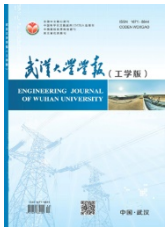


武汉大学学报(工学版)
Engineering Journal of Wuhan University
ISSN 1671-8844,CN 42-1675/T

《武汉大学学报(工学版)》网络首发论文

题目：基于深度强化学习 TD3 的 PID 参数自整定算法
作者：梁杰，专祥涛，严家政
网络首发日期：2023-04-13
引用格式：梁杰，专祥涛，严家政．基于深度强化学习 TD3 的 PID 参数自整定算法[J/OL]．武汉大学学报(工学版).
https://kns.cnki.net/kcms/detail/42.1675.T.20230412.1609.002.html

网络首发时间：2023-04-13 10:20:36
网络首发地址：https://kns.cnki.net/kcms/detail/42.1675.T.20230412.1609.002.html

基于深度强化学习 TD3 的 PID 参数自整定算法

梁杰[1]，专祥涛[1,2]，严家政[1]

(1.武汉大学电气与自动化学院，湖北 武汉 430072；2.武汉大学深圳研究院，广东 深圳 518057)

摘要：传统的 PID（Proportional Integral Differential，PID）算法在用于控制一些模型复杂、参数时变的对象时存在参数整定过程繁琐、控制性能不佳、无法解决控制对象实时变化的状态的影响等问题。针对以上问题，提出了一种基于双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）算法的 PID 参数自整定算法。该算法将 TD3 算法与 PID 算法相结合，对 TD3 算法中的神经网络结构、奖励函数进行设计，能够实现控制器参数的自整定。以两轮直立车为实验对象，对直立车的角度 PID 控制器进行参数整定实验。实验结果表明，与传统的参数整定算法（Z-N 参数整定法）和基于强化学习的动态 PID 参数自整定算法相比，所提出的算法具有更优的控制效果，可以通过神经网络学习、拟合更优的控制策略，提升控制器的动态响应性能和鲁棒性。

关键词：深度强化学习；TD3 算法；整定；控制器；直立车

中图分类号：TP273+.2　　**文献标志码：**A

PID parameter self-tuning algorithm based on deep reinforcement learning TD3

LIANG Jie[1], ZHUAN Xiangtao[1,2], YAN Jiazheng[1]

(1. School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, Hubei, China; 2. Shenzhen Research Institute, Wuhan University, Shenzhen 518057, Guangdong, China)

Abstract: When the traditional PID (Proportional Integral Differential, PID) algorithm is used to control some objects with complex models and time-varying parameters, there are some problems, such as cumbersome parameter setting process, poor control performance, and unable to solve the influence of the real-time changing state of the control object. To solve the above problems, a PID parameter self-tuning algorithm based on twin delayed deep deterministic policy gradient algorithm was proposed. The algorithm combines TD3 algorithm with PID algorithm, and designs the neural network structure and reward function in TD3 algorithm, which can realize the self-tuning of controller parameters. Taking a two-wheel upright vehicle as the experimental object, the Angle PID controller of the upright vehicle is used for parameter setting experiment. Experimental results show that compared with the traditional parameter tuning algorithm (Z-N parameter tuning method) and dynamic PID parameter self-tuning algorithm based on reinforcement learning, the proposed algorithm has better control effect, and can improve the dynamic response performance and robustness of the controller by learning and fitting better control strategy through neural network.

Key words: deep reinforcement learning; TD3 algorithm; tuning; controllers; upright car

作者简介:梁杰(1997-)，男，硕士研究生，主要从事系统建模与最优控制方面的研究，E-mail: 2016301470076@whu.edu.cn
通讯作者:专祥涛(1975-)，男，教授，主要从事系统建模与最优控制方面的研究，E-mail: xtzhuan@whu.edu.cn
基金项目:深圳市知识创新计划项目(JCYJ20170818144449801)。

控制器的参数整定是现代工业控制过程中的重要环节，参数整定的目的是为了获得更好的控制性能指标，以保证系统稳定高效地运行[1]。工业控制系统通常具有非线性、多变量的复杂特性，对此专家学者们提出了模糊 PID 控制[2,3]、滑模控制[4,5]、自抗扰控制[6,7]等算法来提升控制器的性能。然而，这些算法的控制器参数整定需要经验丰富的专业人员耗费大量的时间精力来进行，参数整定优化过程繁琐、耗时耗力。除此之外，也能通过对控制系统进行建模，根据控制对象的模型进行推理计算得出控制器的参数。但对于某些模型复杂、参数时变的系统而言，建立其模型十分困难，并且由于其参数时变，模型也要即时地进行调整，建立模型对其进行控制器参数整定工程量巨大。随着计算机技术的不断发展，强化学习、深度学习等人工智能技术被越来越多地应用于控制领域[8,9]，例如车辆自动驾驶控制[10]、无人机的自主飞行控制[11]、机械臂的目标点到达控制[12]等。同时，由于深度学习算法的数据特征提取能力[13]、强化学习算法的决策优化能力[14,15]，深度学习和强化学习算法也能用于控制器参数的整定和动态调节，以提升控制系统的性能。目前已有一些学者将深度强化学习算法中的 DDPG（Deep Deterministic Policy Gradient，DDPG）算法用于 PID 参数的整定和优化[16,17]。但 DDPG 算法容易产生 Q 值过度估计的问题，即算法中价值网络预测的期望 Q 值与真实值相比存在较大的差别，这可能会影响算法中价值网络参数学习的准确性，导致策略网络参数难以收敛到最优目标。此外，这些研究大多处于理论和仿真阶段，少有实际工程实验的验证。

