基于滚动时域强化学习的智能车辆侧向控制算法

张兴龙 [1] 陆 阳 [1] 李文璋 [1] 徐 昕 [1]

摘 要 针对智能车辆的高精度侧向控制问题, 提出一种基于滚动时域强化学习 (Receding horizon reinforcement learning, RHRL) 的侧向控制方法. 车辆的侧向控制量由前馈和反馈两部分构成, 前馈控制量由参考路径的曲率以及动力学模型直接计算得出; 而反馈控制量通过采用滚动时域强化学习算法求解最优跟踪控制问题得到. 提出的方法结合滚动时域优化机制, 将无限时域最优控制问题转化为若干有限时域控制问题进行求解. 与已有的有限时域执行器−评价器学习不同, 在每个预测时域采用时间独立型执行器−评价器网络结构学习最优值函数和控制策略. 与模型预测控制 (Model predictive control, MPC) 方法求解开环控制序列不同, RHRL 控制器的输出是一个显式状态反馈控制律, 兼具直接离线部署和在线学习部署的能力. 此外, 从理论上证明了 RHRL 算法在每个预测时域的收敛性, 并分析了闭环系统的稳定性. 在仿真环境中完成了结构化道路下的车辆侧向控制测试. 仿真结果表明, 提出的 RHRL 方法在控制性能方面优于现有先进算法, 最后, 以红旗 E-HS3 电动汽车作为实车平台, 在封闭结构化城市测试道路和乡村起伏砂石道路下进行了侧向控制实验. 实验结果显示, RHRL 在结构化城市道路中的侧向控制性能优于预瞄控制, 在乡村道路中具有较强的路面适应能力和较好的控制性能.

关键词 滚动时域, 强化学习, 智能汽车, 侧向控制

引用格式 张兴龙, 陆阳, 李文璋, 徐昕. 基于滚动时域强化学习的智能车辆侧向控制算法. 自动化学报, 2023, **49**(12): 2481−2492

DOI 10.16383/j.aas.c210555

Receding Horizon Reinforcement Learning Algorithm for Lateral Control of Intelligent Vehicles

ZHANG Xing-Long[1] LU Yang[1] LI Wen-Zhang[1] XU Xin[1]

Abstract This paper presents a receding horizon reinforcement learning (RHRL) algorithm for realizing high-accuracy lateral control of intelligent vehicles. The overall lateral control is composed of a feedforward control term that is directly computed using the curvature of the reference path and the dynamic model, and a feedback control term that is generated by solving an optimal control problem using the proposed RHRL algorithm. The proposed RHRL adopts a receding horizon optimization mechanism, and decomposes the infinite-horizon optimal control problem into several finite-horizon ones to be solved. Different from existing finite-horizon actor-critic learning algorithms, in each prediction horizon of RHRL, a time-independent actor-critic structure is utilized to learn the optimal value function and control policy. Also, compared with model predictive control (MPC), the control learned by RHRL is an explicit state-feedback control policy, which can be deployed directly offline or learned and deployed synchronously online. Moreover, the convergence of the proposed RHRL algorithm in each prediction horizon is proven and the stability analysis of the closed-loop system is peroformed. Simulation studies on a structural road show that, the proposed RHRL algorithm performs better than current state-of-the-art methods. The experimental studies on an intelligent driving platform built with a Hongqi E-HS3 electric car show that RHRL performs better than the pure pursuit method in the adopted structural city road scenario, and exhibits strong adaptability to road conditions and satisfactory control performance in the country road scenario.

Key words Receding horizon, reinforcement learning, intelligent vehicles, lateral control

Citation Zhang Xing-Long, Lu Yang, Li Wen-Zhang, Xu Xin. Receding horizon reinforcement learning algorithm for lateral control of intelligent vehicles. *Acta Automatica Sinica*, 2023, **49**(12): 2481−2492

收稿日期 2021-06-20 录用日期 2021-11-02
Manuscript received June 20, 2021; accepted November 2, 2021
国家重点研究发展计划 (2018YFB1305105), 国家自然科学基金 (62003361, 61825305) 资助
Supported by National Key Research and Development Program of China (2018YFB1305105) and National Natural Science Foundation of China (62003361, 61825305)
本文责任编委 吕宜生
Recommended by Associate Editor LV Yi-Sheng

作为智能驾驶中的一个重要模块, 运动控制器通过控制刹车、油门、档位、方向盘等执行机构使车辆安全、平稳地跟踪参考路径. 智能车辆在行驶中主要涉及两种运动形式: 纵向运动和侧向运动. 为

1. 国防科技大学智能科学学院 长沙 410073
1. College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073

了简化控制器的设计, 通常将运动进行解耦并分别设计纵向和侧向控制器. 与纵向控制中的舒适性、平滑性控制需求不同, 跟踪精度是侧向控制器的核心考量. 由于车辆本身是一个复杂的高阶非线性系统, 同时又受到行驶环境的影响, 因此如何提高跟踪精度是运动控制中的难题[1−3]. 本文主要针对智能车辆的高精度侧向控制问题开展研究.

目前, 常见的侧向控制方法包括比例−积分−微分 (Proportional-integral-derivative, PID) 控制方法[4−8]、模糊控制方法[9−12]、反馈控制方法[13−16]、模型预测控制 (Model predictive control, MPC) 方法、基于强化学习 (Reinforcement learning, RL) 的控制方法. 在上述方法中, PID 的优势在于不需要对车辆进行建模, 控制器的鲁棒性较强、容易实现, 但难以保证性能指标的最优性; 模糊控制器可以推理并产生专家行为, 但是由于驾驶环境的复杂性导致了基于驾驶员行为的模糊规则较难制定.

典型的反馈控制器根据智能车辆与参考路径之间的几何关系计算出航向偏差与侧向偏差, 并计算出方向盘转角直接用于转向控制. 根据选取的路径参考点与车辆位置之间的关系, 可以分为单点跟踪法、预瞄距离法、Stanley 法、点跟踪法[13] 和预瞄距离法[14−15], 具有算法简单、易于实现的特点, 但预瞄距离的选取完全依赖于设计者的经验; Stanley 方法[16] 由美国斯坦福大学的无人车队率先提出, 该方法适用于较低的车速, 并且要求参考轨迹的曲率具有连续性.

将 MPC 方法用于车辆运动控制的研究成果颇多[17−24]. 在上述成果中, Falcone 等[18] 提出了基于连续线性化模型的 MPC 运动控制器, 仿真的结果表明, 连续线性化的 MPC 设计方法能够降低计算代价. Carvalho 等[19] 研究了采用局部线性化 MPC 的局部路径规划算法, 并对非线性的避障边界进行了线性化和凸逼近处理. Beal 等[20] 考虑了车辆的处理极限, 通过引入摩擦力圆来分配车辆的纵向与侧向加速度, 使车辆在控制过程中最大程度地利用地面摩擦力. 在计算车辆与参考路径之间的航向与侧向偏差时需要求出车辆在参考路径上的投影点, 计算过程十分复杂. Liniger 等[21] 提出一种模型预测轮廓控制 (Model predictive contouring control, MPCC) 的侧向运动方法, 该方法通过估计投影点的位置来计算侧向偏差, 一定程度上降低了计算复杂度. Kabzan 等[22] 基于输入输出数据构建了赛车的非参数化动力学模型, 然后采用 MPC 方法同时控制赛车的速度与转向. Ostafew 等[23] 采用高斯过程回归构建移动机器人的非参数化模型, 并设计了鲁棒的非线性 MPC 算法, 实现机器人在越野环境下的避障与跟踪控制. 总的来说, 基于 MPC 方法的车辆运动控制器一般需要采用数值计算的方法实时求解一个开环控制序列, 其性能可能会受到模型准确度的影响. 另外, 在线计算复杂度也是一个无法回避的问题.

近年来, 由于其高效求解优化问题的能力和自适应学习能力, 强化学习和近似动态规划方法 (Approximate dynamic programming, ADP) 广泛应用于机器人决策与控制算法的设计[25−26]. Oh 等[27] 采用对偶启发式 (Dual heuristic programming, DHP) 方法设计了车辆侧向控制器. 杨慧媛等[28] 针对轮式移动机器人的跟踪控制问题, 提出了一种学习型 PID 控制方法, 以优化机器人的跟踪偏差为目标, 采用 DHP 算法实时调整 PID 参数以提高路径跟踪精度. 连传强等[29] 提出一种基于核特征的 DHP (Kernel-based DHP, KDHP) 算法, 并设计了车辆侧向运动控制器, 通过在城市道路、高速公路等驾驶环境下的仿真测试证明了基于 KDHP 算法的有效性. 黄振华等[30] 设计了基于同步迭代的 DHP (Synchronous iterative DHP, SI-DHP) 算法的车辆侧向运动控制器.

为了解决强化学习算法学习效率低的问题, Lian 等[31] 针对轮式移动机器人对象提出了一种基于滚动时域的对偶启发式规划方法 (Receding horizon DHP, RH-DHP). 仿真结果表明, RH-DHP 算法在控制效果上优于传统 DHP 和 MPC 的控制效果, 并且相比于 MPC 具有更短的运算时间. 但是上述方法还存在以下三个方面的问题: 1) 其执行器−评价器网络需要将时间作为额外的输入信号, 增加了网络设计的复杂度; 2) 该工作没有分析执行器−评价器学习算法的收敛性以及在此基础上的闭环稳定性; 3) 该方法仅在小型轮式仿真平台中进行了验证, 目前, 尚未见到其在实际智能车辆平台中进行应用验证的相关报道.

