Journal Pre-proof

Nash double Q-based multi-agent deep reinforcement learning for interactive merging strategy in mixed traffic

Lin Li, Wanzhong Zhao, Chunyan Wang, Abbas Fotouhi, Xuze Liu

PII: S0957-4174(23)01960-7
DOI: https://doi.org/10.1016/j.eswa.2023.121458
Reference: ESWA 121458

To appear in: *Expert Systems With Applications*

Received date : 18 April 2023
Revised date : 2 September 2023
Accepted date : 2 September 2023

Please cite this article as: L. Li, W. Zhao, C. Wang et al., Nash double Q-based multi-agent deep reinforcement learning for interactive merging strategy in mixed traffic. *Expert Systems With Applications* (2023), doi: https://doi.org/10.1016/j.eswa.2023.121458.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier Ltd.

Journal Pre-proof

Expert Systems with Applications, Volume 237, Part B, March 2024, Article number 121458
DOI:10.1016/j.eswa.2023.121458

Nash Double Q-based Multi-Agent Deep Reinforcement Learning for Interactive Merging Strategy in Mixed Traffic

Lin Li[a], Wanzhong Zhao[a], Chunyan Wang[a], Abbas Fotouhi[b], Xuze Liu[b]

[a] Department of Vehicle Engineering, College of Energy and Powder, Nanjing University of Aeronautics and Astronautics, Yudao29, Nanjing, 210016, China
[b]Advanced Vehicle Engineering Centre, School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, Bedfordshire, MK43 0AL, UK

Abstract

The interaction between ramp and mainline vehicles plays a crucial role in merging areas, especially in the mixed-traffic environment. The driving behaviours of human drivers are uncertain and diverse, and the uncertainty makes it more complex for connected automated vehicles (CAV) to plan trajectories and merge into the mainline. To overcome this problem, a interactive merging strategy based on multi-agent deep reinforcement learning (MADRL) is designed, enabling the ramp vehicle (CAV) to consider the dynamic reaction of mainline vehicles. There are two agents in our interactive strategy, one of which is to predict and analyse the behaviour of mainline vehicles (human-driven vehicles, HDV, or non-connected vehicles). The other is created for exploring optimal merging actions of ramp vehicles. Firstly, game theory is used to model the competitive behaviours between ramp and mainline vehicles, and the Nash equilibrium of joint actions guides the ramp vehicle to learn best response to the mainline vehicle. Secondly, the Nash double Q algorithm is developed to ensure the outputs of Q networks are trained to efficiently converge to the Nash equilibrium point. The trained Q networks are then used for online control. Finally, our strategy is compared with single RL and existing MADRL algorithms in real on-ramp scenarios. Simulations show our strategy to be successful in coordinating both vehicles via analysis of human drivers, resulting in improved driving performance in terms of global safety, efficiency, and comfort.

Keywords:
Reinforcement learning, Trajectory planning, Deep reinforcement learning, Autonomous driving

1. Introduction

In the merging area on highways, the competition among ramp and mainline vehicles regarding right of way is the major cause of congestion and even collision. Previous research has suggested that connected and automated vehicles (CAVs) with merging control may improve traffic safety and efficiency for on-ramps (Chen and Englund, 2016). In a typical merging area, the ramp vehicle observes gaps in the mainline and seek to merge into the platoon, while the mainline vehicle assesses the situation to decide whether to yield or accelerate to pass. These two vehicles have multiple choices, and different combinations of their decisions will result in different situations, including serious crashes. For example, when the mainline and ramp vehicles both decide to accelerate to pass the merge area, without realizing the hazardous situations, they will collide . Existing studies have demonstrated that cooperative merging control can coordinate globally optimal merge sequence and trajectory to ensure a balance of benefit to both ramp and mainline vehicles (Fan et al., 2020). However, it will still take considerable time to achieve a fully connected traffic environment. Accordingly, in the very near future, it will be necessary to deal with the mixed traffic composed of CAVs and human-driven vehicles (HDVs) (Zhang et al., 2023). In contrast to autonomous driving systems which obey specific rules, human drivers tend to dynamically and randomly make their decisions dependent on the response of surrounding vehicles. When coupled with uncertain human driver behaviour , the interaction between CAVs and HDVs becomes more complex to model. However, without the consideration of interaction, the collision safety and global traffic efficiency are hard to guarantee. Taking this interaction into account when designing a cooperative control strategy, especially in mixed traffic remains a current challenge.

The on-ramp on highways is one of the major traffic bottlenecks, for the serious accidents, congestion and pollution it causes. The developments on merging control have attracted intensive attention (Rios-Torres and Malikopoulos, 2017). In the early stage, ramp metering was designed to control the ramp vehicle to merge only when the mainline is empty (Abdel-Aty et al., 2007). However, ramp metering displayed limited benefits since it sacrifices ramp efficiency. Recent works have shown an increased interest in cooperative merging control (CMC) which is aimed at cooperating the two traffic streams to safely and efficiently

Published by Elsevier. This is the Author Accepted Manuscript issued with: Creative Commons Attribution Non-Commercial No Derivatives License (CC:BY:NC:ND 4.0).
The final published version (version of record) is available online at DOI:10.1016/j.eswa.2023.121458. Please refer to any applicable publisher terms of use.

merge (Wang et al., 2013; Ntousakis et al., 2016). Compared with cooperative adaptive cruise control (CACC), which only considers longitudinal motion in the formulation of control, CMC couples lateral and longitudinal motion. This results in nonlinear characteristics in formulation of the control problem, which makes it hard to solve. To resolve the problem, the concept of "virtual vehicle" is proposed in the literature, e. g. (Scholte et al., 2022), which maps the ramp vehicle to the mainline. In this way, the merging control could be converted into a "car-following" problem. Another way is to divide the problem into the merging sequence and trajectory planning. Some studies (Yao et al., 2020) use the simple rules like "first-in-first-out" to schedule the merging sequences, without exploring the optimal sequence during merging control. To ensure optimal state and time of arrival of the vehicle at the merging point, (Jing et al., 2022) developed a hierarchical framework composed of independent lane control and trajectory planning control. The lane controller determines the merging sequence and merging states (final position, velocity and acceleration) according to the traffic flow, while dynamic trajectory planning relies on the lane decision and merging states, using a sigmoid function to represent the lane-changing trajectory. The merging control is easy to implement and validate by adjusting the specific parameters of the pre-defined function, whereas it is difficult to be adaptive with the actual traffic states. In (Zhou et al., 2019), two related optimal control problems are formulated to plan vehicle trajectories for automated on-ramp merging. Assuming that the mainline is congested, the first is to derive an optimal control for the mainline vehicle to develop a suitable gap to facilitate the ramp vehicle. The second problem is the trajectory planning controller to govern the ramp vehicle to merge into this gap. This method is more flexible with traffic management through treating the time and location of the merge point as the decision variables. To deal with the coupling relationship between merging sequence and trajectory planning, the planner in (Chen et al., 2022) first makes a prediction on ramp vehicles and decides dynamic sequences in each lane by minimizing overall disturbances to upstream mainline traffic. The dynamic sequences are then sent to the operational controller for ramp and mainline vehicles respectively. The above research attempted to separately consider the merging process, in this way, cooperative merge control can be formulated and solved based on existing optimal theory. However, as these methods required the speed be constant during lane-changing, the lateral and longitudinal operations cannot be planned at the same time. Moreover, the increase in number of vehicles will aggravate the computation burden with. Also, only the cost of either vehicle could can be considered in a controller, without equilibrium of the competing vehicles.

Another area of interest in current cooperative merging control is the design of objective functions. The objectives in these studies can be classified into two categories: local (for the single vehicle) and global(for multiple vehicles). The first focuses on local objectives, such as safety (Jing et al., 2021; Xiao et al., 2021; Peng et al., 2020), fuel consumption and emission (Heppeler et al., 2017), and passenger comfort. To enhance the safety in a dynamic traffic environment, temporal and spatial risk identification was considered and then minimised by the controller in (Wu et al., 2023; You et al., 2015). (Jeon et al., 2022) assumes that it is challenging but important to maximize efficiency while guaranteeing safety. Hence they use conditional state machines to classify the safe manoeuvres and set them as the safety constraints and a trajectory planner based on chance constrained model predictive control is formulated to optimize efficiency within the safety constraints. In (Li et al., 2022), different driving styles are evaluated through LSTM network and collision detection depends on the driver's risk acceptance, trajectory planner selects the optimal from trajectory cluster to maximize driver comfort. The major challenge in local optimization is reliant on how to coordinate multiple objectives and the above studies have proposed effective methods to deal with different driving performance indexes. However, these studies only investigated the benefit of a single vehicle, resulting in inequity of the other vehicles. Also they ignore the fact that a connected environment has the advantage of providing information about surrounding vehicles, which should be made best use of to find a globally optimal trajectory. Therefore, the second type tries to address the global environmentally sustainable, efficient and equitable coordinated merging control strategy. (Kotsialos, 2021) discussed a newly multi-class second-order macroscopic traffic model coupled with a pollutant emissions model. The global network-wide emission is treated as the objective, while local environmental requirements are set as the constraints. In terms of global traffic efficiency, (Uno et al., 1999) has adopted the sum of waiting and travel time of all vehicles as the optimization objectives based on model predictive control to derive the efficient ramp metering rates. (Liu et al., 2019) noted that although these works pay attention to the global traffic efficiency and environmental emissions using the traffic flow model, they fail to guarantee the pre-conditional safety. Thus, they proposed a cooperative scheduling mechanism to ensure safe driving while minimizing waiting time. A window searching algorithm is performed to find a collision-free trajectory with minimal delay, besides backup windows.

As shown in the above literature, though the local and global traffic performance are studied in these works, the conflicts between the competing vehicles are not explicitly modeled and analyzed. In most cases, irrespective of whether a ramp or mainline vehicle executes the maximum action, it will impact the driving utility of other vehicles and even the overall traffic.

Additionally, the literature are limited to a pure connected environment. Since the human-driven vehicle is not controllable, the two vehicles are hard to cooperate. However, the interactions among them are inevitable during planning. In mixed traffic, the ramp vehicle is required to

Journal Pre-proof

make decisions in uncertain environments, which is caused by human-driven vehicles and the decisions they make. The uncertain behaviors of the mainline vehicle make the interaction complex, and increase the difficulty for ramp vehicles to make predictions and analyze the utilities of both vehicles.