基于此，本文提出了一种基于深度强化学习 TD3 的 PID 参数自整定算法，将 TD3 算法与 PID 算法相结合，对 TD3 算法中的神经网络结构、奖励函数进行设计，以两轮直立车为实验对象，通过改进 TD3 算法对直立车的角度 PID 控制器进行参数整定，利用整定得到的控制器对直立车的俯仰角进行控制，并与传统的 Z-N 参数整定法、基于强化学习的动态 PID 参数自整定算法对比，验证了所提算法的可行性、有效性。

1 深度强化学习

深度强化学习[18,19]（Deep reinforcement learning，DRL）将深度学习与强化学习结合，同时利用了深度学习的感知能力和强化学习的决策能力。其中，强化学习作为机器学习的分支之一，其思想源于人们对于行为心理学的研究。由于其在学习过程中不用具体地了解控制对象的系统结构，也不用对学习过程中的每个状态标注监督信号，所以可以被用于解决一些繁杂的优化决策问题。例如，采用强化学习算法对某些系统的控制参数进行自整定，从而避免繁杂的人为整定参数过程。然而，对于某些模型复杂，参数时变的控制对象来说，仅以强化学习为基础的参数自整定算法[20]无法很好地解决控制对象实时变化的状态的影响。所以，需要采用深度强化学习算法将强化学习和深度学习相结合，对系统的控制参数进行自整定。在结合了深度学习的提取多维数据特征的能力后，深度强化学习算法只需要给出初始的数据，而不用通过人工特征来生成输出，进而避免了每个独立学习模块执行任务之前都需要做数据标注的繁琐过程，真正实现了端到端的学习。

双延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）算法适合解决动作空间连续且多维的系统控制问题，并且其能够有效解决其他算法中容易出现的 Q 值过度估计问题，使得训练过程更加稳定。所以，本文采用 TD3 算法，以两轮直立车为实验对象，对直立车的直立角度进行控制。

2 算法设计

两轮直立车的实物图如图 1 所示，其简化的侧面模型图如图 2 所示。图 2 中 R 为直立车车轮半径，M 为电机转矩，γ为车轮转角，L 为车质心到车轴的距离，θ为直立车的直立角

度，也就是本文实验需要控制的变量。

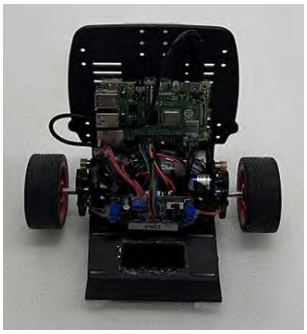


图 1 直立车实物图

Fig.1 Physical map of upright vehicle

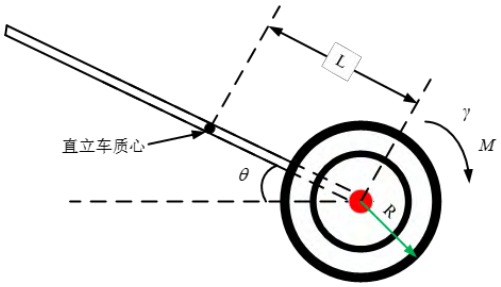


图 2 直立车简化模型图

Fig.2 Simplified model diagram of upright vehicle

2.1 算法的主要框架

在本文的直立车角度控制实验中，若直接采用常规的 TD3 算法对直立车进行控制存在以下问题：

（1）若直接使用 TD3 算法对直立车的角度姿态进行控制，即 TD3 算法的 Actor 策略网络直接输出电机的占空比来控制直立车的角度，其动作空间的搜索范围太大，这会使得算法训练过程较长。此外，在算法的训练阶段，如果直接控制电机的输出会导致直立车状态不稳定，容易产生较多的低性能样本，使得训练过程的收敛性变差。

（2）需要合理地设计深度强化学习 TD3 算法中的神经网络结构。因为若神经网络设置地较为简单，可能无法完全体现动作与价值之间的函数映射关系；而神经网络设置地过于复杂则会加大计算成本，降低智能体的学习效率。

（3）深度强化学习 TD3 算法中的奖励函数体现了监督信号的作用，是 TD3 算法中最为重要的环节之一。若奖励函数设计的不合理会给智能体带来错误的引导，将导致策略最终收敛至局部最优解。

针对以上三个问题，本文对常规 TD3 算法进行了如下的改进：

（1）将 TD3 算法和 PID 控制算法相结合，设计了一种复合控制算法，其算法的框架结构如图 3 所示。TD3 算法的策略网络接收状态观测数据输入，其输出动作为 PID 控制器参数的增量值，即利用 TD3 算法学习并自整定 PID 控制器的参数。这种复合控制算法能够提高强化学习训练过程的安全性和可控性，减少低性能样本数量，提高训练过程的收敛性。

