DOI:10.11918/202310026

依托平滑强化学习的铰接车轨迹跟踪方法

陈良发[1],宋绪杰[2],肖礼明[1],高路路[1],张发旺[3],李升波[2],马 飞[1],段京良[1]

(1.北京科技大学 机械工程学院,北京 100083; 2. 清华大学 车辆与运载学院,北京 100084;
3.北京理工大学 机械与车辆学院,北京 100081)

摘 要:为解决现有铰接车轨迹跟踪控制面临的动作波动问题,提高铰接车轨迹跟踪控制的精度以及平滑性,提出了一种考虑轨迹预瞄的平滑强化学习型跟踪控制方法。首先,为保证控制精度,通过将参考轨迹信息作为预瞄信息引入强化学习策略网络和值网络,构建了预瞄型强化学习迭代框架。然后,为保证控制平滑性,引入 LipsNet 网络结构近似策略函数,从而实现策略网络 Lipschitz 常数的自适应限制。最后,结合值分布强化学习理论,建立了最终的平滑强化学习型轨迹跟踪控制方法,实现了铰接车轨迹跟踪的控制精度和控制平滑性的协同优化。仿真结果表明,本研究提出的平滑强化学习跟踪控制方法(SDSAC)在6种不同噪声等级下均能保持良好的动作平滑能力,且具备较高跟踪精度;与传统值分布强化学习(DSAC)相比,在高噪声条件下,SDSAC 动作平滑度提升超过5.8倍。此外,与模型预测控制相比,SDSAC 的平均单步求解速度提升约60倍,具有较高的在线计算效率。

关键词:自动驾驶;铰接车;轨迹跟踪;强化学习;动作平滑

中图分类号:TP273+.1 **文献标志码**:A **文章编号**:0367-6234(2024)12-0116-08

Smooth reinforcement learning-based trajectory tracking for articulated vehicles

CHEN Liangfa[1], SONG Xujie[2], XIAO Liming[1], GAO Lulu[1], ZHANG Fawang[3],
LI Shengbo[2], MA Fei[1], DUAN Jingliang[1]

(1. School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China;
2. School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China; 3. School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China)

Abstract: This research tackles the challenge of action fluctuation in articulated vehicle trajectory tracking control, aiming to enhance both accuracy and smoothness. It introduces a novel approach: a smooth tracking control methodology grounded in reinforcement learning (RL). Firstly, to improve the control accuracy, we incorporate trajectory preview information as input to both the policy and value networks and establish a predictive policy iteration framework. Then, to ensure control smoothness, we employ the LipsNet network to approximate the policy function, to realize the adaptive restriction of the Lipschitz constant of the policy network. Finally, coupled with distributional RL theory, we formulate an articulated vehicle trajectory tracking control method, named smooth distributional soft actor-critic (SDSAC), focusing on achieving synergistic optimization of both control precision and action smoothness. The simulation results demonstrate that the proposed method can maintain good action smoothing ability under six different noise levels, and has strong noise robustness and high tracking accuracy. Compared with traditional value distribution reinforcement learning distributional soft actor-critic (DSAC), SDSAC improves action smoothness by more than 5.8 times under high noise conditions. In addition, compared with model predictive control, SDSAC's average single-step solution speed is improved by about 60 times, and it has higher online computing efficiency.

Keywords: automatic drive; articulated vehicle; trajectory tracking; reinforcement learning; action smoothing

铰接式车辆(铰接车)具有转弯半径小、通过性强以及使用成本低的优点,在矿山、山地等复杂地形环境中有着广泛的应用。然而,特殊的转向形式使得铰接车的运动控制相较于一般刚体车辆更为复杂,自动驾驶实现难度更大。

轨迹跟踪控制作为铰接车自动驾驶的关键技术之一,近年来得到了国内外学者的广泛研究[1]。现有的铰接车轨迹跟踪控制方法有 PID (proportional

收稿日期:2023-10-14;**录用日期**:2023-12-15;**网络首发日期**:2024-10-08
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20240930.1739.005
基金项目:国家自然科学基金(52202487);汽车安全与节能国家重点实验室开放基金(KF2212)
作者简介:陈良发(1999—),男,硕士研究生;李升波(1982—),男,长聘教授,博士生导师;马 飞(1968—),男,教授,博士生导师
通信作者:段京良,duanjl@ustb.edu.cn

integral derivative)[2]、线性二次调节控制(linear quadratic regulator, LQR)[3-4]、滑模控制[5]以及模型预测控制(model predictive control, MPC)[6]等。其中,PID 虽然结构简单,使用方便,但是无法处理复杂的系统约束和实现对车辆的横、纵向联合控制。而 LQR 应用于非线性系统时,由于需要对系统进行线性化处理,因而在实际应用中难以实现对参考轨迹的准确跟踪。MPC 作为一种解决有限时域优化控制问题的常用方法,具有状态约束处理、滚动时域优化的优势,且具备理论最优性的保障[7-9]。但是在实际应用中,MPC 控制器需在每个控制周期内在线迭代求解最优控制动作序列,在系统非线性强或车载计算资源受限时,其在线求解速度通常难以满足控制实时性要求[10]。

为提高控制量的在线求解效率,近年来一些依托强化学习的高实时性离线求解在线应用的控制模式得到了广泛的研究和应用[11-17],典型的算法有 RMPC (recurrent MPC)[18]、DSAC (distributional soft actor-critic)[19-20]等。然而,强化学习在实际应用中面临着动作波动难题,轻微的状态差异会引起动作的大幅变化,而不平滑的控制动作会加快机械部件的磨损并影响实际驾乘体验[21]。Mysore 等[22]发现,直接对策略网络输出动作进行滤波会改变系统的动态响应且违背马尔可夫假设,从而严重损害控制性能。为兼顾控制的最优性和平滑性,需在策略网络训练时对动作的波动进行限制。Kobayashi 等[23]将动作平滑性惩罚项添加到策略网络的损失函数中,实现了对动作波动的抑制,然而该方法性能对参数的调整较为敏感。Takase 等[24]通过在策略网络层级上应用谱归一化技术,实现了控制抖动的抑制,但是谱归一化技术会带来严重的性能损失。Song 等[25]提出了一种名为 LipsNet 的神经网络架构,可以根据状态自动调整策略网络的 Lipschitz 常数,实现动作波动的动态抑制。由于该方法不需要对算法结构进行修改,大大降低了其应用的难度。