Game theory is a major method for modeling competing behaviors of interacting agents and offers suggestions based on instantaneous utility (Koller and Pfeffer, 1997). On the other hand, reinforcement learning is based on the Markov decision model, whose objective is to optimize the long-term return of the agents. Considering the motion planning problem has the dynamic and sequential characteristic, the action at current time has unignorable effects on the following decision. Hence, we tried to integrate the advantage of the game theory and reinforcement, to maximize the system utility through the interacting utility of each agent.

The main contribution of this work is its design of a novel interactive merging strategy for the connected automated vehicles (CAV) based on multi-agent deep reinforcement learning (MADRL). The main contribution of this paper is given as follows: 1) Aiming at uncertain and diverse driving behaviours of human drivers, a novel interactive merging strategy is proposed to enhance individual and global safety and efficiency in mixed traffic with connected automated vehicles and human driven vehicles. 2) Traditionally, the motion prediction and decision-making modules are independently treated, which ignores the interconnection between these two modules. In this paper, based on the multi-agent deep reinforcement learning, the analysis and prediction of other vehicles is integrated in the ego vehicle's decision-making framework. 3) Different from single agent q learning, the joint optimal strategy of multi-agent requires to consider response of the other agent. Nash double q learning algorithm is proposed to estimate the state-joint action values of vehicles, and then ensure that the value functions could converge to Nash equilibrium point. 4) the simulation of single RL and existing MADRL algorithms are conducted to compare the merging performance in real on-ramp scenarios.

The overall framework of the study is shown in Figure 1. A multi-agent reinforcement learning is formulated, where two agents are created, one for predicting and analysing the behaviours of mainline vehicle (human-driven vehicle), and the other for exploring optimal merging actions of ramp vehicle (connected automated vehicle). The agent in our study is cognitive, using game theory to model the competitive behaviours between ramp and mainline vehicles.

2. Literature Reviews

2.1. Mixed traffic

In this study, the mix traffic is composed of the CAVs and HDVs. The characteristic of the mixed traffic is different from pure CAV traffic. To fill this gap, this work (An et al., 2022) is aimed to model system dynamics for mixed traffic, which describes the impedance of CAVs' speed reduction. Based on the modelling, the quantified impacts of the speed changes on the overall mix traffic could be achieved. Considering the vacancy of comprehensive properties of the mixed traffic, (Wang et al., 2021) was the first to prove that a ring-road mixed traffic system with one CAV and multiple heterogeneous human-driven vehicles is not completely controllable, but is stabilizable under a very mild condition, based on Popov-Belevitch-Hautus (PBH) criterion. (Yao et al., 2023) presents the capacity model and stability condition of mixed traffic flow, based on the probability distribution of the CAVs platoon size, and also analyses how the maximum platoon size influences the capacity and stability. In (Khattak et al., 2022), a lane-management based strategy is proposed and then the outperformance of the strategy in terms of the operations, fuel efficiency, and safety are illustrated. It is observed that the mixed traffic in various scenarios has been investigated in previous research. In (Al-Turki et al., 2022), a cooperative-based method for signalised intersection control under an AV traffic environment is proposed to overcome stochastic driving behaviour of humans. For on and off ramp scenarios, this study (Zhu et al., 2022) evaluates how the different driving strategies impact on the congested patterns, safety, efficiency, and comfort. The existing works have dealt with the modelling and analysis of the mixed traffic, based on that the mixed traffic has also been studied in the various scenarios. However, the external disturbance of the human driver remains the biggest challenge in mixed traffic. In (Zhongcheng et al., 2022), a novel H∞ pinning approach is designed to deal with the delay of the network and the disturbance from HDVs. One major issue in the existing method is that human behaviour is modelled as a random variable, which ignores the cognition and intelligence of human drivers. That is the reason why we introduced multi-agent reinforcement learning in our studies.

2.2. Multi-agent reinforcement learning

The MARL is proposed to resolve the non-stationarity of the environment, especially when the number of agents grows (Bautista-Montesano et al., 2022; Madridano et al., 2021). Understanding and modelling the mainline vehicle dynamic and their interactive driving behaviour at on-ramp areas are crucial for the ramp vehicle to merge. Thus, we formulate the merging problem based on MARL framework. Several attempts have been made to apply the MARL on the intelligent transport system. For example, (Shou et al., 2022) develop a MARL-based enroute path choice policy considering interactions in the transportation networks. By introducing the knowledge-sharing protocol, a MARL-based traffic signal control method is proposed to speed up the convergence of the model (Li et al., 2021). To address the coupled vehicle navigation and traffic signal control, a multi-agent advantage actor–critic method with policy sharing is adopted via a design of multi-class RL

Journal Pre-proof

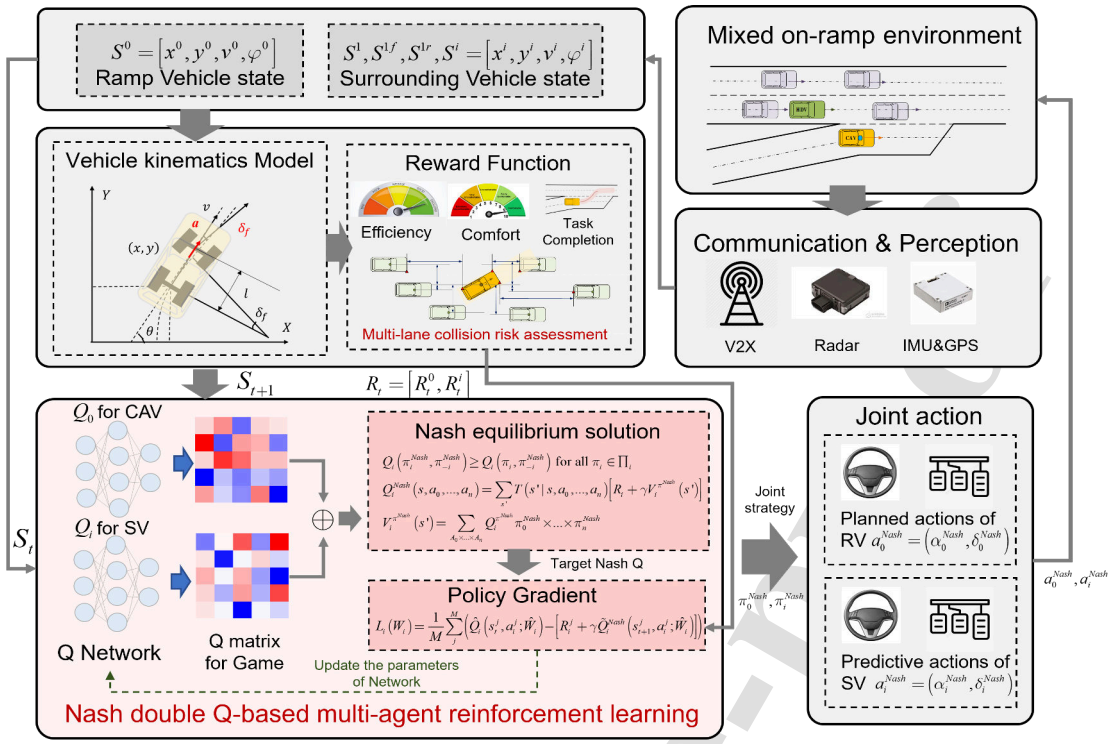


Figure 1: The Nash double Q-based multi-agent RL framework for on-ramp scenarios in mixed traffic.

agents and a pressure-based reward function (Su et al., 2022). For online parking systems, to generate efficient online parking assignment strategies with partial observations of parking demand, (Zhang et al., 2022) uses MARL to solve availability constraints imposed by the uncertain arrivals and departures of non-connected vehicles. There are also some works focusing on the on-ramp scenarios, such as (Schester and Ortiz, 2021; Zhou et al., 2022; Deng et al., 2019). Recent research has shown that the MARL is helpful to model and analyse the interaction among agents in the traffic environment. However, most existing works adopt a shared policy for multi-agents, which ignore the competition and possible cooperation of agents in a traffic environment. Therefore, we consider introducing game analysis in our MARL framework.

2.3. Game theory

Game theory is powerful tool to model interaction among players. The previous approach is limited to the one-way interaction where each vehicle adapts to the putative leader over lanes. This way may induce a drawback of larger deceleration. In ref (Fukuyama, 2020), a dynamic game-based vehicle trajectory optimization approach is proposed for CAV technologies. This paper (Talebpour et al., 2015) presents a lane-changing model that considers the dynamic interactions among drivers and the stochastic nature of lane-changing maneuver. To solve the failed double lane-changing (Ali et al., 2021), the gap acceptance behavior using a game theory approach is integrated into a target lane selection. In term of the ramp metering problem, based on a bargaining game approach (Heshami and Kattan, 2021), a local optimization problem based on expected local costs and constraints is formulated to achieve the control strategy. In the control and optimization of the intelligent vehicles, game theory is widely used to analyse the interaction among vehicles, driver-vehicles etc. However, game theory is a mathematical tool, which is not easy to learn and applied. For solving these problems, we attempt to use game theory to model and analyse the interaction among vehicles and employ reinforcement learning framework to simulate the environment and learning the optimal strategy.

