中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于深度强化学习的网联车辆队列纵向控制

李永福, 周发涛, 黄龙旺, 于树友, 施树明

引用本文:

李永福,周发涛,黄龙旺,于树友,施树明. 基于深度强化学习的网联车辆队列纵向控制[J]. 控制与决策, 2024, 39(6): 1879−1887.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.2094

基于深度强化学习的网联车辆队列纵向控制

李永福[1†], 周发涛[1], 黄龙旺[1], 于树友[2], 施树明[3]

(1. 重庆邮电大学 智能空地协同控制重庆市高校重点实验室，重庆 400065；
2. 吉林大学 控制科学与工程系，长春 130012；3. 吉林大学 交通学院，长春 130012)

摘 要: 针对车辆队列中多目标控制优化问题,研究基于强化学习的车辆队列控制方法. 控制器输入为队列各车辆状态信息以及车辆间状态误差,输出为基于车辆纵向动力学的期望加速度,实现在V2X通信下的队列单车稳定行驶和队列稳定行驶. 根据队列行驶场景以及采用的间距策略、通信拓扑结构等特性,建立队列马尔科夫决策过程(Markov decision process, MDP)模型. 同时根据队列多输入-多输出高维样本特性,引入优先经验回放策略,提高算法收敛效率. 为贴近实际车辆队列行驶工况,仿真基于PreScan构建多自由度燃油车动力学模型,联合Matlab/Simulink搭建仿真环境,同时引入噪声对队列控制器中动作网络和评价网络进行训练. 仿真结果表明基于强化学习的车辆队列控制燃油消耗更低,且控制器实时性更高,对车辆的控制更为平滑.

关键词: 车辆队列；深度确定性策略梯度；强化学习；燃油消耗；V2X；PreScan

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.2094

引用格式: 李永福, 周发涛, 黄龙旺, 等. 基于深度强化学习的网联车辆队列纵向控制[J]. 控制与决策, 2024, 39(6): 1879-1887.

Longitudinal control of connected vehicle platoon based on deep reinforcement learning

LI Yong-fu[1†], *ZHOU Fa-tao*[1], *HUANG Long-wang*[1], *YU Shu-you*[2], *SHI Shu-ming*[3]

(1. Key Laboratory of Intelligent Air-Ground Cooperative Control for Universities in Chongqing，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；2. Department of Control Science and Engineering，Jilin University，Changchun 130012，China；3. Transportation College，Jilin University，Changchun 130012，China)

Abstract: This paper presents a vehicle platoon control method based on reinforcement learning (RL) to solve the multi-objective optimization problem. The actor network is designed to receive the state information of each vehicle in the platoon and the inter-vehicle state error, and outputs the desired acceleration based on the longitudinal dynamics of the vehicle. The proposed approach ensures both the individual vehicle stability and the string stability of the platoon under V2X communication. To model the platoon driving scenario with the spacing policy and communication topology, the Markov decision process (MDP) model of the platoon is established. In addition, considering the multi-input and multi-output high-level sample characteristics of the platoon, the deep deterministic policy gradient (DDPG) algorithm is adopted with the priority experience replay strategy to improve the convergence efficiency. To better approximate the actual platoon vehicle fuel consumption, the simulation is based on PreScan to build a high-degree fuel vehicle dynamics model. A co-simulation environment is created using Matlab/Simulink to train the actor network and critic network in the platoon controller by adding noise. The simulation results demonstrate that the reinforcement learning-based vehicle platoon control approach reduces fuel consumption and achieves faster and smoother vehicle control.

Keywords: vehicle platoon；deep deterministic policy gradient；reinforcement learning；fuel consumption；V2X；PreScan

0 引 言

车辆队列控制是指队列行驶过程中各车车速、车间距等状态保持一致,该研究对于减少能源消耗、增强道路行车安全、提升道路通行效率具有重要意义[1]. 现有对于车辆队列控制算法的研究大都集中于经典控制理论,以分布式控制方式完成基于车

收稿日期: 2022-12-03；**录用日期:** 2023-03-12.

基金项目: 国家自然科学基金项目(U1964202, 62273027)；重庆市自然科学基金创新发展联合基金项目(CSTB2022NSCQ-LZX0025)；重庆市教育委员会科学技术研究项目(KJZD-M202300602).

责任编委: 郭戈.

†通讯作者. E-mail: laf1212@163.com.

辆纵向动力学的控制器设计和分析[2-3]. 文献[4]在自适应巡航控制(adaptive cruise control, ACC)基础上引入车-车(vehicle-to-vehicle, V2V)通信技术,实现了协同式自适应巡航控制(cooperative adaptive cruise control, CACC),减少队列中由于前车速度变化引起的振荡. 文献[5]基于CACC系统提出一种线性前馈控制算法,提升了车辆在加减速时的安全性. 文献[6]提出了一种基于自适应容错控制(adaptive fault-tolerant control)的异质车辆队列控制算法,以解决非线性车辆动力学的执行器饱和问题. 文献[7]基于滑模控制(sliding model control)建立分布式框架解决了车辆队列轨迹优化与控制问题. 此类典型线性或非线性控制器,其控制器增益参数往往需要手动调整,对于在多个控制目标的影响下鲁棒性较差. 文献[8]提出一种基于模型预测控制(model predictive control, MPC)的异质车辆队列控制算法,证明了在车辆动力学特性未知情况下的队列串稳定性. 文献[9]结合分布式模型预测(distributed model predictive control, DMPC)实现针对单向拓扑结构异质车辆队列控制,并对算法进行了仿真验证. MPC控制器通过其参数估计以及约束控制特性,能够将队列控制的多个目标以成本函数的方式平衡,但MPC优化过程计算复杂度较大,实时性较弱.