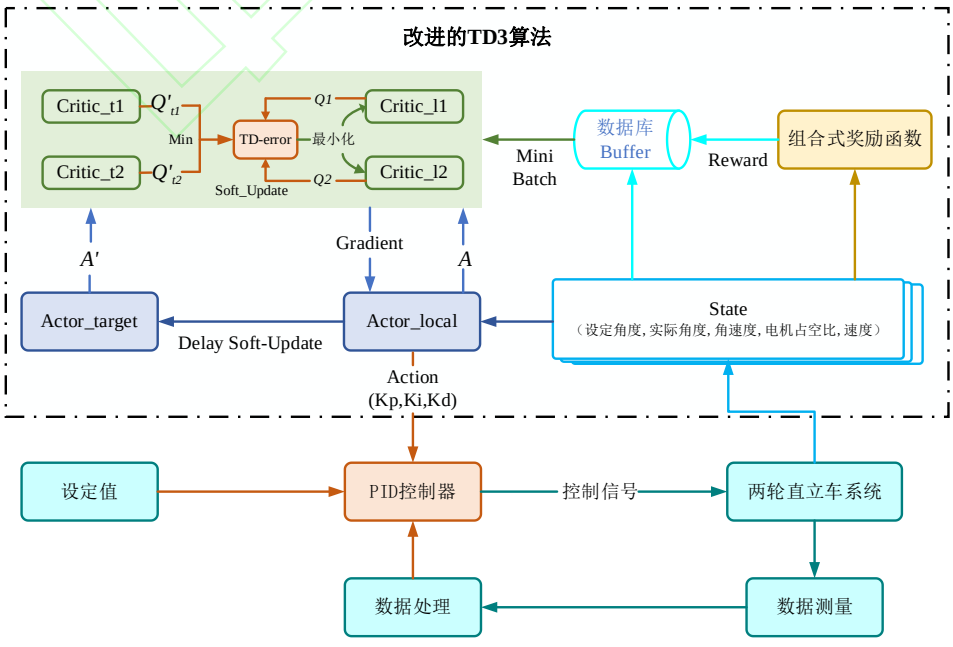


图 3 基于 TD3 的复合控制算法结构框图

Fig.3 Structure block diagram of composite control algorithm based on TD3

（2）根据本文控制对象的状态特征，设计了一种神经网络结构。通过改变神经网络的节点和隐藏层数量，并选择恰当的激活函数，以提升深度强化学习算法中模型的学习效率。

（3）针对本文实验对象两轮直立车的角度控制目标，设计了一种组合式的奖励函数，通过连续性的奖励，提供学习最优控制策略的价值估计引导，提高训练的收敛速度。

2.2 动作和状态的设计

对应着两轮直立车的三个 PID 控制参数K_p、K_i、K_d，本实验中的 TD3 算法其空间维度为三维，动作$a_t = [U_p, U_i, U_d]$。其中，每个动作的取值均在-1~1 之间。PID 控制器的三个参数更新公式如下所示：

$$\begin{cases} K_p = B_p + N_p U_p \\ K_i = B_i + N_i U_i \\ K_d = B_d + N_d U_d \end{cases} \tag{1}$$

上式中，K_p、K_i、K_d是 PID 控制器的实际参数，B_p、B_i、B_d为 PID 控制器参数的初始值，N_p、N_i、N_d为控制器动态参数的偏移量。其中B_p、B_i、B_d，N_p、N_i、N_d需根据系统的实际情况进行调整。

本文的实验以两轮直立车为控制对象，采用改进的 TD3 算法进行 PID 参数的自整定，控制目的为对直立车进行稳定的角度控制。为达到这一目的，选择合适的状态观测量（即状态S_t中包含的数据）至关重要。对于直立车系统，车的俯仰角度（控制目标）、车模的运行速度等都可以作为状态观测量，本文所选取的状态观测量及其说明如表 1 所示。

表 1 实验选取的状态观测量

Table 1 Experimental selection of state observations

变量名	单位	范围	说明
speed	cm/s	-500~500	当前时刻的直立车运行速度
angle_set	°	35~50	直立车需要达到的角度设定值
pitch	°	0~80	当前时刻的直立车俯仰角
pitch_dot	(°)/s	-125~125	当前时刻的直立车角速度(俯仰角)
pwm	%	0~100	上一时刻的电机占空比

2.3 神经网络结构和奖励函数的设计

为了提升算法的学习效率、收敛性、增强神经网络参数的稳定性，本文对 TD3 算法中的神经网络结构、奖励函数进行设计。

2.3.1 神经网络结构的设计

由于本文控制对象的状态特征相对简单，故选取较为一般性的深度神经网络（Deep Neural Networks，DNN）作为智能体策略网络和价值网络的神经网络基础。通过简化 DNN 的神经元个数及层数来降低算法训练过程中的计算量，同时提高算法的训练效率。本文所建立的 TD3 策略网络包括 4 层：状态输入层、隐藏层 A1、隐藏层 A2、动作输出层，神经元个数分别为[5，64，48，3]；价值网络包括 5 层：输入层、隐藏层 C1、隐藏层 C2、隐藏层 C3、Q 值输出层，神经元个数则设计为[8，48，48，24，1]，神经网络的结构图如图 4 所示。

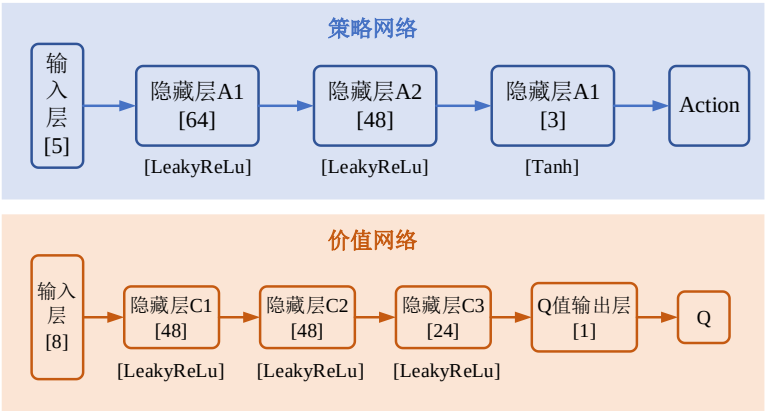


图 4 TD3 的神经网络结构图

Fig.4 Neural network structure diagram of TD3

为提高网络模型对非线性系统特征的学习能力，本文选取 LeakyReLU 函数作为输入层和隐藏层的激活函数，其表达式如下：

$$LeakyReLU(x)=\begin{cases}x, x\geq 0\\ \lambda x, x<0\end{cases} \tag{2}$$

上式中，λ为超参数，当λ的值为 0 时，式（2）就变成了 ReLu 函数，ReLu 函数会出现“神经元死亡”问题，即当 ReLu 函数的输入一直为负的时候，反向传播过程经过该处的梯度一直为 0，导致神经元无法更新参数，不再学习。和 ReLU 函数相比，LeakyReLU 函数在输入值为负的时候保留了较小的梯度值，其超参数λ不为 0 而是一个较小的值，有效解决了“神经元死亡”的问题。超参数λ的取值通常为 0.001~0.1，在本文的实验中取为 0.001。

