

基于增强型 PPO-PID 的机械臂跟踪控制

周凌云¹, 关哲^{1,3†}, 华长春¹, 曾祥瑞²

(1. 燕山大学 电气工程学院, 河北 秦皇岛 066000; 2. 石家庄铁道大学 机械工程学院, 石家庄 050043;
3. 河北省工业计算机控制工程重点实验室, 河北 秦皇岛 066000)

摘要: 针对多自由度机械臂的轨迹跟踪问题, 提出一种基于增强型近端策略优化的 PID (enhanced PPO-PID) 控制算法. 首先, 该方法构建基于共享特征提取层的统一网络架构, 在提升策略与价值函数协同优化能力的同时, 显著减少模型参数量, 提高收敛速度与学习效率; 其次, 提出一种基于性能回报动态调节训练迭代次数的机制, 从而实现训练初期快速收敛与后期策略稳定之间的平衡; 再次, 引入价值函数裁剪, 通过限制单次更新幅度, 有效平滑学习曲线, 增强高方差环境下的训练鲁棒性; 最后, 在 PandaReach-v3 机械臂仿真平台上, 通过对比实验验证了所提出方法在轨迹跟踪性能和鲁棒性方面的优越性.

关键词: 机械臂; 增强型近端策略优化; 数据驱动; PID 控制器

中图分类号: TP241.3 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1341

引用格式: 周凌云, 关哲, 华长春, 等. 基于增强型 PPO-PID 的机械臂跟踪控制 [J]. 控制与决策.

Enhanced PPO-PID for tracking control of robotic manipulators

ZHOU Ling-yun¹, GUAN Zhe^{1,3†}, HUA Chang-chun¹, ZENG Xiang-duan²

(1. School of Electrical Engineering, Yanshan University, Qinhuangdao 066000; 2. School of Mechanical Engineering, Shijiazhuang Railway University, Shijiazhuang 050043; 3. Hebei Province Key Laboratory of Industrial Computer Control Engineering, Qinhuangdao 066000)

Abstract: This paper proposes a PID control algorithm based on enhanced proximal policy optimization (enhanced PPO-PID) to address the trajectory tracking problem of multi-degree-of-freedom robotic arm system. First, the proposed method constructs a unified network architecture based on a shared feature extraction layer, which significantly reduces the number of model parameters while enhancing the synergistic optimization capability of the policy and value functions. The new architecture can improve the convergence speed and learning efficiency. Second, we propose a mechanism to dynamically adjust the number of training iterations based on reward, which results in achieving a balance between rapid convergence in the early stages of training and policy stability in the later stages. Third, we introduce a value function clipping to effectively smooth the learning curve by limiting the amplitude of a single update, thereby enhancing the robustness of training in high-variance environments. Finally, comparative experiments are conducted on the PandaReach-v3 robotic arm system to verify the superiority of the proposed method in trajectory tracking performance and robustness.

Keywords: robotic arm; enhanced proximal policy optimization; data-driven; PID controller

0 引言

在工业制造, 医疗手术以及物流运输等诸多领域, 多自由度机械臂都有着广泛的应用. 在复杂动态环境下, 机械臂通常表现出高度非线性、强耦合以及快速时变等特性. 如何提升控制系统的在线调节能

力与鲁棒性已成为该领域的重要研究方向^[1-2]. 在机械臂的跟踪任务中, PID 控制器因结构简单, 鲁棒性强而被广泛应用^[3-4]. 然而, PID 控制器的参数整定依赖于人工经验, 方法存在适应性不足的问题. 传统整定方法如 Ziegler-Nichols 法^[5]与 Chien-Hrones-

收稿日期: 2025-12-26; 录用日期: 2026-03-17.

基金项目: 国家自然科学基金项目 (62303398); 骨干人才项目 (留学回国平台)(C2025006); 省级重点实验室绩效补助经费项目 (22567612H); 河北省自然科学基金项目 (E2025210100); 河北省高等学校科学研究项目 (QN2025159).

责任编辑: 孙宗耀.

†通信作者. E-mail: guan@ysu.edu.cn.

Reswick 法^[6]在复杂系统中难以兼顾控制精度与系统稳定性. 当系统面对快速变化信号时可能会产生微分冲击 (Derivative Kick) 的现象, 并对系统的安全造成潜在的威胁^[7-9]. 研究者为解决上述方法的局限性提出在线调整 PID 参数的控制方法从而提升控制性能, 但仍存在对系统模型依赖程度较高的问题^[10], 并且难以有效应对机械臂的非线性和时变特性^[11]. 基于神经网络的控制方法, 通过监督学习优化 PID 参数能够有效的缓解建模难题, 但需要大量训练数据并且存在泛化能力不足的问题^[12-13].

不同于上述方法, 基于数据驱动的强化学习方法通过智能体与环境进行交互, 实现自主优化控制策略, 能够显著的降低对高精度模型的依赖为解决复杂控制任务提供了新的技术路径^[14]. 深度强化学习 (DRL) 通过融合深度神经网络的感知能力与强化学习的决策优势, 在 PID 参数的在线整定方面取得了显著进展^[15]. 典型方法包括: 近端策略优化 (PPO), 其通过引入剪裁机制提升了训练过程的稳定性^[16]; 深度 Q 网络 (DQN), 能够动态调整 PID 增益, 从而改善轨迹跟踪的响应速度与抗干扰能力^[17]; 双延迟深度确定性策略梯度 (TD3), 利用双评论网络与延迟更新机制, 有效抑制了过冲并加快收敛速度^[18]; 柔性演员-评论家算法 (SAC), 通过熵正则化实现了高精度的姿态控制^[19].