综上所述,本文针对铰接车轨迹跟踪控制问题,提出了一种依托平滑强化学习的预瞄型铰接车轨迹跟踪算法。该算法所求得的跟踪策略不仅考虑了轨迹预瞄信息,而且具备较好的控制平滑性和高在线控制实时性。

1　问题描述

1.1　铰接车运动学模型

铰接车运动学模型的构建是轨迹跟踪优化求解的基础。如图 1 所示,设铰接车前、后车体重心位于前、后桥中心点上,坐标分别为 $p_f=(x_f,y_f)$, $p_r=(x_r,y_r)$,前、后车体铰接点到前、后轴的距离分别为 l_f 和 l_r。使用铰接车前轴中心点 p_f 作为参考点,根据车辆的几何特征,可得

$$\begin{cases}\dot{x}_f = v_f\cos\varphi_f\\ \dot{y}_f = v_f\sin\varphi_f\end{cases} \tag{1}$$

式中: $\dot{x}_f$ 为车辆前车体沿 x 轴方向的速度, m/s; $\dot{y}_f$ 车辆前车体沿 y 轴方向的速度,m/s; v_f 为前车体参考速度,m/s; φ_f 为航向角。

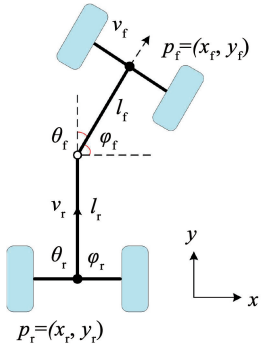


图 1　铰接车几何结构示意

Fig. 1　Geometric structure diagram of articulated vehicle

由于前、后车体通过铰链机构连接,因此前、后车体速度 v_f 和 v_r 均满足:

$$\begin{cases}v_f = v_r\cos\theta_f + \dot{\varphi}_r l_r\sin\theta_f\\ v_r\sin\theta_f = \dot{\varphi}_f l_f + \dot{\varphi}_r l_r\cos\theta_f\\ \theta_f = \varphi_f - \varphi_r\end{cases} \tag{2}$$

式中: $\dot{\varphi}_f$、$\dot{\varphi}_r$ 分别为前、后车体的航向角变化率, rad/s。将铰接车前车体中心横坐标 x_f、纵坐标 y_f、航向角 φ_f、车速 v_f 和铰接角 θ_f 作为铰接车运动学模型状态向量 $\boldsymbol{X}$,前车体加速度 a 和铰接角角速度 ω 作为输入向量 $\boldsymbol{u}$,即

$$\begin{cases}\boldsymbol{X} = [x_f, y_f, \varphi_f, v_f, \theta_f]^{\mathrm{T}}\\ \boldsymbol{u} = [a, \omega]^{\mathrm{T}}\end{cases} \tag{3}$$

式中: $\dot{v}_f = a$, $\dot{\theta}_f = \omega$。

进而,铰接车运动学模型可表示为

$$\dot{\boldsymbol{X}} = \tilde{\boldsymbol{A}} + \tilde{\boldsymbol{B}}\boldsymbol{u} \tag{4}$$

其中:

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} v_f\cos\varphi_f \\ v_f\sin\varphi_f \\ \dfrac{v_f\sin\theta_f}{l_f\cos\theta_f + l_r} \\ 0 \\ 0 \end{bmatrix}, \tilde{\boldsymbol{B}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \dfrac{l_r}{l_f\cos\theta_f + l_r} \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

利用前向欧拉方法将连续时间运动学模型进行离散化,可得

$$X_{t+1} = \tilde{\boldsymbol{A}}/f + \tilde{\boldsymbol{B}}u_t/f + X_t \tag{5}$$

式中 f 为控制频率。

1.2 轨迹跟踪任务描述

1.2.1 强化学习问题描述

强化学习本质在于智能体通过与环境的不断交互,自主学习到一个使得未来累计损失最小化的控制策略。在强化学习过程中,智能体在 t 时刻观测得到环境状态 X_t,通过采取动作 u_t 与环境发生交互,环境的状态转移为 X_{t+1},智能体同时获得一个损失信号 l_t,损失信号可用于评价智能体在状态 X_t 采取动作 u_t 的好坏。

将强化学习应用于铰接车轨迹跟踪控制任务时,智能体的控制目标在于找到一个最优控制策略 π^*,在满足 $u_t=\pi^*(X_t)$ 的条件下使得轨迹跟踪的期望累计损失最小,即

$$\pi^* = \min_{\pi} V_\pi(x) \tag{6}$$

其中

$$V_\pi(x) = E_\pi[G_t \mid X_t = x] = E_\pi\left[\sum_{k=0}^{\infty}\gamma^k l_{t+k} \mid X_t = x\right]$$

式中:状态价值函数 $V_\pi(x)$ 为从状态 x 开始执行策略 π 得到的累计期望损失,γ 为折扣因子。

1.2.2 损失函数设计

在铰接车轨迹跟踪控制任务中,损失函数由跟踪误差损失和动作正则损失构成。设损失函数 l 为状态跟踪误差和动作正则的二次型加权求和:

$$l(X_t, \boldsymbol{X}_t^{\mathrm{ref}}, u_t) = \|X_t - \boldsymbol{X}_t^{\mathrm{ref}}\|_{\boldsymbol{Q}} + \|u_t\|_{\boldsymbol{R}} \tag{7}$$

式中:$\boldsymbol{Q}$、$\boldsymbol{R}$ 分别为状态惩罚矩阵和动作惩罚矩阵,数学形式上均为正定对角矩阵。其中,$\boldsymbol{X}_t^{\mathrm{ref}} = [x_t^{\mathrm{ref}}, y_t^{\mathrm{ref}}, \varphi_t^{\mathrm{ref}}, v_t^{\mathrm{ref}}, \theta_t^{\mathrm{ref}}]^{\mathrm{T}}$ 为参考状态向量。