2.3.2 *奖励函数的设计*

奖励函数的作用是引导智能体学习预先期望的价值函数，进而学习获得符合任务需求的最优化策略，是深度强化学习算法中最重要的环节。奖励函数主要可分为离散型奖励函数和连续型奖励函数。离散型奖励函数易于实现，但是只适用于解决简单动作的强化学习问题，对于复杂的模型对象，离散型的奖励分布比较稀疏，会出现智能体的动作探索时间过长、策略陷入局部最优等问题。故本文采取连续型的奖励函数，但连续型奖励函数的设计方案比较复杂，所以需要根据实际问题进行调整。据此，本文设计了一种组合式的奖励函数，以两轮直立车的角度控制为奖励函数基础，附加安全状态奖励、动作奖励等奖励函数，提供学习最优控制策略的价值估计引导。基于观测的状态S_t、S_{t+1}与执行的动作A_t，本实验对奖励函数做如下设计：

（1）角度控制经历函数r_1。本文实验的控制目标时实现直立车俯仰角的稳定和准确控制。据此，将直立车俯仰角的角度误差 e 和角速度ω作为奖励函数的自变量，如下：

$$r_1=-c_0*e^2-c_1*\omega^2 \tag{3}$$

上式中，其中c_0、c_1为角度权重和角速度权重，表示直立车角度和角速度对整体奖励值的影响权重。

（2）安全奖励函数r_2。对直立车处于安全区间的角度状态给予确定值的奖励，以提升直立车训练过程中的稳定性、安全性，如下：

$$r_2=c_2 \tag{4}$$

上式中，c_2是安全状态奖励权重，此函数表明直立车处于安全角度区间的时间越长以获得的累计奖励值就越大。

（3）动作奖励函数r_3。本文实验的控制器结合了 PID 控制算法，将控制器的输出u(直立车电机 PWM 占空比)作为奖励函数自变量，奖励函数公式如下所示：

$$r_3=-c_3*u^2 \tag{5}$$

上式中c_3为动作奖励权重，表示的是控制器的输出u对最终奖励值的影响程度。

（4）速度奖励函数r_4。直立车的模型参数不仅受其俯仰角大小的影响，同时也受车模行驶速度v的影响。为提升角度控制的稳定性，将速度变量也加入奖励函数的评估。在角度控制实验中，期望的车模行驶速度应接近于零，据此设计的奖励函数如下：

$$r_4 = -c_4 * |v| \tag{6}$$

上式中，c_4为速度奖励权重，表示了直立车速度对整体奖励值的影响程度，本文实验中将其设置地很小。

本文的实验为持续性的强化学习控制问题，将安全边界条件（即直立车的俯仰角大于75°或者小于0°）作为训练的终止状态S_{end}。本文设计的奖励函数非终止状态的最差奖励极限接近-10，为了进行一种非线性区分，促进智能体学习并保持安全状态，当直立车处于终止状态时，给予惩罚性的奖励值-20。综上，本文设计的组合式奖励函数 R 如下：

$$R = r_1 + r_2 + r_3 + r_4 = \begin{cases} -(c_0 e^2 + c_1 \omega^2 + c_3 u^2 + c_4 |v|) + c_2, S_t \neq S_{end} \\ -20 \end{cases}, S_t = S_{end} \tag{7}$$

3 实验及结果分析

3.1 改进 TD3 整定 PID 参数

采用本文设计的基于改进 TD3 的参数自整定算法对两轮直立车的角度 PID 控制器进行参数整定的实验。实验采用正弦形式的信号作为直立车角度跟随的期望值。算法通过采集经验样本池中的数据进行训练，为了探索更多的状态空间及经验数据，限制每回合的训练时间最多为 10 秒，超过这个时间立刻对直立车进行环境初始化，并进入下一回合的训练。神经网络的梯度更新计算需耗费一定的时间，为保证实验过程中控制程序稳定运行，直立车的角度控制程序和深度强化学习算法的学习程序是并行、异步进行的。其中，控制程序的执行周期是 0.02s，深度强化学习程序的更新周期是 0.1s。本文提出的改进 TD3 算法仍然采用了目标网络软更新策略以限制更新幅度，软更新系数能够降低系统训练过程中的数据噪声的影响，其值应在 0~1 之间，且比 1 小得多，在本文的实验中，软更新系数τ设置为常用的通用值 0.005。奖励折扣因子是对未来奖励的考量，本文选取常用的值 0.99。奖励权重系数c_0~c_4是根据相应的实验状态变量的量程范围确定的，状态量的量程如表 1 所示，奖励权重系数c_0~c_4确定后，实验中的最差奖励极限接近-10，最好奖励为 1。为保证系统的安全性，对策略网络的动作输出和 PID 控制参数的范围进行限制。具体训练参数如下表所示：