随着深度学习(deep learning, DL)的发展,强化学习(reinforcement learning, RL)辅以深度神经网络求解最优动作的方式,为解决不确定环境下控制问题提供了一种新的技术路线[10-11],且在目前已有利用深度强化学习(feep reinforcement learning, DRL)进行车辆控制的一些进展. 文献[12]提出了一种遗传算法(genetic algorithm, GA)与DRL结合用于自动驾驶车辆(autonomous vehicles, AVs)形成智能车队的算法,利用DRL解决高维度动态环境的复杂性计算,结合GA解决DRL收敛速度慢的问题,最后控制不同车队的领航车实现AVs智能车队的形成、 控制和管理. 文献[13]提出一种基于DRL的车辆自主避碰决策控制模型,该模型以安全性、舒适性以及效率为奖励影响因素,利用深度确定性策略梯度(deep deterministic policy gradient, DDPG)建立了端到端的车辆自主避碰模型. 文献[14]研究基于DRL的车辆跟随决策控制算法,通过改进DDPG算法原有的经验回放机制,提升算法性能,验证了算法在不同测试环境下依旧能保持车辆跟随决策. 文献[15]在使用PID控制实现车辆队列纵向跟踪控制的基础上,利用DDPG算法求解PID控制参数,在保证队列串稳定性的同时实现了相比于单PID控制更小的跟踪误差. 上述基于DRL在车辆决策控制上的研究多建立在运动学模型或较低自由度的动力学模型上,且在整个车辆队列多输入变量-多输出变量的控制上较少应用.

本文旨在实现车辆队列控制并满足队列稳定性,主要贡献如下: 1)基于DDPG算法提出了一种新的车辆队列控制方法; 2)基于构建的车辆动力学,通过设计合理的多目标奖励函数实现队列的多输入-多输出控制; 3)通过引入优先经验回放策略提高算法的收敛效率,并在多场景下验证改进后算法的有效性.

1 问题描述

1.1 场景描述

如图1所示场景,考虑由1辆领航车和N辆跟随车组成的匀质队列进行控制器设计和分析. 车辆队列通信拓扑采取广泛使用的前车-领航车跟随式(predecessor-leader follower type, PLF)拓扑[2-9]. 假设队列中车辆均可通过V2V通信或传感器获取领航车广播信息和前车位置、速度信息[16]. 基于此假设,队列中跟随车辆获取的状态信息可以分为全局信息(global)和本地信息(local). 同时,本文主要关注于提出一种DRL控制框架,寻求最优的车辆队列控制策略,对于通信失败问题并未考虑. 控制目标表示如下:

$$\begin{cases} \lim\limits_{t\to\infty} \|p_{i-1}(t)-p_i(t)-d_{\rm des}\|=0,\\ \lim\limits_{t\to\infty} \|v_{i-1}(t)-v_i(t)\|=0,\\ \lim\limits_{t\to\infty} \|a_{i-1}(t)-a_i(t)\|=0. \end{cases} \tag{1}$$

其中: $p_i(t)$、$v_i(t)$、$a_i(t)$为车辆i在t时刻的位置、速度、加速度, $d_{\rm des}$为相邻两车之间的期望间距. 同时,采用固定时距策略(CTH)决定队列的几何构型[16].

1.2 车辆动力学模型

区别于传统车辆控制采用线性或非线性简化动力学模型[12-15],本文采用如图2所示的传动系统构建下层车辆纵向动力学模型. 下层动力需求主要克服行驶时空气阻力、滚动阻力、加速阻力,表示如下:

$$F_{\rm tr}=F_{\rm aero}+F_{\rm rr}+F_G, \tag{2}$$

$$F_{\rm aero}=2^{-1}\rho_{\rm air}C_dA_fv^2, \tag{3}$$

$$F_{\rm rr}=mgC_{\rm rr}\cos\alpha, \tag{4}$$

$$F_G=m_{\rm in}a\sin\alpha. \tag{5}$$

其中: $F_{\rm tr}$为车辆牵引力, $\rho_{\rm air}C_d$为空气阻力系数, A_f为迎风面积, v为车辆风中速度, m为车辆质量, g为重力系数, $m_{\rm in}$为车辆惯性质量, $C_{\rm rr}$为滚动阻力系数, α为道路坡度.

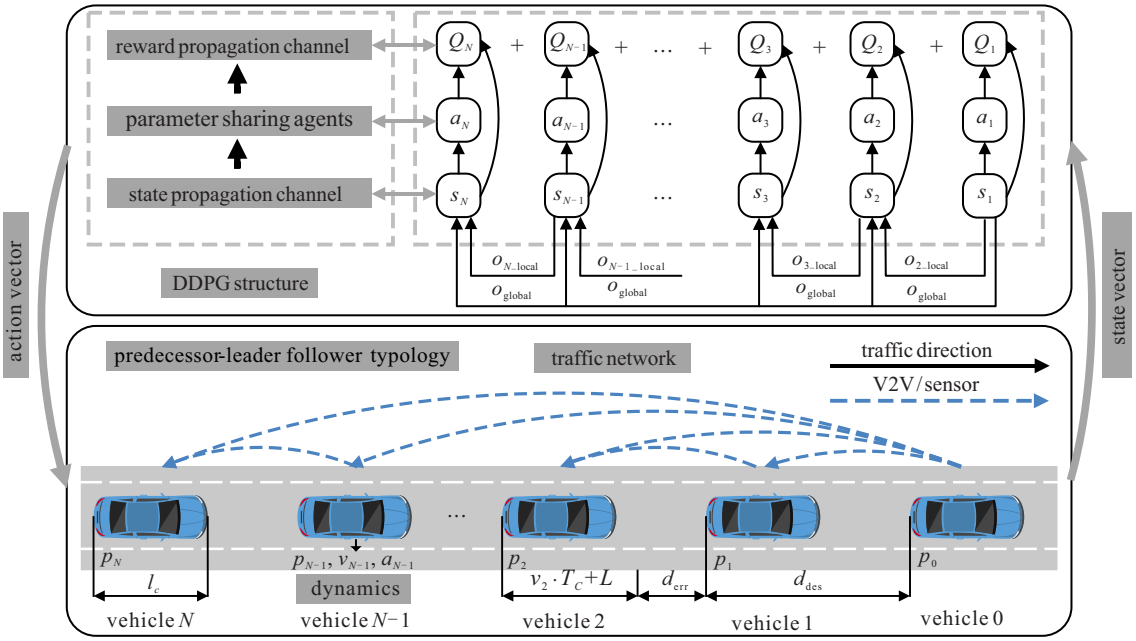


图 1　队列行驶场景

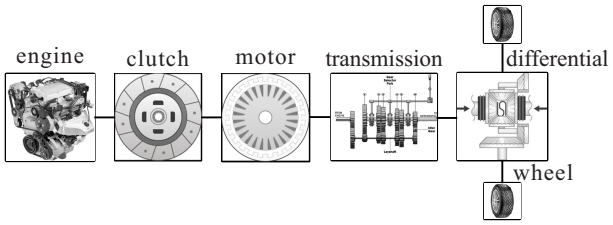


图 2　车辆传动结构

同时, 引擎采用准静态模型, 通过瞬时引擎转速和转矩插值获得燃油消耗, 计算公式如下:

$$\text{Fuel} = \int_0^T m_f(T_{\text{eng}}, S_{\text{eng}})\mathrm{d}t. \tag{6}$$