基于 DRL 的 PID 参数整定方法已取得一定成果, 但在效率与稳定性方面仍存在不足. 多数 Actor-Critic 框架采用策略网络与价值网络相互独立的结构设计, 导致了参数冗余和计算开销增加等问题. 并且, 当面临奖励信号方差较大的环境时, 若缺乏有效的价值网络更新约束机制, Critic 网络输出易产生剧烈波动, 将影响优势函数计算的准确性以及策略更新的稳定性. 此外, 训练策略同样是影响强化学习算法性能的关键因素. 现有方法通常采用固定训练周期 (Epoch). 但在训练初期迭代次数的不足可能会导致收敛速度缓慢. 而在训练后期, 过度迭代则易引发过拟合并造成性能的退化和计算资源浪费^[20]. 为解决上述问题, 本文提出一种基于增强型近端策略优化的 PID(Enhanced PPO-PID) 控制算法, 以提升控制策略的学习效率与稳定性, 从而更好地满足高精度轨迹跟踪任务的需求. 本文的主要贡献总结如下:

1) 针对策略网络与价值网络分离所引发的学习不一致与参数冗余问题, 设计了一种共享基础网络架构, 实现了策略与价值函数在统一特征空间下的协同优化, 显著减少了模型参数量, 提高了收敛速度与学习效率.

2) 为提升训练过程的稳定性, 本文采用价值函数剪裁机制. 通过限制单次价值函数更新的幅度, 有效平滑了价值估计的学习曲线, 为策略优化提供更加稳定的基线, 在奖励方差较大的任务中表现出显著优势.

3) 针对固定训练周期 (Epoch) 可能导致的训练不足或过拟合问题, 本文提出了一种动态调整训练迭代次数机制, 以在训练初期实现快速收敛, 并在训练后期保持策略稳定性.

1 本文方法

针对机械臂轨迹跟踪精度问题, 本文提出一种 Enhanced PPO-PID 方法. 该方法基于标准 PPO 框架下, 通过改进网络结构、更新机制与训练策略, 实现 PID 控制参数的在线自整定.

1.1 共享基础网络

在深度强化学习算法中, Actor-Critic 方法的效率与稳定性依赖于策略网络与价值网络的协同作用. 标准 PPO 算法的分离式网络结构在高维连续控制任务中主要存在两方面不足^[21-22]. 一方面是参数和计算的冗余. 独立的网络结构需要分别学习相似的底层状态特征, 将会增加训练开销和参数规模. 另一方面为学习目标冲突. 策略网络侧重于最优动作概率分布的学习, 而价值网络则聚焦于状态价值的精确估计. 独立网络结构进行优化可能导致策略梯度方向不稳定并造成训练振荡. 为缓解上述问题, 本文提出了一种 Actor-Critic 共享基础网络 (Shared Backbone Network) 架构, 如图 1 所示. 在该架构中, Actor 与 Critic 共享统一的底层特征提取网络, 仅在高层结构处分化为各自的输出分支. 该设计将策略学习与价值评估统一在同一特征空间中, 使价值函数对状态的估计能够更加直接且准确地指导策略梯度的计算. 同时, 参数共享不仅使模型结构更加紧凑, 还有效降低了过拟合风险, 并显著提升了训练效率. 具体网络配置如下: 共享特征提取层包含两层全连接层 (节点数分别为 256 和 128) 激活函数 (ReLU);

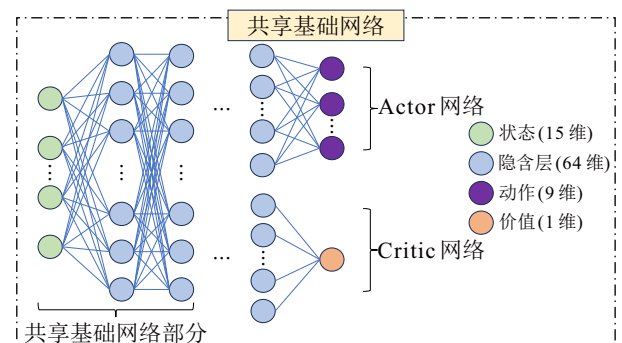


图1 共享网络架构

Actor 和 Critic 分支在共享层后各接两层全连接层(节点数均为 256), 最后分别输出 PID 参数和价值估计.