最近, 也有一些重要的工作采用深度学习和深度强化学习基于图像或状态信息设计控制器实现车辆的侧向控制[32−34]. 这类方法的主要优点是利用深度网络来提高强化学习或监督学习的特征表示能力, 训练过程中完全由数据驱动, 不需要动力学模型信息. 其不足之处在于: 1) 由于深度网络过于复杂, 一般只能离线训练控制策略用于在线部署, 其控制性能容易受训练样本数量和分布的影响; 2) 针对深度网络学习的收敛性和鲁棒性等理论特性分析仍是目前学术界需要解决的一个重要难点问题.

由上述问题驱动, 本文针对智能车辆的高精度侧向控制问题, 提出了一种基于滚动时域强化学习的侧向控制方法. 首先构建了智能车动力学四阶偏

差模型. 车辆的转向控制量由前馈和反馈两部分构成. 前馈控制量由参考路径的曲率以及偏差模型直接计算得出; 而反馈控制量通过采用本文提出的滚动时域强化学习 (Receding horizon RL, RHRL) 算法求解最优跟踪控制问题得到. 有别于传统基于强化学习的最优控制方法, RHRL 采用滚动时域优化机制, 将无限时域的最优控制问题转化为一系列有限时域的启发式动态规划 (Heuristic dynamic programming, HDP) 问题进行求解. 与已有的有限时域执行器−评价器学习算法[31, 35] 不同, 在每个预测时域, 我们采用时间独立型执行器−评价器结构在线学习逼近最优值函数和控制函数. 与 MPC 方法求解开环控制序列不同, 该方法学习得到的策略是一个显式状态反馈控制律, 具有离线直接部署和在线学习部署的能力. 此外, 本文从理论上分析了提出的 RHRL 算法在每个预测时域内的收敛性和闭环稳定性. 最后, 基于 RHRL 算法进行了侧向控制的大量的仿真对比实验和实车验证. 在结构化城市道路下的仿真和实车实验结果表明, RHRL 算法在仿真和实验中的控制性能均优于预瞄控制; 在仿真测试中, 其控制性能与 MPC 相当并在计算效率方面具有优势, 与最近流行的软执行器−评价器 (Soft actor-critic, SAC) 算法和深度确定性策略梯度 (Deep deterministic policy gradient, DDPG) 算法相比, 控制性能更好, 且具有更低的样本复杂度和更高的学习效率. 在乡村砂石道路下的实验结果表明, RHRL 具有较强的路面适应能力和较好的控制性能.

需要强调的是, 与最近发展的基于深度学习和深度强化学习的方法[32−34] 相比, 本文提出的 RHRL 算法采用简单的网络结构, 计算效率更高, 可以在线同步训练和部署, 具有较强的环境适应能力; 而且, RHRL 算法通过引入滚动时域优化思想来提高强化学习的实时学习效率和稳定性. 更重要地, 我们分析证明了 RHRL 中执行器−评价器学习算法的收敛性以及闭环稳定性, 并在实际平台中进行了应用验证. 实验结果证明了 RHRL 算法的有效性.

本文的结构如下:第 1 节首先介绍智能驾驶车辆的侧向动力学模型和控制问题描述; 第 2 节主要介绍基于滚动时域强化学习的车辆侧向控制算法及其收敛性分析; 第 3 节和第 4 节分别给出仿真和实验验证结果以及本文的结论.

本文符号定义如下: 对于一个普适变量 $z \in \mathbf{R}^p$, 定义 $\Delta z(l+1) = z(l+1) - z(l)$, 其中 l 是离散时间指针; 定义 $\|z\|_Q^2 = z^{\mathrm{T}}Qz$, 其中矩阵 $Q \in \mathbf{R}^{p\times p}$. 在一个预测时域 $[k, k+N]$ 内, 采用变量 z 简化表示 $z(l)$, 其中时间指针 $l \in [k, k+N-1]$, 采用 z^+ 表示其下一个时间步的变量值, 也就是 $z^+ = z(l+1)$; 采用 z_f 表示其预测时域 $[k, k+N]$ 的终端变量值 $z(k+N)$. 对于一个关于变量 x 的函数 $f(x)$, 定义 $\nabla f(x)$ 为其关于 x 的梯度. 给定一个矩阵 $B \in \mathbf{R}^{p\times p}$, 采用 $\lambda_{\min}(B)$ 表示 B 的最小特征值.

1　车辆侧向动力学模型和控制问题描述

由于车辆本身的运动较为复杂并且在运动过程中还要受到环境因素的影响, 为了降低建模工作的难度, 将原来车辆的四轮侧向模型简化为如图 1 所示的二自由度侧向模型, 即自行车模型.

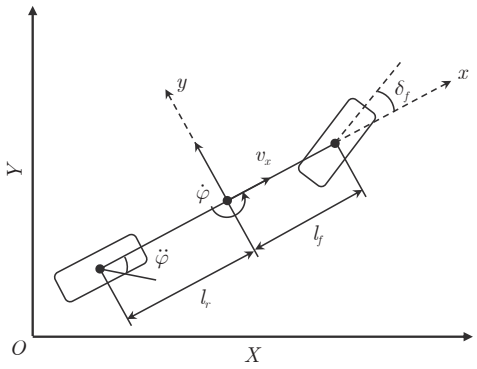

图 1　智能车辆二自由度侧向模型

Fig. 1　Two-degree-of-freedom lateral model of intelligent vehicle

根据牛顿运动定律, 车辆的运动满足如下动力学方程

$$\begin{cases} \dot{v}_y = \dfrac{1}{m}\left(F_{yf} + F_{yr}\right) - \dot{\varphi} v_x \\ \ddot{\varphi} = \dfrac{1}{I_z}\left(l_f F_{yf} - l_r F_{yr}\right) \end{cases} \tag{1}$$

其中, v_x 和 v_y 分别表示在车体坐标系 XOY 下车辆的纵向与横向速度, φ 表示车辆的偏航角, $\dot{\varphi}$ 表示车辆的横摆角速度, δ_f 表示前轮的偏转角, m 和 I_z 分别表示车身的质量以及绕 z 轴的转动惯量, l_f 和 l_r 分别表示质心到车辆前后轴的距离, F_{yf} 和 F_{yr} 分别表示车辆前轮与后轮的侧向轮胎力.

假设车辆行驶中轮胎侧滑角很小, 轮胎的侧向力可以按照式 (2) 近似计算:

$$\begin{cases} F_{yf} = 2C_f\left(\delta_f - \dfrac{v_y + l_f\dot{\varphi}}{v_x}\right) \\ F_{yr} = -2C_r\dfrac{v_y - l_r\dot{\varphi}}{v_x} \end{cases} \tag{2}$$

其中, C_f 和 C_r 分别表示车辆前后轮的侧偏刚度.

考虑车体坐标系与全局坐标系的相对位置关系, 可以得到如下方程:

$$\begin{cases}\dot{Y}=v_x\sin(\varphi)+v_y\cos(\varphi)\\ \dot{X}=v_x\cos(\varphi)-v_y\sin(\varphi)\end{cases}\tag{3}$$

选取 $Z=[X,\ Y,\ \varphi,\ \dot{\varphi},\ v_y]$ 作为系统的状态变量, 前轮偏转角 δ_f 作为控制量, 联立式 (1) ~ (3), 可以得到车辆的动力学方程

$$\dot{Z}=F(Z)+G(Z)\delta_f\tag{4}$$

在进行跟踪控制时, 有必要描述车辆与期望路径之间的相对位置关系, 如图 2 所示, P 点表示车辆处于当前位置时距离道路中心线的最近点, 我们称其为道路投影点. 记 $P(X_p,\ Y_p,\ \varphi_d,\ \kappa)$ 为投影点处的道路信息, 其中, $X_p,\ Y_p$ 是投影点 P 的全局坐标; φ_d 是 P 的切线与 X 轴的夹角, 也称为道路的方向; κ 是 P 点处道路的曲率.

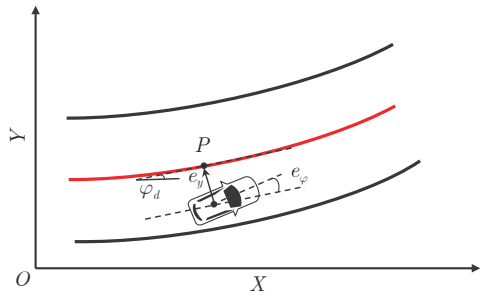

图 2 侧向误差模型

Fig. 2 Lateral error model

从投影点 P 到车辆质心之间的距离称为侧向偏差 e_y, 并且规定沿着行进方向车辆位于道路中心线左侧时 $e_y>0$; 车辆位于道路中心线右侧时 $e_y<0$. 因此, 侧向偏差可以表示为 $e_y=-(X-X_p)\sin(\varphi_d)+(Y-Y_p)\cos(\varphi_d)$. 定义车辆的航向偏差 e_φ 为航向与道路方向之差, 即 $e_\varphi=\varphi-\varphi_d$. e_y 和 e_φ 对时间的一阶导数为

$$\begin{cases}\dot{e}_y=v_y\cos(e_\varphi)+v_x\sin(e_\varphi)\\ \dot{e}_\varphi=w-\kappa(v_x\cos(e_\varphi)-v_y\sin(e_\varphi))\end{cases}\tag{5}$$