表 2 改进 TD3 算法角度控制实验训练参数

Table 2 Improved TD3 algorithm angle control experiment training parameters

参数名	数值	参数	数值
最大训练轮次	300	角度奖励系数c_0	0.001
经验回放池容量	100000	角速度奖励系数c_1	0.00012
奖励折扣因子	0.99	安全状态奖励系数c_2	1
软更新系数 τ	0.005	动作奖励系数c_3	0.0005
单回合最大时间(秒)	10	速度奖励系数c_4	0.001
期望的平均奖励值	170	动作 Kp 范围	1.8~2.3
小批量梯度的样本数量	128	动作 Ki 范围	0.6~1.1
自适应学习因子初值α_0	0.002	动作 Kd 范围	0.07~0.13

3.2 性能测试实验

（1）角度跟随性能测试。改进的 TD3 算法能够保留训练过程中的网络参数和经验样本池等数据，以便于算法模型的再训练。整定完成后的 TD3 策略网络保存为参数文件，以离

线神经网络函数的形式被控制程序调用。基于该策略网络的两轮直立车控制系统如图 5 所示。为验证整定结果的有效性，设定与训练过程相同角度期望值，进行角度跟随实验，TD3-PID 控制器响应过程中的 PID 参数曲线如图 6 所示，角度跟随的实验结果如图 7 所示。

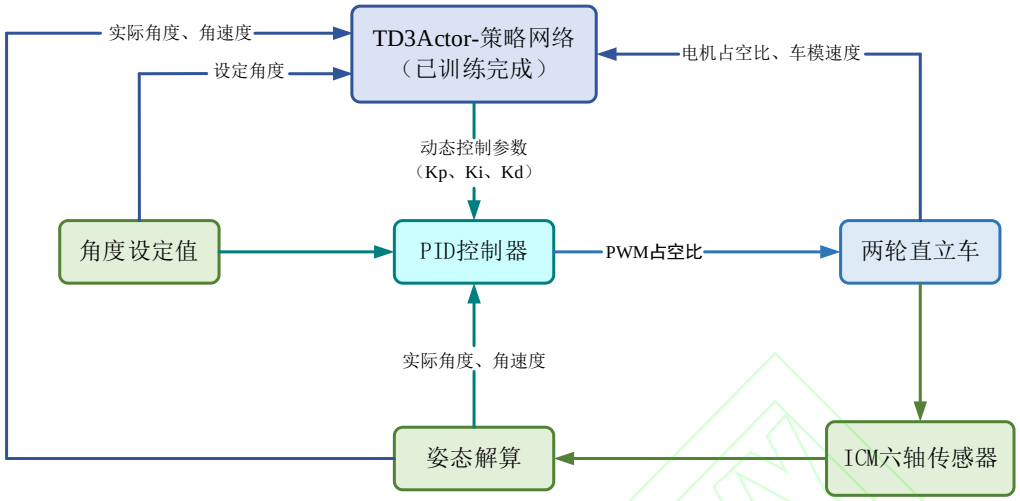
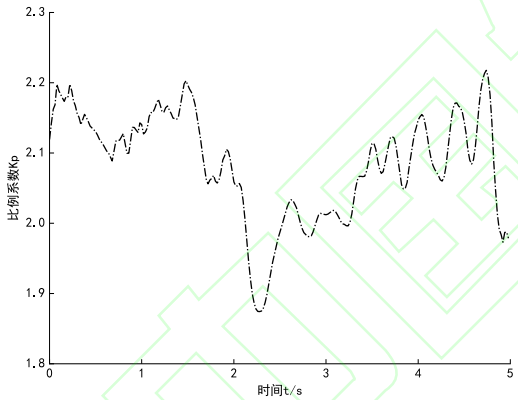
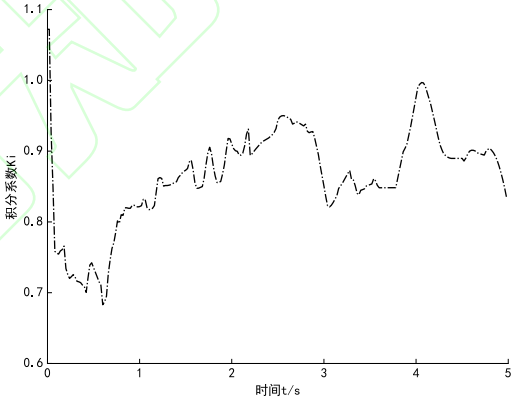


图 5 改进 TD3-PID 算法整定完成后的直立车控制系统框图

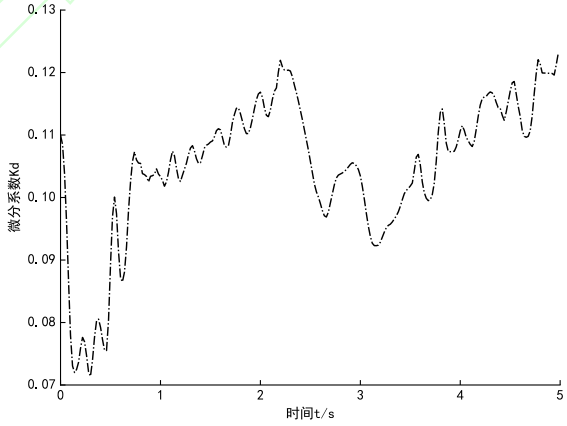
Fig.5 The block diagram of the upright vehicle control system after the improved TD3-PID algorithm is tuned



(a)比例系数变化曲线



(b)积分系数变化曲线



(c)微分系数变化曲线

图 6 改进 TD3-PID 控制器响应过程的动态 PID 参数曲线

Fig.6 Dynamic PID parameter curve of improved TD3-PID controller response process

