

基于行为规约近端策略优化的 自主交叉口管理方法

高镇海¹, 郝鹤声², 高 菲¹, 赵 睿²

(1. 吉林大学 汽车仿真与控制国家重点实验室, 长春 130022; 2. 吉林大学 汽车工程学院, 长春 130022)

摘 要: 针对当前集中式协同控制方法存在计算效率低和无安全保障的问题, 本文首先提出了一种基于强化学习的群体协同算法, 将单智能体近端策略优化扩展到多智能体协同合作的复杂交互环境中, 以解决多智能体系统的复杂合作问题。其次, 将无信号交叉口车辆集中式协同控制形式化为多智能体强化学习问题, 并提出一种安全增强的交叉口集中式协同控制方法——行为规约近端策略优化。该方法将形式化安全验证及行为规约融入群体协同算法, 以指导策略安全迭代优化和避免非安全驾驶行为, 进一步保障未知场景下的通行安全。最后, 通过仿真软件 Carla 进行模拟实验。仿真结果表明: 行为规约的纳入牺牲了 8.06% 的通行效率, 获得了 100% 的安全提升; 相较典型的模型预测控制方法, 本文方法将计算时间缩短到 1/326 倍, 交通效率提高了 67.0%, 碰撞率从 63.5% 降低到 0, 舒适性提升了 26.5%。

关键词: 车辆工程; 自主交叉口管理; 智能网联汽车; 强化学习; 形式化验证

中图分类号: U491 **文献标志码:** A **文章编号:** 1671-5497(2025)10-3151-11

DOI: 10.13229/j.cnki.jdxbgxb.20231335

Behavior-constrained proximal policy optimization for autonomous intersection management

GAO Zhen-hai¹, HAO He-sheng², GAO Fei¹, ZHAO Rui²

(1. State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun 130022, China; 2. College of Automotive Engineering, Jilin University, Changchun 130022, China)

Abstract: Aiming at the low computational efficiency and absence of safety guarantees in existing centralized cooperative control schemes, a swarm-coordination algorithm based on reinforcement learning is first proposed. Single-agent proximal policy optimization is extended to multi-agent interactive environments so that complex cooperation in multi-agent systems can be addressed. Secondly, the cooperative control of vehicles at unsignalized intersections is formulated as a multi-agent reinforcement-learning problem, and a safety-augmented centralized cooperative control method—behavior-restricted proximal policy optimization—is developed. Formal safety verification and behavior restrictions are

收稿日期: 2023-12-02.

基金项目: 国家自然科学基金项目 (52202494, 52202495).

作者简介: 高镇海 (1973-), 男, 教授, 博士. 研究方向: 汽车驾驶辅助系统, 无人驾驶与驾驶行为分析.

E-mail: gaozh@jlu.edu.cn

通信作者: 赵睿 (1986-), 女, 副教授, 博士. 研究方向: 自动驾驶主动安全. E-mail: rzhao@jlu.edu.cn

integrated into the swarm-coordination algorithm, whereby the policy is guided to be iteratively optimized in a safe manner and unsafe driving behaviors are avoided, so that traffic safety in unknown scenarios can be further guaranteed. Finally, simulation experiments are conducted with the Carla platform. It is shown that the incorporation of behavior restrictions causes an 8.06 % loss in traffic efficiency, yet a 100 % safety improvement is achieved. Compared with the representative model predictive control approach, the proposed method reduces the computation time to 1/326 of the original, increases traffic efficiency by 67.0 %, lowers the collision rate from 63.5 % to 0, and improves ride comfort by 26.5 %.

Key words: automotive engineering; autonomous intersection management; autonomous and connected vehicles; reinforcement learning; formal verification

0 引 言

随着汽车保有量的急剧增加和通行需求的迅速扩大,交通拥堵问题日益严重。交叉口作为城市路网的关键节点,对整体路网的效率和安全有深刻影响。

最早的交叉口控制系统使用固定信号配时方案,存在信号配时不合理问题,缺乏自适应能力。一些研究^[1,2]聚焦于可变信号配时以改进信号控制系统。然而,这些改进仍未能充分利用交叉口的时空资源^[3,4]。

智能网联汽车(Connected and autonomous vehicle, CAV)和高效通信技术(如 V2X、5G)的出现,推动了自主交叉口管理(Autonomous intersection management, AIM)的广泛发展^[5,6]。AIM旨在有效管理CAVs在无信号交叉口的安全高效通行。当前 AIM 相关研究可分为基于规则和基于优化的方法。

基于规则的方法通常基于预定义的规则来为CAVs分配优先级或进入交叉口的时间^[7-11]。先到先得服务(First come first service, FCFS)^[7-9]是一种经典的基于规则的算法,根据车辆到达交叉口的顺序对车辆的通行优先级进行排序。其他方法则基于交通法规或预约系统的固定规则。Lu等^[10]提出了一套基于交通安全法规的交叉口车辆引导策略,允许某些车辆抢先或礼让其他车辆,来建立无信号交叉口的通行序列。Zhang等^[11]提出了一套基于预约请求的车辆调度规则,以确保高级别的预约请求得到优先响应。在稀疏交通流场景中,相较传统的信号控制,基于规则的方法实现了更高的交叉口吞吐量、通行效率、通行公平性。然而,这种方法也有明显的缺点,随着进入交叉口的车辆增多,通行效率迅速降低。在密集交

通流量场景下,如果预约请求失败,车辆必须以更低的速度重新提出请求,导致交通堵塞问题加剧^[12]。

基于优化的 AIM 方法主要采用博弈论、最优控制(Optimal control, OC)和深度强化学习(Deep reinforcement learning, DRL)方法。

在博弈论方法中,通常将CAVs视为博弈游戏的参与者,目的是建立车辆通行决策的纳什均衡。Li等^[13]提出了一种基于博弈的通行方法,模拟人类驾驶员在无信号交叉口的决策交互过程。类似地,Wang等^[14]利用纳什博弈来模拟两辆CAV在无信号交叉口竞争的通行策略。Elhenawy等^[15]提出了一种交叉口通行博弈方法,用于管理CAVs的运动,以替代传统的信号控制系统。然而,基于博弈论的方法存在一些局限性:①在需要实时响应的复杂问题中,计算复杂度较高;②寻找到稳定的纳什均衡解可能具有挑战性。

基于OC的方法通常将交叉口的车辆运动建模为一个动态系统,建立考虑车辆动力学约束的目标函数,然后计算车辆控制指令以优化车辆通行。Zhao等^[16]开发了一套集中式管理机制,将信号配时控制和车辆轨迹优化相结合,采用混合整数线性规划来最小化总行驶时间。Nair等^[17]提出了一种应用于交叉口自动驾驶的随机模型预测控制(Model predictive control, MPC)方案,使用高斯混合模型对周围车辆进行多模态预测。Kamal等^[18]提出了一种基于MPC框架的交叉口车辆协同方案(Vehicles-intersection coordination scheme, VICS),将交叉口通行问题转化为受约束的非线性优化问题。数值模拟结果表明,VICS显著降低了车辆的停车延迟,提高了交叉口的通行效率。MPC方法相较其他OC方法更具有预测性和鲁棒性,能迅速适应环境变化^[19]。然而,MPC方法