2 平滑强化学习轨迹跟踪算法

2.1 预瞄型强化学习迭代框架

为减少铰接车的轨迹跟踪误差,本文拟将轨迹预瞄点作为前馈信息,从而提高跟踪精度。然而,由于车辆行驶时实际速度会受到道路曲率的影响,因而预瞄点的位置关系通常难以给定。为此,如图 2 所示,本文在假设已知期望速度 v^{ref} 以及参考轨迹曲线的情况下,通过利用期望速度对时间的积分获取预测时域内各预瞄点的横坐标,得到预瞄点对应的状态参考向量,并将预测时域内各预瞄点信息作为策略输入,即

$$u_t = \pi(X_t, X_{t:t+N-1}^{\mathrm{ref}}) \tag{8}$$

式中:$X_{t:t+N-1}^{\mathrm{ref}}$ 为 N 步预瞄信息,即 $X_{t:t+N-1}^{\mathrm{ref}} = [X_t^{\mathrm{ref}}, X_{t+1}^{\mathrm{ref}}, \cdots, X_{t+N-1}^{\mathrm{ref}}]$。参考轨迹上横坐标可由下式求得:

$$x_j^{\mathrm{ref}} = \int_{t+\frac{j}{\mathrm{f}}}^{t+\frac{j+1}{\mathrm{f}}} v^{\mathrm{ref}}(j+1)\,\mathrm{d}t \tag{9}$$

式中:$j\in 0,\cdots,N-1$,$j=0$ 即第 1 个参考点;N 为预测时域。

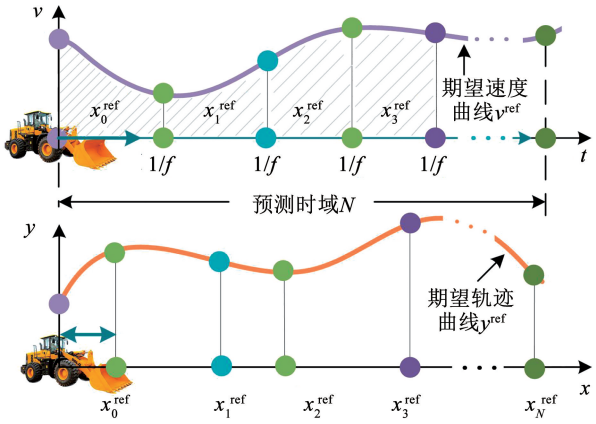


图 2 预瞄点获取示意

Fig. 2 Schematic diagram of obtaining preview points

为求解最优控制策略 π^*,本文将铰接车轨迹跟踪控制构建为无穷时域最优控制问题。为此,定义 Q 为从状态轨迹对 $(X_t, X_{t:t+N-1}^{\mathrm{ref}})$ 出发以 u_t 为初始动作,执行策略 π 到无穷时刻得到的累计期望损失:

$$Q_\pi(X_t, X_{t:t+N-1}^{\mathrm{ref}}, u_t) \triangleq E_\pi\left\{\sum_{i=0}^{\infty}\gamma^i l(X_{t+i}, X_{t+i:t+N+i-1}^{\mathrm{ref}}, u_{t+i})\right\} \tag{10}$$

由式(10)可知,Q_π 与铰接车状态量 X_t、轨迹预瞄信息 $X_{t:t+N-1}^{\mathrm{ref}}$ 和控制动作 u_t 有关。式(10)可进一步展开为

$$Q_\pi(X_t, X_{t:t+N-1}^{\mathrm{ref}}, u_t) = l(X_t, X_t^{\mathrm{ref}}, u_t) + E_\pi\{\gamma Q_\pi(X_{t+1}, X_{t+1:t+N}^{\mathrm{ref}}, u_{t+1})\} \tag{11}$$

式(11)表明铰接车轨迹跟踪控制问题可以利用强化学习进行求解。即利用策略迭代框架对 Q 函数以及策略 π 进行交替优化求解,使其逐步迭代收敛至最优策略 π^*。其中,策略迭代分为策略评估和策略改进两个环节。基于式(11)策略评估得到的 Q_π,利用下式可求得改进的策略 π_{k+1},即

$$\pi_{k+1}(X_t, X_{t:t+N-1}^{\mathrm{ref}}) = \underset{u_t}{\operatorname{argmin}}[Q_{\pi_k}(X_t, X_{t:t+N-1}^{\mathrm{ref}}, u_t)] \tag{12}$$

2.2 平滑策略网络

为保障策略的平滑性,本文采用 LipsNet 网络近似策略函数。如图 3 所示,通过将 Lipschitz 常数 k 作为可学习的参数,实现策略网络 Lipschitz 常数的自动调整,从而对策略网络的输出的波动抑制。其中,平滑策略网络可表示为

$$\pi_{\mathrm{s}}(X_t, X_{\mathrm{t:t+N-1}}^{\mathrm{ref}}; \omega, \phi) = k(X_t, X_{t:t+N-1}^{\mathrm{ref}}; \phi)\frac{\pi(X_t, X_{t:t+N-1}^{\mathrm{ref}}; \omega)}{\|\nabla_{X_t, X_{t:t+N-1}^{\mathrm{ref}}}\pi(X_t, X_{t:t+N-1}^{\mathrm{ref}}; \omega)\| + \varepsilon} \tag{13}$$

式中:$\pi_{\mathrm{s}}(X_t, X_{t:t+N-1}^{\mathrm{ref}}; \omega, \phi)$ 为平滑策略网络,输出

数度与控制动作维数相同; $k(X_t, X^{\mathrm{ref}}_{t:t+N-1};\phi)$ 为单输出 Lipschitz 乘子网络; ϕ 为乘子网络参数,该网络可根据自车状态以及参考轨迹信息自动调整输出的 Lipschitz 常数大小,实现控制动作波动的动态抑制; $\pi(X_t, X^{\mathrm{ref}}_{t:t+N-1};\omega)$ 为原始策略网络,网络参数为 ω,输出维数与控制动作维数相同; $\|\cdot\|$ 为矩阵2范数; $\nabla_{X_t, X^{\mathrm{ref}}_{t:t+N-1}}\pi(X_t, X^{\mathrm{ref}}_{t:t+N-1};\omega)$ 为 Jacobian 矩阵; ε 为正数小量。