3. Problem Statement

The aim of the interactive merging strategy in this paper is to plan the longitudinal and lateral action for ramp vehicles that try to merge into the mainline in mixed traffic. The studied scenario is shown in Figure 2. There is one on-ramp lane 0, acceleration lane and the mainline with multi-lanes, noted as 1, 2, ..., l. Considering the multi-lane scenarios, the mainline vehicles have more options, e. g. change to another lane to facilitate the ramp vehicle, not limited to longitudinal motion. The simulation starts when the ramp vehicle arrives at the conjunction point of the ramp and acceleration lane. The length of the acceleration lane is 100 meters. The width of the lanes is 3.5 meters, and the ramp vehicles need to complete merging

Journal Pre-proof

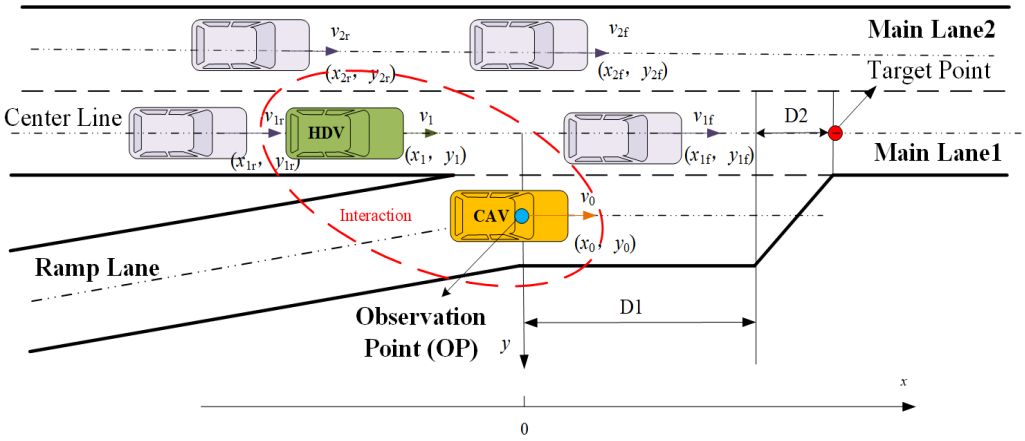


Figure 2: Research Scenario Description.

at 50 meters after the end of the acceleration lane. The first ramp vehicle v_0 is considered and controlled by our proposed algorithm. And the following vehicle on the lane l, closest to the ramp vehicle, is viewed as the competing v_1 vehicle in the mainline. The front and rear vehicles of v_1 in all mainline are regarded as the surrounding vehicle $v_{1,f}$ or $v_{1,r}$. The proposed algorithm is formulated in multi-agent reinforcement learning (MARL) framework, where one agent is responsible for exploring optimal strategy of ramp vehicles, and the motions of the mainline vehicles are predicted using other agents. Our algorithm is implemented based on the following assumptions: (1)Vehicle state information (e.g., speed, position, gap) can be shared and obtained by local CAVs through V2X communication or on-board sensors in real-time. (2)The sensing error (e.g., position error and velocity error) is not considered in this paper. (3)The communication delay of V2X is negligible in this study by the increasing maturity of the future generation of communication technology. (5)Lane-changing maneuvers are permitted for mainline vehicles.

4. Interactive merging strategy in mixed traffic based on Multi-agent reinforcement learning framework

This section moves on to the description of our interactive merging strategy based on multi-agent reinforcement learning (MARL). In motion planning problems, the vehicle states depend on current action and previous vehicle states. It is assumed that vehicle dynamics is consistent with the Markov property (Bellman, 1977). At each time step, the motion state is influenced by chosen driving action and current motion states. Assume the given current vehicle motion state is conditionally independent of all previous states. Inspired by single-vehicle reinforcement learning, MARL is modelled as an extension of Markov decision process, called Markov game decision process. What is different is that the vehicle must track other vehicle's states, driving actions and rewards to achieve a interactive driving action. Thus, the interactive merging decision process can be represented by a tuple,

$$G = (\mathcal{N}, \gamma, \mathcal{S}, \mathcal{Z}, \mathcal{O}, \mathcal{A}_0, , \mathcal{A}_n, R_0, , R_n, T) \tag{1}$$

where $\mathcal{N} := \{v_0, ..., v_i ..., v_n\}$ is the set of vehicles. Herein, 0 denotes ramp vehicle and i represents mainline vehicles. $\gamma \in [0, 1]$ is discount factor, indicates the influence of the future on the current decision. $s \in \mathcal{S}$ describes the true state of the on-ramps situations. At each time step t, the ramp vehicle chooses an action $a_0 \in \mathcal{A}_0$, and the mainline vehicle v_i is predicted to choose an action $a_i \in \mathcal{A}_i$, forming a joint action $(a_0, ..., a_n) \in \mathcal{A} := \mathcal{A}_0 \times ...\mathcal{A}_i \times ...\mathcal{A}_n$. This causes a transition on the traffic environment according to the state transition function $T(s_{t+1}|s_t, (a_0, ..., a_n)) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. All vehicles receive a reward $R_i : \mathcal{S} \times \mathcal{A} \to R$.

4.1. On-ramp States Description and Observations

In the decision process, the states of on-ramps are required to hold sufficient information for the ramp vehicle to make decisions. The system state of on-ramp scenarios S_t includes the motion states of ramp vehicle, and competing and surrounding vehicles at mainline,

$$S_t = \left[S_t^0, S_t^1, \left\{S_t^{l,f}, S_t^{l,r}\right\}_{l \in L}\right] \tag{2}$$

$$S_t^i = \left[x_t^i, y_t^i, v_t^i, \varphi_t^i\right] \tag{3}$$

where S^0 is the motion states of ramp vehicle. S^1 is the motion states of competing vehicle at mainline, defined as the closet vehicle at outermost lane. $S^{l,f}$, $S^{r,f}$ represent the front and rear vehicle on the lane l, respectively. L is a set of lanes at mainline. For each vehicle v_i, the motion states is composed of vehicle position $\left(x_t^i, y_t^i\right)$, velocity v_t^i, and heading angle φ_t^i.

4.2. Mixed Actions and Motion Transition Model

To couple the longitudinal and lateral motion, our action space is two-dimensional and continuous, consisting of

acceleration α_t and steering angle δ_t at time t. The action of vehicle at time t denotes as,

$$a_t^i = \left(\alpha_t^i, \delta_t^i\right) \tag{4}$$

where the acceleration ranges from -3 to $2m/s^2$. The steering angle ranges from -15° to 15°.

Actually, game theory is hard to be applied in continuous space. Thus, the final action space needs to be transferred into discrete space during the process. The expectation of the mix strategy is tried as the final action. At every time step, the action can be decomposed into five basis vectors $\varepsilon_{i,k} \in \mathcal{A}_i, k = 1, 2, 3, 4, 5$, corresponding to acceleration, deceleration, keeping, left-turn, right-turn.

Then we use strategy $\sigma_{i,k} : \varepsilon_{i,k} \rightarrow [0, 1]$ to represent the length of the projection of the final action onto each basis vector $\varepsilon_{i,k}$. Some of the $\sigma_{i,k}$ could be equal to 0, but the sum of all the $\sigma_{i,k}$ is equal to 1. The final action can be represented as the vector product of basis vector ε_k and length vector $\sigma_{i,k}$,

$$a_t^i = \sigma_{i,k} \times \varepsilon_{i,k}, \sigma \in \mathbb{R}^{1\times5}, \varepsilon \in \mathbb{R}^{5\times2} \tag{5}$$

Different from most literatures (Jeon et al., 2022), our vehicle is no longer treated as a point mass, improving the precision of the trajectory and ensuring safety. Since the time interval Δt of our planning is less than 1s, the kinematic equation can describe the motion transition precisely. When received an action a_t^i at time t, the next motion state of vehicle v_i is given as,

$$\begin{cases} x_{t+\Delta t} = x_t + v_{t+\Delta t} \cos\left(\theta_{t+\Delta t}\right) \Delta t \\ y_{t+\Delta t} = y_t + v_{t+\Delta t} \sin\left(\theta_{t+\Delta t}\right) \Delta t \\ v_{t+\Delta t} = v_t + \alpha_t \Delta t \\ \theta_{t+\Delta t} = \theta_t + v_{t+\Delta t} \tan\left(\delta_t\right) / l \Delta t \end{cases} \tag{6}$$

4.3. Rewards

The reward function reflects feedback of the environment by choosing an action from current state to next state, to measure the benefit of the chosen action. Many indicators should be considered in the reward function, and the most fundamental indicator to evaluate a driving action is collision safety. In terms of on-ramps scenario, the ramp vehicles have to complete merging tasks, reaching to the destination quickly and comfortably. Therefore, a driving task completion, efficiency and comfort indicators are designed.

4.3.1. Multi-Lane Collision Risk Assessment

The commonly used safety assessment indicators include Time-to-Collision (TTC), Time Headway (THW), Margin to collision (MTC) etc. (Dahl et al., 2019). Though these indicators are demonstrated to be effective in most work, they fail in our merging areas, since the existing indicators assume the vehicles are in the same lane and they ignore the lateral behaviours. To overcome this limitation, the longitudinal and lateral temporal safety indicators TTC, TTL are first used to measure the correlation degree k, and then the spatial indicators MTC are integrated to estimate the final multi-lane collision risk.

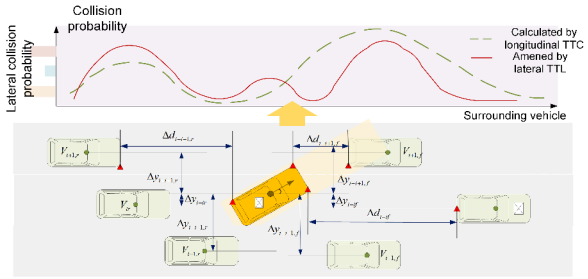


Figure 3: Multi-Lane Collision Risk Assessment Method.

Our designed collision risk assessment method is also illustrated in Figure 3. The dot line represents the collision probability that calculated by the TTC. While the solid line represents the amended collision probability considering the lateral TTL. It means that even if the longitudinal safety distance could be guaranteed, when the vehicles are very close in lateral, the collision probability will increase.

The safety reward is calculated as follows,

$$R_{i,t}^{safe} = \sum_j k_{i,j,t}^{lon.} k_{i,j,t}^{lat.} MTC_{i,j,t} \tag{7}$$

When the front vehicle decelerates to stop and the rear vehicle reacts to decelerate, MTC is defined as the sum of deceleration distance of proceeding vehicle and initial distance between proceeding and following vehicles, divide by the deceleration distance of the following vehicle,

$$MTC_{i,t} = \left(\frac{v_{f,t}^2}{2a} + \Delta d_{f,t}\right) \times \left(\frac{V_{r,t}^2}{2a} + v_{r,t} t_r\right)^{-1} \tag{8}$$

And then, the temporal indicators in the longitudinal and lateral directions are used to calculate the $k^{lon.}, k^{lat.}$, to evaluate how the surrounding vehicle is related to the ego vehicle. Time to collision, $TTC_{i,j,t}$, is defined as the time until a collision between vehicle v_i and v_j occurs at time t, to represent the longitudinal collision risk.

$$k_{i,j,t}^{lon.} = SoftMax(TTC_{i,j,t}) \tag{9}$$

$$TTC_{i,j,t} = \begin{cases} \frac{\Delta d_{i,j,t}}{v_{f,t} - v_{p,t}}, v_{f,t} \geq v_{p,t} \\ +\infty, v_{f,t} < v_{p,t} \end{cases} \tag{10}$$

Where $\Delta d_{i,j,t}$ is the initial distance between vehicle v_i and v_j at time t. $v_{p,t}$ is velocity of the proceeding vehicle between v_i and v_j, while $v_{f,t}$ is the velocity of following vehicle. Here, the subscription is not the same as previous. p means the proceeding vehicle between v_i and v_j. SoftMax is a normalised function, to achieve the risk probability of the surrounding vehicle related to the ego vehicle. Based on the definition of time to lane, herein, the $TTL_{i,j,t}$ is defined as the time to collide in lateral direction between v_i

and v_j, if they maintain lateral velocity at time t.