存在明显的计算效率问题。计算所有车辆未来运动状态并寻找最优解需要消耗大量计算资源,随着车辆数量的增加,计算时间呈指数级增长,因此,MPC无法满足自动驾驶技术对控制指令的高实时性要求。

近年来,DRL在单智能体系统中得到了广泛应用,并取得了积极成果^[20-23]。DRL结合强化学习(Reinforcement learning, RL)和深度神经网络,具有模拟复杂行为策略和显著提高计算速度等优势。DRL采用试错机制来最大化奖励与环境交互,从而学习最佳策略。在信号灯交叉口研究方面,Zhou等^[20]设计了一个基于RL的汽车跟随模型,以提高交叉口的通行效率、燃油经济性和安全性。Guo等^[21]开发了一种利用神经网络进行Q函数近似的交叉口控制方法。在无信号交叉口方面,Guan等^[22]提出了一种基于策略梯度的交叉口管理方法来控制车辆在交叉口的运动,利用先验模型提高了算法的训练效率,但未验证算法的安全性。Wang等^[23]提出了一种基于神经网络的策略来管理CAV在交叉口右转场景下的通行速度,提升了通行成功率,缩短了平均通行时间。然而,在交叉口协同控制问题中,DRL的实施存在一些挑战和困难。第一,对于安全关键的车辆通行问题,基于奖励驱动的RL将安全性作为评价指标而不是强约束,从而导致系统存在安全风险。为提高交通安全性,需要在训练过程中进行形式化安全验证,以精细化指导策略的低行车风险更新;此外,在部署过程中仍需对驾驶行为进行规约,以保证高风险工况下的安全行驶。第二,在多智能体交互系统中,离散化每个智能体并对其策略进行训练,会导致训练难以收敛。训练过程中其他智能体的策略也在不断更新,导致训练环境不稳定。因此,单智能体RL方法不足以开发性能优秀且稳定的AIM方法。第三,一些研究的场景设定过于简化,难以应用到真实世界中,如Wang等^[23]仅考虑了车辆在交叉口右转的单一场景,导致算法在实际应用中存在较大的局限性。

基于上述问题,本文首先提出基于多智能体RL的群体协同控制算法——多智能体近端策略优化(Multi-agent proximal policy optimization, MAPPO),来解决多智能体间竞争合作的复杂交互问题。然后,将交叉口车辆协同控制问题形式化为多智能体RL问题,并提出一种基于行为规

约的近端策略优化(Behavior-constrained proximal policy optimization, BCPPO)的AIM方法,在MAPPO训练期间对驾驶行为进行形式化安全验证,以指导策略迭代向较低行车风险的方向更新,并在策略部署期间将行为规约融入MAPPO,进一步确保驾驶安全。最后,在仿真环境中训练集中式车辆协同控制策略,并将计算效率、通行安全、交通效率和舒适性与无行为规约的MAPPO和基于MPC的协同方法进行了比较。

1 交叉口通行问题分析

1.1 交叉口场景模型

本节建立了城市道路十字形交叉口的道路拓扑模型,如图1所示。交叉口车辆通行驾驶路线

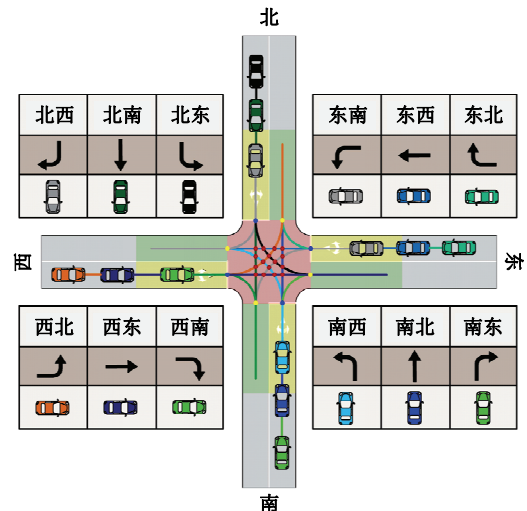


图1 行驶区域划分、驾驶路线及潜在冲突点

Fig. 1 Driving area division, driving routes and potential conflict points

包括12种情况。车辆通过交叉口的行驶区域可分为缓冲区域、核心区域和驶离区域。缓冲区域(黄色阴影)指车辆进入交叉口之前的行驶区域。核心区域(红色阴影)指各车道交汇相聚形成的区域,包含潜在的行驶冲突点。驶离区域(绿色阴影)指车辆驶出交叉口后恢复到巡航速度的区域。车辆在交叉口可能发生的碰撞类型分为交叉碰撞、合流碰撞和分流碰撞,分别如图1中红色点、黄色点和蓝色点所示。

1.2 车辆运动状态模型

车辆运动状态模型如图2所示,其中 d^b 为交叉口缓冲区的长度。车辆 i 的状态特征使用元组 $(v_i(t), a_i(t), j_i(t), \Psi_i(t, p), d_i^c(t), d_i^e(t), d_i^l, d_i^r)$ 表示。其中 $v_i(t)$ 为车辆 i 在时刻 t 的行驶速度,

$a_i(t)$ 和 $j_i(t)$ 分别为车辆 i 的加速度和加加速度,计算公式如下:

$$\begin{aligned} a_i(t) &= \dot{v}_i(t) \\ \text{s. t. } a_i(t) &\in (a_i^{\min}, a_i^{\max}) \end{aligned} \quad (1)$$

$$j_i(t) = \ddot{v}_i(t) \quad (2)$$

式中: $a_i^{\min}$ 和 $a_i^{\max}$ 分别为车辆动力学约束的最小和最大加速度,设置为 -5 m/s^2 和 5 m/s^2 。

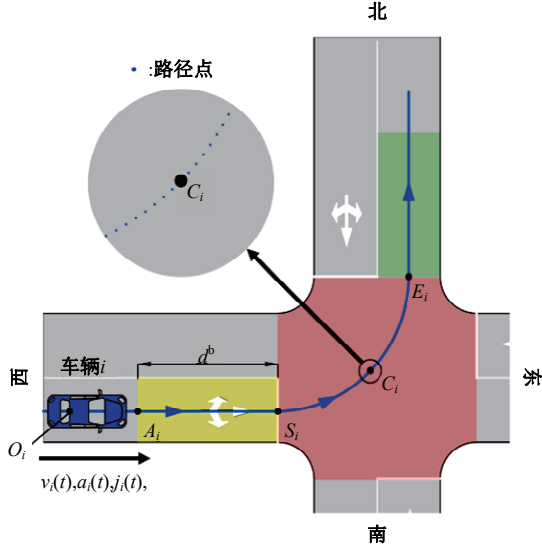


图 2 车辆运动模型

Fig. 2 Vehicle motion model

本文引入了一个一维坐标系 $\Psi_i(t, p)$ 来表示时刻 t 路径点 p 在车辆 i 路径上的位置。该坐标系以交叉口入口为坐标原点,并沿着车辆路径延伸。根据上述定义,从车辆 i 的中心点 O_i 到交叉口内路径中点 C_i 的距离可表示为 $d_i^c(t) = \Psi_i(t, C_i) - \Psi_i(t, O_i)$; O_i 到交叉口出口 E_i 的距离可表示为 $d_i^e(t) = \Psi_i(t, E_i) - \Psi_i(t, O_i)$ 。另外, d_i^f 和 d_i^r 分别表示 O_i 到车辆 i 车身的前端和后端的距离。定义 $d_{ij}^c(t)$ 为车辆 i 的中心点 O_i 到其与车辆 j 的冲突点的距离。定义 $T_{ij}(t) = (d_{ij}^c(t) - d_i^f)/v_i(t)$ 为车辆 i 的车身前端到达冲突点的时间。进一步地,本文定义在时刻 t 一对冲突车辆 p_{ij} 的碰撞时间差为 $T_{ij}^d(t)$,计算公式如下:

$$T_{ij}^d(t) = |T_{ij}(t) - T_{ji}(t)| \quad (3)$$

1.3 问题描述

将问题描述如下:有一组数量为 N_a 的车辆驶近交叉口,每辆车辆 $i(i=0:N_a)$ 均在其特定的车道上行驶。对于每辆车辆 i ,一旦其驶进缓冲区,交叉口控制器就会开始接管控制它的运动。在每一个离散的时间步,根据车辆的静态信息(车身尺寸 d_i^f)和动态信息(位置 $d_i^c(t)$ 和速度 $v_i(t)$)实时

计算发送给车辆 i 的纵向加速度,来实现多辆CAVs的协同通行。当车辆 i 驶出交叉口($d_i^e(t) < 0$),车辆 i 在比例积分微分方法的控制下加速至期望速度,恢复正常行驶。

2 多智能体强化学习

首先,将多智能体RL问题形式化为马尔可夫博弈架构,然后提出MAPPO方法来解决复杂的多智能体系统中协同竞争合作的严重挑战。

2.1 马尔可夫博弈

马尔可夫博弈(Markov game, MG)形式化了多个智能体与环境之间交互的RL过程。定义包含了 N_a 个智能体的MG为一个五元素的元组 $\{S, A, R, P, \gamma\}$,元组的详细描述如下:

- $S: S_1 \times S_2 \times \dots \times S_{N_a}$ 表示多智能体联合状态空间,代表所有智能体可能的特征集合,其中 S_i 为智能体 i 的状态空间。

- $A: A_1 \times A_2 \times \dots \times A_{N_a}$ 表示多智能体的联合动作空间,其中 A_i 为智能体 i 的动作空间。

- $R: S \times A_1 \times A_2 \times \dots \times A_{N_a} \times S \rightarrow R$ 表示联合奖励函数,描述了在时刻 t 从联合状态 $s_t = \{s_{1,t}, s_{2,t}, \dots, s_{N_a,t}\}$ 采用联合动作 $a_t = \{a_{1,t}, a_{2,t}, \dots, a_{N_a,t}\}$ 到下一个状态 s_{t+1} 的即时奖励。其中 $s_{i,t} \in S_i$ 和 $a_{i,t} \in A_i$ 分别是智能体 i 在时刻 t 的状态和行为。

- $P: S \times A_1 \times A_2 \times \dots \times A_{N_a} \times S \rightarrow [0, 1]$ 为状态转移函数,表示在时刻 t 采取联合动作 a_t ,从当前状态 s_t 转移到下一个状态 s_{t+1} 的概率。

- $\gamma \in (0, 1]$ 指奖励衰减系数,是一个超参数,代表未来奖励相对于近期奖励的重要性。

此外,RL还引入了一个随机策略 $\pi: S \times A \rightarrow [0, 1]$ 来表示在某个状态下采取每个动作的概率分布。回合是指从初始状态到最终状态的 (s_t, a_t) 序列。RL与环境交互得到轨迹 τ :

$$\tau = \{s_1, a_1, R_1, \dots, s_{T-1}, a_{T-1}, R_{T-1}, s_T\} \quad (4)$$

G_t 指在一个回合内从状态 s_t 到最终状态 s_T 所获得的累积折扣奖励,数学公式如下:

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots = \sum_{r=0}^{T-t-1} \gamma^r R_{t+r+1} \quad (5)$$

定义状态价值函数 $V_\pi(s)$ 为从状态 s 开始,遵循策略 π 直至最终时间步的期望累积折扣奖励,即:

$$V_\pi(s) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k} | s_t = s \right\} \quad (6)$$

相似地,定义动作价值函数 $Q_\pi(s, a)$ 为在状态 s 执行动作 a , 随后遵循策略 π 直至最终时间步的期望累积折扣奖励, 即:

$$Q_\pi(s, a) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k} | s_t = s, a_t = a \right\} \quad (7)$$

2.2 多智能体近端策略优化

在 MG 架构的基础上, 本文提出多智能体近端策略优化 (Multi-agent proximal policy optimi-

zation, MAPPO) 算法, 将单智能体近端策略优化扩展到多智能体竞争合作的复杂交互环境。

2.2.1 网络架构

MAPPO 采用图 3 所示的行动家-评论家网络架构。行动家网络代表智能体的策略, 旨在促进智能体与环境的交互。评论家网络用于近似状态价值函数, 以评价行动家策略的性能并指引行动家网络的更新方向。

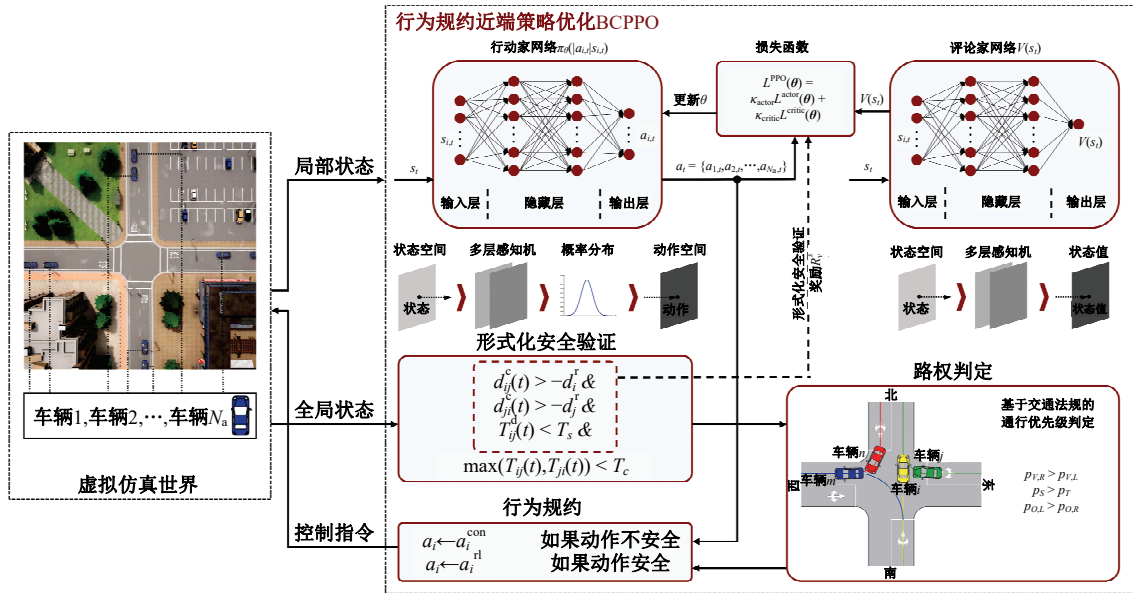


图 3 BCPPO 系统架构

Fig. 3 Architecture diagram of BCPPO

行动家网络 $\pi_\theta(a_i|s_i): S_i \times A_i \rightarrow [0, 1]$ 以 θ 为网络参数, 利用多层感知机来进行多智能体状态 s_i 到联合动作 a_i 的映射。首先, 将所有智能体的状态 s_i 输入策略网络的输入层, 随后经过两层隐藏的全连接 (Fully connect, FC) 层的映射, 将隐藏信息前向传播到神经网络的输出层, 得到智能体动作的高斯概率分布 $N^i(\mu^i, \sigma^i)_{i=1:N_s}$, 并在此分布中选择每个智能体的动作 $a_i \sim N^i(\mu^i, \sigma^i)_{i=1:N_s}$ 。