本文采用动作波动率定量表征控制动作的波动情况,定义为

$$\xi(\pi_s) \triangleq E_{\tau\sim\rho_{\pi_s}}\left[\frac{1}{T}\sum_{t=1}^{T}\|u_t - u_{t-1}\|\right] \tag{14}$$

式中: ρ_{π_s} 为策略 π_s 产生的动作 – 状态分布, T 为终止时间, $\|\cdot\|$ 为动作向量的2范数。 $\xi(\pi_s)$ 越小,表示策略 π_s 输出的动作越平滑。

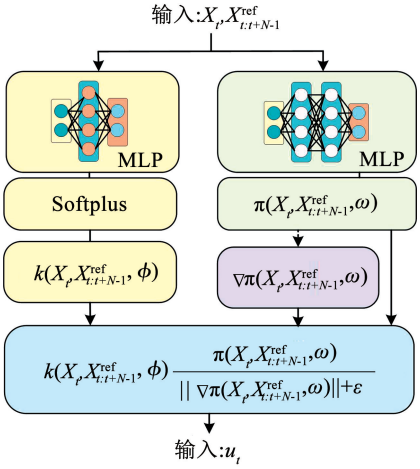


图 3　LipsNet 网络结构

Fig. 3　LipsNet network structure

2.3　考虑预瞄信息的平滑 DSAC 算法

依托式(11)的预瞄型自洽条件和平滑策略网络结构,本文提出了面向铰接车轨迹跟踪控制的平滑 DSAC (smooth distributional soft actor-critic, SDSAC)算法。与传统算法不同,平滑 DSAC 的值函数对应的是随机累计损失的分布而非单纯的期望值,也被称为值分布函数。策略 π_s 产生的随机累计损失定义为

$$Z_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t) = l(X_t, X^{\mathrm{ref}}_t, u_t) + \gamma G_{t+1} \tag{15}$$

式中, $G_t = \sum_{k=0}^{\infty}\gamma^k l_{t+k}$。

定义随机累计损失 $Z_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t)$ 的概率密度为分布函数 $\boldsymbol{Z}_{\pi_s}(\cdot\mid X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t)$, $\boldsymbol{Z}_{\pi_s}(\cdot\mid X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t)$ 表示给定 $(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t)$ 时 $Z_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t)$ 的概率密度,即 $Z_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t) \sim \boldsymbol{Z}_{\pi_s}(\cdot\mid X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t)$。根据随机累计损失的定义,其对应的值分布自洽条件为

$$\Gamma_{\pi_s} Z_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t) = l(X_t, X^{\mathrm{ref}}_t, u_t) + \gamma(Z_{\pi_s}(X_{t+1}, X^{\mathrm{ref}}_{t:t+N}, u_{t+1}) + \alpha\log\pi_s(u_t\mid X_t, X^{\mathrm{ref}}_{t:t+N-1})) \tag{16}$$

式中 Γ_{π_s} 为值分布自洽算子。

平滑 DSAC 算法采用 Actor-Critic 结构学习独立的值分布网络以及随机策略网络。其中,策略网络的输入为铰接车的状态量和参考轨迹信息,输出为该状态量下对应动作的均值 u 和标准差 σ。利用平滑 DSAC 算法求解铰接车轨迹跟踪控制问题时,策略评估通过最小化目标损失分布与当前损失分布之间的差异来实现,具体目标函数为

$$J_Z(\theta) = \mathrm{E}[D_{\mathrm{KL}}(\Gamma_{\pi_s}\boldsymbol{Z}(\cdot\mid X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t;\theta), \boldsymbol{Z}(\cdot\mid X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t;\theta))] \tag{17}$$

式中: $J_Z(\theta)$ 为值分布网络更新目标, θ 为值网络参数, D_{KL} 为 Kullback-Leibler(KL)散度,用于度量两分布之间的距离; α 为策略熵系数,其更新规则为

$$\alpha = \alpha - \beta_\alpha(\mathrm{E}[-\log\pi_s(u_t\mid X_t, X^{\mathrm{ref}}_{t:t+N-1};\omega,\phi)] - \overline{H}) \tag{18}$$

式中: β_α 为学习率, $\overline{H}$ 为策略熵目标值。

平滑策略网络通过最小化期望累计损失和 Lipschitz 常数 k 正则的加权进行优化,其目标函数为

$$J_{\pi_s}(\omega,\phi) = \lambda\|k(X_t, X^{\mathrm{ref}}_{t:t+N-1};\phi)\|^2 + \mathrm{E}[Q_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t;\theta) - \alpha\log\pi_s(u_t\mid X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t;\omega,\phi)] \tag{19}$$

式中 λ 为平滑项权重。其中 Q_{π_s} 可由值分布网络可得

$$Q_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t;\theta) = \mathrm{E}[Z_{\pi_s}(X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t)] \tag{20}$$

算法具体更新过程如下:

1)给定自车初始状态 X_t、参考轨迹 $y^{\mathrm{ref}}(x)$、期望速度 v^{ref},利用期望速度对时间的积分获取预测时域内大地坐标系下的 N 个参考轨迹点 $X^{\mathrm{ref}}_{t:t+N-1}$,见图 4。

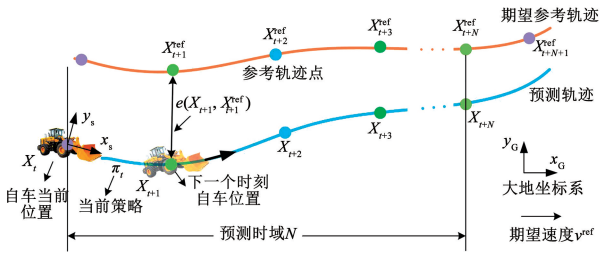


图 4　铰接车轨迹跟踪示意

Fig. 4　Trajectory tracking diagram of articulated vehicle

2)在当前状态 X_t 下使用策略 π_s 与环境交互采样,得到损失 l_t 以及观测下一时刻状态 X_{t+1},同时获取新的 N 个参考点 $X^{\mathrm{ref}}_{t+1:t+N}$,将 $\{X_t, X^{\mathrm{ref}}_{t:t+N-1}, u_t, l_t, X_{t+1}, X^{\mathrm{ref}}_{t+1:t+N}\}$ 组成一个经验样本,并存入经验池 B。