其中, $w=\dot{\varphi}$. 假设在运动过程中车辆的纵向速度 v_x 保持不变且不出现侧滑现象, 车辆的参考路径的期望横摆角速度是恒定的, 那么当车辆稳定跟踪期望道路时的侧向加速度为 $a_y=v_x^2\kappa$. 假设航向偏差 e_φ 较小, 根据小角度定理, 有 $\sin(e_\varphi)\approx e_\varphi,\ \cos(e_\varphi)\approx 1$, 那么, 侧向偏差对时间的二阶导数可以表示为

$$\ddot{e}_y=(\dot{v}_y+v_xw)-v_x^2\kappa\tag{6}$$

其一阶导数可以近似表示为

$$\dot{e}_y=v_y+v_xe_\varphi\tag{7}$$

将式 (6) 和式 (7) 代入式 (1)[36] 和式 (2) 中, 得

$$\dot{e}=A_ce+B_{c1}u+B_{c2}w_d\tag{8}$$

其中, $w_d=\dot{\varphi}_d,\ e=[e_y,\ \dot{e}_y,\ e_\varphi,\ \dot{e}_\varphi]^{\mathrm{T}}$, 控制量 $u=\delta_f$,

$$A_c=\begin{bmatrix}0 & 1 & 0 & 0\\ 0 & -\frac{2(C_f+C_r)}{mv_x} & \frac{2(C_f+C_r)}{m} & -\frac{2(C_fl_f-C_rl_r)}{mv_x}\\ 0 & 0 & 0 & 1\\ 0 & -\frac{2(C_fl_f-C_rl_r)}{I_zv_x} & \frac{2(C_fl_f-C_rl_r)}{I_z} & -\frac{2\left(C_fl_f^2+C_rl_r^2\right)}{I_zv_x}\end{bmatrix}$$

$$B_{c1}=\begin{bmatrix}0\\ \frac{2C_f}{m}\\ 0\\ \frac{2C_fl_f}{I_z}\end{bmatrix},\quad B_{c2}=\begin{bmatrix}0\\ -\frac{2(C_fl_f-C_rl_r)}{mv_x}-v_x\\ 0\\ -\frac{2\left(C_fl_f^2+C_rl_r^2\right)}{I_zv_x}\end{bmatrix}$$

给定一个采样周期 Δt, 可以离散化得到式 (8) 的离散时间模型为

$$e(k+1)=Ae(k)+B_1u(k)+B_2w_d(k)\tag{9}$$

其中, $A=I+\Delta tA_c,\ B_1=\Delta tB_{c1},\ B_2=\Delta tB_{c2},\ k$ 是离散时间指针. 在控制过程中, 由于前轮转角所对应的执行机构有限幅, 因此我们假设反馈控制量满足输入约束 $|u|\le\bar{u}$, 其中 $\bar{u}$ 表示前轮最大偏转角.

针对上述模型 (9), 假设给定参考的路径信息 $(X_i,\ Y_i)_{i=1}^M$, 本文的控制目标是设计一个基于滚动时域强化学习的侧向控制算法 (如图 3 所示), 使得在控制过程中, 上述侧向误差状态量逐渐收敛至 0, 即 $e\to 0$, 同时需要满足控制约束 $|u|\le\bar{u}$.

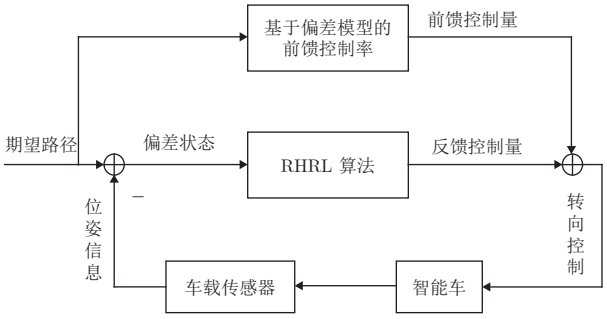

图 3 智能车侧向控制框图

Fig. 3 Lateral control diagram of intelligent vehicle

2 基于滚动时域强化学习的智能车辆侧向控制算法

本节详细给出基于滚动时域强化学习的侧向控制算法. 我们首先设计智能车辆有限时域侧向控制问题的性能指标, 在此基础上给出滚动时域强化学习算法的主要思想和基于执行器–评价器的设计实

现及其收敛性分析.

2.1　有限时域侧向控制问题的性能指标设计

对于系统偏差模型 (9), 我们将控制量拆分成前馈量 u_f 加反馈量 u_b 的形式, 即 $u = u_f + u_b$ (如图 3 所示). 前馈控制量是车辆处于稳态行驶中的期望控制量. 当车辆稳定跟踪参考路径时, 有 $e(k) = e(k+1) = 0$ 成立, 而且 $u_b = 0$, 可以求得前馈控制量 u_f, 使得

$$\sum_{j=0}^{\infty} A^j B_1 u_f \approx -\sum_{j=0}^{\infty} A^j B_2 w_d \tag{10}$$

其中, w_d 的值也可以通过 $w_d = v_x \kappa$ 计算得到.

由于在任意当前时刻 k, u_f 可以很容易求解得到, 我们假设 u_f 在整个预测时域 $[k, k+N]$ 保持恒定不变, 那么需要求解的反馈控制量 u_b 应满足以下约束条件

$$u_b \in \mathcal{U}_b = \{u \in \mathbf{R} | \underline{u}_b \le u \le \bar{u}_b\} \tag{11}$$

其中, $\bar{u}_b = -\bar{u} - u_f$, $\underline{u}_b = \bar{u} - u_f$. 本文提出的滚动时域强化学习算法, 在每个预测时域通过优化控制量 $u_b \in \mathcal{U}_b$ 最小化如下性能指标函数:

$$V(e(k)) = \sum_{l=k}^{k+N-1} L\left(e(l), u_b(l)\right) + V_f\left(e(k+N)\right) \tag{12}$$

其中,代价函数 $L\left(e(l), u_b(l)\right) = e^{\mathrm{T}}(l) Q e(l) + R(u_b(l))^2$, $Q \in \mathbf{R}^{4\times 4}$ 是正定矩阵, R 是正实数, 预测时域终端的代价函数为

$$V_f(e(k+N)) = e^{\mathrm{T}}(k+N) \bar{P} e(k+N) \tag{13}$$

其中, 惩罚矩阵 $\bar{P} \in \mathbf{R}^{4\times 4}$ 是正定矩阵, 可通过如下 Lyapunov 方程求解得到

$$F^{\mathrm{T}} \bar{P} F - \bar{P} = -Q - K^{\mathrm{T}} R K \tag{14}$$

其中, $F = A + B_1 K$, $K \in \mathbf{R}^{1\times 4}$ 是反馈增益矩阵, 满足 F 是 Schur 稳定的.

注 1. 需要强调的是, 另一种供选择的设计方法是将计算得到的前馈控制量用作整体控制量的参考信号, 由此可以设计一个新的代价函数 $L(e(l), u(l)) = e^{\mathrm{T}}(l) Q e(l) + R(u(l) - u_f(l))^2$. 与本文中的设计不同, 这里整体控制量 u 变成了待优化的变量, 其通过优化得到的控制量可以直接应用到系统中.

2.2　基于滚动时域强化学习的侧向路径控制算法

首先, 根据式 (12), 对任意 $l \in [k, k+N-1]$, 可以将值函数表示成差分形式, 即

$$V(e(l)) = L\left(e(l), u_b(l)\right) + V\left(e(l+1)\right) \tag{15}$$

其中, $V\left(e(k+N)\right) = V_f(e(k+N))$. 在第 l 个预测时刻, 定义 $V^*(e(l))$ 为最优值函数, 给出上述有限时域优化控制问题的 HJB 方程, 即

$$V^*(e(l)) = \min_{u_b(l)\in\mathcal{U}_b} L\left(e(l), u_b(l)\right) + V^*\left(e(l+1)\right) \tag{16}$$

以及最优控制策略

$$u^*(e(l)) = \arg \min_{u_b(l)\in\mathcal{U}_b} L\left(e(l), u_b(l)\right) + V^*\left(e(l+1)\right) \tag{17}$$

实际上, 由于存在控制约束, 通过式 (16) 和式 (17) 很难求解得到 V^* 和 u^* 的解析解. 原则上, 可以通过值迭代的方法近似求解其值函数和控制策略的最优解. 对任意 $l \in [k, k+N-1]$, 给定初始值 $V^0(e(l)) = 0$, 迭代步数 $i = 0, 1, 2\cdots$, 需要重复求解如下两个步骤, 直至 $V^{i+1}(e(l)) - V^i(e(l)) \to 0$.

1) 策略更新

$$u^i(e(l)) = \arg \min_{u_b(l)\in\mathcal{U}_b} L\left(e(l), u_b(l)\right) + V^i\left(e(l+1)\right) \tag{18a}$$

2) 值更新

$$V^{i+1}\left(e(l)\right) = L\left(e(l), u_b^i(e(l))\right) + V^i\left(e(l+1)\right) \tag{18b}$$

引理 1. 基于上述算法步骤 (18a) 和 (18b), $V^i(e(l)) \le V^{i+1}(e(l))$, 且 $V^\infty(e(l)) \to V^*(e(l))$, $l \in [k, k+N]$.