其中: m_f 为瞬时油耗, T_{eng} 为引擎转矩, S_{eng} 为引擎转速.

2　算法设计与实现

2.1　传统控制算法

为方便后续对比分析 DRL 控制方法的性能, 选取传统车辆队列控制方法中 MPC 控制策略作为基准, 在相同场景下进行应用. 参考文献 [17] 分布式模型预测控制策略, 车辆队列状态空间模型表示如下:

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{7}$$

其中 A、B 为系数矩阵. 状态量与控制量表示如下:

$$x(t) = \begin{bmatrix} p_{i-1} - p_i - (v_i T_C + L) \\ v_{i-1} - v_i \end{bmatrix},\ u(t) = a_i. \tag{8}$$

其目标函数为

$$J = \sum_{k=t}^{t+Y_N} \{x(k)'Wx(k) + u(k)'Iu(k) + \omega f(k)_v\}. \tag{9}$$

其中: W、I 为加权系数矩阵, f 为终端约束. 其状态约束为

$$\begin{aligned} &0 \leqslant v \leqslant v_{\max},\ a_{\min} \leqslant a \leqslant a_{\max}, \\ &p_{i-1} - p_i > l_c + L. \end{aligned} \tag{10}$$

其中: $a_{\min}$、$a_{\max}$、$v_{\max}$、L 分别为最小加速度、最大加速度、最大速度和 CTH 间距策略中最小静止距离; T_C、Y_N 分别为采用 CTH 间距策略时间常数以及 MPC 预测步长.

2.2　基于 DRL 控制器设计

相比传统控制方法, RL 是在没有任何先验知识的情况下, 在智能体 (agent) 通过与环境 (environment) 不断交互过程中获得对于执行动作反馈 (奖励或惩罚) 的方法, 以此迭代更新获取最优动作的策略. 由第 2.1 节可知, 车辆队列控制的状态转移过程可用马尔科夫决策过程 (Markov decision process, MDP)[18] 描述为

$$\begin{aligned} &P(s(t+1)) = \\ &E(S(t+1) = s(t+1) | S(t) = s(t), A(t) = a(t)), \end{aligned} \tag{11}$$

即当前状态转移到下一动作的概率只与当前状态和动作有关, 其中 $s(t)$、$a(t)$ 分别为当前时刻车辆队列行驶过程状态和动作的集合.

在状态转移过程中, 智能体获得的奖励 r_t 以及执行的策略 π 满足下式:

$$\pi^* = \arg\max_{a_t \in A} E\Big[\sum_{t=0}^{T} \gamma^t r(t)\Big]. \tag{12}$$

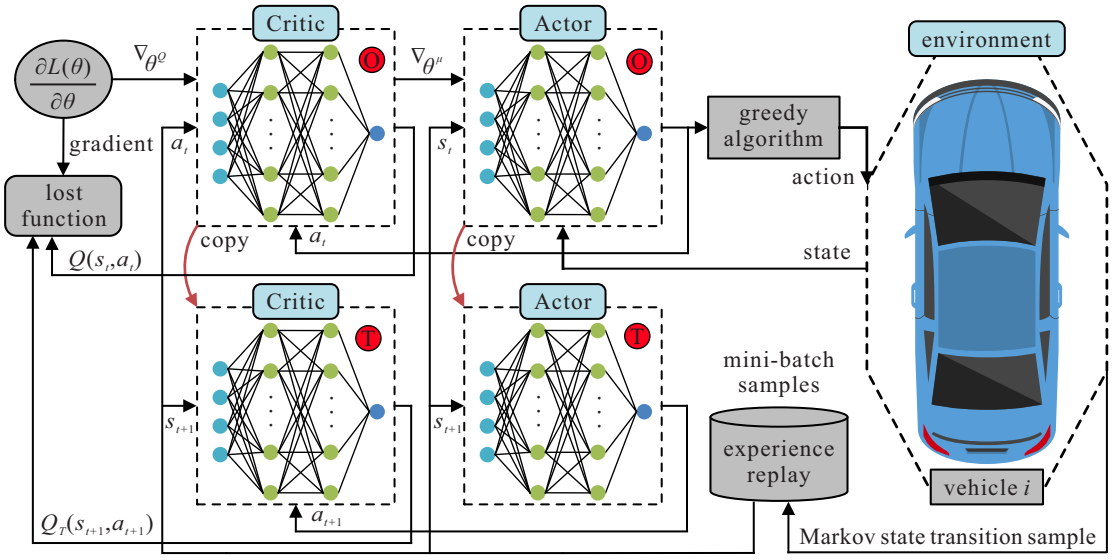


图 3 DDPG算法控制框架

其中: 折扣因子$\gamma \in (0,1)$, π^*为获取最大回报的期望策略.

DDPG是基于演员-评论家(actor-critic, AC)模型的算法[19], 算法控制框架如图3所示.