$$k_{i,j,t}^{lat.} = SoftMax(TTL_{i,j,t}) \tag{11}$$

$$TTL_{i,j,t} = \begin{cases} \frac{\Delta y_{i,j,t}}{v_{y,f,t} - v_{y,p,t}}, v_{y,f,t} \geq v_{y,p,t} \\ +\infty, v_{y,f,t} < v_{y,p,t} \end{cases} \tag{12}$$

where $\Delta y_{i,j,t}$ is the initial lateral distance of vehicle v_i and v_j. $v_{y,p,t}$ is lateral velocity of the proceeding vehicle between v_i and v_j, while $v_{y,f,t}$ is the lateral velocity of following vehicle. To better understand the effectiveness of multi-lane collision risk assessment, some simulation results are presented in the Figure 4. The first row contains safety rewards considering only the longitudinal distance. The x axis is the longitudinal position of the following vehicle, relative to the ego vehicle. and along the y axis is the longitudinal position of the proceeding vehicle. Figure 4(a) represents the situation where the velocity of the proceeding vehicle is less than the ego vehicle, meanwhile the velocity of the following vehicle is larger than the ego vehicle. It means the ego vehicle should keep the distance from the proceeding and following vehicles at the same time, as shown in the green area (risk less than 0.5) of Figure. 4(a). to ensure safety. Figure 4(b) represents the situation where the velocity of the proceeding vehicle is larger than the ego vehicle, meanwhile the velocity of the following vehicle is also larger than the ego vehicle. Therefore, the risk is only dependent on the following vehicle. Figure 4(c) represents the situation where the velocity of the proceeding vehicle is less than the ego vehicle, meanwhile the velocity of the following vehicle is also less than the ego vehicle. In this situation, the collision with the proceeding vehicle is more likely to happen if the proceeding vehicle is very close, less than 12 meters shown in the Figure 4(c). The velocity difference is larger in Figure 4(b), so the risk area is smaller in Figure 4(b). And then, the second row contains our designed safety rewards. When the obstacle vehicle is offset in lateral direction, the reward is adjusted. From this, the vehicles in multi-lanes could be assessed at the same time. Figure 4(d) shows the risk of the following vehicle which is right offset to the ego vehicle. And the risks of the proceeding vehicle are shown in Figure 4(e), (f). It is observed that the lateral position of the proceeding vehicle is equal to the ego vehicle, so that the risk is symmetrical in Figure 4(e). While the proceeding vehicle in the Figure 4(f) is left offset relative to the lateral position of the ego vehicle.

4.3.2. Driving Efficiency

The vehicles always expect to reach the destination with least time. Therefore, on the premise of safety, the decision-making system should also be encouraged to improve the velocity as much as possible. The efficiency of the selected action is evaluated by considering the time that the vehicle will take to reach the target point maintaining current action. The target point is selected as TP shown in the Figure 2. And the reward of efficiency can be obtained by the following equations.

$$R_{i,t}^{eff} = exp\left(-T_i^a\right) \tag{13}$$

$$T_i^a = \frac{-v_i + \sqrt{v_i^2 + 2a\left(D1 + D2 - x_i\right)}}{a} \tag{14}$$

where a is acceleration. v is actual velocity. x is current longitudinal position. D is the distance to target point.

4.3.3. Driving Comfort

Considering that the more stable the vehicle is, the higher the ride comfort of passengers is. Therefore, the change value of heading angle is used to quantify the driving comfort. The comfort function is defined as follows,

$$R_{i,t}^{com} = \exp\left(-m * abs\left(\dot{\varphi}_t^i\right)\right) \tag{15}$$

where $\dot{\varphi}_t^i$ is the yaw rate of vehicle v_i. m is the adjustment coefficients of yaw rate.

4.3.4. Driving Task Completion

Different from free lane-changing on highways, the merging behaviours in on-ramps are mandatory. If the ramp vehicles fail to complete lane-changing at the specified position, they will hit the road boundary. Since the driving effiency reward only cares about the longitudinal states. Therefore, it is necessary to design an indicator in the reward function to guide ramp vehicles to merge into the mainline. The errors between the lateral position, heading angle of ramp vehicles and the centerline of the main lane are used to judge whether the ramp vehicles have successfully merged into the mainline. The closer the ramp vehicles are to the mainline and the closer the heading angle is to zero, the greater the rewards are.

$$R_{i,t}^{task} = exp\left(-E_i^{lane}\right) \tag{16}$$

$$E_i^{lane} = \omega_y(y_i - y_{tar})^2 + \omega_\varphi(\varphi_i)^2 \tag{17}$$

Totally, the whole reward is given by the weighted sum of all the reward,

$$R_{i,t} = \omega_i^{safe} R_{i,t}^{safe} + \omega_i^{task} R_{i,t}^{task} + \omega_i^{eff} R_{i,t}^{eff} + \omega_i^{com} R_{i,t}^{com} \tag{18}$$

where $R_{i,t}^{safe}$ is the reward of collision safety; $R_{i,t}^{task}$ is the reward of driving task completion;$R_{i,t}^{com}$ is the reward of comfort; $R_{i,t}^{eff}$ is the reward of efficiency. And the weighting coefficients of the four different driving characteristics can be adjusted according to the driving style. The analysis of the efficiency and task completion rewards is also provided in the Figure 5. It shows that the task reward increases with the decrease of the error to the target lane centre. In terms of the efficiency reward, when the ego vehicle is closer to the target position and the velocity is larger, the efficiency is higher.

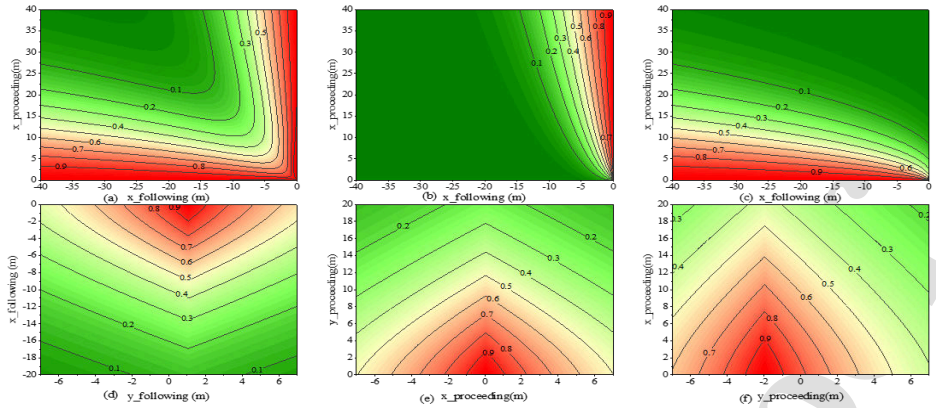
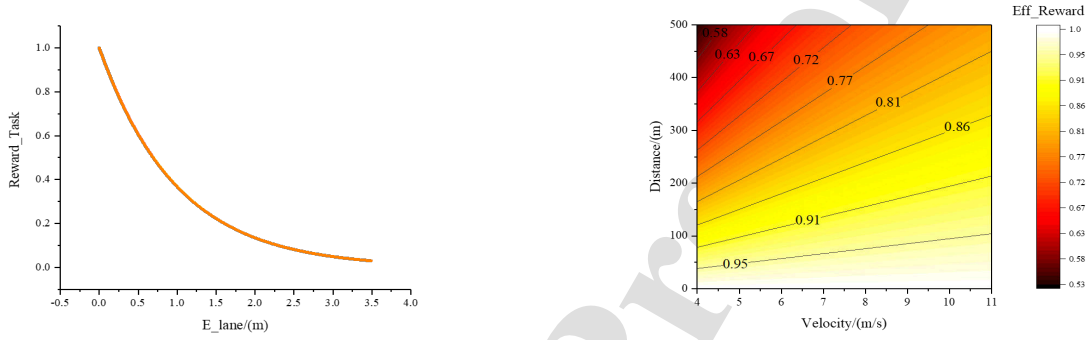


Figure 4: The analysis of the multi-lane collision assessment model.



Task Completion Driving Efficiency

Figure 5: Reward Map.

5. Optimal Merging Strategy based on Nash Double Q-Learning

The previous section defines the way ramp vehicles interact with the traffic environment and other surrounding vehicles. After numerous simulations of the interaction model, the experience tuples could be collected for learning, in the form,

$$(S_t, a_0, ...a_n, S_{t+1}, R_0, ...R_n) \tag{19}$$

In this section, we will investigate how to learn a interactive merging strategy in the presence of competing vehicles from a long-term perspective. In our merging strategy, the ramp and mainline vehicle are both viewed as agents, they have their own objectives, and these objectives tend to be contradicted. Thus, game theory is introduced to enable the agent to learn to coordinate by themselves.

5.1. Interactive strategy and long-term value assessment

In multi-agent system, the interactive strategy π is defined as the probability distribution of joint actions $(a_0, ...a_n)$, given certain states.

$$\pi = \pi_0 \times ... \times \pi_n : S \times \mathcal{A}_0 \times ... \times \mathcal{A}_n \rightarrow [0, 1] \tag{20}$$

For vehicle v_i, to measure the long-term quality of states and actions, Q-function is established, known as future expected return of current joint action $(a_0, ...a_n)$.

$$\begin{aligned} Q_i^{\pi}(s, a_0, ...a_n) &= E\left[R_{i,t}^{\pi} | S_t, a_{0,t}, ...a_{n,t}\right] \\ &= E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} R_{i,\tau} | S_t, a_{0,t}, ...a_{n,t}\right] \end{aligned} \tag{21}$$

