Journal of System Simulation

Volume 36 | Issue 6 Article 15

6-28-2024

Adaptive PID Control Algorithm Based on PPO

Zhiyong Zhou
School of Mechanical Engineering, Shanghai Dianji University, Shanghai 201306, China,
zhouzhiyong789@tom.com

Fei Mo
School of Mechanical Engineering, Shanghai Dianji University, Shanghai 201306, China

Kai Zhao
Shanghai Aerospace Equipment Manufacturing General Factory, Shanghai 200245, China

Yunbo Hao
Shanghai Aerospace Equipment Manufacturing General Factory, Shanghai 200245, China

See next page for additional authors

Adaptive PID Control Algorithm Based on PPO

Abstract

Abstract: A six-axis robotic arm is built and simulated in a complex control environment with disturbances by using MATLAB physics engine and Python, which provides a trial-and-error environment for the robotic arm training that could not be provided in reality. Proximal policy optimization(PPO) algorithm in reinforcement learning is proposed to improve the traditional PID control algorithm. By introducing the multi-agent idea and on the basis of the different effects of the three parameters of PID on control system and the characteristics of the six-axis robotic arm, the three parameters are separately trained as different intelligent individuals to achieve a new multi-agent adaptive PID algorithm with multi-agent adaptive adjustment of parameters. Simulation results show that the algorithm outperforms MA-DDPG and MA-SAC algorithms in training convergence. Compared with the traditional PID algorithm, the algorithm can effectively suppress the disturbances and oscillations, and has lower overshoot and adjustment time, which makes the control process smoother and effectively improves the control accuracy of the robotic arm. The robustness and effectiveness is proved.

Keywords

RL, PPO algorithm, adaptive PID tuning, robotic arm, multi-agent

Authors

Zhiyong Zhou, Fei Mo, Kai Zhao, Yunbo Hao, and Yufeng Qian

Recommended Citation

Zhou Zhiyong, Mo Fei, Zhao Kai, et al. Adaptive PID Control Algorithm Based on PPO[J]. Journal of System Simulation, 2024, 36(6): 1425-1432.

第 36 卷第 6 期 系统仿真学报© Vol. 36 No. 6
2024 年 6 月 Journal of System Simulation Jun. 2024

基于PPO的自适应PID控制算法研究

周志勇[1]，莫非[1]，赵凯[2]，郝云波[2]，钱宇峰[1]
(1. 上海电机学院 机械学院，上海 201306；2. 上海航天设备制造总厂有限公司，上海 200245)

摘要：采用MATLAB物理引擎联合Python搭建了一个六轴机械臂，并模拟带有扰动的复杂控制环境，为机械臂训练提供现实中无法提供的试错环境。*使用强化学习中近端优化算法(proximal policy optimization，PPO)算法对传统PID控制算法进行改进，引入多智能体思想，根据PID三个参数对控制系统的不同影响及六轴机械臂的特性，将三个参数分别作为不同的智能个体进行训练，实现多智能体自适应调整参数的新型多智能体自适应PID算法。*仿真结果表明：该算法的训练收敛性优于MA-DDPG与MA-SAC算法，与传统PID算法的控制效果相比，在遇到扰动及振荡的情况下，能够更有效地抑制振荡，并具有更低的超调量和调整时间，控制过程更为平缓，有效提高了机械臂的控制精度，证明了该算法的鲁棒性及有效性。

关键词：强化学习；近端优化算法；自适应PID整定；机械臂；多智能体
中图分类号：TP242.2 文献标志码：A 文章编号：1004-731X(2024)06-1425-08
DOI: 10.16182/j.issn1004731x.joss.23-0137
引用格式: 周志勇, 莫非, 赵凯, 等. 基于PPO的自适应PID控制算法研究[J]. 系统仿真学报, 2024, 36(6): 1425-1432.
Reference format: Zhou Zhiyong, Mo Fei, Zhao Kai, et al. Adaptive PID Control Algorithm Based on PPO[J]. Journal of System Simulation, 2024, 36(6): 1425-1432.