3)从经验回放池 B 中采样得到的多个经验样本作为前向求解过程的初值,利用式(17)实施梯度下降,其更新规则为

$$\theta_{k+1} = -\beta_Z \nabla_\theta J_Z(\theta_k) + \theta_k \tag{21}$$

式中 β_Z 为值分布函数学习率。

4)值网络进行若干次更新后,依托式(19)利用梯度下降分别对策略网络和 k 网络参数进行更新,更新规则为:

$$\omega_{k+1} = -\beta_\pi \nabla_\omega J_{\pi_s}(\omega_k, \phi_k) + \omega_k \tag{22}$$

$$\phi_{k+1} = -\beta_k \nabla_\phi J_{\pi_s}(\omega_k, \phi_k) + \phi_k \tag{23}$$

平滑强化学习算法更新伪代码如下。

平滑强化学习算法:

初始化值分布网络参数 θ、策略网络参数 ω,ϕ、策略熵系数 α

设置学习率 β_Z、β_π、β_k、β_α

初始化迭代步数 $k=0$

给定期望速度、参考轨迹、自车初始状态 X_0 和预测时域 N

Repeat

获取预测时域内各预瞄点参考信息

根据策略选择动作 $u_t \sim \pi_s(\cdot | X_t, X^{\text{ref}}_{t:t+N-1}; \omega, \phi)$

与环境交互得到 $X_{t+1}, X^{\text{ref}}_{t+1:t+N}$ 以及损失信号 l_t

将样本 $\{X_t, X^{\text{ref}}_{t:t+N-1}, u_t, l_t, X_{t+1}, X^{\text{ref}}_{t+1:t+N}\}$ 存入经验池 B

从 B 中随机选择批量样本 $\{X_t, X^{\text{ref}}_{t:t+N-1}, u_t, l_t, X_{t+1}, X^{\text{ref}}_{t+1:t+N}\}$

基于式(21)计算值分布网络梯度并更新参数 θ

if k 能被整数 m 整除

基于式(22)更新策略网络参数 ω

基于式(23)更新 Lipschitz 常数网络参数 ϕ

基于式(18)更新策略熵系数 α

end if

$k = k+1$

until 收敛

平滑强化学习的跟踪控制算法框图见图 5。

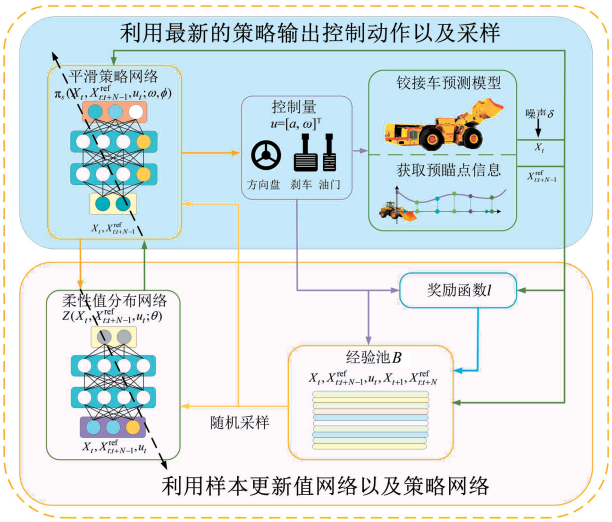


图 5 依托平滑强化学习的跟踪控制算法框架

Fig. 5 Tracking control framework based on smooth reinforcement learning

3 仿真分析

本文通过仿真实验验证所提出平滑 DSAC 算法的轨迹跟踪性能。实验首先在 PC 机上依托 GOPS (general optimal control problem solver)平台[26]对平滑策略网络进行离线预训练,然后利用训练得到的平滑策略网络根据自车状态及预瞄信息直接输出控制信号至仿真模型,模型执行相应控制动作后将所得新的自车状态信息以及预瞄信息反馈至平滑策略网络,实现闭环控制过程。

仿真实验中,铰接车期望行驶速度设为 5 m/s 参考轨迹选用角度为 60°的三角波曲线来模拟铰接车巷道内的转弯工况。仿真时各算法的预测时域均相同,仿真时间为 25 s。实验平台基于 Windows 操作系统,搭载 3.6 GHz、12 核心 20 线程的英特尔 i7 处理器。

3.1 算法参数设计

值网络以及 π 网络均采用双隐层结构,单层 256 个神经元,激活函数为 Relu。k 网络采用单隐层结构,激活函数为 Relu。各网络均通过 Adam 方法更新参数。铰接车轨迹跟踪任务关键参数和算法的超参数分别见表 1、2。

表 1 铰接车轨迹跟踪任务关键参数

Tab. 1 Key parameters of trajectory tracking task of articulated vehicle

前车体长度 l_f/m	后车体长度 l_r/m	预测时域 N
2	2	30
状态权重系数 Q	动作权重系数 R	控制频率 f/Hz
[0.5,0.5,0.1,0.1,0.1]	[0.1,0.1]	10

表 2 算法超参数

Tab. 2 Algorithm hyperparameters

值网络学习率 β_Z	π 网络学习率 β_π	策略熵学习率 β_α	k 网络学习率 β_k
1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-4}
平衡权重	k 初始值	目标策略熵	正数小量
0.2	5	−2	1×10^{-4}

为检验算法的控制平滑性和轨迹跟踪性能,本文选用 DSAC、TD3[27]、SAC[28]和 MPC 等主流算法与本文提出的平滑 DSAC 在相同工况下进行对比分析。此外,为模拟实际工况下传感器的测量误差对控制动作的影响,实验选用 6 组满足高斯分布的随机噪声进行测试,设为 6 个(0 ~ 5)噪声等级。其中每个等级的高斯分布的均值 μ_n 均设为 0,标准差 σ_n