In the problem with finite states, the Q table is constructed to store the Q value corresponding to the states and actions. However, the vehicle state is continuous, the states and action are numerous. It will be time-consuming to maintain or query the q value, if we still rely on the Q table. Thus, the Q value is estimated based on the neural network,

$$\hat{Q}_i(s, (a_0, ...a_n); W_i) \approx Q_i^{\pi}(s, a_0, ...a_n) \tag{22}$$

5.2. Nash Equilibrium based optimal strategy

In single-agent RL, the vehicles only care about their own return. Thus, given current states, the action corresponding to maximum Q value will be selected. However, in multi-agent system, if each vehicle pursues to maximizing its own return, the globally optimality is not guaranteed. We hope to see the vehicles could coordinate with

each other and drive in a interactive way. Nash equilibrium, proposed by (Nash, 1950), defines the best response to the current situation, where no one has anything to increase his return by change his own strategy. Similar to the Bellman optimal theory (Bellman, 1977), it refers to the expected sum of future discount returns when all vehicles start to implement the Nash equilibrium strategy from the next step. As mentioned in previous, our merging game needs to formulate under mix strategy. Formally, let us define the Nash equilibrium under mix strategy in the Q learning. Firstly, let $\prod_i$ be the set of all possible strategies for vehicle v_i, where $i = 0, 1, ..., n$. Let $\pi^{nash} = \left(\pi_i^{nash}, \pi_{-i}^{nash}\right)$ be a strategy profile, a set consisting of one strategy for each vehicle, where π_{-i}^{nash} denotes the n-1 strategies of all the vehicles except v_i. The strategy profile π^{nash} is a Nash equilibrium if,

$$Q_i\left(\pi_i^{nash}, \pi_{-i}^{nash}\right) \geq Q_i\left(\pi_i, \pi_{-i}^{nash}\right) \text{for all } \pi_i \in \prod_i \tag{23}$$

According to Bellman (Bellman, 1977), for vehicle vi, when all the vehicles execute the joint Nash equilibrium strategy $\pi^{nash} = \left(\pi_0^{nash}, ..., \pi_i^{nash}, ..., \pi_n^{nash}\right)$, the Nash-q value is defined as the sum of current reward and future expected return,

$$\begin{aligned} Q_i^{nash}(s, a_0, ...a_n) &= E\left[R_{i,t}^{\pi^{nash}} | S_t, a_{0,t}, ...a_{n,t}\right] \\ &= E\left[R_{i,t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{i,t+k+2} | S_t, a_{0,t}, ...a_{n,t}\right] \\ &= \sum_{s'} T(s'|s, a_0, ...a_n)\left[R_i(s', a_0, ...a_n) + \gamma V_i^{\pi^{nash}}(s')\right] \end{aligned} \tag{24}$$

where $V_i^{\pi^{nash}}(s')$ is the state value after Nash equilibrium strategy,

$$V_i^{\pi^{nash}}(s') = \sum_{A_0 \times ... A_n} \begin{array}{l} Q_i^{\pi^{nash}}(s', a_0, ...a_n)\, \pi_0^{nash}(s', a_0) \\ \times ... \pi_n^{nash}(s', a_n) \end{array} \tag{25}$$

Prior to discuss the existence of the Nash equilibrium, the following theorems are given.

Theorem 1. Every game with finite strategy has a mixed strategy equilibrium (Nash, 1950).

Theorem 2. For the Nash-Q learning algorithm, if the global optimum can be found in a certain state, then the strategy will converge to the Nash equilibrium strategy. Based on these two theorems, it can be guaranteed that the Nash-Q learning algorithm can converge to Nash equilibrium under mix strategy. In our method, the Lemke–Howson (LEMKE et al., 1964) algorithm is used to compute the Nash equilibrium strategy.

Considering the game $G_{i,j}$ formulated by vehicle v_i, v_j, and their mix strategy $\pi_i.\pi_j$, Q matrix $Q_i(s), Q_j(s) \in \mathbb{R}^{n(\mathcal{A}_i) \times n(\mathcal{A}_j)}$, Two polyhedron V, W related to the game $G_{i,j}$ are introduced,

$$\begin{aligned} V &= \left\{\pi_i | \pi_{i,k} \geq 0, \pi_i^T Q_j(s) \leq 1\right\}, \pi_i \in \mathbb{R}^{n(\mathcal{A}_i) \times 1}, \pi_{i,k} \in \pi_i \\ W &= \left\{\pi_j | Q_i(s)\, \pi_j \leq 1, \pi_{j,r} \geq 0\right\}, \pi_j \in \mathbb{R}^{n(\mathcal{A}_j) \times 1}, \pi_{j,r} \in \pi_j \end{aligned} \tag{26}$$

where V, W are respectively the polyhedron in the $\mathbb{R}^{n(\mathcal{A}_i) \times 1}, \mathbb{R}^{n(\mathcal{A}_j) \times 1}$, they are limited by $n(\mathcal{A}_i) + n\left(\mathcal{A}_j\right)$ inequality. It is easy to see that each Nash equilibrium strategy $\left(\pi_i^{Nash}, \pi_j^{Nash}\right)$ is a vertex of the polyhedron. The following parts will illustrate how to find the vertex corresponding to the Nash equilibrium. Firstly, label these inequalities from 1 to $n(\mathcal{A}_i) + n\left(\mathcal{A}_j\right)$,

$$L = \left\{1, 2, ...n(\mathcal{A}_i), n(\mathcal{A}_i) + 1, ...n(\mathcal{A}_i) + n\left(\mathcal{A}_j\right)\right\} \tag{27}$$

Remark. Each vertex v of V is associated with the label set L as follows. For $k \in n(\mathcal{A}_i)$, vertex v gets the label k if $\pi_{i,k} = 0$. For $r \in n\left(\mathcal{A}_j\right)$, vertex v gets the label $n(\mathcal{A}_i) + r$ if $Q_{j,1,r}(s)\, \pi_{i,1} + ... Q_{j,n(\mathcal{A}_i),r}(s)\, \pi_{i,n(\mathcal{A}_i)} = 1$. Similarly, each vertex w of W is associated with the label set L as the same rule. Note that the origin, which is a vertex of both V and W, has the labels $\{1, ..., n(A_i)\}$ and $\left\{n(\mathcal{A}_i) + 1, ..., n(\mathcal{A}_i) + n\left(\mathcal{A}_j\right)\right\}$.

Definition 1. [completely labeled] (Herings and Peeters, 2009) Consider pairs of vertices $(v, w), v \in V, w \in W$. We say that (v, w) is completely labeled if the sets associated with v and w contain all labels in L.

Definition 2. [Drop] (Herings and Peeters, 2009) A pivot operation consists of taking some pair (v, w) and replacing v with some vertex adjacent to v in V, or alternatively replacing w with some vertex adjacent to w in W. This has the effect (in the case that v is replaced) of replacing some label of v with some other label. The replaced label is said to be dropped.

Based on the above concept, the algorithm then starts at the completely labeled pair (v, w) consisting of the pair of origins. Through the pivot and drop operation, a completely labeled pair (v^*, w^*) can be find, except the origin. The found (v^*, w^*) corresponds to a pair of unnormalized probability distributions. Normalizing these values to probability distributions, we have a Nash equilibrium strategy $\left(\pi_i^{\text{Nash}}, \pi_j^{Nash}\right)$.

5.3. Nash double Q-Learning process

Unlike supervised learning, our learning algorithm does not rely on human labeling. Instead, to train the Q network, we acquire the driving samples by interacting with the traffic environment and surrounding vehicles. The flowcharts of the online control and offline training are presented in the Figure 6. And the pseudo-code is given in Table 1. Every time when all the agents execute the joint driving action $\left(a_0^j, ..., a_n^j\right)$, a piece of driving sample will be generated and then be stored into a driving experience pool.

$$M = \left\{S_t^j, a_0^j, ..., a_n^j, S_{t+1}^j, R_0^j, ..., R_n^j\right\}_{j \in N} \tag{28}$$

where j-th sample records one transition in which all the vehicles execute. N is the maximum number of the experience pool.

During training, a batch of driving samples are taken from the experience pool M, used for the following Q network updating accordingly. In addition, the Q network is

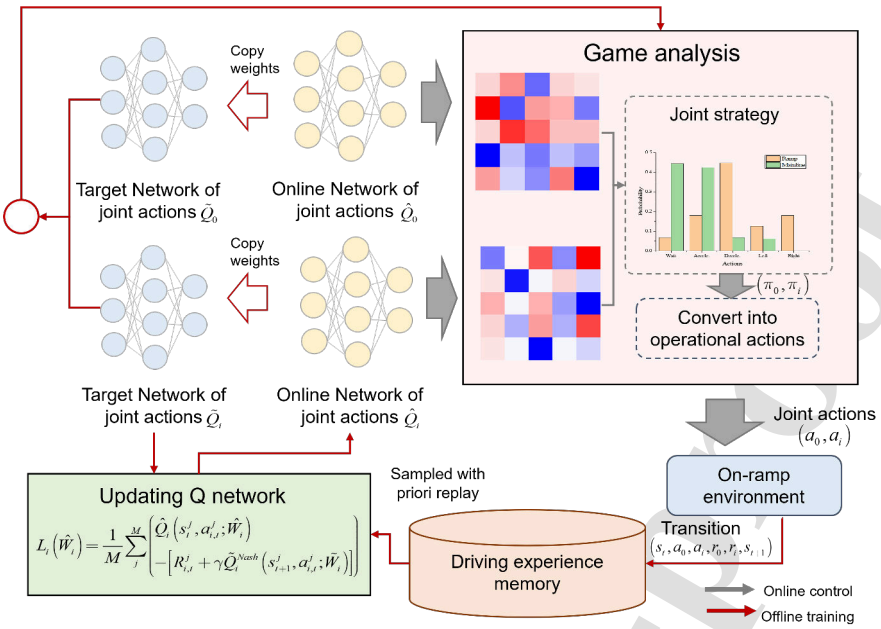


Figure 6: The algorithm structure of Nash double Q.

