor at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. Her research interests include system identification and intelligent mixed-line assembly.)



郑一力 北京林业大学工学院教授. 主要研究方向为智慧林业监测和机器人控制.
E-mail: zhengyili@bjfu.edu.cn
(**ZHENG Yi-Li** Professor at the School of Technology, Beijing Forestry University. His research interests include smart forestry monitoring and robot control.)