1.2 价值函数裁剪

共享网络架构提升了算法的效率, 在高效网络架构的基础上, 保障训练过程的动态稳定性是实现高性能控制的关键. PPO 算法通过裁剪机制限制策略更新的步长, 减少了破坏性的策略退化. 但在机械臂控制等奖励信号方差较大的任务中, 仅对策略施加约束难以保证整体训练的稳定性. 为此, 本文在保留 PPO 原有策略裁剪机制的基础上, 引入了价值函数裁剪 (Value Function Clipping) 方法, 构建双重裁剪的更新体系. 其中, 策略裁剪通过对概率比进行截断以限制策略更新幅度, 从而保持 PPO 算法的核心优势并保证学习过程的平稳性^[23-25]. 具体表达式如下:

$$L^{clip}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)], \quad (1)$$

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, \quad (2)$$

其中, $r_t(\theta)$ 重要性采样比率, 表示新策略与旧策略在状态 s_t 下选择动作 a_t 的概率比, $\pi_\theta(a_t|s_t)$ 表示在状态 s_t 下, 当前策略 π_θ 选择动作 a_t 的概率, $\pi_{\theta_{old}}$ 表示更新前的旧策略在相同状态下选择同一动作的概率. ϵ 为策略裁剪阈值.

价值函数裁剪在损失函数计算中引入约束^[26], 通过限制单次价值函数更新的幅度, 有效抑制了由高方差奖励信号引起的价值估计突变, 为策略优化提供了稳定的基准, 并在高方差环境下提升了算法的鲁棒性. 其更新规则可表示为:

$$V_{clip}(s_t) = V_{\theta_{old}}(s_t) + \text{clip}(V_\theta(s_t) - V_{\theta_{old}}(s_t), -\epsilon, +\epsilon), \quad (3)$$

$$L_V(\theta) = \mathbb{E}_t[\max((V_\theta(s_t) - G_t)^2, (V_{clip}(s_t) - G_t)^2)], \quad (4)$$

其中, $V_\theta(s_t)$ 表示当前价值网络的预测值, $V_{\theta_{old}}(s_t)$ 为上一轮迭代中的预测值, G_t 为目标回报, ϵ 为值裁剪阈值.

1.3 动态 Epoch 机制

为进一步提升训练效率, 本文摒弃了传统强化学习中固定训练周期 (Epoch) 的刚性策略. 固定训练周期难以适应训练过程的动态变化, 在训练初期, 算法需要充分利用有限的数据进行迭代, 而在训练后期, 策略已接近收敛过多迭代不仅造成计算资源浪

费, 还容易对当前批次数据产生过拟合, 从而削弱策略的泛化能力^[27-28].

针对上述问题, 本文提出了一种基于性能回报的动态 Epoch 调整机制. 该机制根据学习进程确定训练深度, 通过实时监测智能体在最近若干回合中的平均回报并与历史最优回报进行比较, 对迭代次数进行动态调整. 其基本规则可表示为:

$$k_{epoch} = \text{clip}((R - m_{best}) \cdot \alpha + k_{base}, k_{min}, k_{max}), \quad (5)$$

$$m_{best} = \max(m_{best}, R), \quad (6)$$

其中, R 表示近期平均回报, m_{best} 为历史最佳平均回报, α 为缩放因子, k_{base} 为基础 Epoch 数.

当近期性能超过历史最优水平时, 表明算法当前探索方向具有有效性, 系统将增加训练 Epoch 数, 通过高质量的样本加快策略的收敛速度. 当性能出现停滞或下降时则减少 Epoch 数, 避免在低质量数据上过拟合, 同时促进智能体在后续交互中进行更广泛的探索.

1.4 奖励函数

奖励函数是引导强化学习智能体学习行为策略的核心. 针对机械臂轨迹跟踪任务, 其根本目标在于最小化末端执行器与目标轨迹之间的偏差. 基于此, 本文设计了以负平方欧氏距离为核心的奖励函数, 其形式如下:

$$r = -\sum(e_t^2), \quad (7)$$

其中 e_t 表示当前时刻的轨迹跟踪误差. 采用误差平方形式不仅能够有效放大较大偏差所带来的惩罚, 引导智能体优先修正严重偏差. 同时, 该函数具备连续且稠密的反馈特性, 使智能体在每个时间步均可获得有效激励, 从而有助于提升学习效率并增强策略稳定性.

1.5 改进的速度型 PID 控制器设计

Enhanced PPO 算法旨在实现对 PID 控制器参数的在线整定, Actor 网络的输出即为比例、积分与微分增益. 其中网络的原始输出经过一个裁剪 (Clipping) 映射函数, 将其限制在预设的正值区间内, 参考基于动作裁剪 (Action Trimming) 的约束处理方法^[29], 从而保证 PID 控制参数的物理可行性. 针对微分冲击的问题, 本文采用了一种基于过程变量微分的速度型 PID 控制器结构^[30-31]. 其控制律的增量形式定义如下:

$$u(t) = u(t-1) + k_i(t) \cdot e(t) - k_p(t) \cdot \Delta y(t) - k_d(t) \cdot \Delta^2 y(t), \quad (8)$$

其中, $e(t)$ 表示时刻的跟踪误差, $\Delta y(t)$ 和 $\Delta^2 y(t)$ 分

别为过程变量 (即可测实际位置) 的一阶与二阶差分, $k_p(t)$, $k_i(t)$, $k_d(t)$ 分别代表比例、积分和微分增益. 速度型 PID 控制器能够有效的避免误差信号突变导致的控制冲击问题, 从而进一步提升轨迹跟踪的精度与平稳性.