Table 1: The pseudo-code of Nash Double Q-learning

Algorithm 1 Nash Double Q-learning
1. Initialize replay buffer: M
2. Initialize online network $\hat{Q}_i(S)$, target network $\tilde{Q}_i(S)$ and parameters θ_i, θ_i^- for agent i
3. Start
4. For episode = 1, M do
5. Initialized
6. For t=1,T do
7. For $i \in N$ do
8. Sample joint action $(a_{0,t}, a_{1,t})$ in Nash equilibrium joint strategy $(\pi_0^{Nash}, \pi_1^{Nash})$ from current state
9. Execute joint action $(a_{0,t}, a_{1,t})$ and observe next state and reward $(R_{0,t}, R_{1,t})$
10. Store driving experience $S_t, a_{0,t}, a_{1,t}, S_{t+1}, R_{0,t}, R_{1,t}$ in reply buffer M
11. End
12. $t \leftarrow t+1, k \leftarrow k+1, S_t \leftarrow S_{t+1}$
13. If $t = T$ then
14. Initialzed $S_0, t \leftarrow 1$
15. End
16. If k>|M| then
17. Sample batch $S_t^j, a_{0,t}^j, a_{1,t}^j, S_{t+1}^j, R_{0,t}^j, R_{1,t}^j$ from M
18. $\tilde{Q}_i(S)$ = TargetNetwork($S_{t+1}^j; \theta_i^-$)
19. Compute Nash equilibrium value $\tilde{Q}_i^{Nash}$ and strategy π_i^{Nash}
20. $Q_{i,j}^* = R_i^{j+1}$, if $j$isterminaltime; otherwise :$R_i^{j+1} + \gamma \tilde{Q}_i^{Nash}$
21. Update online network by minimizing the loss: $\left(Q_{i,j}^* - \hat{Q}_i(S_i; \theta_i)\right)^2$
22. Every $k\%C = 0 : \theta_i^- \leftarrow \theta_i$
23. End

Journal Pre-proof

trained based on Epsilon-greedy strategy, to integrate the exploration and exploits. Specifically, at first, the driving action is totally generated at random. As the training continues, this probability will continue to decay, and finally decay to threshold probability, such as ten percent. That is to say, there is a ninety percent probability to execute the equilibrium strategy according to the learned Q network. The objective of the learning is to make the output of the Q network converge to the Nash equilibrium, For the driving data batch j, the target value of the Q network is,

$$y_i^j = R_i^j + \gamma \hat{Q}_i^{nash}\left(S_{t+1}^j, a_i^j; W_i\right) \tag{29}$$

$$\begin{aligned} L_i(W_i) &= \sum_j^{batchsize}\left(\hat{Q}_i\left(S_t^j, a_i^j; W_i\right) - y_i^j\right) \\ &= \sum_j^{batchsize}\left(\hat{Q}_i\left(S_t^j, a_i^j; W_i\right) - R_i^j - \gamma \hat{Q}_i^{nash}\left(S_{t+1}^j, a_i^j; W_i\right)\right) \end{aligned} \tag{30}$$

It can be seen from the loss function (30) that the Q network is either used as target or current output. During the practice, it will result in the instability of the sample data and Q-network. To relieve the problem, we introduce another network that has the same structure as Q-network, denoted as Target Q network $\tilde{Q}_i$. And the original Q network is called Online Q network $\hat{Q}_i$. At the beginning of training, the network $\tilde{Q}_i, \hat{Q}_i$ are assigned with the same parameters $\tilde{W}_i, \hat{W}_i$. For each sample data j, the online network $\hat{Q}_i$ is used to estimate the quality of sampled states and actions. While target network $\tilde{Q}_i$ is used to calculate the Nash equilibrium. And then, according to the following loss function, the parameters of the online network $\hat{Q}_i$ are updated. The parameters of the target network $\tilde{Q}_i$ are fixed, and they are only updated at certain episodes. Therefore, the following equation is used as our loss function,

$$L_i(W_i) = \frac{1}{M}\sum_j^M\left(\begin{array}{l}\hat{Q}_i\left(S_t^j, a_i^j; \hat{W}_i\right) - \\ \left[R_i^j + \gamma \tilde{Q}_i^{nash}\left(S_{t+1}^j, a_i^j; \tilde{W}_i\right)\right]\end{array}\right) \tag{31}$$

6. Simulations

The main purpose of the simulation is to verify the efficacy of Nash Double Q-learning based MARL algorithm and driving performance tested in real on-ramps scenarios. All algorithms are implemented in PyTorch on the ASUS computer (RAM: 16GB, processor: AMD Ryzen 9 6900HX, GPU: NVIDIA GeForce 3060 Laptop, Operating frequency: 3.30 GHz). We first verify the established interactive merging model and provide some game analysis. Then, the convergence of the Nash Double Q-learning MARL algorithm is compared with the baselines. We also present the planned results and demonstrate the driving performance when our methods are applied in real on-ramp scenarios.

6.1. Game for one step

The simplest merging scenario is used to verify the interactive merging model, proposed in section 4. The scenario is shown in Figure 7. where only one mainline and a ramp lane with acceleration lane connected are included. There are two vehicles set in the scenario, one mainline vehicle and one ramp vehicle. In verification, we only consider one step evolution, when these two vehicles execute different pairs of actions. Specifically, provide that the mainline vehicle chooses to accelerate, the situation for each action taken by ramp vehicle is shown in Figure 8,9,10,11,12, together with the reward analysis.

In order to explain the work of this part more clearly, five kinds of game situations are listed below, in which the main road vehicle accelerates, while the ramp vehicles adopt other strategies, and the game results are analysed according to the mentioned efficiency, task, safety and comfort as indicators. Figure 8 represents the waiting behaviour of the RV, Figure 9 represents the acceleration behaviour of the RV, Figure 10 represents the deceleration behaviour of the RV, Figure 11 represents the right lane change behaviour of the RV, and Figure 12 represents the left lane change behaviour of the RV. As can be seen from Figure 8, 9, 10, the reward trend of the two vehicles in various indicators is generally the same. It can be seen from Figure 8, the RV's waiting behaviour has led to a slight decrease in its task index, and the overall income of both vehicles is relatively similar; In Figure 10, the safety indicators of the two vehicles are significantly reduced due to the deceleration behaviour of the RV, resulting in close distance of these two vehicles, and other rewards are almost the same. On the whole, the total income of the two vehicles is relatively high, reaching about 17.19, 17.65, 16.42 for RV, 18.67,19.05, 18.03 for MV respectively. In Figure 11, due to the left turn operation of the RV, the actual efficiency and purpose indicators of the two vehicles have remained unchanged, and the safety indicators have significantly decreased. It is obvious that, in Figure 11, the distance of these two vehicles is less than 4 meters, which means that the collision occurs. At the same time, the RV has not slowed down in time, and the left turn action has also led to a decline in comfort. The overall indicators of the two vehicles have dropped to about 2.9 for RV, and 12.8 for MV. In Figure 12, the RV's right turn operation leads to the collision with road boundary and contradiction with the merging task. While the MV in this situation is not affected. The overall indicators of the two vehicles are -8.11, 20 respectively.

Through the above analysis, MV can also adopt five driving behaviours. Analyse the driving behaviours that can be used by the two vehicles and analyse their respective benefit results. The results are shown in Figure 13. Through solving the Nash equilibrium, the best reaction of both vehicles of the merging game is found and marked in the Figure with a star. Figure 14 shows the reward radar of MV and RV at the marked point. At this time, the MV takes deceleration action and the RV takes acceleration

Journal Pre-proof

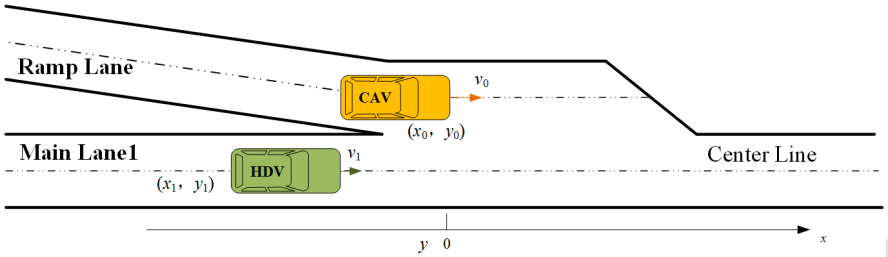


Figure 7: The on-ramp scenario: a ramp vehicle and a competitive vehicle on mainline.

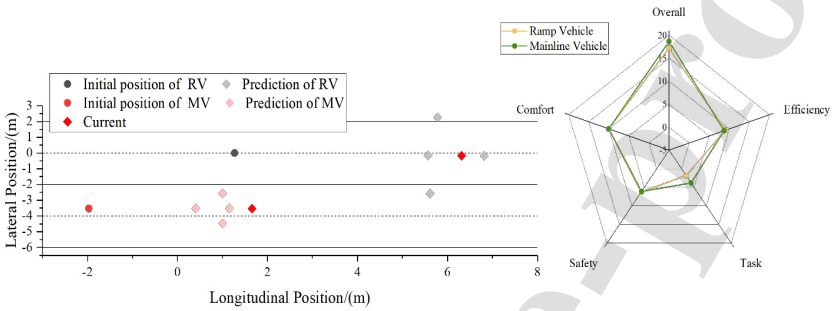


Figure 8: The first situation: positions after one step and reward.

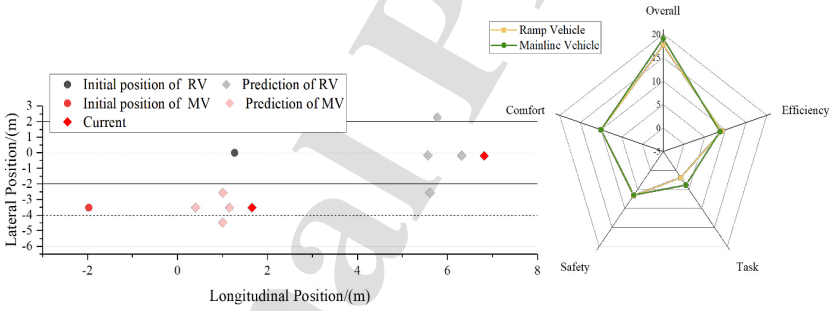


Figure 9: The second situation: positions after one step and reward.

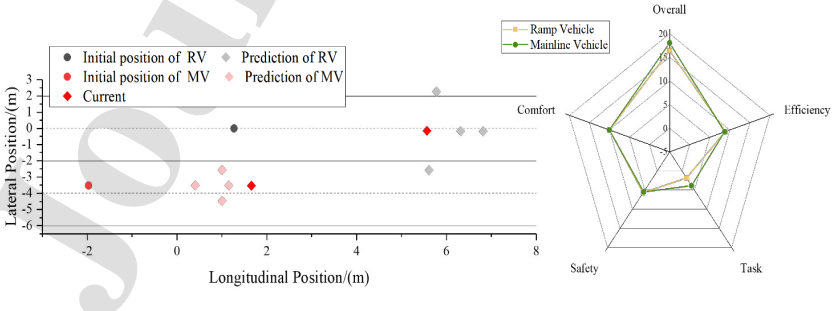


Figure 10: The third situation: positions after one step and reward.

Journal Pre-proof

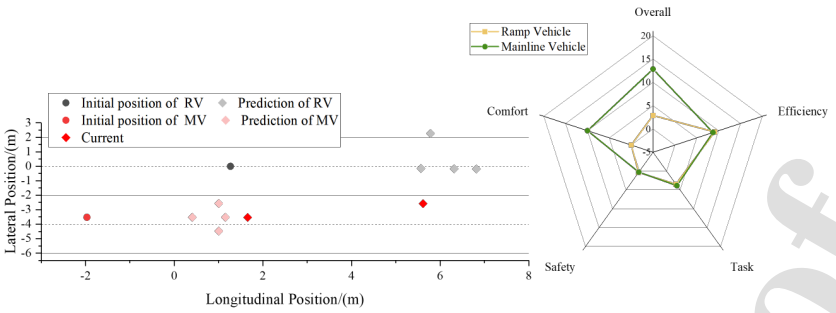


Figure 11: The forth situation: positions after one step and reward.