评论家网络 $V_\phi(s)$ 以 ϕ 为网络参数, 具有与行动家网络相同的网络结构, 即两层隐藏 FC 层。评论家网络 $V_\phi(s)$ 将状态 s_i 映射到状态值 $V(s_i)$ 。

2.2.2 策略优化

策略优化的目标是使 RL 的目标函数 $J(\theta)$ 最大化, 得到最优策略的网络参数 $\theta^* = \arg \max_\theta J(\theta)$ 。目标函数 $J(\theta)$ 最大化意味着期望累积折扣奖励的最大化, 可表示为:

$$J(\theta) = E_{a_i \sim \pi_\theta(a_i|s_i), s_{t+1} \sim p(s_{t+1}|s_i, a_i)} \left\{ \sum_{t=0}^{\infty} \gamma^t R_t \right\} \quad (8)$$

为了寻找到 θ^* , MAPPO 以目标函数作为损失函数, 通过策略梯度 (Policy gradient, PG) 对策略网络进行优化, 在每次迭代中提高获得更大奖励的概率。经过数千次迭代, 训练最终收敛, $J(\theta)$ 取得最大值。利用先前定义的动作价值函数 $Q_\pi(s, a)$, 行动家网络的梯度可计算如下:

$$\begin{aligned} \nabla_\theta J(\theta) &\approx \\ E_{a_i \sim \pi_\theta(a_i|s_i), s_{t+1} \sim p(s_{t+1}|s_i, a_i)} \{ \nabla_\theta \log \pi_\theta(a_i|s_i) Q_{\pi_\theta}(s, a) \} \end{aligned} \quad (9)$$

为了更好地利用高回报和低回报的样本数据进行策略更新, 引入了一个基线 $b_v = V_\phi(s)$ 来减小方差, 提高训练性能。 $\nabla_\theta J(\theta)$ 可更新为:

$$\begin{aligned} \nabla_\theta J(\theta) &\approx \\ E_{a_i \sim \pi_\theta(a_i|s_i), s_{t+1} \sim p(s_{t+1}|s_i, a_i)} \{ \nabla_\theta \log \pi_\theta(a_i|s_i) A_{\pi_\theta, V_\phi}(s, a) \} \end{aligned} \quad (10)$$

式中: $A_{\pi_\theta, V_\phi}(s, a) = Q_{\pi_\theta}(s, a) - V_\phi(s)$ 为优势函数, 旨在量化行动家网络所采取动作的性能。

MAPPO 采用广义优势估计 (Generalized ad-

vantage estimation, GAE) 方法对优势函数进行估计, 公式如下:

$$\hat{A}_{\pi_{\theta}, V_{\phi}}(s, a) = \hat{Q}_{\pi_{\theta}}(s, a) - V_{\phi}(s) \quad (11)$$

式中: $\hat{Q}_{\pi_{\theta}}(s, a)$ 为利用样本估计的动作价值函数。GAE 通过 n 步自举, 优化了神经网络的低偏差和低方差性能。

$\hat{Q}_{\pi_{\theta}}(s, a)$ 的计算公式如下:

$$\hat{Q}_{\pi_{\theta}}(s_t, a_t) = \sum_{l \geq t}^{T-1} (\lambda \gamma)^{l-t} \delta_l + V_{\phi}(s_t) \quad (12)$$

式中: $\delta_t = -V_{\phi}(s_t) + R_t + \gamma V_{\phi}(s_{t+1})$ 为 TD 误差; λ 为回报衰减系数。

此外, MAPPO 利用重要性采样来提高样本的训练效率, 并利用裁切函数来防止行动家网络参数的过度更新。行动家网络的损失函数构造如下:

$$L^{\text{actor}}(\theta) = E_t \{ \min(d_t^{\theta} \hat{A}_{\pi_{\theta}, V_{\phi}}, \text{clip}(d_t^{\theta}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{\pi_{\theta}, V_{\phi}}) \} \quad (13)$$

式中: $d_t^{\theta} = \frac{\pi_{\theta}(a_{i,t}|s_t)_{i=1:N_s}}{\pi_{\theta}^{\text{old}}(a_{i,t}|s_t)_{i=1:N_s}}$ 为新旧策略的动作概率比值。

评论家网络的损失函数为:

$$L^{\text{critic}}(\phi) = (V_{\phi}(s) - \hat{V}_{\phi}(s_t))^2 \quad (14)$$

式中: $\hat{V}(s_t) = \hat{A}_{\pi_{\theta}, V_{\phi}} + V_{\phi}(s_t)$ 为价值目标的估计。

最后, 通过梯度下降法对参数向量 θ 和 ϕ 进行更新, 即:

$$\begin{cases} \theta \leftarrow \theta + \alpha_{\theta} \nabla_{\theta} L^{\text{actor}}(\theta) \\ \phi \leftarrow \phi + \alpha_{\phi} \nabla_{\phi} L^{\text{critic}}(\phi) \end{cases} \quad (15)$$

式中: α 为神经网络的学习率。

3 集中式车辆协同控制

首先通过定义状态空间、动作空间和详细的奖励函数设计, 将交叉口的车辆协同控制问题形式化为 MG。随后, 提出结合 MAPPO 与形式化安全验证与行为规约的 BCPPO 方法, 以解决安全关键的交叉口车辆协同控制问题。

3.1 MG 形式化

通过定义 MG 的基本元素——状态空间、动作空间和奖励函数, 将交叉口车辆协同控制问题转化为 RL 问题。

3.1.1 状态和动作空间

定义车辆数量为 N_a 的车辆协同控制问题的

状态空间为 $S = \{S_1, S_2, \dots, S_{N_s}\}$ 。时刻 t 车辆 i 的状态 $s_{i,t} = \{d_i^c(t), v_i(t)\} \in S_i$ 包含了动态位置和速度信息。定义动作空间为 $A = \{A_1, A_2, \dots, A_{N_s}\}$, 时刻 t 车辆 i 的动作 $a_{i,t}$ 代表纵向加速度控制指令。因此, 构造了 $2N_a$ 维的状态空间和 N_a 维的动作空间。形式化安全验证流程如图 4 所示。