证明. 参见文献 [37]. □

2.3　滚动时域执行器–评价器学习实现

本节采用执行器–评价器结构来实现上述有限时域值函数迭代算法. 在已有的有限时域强化学习控制算法中[31, 35], 预测时域内的值函数被认为是一个时间依赖函数. 因此, 在设计执行器和评价器时不仅需要把时间作为额外输入信号, 而且还会因此增加网络结构的复杂度. 接下来将证明, 对于线性系统而言, 值函数 $V\left(e(l)\right)$ 在一定条件下是一个与时间无关的函数.

假设 1 (控制策略). 存在一个控制策略 $u_b(e) = \Gamma(v(e))$, 使得系统 (9) 在控制策略 $u = u_f + u_b$ 驱动下是渐近稳定的, 其中, $\Gamma(v(e))$ 是一个连续函数, 使得 $u_b(e) \in \mathcal{U}_b$, $\forall v(e) \in \mathbf{R}$.

注 2. 上述假设条件实际上是系统 (9) 可镇定性的另一种表现形式. 本文所述的动力学模型 (9) 是可控的, 因此肯定存在连续函数 $u_b(e) \in \mathcal{U}_b$, 使得式 (9) 在控制策略 $u = u_f + u_b$ 驱动下是渐近稳定的. 因此, 上述假设条件是合理的.

我们定义 $\mathcal{X}_f$ 为控制律 $u_b = Ke \in \mathcal{U}_b$ 下的一个控制不变集, 由此得到定理 1.

定理 1 (时间独立值函数). 如果预测时域 N 的

取值满足: 在任意预测时域 $[k, k+N]$ 内, 对于任意初始状态 $e(k) \in \mathbf{R}^4$, 系统 (9) 在控制策略 $u(e(l))$, $l \in [k, k+N-1]$ 驱动下的终端状态 $e(k+N) \in \mathcal{X}_f$, 那么, 存在控制策略 $u_b(e(l)) \in \mathcal{U}_b$, 使得 $V(e(l))$, $\forall l \in [k, k+N-1]$ 是与时间无关的函数.

证明. 1) 对于 $e(k) \in \mathcal{X}_f$ 的情况, 根据 $\mathcal{X}_f$ 的定义, 存在控制律 $u_b = Ke = \Gamma(Ke) \in \mathcal{U}_b$, 使得未来任意时刻的状态量都满足 $x(l) \in \mathcal{X}_f$. 据此, 可以求解得到

$$V(e(l)) = \sum_{i=l}^{k+N-1} L\left(e(i),\, u_b(i)\right) + V_f\left(e(k+N)\right) = e^{\mathrm{T}}(l)\bar{P}e(l)$$

2) 对于 $e(k) \notin \mathcal{X}_f$ 的情况, 根据假设 1, 存在一个控制策略 $u_b = \Gamma(v(e))$ 和有限的预测步长 N, 使得 $e(k+N) \in \mathcal{X}_f$. 特别地, 令 $v = Ke$, 则

$$V(e(l)) = \sum_{i=l}^{k+N-1} L\left(e(i),\, u_b(i)\right) + V_f\left(e(k+N)\right) = \sum_{i=l}^{+\infty} L\left(e(i),\, u_b(i)\right)$$

其中, $u_b = \Gamma(v(e))$.

因此, 存在一个与时间无关的值函数和策略. □

受此启发, 我们采用时间独立的执行器−评价器结构来实现上述有限时域值函数迭代过程. 首先, 设计一个评价器网络来逼近值函数

$$\hat{V}(e) = \hat{W}_c^{\mathrm{T}}\phi(e) \tag{19}$$

其中, $\hat{W}_c \in \mathbf{R}^{N_c}$ 表示评价器网络的权重, N_c 是网络节点数; $\phi(e)$ 是网络的基函数. 根据评价器网络的定义, 其所产生的误差 E 和终端误差 E_f 可以表示为

$$\begin{cases} E(l) = \hat{W}_c^{\mathrm{T}}\phi(l) - L\left(e(l),\, \hat{u}_b(l)\right) - \hat{W}_c^{\mathrm{T}}\phi(l+1) \\ E_f = \hat{W}_c^{\mathrm{T}}\phi\left(e_f\right) - e_f^{\mathrm{T}}\bar{P}e_f \end{cases} \tag{20}$$

其中, $e_f = e(k+N)$ 可随机在 0 点附近取值. 通过最小化 $E_c(l) = (E(l))^2 + E_f^2$ 可以得到评价器网络权重的更新规则为

$$\hat{W}_c(l+1) = \hat{W}_c(l) + \eta_c(\Delta\phi(e(l+1))E(l) - \phi(e_f)E_f) \tag{21}$$

其中, $\eta_c > 0$ 是评价器网络的学习率.

接下来, 为了处理控制约束, 我们构造执行器网络为

$$\hat{u}_b(l) = \tilde{u}_1 \tanh\left(\hat{W}_a^{\mathrm{T}}\psi(e(l))\right) + \tilde{u}_2 \tag{22}$$

其中, $\tilde{u}_1 = 0.5(\bar{u}_b - \underline{u}_b)$, $\tilde{u}_2 = 0.5(\bar{u}_b + \underline{u}_b)$, $\hat{W}_a \in \mathbf{R}^{N_a}$ 是执行器网络权重; $\psi(e)$ 是网络的基函数向量; N_a 表示网络的节点数. 由于执行器网络的目标是逼近最优控制策略, 我们定义如下控制量偏差, 即

$$E_a(l) = \hat{W}_a^{\mathrm{T}}\psi(e(l)) + \frac{1}{2}R^{-1}B_1^{\mathrm{T}}\nabla\phi(e(l))\hat{W}_c(l) \tag{23}$$

最小化 E_a^2 可以得到网络权值的更新规则为

$$\hat{W}_a(l+1) = \hat{W}_a(l) - \eta_a\frac{\partial E_a^2(l)}{\partial \hat{W}_a(l)} \tag{24}$$

其中, $\eta_a > 0$ 是执行器网络的学习率.

下面给出采用执行器−评价器实现上述有限时域强化学习算法的主要步骤.

步骤 1. 初始化权值 $\hat{W}_c$ 和 $\hat{W}_a$, 并获取初始状态 $Z(0)$.

步骤 2. 在 $t = k\Delta t$ 时刻, 根据状态 $Z(t)$ 找到投影点 P, 并计算出偏差状态 $e(t)$.

步骤 3. $\forall l \in [k, k+N-1]$, 重复步骤 3.1 ~ 3.3:

步骤 3.1. 根据式 (10) 和式 (22), 分别计算出 $u_f(l)$ 和 $\hat{u}_b(l)$.

步骤 3.2. 根据式 (21) 和式 (24), 更新 $\hat{W}_c(l)$ 和 $\hat{W}_a(l)$.

步骤 3.3. 根据式 (10) 和式 (22), 计算 $u(l) = u_f(l) + \hat{u}_b(l)$, 并应用到预测模型, 得到 $e(l+1)$.

步骤 4. 根据式 (10) 和式 (22), 分别计算 $u_f(k)$ 和 $\hat{u}_b(e(k))$.

步骤 5. 在时间周期 $[k\Delta t, (k+1)\Delta t]$ 将控制量 $u(t) = u(k\Delta t)$ 作用到智能车上, 并更新系统状态 $Z((k+1)\Delta t)$.

步骤 6. 设定 $k \leftarrow k+1$, 基于滚动时域优化策略, 重复操作步骤 2 ~ 5.

2.4 有限时域执行器−评价器权值收敛性分析

本节给出上述滚动时域强化学习算法在每个预测时域 $[k, k+N-1]$ 内的收敛性分析. 首先, 可以将 (局部) 最优值函数和控制策略表示成网络的形式, 即

$$V^*(e) = W_c^{\mathrm{T}}\phi(e) + \kappa_c$$

$$u_b^* = \tilde{u}_1 \tanh\left(W_a^{\mathrm{T}}\psi(e) + \kappa_a\right) + \tilde{u}_2$$

其中, W_c 和 W_a 是权值矩阵, κ_c 和 κ_a 是重构误差.

假设 2 (网络重构误差).

1) $\|W_c\| \le W_{c,\,m}$, $\|\phi\| \le \phi_m$, $\|\nabla\phi\| \le \bar{\phi}_m$, $\|\kappa_c\| \le \kappa_{c,\,m}$, $\|\nabla\kappa_c\| \le \bar{\kappa}_{c,\,m}$;

2) $\|W_a\| \le W_{a,\,m}$, $\|\psi\| \le \psi_m$, $\|\kappa_a\| \le \kappa_{a,\,m}$.

假设 3 (持续激励). 存在正实数 q_1, q_2, $q_1 < q_2$, 使得

$$q_1 \le \bar{\phi},\ \bar{\phi}_f \le q_2 \tag{25}$$

其中, $\bar{\phi}=\Delta\phi^{\mathrm{T}}\Delta\phi$, $\bar{\phi}_f=\phi_f^{\mathrm{T}}\phi_f$, $\phi_f=\phi(e_f)$.

为了更紧凑地描述下述定理, 定义 $\gamma_1=4-4\bar{\psi}\eta_a-(4-8\bar{\psi}\eta_a)(\beta_1+\beta_3)$, $\bar{\psi}=\psi^{\mathrm{T}}\psi$, $\tilde{\phi}=\bar{\phi}(l+1)+\bar{\phi}_f$, $\alpha=2\tilde{\phi}-2\eta_c\tilde{\phi}^2-\beta_0$, $\gamma_2=1/\beta_1+(8\bar{\psi}\beta_2+4\bar{\psi})\eta_a$, β_0, β_1, β_2, β_3 是可调正实数.