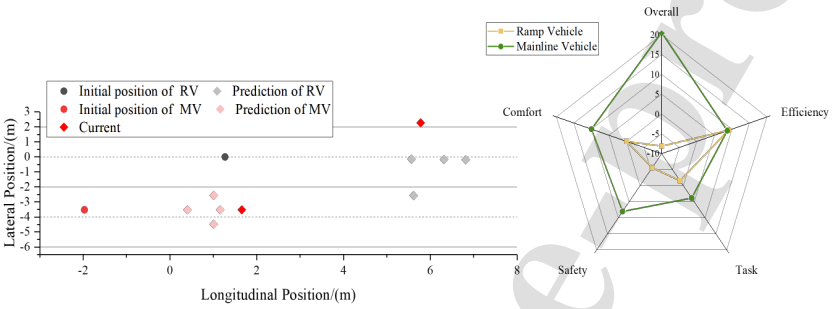


Figure 12: The fifth situation: positions after one step and reward.

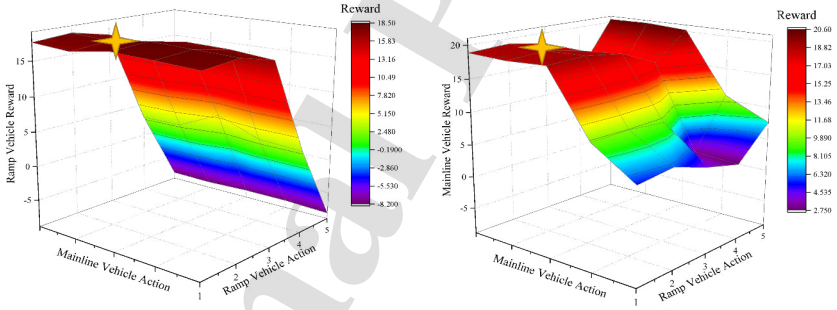


Figure 13: The reward maps.(a)Ramp vehicle(b)Mainline vehicle.

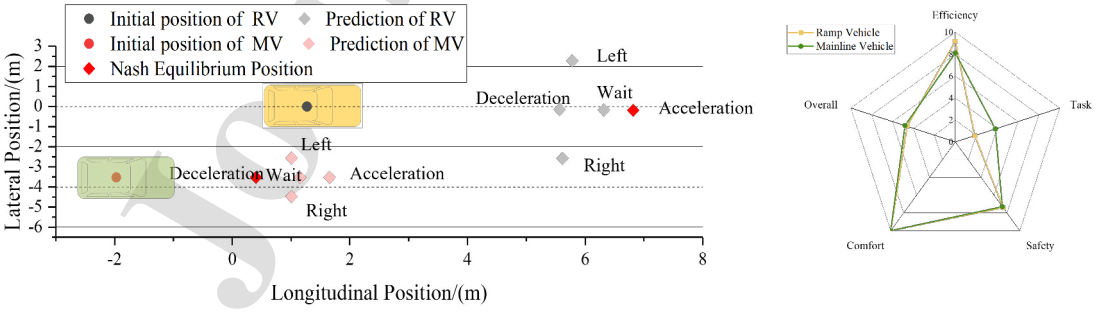


Figure 14: The Nash equilibrium scenario.

action. From Figure 14, it can be found that the comfort and safety reward of the two vehicles are almost the same, and the efficiency reward of both vehicles are almost the highest, while the task reward of RV are lower due to the situation at the beginning of the merging. Overall, the overall income of the two vehicles reached 18.4 and 19.2 respectively.

6.2. Learning performance

To illustrate the training performance of the proposed Nash double Q learning algorithm, we compared the average episode reward with the baselines. The successful single reinforcement learning algorithms and the normal multi-agent reinforcement learning are used as the baselines.

DDPG and DDPG-Priority: DDPG (Lillicrap et al., 2016) is a single agent reinforcement learning algorithm, based on the actor-critic framework. To operate in the continuous action spaces, the actor and critic is established using a neural network, and updated based on deterministic policy gradients. The competing vehicle on the mainline is assumed to keep its motion during control time horizon. While in DDPG-Priority, prioritised experience replay (Schaul et al., 2016) is introduced in the original DDPG, so as to replay importance transitions more frequently, and therefore learn more efficiently.

RDPG: This is also a single agent reinforcement learning algorithm (Heess et al., 2015), which is an extension of the DDPG. In this method, recurrent neural network is used to build the actor and critic, so that the information over many time steps could be preserved. The RDPG is trained with backpropagation through time.

MADDPG: MADDPG (Baltru, 2019) extends DDPG into a multi-agent policy gradient algorithm where ramp and mainline vehicles learn a centralised critic based on the observations and actions of all agents.

MATD3: MATD3 (Ackermann et al., 2019) change the DDPG into Twin Delayed Deep Deterministic policy gradient (TD3) in the MADDPG. An approach that reduces this bias by using double centralized critics.

Nash Double Q learning: This is our proposed algorithm in this work. Our learning algorithm also involves two agents, the ramp and mainline vehicles. In MADDPG, the agents make decisions only care about its own reward. Different from that, our algorithm uses the Q value of each agent at current states to derive the Nash equilibrium strategy. The parameters of the training are listed in the TABLE 2.

The training results are shown in Figure 15. Almost all the rewards of the algorithms increase with the training steps. Except for the DDPG, the reward gradually decreases at around 1500th step. During the training, the rewards of RDPG and DDPG fluctuate around 0.25, -7.19, respectively, in a relatively wide range. However, the reward of RDPG seems hard to converge. As for DDPG-Priority, the convergence is fastest among these five algorithms and it converges at about training step 200, since the use of the replay of prioritised driving experiences. However, the stable reward is not too high, reaching around 0.68. It is obvious that the stable rewards of the single RL algorithms are relatively low, under zero or around zeros. The purple line represents the average reward of MADDPG with training steps. The reward fluctuates at first but gradually stabilises at 6.46 from step 2000. MATD3 is another baseline algorithm as multi-agent RL, which is also the improvement of MADDPG. However, in our simulation, it seems not so satisfactory. The average rewards are always lowest, though they began to rise at about step 2500. What stands out in Figure 15 is the red line, representing our proposed learning algorithm, converges to the highest point, at 18.46, though from step 400. In addition, the fluctuation range is relatively small. Overall, the stable reward of our proposed Nash double Q is highest compared with the state-of-the-art and shows better convergence.

To further illustrate the training performance, the speed and acceleration distributions of ramp vehicles during training are presented in Figure 16 (a), (b). The second best algorithm, MADDPG, is chosen to compare with our method. The mean speed of our training is 36 km/h, while the mean speed of MADDPG is 22 km/h. The mean acceleration of our method during training is -0.2 m/s^2. The mean acceleration of is -1.8 m/s^2. It is obvious that our training results own higher speed, which means the efficiency is higher. Also, in Figure 16 (c), the average collision of our method is slightly higher than MADDPG.

To evaluate whether the Q network is able to fit the Nash Q values, the error distribution of the target Nash Q value and actual Q from Q network are shown in the Figure 17, as well as the box plot. The upper edge of each box represents the 1st quartile, and the lower edge represents the 3rd quartile. The squares in each bin represent the mean of the errors. The x-axis represents the error of the mainline vehicle and y-axis represents the ramp vehicle. As shown in the Figure 17, the error of ramp vehicles ranges from 0.4-0.6 meters, and the error range of mainline vehicles is between 0.9 and 1.0 meter. The mean errors of the ramp and mainline vehicles are 0.53 and 0.94 meters respectively. The Q network can fit the Nash Q of the ramp vehicle better. Though the error of the mainline vehicle is relatively high, it is acceptable.

6.3. Tests in real on-ramp scenarios

The previous results have proved the effectiveness of the proposed algorithm. In this section, the trained model will be applied in real on-ramp scenarios to show the performances. The real on-ramps scenarios are extracted from NGSIM, I-80 Freeway dataset (Information, 2016). The trajectory data was collected on the east-bound (northbound) direction of Interstate 80 in Emeryville, California on April 13, 2005. We used the transcribed data from 4:00 p.m. to 4:15 p.m. The data provides X, Y coordinates of each vehicle, every 1/10th of a second in relative space and

Table 2: The learning parameters of used RL algorithms.

	Nash double Q	MADDPG	MATD3	RDPG	DDPG
Batchsize	64	64	64	64	64
Memory capacity	2000	2000	2000	2000	2000
Discount factor	0.9	0.95	0.95	0.85	0.85
Hidden unit of actor	64*32	64*32	64*32	64*32	40*30
Hidden unit of critic	/	32*16	32*16	128*64	48*24
Learning rate of actor	0.005	0.005	0.005	0.0001	0.0001
learning rate of critic	/	0.005	0.005	0.001	0.001

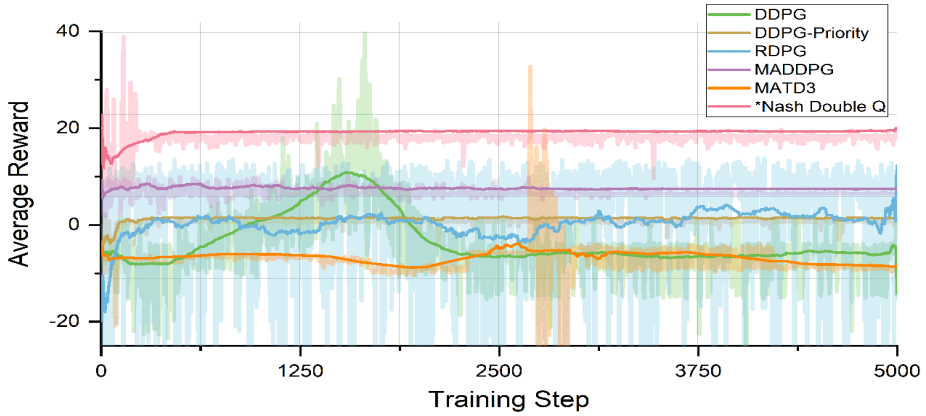


Figure 15: The average reward per episode changing with training under different RL algorithms.