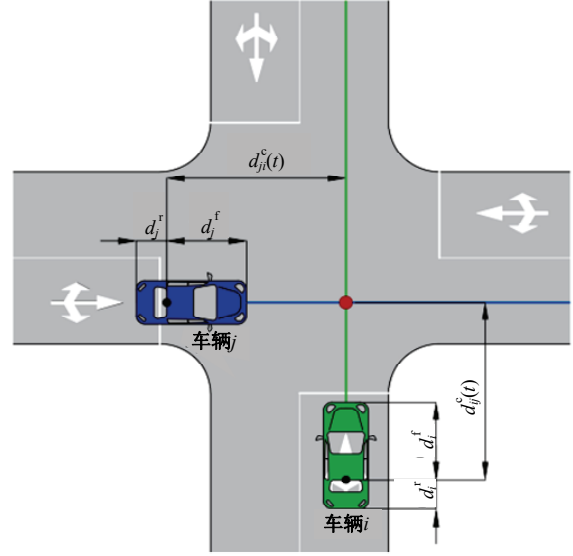


图 4 形式化安全验证流程

Fig. 4 Process of formal safety verification

3.1.2 奖励函数

MAPPO 根据回合的时间步累积奖励来指导策略进行迭代更新。奖励函数的设计是提高算法收敛速度和性能的关键。设置积极奖励可以引导算法向回合更长, 积累更多奖励的方向优化, 而消极奖励则可以引导算法尽快结束回合, 避免奖励积累。

总回合奖励 R 包括 3 个子项: 安全奖励 R^S 、效率奖励 R^E 和舒适奖励 R^C , 有:

$$R = R^S + R^E + R^C \quad (16)$$

安全奖励 R^S 包括 3 个元素: 碰撞奖励 R_c^S 、通过奖励 R_p^S 和形式化安全验证奖励 R_v^S 。

$$R^S = R_c^S + R_p^S + R_v^S \quad (17)$$

碰撞奖励 R_c^S 为车辆碰撞事件相关的奖励。如果回合结束时车辆发生碰撞, RL 会得到严重的消极奖励 $R_c^S = k_c$; 否则碰撞奖励为 0, 即 $R_c^S = 0$ 。

通过奖励 R_p^S 为安全通过交叉口的车辆数量相关的奖励。在回合结束时, 赋予 RL 一个积极奖励 $R_p^S = k_p^{\text{one}} N_p$, 其中 k_p^{one} 为单辆车的通过奖励, N_p 为成功通过交叉口的车辆数量。此外, 如果所有车辆均无碰撞安全地通过了交叉口, RL 将获得

额外的奖励 k_p^{all} 。通过设计渐进增大的通过奖励,引导策略向更多车辆安全通过交叉口的方向更新。

形式化安全验证奖励 R_v^s 量化了车辆交互过程中的碰撞风险,如图 4 所示。当路径冲突车辆对 p_{ij} 的运动状态满足以下条件时,赋予 RL 一个消极奖励:①两辆车均未通过潜在冲突点;②碰撞时间差 $T_{ij}^d(t)$ 小于安全时间阈值 T_s 。在每个时间步,RL 得到车辆对累积的形式化安全验证奖励 R_v^s ,计算公式如下:

$$R_v^s = \begin{cases} \frac{1}{M} \sum_t \sum_{p_{ij}=1}^M k_t e^{-T_{ij}^d(t)+b_t}, & d_{ij}^c(t) > -d_i^r \& d_{ji}^c(t) > -d_j^r \& T_{ij}^d(t) < T_s \\ 0, & \text{其他} \end{cases} \quad (18)$$

式中: M 为存在路径冲突的车辆对数量; k_t 和 b_t 分别为形式化安全验证的权重和偏差,用于调整奖励值。

当形式化验证某个行为不安全时,算法对奖励施加一个呈指数级增加的惩罚,以指导策略朝着车辆不同时接近其冲突点的方向优化,从而加强车辆的碰撞避免性能。相反,验证到某个行为安全时,RL 得到的形式化安全验证奖励可以忽略不计。

效率奖励 R^E 取决于回合的长度,并与之成反比,即:

$$R^E = l_e k_s \quad (19)$$

式中: l_e 为回合长度; k_s 为时间步奖励。

通过设计效率奖励,引导 MAPPO 不断提高通行效率。

舒适奖励 R^C 用于表征车辆行驶的舒适性,引导策略向行驶更舒适的方向优化,计算如下:

$$R^C = \begin{cases} 0, & N_p = N_a \\ \frac{1}{N_a - N_p} \sum_t \sum_{i=1}^{N_a - N_p} k_j j_i(t)^2, & N_p < N_a \end{cases} \quad (20)$$

式中: k_j 为舒适性奖励的权重。

3.2 形式化验证与规约

尽管在 MAPPO 的训练过程中引入了形式化安全验证,对高风险行为施加了惩罚,但由于交叉口通行系连续无限状态的环境,RL 部署中不可避免会面临未知情况。因此,本文提出 BCPPO 方法,将车辆行为规约集成到 MAPPO 部署过程中,进一步保障罕见未知情况下的行驶安全性。

行为规约是指基于特定的规则,以数字化的

形式建立智能体之间的安全规约,以保障实际应用中系统的安全性。对于交叉口车辆通行问题,当路径冲突车辆对即将相撞时,需对车辆施加行为规约,触发条件如下:①经形式化安全验证,当前 RL 行为是不安全行为;②两车到冲突点的时间均小于干预时间阈值 T_c 。

当触发行为规约时,BCPPO 对无道路通行权(以下简称路权)的车辆施加规约。本文参照《中华人民共和国道路交通安全法实施条例》来判定车辆的路权,如图 3 所示。例如,根据“转弯车辆让行直行车辆”的原则,以“南北”为路线的车辆 i 相对于以“东北”为路线的车辆 j 具有路权。然而,在不同国家或地区交通规则可能会有所不同^[24]。类似于本文的描述,可以轻易地建立路权判定规则。

基于上述路权判定规则进行形式化规约。车辆 i 和车辆 j 遵循 MAPPO 的控制命令 a_i^{rl} 和 a_j^{rl} 行驶接近交叉口。假设判定得到车辆 j 相对车辆 i 具有路权,当检测到两车即将碰撞时,立即对车辆 i 施加值为 -5 m/s^2 的规约加速度 a_i^{con} ,以确保车辆安全行驶。直至车辆 j 的车身后端通过冲突点,两车的潜在碰撞消失,解除对车辆 i 的行为规约,MAPPO 恢复接管车辆 i 的通行。对无路权车辆 i 的行为规约如下:

$$a_i = \begin{cases} d_{ij}^c(t) > -d_i^r \& d_{ji}^c(t) > -d_j^r \& T_{ij}^d(t) < T_s \& \max(T_{ij}(t), T_{ji}(t)) < T_c \\ a_i^{\text{con}}, & \\ a_i^{\text{rl}}, & \text{其他} \end{cases} \quad (21)$$

3.3 基于 BCPPO 的车辆协同控制

BCPPO 方法的实现如算法 1 所示。根据相关研究^[25,26],共享网络参数可以获得更好的训练效果。因此,本文共享行动家网络和评论家网络的参数。RL 的损失函数可更新为:

$$L^{\text{PPO}}(\theta) = [\mathbf{k}_{\text{actor}} \quad \mathbf{k}_{\text{critic}}] [L^{\text{actor}}(\theta) \quad L^{\text{critic}}(\theta)]^T \quad (22)$$