算法1 Enhanced PPO-PID训练流程

初始化: Actor网络 π_θ , Critic网络 V_ϕ ;
 状态/奖励归一化统计量 $\mu_s, \sigma_s, \mu_r, \sigma_r$;
 动态Epoch参数 $m_{best}, k_{base}, s, k_{min}, k_{max}$.
 //在策略更新前基于近期回报计算动态Epoch数

for 更新轮次 $u = 1, \dots, M_{updates}$ **do**
 收集一批轨迹数据 B (大小为 T_c).
 在每步 t 中:
 采集状态 s_t , 更新 μ_s, σ_s 并归一化: $\hat{s}_t \leftarrow (s_t - \mu_s) / (\sigma_s + \epsilon)$.
 采样PID增益 $a_t \sim \pi_\theta(\cdot | \hat{s}_t)$.
 计算速度型PID控制器: $u(t) = u(t-1) + k_i(t) \cdot e(t) - k_p(t) \cdot \Delta y(t) - k_d(t) \cdot \Delta^2 y(t)$.
 执行控制, 获得 r_t, s_{t+1} , 并存储转换至 B .
 更新 μ_r, σ_r 并对 B 中的奖励进行归一化.
 计算GAE优势 \hat{A}_t 和回报 G_t .
 计算当前平均回报 R , 并更新Epoch基准: $m_{best} \leftarrow \max(m_{best}, R)$.
 动态调整训练轮数:
 $k_{Epoch} \leftarrow \text{clip}((R - m_{best}) \cdot \alpha + k_{base}, k_{min}, k_{max})$.

for $k = 1, \dots, k_{Epoch}$ **do**
 从 B 中随机采样小批量数据.
 更新Critic ϕ 以最小化剪裁价值损失 $\mathcal{L}_V(\phi)$.

更新Actor θ 以最大化PPO-Clip目标 $\mathcal{L}^{clip}(\theta)$.

return 训练好的Actor网络 π_θ .

1.6 算法总结

基于前述改进策略, 本文提出 Enhanced PPO-PID 方法, 旨在实现对速度型 PID 控制器参数的高效在线整定. 该算法的整体流程如图 2 所示, 其在机械臂轨迹跟踪任务中的具体训练步骤详见算法 1. 在训练过程中, 机械臂首先通过与环境交互采集状态、动作与奖励数据, 对采样结果进行在线归一化处理并计算优势函数^[32]; 随后结合动态 Epoch 调整机制, 在多个训练周期内对共享特征网络结构的参数进行迭代更新, 从而实现控制策略的持续优化.

2 数值模拟

2.1 实验环境

本文基于 Python 平台构建了 PandaReach-v3 强化学习环境, 依托 panda-gym 库 (版本 3.0.7) 与 PyBullet 实现 Franka Emika Panda 七自由度机械臂末端执行器的轨迹跟踪仿真^[33]. 实验运行平台为一台配备 Intel Core i9-14900K 处理器、64 GB 内存以及 NVIDIA GeForce RTX 4090 显卡 (24 GB 显存) 的工作站. 软件环境包括 Python 3.8.20、NVIDIA 显卡驱动版本 560.94 和 CUDA 12.6. 完成对比实验的总训练时长约为 4 小时. 该任务的目标是控制机械臂末端从随机初始位置平稳跟踪至随机生成的目标轨迹. 其状态观测向量是一个 15 维的复合向量, 具体由末端执行器的实时位置 (3 维)、末端执行器的线速度 (3 维)、当前已达成的目标位置 (3 维)、期望的目标位置 (3 维) 以及位置误差的累积积分项