Adaptive PID Control Algorithm Based on PPO

Zhou Zhiyong[1], *Mo Fei*[1], *Zhao Kai*[2], *Hao Yunbo*[2], *Qian Yufeng*[1]
(1. School of Mechanical Engineering, Shanghai Dianji University, Shanghai 201306, China;
2. Shanghai Aerospace Equipment Manufacturing General Factory, Shanghai 200245, China)

Abstract: A six-axis robotic arm is built and simulated in a complex control environment with disturbances by using MATLAB physics engine and Python, which provides a trial-and-error environment for the robotic arm training that could not be provided in reality. *Proximal policy optimization(PPO) algorithm in reinforcement learning is proposed to improve the traditional PID control algorithm. By introducing the multi-agent idea and on the basis of the different effects of the three parameters of PID on control system and the characteristics of the six-axis robotic arm, the three parameters are separately trained as different intelligent individuals to achieve a new multi-agent adaptive PID algorithm with multi-agent adaptive adjustment of parameters*. Simulation results show that the algorithm outperforms MA-DDPG and MA-SAC algorithms in training convergence. Compared with the traditional PID algorithm, the algorithm can effectively suppress the disturbances and oscillations, and has lower overshoot and adjustment time, which makes the control process smoother and effectively improves the control accuracy of the robotic arm. The robustness and effectiveness is proved.

Keywords: RL; PPO algorithm; adaptive PID tuning; robotic arm; multi-agent

收稿日期：2023-02-14 修回日期：2023-04-21
基金项目：上海市闵行区重大产业技术攻关计划(2022MH-ZD20)
第一作者：周志勇(1984-)，男，副教授，博士，研究方向为创新设计理论与方法。E-mail：zhouzhiyong789@tom.com

0 引言

在“工业4.0”的大环境下，人工智能算法、云计算、机器人智能化控制等迅速发展，机械臂作为工业系统中的重要角色，其控制算法的研究对工业产品质量的提升有着重大意义。大多数机械臂是一个多输入多输出的强非线性控制系统，在完成任务时，机械臂的六个轴互相配合，使机械臂末端的位姿达到目标要求，尤其在完成高精度要求的任务时，控制期间发生振荡或超调现象都会导致生产作业上的误差，一个快速、稳定、高精度的控制策略对机械臂的控制来说尤为重要。

文献[1]提出了一种新的非奇异终端滑模算法，能在复杂环境下实现精确的跟踪控制。文献[2]提出的多变量光滑二阶滑模轨迹跟踪控制方法，也实现了所期望的控制效果。以上两种算法虽然都能达到高精度控制，但都存在建模复杂及计算繁琐的问题。

随着人工智能领域的发展，出现了很多强化学习算法，如DQN[3]、TRPO(trust region policy optimization)[4]、PPO(proximal policy optimization)[5]、SAC(soft actor-critic)[6]等。强化学习[7]因其在复杂环境下良好的鲁棒性[8]得到广泛应用。

为了减少建模的时间以及开发成本，强化学习控制算法在机械臂领域得到广泛应用，常被用于难以准确建模的场景以及一些强非线性系统，可有效减少前期的模型设计时间，加快研究进程。文献[9]利用强化学习中的DDPG(deep deterministic policy gradient)算法作为机械臂控制算法核心，实现了机械臂的控制。文献[10]使用强化学习并采用BP算法对机械臂进行建模，实现了对机械臂的自适应跟踪控制。

机械臂的控制方法种类很多，如神经网络控制[11-13]、模糊控制[14-16]、滑模控制[17-19]、鲁棒控制[20-21]等。虽然控制方法种类繁多，但是传统PID控制及其改进型因为简单、灵活、易调节等特点，依旧是大部分机械臂控制方法的首选。传统PID的参数调整一般是通过试凑法及专业人员的经验，不断地调整参数以达到良好的控制效果，这样的调试方法充满随机性和盲目性，不仅工作量大、繁琐，且在面对复杂多目标非线性系统时，难以找到一套合适参数适配每一个控制阶段。

本文利用强化学习中的PPO算法，解决传统PID调节繁琐问题的同时对PID进行改进，并引入多智能体思想，在实现抓握和控制的基础上，对控制精度做进一步提升，克服了建模复杂的缺点，并加入自适应功能，实现了高精度控制，整体系统如图1所示。

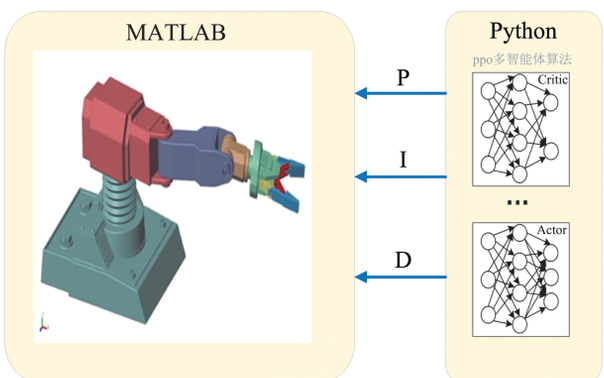

图1 机械臂三维模型
Fig. 1 3D model of robotic arm

1 基于PPO自适应PID算法

1.1 强化学习理论框架

强化学习的框架一般包含5个构成要素：环境、智能体、状态、动作、奖励，框架如图2所示，智能体对系统环境进行状态观察后产生行动，从系统环境中获得相应的奖励，智能体观察到系统对自己上一次行动的奖励信号后，重新调整自己下一次的行动策略。