Actor网络负责选择动作输出, Critic网络负责评估动作的优劣, 两个网络的参数分别用θ^μ和θ^Q表示. 同时, 每个网络又分为当前网络(online)和目标网络(target), 有

$$\text{Actor : policy net}\begin{cases}\text{online : } \mu_\theta(s|\theta^\mu);\\ \text{target : } \mu_{\theta'}(s|\theta^{\mu'}).\end{cases} \tag{13}$$

$$\text{Critic : } Q \text{ net}\begin{cases}\text{online : } Q_\theta(s|\theta^Q);\\ \text{target : } Q_{\theta'}(s|\theta^{Q'}).\end{cases} \tag{14}$$

回合网络训练中, 从经验池M中随机选取若干序列$\{s(t),\ a(t), r(t), s(t+1)\}$作为小批量样本, 分别将$s(t)$和$s(t+1)$作为当前策略网络和目标策略网络的输入, 输出为$\mu(s(t)|\theta^\mu)$和$\mu'(s(t+1)|\theta^{\mu'})$, 用于计算当前和目标$Q$网络的状态价值函数. 将下一状态$s(t+1)$和$\mu'(s(t+1))$作为目标$Q$网络的输入, 得到目标$Q$网络的状态价值函数$Q'$, 计算其更新目标$y(t)$, 有

$$y(t) = r(t) + \gamma Q'(s(t+1), \mu'(s(t+1)|\theta^{\mu'})|\theta^{Q'}). \tag{15}$$

Critic网络部分利用最小化损失函数对参数θ^Q进行更新如下:

$$L(\theta^Q) = \frac{1}{N}\sum_t [y(t) - Q(s(t), a(t)|\theta^Q)]^2, \tag{16}$$

其中N为随机采样样本个数.

Actor网络部分根据确定性策略梯度求解方式对参数θ^μ进行更新如下:

$$\nabla_{\theta^\mu} J(\theta) = \frac{1}{N}\sum_t [\nabla_a Q(s(t), \mu(s(t)|\theta^\mu)|\theta^Q)\nabla_{\theta^\mu}\mu(s(t)|\theta^\mu)]. \tag{17}$$

同时, 为避免训练过程中计算网络梯度时出现震荡和发散, 保证参数波动较小且易于收敛, 按照软更新(soft update)方式更新两个目标网络参数, 即

$$\begin{aligned}\theta^{Q'} &\leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'},\\ \theta^{\mu'} &\leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}.\end{aligned} \tag{18}$$

进一步, 针对DDPG原有的经验回放策略, 在队列多输入-多输出样本空间中进行随机采样进而训练网络, 此方法会存在高维样本空间利用率不高且影响网络收敛效率的问题[19], 本文引入优先经验回放策略加以改进.

step 1: 针对样本j, 计算其样本价值$D(j)$为

$$D(j) = r(j) + \gamma Q'(s(j+1)), \mu'(s(j+1) - Q(s(j), a(j))). \tag{19}$$

step 2: 对经验池中样本按照$D(j)$绝对值进行降序排列, 得到其抽取优先级

$$p(j) = \frac{1}{\text{rank}(j)}, \tag{20}$$

其中rank(j)为样本j排序后队列序号, 且最高优先级为1, 限制了噪声样本存在的影响.

step 3: 获取优先级后进行可调采样, 有

$$P(j) = \frac{p(j,\lambda)}{\sum\limits_{N_k} p(N_k,\lambda)}. \tag{21}$$

其中: N_k 为新样本个数; $\lambda \in [0,1]$ 为可调样本在总样本中比例, 以确保样本多样性, $\lambda = 0$ 表示随机采样.

为了保证价值越高的样本在抽样过程中优先级越高, 对于可调样本一般使用高价值样本填充.

模型中, Actor网络和Critic网络输入状态为

$$s(t) = \{\Delta p_{i-1,i} - d_{\text{des}}, \Delta p_{i,0} - id_{\text{des}}, v_i, a_i(t-1)|T_s\}. \quad (22)$$

各通道状态维度共计20, 分别为5维度跟随车辆与前车间距误差 $\Delta p_{i-1,i} - d_{\text{des}}$、5维度与领航车之间间距误差 $\Delta p_{i,0} - id_{\text{des}}$、5维度跟随车辆速度 v_i、5维度跟随车加速度 $a_i(t-1)$. 此处加速度为上一状态动作输出, 且本文采样时间 $T_s = 0.1$ s.

Actor网络输出动作维度为5, 为避免过大或过小的加速度影响学习效率, 队列车辆加速度 a_i 同样应满足 $a_i \in [a_{\min}, a_{\max}]$, 分别为各跟随车辆的期望加速度, 即

$$a(t) = \{a_i\}. \quad (23)$$

注1　为控制对比变量, 此处控制器输出加速度采用相同PI控制器转换为油门踏板开度和刹车踏板开度.

在不同状态误差条件下, 通过设置合理的奖励函数能够引导智能体达到期望的加速度. 此处, 本文模仿MPC控制器的目标函数, 将奖励函数设置如下:

$$r(t) = \begin{cases} \beta_1 \sum \dfrac{1}{1 + (\Delta p_{i-1,i} - d_{\text{des}})^2}, \\ \beta_2 \sum \left(a_i(t-1)^2 + \dfrac{1}{(\Delta v_{i-1,i})^2} \right), \\ \beta_3 \text{Overshoot}(\Delta p_{i-1,i}, d_{\text{des}}) - 10c(t). \end{cases} \quad (24)$$

对比MPC控制器实际物理意义, 各通道奖励函数中, β 为通道引导系数. 通道1: β_1 为车辆期望间距误差累积和奖励引导系数, 即队列中车辆之间间距误差应趋向0时给予奖励; 通道2: β_2 为车辆前一时刻加速度 $a_i(t-1)$ 与前后车辆速度误差 $\Delta v_{i-1,i}$ 累积和奖励系数, 前两项通道通过求取累积和同时采用引入较小通道系数的方式避免过大的误差影响学习效率; 通道3: β_3 为相邻车辆间距 $\Delta p_{i-1,i}$ 与期望间距 d_{des} 过冲惩罚系数, 惩罚项表示为

$$\max(\Delta p_{i-1,i} - d_{\text{des}}). \quad (25)$$

$c(t)$ 表示碰撞检测项, 有

$$c(t) = \text{any}(\Delta p_{i-1,i} - d_{\text{des}}) \leqslant 0. \quad (26)$$

$c(t)$ 触发给予较大惩罚, 同时中止当前回合训练, 进入下一回合.