大小见表 3。由采样得到的所有噪声均直接添加到原有观测量上,同时为简化问题,假设噪声只影响自车状态的观测值,预瞄的观测量不受影响。此外,为降低随机因素的影响,每组实验均进行 100 次随机初始点的测试,初始点各状态范围见表 4。

表 3　噪声等级与各观测噪声标准差

Tab. 3　Noise level and its corresponding standard deviation

噪声等级	x_f/m	y_f/m	φ_f/(°)	v_f/(m·s^{-1})	θ_f/(°)
0	0	0	0	0	0
1	0.05	0.05	2	0.05	1
2	0.10	0.10	3	0.10	2
3	0.15	0.15	4	0.15	3
4	0.20	0.20	5	0.20	4
5	0.25	0.25	6	0.25	5

表 4　初始点随机范围

Tab. 4　Random range of initial points

状态量	x_0/m	y_0/m	φ_0/(°)	v_0/(m·s^{-1})	θ_0/(°)
范围	±0.5	±0.5	±10	±0.5	±5

3.2　结果分析

图 6、表 5 分别展示了不同等级噪声影响下,平滑 DSAC(SDSAC)算法与传统算法在跟踪过程中的动作波动率以及横向位置误差、速度误差的变化情况。其中,图 6 为 100 个随机初始点对应的控制仿真过程的统计均值,误差棒为 95% 的置信区间。可以看出,在相同噪声等级下,TD3、DDPG 算法相较于 SAC、DSAC 算法,虽然拥有较低的控制波动率,但位置和速度的跟踪误差均出现了明显增加。而采用了策略熵技术的 SAC、DSAC 算法,所学策略能更好跟踪预期轨迹,跟踪精度相较 TD3、DDPG 算法提升近 1 倍,但同时也伴随着较高的控制波动率。相比之下,本文提出的平滑 DSAC 算法在不同噪声等级下均表现出了良好的波动抑制能力,且随着噪声等级的增大,平滑 DSAC 控制波动率增长显著低于传统强化学习算法和 MPC 算法。在 1 ~5 组不同等级的噪声影响下,平滑 DSAC 的动作平滑度相较于 DSAC 算法分别实现了 2.7、3.6、4.7、5.2、5.8 倍的性能提升,而仅损失了 3 cm 左右的位置跟踪误差。上述结果表明,平滑 DSAC 在噪声影响下,仍然具备较高的轨迹跟踪精度及动作平滑能力。

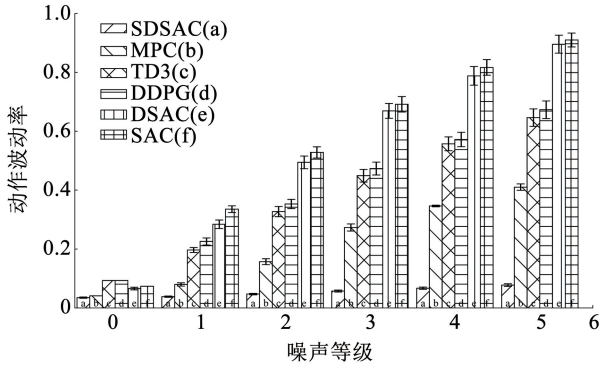


图 6　不同算法的动作波动率

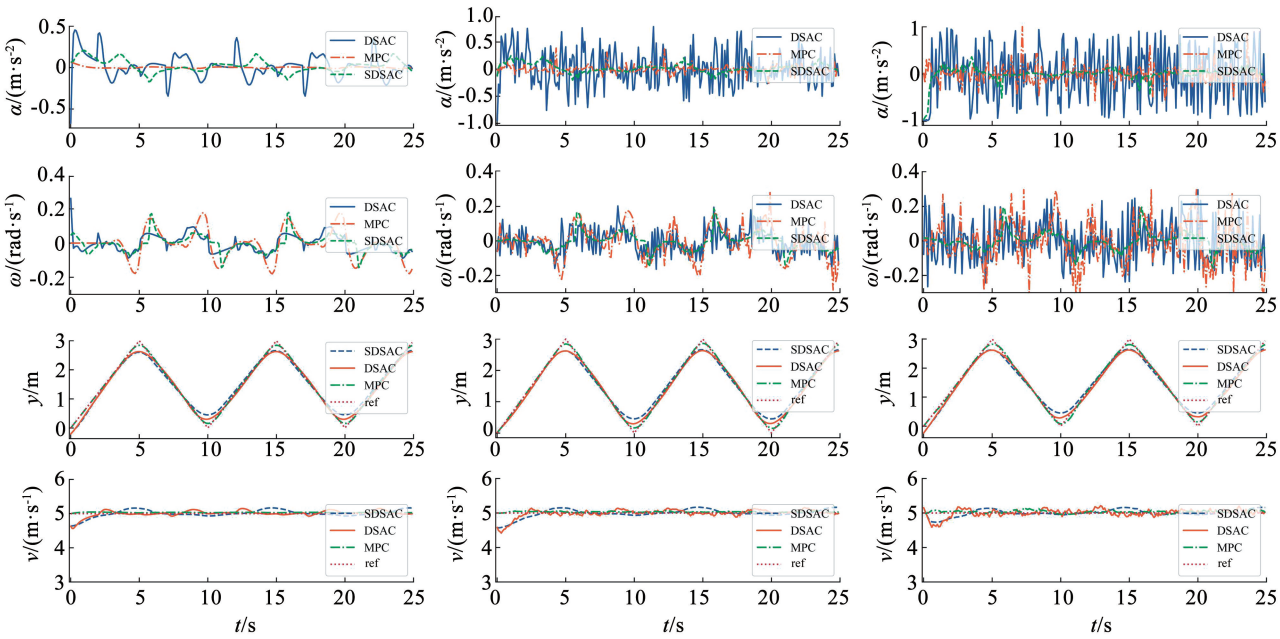
Fig. 6　Action fluctuation of different algorithms

图 7 展示了 3 组不同噪声等级影响下,铰接车跟踪三角波曲线的控制动作、前车体横向位置以及速度曲线的变化情况。受噪声影响,3 种算法的控制动作均出现了一定程度的波动,但相较于 DSAC、MPC,平滑 DSAC 的控制动作曲线更为平滑。进一步说明本文提出的平滑 DSAC 算法具有较强协同优化能力,即在噪声影响下,依然可以有效降低轨迹跟踪过程中的动作波动,同时保证轨迹跟踪的性能,可以实现控制精度和动作平滑性的协同优化。