定理 2. 在假设 2 和假设 3 下, 如果选择合适的学习律 η_c 和 η_a 以及 $\{\beta_i\}_{i=0}^3$, 使得 $\gamma_1>0$, $\alpha-\gamma_2>0$, 那么采用上述策略更新律 (21) 和 (24) 的网络权值 $\hat{W}_c$ 和 $\hat{W}_a$ 将渐近收敛至如下区域:

$$\|\tilde{W}_c\|\le\frac{\sqrt{\mathrm{error}_t}}{\sqrt{\gamma_1}} \tag{26a}$$

$$\|\xi_a\|\le\frac{\sqrt{\mathrm{error}_t}}{\sqrt{\alpha-\gamma_2}\lambda_{\min}(\bar{g})} \tag{26b}$$

其中, $\tilde{W}_c=W_c-\hat{W}_c$, $\xi_a=\tilde{W}_a^{\mathrm{T}}\psi$, $\tilde{W}_a=W_a-\hat{W}_a$, error_t 的定义将在证明中给出.

更进一步地, 如果 $\kappa_{c,\,m}$, $\bar{\kappa}_{c,\,m}$, $\kappa_{a,\,m}\to 0$, 那么 $\tilde{W}_c$ 和 ξ_a 将渐近收敛至 0.

证明. 定义如下 Lyapunov 函数

$$L(l)=L_c(l)+L_a(l)$$

其中, $L_c=\mathrm{tr}(\tilde{W}_c^{\mathrm{T}}\eta_c^{-1}\tilde{W}_c)$, $L_a=\mathrm{tr}(\tilde{W}_a^{\mathrm{T}}\eta_a^{-1}\tilde{W}_a)$. 根据式 (20), 可计算

$$\begin{aligned}E(l)=&\hat{W}_c^{\mathrm{T}}\phi(l)-\hat{W}_c^{\mathrm{T}}\phi(l+1)+\Delta V^*(l+1)=\\&\tilde{W}_c^{\mathrm{T}}\Delta\phi(l+1)+\Delta\kappa_c(l+1)\end{aligned} \tag{27}$$

其中, $\Delta V^*(l+1)=V^*(l+1)-V^*(l)$, $\Delta\kappa_c(l+1)=\kappa_c(l+1)-\kappa_c(l)$,

$$\begin{aligned}E_f=&\hat{W}_c^{\mathrm{T}}\phi_f-W_c^{\mathrm{T}}\phi_f-\kappa_{c,\,f}=\\&-\tilde{W}_c^{\mathrm{T}}\phi_f-\kappa_{c,\,f}\end{aligned} \tag{28}$$

其中, $\kappa_{c,\,f}=\kappa_c(k+N)$. 则根据式 (21), (27), (28), 可得

$$\begin{aligned}\Delta L_c(l+1)=&L_c(l+1)-L_c(l)=\\&2\tilde{W}_c^{\mathrm{T}}(-\tilde{\phi}\tilde{W}_c+\bar{\kappa}_c)+\\&\eta_c(-\tilde{\phi}\tilde{W}_c+\bar{\kappa}_c)^{\mathrm{T}}(-\tilde{\phi}\tilde{W}_c+\bar{\kappa}_c)\le\\&-\alpha\|\tilde{W}_c\|^2+\mathrm{error}_c\end{aligned}$$

其中, $\bar{\kappa}_c=-\Delta\phi(l+1)\Delta\kappa_c(l+1)-\phi_f\kappa_{c,\,f}$, $\mathrm{error}_c=(2\eta_c+1/\beta_0)\|\bar{\kappa}_c\|^2$.

类似地, $\Delta L_a(l+1)$ 可以表示为

$$\Delta L_a(l+1)=\mathrm{tr}\left(2\tilde{W}_a^{\mathrm{T}}(l)\frac{\partial E_a^2(l)}{\partial\hat{W}_a(l)}+\eta_a\left(\frac{\partial E_a^2(l)}{\partial\hat{W}_a(l)}\right)^{\mathrm{T}}\frac{\partial E_a^2(l)}{\partial\hat{W}_a(l)}\right)$$

考虑到 $\frac{\partial E_a^2}{\partial\hat{W}_a}=2\psi E_a$, 以及 $E_a=-\xi_a-g\tilde{W}_c+\bar{\kappa}_a$, $g=g_1\nabla\phi$, $g_1=\frac{1}{2}R^{-1}B_1^{\mathrm{T}}$, $\bar{\kappa}_a=-\kappa_a-g_1\nabla\kappa_c$, 那么

$$\begin{aligned}\Delta L_a=&-(4-4\bar{\psi}\eta_a)\|\xi_a\|^2-8\bar{\psi}\eta_a g\tilde{W}_c\bar{\kappa}_a-\\&(4-8\bar{\psi}\eta_a)\xi_a^{\mathrm{T}}g\tilde{W}_c+4\bar{\psi}\eta_a\|\tilde{W}_c\|_{\bar{g}}^2+\\&(4-8\bar{\psi}\eta_a)\xi_a\bar{\kappa}_a\end{aligned}$$

其中, $\bar{g}=g^{\mathrm{T}}g$. 应用 Young 不等式定理, 可得

$$\Delta L_a(l+1)\le-\gamma_1\|\xi_a\|^2+\gamma_2\|\tilde{W}_c\|_{\bar{g}}^2+\mathrm{error}_a$$

其中, $\mathrm{error}_a=(1/\beta_2+1/\beta_3)\|\bar{\kappa}_a\|^2$. 因此, 考虑到

$$\begin{aligned}\mathrm{error}_c\le&\left(2\eta_c+\frac{1}{\beta_0}\right)\times\\&(2q_2\kappa_{c,\,m}+\phi_m\kappa_{c,\,m})^2=\mathrm{error}_{c,\,m}\end{aligned}$$

$$\begin{aligned}\mathrm{error}_a\le&\left(\frac{1}{\beta_2}+\frac{1}{\beta_3}\right)\times\\&(\kappa_{a,\,m}+\|g_1\|\bar{\kappa}_{c,\,m})^2=\mathrm{error}_{a,\,m}\end{aligned}$$

那么定义 $\mathrm{error}_t=\mathrm{error}_{c,\,m}+\mathrm{error}_{a,\,m}$, 可以得到

$$\Delta L=-\gamma_1\|\xi_a\|^2-(\alpha-\gamma_2)\|\tilde{W}_c\|_{\bar{g}}^2+\mathrm{error}_t \tag{29}$$

因此, 可以得到结论 (26). 在此基础上, 如果 $\kappa_{c,\,m}$, $\bar{\kappa}_{c,\,m}$, $\kappa_{a,\,m}\to 0$, 可得 $\mathrm{error}_t\to 0$, 那么 $\tilde{W}_c$ 和 ξ_a 渐近收敛至 0. □

注 3. 定理 2 的结论表明, 可以通过增加执行器和评价器的基函数节点数使得 u 能够以任意小误差收敛至 u_b^*. 因此, 在假设 1 成立的前提下, 如果选择预测时域 N 足够大[38], 使得系统 (9) 在预测时域 $[k,\,k+N-1]$ 内由控制策略 $u_b^*(k|k),\cdots,u_b^*(k+N-1|k)$ 驱动下满足终端状态 $e(k+N)\in\mathcal{X}_f$, 那么, 在下一个预测时域 $[k+1,\,k+N]$, $u_b^*(k+1|k),\cdots,u_b^*(k+N-1|k)$, $Ke(k+N|k)$ 是一个可行的控制策略. 我们定义由上述可行策略产生的损失函数为 $V^f(k+1|k)$, 并参考文献 [39] 的证明思路, 可得 $V^f(k+1|k)-V^*(k|k)\le-L\left(e(k|k),\,u_b(k|k)\right)$. 由于 $Ke(k+N|k)$ 是次优的, 我们可以得出 $V^*(k+1|k+1)-V^*(k|k)\le V^f(k+1|k)-V^*(k|k)\le-L\left(e(k|k),\,u_b(k|k)\right)$, 从而可以借助李雅普诺夫稳定性分析得到闭环系统的稳定性. 对上述分析的详细推导过程可以参考文献 [37−39], 由于篇幅限制, 这里不再赘述. 至于学习逼近得到的策略存在较大误差的情况, 我们将在以后的研究中借助鲁棒 MPC[40−41] 的思想进一步分析和证明.

3　仿真和实车实验验证

在本节中通过仿真和实车实验验证本文提出的 RHRL 算法的控制性能.

3.1　仿真验证结果

在控制器设计中车辆的相关参数设置如表 1 所

示, 本文在如图 4 所示的道路环境下进行了仿真实验, 图 4 中, 黑色实线表示道路边界, 黑色点划线表示道路中心线, 红色实线表示期望的参考路径, 蓝色边框表示初始位置下的智能车辆.