1.2 PPO算法

PPO算法是一种新型的策略梯度算法，是对策略梯度算法(policy gradient, PG)的一种改进，通过策略critic与动作actor相互配合学习使智能体agent得到的奖励最大化，即actor-critic算法，如图3所示。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss6/15
DOI: 10.16182/j.issn1004731x.joss.23-0137

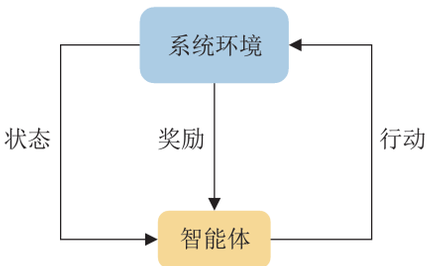

图2　强化学习框架
Fig. 2　RL-structure

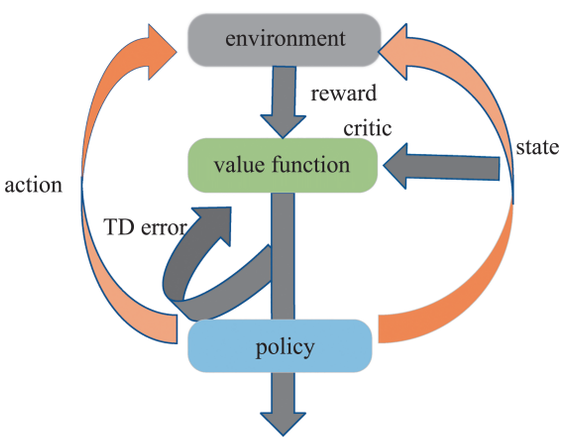

图3　Actor-critic结构
Fig. 3　Actor-critic structure

在PG算法中，不同训练步长对训练效果影响较大。在训练时，较小的步长会导致训练过程出现局部难以收敛的情况，反之，就会导致梯度增长过快而丢失训练细节，因此，PPO修改了最初的PG公式，不再使用PG算法而是加入了新的目标函数，并使用了import sampling重要采样方法，解决了PG算法对于训练步长难以确定的缺点，提升了训练效果和速度。本文使用的算法是PPO算法的变体之一PPO-Clip，该算法更新策略通过式(1)更新神经网络参数θ_k，通常采取多步SGD更新(mini-batch)来最大化目标的期望E。

$$\theta_{k+1}=\arg\max_{\theta}\underset{s,a\sim\pi_{\theta_k}}{E}[L(s,a,\theta_k,\theta)] \tag{1}$$

$$[L(s,a,\theta_k,\theta)]=\min\left(\frac{\pi_\theta(a|s)}{\pi_\pi(a|s)}A^{\pi\theta_k}(s,a),g(\epsilon,A^{\pi\theta_k}(s,a))\right) \tag{2}$$

$$A^{\pi\theta_k}(s,a)=Q(s_t,a_t)-V(s_t) \tag{3}$$

$$g(\epsilon,A)=\begin{cases}(1+\epsilon)A, & A\geqslant 0\\(1-\epsilon)A, & A<0\end{cases} \tag{4}$$

式中：$\frac{\pi_\theta(a|s)}{\pi_\pi(a|s)}$为新旧策略根据状态$s$采取动作$a$的概率的比值；$Q(s_t,a_t)$为神经网络预测值；$V(s_t)$为价值函数$V$的输出值；$\epsilon$为可调节的超参数。利用式(1)可有效减少训练时的方差，提高学习效率，PPO更新时进行Clip操作，进而对步长进行调控，防止梯度爆炸的发生。