表 5　不同算法的跟踪误差对比

Tab. 5　Tracking error comparison of different algorithms

噪声等级		0		1		2		3		4		5	
		Δy/m	Δv/(m·s^{-1})	Δy/m	Δv/(m·s^{-1})	Δy/m	Δv/(m·s^{-1})	Δy/m	Δv/(m·s^{-1})	Δy/m	Δv/(m·s^{-1})	Δy/m	Δv/(m·s^{-1})
算法名称	SDSAC	0.088	0.087	0.086	0.089	0.085	0.085	0.090	0.090	0.087	0.087	0.082	0.082
	DSAC	0.069	0.048	0.068	0.044	0.068	0.050	0.070	0.050	0.067	0.058	0.065	0.059
	SAC	0.082	0.051	0.083	0.051	0.084	0.056	0.085	0.058	0.186	0.060	0.088	0.061
	TD3	0.182	0.072	0.182	0.072	0.183	0.073	0.184	0.074	0.186	0.074	0.190	0.075
	DDPG	0.191	0.072	0.186	0.075	0.194	0.076	0.189	0.078	0.188	0.078	0.191	0.080
	MPC	0.021	0.031	0.022	0.032	0.024	0.033	0.026	0.035	0.026	0.035	0.028	0.036



(a) 噪声等级=0 (b) 噪声等级=2 (c) 噪声等级=5

图 7 不同噪声下跟踪三角波的动作及状态变化曲线

Fig. 7 Action and state curves of tracking triangular wave under different noises

最后,本文基于搭载英特尔 12 700 KF 处理器的 Windows 实验平台,对上述实验中噪声等级为 0 场景下的 MPC 及 SDSAC 算法的平均单步求解时间进行了统计。实验在相同工况下进行,结果见表 6。当预测时域在 30 步时,MPC(依托 IPOPT 优化器[29])的平均单步求解时间约为 33.0 ms,平滑 DSAC 的平均单步求解时间均约为 0.49 ms,在线求解速度提升超过 60 倍,且求解用时更为稳定,而 MPC 受在线优化问题复杂度和计算资源分配的影响,求解时间出现了较大程度的波动。

表 6 平均单步求解时间对比

Tab. 6 Comparison of average single-step solution time ms

算法名称	均值	标准差
SDSAC	0.49	0.143
MPC	33.00	83.749

综上所述,验证结果表明本文提出的平滑 DSAC 算法具有较高的跟踪精度和在线计算效率,且在不同噪声环境下均能保持良好的控制平滑性。

4 结 论

1)本文针对铰接车轨迹跟踪问题,通过将参考轨迹信息作为预瞄信息引入策略网络和值网络,构建了预瞄型强化学习迭代框架,并结合 LipsNet 网络提出了平滑 DSAC 算法。高噪声环境下,铰接车轨迹跟踪横向误差小于 9 cm。

2)从控制平滑度上看,平滑 DSAC 在不同噪声等级下均能保持良好的动作平滑能力,具有较强的噪声鲁棒性以及较高的跟踪精度。与传统 DSAC 相比,高噪声条件下平滑 DSAC 动作平滑度提升超过 5.8 倍,实现了跟踪控制精度和动作平滑性的协同优化。

3)从控制实时性上看,平滑 DSAC 平均单步求解速度相较于 MPC 提升约 60 倍,具有较高的在线计算效率。且平滑 DSAC 求解用时更为稳定,而 MPC 受在线优化问题复杂度和计算资源分配的影响,求解时间出现了较大程度的波动。

4)然而,平滑 DSAC 在提高控制平滑性的同时,引入了额外的超参数 λ(式(19)),以对控制平滑项和性能项进行平衡。

参考文献

[1]于向军, 槐元辉, 姚宗伟, 等. 工程车辆无人驾驶关键技术[J]. 吉林大学学报(工学版), 2021, 51(4): 1153
YU Xiangjun, HUAI Yuanhui, YAO Zongwei, et al. Key technologies in autonomousvehicle for engineering[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(4): 1153. DOI: 10.13229/j.cnki.jdxbgxb20210038

[2]TAN Senqi, ZHAO Xinxin, YANG Jue, et al. A path tracking algorithm for articulated vehicle: development and simulations[C]// 2017 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific). Harbin: IEEE, 2017: 1. DOI: 10.1109/ITEC-AP.2017.8080797

[3]MENG Yu, GAN Xin, WANG Yu, et al. LQR-GA controller for articulated dump truck path tracking system[J]. Journal of Shanghai Jiaotong University (Science), 2019, 24(1): 78. DOI:10.1007/s12204-018-2012-z

[4]YAKUB F, MORI Y. Comparative study of autonomous path-following vehicle control via model predictive control and linear

quadratic control[J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2015, 229(12): 1695. DOI:10.1177/0954407014566031

[5]TIAN Haiyong, SHEN Yanhua, ZHANG Wenming, et al. Slip ratio control for articulated dump truck based on fuzzy sliding mode[C]// 2011 International Conference on Consumer Electronics, Communications and Networks (CECNet). Xianning: IEEE, 2011: 4404. DOI: 10.1109/CECNET.2011.5768560.