式中: $\mathbf{k}_{\text{actor}}$ 和 $\mathbf{k}_{\text{critic}}$ 分别为行动家网络和评论家网络的损失函数的权重。

算法 1 基于 BCPPO 的交叉口车辆协同控制

- 1 初始化行动家-评论家网络 π 为 π_0 ,网络参数 θ 为 θ_0 。设置学习率 α ,奖励衰减系数 γ ,裁剪范围 ϵ ,总时间步 T ,批大小 B ,最小批大小 MB ,世纪 U
- 2 for $k = 1, 2, \dots, T/B$ (对于每次迭代)

```

3   初始化缓冲区  $D_{\text{batch}} = \emptyset$ 
4   计算状态值  $V_{\theta_k}(s_t)$  和奖励  $R_t$ 
5   在状态  $s_t$ , 执行联合动作  $a_t = \{a_{i,t}\}_{i=1:N_s}$ , 得到下一个状态  $s_{t+1}$ 
6   采集样本轨迹  $\tau_k$ 
7   计算 TD 误差  $\delta_t$ 
8   计算优势估计  $\hat{A}_{\pi_{\theta_k}}$ 
9   计算值目标  $\hat{V}(s_t) = \hat{A}_{\pi_{\theta_k}} + V_{\theta_k}(s_t)$ 
10  存储  $\tau_k, V_{\theta_k}(s_t), \hat{A}_{\pi_{\theta_k}}, \hat{V}(s_t)$  到缓冲区  $D_{\text{batch}}$  中
11  更新  $\pi_{\theta_k}$  为  $\pi_{\theta_k}^{\text{old}}$ 
12  for epoch = 1, 2, ..., U (对于每个世纪)
13      随机排序缓冲区  $D_{\text{batch}}$  内的数据批, 并将其分成大小为 MB 的  $D_m^{\text{mini}}$ 
14      for m = 1, 2, ..., B/MB (对于每个最小批)
15          利用  $D_m^{\text{mini}}$  计算新旧策略差异  $d_t^{\theta_k}$ 
16          计算策略网络的替代目标  $L^{\text{actor}}(\theta_k)$ 
17          计算值函数误差  $L^{\text{critic}}(\theta_k)$ 
18          计算行动家-评论家网络的替代目标  $L^{\text{PPO}}(\theta_k)$ 
19          利用  $\nabla_{\theta_k} L^{\text{PPO}}(\theta_k)$  更新  $\theta_k$ 
20      end for
21  end for
22 end for
23 for t = 0, 1, ..., n (对于策略部署中每个时间步)
24     在状态  $s_t$ , 遵循策略  $\pi_{\theta}$  选择动作  $a_{i,t}$ 
25     for  $p_{ij} = 1, 2, \dots, M$  (对于每个冲突车辆对)
26         if 形式化验证到车辆 i 和车辆 j 即将碰撞
27             if  $p_i > p_j$ 
28                 更新  $a_j$  为  $a_j^{\text{con}}$ 
29             else
30                 更新  $a_i$  为  $a_i^{\text{con}}$ 
31             end if
32     end for
33     执行联合动作  $a_t = \{a_{i,t}\}_{i=1:N_s}$  并得到下一个状态  $s_{t+1}$ 
34 end for

```

算法的第一行初始化所构造的行动家-评论家网络 π 为 π_0 , 并设置超参数 $\alpha, \lambda, \gamma, \epsilon, T, B, MB, U$ 。第二行开启策略训练迭代的主循环, 它分为采样和更新两个阶段。第一阶段使用集中式策略 π_k 收集数据样本, 其中下角标 k 为当前的迭代次数, 以便清楚地表达训练过程(行 3~11)。首先, 算法清空缓冲区 D_{batch} (行 3)。其次, 计算价值函数 $V_{\theta_k}(s_t)$ 和时间步奖励 R_t , 即安全奖励、效率奖励和舒适奖励之和(行 4)。基于状态 s_t , 对所有智能体执行联合动作 $\{a_{i,t}\}_{i=1:N_s}$, 与环境交互得到下一个状态 s_{t+1} , 直到轨迹 τ_k 的长度达到 B (行 5~6)。接下来, 计算 TD 误差 δ_t 、优势估计 $\hat{A}_{\pi_{\theta_k}}$ 和值

目标 $\hat{V}(s_t)$, 然后存储 $\tau_k, V_{\theta_k}(s_t), \hat{A}_{\pi_{\theta_k}}, \hat{V}(s_t)$ 到 D_{batch} 中(行 7~10)。算法第二阶段的目标是更新策略 π_{θ_k} 。 D_{batch} 中的所有样本数据将使用 U 次(行 12~21)。在每个世纪中, 算法对 D_{batch} 中的数据批进行随机排序, 并将其分成大小为 MB 的最小批 $D_m^{\text{mini}}, m = 1, 2, \dots, B/MB$ (行 13)。随后, 通过 D_m^{mini} 对策略 π_{θ_k} 进行优化。计算新旧策略的差异 $d_t^{\theta_k}$ 和策略的替代目标函数(行 15~16)。接下来, 使用值目标 $\hat{V}(s_t)$ 和状态值 $V_{\theta_k}(s_t)$ 来计算评论家的损失(行 17)。最后, 计算行动家-评论家网络 π_{θ_k} 的损失函数 $L^{\text{PPO}}(\theta_k)$, 并利用 Adam 优化器更新网络参数向量 θ_k (步骤 18~19)。在 BCPPO 的部署过程中, 在每个时间步对每个车辆对 p_{ij} 进行形式化安全验证(行 23~25)。然后, 根据预定义的规则判定车辆的路权。对无路权车辆施加行为规约, 以避免碰撞事件发生(行 26~31)。随后, 所有车辆执行联合动作 $a_t = \{a_{i,t}\}_{i=1:N_s}$ 以获得下一个状态 s_{t+1} (行 32)。

4 实验设计

4.1 环境与参数设置

本文进行了两组实验。第一组实验利用 MAPPO 训练交叉口车辆协同控制策略, 对训练过程进行分析。第二组实验将 MAPPO、BCPPO 和基于 MPC 的基线方法 VICS^[18] 在计算效率、安全性、通行效率和舒适性方面进行比较。本文实验基于 Carla-0.9.13 (Windows 10) 仿真环境进行, 使用 i7-11700 CPU 训练神经网络和运行优化求解器。

为了高覆盖度还原真实的交叉口场景, 车辆的初始速度和位置是在一定范围内随机分布的。当有车辆发生碰撞或所有车辆均通过交叉口时, 该回合结束, 开启下一回合。车辆的期望行驶路线由高级的全局路径规划算法确定。车辆的横向运动由独立的路径跟踪模块控制。车辆均配备了 V2X 通信技术, 允许实时查询其位置和速度。仿真实验的参数设置见表 1。

4.2 训练过程分析

使用 MAPPO 训练交叉口车辆协同控制策略, 取 3 次训练的平均值作为实验结果。通过奖励值, 回合长度和碰撞率 3 个指标来对策略的训练过程进行分析, 如图 5 所示。