3　仿真实验与分析

3.1　仿真设置

为验证控制器工作性能, 强化学习训练环境采用Matlab/Simulink强化学习工具箱, 车辆动力学模型基于ADAS仿真平台PreScan搭建内燃机动力系统模型[20-21], 其发动机模型、变速器模型、换挡控制策略使用Siemens公开数据. 选取由1辆领航车和5辆跟随车组成的车辆队列进行仿真研究, 队列初始位置设置为 $p(0) = [110, 86, 63, 41, 20, 0]$ m, 各车初始速度均设置为15 m/s. 且在MPC控制器中, 有

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix},\ B = \begin{bmatrix} -T_C \\ -1 \end{bmatrix},$$
$$W = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix},\ I = 2, \omega. \quad (27)$$

其他控制器参数及车辆相关物理和空气动力学特性参数见表1.

表1　控制器参数及车辆动力学参数

参数	取值	参数	取值
$a_{\min}$/(m/s^2)	−3	$a_{\max}$/(m/s^2)	3
$v_{\max}$/(m/s)	33	ω	10
T_C/s	1.1	L/m	3.5
Y_N	10	l_c/m	3.8
β_1	0.1	β_2	0.2
β_3	−1	−	−
m/kg	1 575	g/(m/s^2)	9.8
m_{in}/m	1.04	α/(°)	0
ρ_{air}/(g/m^3)	1.225 6	A_f/m^2	2.5
C_d	0.3	C_{rr}	0.01

算法网络结构中, AC网络均采用全连接方式设计, 各隐含层之间采用ReLu函数激活, Actor网络均值激活函数为tanh函数, 方差激活函数为softmax函数.　两者损失函数均采用Adam方法更新梯度. 算法训练中神经网络超参数设计如表2所示.

表2　主要训练参数

超参数	取值	超参数	取值
Actor学习率	1e-4	Critic学习率	1e-3
折扣系数γ	0.99	软更新因子τ	1e-3
内存批大小	64	经验池 M	1e6
噪声方差	0.6	噪声衰减率	1e-5
最大步数	2 000	最大回合数	2 000 Episode

3.2　模型训练

训练阶段, 为增强模型的泛化能力, 每回合训练时, 在领航车参考轨迹、车辆队列初始状态、间距策略上均引入噪声, 如下所示:

$$v_0'(t) = v_0(t) + \text{amp} \cdot \sin(\text{fre} \cdot t), \quad (28)$$

$$p_i'(0) = p_i(0) + 5 \times \text{randn}(1,5), \tag{29}$$

$$v_i'(0) = v_i(0) + 1 \times \text{randn}(1,5), \tag{30}$$

$$L' = L + 3\text{randn}. \tag{31}$$

其中: sin 函数振幅 $\text{amp} = \max(2 + 0.1\text{randn}, 0.1)$, 频率 $\text{fre} = \max(1 + 0.1\text{randn}, 0.1)$, $\max(\cdot)$ 表示取其中的最大元素, randn$(\cdot)$ 表示产生的一个服从正态分布的随机数.

DRL 中, 通常采用回合奖励 (episode reward, ER) 与平均奖励 (average reward, AR) 反映模型训练的收敛水平和学习效果, 图 4 为模型训练过程中奖励值变化对比.

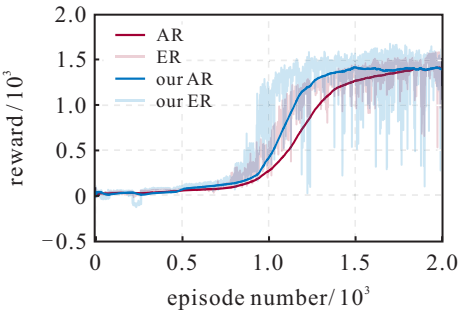


图 4 训练过程奖励值变化对比

在训练前 800 回合左右, 引入优先经验回放策略 (our) 算法与原有算法两者奖励值呈现较低水平, 方差基本恒定, 此时智能体从零开始学习, 未针对特定状态选取特定动作的经验, 处于随机探索的状态; 随着训练回合 (800 ~ 1 200) 的增加, 两者奖励值均值开始增加, 且方差也在增大, 表明此时智能体可能转移到局部最优附近, 此处具有较高探索意义. 此时, 引入优先经验回放策略后整体上升过程更快且更稳定, 在 1 500 回合左右基本收敛, 最终在 1800 回合后, 两者奖励值均值基本稳定且相差不大, 智能体不再获取奖励, 表明两者基本已收敛至全局最优, 智能体已学习到有效策略.

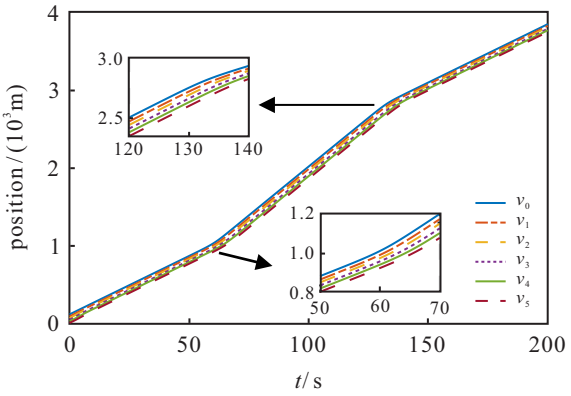
3.3 模型测试

3.3.1 场 景 1

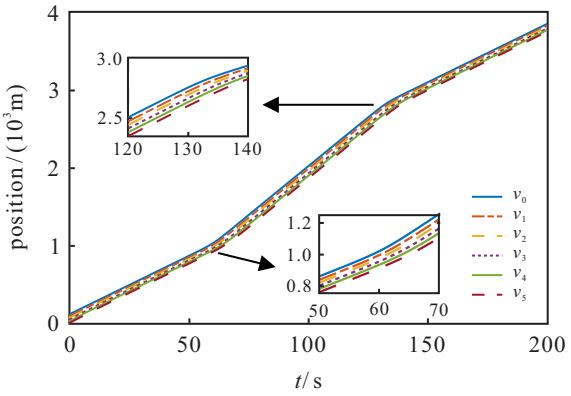
第 1 组测试场景主要测试队列单车稳定性和队列稳定性, 测试场景领航车的速度如下所示:

$$v_0(t) = \begin{cases} 15\,\text{m/s},\ 0\,\text{s} \leqslant t < 40\,\text{s}; \\ 15 + \dfrac{10}{1 + \text{e}^{-0.55t+33}}\,\text{m/s},\ 40\,\text{s} \leqslant t < 80\,\text{s}; \\ 25\,\text{m/s},\ 80\,\text{s} \leqslant t < 110\,\text{s}; \\ 15 + \dfrac{10}{1 + \text{e}^{0.55t-73}}\,\text{m/s},\ \text{otherwise}. \end{cases} \tag{32}$$

场景 1 下, MPC 控制器和 DRL 控制器测试过程中车辆队列各车位置、速度、加速度以及车辆间间距误差如图 5 ~ 图 8 所示.