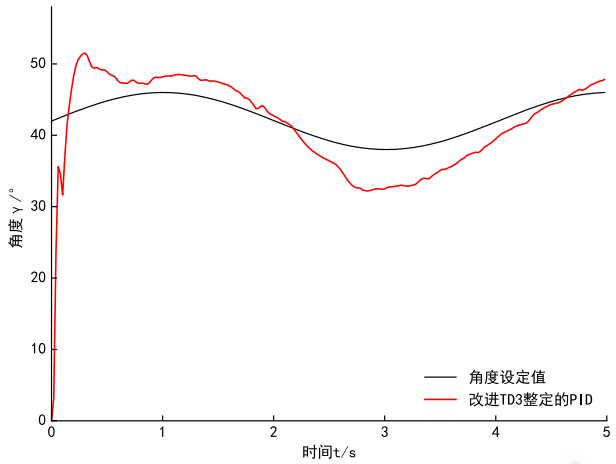


图 7 改进 TD3-PID 控制器的正弦形式角度跟随曲线

Fig.7 Improved TD3-PID controller sine form angle following curve

图 6 显示了角度跟随实验过程中 TD3-PID 控制器的各项参数，由图 6 可知，在实验过程中 TD3-PID 控制器的K_p、K_i、K_d参数随时间不断变化。由于在实验中直立车的各项模型参数会随着时间及直立车的状态发生改变，TD3 算法的策略网络可以基于直立车不同的瞬时状态，对控制器的 PID 参数进行动态调节，通过实时整定的参数对直立车的角度进行控制。图 7 显示了在设定期望角度值下，直立车角度跟随的效果曲线，由图 7 可知，初始时直立车处于平放状态，直立角度为 0，与设定的期望值偏差较大，在 TD3-PID 控制器的控制下，直立车角度迅速达到期望值附近，之后随着期望角度值的变化，直立车的实际角度值能够较好地跟随设定期望角度，验证了本文所提出的改进 TD3 算法能够实时调整两轮直立车控制器的 PID 参数，实现直立车的角度跟随。改进 TD3 算法在控制过程中针对不同时刻的直立车状态，对 PID 参数进行动态调节，可以提升控制系统的响应性能和鲁棒性。

（2）抗干扰能力测试。在两轮直立车角度姿态稳定的前提下，给车加上不同大小、不同方向的暂态干扰，改变直立车的实际角度，以测试控制器的抗干扰能力。直立车受到干扰和恢复至稳定状态的响应曲线如图 8 所示。本文设定直立车状态恢复为稳定的边界条件为设定角度的 5%左右误差范围内，即稳定区间为设定角度±2°。如图 8，直立车在干扰消失后，经过大约 0.5 秒的时间就回到了近似稳定的状态，这说明了改进 TD3 算法整定的 PID 控制器具有较好的抗干扰能力。

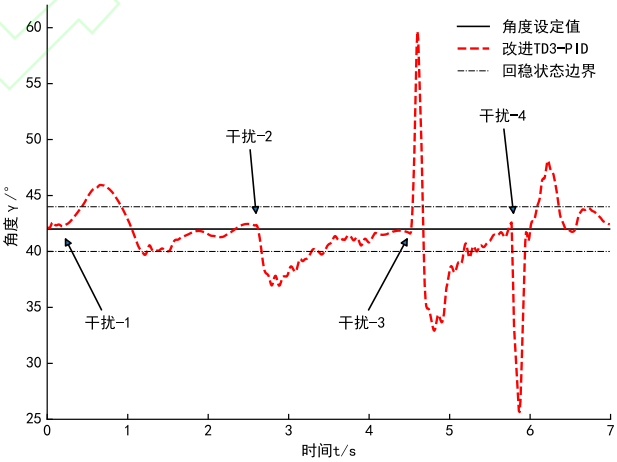


图 8 改进 TD3-PID 控制器的扰动响应曲线

Fig.8 The disturbance response curve of the improved TD3-PID controller

3.3 性能对比实验

为进一步测试 TD3 算法整定 PID 控制器的性能，设计对比实验。首先采用 Z-N 临界比例度法整定两轮直立车的 PID 参数进行对比，得到的固定 PID 参数如下：

$$\begin{cases} kp = 2.1 \\ ki = 0.9 \\ kd = 0.11 \end{cases} \tag{8}$$

此外，采用基于强化学习的动态 PID（Reinforcement learning dynamic PID，RLD-PID）参数自整定算法[18]对直立车的角度控制器参数进行整定，动态 PID 系数的计算公式如下：

$$\begin{cases} K_p = 2.1 + P_1 e + P_2 \omega + P_3 e^3 \\ K_i = 0.9 + I_1 e + I_2 \omega + I_3 e^3 \\ K_d = 0.1 + D_1 e + D_2 \omega + D_3 e^3 \end{cases} \tag{9}$$

上式中，K_p、K_i、K_d为 PID 控制器的参数，e 为归一化的角度误差，ω为直立车的角速度，P_i、I_i、$D_i(i = 1,2,3)$为待整定的参数。基于文献[18]中的参数自整定算法对公式中的P_i、I_i、$D_i(i = 1,2,3)$进行整定，得到的整定结果如下：

表 3 基于强化学习的动态 PID 参数自整定算法（RLD-PID）整定结果

Table 3 Tuning results of dynamic PID parameter self-tuning algorithm based on reinforcement learning

参数名	P_1	P_2	P_3	I_1	I_2	I_3	D_1	D_2	D_3
数值	-0.142	-0.082	-0.064	-0.035	-0.092	-0.066	0.039	-0.073	0.014