[6]BAI Guoxing, LIU Li, MENG Yu, et al. Path tracking of mining vehicles based on nonlinear model predictive control[J]. Applied Sciences, 2019, 9(7): 1372. DOI:10.3390/app9071372

[7]李斯旭, 徐彪, 胡满江, 等. 基于动力学模型预测控制的铰接车辆多点预瞄路径跟踪方法[J]. 汽车工程, 2021, 43(8): 1187
LI Sixu, XU Biao, HU Manjiang, et al. A dynamic model predictive control approach for multipoint preview path tracking of articulated vehicles[J]. Automotive Engineering, 2021, 43(8): 1187. DOI: 10.19562/j.chinasae.qcgc.2021.08.009

[8]LIN Fen, WANG Shaobo, ZHAO Youqun, et al. Research on autonomous vehicle path tracking control considering roll stability [J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2021, 235(1): 199. DOI:10.1177/0954407020942006

[9]刘清河, 王泽文, 赵立军. 自适应 LOS 制导结合 MPC 控制的车辆循迹优化[J]. 哈尔滨工业大学学报, 2022, 54(1): 96
LIU Qinghe, WANG Zewen, ZHAO Lijun. Vehicle tracking optimization based on adaptive LOS guidance and MPC control[J]. Journal of Harbin Institute of Technology, 2022, 54(1): 96. DOI: 10.11918/202012053

[10]GE Qiang, SARTORETTI G, DUAN Jingliang, et al. Distributed model predictive control of connected multi-vehicle systems at unsignalized intersections[C]//2022 IEEE International Conference on Unmanned Systems (ICUS). Guangzhou: IEEE, 2022: 1466. DOI: 10.1109/ICUS55513.2022.9986954

[11]LI Shengbo. Reinforcement learning for sequential decision and optimal control[M]. Singapore: Springer, 2023. DOI:10.1007/978-981-19-7784-8

[12]GUAN Yang, REN Yangang, SUN Qi, et al. Integrated decision and control: toward interpretable and computationally efficient driving intelligence[J]. IEEE Transactions on Cybernetics, 2023, 53(2): 859. DOI:10.1109/TCYB.2022.3163816

[13]DUAN Jingliang, LI Jie, GE Qiang, et al. Relaxed actor-critic with convergence guarantees for continuous-time optimal control of nonlinear systems[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(5): 3299. DOI: 10.1109/TIV.2023.3255264

[14]李永丰, 史静平, 章卫国, 等. 深度强化学习的无人作战飞机空战机动决策[J]. 哈尔滨工业大学学报, 2021, 53(12): 33
LI Yongfeng, SHI Jingping, ZHANG Weiguo, et al. Maneuver decision of UCAV in air combat based on deep reinforcement learning[J]. Journal of Harbin Institute of Technology, 2021, 53(12): 33. DOI: 10.11918/202005108

[15]温广辉, 杨涛, 周佳玲, 等. 强化学习与自适应动态规划:从基础理论到多智能体系统中的应用进展综述[J]. 控制与决策, 2023, 38(5): 1200
WEN Guanghui, YANG Tao, ZHOU Jialing, et al. Reinforcement learning and adaptive/approximate dynamic programming: a survey from theory to applications in multi-agent systems[J]. Control and Decision, 2023, 38(5): 1200. DOI:10.13195/j.kzyjc.2022.1933

[16]李升波, 占国建, 蒋宇轩, 等. 类脑学习型自动驾驶决控系统的关键技术[J]. 汽车工程, 2023, 45(9): 1499
LI Shengbo, ZHAN Guojian, JIANG Yuxuan, et al. Key technologies of brain-inspired decision and control intelligence for autonomous driving systems[J]. Automotive Engineering, 2023, 45(9): 1499. DOI: 10.19562/j.chinasae.qcgc.2023.ep.006

[17]段京良, 陈良发, 王文轩, 等. 智能汽车主动避撞工况的高实时预测控制[J]. 汽车安全与节能学报, 2023, 14(5): 580
DUAN Jingliang, CHEN Liangfa, WANG Wenxuan, et al. High real-time predictive control for active collision avoidance of intelligent vehicles[J]. Journal of Automotive Safety and Energy, 2023, 14(5): 580. DOI: 10.3969/j.issn.1674-8484.2023.05.007

[18]LIU Zhengyu, DUAN Jingliang, WANG Wenxuan, et al. Recurrent model predictive control: learning an explicit recurrent controller for nonlinear systems[J]. IEEE Transactions on Industrial Electronics, 2022, 69(10): 10437. DOI: 10.1109/TIE.2022.3153800

[19]DUAN Jingliang, GUAN Yang, LI Shengbo, et al. Distributional soft actor-critic: off-policy reinforcement learning for addressing value estimation errors[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(11): 6584. DOI: 10.1109/TNNLS.2021.3082568

[20]DUAN Jingliang, WANG Wenxuan, XIAO Liming, et al. DSAC-T: distributional soft actor-critic with three refinements [EB/OL]. 2023: 2310.05858. https://arxiv.org/ abs/2310.05858v4

[21]YU Haonan, XU Wei, ZHANG Haichao. TAAC: temporally abstract actor-critic for continuous control[EB/OL]. 2021: 2104.06521. https://arxiv.org/abs/2104.06521v3

[22]MYSORE S, MABSOUT B, MANCUSO R, et al. Regularizing action policies for smooth control with reinforcement learning[C]// 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an: IEEE, 2021: 1810. DOI: 10.1109/ICRA48506.2021.9561138

[23]KOBAYASHI T. L2C2: locally lipschitz continuous constraint towards stable and smooth reinforcement learning[C]//2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Kyoto: IEEE, 2022: 4032. DOI: 10.1109/IROS47612.2022.9981812

[24]TAKASE R, YOSHIKAWA N, MARIYAMA T, et al. Stability-certified reinforcement learning control via spectral normalization [J]. Machine Learning with Applications, 2022, 10: 100409. DOI: 10.1016/j.mlwa.2022.100409

[25]SONG Xujie, DUAN Jingliang, WANG Wenxuan, et al. LipsNet: a smooth and robust neural network with adaptive Lipschitz constant for high accuracy optimal control[C]//Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu: ML Research Press, 2023: 32253

[26]WANG Wenxuan, ZHANG Yuhang, GAO Jiaxin, et al. GOPS: a general optimal control problem solver for autonomous driving and industrial control applications [J]. Communications in Transportation Research, 2023, 3: 100096. DOI: 10.1016/j.commtr.2023.100096

[27]FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm: International Machine Learning Society (IMLS), 2018: 2587

[28]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 (ICML). Stockholm: International Machine Learning Society (IMLS), 2018: 2976

[29]ANDERSSON J A E, GILLIS J, HORN G, et al. CasADi: a software framework for nonlinear optimization and optimal control [J]. Mathematical Programming Computation, 2019, 11(1): 1. DOI:10.1007/s12532-018-0139-4

(编辑　张　红)