1.3 控制器设计

PID算法由于其易用性和可靠性，被广泛应用于各种工业控制场景，是控制中十分经典的算法，PID算法由P(比例环节)、I(积分环节)、D(微分调节)组成。在调节误差e时，P环节越大调节速度越快，但会给系统带来振荡；I环节可对输入量进行调整，用于消除静态误差，值越大调节速度越快，同样也会带来振荡；D环节可对误差的趋势进行预测，提前做出预判性调整，可减小振荡，但会引入高频噪声。因为PID算法的上述特点，在一些非线性系统下，会表现出时滞性和振荡性且依赖于工作人员的专业经验，因此，本文在PID的基础上加入PPO算法进行PID参数整定以提升控制性能并达到自适应效果，克服传统PID算法单一参数的缺点。图4为本文的PID控制器的整体结构。PID根据时间t的计算式为

$$u=K_{\mathrm{P}}e(t)+K_{\mathrm{I}}\int e(t)\mathrm{d}t+K_{\mathrm{D}}\frac{\mathrm{d}e(t)}{\mathrm{d}t} \tag{5}$$

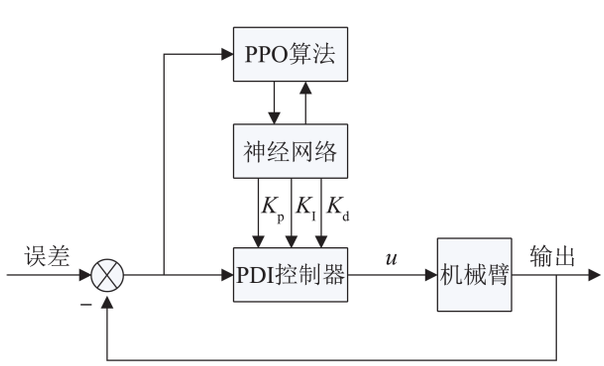

图4　控制器设计
Fig. 4　Controller design

1.4 PPO 多智能体算法结构设计

本文提出一种基于 PPO 的多智能体新型自适应算法，该算法采用去中心化训练+去中心化决策，每个智能体配置一个独立策略函数 π 和价值函数 V，并同时共享同一个环境，进行训练的多智能体结构设计如图 5 所示。

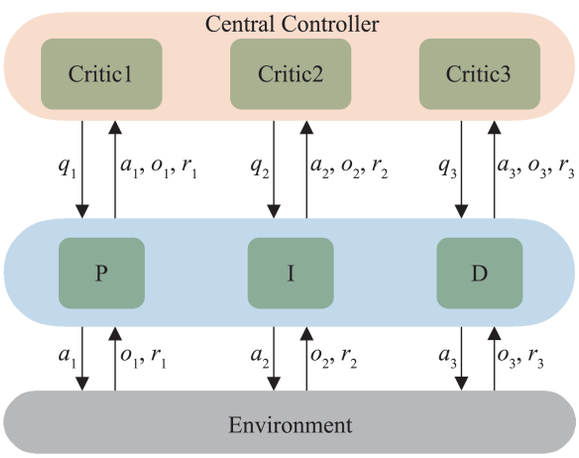

图 5　多智能体结构
Fig. 5　Multi-agent structure

PID 的 3 个参数中，每个参数控制的变量各不相同，P 参数对角度的误差做积，I 则针对误差量的累计，D 参数针对角度的变化率，3 个参数对于控制的影响也各不相同，如图 6 所示。

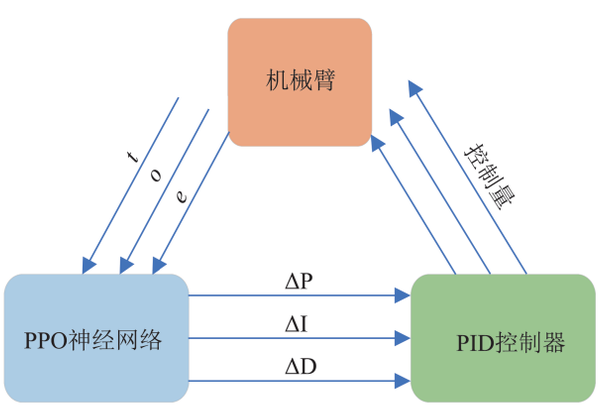

图 6　输入/输出解析
Fig. 6　Input/output analysis

针对这一特性，为 3 个参数分别设计 3 个价值网络进行神经网络的训练，3 个价值网络的训练使策略函数 π 不断优化，如图 7 所示，同时 3 个价值网络的奖励值 r_1、r_2、r_3 也不断增加，最终策略函数 π 会根据当前状态 o 输出最佳的动作 a_1、a_2、a_3，即 PID 的 3 个参数。