（1）角度跟随性能对比。将设定值设置为正弦信号的形式分别采用 Z-N 法、RLD-PID 算法、以及本文提出的改进 TD3-PID 算法整定获得的 PID 参数及控制器，进行两轮直立车的角度姿态跟随实验，实验的结果如图 9 及表 4 所示：

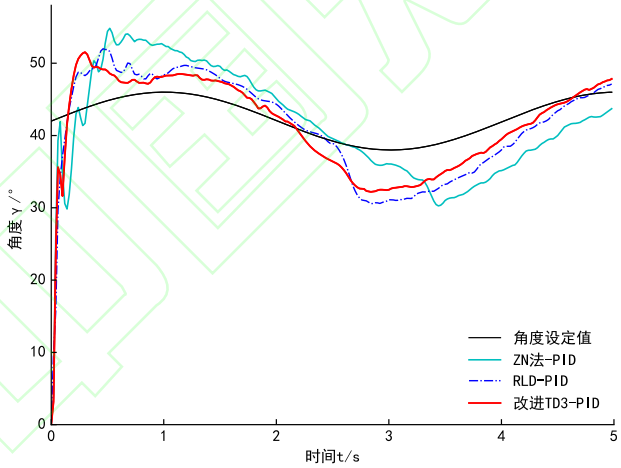


图 9 三种算法整定 PID 参数的控制器角度跟随实验曲线

Fig.9 The controller angle of the three algorithms for tuning PID parameters follows the experimental curve

由图 9 和表 4 可知，与 ZN 法和 RLD-PID 算法相比，基于改进 TD3 算法整定的 PID 控制器能够利用更多的状态特征信息，学习更优的控制策略，实现了误差绝对值均值更小、响应速度更快的角度控制，并且具有更强的鲁棒性。

表 4 三种参数整定方法的两轮直立车角度跟随实验数据

Table 4 Two-wheel upright vehicle angle following experimental data with three parameter tuning methods

参数整定方法	方法依据	PID 参数类型	误差绝对值标准积分	角度误差标准差
Z-N 临界比例度法	阶跃响应|经验公式	定值	1223.750	6.360
RLD-PID 算法	超调量|上升时间|调节时间	动态	974.268	5.502

改进 TD3-PID 算法	每一时刻的状态|奖励函数	动态	885.146	5.191

（2）抗干扰能力对比。对三种方法整定参数的控制器，在直立车角度稳定的状态下，人为地施加尽可能相同的干扰，使得直立车的角度从 45°降至 30°，对比三种控制器在扰动消失后恢复响应的过程。抗干扰性能对比实验的结果如图 10 所示。

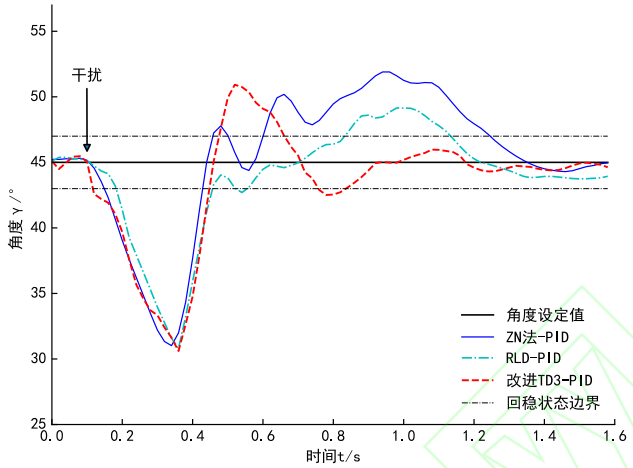


图 10 三种方法整定 PID 参数的控制器抗干扰性能对比曲线

Fig.10 Comparison curve of anti-interference performance of controllers for tuning PID parameters by three methods

图 10 记录了三种整定方法所获得的控制器的抗干扰实验结果，仍然以设定角度±2°作为恢复至稳定的边界判断条件。由图可知，与 Z-N 法相比，RLD-PID 算法整定的 PID 控制器显著降低了恢复过程中的超调量，在 1.2s 时刻基本恢复稳定状态；基于改进 TD3 算法整定的 PID 控制器，在 0.9s 的时刻基本恢复稳定状态，与另外两种方法相比，有效减少了恢复过程的时间，其超调时间相对较为提前，而超调量没有明显改善，但仍然在可接受的角度范围内。

4 结论

本文针对数据特征较为复杂的控制问题，选择了深度强化学习算法中的双延迟深度确定性策略梯度算法作为控制器自整定算法的基础，并对其进行了多个方面的改进，包括：结合 PID 控制算法，对智能体策略的动作值进行调整，以提升算法训练过程中的安全性和可控性；对深度神经网络结构进行优化，选择合适的隐藏层个数和激活函数等参数；设计提出了一种组合式的奖励函数，引导智能体向最优策略学习。基于两轮直立车的角度控制实验，验证了基于改进 TD3 的参数自整定算法应用于随动控制系统的 PID 控制器的可行性和有效性，对比人工整定的 PID 控制器（Z-N 法整定）和基于强化学习的动态 PID（RLD-PID）参数自整定算法所获得的控制器，证明了深度强化学习算法可以通过神经网络，学习、拟合更优的控制策略，在控制过程中基于系统的状态自适应地调整 PID 参数，提升控制器的动态响应性能和鲁棒性。

本文对两轮直立车的控制器参数整定实验，仅针对其角度跟随控制这一单一问题，实际的控制任务中还包括直立车的速度控制和方向控制等多输入多输出的控制问题。因此，关注复杂的耦合系统，提高算法的广泛使用性，是未来的重要研究内容。

参考文献

[1] 李国林. PID 控制器参数整定技术研究与优化设计[D]. 大连理工大学, 2010.
Li Guolin. Research and optimization design of PID controller parameter tuning technology [D]. Dalian University of Technology, 2010.