表 1 车辆动力学参数

Table 1 The parameters of the vehicle dynamics

符号	物理意义	数值	单位
m	车身质量	1723	kg
I_z	转动惯量	4175	kg·m^2
l_f	质心到前轴距离	1.232	m
l_r	质心到后轴距离	1.468	m
C_f	前轮侧偏刚度	66900	N/rad
C_r	后轮侧偏刚度	62700	N/rad

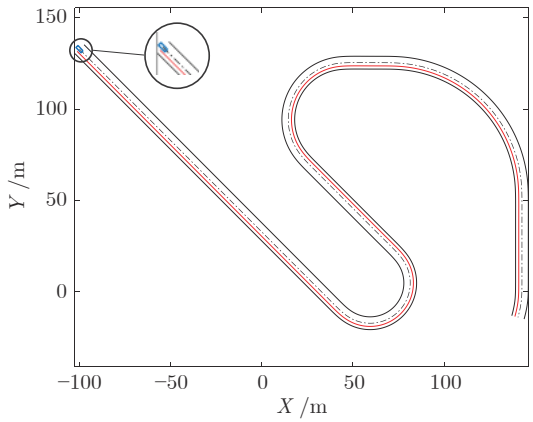

图 4 参考路径

Fig. 4 Reference path

在控制过程中, 各偏差状态需要满足如下约束: $e_y \in [-5\ \text{m}, 5\ \text{m}]$, $\dot{e}_y \in [-10\ \text{m/s}, 10\ \text{m/s}]$, $e_\varphi \in [-\pi/3\ \text{rad}, \pi/3\ \text{rad}]$, $\dot{e}_\varphi \in [-\pi\ \text{rad/s}, \pi\ \text{rad/s}]$. 在设计基于 RHRL的侧向运动控制方法时, 选取 $Q = I_4$, $R = 1$, $N = 50$, $\Delta t = 0.02$ s. 将执行器网络的基函数向量 $\psi(e)$ 选取为 $\psi(e) = [e_1^2, e_2^2, e_3^2, e_4^2, e_1e_2, e_1e_3, e_1e_4, e_2e_3, e_2e_4, e_3e_4]^{\mathrm{T}}$, 评价器网络的基函数向量 $\phi(e)$ 选取为 $\phi(e) = [e_1, e_2, e_3, e_4, e_1^2, e_2^2, e_3^2, e_4^2, e_1e_2, e_1e_3, e_1e_4, e_2e_3, e_2e_4, e_3e_4]^{\mathrm{T}}$, 其中, $[e_1, e_2, e_3, e_4] = [e_y, \dot{e}_y, e_\varphi, \dot{e}_\varphi]$. 在学习开始前, 网络权重 $\hat{W}_a$ 和 $\hat{W}_c$ 在 $[-1, 1]$ 之间随机初始化. 在学习过程中, 评价器和执行器网络的学习率分别设置为 $\eta_c = 0.08$, $\eta_a = 0.06$, 执行器和评价器的权值的更新方式为增量式. RHRL 每次训练的轮数设置为 5.

在仿真验证实验中, 主要对比了软执行器–评价器 (SAC) 算法[42]、深度确定性策略梯度 (DDPG)[43]、HDP 方法 (执行器–评价器结构与本文相同)、纯点预瞄方法[44] 和 MPC 控制方法. 在采用 SAC 和 DDPG 算法训练前, 利用本文构建的模型 (9) 生成 100 万个动作–状态 (u, e) 的数据对 (即样本) 用于离线训练. SAC 训练过程中的所有参数设置与文献 [42] 保持一致, 其训练中使用的样本数量级为 40 万个. DDPG 算法训练时的参数设置与文献 [43] 保持一致, 训练中使用的样本数量级为 40 万个. 在仿真实验中, 分别采用 SAC 和 DDPG 算法进行了 5 次重复训练, 每次训练的轮数为 2 000. 在训练完成后, 我们利用 5 次训练得到的执行器网络分别生成控制策略用于直接控制系统 (9), 并选取性能表现最好的一组数据与 RHRL 对比. 由于 HDP 对比算法的执行器–评价器结构与本文相同, 其控制器参数设置、仿真测试设计与 RHRL 算法保持一致; 其权值训练方式为增量式、训练轮数为 30. 对于纯点预瞄方法, 根据文献 [44], 可以得到相应的控制器表达式为 $\delta(t) = \arctan\left(2\left(l_f + l_r\right)\sin(\theta(t))/l_d\right)$, 其中, l_d 是控制器的预瞄距离, 一般与车速相关, 仿真实验中设置 $l_d = 0.55v_x$; $\theta(t)$ 是车身和预瞄点之间的夹角. 在离散时间 MPC 控制器中, 我们设置参数 Q, R 与 RHRL 算法保持一致. 在纵向速度 v_x 分别为 30 km/h 和 50 km/h 下, 智能车在运行过程中的侧向误差和航向角偏差结果如图 5 和图 6 所示, 其均方根误差 (Root mean square error, RMSE) 如表 2 所示. 仿真结果显示, 本文提出的 RHRL 与 MPC 相比, 跟踪控制性能相当, 但在采用的 Inter (R) Core (TM) i7-7700HQ CPU @2.80 GHz 笔记本中, MPC (采用 QuadProg 求解器) 平均计算时间为 0.0397 s, 而 RHRL 的平均计算时间为 0.0160 s. 另外, RHRL 算法的控制性能在 30 km/h 和 50 km/h 下优于预瞄控制、HDP、SAC 和 DDPG. RHRL 算法的性能表现之所以优于深度强化学习算法 SAC

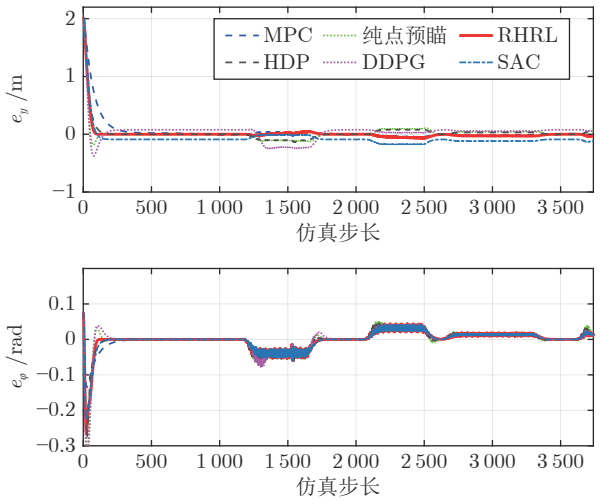

图 5 30 km/h 下智能车跟踪控制侧向偏差对比

Fig. 5 Comparison of lateral tracking error of intelligent vehicles under $v_x = 30$ km/h

表 2　各控制器的均方根误差对比
Table 2　The RMSE comparison among all the controllers

方法	$v_x = 30$ km/h		$v_x = 50$ km/h	
	e_y (m)	e_φ (rad)	e_y (m)	e_φ (rad)
RHRL	**0.156**	0.030	**0.246**	0.020
HDP	0.165	0.030	0.315	0.019
SAC	0.189	0.029	0.283	0.017
DDPG	0.172	0.037	0.319	0.017
MPC	0.212	**0.025**	0.278	**0.015**
纯点预瞄	0.159	0.036	0.286	0.030

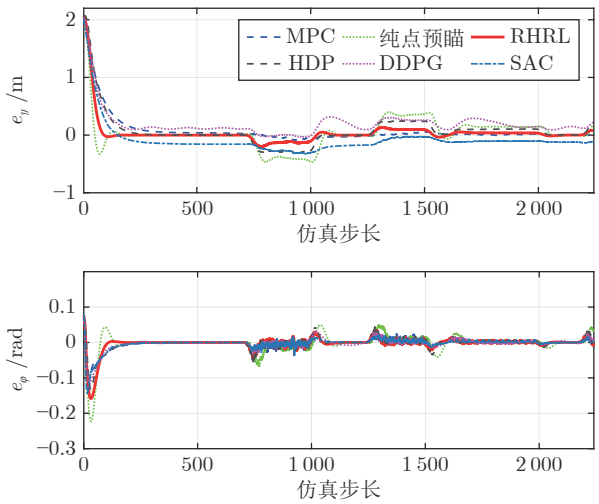

图 6　50 km/h 下智能车跟踪控制侧向偏差对比
Fig. 6　Comparison of lateral tracking error of intelligent vehicles under $v_x = 50$ km/h

和 DDPG, 其原因在于 RHRL 算法采用了滚动时域优化机制来提升学习效率, 并在每个预测时域利用模型信息产生预测; 而且, RHRL 算法的实现方式是在线同步增量式学习和部署.

3.2　结构化城市场景中的实验验证结果

为了更进一步验证 RHRL 在实际车辆系统控制问题中的有效性, 我们利用红旗 E-HS3 智能驾驶平台 (如图 7 所示) 首先在城市场景中进行实车实验. 在实验设计中, 采用离线仿真训练得到的权值作为初始权值. 其他参数设置, 如学习率、基函数等与仿真实验相同. 在实验过程中, RHRL 算法以 50 Hz 的工作频率, 通过在线学习不断优化策略以适应动态路面环境. RHRL 算法的在线增量式学习部署过程实现方式如下. 在每个学习 (计算) 时刻, 根据车辆装配的卫星和惯性组合导航系统 (如图 7 所示) 实时测量得到车辆状态信息对 (X, Y, v_x, v_y, φ, w), 由此在车载计算机 (工控机) 中计算当前误差状态信息 e. 在此基础上, 将求解得到的 e 的值作为初始状态值, 利用预测模型 (9) 在当前预测时域内实时更新执行器和评价器的权值. 接下来, 通过学习得到的执行器权值和前馈控制量求解得到当前的控制量 u, 也就是车辆前轮转角. 据此, 可以利用前轮转角和方向盘转角的经验比例关系计算得到当前时刻方向盘的期望转角为 $u_w = 15u$, 也就是车辆的控制量. 在后面的每个采样时刻, 通过不断重复上述步骤实现整个学习控制过程.