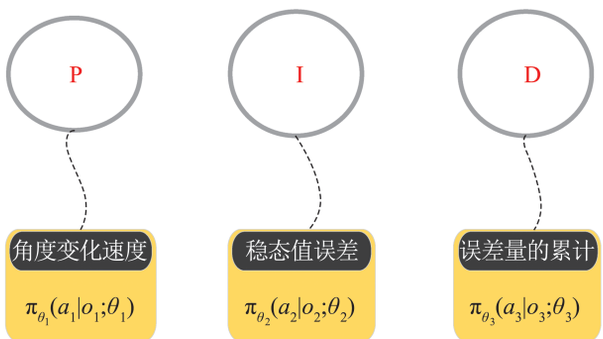

图 7　参数解析
Fig. 7　Parameter analysis

对 3 个价值网络进行不同的奖励设置和赋值，然后进行合并分析，3 个参数同时根据不同的方案进行调节，最终达到高精度控制的目的。神经网络的输入状态 s 包括控制量的时间 t，机械臂目标值与当前值误差 e，根据状态的输入，最终会训练出在具体时间下最合适的 PID 参数。

根据以上信息，使用一种基于 PPO 的新型合作学习奖励方案，定义为

$$r_1 = r_{\mathrm{P}} + t_r\alpha_1 - t_{\mathrm{d}} - \alpha_2 y_{\mathrm{m}} \tag{6}$$

$$r_2 = r_{\mathrm{I}} - \Delta e \tag{7}$$

$$r_3 = t_{\mathrm{D}} - \alpha_3 y_{\mathrm{m}} \tag{8}$$

式中：t_{d} 为上升时间；Δe 为稳态值误差；y_{m} 为最大峰值；α_1、α_2、α_3 为可调节系数；r_{P}、r_{I}、r_{D} 为 PID 参数对应 3 个价值网络的原始奖励，最终经过该算法求出 r_1、r_2、r_3。

在该系统中，本文设计了三层全连接结构的神经网络作为价值网络(critic)，并将系统控制的上升时间、稳态误差幅值、动态误差幅值作为奖励的主要依据，再根据上述定义的奖励规则及训练模式对整个控制系统进行训练，构建一个 PPO 神经网络训练系统。