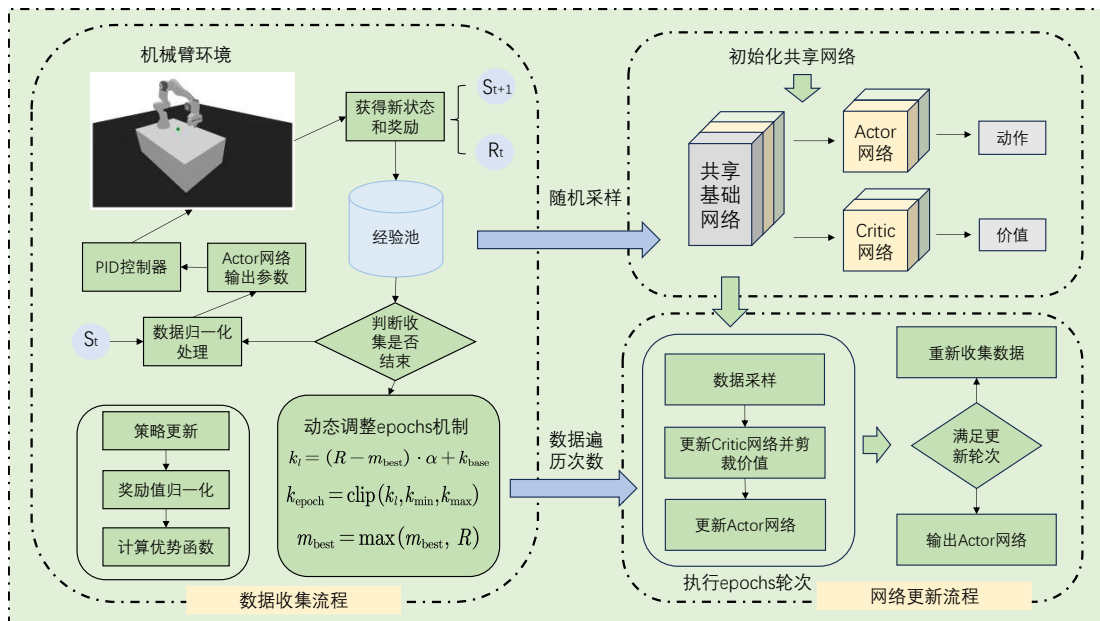


图2 Enhanced PPO-PID 算法流程图

(3 维) 组成. 通常情况下, 实时位置和达成的目标位置相同. 为验证所提出的 Enhanced PPO-PID 控制器的有效性, 本文设计了正弦曲线与 3D 不规则曲线的轨迹跟踪任务, 并与传统 Proximal Policy Optimization PID (PPO-PID) 控制器进行了对比实验. 图 3 展示了 PandaReach-v3 测试环境的结构示意图, 表 1 给出了主要实验参数配置.

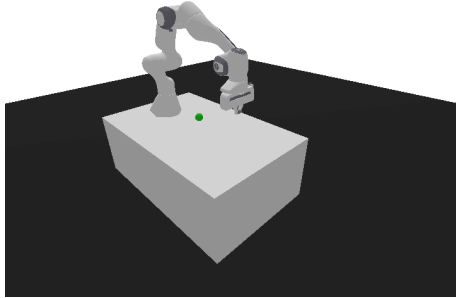


图3 PandaReach-v3

表1 实验参数

参数	含义	取值 / 范围
学习率 (<i>Actor/Critic</i>)	PPO学习率	3e-4
γ	折扣因子	0.99
λ (GAE)	GAE参数	0.95
策略切值 ϵ	策略更新剪切范围	0.2
价值切值 ϵ	价值更新剪切范围	0.2
基础 k_{base}	动态 <i>Epoch</i> 基准	20
动态 <i>Epoch</i> 范围	[min_Epoch, max_Epoch]	[20, 100]
<i>batchsize</i>	每次更新批量大小	64
T_c	每轮收集步数	2048
α	调节模型响应程度	1.2
<i>Optimizer</i>	优化器类型	Adam
<i>Mini_batches</i>	单次更新数据切分份数	8
归一化 ϵ	归一化防除零项	1e-6

2.2 实验结果分析

在多自由度机械臂仿真环境中, 本文针对正弦信号轨迹跟踪与 3D 不规则曲线跟踪分别开展了两组对比实验. 每组实验均独立重复五次, 以减小偶然性因素对结果的影响.

图 4 对比了 PPO-PID(蓝色) 与 Enhanced PPO-PID(红色) 在正弦曲线跟踪任务中五次独立训练下的平均回报 (实线) 及其标准差 (阴影区域). 结果表

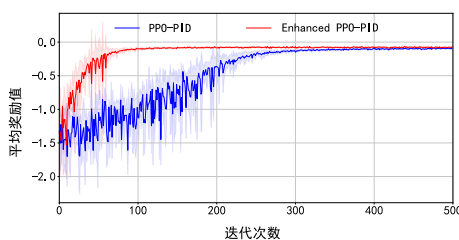


图4 奖励值对比

明, Enhanced PPO-PID 在各个训练阶段均取得了更高的平均回报, 且标准差范围明显小于 PPO-PID, 说明其具有更好的训练稳定性.

图 5 和图 6 展示了在训练早期 (episode 250) 和训练结束时 (episode 500) 机械臂跟踪正弦曲线的效果. 在早期 Enhanced PPO-PID 的跟踪效果明显优于 PPO-PID, 在最终的轮次的对比, Enhanced PPO-PID 的跟踪精度也有明显的优势, 证明了 Enhanced PPO-PID 在收敛速度和整体控制性能上具有明显优势.

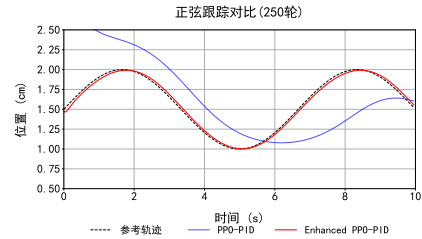


图5 第 250 轮正弦信号跟踪对比

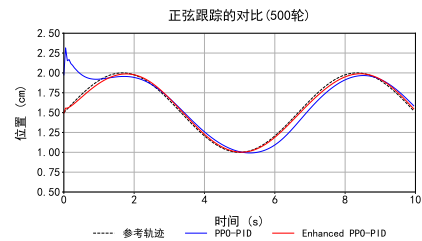


图6 第 500 轮正弦信号跟踪对比

下述图 7 与图 8 分别给出了在 3D 不规则曲线跟踪任务下的对比表现. 如图 7 所示, 在训练早期 (episode 200), Enhanced PPO-PID(红线) 能够更紧密地跟踪由虚线表示的目标轨迹, 且跟踪误差显著小于 PPO-PID(蓝线). 图 8(episode 500) 显示 Enhanced PPO-PID 在收敛后依然保持较高的跟踪精度与稳定性, 体现出较强的鲁棒性.