图 7　红旗 E-HS3 智能驾驶平台
Fig. 7　Hongqi E-HS3 intelligent driving platform

在实车实验中, 还与纯点预瞄控制方法进行了对比, 纯点预瞄控制的参数设置与仿真实验中相同. 对纯点预瞄方法进行测试时, 采用恒定的期望车辆速度, 为 20 km/h; 而对 RHRL 算法进行测试时, 令车辆始终跟踪当前期望的动态参考速度, 平均速度达到约 30 km/h, 最高速度达到 38.988 km/h. 图 8 为两种方法在用于控制实车后所生成的路径图; 图 9 展示了 RHRL 和纯点预瞄方法下红旗 E-HS3 的车辆侧向偏差. 实车实验结果表明, RHRL 算法的控制性能优于纯点预瞄控制算法.

需要指出的是, 预瞄方法由于采用的是动态预瞄距离的方法, 因此在车辆起步阶段由于惯导和较大侧向偏差的情况下, 智能车会产生较大的侧向偏差, 而 RHRL 却可以快速优化, 具有较小的侧向跟踪控制误差.

3.3　乡村起伏砂石道路中的实验验证结果

为了验证本文提出的算法对路面的适应能力, 我们还在乡村起伏砂石路面上进行了控制性能的验证, 其测试场景如图 10 所示. 车辆首先从 C 点出发, 经过 B 点所在的直角弯, 再行驶至终点 A. 在从 B 至 A 段的行驶过程中, 车辆首先要经过一个明显的下坡, 在终点附近需要经过一个狭窄的通道

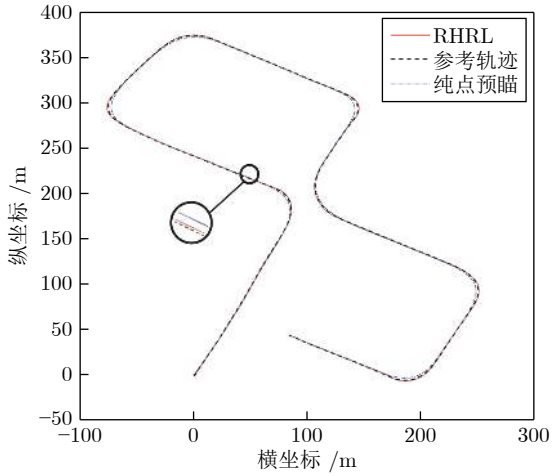

图 8 基于 RHRL 和纯点预瞄方法的红旗 E-HS3 行驶路径

Fig. 8 Path of Hongqi E-HS3 vehicle controlled by RHRL and pure pursuit methods

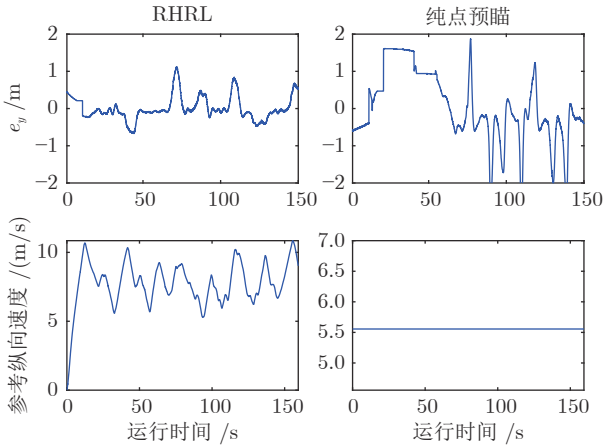

图 9 RHRL 与纯点预瞄方法的车辆实测侧向偏差对比

Fig. 9 Comparison of experimental lateral tracking error of the RHRL and pure pursuit methods

(由路桩铺设构成). 车辆在初始位置 C 点由静止状态出发, 在行驶中平均速度为 4.19 m/s, 最高速度为 4.94 m/s. 实验中车辆在不同行驶阶段的状态如图 10 所示, 其表明车辆能够在起伏砂石路面上实现平稳的转弯和下坡, 而且还实现了狭窄通道下的高精度控制 (如图 11 所示).

4 结束语

提出了一种基于滚动时域强化学习的智能驾驶车辆侧向控制算法. 该算法将强化学习与滚动时域优化机制融合, 把无限时域自学习优化问题转化为一系列有限时域优化问题, 并通过执行器–评价器算法进行求解. 该设计思想通过滚动时域机制提高

图 10 乡村砂石道路地图和车辆行驶中各阶段状态

Fig. 10 The route map in the country sand and gravel road, and the status of different stages in the control process

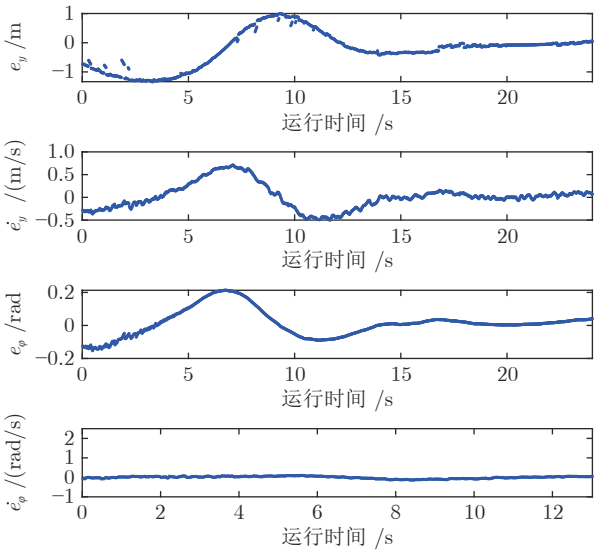

图 11 侧向误差曲线

Fig. 11 Curves of the lateral error

了强化学习算法的学习效率; 与 MPC 相比, 采用执行器–评价器的优化方式能够提高计算效率. 因此, 本文提出的 RHRL 可以看作是一种介于强化学习和 MPC 之间的控制算法. 此外, 与大多数已有的有限时域执行器–评价器学习算法不同, 本文提出的 RHRL 采用时间独立的网络结构, 降低了网络的设计和在线计算复杂度, 而且本文还从理论上分析了其在每个预测时域内的收敛性以及闭环系统的稳定性. 在仿真场景中与典型传统算法和深度强化学习算法的对比实验结果验证了 RHRL 算法的有效性. 另外, 从结构化道路场景中的实验结果可以看出, 即使在变速控制条件下, RHRL 依然比恒速条件下

的纯点预瞄控制方法具有更好的控制性能. 从乡村起伏砂石道路中的实际实验结果可以看出, RHRL 具有良好的路面适应能力和控制性能.

References

1 Xiong Lu, Yang Xing, Zhuo Gui-Rong, Leng Bo, Zhang Ren-Xie. Review on motion control of autonomous vehicles. *Journal of Mechanical Engineering*, 2020, **56**(10): 127−143
(熊璐, 杨兴, 卓桂荣, 冷搏, 章仁燮. 无人驾驶车辆的运动控制发展现状综述. 机械工程学报, 2020, **56**(10): 127−143)

2 Chen Hong, Guo Lu-Lu, Gong Xun, Gao Bing-Zhao, Zhang Lin. Automative control in intelligent era. *Acta Automatica Sinica*, 2020, **46**(7): 1313−1332
(陈虹, 郭露露, 宫洵, 高炳钊, 张琳. 智能时代的汽车控制. 自动化学报, 2020, **46**(7): 1313−1332)

3 Tian Tao-Tao, Hou Zhong-Sheng, Liu Shi-Da, Deng Zhi-Dong. Model-free adaptive control based lateral control of self-driving car. *Acta Automatica Sinica*, 2017, **43**(11): 1931−1940
(田涛涛, 侯忠生, 刘世达, 邓志东. 基于无模型自适应控制的无人驾驶汽车横向控制方法. 自动化学报, 2017, **43**(11): 1931−1940)

4 Ahmed A A, Alshandoli A F S. Using of neural network controller and fuzzy PID control to improve electric vehicle stability based on A14-DOF model. In: Proceedings of the International Conference on Electrical Engineering (ICEE). Istanbul, Turkey: IEEE, 2020. 1−6

5 Meng J, Liu A B, Yang Y Q, Wu Z, Xu Q Y. Two-wheeled robot platform based on PID control. In: Proceedings of the 5th International Conference on Information Science and Control Engineering (ICISCE). Zhengzhou, China: IEEE, 2018. 1011−1014

6 Farag W. Complex trajectory tracking using PID control for autonomous driving. *International Journal of Intelligent Transportation Systems Research*, 2020, **18**(2): 356−366

7 Zhao P, Chen J J, Song Y, Tao X, Xu T J, Mei T. Design of a control system for an autonomous vehicle based on adaptive-PID. *International Journal of Advanced Robotic Systems*, 2012, **9**(2): Article No. 44

8 Han G N, Fu W P, Wang W, Wu Z S. The lateral tracking control for the intelligent vehicle based on adaptive PID neural network. *Sensors*, 2017, **17**(6): Article No. 1244

9 Fraichard T, Garnier P. Fuzzy control to drive car-like vehicles. *Robotics and Autonomous Systems*, 2001, **34**(1): 1−22

10 Pérez J, Milanés V, Onieva E. Cascade architecture for lateral control in autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2011, **12**(1): 73−82

11 Li H M, Wang X B, Song S B, Li H. Vehicle control strategies analysis based on PID and fuzzy logic control. *Procedia Engineering*, 2016, **137**: 234−243