每个智能体程序整体训练过程如图 8 所示。

DOI: 10.16182/j.issn1004731x.joss.23-0137

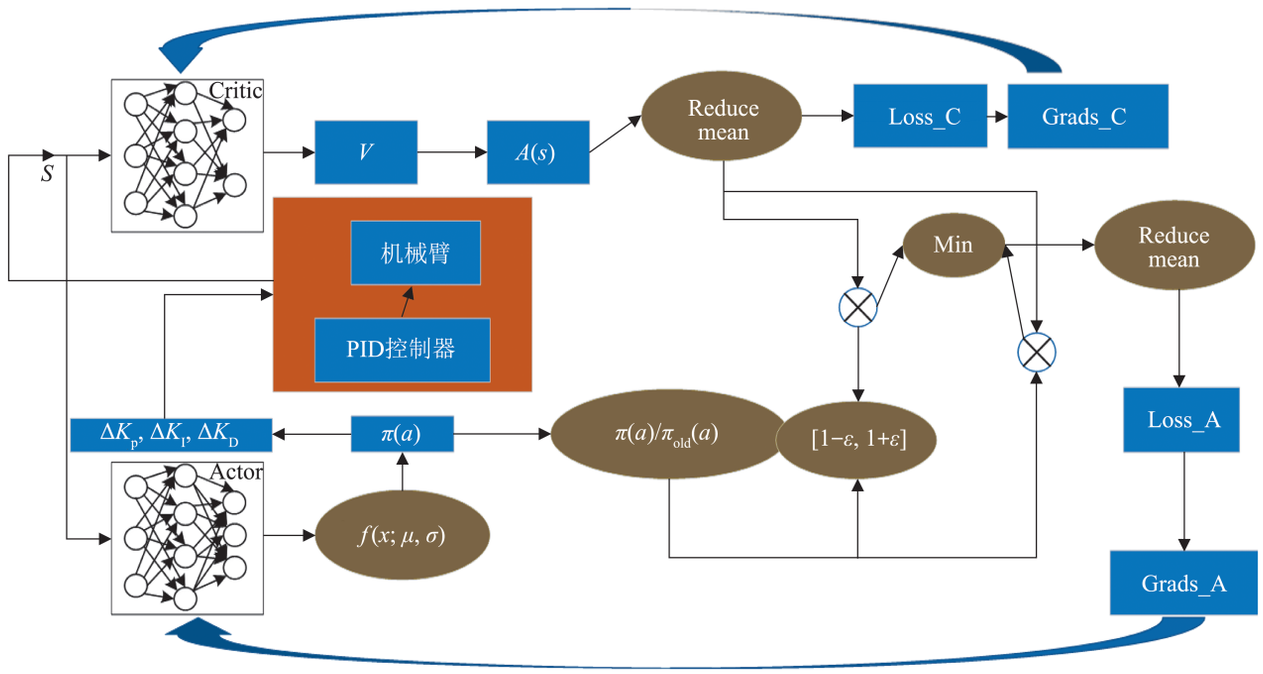

图8　整体训练流程
Fig. 8　Overall training process

2　训练结果对比分析

2.1　仿真试验对比

在MATLAB中搭建机械臂模型，并为后续提供观察、动作等仿真环境。Python提供训练模型的API接口，用于搭建神经网络以及训练神经网络。两个平台实时交互，如图9所示，机械臂按照神经网络输出的PID来控制电机，神经网络从机械臂获取本回合信息进行训练。

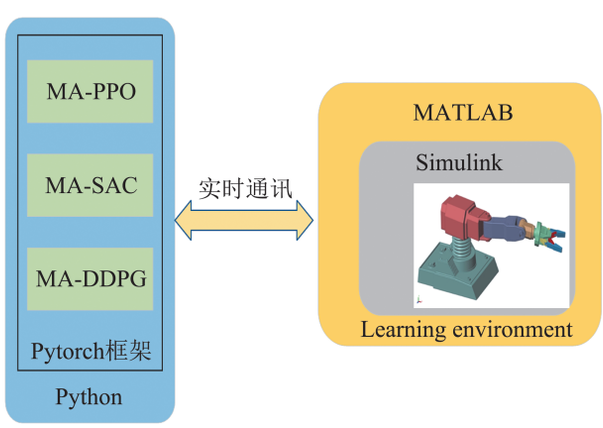

图9　整体模型
Fig. 9　Overall model diagram

本文将MA-PPO、MA-SAC、MA-DDPG算法训练结果进行对比，使用相同随机种子和环境参数，如表1和图10所示，图10中，阴影部分表示该部分的波动，阴影部分越大表示算法稳定性越低。在提供的连续控制任务中，图10所示3种算法均在训练初期逐渐上升，在20 000步时MA-PPO算法的奖励累计逐渐趋于稳定，MA-SAC算法和MA-DDPG算法达到30 000时也趋于稳定，表明3种算法在训练后期都逐渐收敛，MA-PPO算法在收敛速度和稳定性上优于其他2种算法，从而验证了MA-PPO算法的有效性。

2.2　仿真实验验证

为验证本文提出的自适应PID控制算法的优势和有效性，在Simulink模拟环境中进行实验，设置3种不同类型控制信号：正弦信号、脉冲信号和阶跃信号，实验对象为六轴机械臂，比较分析在不同的输入信号下，该六轴机械臂控制系统分别在多智能体自适应MA-PPO算法和经典PID算法控制下的系统响应结果。

表 1　超参数设定

Table 1　Environmental training structure chart training super parameter setting

参数名称	参数解释	参数取值范围	本文取值
epsilon	PPO-clip 算用于控制策略更新时新策略和旧策略的差异范围	0.1~0.3	0.2
learning rate	神经网络优化器的学习率，用于控制神经网络权重的更新速度	0.000 01~0.001	0.000 01
batch size	每个训练步骤中采样的样本数	64~512	320
buffer_size	收集的经验数，包含观测、行为与奖励用于后续训练	2 048~409 600	2 400
clip range	PPO-clip 算法中用于控制策略更新步长的截断范围	0.1~0.3	0.25
Value function coefficient	价值函数在总损失函数中的权重系数	0.5~1.0	0.7
entropy coefficien	策略的熵在总损失函数中的权重系数，用于探索	0.001~0.01	0.01