训练初始阶段, 策略无法成功消解车辆间的

表 1 仿真实验参数设置	
Table 1 Parameters for simulation	
参 数	值
离散时间步长/s	0.1 s
车辆数量 N_a	8
冲突车辆对数量 M	19
速度范围/($\text{m}\cdot\text{s}^{-1}$)	[5.5, 8]
缓冲区长度 d^b/m	25
行动家网络权重 k_{actor}	1
评论家网络权重 k_{critic}	-0.5
隐藏层数量	2
隐藏单元数量	512, 512
种子数	5
总时间步 T	2 048 000
批大小 B	2 048
最小批大小 MB	64
世纪 U	10
裁剪范围 ϵ	0.2
奖励衰减系数 γ	0.99
学习率 α	0.000 03 $\rightarrow$ 0
回报衰减系数 λ	0.95
碰撞事件奖励 k_c	-50
单辆车通过奖励 k_p^{one}	10
全部车通过奖励 k_p^{all}	50
安全时间阈值 T_s	1.2
干预时间阈值 T_c	4
形式化验证权重 k_t	-22
形式化验证偏差 b_f	2
时间步奖励 k_s	-0.1
舒适奖励权重 k_j	-4.5e-2
最大碰撞风险 H	1 000
风险权重 α_v	0.005
通行效率权重 w_v	1
舒适性权重 w_a	5
预测步数	16
期望速度/($\text{m}\cdot\text{s}^{-1}$)	8

冲突(如黄色阴影所示),回合开始后车辆立即发生碰撞,带来了极低的回合奖励和极短的回合长度,碰撞率几乎为100%。此时,算法倾向于控制车辆缓慢驾驶来延缓碰撞。在精细化设计的奖励函数的引导下,策略不断朝着更高奖励和更安全的车辆通行方向优化。经过约100次策略迭代,算法成功地控制车辆避免碰撞,奖励值显著提高,碰撞率迅速降低(如棕色阴影所示)。然而,此时的通行决策过于保守,交通效率较低。训练的中后期,算法继续优化策略,回合长度缓慢缩短,碰撞率进一步降低。最终,训练达到稳定,奖励值收敛到约75,回合长度稳定在约67,但仍存有约0.425%的碰撞率。

4.3 性能对比实验

将BCPPO的实验结果与无行为规约的MAPPO和基于MPC的交叉口协同方案VICS进行比较,以验证BCPPO在计算效率、通行安全性、通行

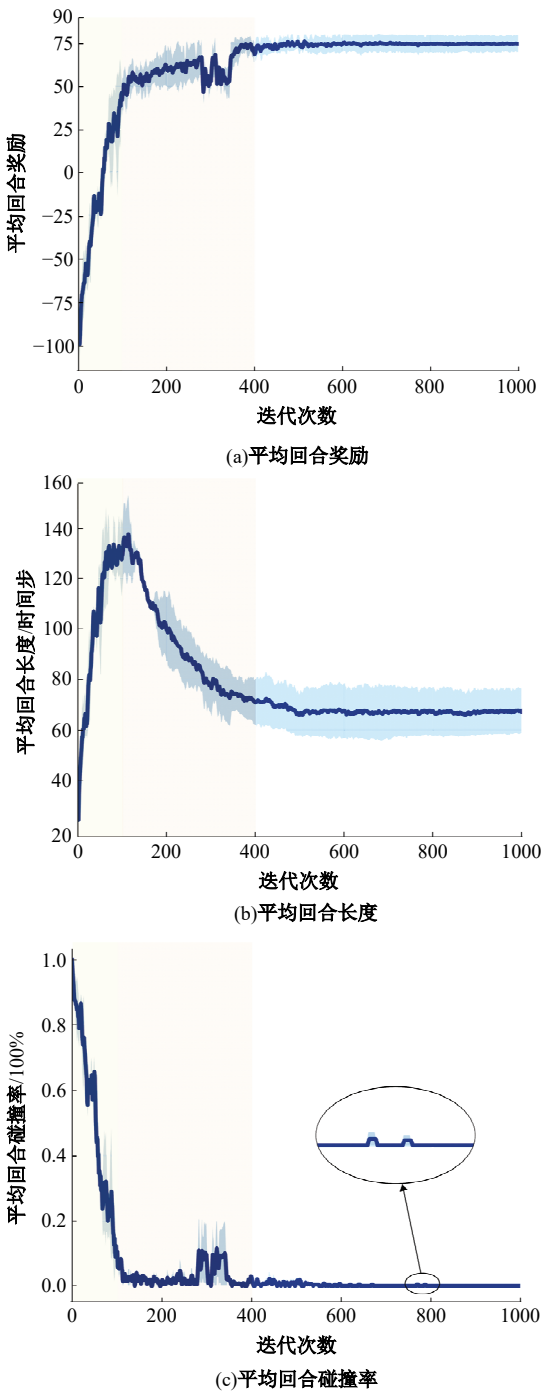


图 5 MAPPO 训练结果

Fig. 5 Results of MAPPO training

效率和舒适性方面的性能提升。由于VICS的求解速度对方法性能造成较大影响,在VICS的测试中,分别在Carla仿真软件的异步模式和同步模式下进行了测试。实验结果如表2所示。

计算效率方面,BCPPO和MAPPO均有着极高的计算效率。VICS由于其求解速度显著受到车辆数量和初始值准确度的影响,在大多数回合中其无法在短时间(如100 ms)内输出控制指令,

直接影响了异步模式下车辆行驶的安全性。

安全性方面,得益于安全性相关的评价指标及形式化行为规约的设定,同步模式下的 VICS 与 BCPPO 均获得了零碰撞率的实验结果。相比 MAPPO,BCPPO 所纳入的行为规约进一步保障了车辆通行的安全性。异步模式下的 VICS 受到求解速度的影响,需要平均 1.63 s 计算出车辆通行的控制指令,无法匹配车辆运动控制的实时性要求,导致车辆运动失控。因此,异步模式下的 VICS 碰撞率较大,达到了 63.5%。

表 2 交叉口车辆通行实验结果比较

Table 2 Comparison of experimental results for vehicle travelling

方 法	计算时间/s	碰撞率/%	通行效率/s	舒适性/($\text{m}\cdot\text{s}^{-2}$)
MAPPO	0.004	0.425	6.7	1.25
BCPPO	0.005	0	7.24	1.33
VICS(异步)	1.63	63.5	21.92	1.81
VICS(同步)	1.61	0	12.86	1.43

通行效率和舒适性方面,安全规约使车辆的驾驶行为变得更加保守,BCPPO 相比 MAPPO 牺牲了 8.06% 的通行效率和 6.4% 的舒适性。在异步模式 VICS 的通行实例中,由于控制指令的计算与发送不及时,车辆在交叉口中花费了大量的时间等待控制指令,导致通行效率和舒适性较差。BCPPO 相比异步模式 VICS 取得了 67.0% 的效率提升和 26.5% 的舒适性提升。在同步模式的 VICS 测试中,忽略了求解延迟对算法性能的影响,通行效率和舒适性相比异步模式取得了较大的提高。但由于 VICS 原论文将安全性视为目标函数中权重更大的子项,VICS 获得了性能有限的通行效率及舒适性。BCPPO 的通行效率和舒适性相比同步模式 VICS 分别提升了 43.7% 和 7.0%。值得指出的是,在真实世界中,系统是以异步模式的方式进行的,因此,异步模式的实验结果更具有实际应用的参考意义。