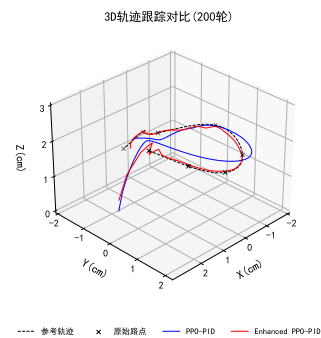


图7 第 200 轮 3D 不规则曲线跟踪对比

2.3 干扰实验

为了验证方法在复杂扰动条件下的鲁棒性, 本文设计了两组含扰动的轨迹跟踪实验^[34]. 分别在二

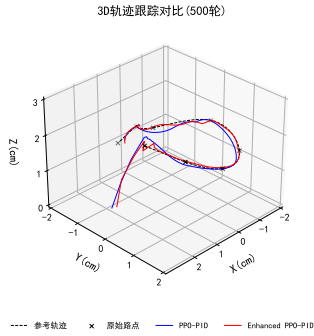


图8 第500轮3D不规则曲线跟踪对比

维正弦轨迹跟踪和3D不规则轨迹跟踪任务中对机械臂随机施加外部扰动。

实验结果如图9,图10所示.相较于标准PPO-PID, Enhanced PPO-PID算法在受到扰动时仍能保持平滑、稳定的轨迹跟踪,并在受到干扰后能够快速的响应,以较短的时间进行轨迹的调节。

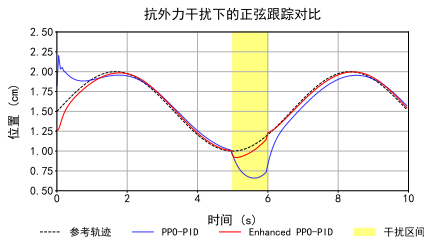


图9 正弦跟踪对比

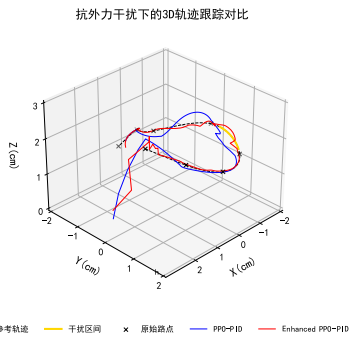


图10 3D跟踪对比

2.4 消融实验

为验证 Enhanced PPO-PID 方法中各项关键改进措施的有效性, 本文分别对共享基础网络、动态 Epoch 机制与值函数裁剪进行了独立消融实验. 实验结果分别如图 11, 图 12, 图 13 所示, 图 14 为不同消融配置下的平均奖励值变化的对比。

图 11 和图 12 分别展示了算法移除共享基础网络和动态 Epoch 机制后, 轨迹跟踪精度显著下降, 说明了共享网络结构可减少冗余计算、提升收敛效率

共享网络对跟踪效果的影响

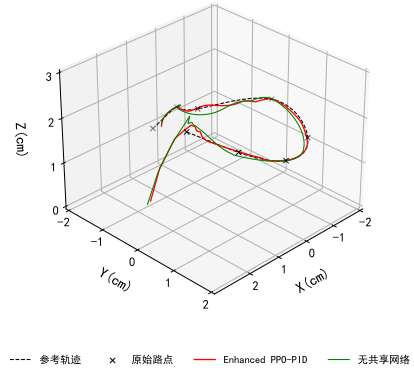


图11 共享基础网络消融实验

动态Epoch对跟踪效果的影响

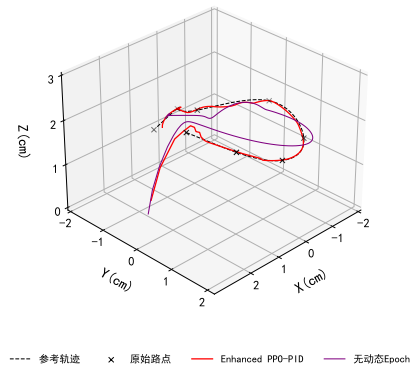


图12 动态 Epoch 机制消融实验

值函数裁剪对跟踪效果的影响

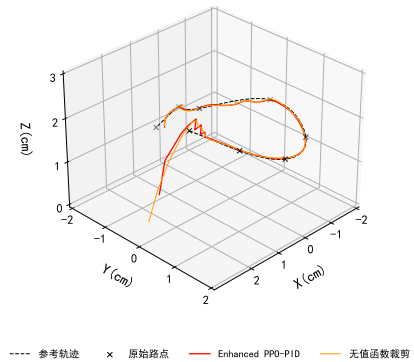


图13 值函数裁剪消融实验

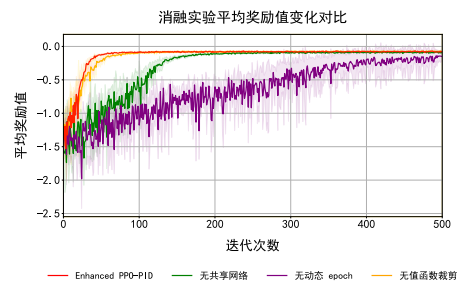


图14 消融实验奖励值对比

与特征一致性以及动态 Epoch 机制调整迭代次数, 提升早期学习效率并维持后期稳定, 合理分配计算

资源。

图 13 与图 14 完整展示了不同消融配置下的跟踪效果与平均奖励变化对比。结果显示, 完整的 Enhanced PPO-PID 算法(红线)在训练初期收敛速度最快, 并最终获得了最高的平均奖励; 相比之下, “无共享网络”变体因特征提取参数冗余导致收敛明显滞后。通过对比奖励曲线可以发现, 带有值函数裁剪机制的算法在前期收敛更迅速且奖励值更加稳定, 而“无值函数裁剪”变体则表现出较大的曲线震荡与末轮跟踪波动, 有力验证了双裁剪机制在抑制训练波动、提升系统鲁棒性方面的有效性。