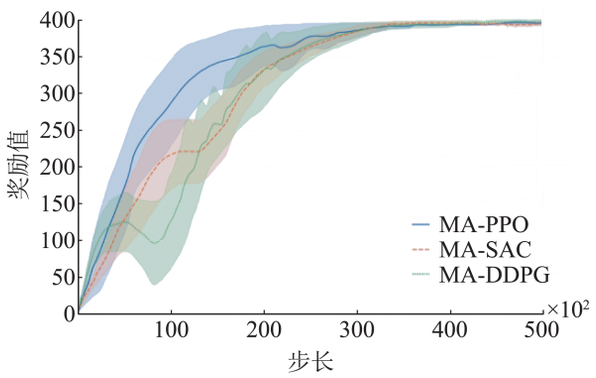

图 10　累计奖励曲线

Fig. 10　Cumulative reward curve

图 11 为加入正弦信号后的响应曲线，可见在正弦信号作用下，2 种算法的控制均能达到较好的跟随效果，没有出现超调和振荡现象。

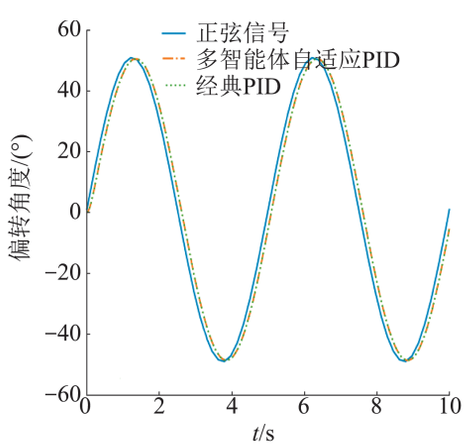

图 11　正弦信号响应曲线

Fig. 11　Sinusoidal signal response curve

图 12 为一个幅值为 60，脉宽为 1.5 s 的脉冲信号，从图中可以看到，经典 PID 算法出现了明显的超调现象，超调量达到 10% 且无法及时恢复，而本文的多智能体自适应 PID 算法并未出现超调现象且上升时间也更短。

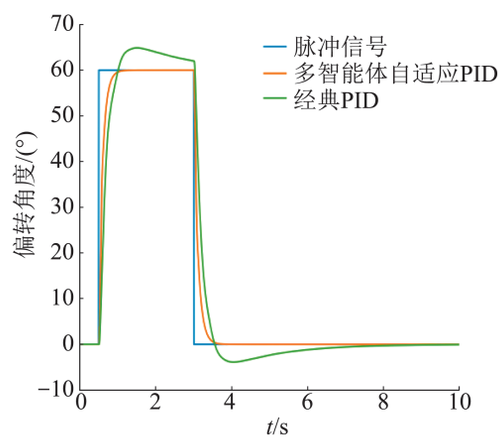

图 12　脉冲信号响应曲线

Fig. 12　Pulse signal response curve

图 13 中加入的是一个幅值为 50 的阶跃信号，传统经典 PID 算法的曲线表现出了明显的振荡，超调量也达到了 16%，而多智能体自适应 PID 算法同样没有出现超调现象也无振荡现象。

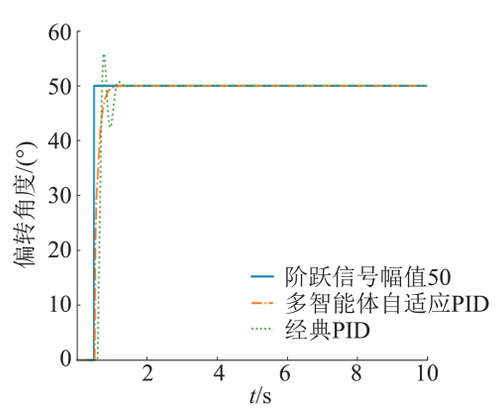

图 13　阶跃信号响应曲线

Fig. 13　Step signal response curve

DOI: 10.16182/j.issn1004731x.joss.23-0137

3 结论

通过仿真实验可以看出，本文基于机械臂的非线性等控制难点提出的PPO多智能体自适应PID算法，控制效果明显高于传统经典PID算法，利用强化学习的优势，高效解决了传统PID调参繁琐问题，同时大幅降低了建模时间，在不同的信号下，无论是正弦信号还是脉冲信号或是阶跃信号，机械臂依旧能快速稳定，实验也证明了该算法显著的鲁棒性和泛用性。

参考文献：

[1] 杜宝林, 朱大昌, 盘意华. 机械臂模糊超螺旋二阶滑模轨迹跟踪控制[J]. 系统仿真学报, 2022, 34(6): 1343-1352.
Du Baolin, Zhu Dachang, Pan Yihua. Fuzzy Super-twisting Second Order Sliding Mode Trajectory Tracking Control for Robotic Manipulator[J]. Journal of System Simulation, 2022, 34(6): 1343-1352.

[2] 张瑞民, 陈巧玉. 基于光滑二阶滑模的机械臂轨迹跟踪控制[J]. 系统仿真学报, 2021, 33(6): 1315-1322.
Zhang Ruimin, Chen Qiaoyu. Trajectory Tracking Control of Robotic Manipulators Based on Smooth Second-order Sliding Mode[J]. Journal of System Simulation, 2021, 33(6): 1315-1322.