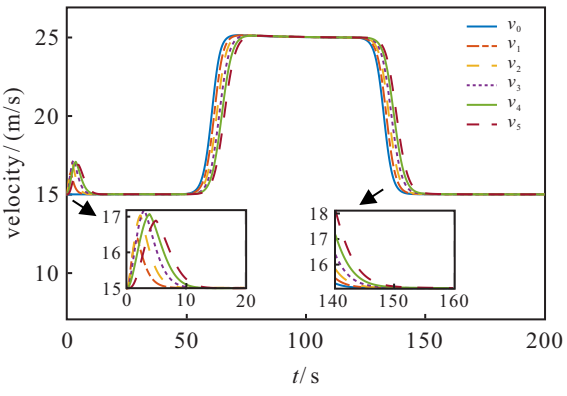


(a) MPC 控制器

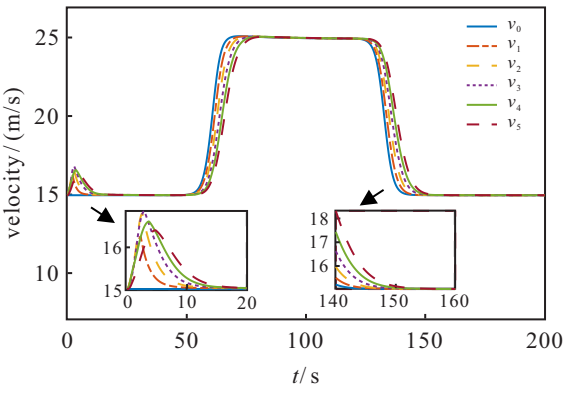


(b) DRL 控制器

图 5 位置曲线

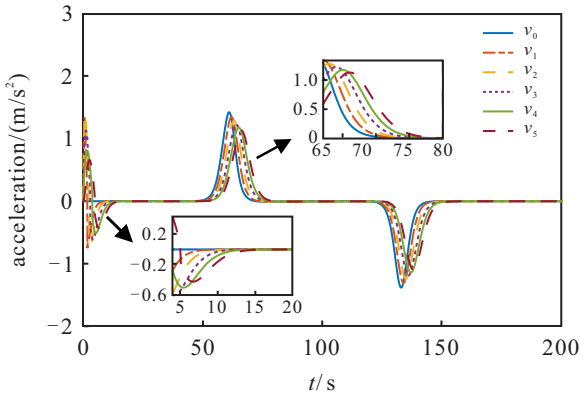


(a) MPC 控制器

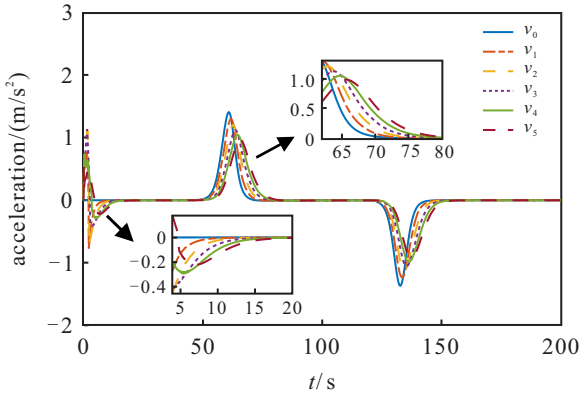


(b) DRL 控制器

图 6 速度曲线

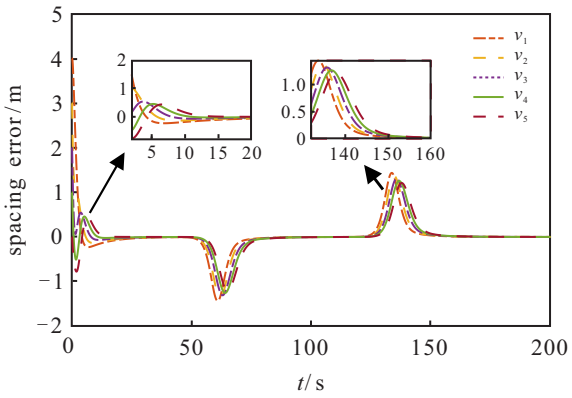


(a) MPC 控制器

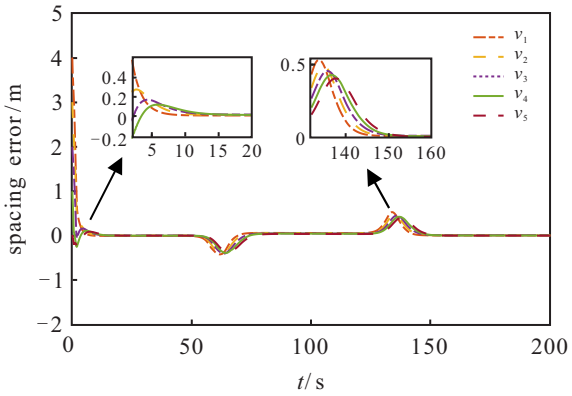


(b) DRL 控制器

图 7　加速度曲线



(a) MPC 控制器



(b) DRL 控制器

图 8　间距误差曲线

表 3　场景 1 队列过程控制性能表现

跟随车辆	与前车最小间距/m		平均速度/(m/s)		50～80 s 加速度峰值/(m/s²)		120～160 s 与前车间距误差峰值/m	
	MPC	DRL	MPC	DRL	MPC	DRL	MPC	DRL
vehicle1(v_1)	20	20	18.66	18.64	1.344	1.307	1.433	0.533
vehicle2(v_2)	20	20	18.71	18.66	1.279	1.211	1.365	0.495
vehicle2(v_3)	20	20	18.76	18.69	1.222	1.127	1.306	0.456
vehicle2(v_4)	20	20	18.82	18.72	1.173	1.051	1.253	0.429
vehicle2(v_5)	20	20	18.87	18.75	1.128	0.986	1.207	0.405

由图 5 车辆位置曲线, 结合表 3 中队列各跟随车辆与前车最小间距均可以保持在 20 m 可以得出, 基于 DRL 控制器同样能够保持期望的距离实现稳定队列行驶, 且在此过程中碰撞检测项 $c(t)$ 并未触发, 单车稳定性得到保证[22].