图 15 与表 2 共同展示了 Enhanced PPO-PID 与 SAC-PID 算法在三维轨迹跟踪任务中的对比分析结果。轨迹跟踪曲线(图 15)显示, 相比于 SAC-PID 算法(绿色曲线)在初始阶段的较大超调及路径转折处的明显震荡, 本文算法(红色曲线)能够迅速收敛并始终紧密贴合参考轨迹(黑色虚线), 全程保持平滑。定量实验数据(表 2)进一步验证了该定性观察: Enhanced PPO-PID 在各项指标上均全面优于 SAC-PID。具体而言, 该算法将均方根误差从 0.3523 cm 降低至 0.1975 cm(提升 44.0%), 显著提高了整体控制精度; 同时, 最大误差与稳态误差分别降低了 32.0% 与 55.9%。上述结果有力证明了 Enhanced PPO-PID 在有效抑制启动瞬间冲击的同时, 在长时间运行任务中具备极高的稳定性与鲁棒性。

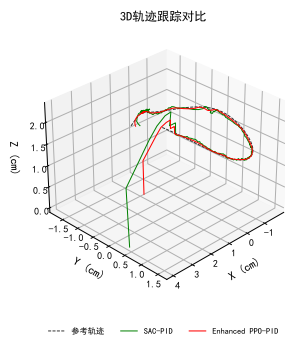


图 15 SAC 对比

表 2 性能指标对比

性能指标	均方根误差 / cm	最大误差 / cm	稳态误差 / cm
SAC-PID	0.3523	2.6250	0.3126
Enhanced PPO-PID (ours)	0.1975	1.7860	0.1379
性能提升	+44.0%	+32.0%	+55.9%

3 结论

在机械臂的轨迹跟踪控制问题中, 保障系统在非线性和动态环境下的控制精度和鲁棒性是关键。

本文提出了一种基于增强型近端策略优化的 PID (Enhanced PPO-PID) 控制方法。通过引入速度型 PID 结构缓解了微分突变问题对系统造成的影响。通过构造共享特征网络与双裁剪机制减少冗余参数并提升收敛效率。利用动态 Epoch 调整机制实现了计算资源的合理分配。实验结果表明, 方法在复杂扰动条件下表现出更优的跟踪精度、稳定性与鲁棒性。未来的研究将进一步探索将元学习策略融入该框架, 以提升模型在不同动力学条件下的快速适应与泛化能力, 从而增强算法的实用价值。

参考文献 (References)

- [1] 陈钢, 叶佩昌, 贾庆轩, 等. 基于速度修正项的机械臂避障路径规划[J]. *控制与决策*, 2015, 30(1): 156-160. (Chen G, Ye P C, Jia Q X, et al. Obstacle avoidance path planning of manipulator based on speed correction term[J]. *Control and Decision*, 2015, 30(1): 156-160.)
- [2] 张安龙, 林志赞, 王博, 等. 基于笛卡尔空间力补偿的柔性关节协作机械臂轨迹跟踪控制[J]. *控制与决策*, 2025, 40(6): 1807-1816. (Zhang A L, Lin Z Y, Wang B, et al. Trajectory tracking control for collaborative robotic arms with SEA based on force compensation in Cartesian space[J]. *Control and Decision*, 2025, 40(6): 1807-1816.)
- [3] Borase R P, Maghade D K, Sondkar S Y, et al. A review of PID control, tuning methods and applications[J]. *International Journal of Dynamics and Control*, 2021, 9(2): 818-827.
- [4] Hantoro R, Mozef E, Pardede H F. Two heuristic PID tuning for a 4-DOF robot arm control[J]. *International Journal of Recent Technology and Engineering*, 2019, 8(4): 5205-5210.
- [5] Ziegler J G, Nichols N B. Optimum settings for automatic controllers[J]. *Transactions of the ASME*, 1942, 64: 759-768.
- [6] Chien K L, Hrones J A, Reswick J B. On the automatic control of generalized passive systems[J]. *Transactions of the ASME*, 1952, 74: 175-185.
- [7] Lendek A, Tan L Z. Mitigation of derivative kick using time-varying fractional-order PID control[J]. *IEEE Access*, 2021, 9: 55974-55987.
- [8] Åström K J, Hägglund T. *Advanced PID control*[M]. Research Triangle Park: ISA — The Instrumentation, Systems and Automation Society, 2006.
- [9] Joyo M K, Raza Y, Ahmed S F, et al. Optimized proportional-integral-derivative controller for upper limb rehabilitation robot[J]. *Electronics*, 2019, 8(8): 826.
- [10] 杨亮, 陈勇, 刘治. 基于参数不确定机械臂系统的自适应轨迹跟踪控制[J]. *控制与决策*, 2019, 34(11): 2485-2490. (Yang L, Chen Y, Liu Z. Adaptive trajectory tracking control for manipulator with uncertain dynamics and kinematics[J]. *Control and Decision*, 2019, 34(11):