综上所述,纳入了必要的车辆行为规约后,BCPPO 相比 MAPPO 实现了更安全的交叉口车辆协同通行,同时只需牺牲很低的通行效率。相比 VICS,BCPPO 在计算效率、安全性、通行效率和舒适性方面均取得了显著的提高,能够更可靠地应用于交叉口车辆通行的现实场景。

5 结束语

本文提出了一种城市道路无信号交叉口

CAV 集中式协同控制方法,以解决当前集中式协同控制方法存在计算效率低和无安全保障的问题。首先,提出了一种多智能体强化学习方法 MAPPO,该方法将单智能体近端策略优化扩展到多智能体协同合作的复杂交互环境中。其次,将无信号交叉口 CAV 集中式协同控制形式化为 RL 问题,将形式化安全验证及行为规约融入 MAPPO,以指导策略安全迭代优化和避免非安全驾驶行为,形成安全增强的交叉口集中式协同控制方法 BCPPO。最后,在十字形交叉口场景的仿真环境中训练了协同控制策略,并将其在计算效率、通行效率和行驶安全方面的性能与基于 MPC 的协同方案 VICS 进行了比较。结果表明,基于 BCPPO 的交叉口车辆协同控制方法显著优于 VICS 方法。然而,此工作仍有发展改进的潜力。目前只考虑了车辆的纵向动力学,但在现实世界中,横向动力学和纵向动力学是耦合的,这可能导致次优的横向行驶舒适性。为了解决这个问题,计划在后续的工作中引入横向舒适性规约和代表车辆横向行为的状态,从而提高策略的行驶舒适性。

参考文献:

- [1] Du Y, Shang G W, Chai L G. A coupled vehicle-signal control method at signalized intersections in mixed traffic environment[J]. IEEE Transactions on Vehicular Technology, 2021, 70(3): 2089-2100.
- [2] 庄伟超,丁昊楠,董昊轩,等. 信号交叉口网联电动汽车自适应学习生态驾驶策略[J]. 吉林大学学报:工学版, 2023, 53(1): 82-93.
Zhuang Wei-chao, Ding Hao-nan, Dong Hao-xuan, et al. Learning based eco-driving strategy of connected electric vehicle at signalized intersection[J]. Journal of Jilin University(Engineering and Technology Edition), 2023, 53(1): 82-93.
- [3] Khoury J, Khoury J, Zouein G, et al. A practical decentralized access protocol for autonomous vehicles at isolated under-saturated intersections[J]. Journal of Intelligent Transportation Systems, 2019, 23(5): 427-440.
- [4] Karthikeyan P, Chen W, Hsiung P. Autonomous intersection management by using reinforcement learning[J]. Algorithms, 2022, 15(9): No. 326.
- [5] Chamideth S, Tarneberg W, Kihl M. A safe and robust autonomous intersection management system using a hierarchical control strategy and V2I communication[J]. IEEE Systems Journal, 2023, 17(1): 50-61.

- [6] Antonio G, Maria-Dolores C. Multi-agent deep reinforcement learning to manage connected autonomous vehicles at tomorrow's intersections[J]. IEEE Transactions on Vehicular Technology, 2022, 71(7): 7033–7043.
- [7] Fajardo D, Au T, Waller S T, et al. Automated intersection control[J]. Transportation Research Record: Journal of the Transportation Research Board, 2011, 2259(1): 223–232.
- [8] Dresner K, Stone P. A multiagent approach to autonomous intersection management[J]. Journal of Artificial Intelligent Research, 2008, 31(1): 591–656.
- [9] Gregoire J, Bonnabel S, Arnaud D. Optimal cooperative motion planning for vehicles at intersections[J/OL]. [2023–11–23]. <https://arxiv.org/abs/1310.7729>.
- [10] Lu G, Li L, Wang Y, et al. A rule based control algorithm of connected vehicles in uncontrolled intersection[C]//The 17th International IEEE Conference on Intelligent Transportation Systems, Qingdao, China, 2014: 115–120.
- [11] Zhang K, Arnaud D, Zhang D, et al. Analysis and modeled design of one state-driven autonomous passing-through algorithm for driverless vehicles at intersections[C]//The 16th International Conference on Computational Science and Engineering, Sydney, Australia, 2013: 751–757.
- [12] Arnaud D. Analysis of reservation algorithms for cooperative planning at intersections[C]//The 13th International IEEE Conference on Intelligent Transportation Systems, Funchal, Portugal, 2010: 445–449.
- [13] Li N, Kolmanovsky I, Girard A, et al. Game theoretic modeling of vehicle interactions at unsignalized intersections and application to autonomous vehicle control[C]//Annual American Control Conference, Milwaukee, USA, 2018: 3215–3220.
- [14] Wang H, Meng Q, Chen S. Competitive and cooperative behaviour analysis of connected and autonomous vehicles across unsignalised intersections: a game-theoretic approach[J]. Transportation Research Part B: Methodological, 2021, 149: 322–346.
- [15] Elhenawy M, Elbery A A, Hassan A A, et al. An intersection game-theory-based traffic control algorithm in a connected vehicle environment[C]//IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 2015: 343–347.
- [16] Zhao W, Liu R, Ngoduy D. A bilevel programming model for autonomous intersection control and trajectory planning[J]. Transportmetrica A: Transport Science, 2021, 17(1): 34–58.
- [17] Nair S H, Govindarajan V, Lin T, et al. Stochastic MPC with multi-modal predictions for traffic intersections[C]//IEEE 25th International Conference on Intelligent Transportation Systems, Macau, China, 2022: 635–640.
- [18] Kamal M A S, Imura J, Hayakawa T, et al. A vehicle-intersection coordination scheme for smooth flows of traffic without using traffic lights[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 16(3): 1136–1147.
- [19] Fink M. Implementation of linear model predictive control-tutorial[J/OL]. [2023–11–06]. <https://arxiv.org/abs/2109.11986>.
- [20] Zhou M, Yu U, Qu X. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: a reinforcement learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 20(1): 433–443.
- [21] Guo M, Wang P, Chan C Y, et al. A reinforcement learning approach for intelligent traffic signal control at urban intersections[C]//IEEE Intelligent Transportation Systems Conference, Auckland, New Zealand, 2019: 4242–4247.
- [22] Guan Y, Ren Y, Li S, et al. Centralized cooperation for connected and automated vehicles at intersections by proximal policy optimization[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 12597–12608.
- [23] Wang S, Wan Q. Right-turn driving decisions of autonomous vehicles at signal-free intersections [J]. Application Research of Computers, 2023, 40(5): 1468–1472.
- [24] Nordfjarn T, Simseloglu, O, Rundmo T. Culture related to road traffic safety: a comparison of eight countries using two conceptualizations of culture[J]. Accident Analysis and Prevention, 2014, 62: 319–328.
- [25] Zheng J, Zhu K, Wang R. Deep reinforcement learning for autonomous vehicles collaboration at unsignalized intersections[C]//IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022: 1115–1120.
- [26] Teh Y W, Bapst V, Czarnecki W M, et al. Distral: robust multitask reinforcement learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 4499–4509.