由图 6 车辆速度曲线可以观察到, 在领航车起步 0～20 s 以及引入波动 40～80 s 后, 5 辆跟随车的速度虽逐渐趋近, 但 DRL 控制器速度响应出现一定下降. 跟随车辆 5 分别在 150 s 前后与领航车速度一致, 原因是与 DRL 奖励函数设计累计误差限制幅度有关. 仿真发现, 累计误差通道系数越小其超调越小, 但响应速度会下降. 结合表 3 中 DRL 控制器平均速度均比 MPC 控制器更小, 进一步反映了 DRL 控制器控制过程相对更稳定, 具有更好的舒适性.

由图 7 车辆加速度曲线可以看出, DRL 控制器起始过程 0～20 s 的加速度变化范围在 $-0.4\ \mathrm{m/s^2}$ ～ $0.2\ \mathrm{m/s^2}$, 而 MPC 控制器加速度变化范围在 $-0.6\ \mathrm{m/s^2}$ ～ $0.4\ \mathrm{m/s^2}$, 且表 3 中 50～80 s 过程加速度峰值降低最多约 10 %, 也可以表明基于 DRL 控制器对车辆控制更为平滑.

由图 8 车辆间间距误差曲线, 结合表 3 数据队列行驶过程中 120～160 s 与前车间距误差峰值可以得出, DRL 控制器下间距误差峰值均降低约 0.8 m, 反映出跟随车辆对期望间距跟随更为紧密, 对于道路占用更小. 且在引入速度波动后两者队列中车辆间距误差在向后传播过程中都在不断减小, 此表现满足队列

串稳定性要求[16,23].

3.3.2 场景2

场景2参考CLTC-P中结合城市、郊区、高速行驶的工况对车辆队列跟随车辆燃油消耗以及两种控制器输出求解时间进行对比. 测试场景领航车的速度如图9所示. 测试场景总计运行1 800 s, 共计里程为14.48 km. 同时基于i7-9700@3.0 GHz, 16G RAM, Matlab R2021a, PreScan8.5环境下测试算法运行时间计算. 队列跟随车辆测试结果如表4所示. 表4中: 燃油消耗计算队列中所有跟随车辆采用相同气动特性; 算法运行时间以联合仿真环境为基准; 单步求解时间以Matlab计算时间为基准, 且表中参数采取了多次运行求取平均值的方式.

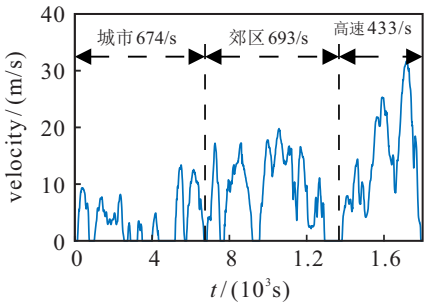


图9 领航车测试行驶场景

表4 场景2燃油消耗与算法运行时间对比

测试场景	燃油消耗/ml		运行时间/s		平均速度/(m/s)		单步求解时间/min	
	MPC	DRL	MPC	DRL	MPC	DRL	MPC	DRL
城市工况	143.78	133.25	679.73	33.99	5.74	5.49	0.62	0.04
郊区工况	356.58	328.65	701.24	36.38	11.02	10.47	0.61	0.039
高速工况	393.74	361.54	464.73	28.61	15.63	14.87	0.41	0.04

由表4可见, 不同工况下随着里程的增加, 两种控制器下的燃油消耗都在提升, 但DRL策略相比MPC策略在3种工况下分别节约近7.9%、8.5%、8.9%的油耗, 且在多次启停、加减速过程中平均速度均降低约5%, 表明DRL策略对于局部最优动作的选取更合适, 具有更好的燃油经济性和舒适性. 对比两者算法运行时间, DRL策略相较于MPC策略节约近90%的时间. 进一步对比两者单步计算时间分析可知, DRL通过训练好的策略网络直接输出动作, 其单步求解时间基本稳定不变, 而MPC单步决策过程中需要计算多个变量以及约束条件的目标函数, 若队列中车辆数目进一步增多, 则其决策变量也会增加, 算法实时性会减弱. 同时值得注意的是, DRL策略还需消耗时间在模型的训练上.

4 结 论

本文提出了一种基于DRL的网联车辆队列控制方法, 实现了不同工况下的单车稳定和队列稳定行驶. 相比于传统控制方法, DRL控制器通过状态-动作方式遍历网络直接输出控制, 同时以寻求局部最优动作达到全局最优动作. 从仿真场景看, 对于车辆控制能耗以及实时性均有一定的提升. 同时, 局限于DRL以交互式试错学习的特点, 难以在实际场景中训练, 但同时随着车辆智能化、网联化的发展, 对于硬件在环、实车验证等却很有必要, 后续将逐步考虑移植现有的模型.

参考文献(References)

[1] 李克强, 戴一凡, 李升波, 等. 智能网联汽车(ICV)技术的发展现状及趋势[J]. 汽车安全与节能学报, 2017, 8(1): 1-14.
(Li K Q, Dai Y F, Li S B, et al. State-of-the-art and technical trends of intelligent and connected vehicles[J]. Journal of Automotive Safety and Energy, 2017, 8(1): 1-14.)

[2] Ge X H, Han Q L, Ding L, et al. Dynamic event-triggered distributed coordination control and its applications: A survey of trends and techniques[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(9): 3112-3125.