[3] Wu Jingda, He Hongwen, Peng Jiankun, et al. Continuous Reinforcement Learning of Energy Management with Deep Q Network for a Power Split Hybrid Electric Bus[J]. Applied Energy, 2018, 222: 799-811.

[4] Schulman J, Levine S, Moritz P, et al. Trust Region Policy Optimization[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2015: 1889-1897.

[5] Zhang Yao, Deng Zhongliang, Gao Yuhui. Angle of Arrival Passive Location Algorithm Based on Proximal Policy Optimization[J]. Electronics, 2019, 8(12): 1558.

[6] 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. Chia Laguna Resort, Sardinia, Italy: PMLR, 2018: 3008-3018.

[7] Morales E F, Zaragoza J H. An Introduction to Reinforcement Learning[M]. IEEE, 2011, 11(4): 219-354.

[8] Nguyen Cong Luong, Dinh Thai Hoang, Gong Shimin, et al. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3133-3174.

[9] 李鹤宇, 赵志龙, 顾蕾, 等. 基于深度强化学习的机械臂控制方法[J]. 系统仿真学报, 2019, 31(11): 2452-2457.
Li Heyu, Zhao Zhilong, Gu Lei, et al. Robot Arm Control Method Based on Deep Reinforcement Learning[J]. Journal of System Simulation, 2019, 31(11): 2452-2457.

[10] 江达, 蔡志勤, 刘忠振, 等. 基于强化学习的连续型机械臂自适应跟踪控制[J]. 系统仿真学报, 2022, 34(10): 2264-2271.
Jiang Da, Cai Zhiqin, Liu Zhongzhen, et al. Reinforcement-learning-based Adaptive Tracking Control for a Space Continuum Robot Based on Reinforcement Learning[J]. Journal of System Simulation, 2022, 34(10): 2264-2271.

[11] Mahmoud Elsisi, Karar Mahmoud, Matti Lehtonen, et al. An Improved Neural Network Algorithm to Efficiently Track Various Trajectories of Robot Manipulator Arms[J]. IEEE Access, 2021, 9: 11911-11920.

[12] Duc-Thien Tran, Hoai-Vu-Anh Truong, Kyoung Kwan Ahn. Adaptive Nonsingular Fast Terminal Sliding Mode Control of Robotic Manipulator Based Neural Network Approach[J]. International Journal of Precision Engineering and Manufacturing, 2021, 22(3): 417-429.

[13] Yang Shichun, Xie Hehui, Chen Fei, et al. Research on Manipulator Trajectory Tracking Based on Adaptive Fuzzy Sliding Mode Control[C]//2020 Chinese Automation Congress (CAC). Piscataway, NJ, USA: IEEE, 2020: 3086-3091.

[14] Saim Ahmed, Wang Haoping, Tian Yang. Adaptive High-order Terminal Sliding Mode Control Based on Time Delay Estimation for the Robotic Manipulators with Backlash Hysteresis[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(2): 1128-1137.

[15] Ma Yajun, Zhao Hui, Li Tao. Robust Adaptive Dual Layer Sliding Mode Controller: Methodology and Application of Uncertain Robot Manipulator[J]. Transactions of the Institute of Measurement and Control, 2022, 44(4): 848-860.

[16] Mohammadi F, Mohammadi-Ivatloo B, Gharehpetian G B, et al. Robust Control Strategies for Microgrids: A Review[J]. IEEE Systems Journal, 2022, 16(2): 2401-2412.

[17] Amit Konar, Indrani Goswami Chakraborty, Sapam Jitu Singh, et al. A Deterministic Improved Q-learning for Path Planning of a Mobile Robot[J]. IEEE Transactions

on Systems, Man, and Cybernetics: Systems, 2013, 43 (5): 1141-1153.

[18] Zhou Changjiu, Meng Qingchun. Dynamic Balance of a Biped Robot Using Fuzzy Reinforcement Learning Agents[J]. Fuzzy Sets and Systems, 2003, 134(1): 169-187.

[19] Wu Hui, Song Shiji, You Keyou, et al. Depth Control of Model-free AUVs via Reinforcement Learning[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 49(12): 2499-2510.

[20] 魏楠哲. 空间机械臂柔性关节高精度控制研究[D]. 北京: 北京邮电大学, 2016.
Wei Nanzhe. Study on Flexible Joint Control System with High Precision for Space Manipulator[D]. Beijing: Beijing University of Posts and Telecommunications, 2016.

[21] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[EB/OL]. (2017-08-28) [2023-01-12]. https://arxiv.org/abs/1707.06347.

DOI: 10.16182/j.issn1004731x.joss.23-0137