[2] 吴雪颖, 严芝健, 颜翰宇, 等. 基于模糊 PID 控制器的电动汽车放电控制研究[J]. 电源技术, 2020, 44(11): 1662-1665+1700.
Wu Xueying, Yan Zhijian, Yan Hanyu, et al. Electric Vehicle Discharge Control Based on Fuzzy PID Controller [J]. Chinese Journal of Power Supply, 2020, 44(11): 1662-1665+1700.

[3] Nguyen Dinh Phu et al. A New Fuzzy PID Control System Based on Fuzzy PID Controller and Fuzzy Control Process[J]. International Journal of Fuzzy Systems, 2020, 22(7): 1-25.

[4] 高苗苗. 基于滑模控制的多机械臂同步控制研究[D]. 浙江工业大学, 2019.
Gao Miaomiao. Research on Synchronous Control of Multi-Manipulator Based on Sliding Mode Control [D]. Zhejiang University of Technology, 2019.

[5] Lal Ioana and Codrean Alexandru and Buşoniu Lucian. Sliding mode control of a ball balancing robot[J]. IFAC PapersOnLine, 2020, 53(2): 9490-9495.

[6] Baquero Suárez Mauro et al. A robust two-stage active disturbance rejection control for the stabilization of a riderless bicycle[J]. Multibody System Dynamics, 2019, 45(1): 7-35.

[7] 吴林. 基于自抗扰控制的风力发电机变速变桨控制系统研究[D]. 南京信息工程大学, 2020.
Wu Lin. Research on Variable Speed and Rotor Control System of Wind Turbine Based on Active Disturbance Rejection Control [D]. Nanjing University of Information Science & Technology, 2020.

[8] Shuti Wang et al. Correction to: Trajectory Tracking Control for Mobile Robots Using Reinforcement Learning and PID[J]. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 2020, 44(prepublish): 1-1.

[9] Gabriel Hartmann and Zvi Shiller and Amos Azaria. Model-based Reinforcement Learning for Time-optimal Velocity Control[J]. IEEE Robotics and Automation Letters, 2020, PP(99): 1-1.

[10] 左思翔. 基于深度强化学习的无人驾驶智能决策控制研究[D]. 哈尔滨工业大学, 2018.
Zuo Sixiang. Research on Intelligent Decision Control of Unmanned Vehicle Based on Deep Reinforcement Learning [D]. Harbin Institute of Technology, 2018.

[11] 林晓波. 小型无人机的强化学习控制[D]. 北京科技大学, 2020.
Lin Xiaobo. Reinforcement Learning Control of Small Uav [D]. University of Science and Technology Beijing, 2020.

[12] 卜令正. 基于深度强化学习的机械臂控制研究[D]. 中国矿业大学, 2019.
Bu Lingzheng. Research on Manipulator Control Based on Deep Reinforcement Learning [D]. China University of Mining & Technology, 2019.

[13] 林伯钧. 基于深度学习算法的实时手机数据分类及其对智慧城市建设的影响研究[J]. 科技创新导报, 2019, 16(28): 240+242.
Lin Bojun. Research on real-time mobile phone data classification based on deep learning algorithm and its impact on smart city construction [J]. Science and Technology Innovation Review, 2019, 16(28): 240+242.

[14] Morcos Amir and West Aaron and Maguire Brian. Multi-Agent Reinforcement Learning for Convex Optimization[J]. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR MULTI-DOMAIN OPERATIONS APPLICATIONS III, 2021, 11746.

[15] Transportation - Self-Driving Cars; Findings in the Area of Self-Driving Cars Reported from Chongqing University (Decision-making Strategy On Highway for Autonomous Vehicles Using Deep Reinforcement Learning)[J]. Journal of Transportation, 2020, PP: 62-.

[16] 张茂盛, 段杰, 肖息, 陈善洛, 欧阳权, 王志胜. 基于深度强化学习-PI 控制的机电作动器控制策略[J]. 应用科技, 2022, 49(4): 18-22.
Zhang Maosheng, Duan Jie, Xiao Xi, Chen Shanluo, Ouyang Quan, Wang Zhisheng. Control strategy of Electromechanical Actuators based on Deep Reinforcement Learn-PI control [J]. Applied Science and Technology, 2022, 49(4): 18-22.

[17] 甄岩, 郝明瑞. 基于深度强化学习的智能 PID 控制方法研究[J]. 战术导弹技术, 2019(5): 37-43.
Zhen Yan, Hao Mingrui. Research on Intelligent PID Control Method Based on Deep Reinforcement Learning [J]. Tactical Missile Technology, 2019(5): 37-43.

[18] 邓海波. 基于深度强化学习的时序差分优化算法研究[D]. 西南大学, 2021.
Deng Haibo. Research on Time-series Differential Optimization Algorithm Based on Deep Reinforcement Learning [D]. Southwest University, 2021.

[19] 尹舸帆. 深度强化学习中探索问题的研究和实现[D].北京邮电大学, 2021.
Yin Gefan. Research and Implementation of exploration problem in Deep Reinforcement Learning [D]. Beijing University of Posts and Telecommunications, 2021.

[20] 严家政, 专祥涛. 基于强化学习的参数自整定及优化算法[J]. 智能系统学报, 2022, 17(2): 341-347.
Yan Jiazheng, Zhuan Xiangtao. Parameter Self-tuning and Optimization Algorithm Based on Reinforcement Learning [J]. Journal of Intelligent Systems, 2022, 17(2): 341-347.