[3] 陈龙, 何德峰, 李壮. 约束非线性车辆队列分布式多目标模型预测控制[J]. 控制与决策, 2022, 37(12): 3122-3128.
(Chen L, He D F, Li Z. Distributed multi-objective model predictive control for constrained nonlinear vehicle platoons[J]. Control and Decision, 2022, 37(12): 3122-3128.)

[4] Milanés V, Shladover S E, Spring J, et al. Cooperative adaptive cruise control in real traffic situations[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(1): 296-305.

[5] Lidström K, Sjöberg K, Holmberg U, et al. A modular CACC system integration and design[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(3): 1050-1061.

[6] Guo G, Li P, Hao L Y. A new quadratic spacing policy and adaptive fault-tolerant platooning with actuator saturation[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(2): 1200-1212.

[7] Guo G, Yang D Q, Zhang R. Distributed trajectory optimization and platooning of vehicles to guarantee smooth traffic flow[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(1): 684-695.

[8] van Nunen E, Reinders J, Semsar-Kazerooni E, et al. String stable model predictive cooperative adaptive cruise control for heterogeneous platoons[J]. IEEE Transactions on Intelligent Vehicles, 2019, 4(2): 186-196.

[9] Zheng Y, Li S E, Li K Q, et al. Distributed model predictive control for heterogeneous vehicle platoons under unidirectional topologies[J]. IEEE Transactions on Control Systems Technology, 2017, 25(3): 899-910.

[10] Arulkumaran K, Deisenroth M P, Brundage M, et al. Deep reinforcement learning: A brief survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[11] 闫超, 相晓嘉, 徐昕, 等. 多智能体深度强化学习及其可扩展性与可迁移性研究综述[J]. 控制与决策, 2022, 37(12): 3083-3102.
(Yan C, Xiang X J, Xu X, et al. A survey on scalability and transferability of multi-agent deep reinforcement learning[J]. Control and Decision, 2022, 37(12): 3083-3102.)

[12] Prathiba S B, Raja G, Dev K, et al. A hybrid deep reinforcement learning for autonomous vehicles smart-platooning[J]. IEEE Transactions on Vehicular Technology, 2021, 70(12): 13340-13350.

[13] 李文礼, 张友松, 韩迪, 等. 基于深度强化学习的车辆自主避撞决策控制模型[J]. 汽车安全与节能学报, 2021, 12(2): 201-209.
(Li W L, Zhang Y S, Han D, et al. Vehicle autonomous collision avoidance decision control model based on deep reinforcement learning[J]. Journal of Automotive Safety and Engergy, 2021, 12(2): 201-209.)

[14] 邓小豪, 侯进, 谭光鸿, 等. 基于强化学习的多目标车辆跟随决策算法[J]. 控制与决策, 2021, 36(10): 2497-2503.
(Deng X H, Hou J, Tan G H, et al. Multi-objective vehicle following decision algorithm based on reinforcement learning[J]. Control and Decision, 2021, 36(10): 2497-2503.)

[15] Yang J R, Liu X L, Liu S D, et al. Longitudinal tracking control of vehicle platooning using DDPG-based PID[C]. The 4th CAA International Conference on Vehicular Control and Intelligence. Hangzhou, 2021: 656-661.

[16] 于晓海, 郭戈. 车队控制中的一种通用可变时距策略[J]. 自动化学报, 2019, 45(7): 1335-1343.
(Yu X H, Guo G. A general variable time headway policy in platoon control[J]. Acta Automatica Sinica, 2019, 45(7): 1335-1343.)

[17] Zhou Y, Wang M, Ahn S. Distributed model predictive control approach for cooperative car-following with guaranteed local and string stability[J]. Transportation Research — Part B: Methodological, 2019, 128: 69-86.

[18] Sutton R S, Barto A G. An introduction[M]. Cambridge: MIT Press, 1998.

[19] 陈亮, 梁宸, 张景异, 等. Actor-Critic框架下一种基于改进DDPG的多智能体强化学习算法[J]. 控制与决策, 2021, 36(1): 75-82.
(Chen L, Liang C, Zhang J Y, et al. A multi-agent reinforcement learning algorithm based on improved DDPG in Actor-Critic framework[J]. Control and Decision, 2021, 36(1): 75-82.)

[20] Rajamani R. Vehicle dynamics and control[M]. The 2nd edition. New York: Springer, 2012.

[21] Bovee K, Hyde A, Trippel T, et al. Rapid vehicle architecture selection with use of autonomie[C]. The 5th Annual Dynamic Systems and Control Conference Joint with the 11th Motion and Vibration Conference. Florida: DSCC, 2013: 119-128.

[22] 秦严严, 王昊, 王炜, 等. 混有CACC车辆和ACC车辆的混合交通流驾驶舒适性[J]. 哈尔滨工业大学学报, 2017, 46(9): 103-108.
(Qin Y Y, Wang H, Wang W, et al. Driving comfort of traffic flow mixed with cooperative adaptive cruise control vehicles and adaptive cruise control vehicles[J]. Journal of Harbin Institute of Technology, 2017, 46(9): 103-108.)

[23] Qin W B, Orosz G. Experimental validation of string stability for connected vehicles subject to information delay[J]. IEEE Transactions on Control Systems Technology, 2020, 28(4): 1203-1217.

作者简介

李永福 (1983−), 男, 教授, 博士生导师, 从事智能网联汽车、空地协同控制等研究, E-mail: laf1212@163.com;

周发涛 (1997−), 男, 硕士生, 从事车辆队列、强化学习等研究, E-mail: zhoufatao_cqupt@163.com;

黄龙旺 (1993−), 男, 讲师, 博士, 从事智能控制、智能网联汽车等研究, E-mail: huanglw@cqupt.edu.cn;

于树友 (1974−), 男, 教授, 博士生导师, 从事模型预测控制、车辆队列控制等研究, E-mail: shuyou@jlu.edu.cn;

施树明 (1965−), 男, 教授, 博士生导师, 从事车辆动力学建模与控制、车辆运行仿真等研究, E-mail: shishuming@jlu.edu.cn.