12 Park M W, Lee S W, Han W Y. Development of lateral control system for autonomous vehicle based on adaptive pure pursuit algorithm. In: Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS). Gyeonggi-do, South Korea: IEEE, 2014. 1443−1447

13 Guo Jing-Hua, Hu Ping, Li Lin-Hui, Wang Rong-Ben, Zhang Ming-Heng, Guo Lie. Study on lateral fuzzy control of unmanned vehicles via genetic algorithms. *Chinese Journal of Mechanical Engineering*, 2012, **48**(6): 76−82
(郭景华, 胡平, 李琳辉, 王荣本, 张明恒, 郭烈. 基于遗传优化的无人车横向模糊控制. 机械工程学报, 2012, **48**(6): 76−82)

14 Leonard J, How J, Teller S, Berger M, Campbell S, Fiore G, et al. A perception-driven autonomous urban vehicle. *Journal of Field Robotics*, 2008, **25**(10): 727−774

15 Rajamani R, Zhu C, Alexander L. Lateral control of a backward driven front-steering vehicle. *Control Engineering Practice*, 2003, **11**(5): 531−540

16 Thrun S, Montemerlo M, Dahlkamp H, Stavens D, Aron A, Diebel J, et al. Stanley: The robot that won the DARPA grand challenge. *Journal of Field Robotics*, 2006, **23**(9): 661−692

17 Gong Jian-Wei, Jiang Yan, Xu Wei. *Model Predictive Control for Self-Driving Vehicles*. Beijing: Beijing Institute of Technology Press, 2014.
(龚建伟, 姜岩, 徐威. 无人驾驶车辆模型预测控制. 北京: 北京理工大学出版社, 2014.)

18 Falcone P, Borrelli F, Asgari J, Tseng H E, Hrovat D. Predictive active steering control for autonomous vehicle systems. *IEEE Transactions on Control Systems Technology*, 2007, **15**(3): 566−580

19 Carvalho A, Gao Y Q, Gray A, Tseng H E, Borrelli F. Predictive control of an autonomous ground vehicle using an iterative linearization approach. In: Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC). The Hague, The Netherlands: IEEE, 2013. 2335−2340

20 Beal C E, Gerdes J C. Model predictive control for vehicle stabilization at the limits of handling. *IEEE Transactions on Control Systems Technology*, 2013, **21**(4): 1258−1269

21 Liniger A, Domahidi A, Morari M. Optimization-based autonomous racing of 1:43 scale RC cars. *Optimal Control Applications and Methods*, 2015, **36**(5): 628−647

22 Kabzan J, Hewing L, Liniger A, Zeilinger M N. Learning-based model predictive control for autonomous racing. *IEEE Robotics and Automation Letters*, 2019, **4**(4): 3363−3370

23 Ostafew C J, Schoellig A P, Barfoot T D. Robust constrained learning-based NMPC enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 2016, **35**(13): 1547−1563

24 You Zhi-Heng. Research on Model Predictive Control-based Trajectory Tracking for Unmanned Vehicles [Master thesis], Jilin University, China, 2018.
(由智恒. 基于 MPC 算法的无人驾驶车辆轨迹跟踪控制研究 [硕士学位论文], 吉林大学, 中国, 2018.)

25 Wang Ding. Research progress on learning-based robust adaptive critic control. *Acta Automatica Sinica*, 2019, **45**(6): 1031−1043
(王鼎. 基于学习的鲁棒自适应评判控制研究进展. 自动化学报, 2019, **45**(6): 1031−1043)

26 Wang D, Ha M M, Qiao J F. Data-driven iterative adaptive critic control toward an urban wastewater treatment plant. *IEEE Transactions on Industrial Electronics*, 2021, **68**(8): 7362−7369

27 Oh S Y, Lee J H, Choi D H. A new reinforcement learning vehicle control architecture for vision-based road following. *IEEE Transactions on Vehicular Technology*, 2000, **49**(3): 997−1005

28 Yang Hui-Yuan. Reinforcement Learning-Based Optimal Control Methods with Applications to Mobile Robots [Master thesis], National University of Defense Technology, China, 2014.
(杨慧媛. 基于增强学习的优化控制方法及其在移动机器人中的应用 [硕士学位论文], 国防科学技术大学, 中国, 2014.)

29 Lian Chuan-Qiang. Optimization Control Methods Based on Approximate Dynamic Programming and Its Applications in Autonomous Land Vehicles [Ph.D. dissertation], National University of Defense Technology, China, 2016.
(连传强. 基于近似动态规划的优化控制方法及在自主驾驶车辆中的应用 [博士学位论文], 国防科学技术大学, 中国, 2016.)

30 Huang Zhen-Hua. Researches on Adaptive Critic Learning Control Approaches for Intelligent Driving Vehicles [Ph.D. dissertation], National University of Defense Technology, China, 2017.
(黄振华. 智能驾驶车辆自评价学习控制方法研究 [博士学位论文], 国防科学技术大学, 中国, 2017.)

31 Lian C Q, Xu X, Chen H, He H B. Near-optimal tracking control of mobile robots via receding-horizon dual heuristic programming. *IEEE Transactions on Cybernetics*, 2016, **46**(11): 2484−2496

32 Kuutti S, Bowden R, Jin Y C, Barber P, Fallah S. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 2021, **22**(2): 712−733

33 Li D, Zhao D B, Zhang Q C, Chen Y R. Reinforcement learning and deep learning based lateral control for autonomous driving [Application Notes]. *IEEE Computational Intelligence Magazine*, 2019, **14**(2): 83−98

34 Chen Y X, Hereid A, Peng H E, Grizzle J. Enhancing the performance of a safe controller via supervised learning for truck lateral control. *Journal of Dynamic Systems, Measurement, and Control*, 2019, **141**(10): Article No. 101005

35 Dong L, Yan J, Yuan X, He H B, Sun C Y. Functional nonlinear model predictive control based on adaptive dynamic programming. *IEEE Transactions on Cybernetics*, 2019, **49**(12): 4206−4218

36 Rajamani R. *Vehicle Dynamics and Control* (Second edition). New York: Springer, 2012.

37 Xu X, Chen H, Lian C Q, Li D Z. Learning-based predictive control for discrete-time nonlinear systems with stochastic disturbances. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(12): 6202−6213

38 Chmielewski D, Manousiouthakis V. On constrained infinite-time linear quadratic optimal control. *Systems and Control Letters*, 1996, **29**(3): 121−129

39 Rawlings J B, Mayne D Q, Diehl M M. *Model Predictive Control: Theory, Computation, and Design* (*Second edition*). Madison: Nob Hill Publishing, 2017.

40 Mayne D Q, Kerrigan E C, van Wyk E J, Falugi P. Tube-based robust nonlinear model predictive control. *International Journal of Robust and Nonlinear Control*, 2011, **21**(11): 1341−1353

41 Zhang X L, Pan W, Scattolini R, Yu S Y, Xu X. Robust tube-based model predictive control with Koopman operators. *Automatica*, 2022, **137**: Article No. 110114

42 Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden: PMLR, 2018. 1856−1865

43 Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. In: Proceedings of the 4th International Conference on Learning Representations (ICLR). San Juan, Puerto Rico: 2016.

44 Snider J M. Automatic Steering Methods for Autonomous Automobile Path Tracking, Technical Report CMU-RI-TR-09-08, Robotics Institute, Carnegie Mellon University, USA, 2009.

张兴龙 国防科技大学智能科学学院副研究员. 2018 年获得意大利米兰理工大学博士学位. 主要研究方向为滚动时域强化学习及其在无人系统中的应用.
E-mail: zhangxinglong18@nudt.edu.cn
(**ZHANG Xing-Long** Associate professor at the College of Intelligence Science and Technology, National University of Defense Technology. He received his Ph.D. degree from Politecnico di Milano, Italy, in 2018. His research interest covers receding horizon reinforcement learning and its application in unmanned systems.)

陆 阳 国防科技大学智能科学学院博士研究生. 2020 年获得国防科技大学硕士学位, 2018 年获得山东大学学士学位. 主要研究方向为强化学习及其在无人系统中的应用.
E-mail: luyang18@nudt.edu.cn
(**LU Yang** Ph.D. candidate at the College of Intelligence Science and Technology, National University of Defense Technology. He received his master and bachelor degrees from National University of Defense Technology in 2020 and Shandong University in 2018, respectively. His research interest covers reinforcement learning and its application in unmanned systems.)

李文璋 2018 年获得北京理工大学学士学位, 2020 年获得国防科技大学硕士学位. 主要研究方向为智能车学习控制.
E-mail: 15624953231@163.com
(**LI Wen-Zhang** Received his bachelor and master degrees from National University of Defense Technology in 2018 and Beijing Institute of Technology in 2020, respectively. His research interest covers learning control of intelligent vehicles.)

徐 昕 国防科技大学智能科学学院研究员. 2002 年获得国防科技大学机电与自动化学院控制科学与工程博士学位. 主要研究方向为智能控制, 强化学习, 近似动态规划, 机器学习, 机器人和智能驾驶. 本文通信作者.
E-mail: xinxu@nudt.edu.cn
(**XU Xin** Professor at the College of Intelligence Science and Technology, National University of Defense Technology. He received his Ph.D. degree in control science and engineering from the College of Mechatronics and Automation, National University of Defense Technology in 2002. His research interest covers intelligent control, reinforcement learning, approximate dynamic programming, machine learning, robotics, and autonomous vehicles. Corresponding author of this paper.)