- 2485-2490.)
- [11] Zhang H G, Liu D R, Luo Y H, et al. Adaptive dynamic programming for control: algorithms and stability[M]. Berlin: Springer Science & Business Media, 2012.
- [12] Goodfellow I, Bengio Y, Courville A, et al. Deep learning[M]. Cambridge: MIT Press, 2016.
- [13] 赵强, 刘敏. 在线神经网络自适应控制在非线性系统中的应用[J]. 计算机仿真, 2020, 37(12): 101-108. (Zhao Q, Liu M. Application of online neural network adaptive control in nonlinear systems[J]. Computer Simulation, 2020, 37(12): 101-108.)
- [14] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529-533.
- [15] Guan Z, Yamamoto T. Design of a reinforcement learning PID controller[C]. International Joint Conference on Neural Networks. Glasgow, 2020: 1-6.
- [16] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J/OL]. 2017, arXiv: 1707.06347.
- [17] Gu S X, Holly E, Lillicrap T, et al. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]. IEEE International Conference on Robotics and Automation. Singapore, 2017: 3389-3396.
- [18] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]. Proceedings of the 35th International Conference on Machine Learning. Stockholm, 2018: 1587-1596.
- [19] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]. Proceedings of the 35th International Conference on Machine Learning. Stockholm, 2018: 1856-1865.
- [20] Zhou Z Y, Mo F, Zhao K, et al. Adaptive PID control algorithm based on PPO[J]. Journal of System Simulation, 2024, 36(6): 1425-1432.
- [21] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]. Proceedings of the 33rd International Conference on Machine Learning. New York 2016: 1928-1937.
- [22] Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive neural networks[J/OL]. 2016, arXiv: 1606.04671.
- [23] Henderson P, Islam R, Bachman P, et al. Deep reinforcement learning that matters[C]. Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, 2018: 3207-3214.
- [24] Schulman J, Levine S, Moritz P, et al. Trust region policy optimization[C]. Proceedings of the 32nd International Conference on Machine Learning. Lille, 2015: 1889-1897.
- [25] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation[C]. Proceedings of the International Conference on Learning Representations. San Juan, 2016: 1-14.
- [26] Engstrom L, Ilyas A, Santurkar S, et al. Implementation matters in deep policy gradients: A case study on PPO and TRPO[C]. Proceedings of the International Conference on Learning Representations. Addis Ababa, 2020: 1-12.
- [27] Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts[C]. Proceedings of the International Conference on Learning Representations. Toulon, 2017: 1-16.
- [28] Wu Y, Tian Y, Wang L. Adaptive epoch adjustment in deep reinforcement learning for efficient training[J]. Neural Computing and Applications, 2021, 33(14): 7789-7802.
- [29] Fu Z M, Wang H C, Tao F Z, et al. Energy management strategy for fuel cell/battery/ultracapacitor hybrid electric vehicles using deep reinforcement learning with action trimming[J]. *IEEE Transactions on Vehicular Technology*, 2022, 71(7): 7171-7185.
- [30] Visioli A. Practical PID control: A modern perspective[M]. Cham: Springer, 2020: 1-25.
- [31] Zheng J, Li Y. Mitigation of derivative kick in PID control for nonlinear systems[J]. IEEE Transactions on Industrial Electronics, 2018, 65(3): 2123-2132.
- [32] Ioffe S, Szegedy C. Batch normalization: Accelerating deep Nnetwork training by reducing internal covariate shift[C]. Proceedings of the 32nd International Conference on Machine Learning. Lille, 2015: 448-456.
- [33] Gallouédéc Q, Cazin N, Dellandréa E, et al. Panda-gym: Open-sourcegoal-conditioned environments for robotic learning[J/OL]. 2021, arXiv: 2106.13687.
- [34] 艾海平, 陈力. 空间机器人捕获航天器操作的避撞柔顺复合自抗扰控制[J]. *控制与决策*, 2021, 36(2): 355-362. (Ai H P, Chen L. Collision avoidance and compliant composite active disturbance rejection control of space robot capture spacecraft[J]. *Control and Decision*, 2021, 36(2): 355-362.)

作者简介

周凌云 (2001-), 男, 硕士生, 主要研究方向为深度强化学习、数据驱动, E-mail: 13332228416@163.com;

关哲 (1988-), 男, 讲师, 博士, 主要研究方向为数据驱动理论及其应用、工程机械控制、过程控制, E-mail: guan@ysu.edu.cn;

华长春 (1979-), 男, 教授, 博士, 博士生导师, 主要研究方向为非线性动力系统的控制及应用、网络化控制系统的分析与综合、基于数据驱动的故障诊断和容错控制, E-mail: cch@ysu.edu.cn;

曾祥瑞 (1993-), 男, 副教授, 博士, 博士生导师, 主要研究方向为机器人控制、智能车辆控制、复合学习, E-mail: xzdysu@163.com.