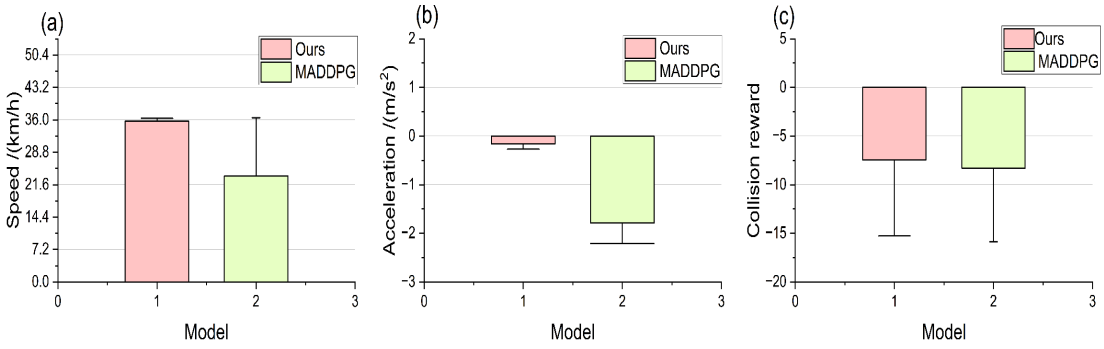


Figure 16: The training results.(a) Speed distribution. (b) Acceleration distribution. (c) Collision reward distribution.

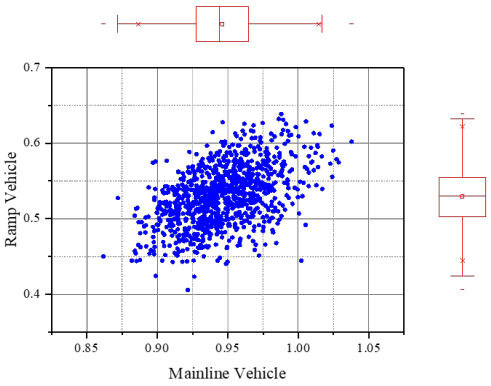


Figure 17: The Q error analysis of ramp and mainline vehicles.

Figure 18: Left picture shows a digital video camera mounted on top of a building that overlooks I-80. Right one is the aerial photograph of the extent of the I-80 study area.

in the California State Plane Coordinate System, Zone 3, NAD83 (the units are US Survey Feet in original dataset, and we transfer them into metres). Figure 18 provides a schematic illustration of the location for the vehicle trajectory dataset. The site was approximately 502 meters in length, with an on-ramp at Powell Street. In our simulation, two scenarios are selected. There is only one lane on the mainline in the first scenario, where only two vehicles are involved, including the ramp vehicle and the competing vehicle on the mainline. The second one is more complex, where there are two lanes on the mainline and four surrounding vehicles in addition to the competing vehicle.

Scenario 1: Single Lane, Single Vehicle The first is a simplified merging scenario extracted from I-80 dataset, where the initial condition is obtained, as shown in the Figure 7. Our task is to consider how to plan the motion of the yellow vehicle to merge into the mainline, when faced with the competing human-driven vehicle (green car). The initial position of the ramp vehicle is (1.27. 0), and the initial velocity is 10.1 m/s. The initial position of the green vehicle is (-1.97, -3.52), and the initial velocity is 6.25m/s. The simulations are conducted using the proposed and baseline RL algorithms. Since there is only one lane on the mainline, the simulation ends when the ramp vehicle reaches the center line. The planned trajectories are presented in the Figure 19 . The grey line is the trajectory of a real human-driven vehicle based on the dataset. And the others are the planned trajectories of RL algorithms. To clearly evaluate the driving performance of our proposed algorithm, only the trajectory points within the control time domain of Nash Double Q are provided. The point on the path represents the position of each time step. The time interval of the points is 0.1s. It is found that the trajectory of DDPG and MADDPG are very close to the boundary of the acceleration lane. Especially for the DDPG, considering the width of the vehicle body, the planned path will result in a collision with the road boundary. It can be explained according to the Figure 20, where the steering angle and acceleration are presented. In the beginning of the simulation, the DDPG steers to the right with acceleration. And then it realised the collision risk and turned left with deceleration. Whereas the collision is still hard to avoid in this situation. The learning of MADDPG teaches the ramp vehicle to avoid the collision. However, in the shown control domain, the ramp vehicle always drives along the acceleration lane. It seems that the ramp vehicle is hard to complete the merging task under the control of MADDPG algorithm. As suggested in the Figure 20, the ramp vehicle also steers to the right and then turn left with a small angle under MADDPG control. The planned trajectories of RDPG and our Nash double Q are more similar. They both have obvious lateral replacements, and they are promising to complete the merging task. However, our planned trajectory is more efficient, and the longitudinal position is about 2 meters ahead of RDPG in the end of the control domain. In this scenario, our proposed method steers to the left at first and then accelerates to arrive at the center line as soon as possible. While the RDPG is too conservative and always decelerates during the merging. The excellence of our results is attributed to the fact that our method can better understand the behaviour of the human driver and the driving task, through introduction of the game analysis. In this case, it seems that the historical information (RDPG) is more helpful than the surrounding states (MADDPG). Moreover, the averages of each reward during the control domain are given in the Figure 21. The efficiency rewards of all used algorithms are very close. Regarding the safety, comfort and the task completion, our algorithm outperforms, that is the reason why our overall reward is highest, reaching to 13.85.

Also, in order to better understand the effectiveness of our algorithm, we provide the game analysis during applications, in Figure 22. At the first step of the control, the Q value of all action pairs are given in Figure 22(a), and the heatmaps of Q value for ramp and mainline vehicles are given in Figure 22(b), (c). In Figure 22(e), the green histogram represents the strategy for ramp vehicles, while the yellow histogram is the predictive strategy of competing vehicles on mainline, by Nash equilibrium analysis.

Journal Pre-proof

Journal Pre-proof

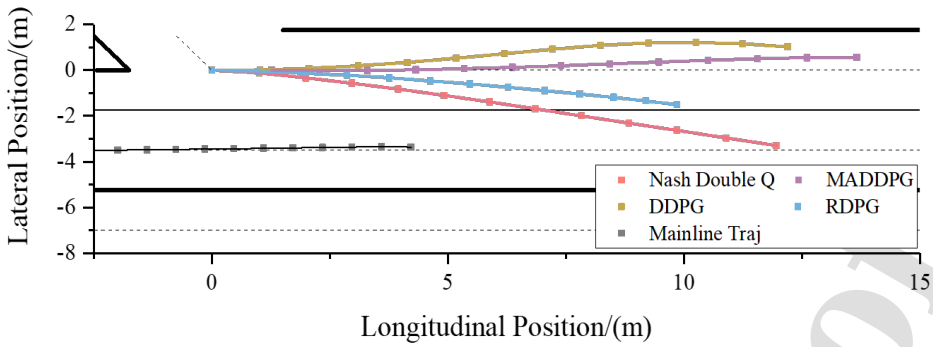


Figure 19: The planned trajectories of different algorithms in the first scenario.

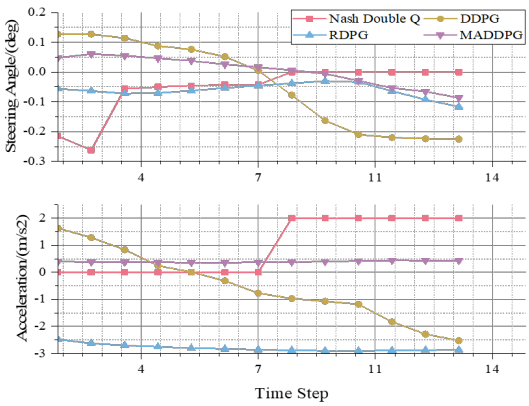


Figure 20: The planned actions in first scenario.

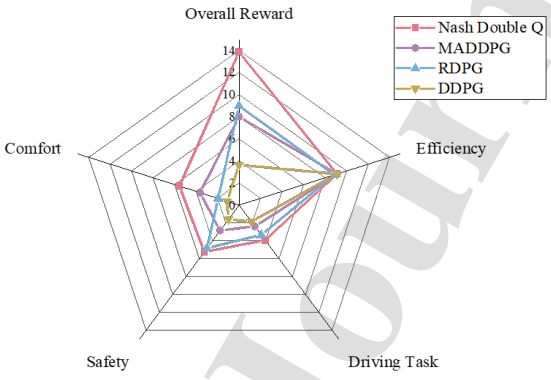


Figure 21: The reward comparisions in first scenario.

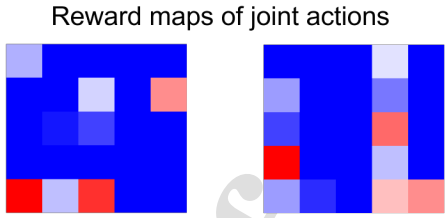
According to the strategy, the ramp vehicle should just execute left steering with 14.25 deg and maintain the speed. The predicted position and the actual traffic states are shown in the Figure 22(d). The actual states is given according to the real-world traffic datasets. The error of the prediction and actual position is 0.037 meters currently. Similarly, the situation at t = 0.5s is provided in the Figure 23. At this time, the ramp vehicle is more confident to steer to the left with 15 deg, since it predicted that the mainline is more likely to yield for the ramp vehicle to merge.

Scenario 2: Multi Lanes, Multi Vehicles The second is the real merging scenarios in the I-80 dataset, shown in the Figure 24. This scenario is a typical multi-lane merging scenario, and more vehicles are involved in the planning. When the ramp vehicle arrives on the acceleration lane, in addition to the completing vehicle, there is another front vehicle on the outermost lane, denoted as v_{1f}. On the second mainline, there are two vehicles, named v_{2f}, and v_{2r}. The initial positions and velocity of these vehicles are given in the TABLE 3.

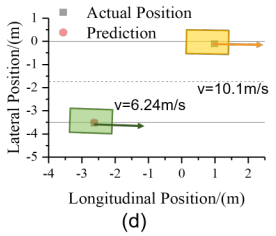
The planned trajectories of the second scenario are presented in the Figure 25 . The total merging costs 1.8 seconds in our proposed method. Unsurprisingly, DDPG could not produce the correct trajectory in this even complex situation. It is found in Figure 26, the planned steering angle of DDPG fluctuates dramatically, and it accelerates all the time. As a result, the DDPG still collides with the road boundary. The RDPG did not show a satisfying result in this scenario. The RDPG controller make the ramp vehicle decelerate all the time, to keep safe in the existence of the surrounding vehicles. However, the MADDPG shows the outperformance in this situation, as well as our proposed method. Though these two methods are able to successfully complete the merging task, the MADDPG controller steers to left initially and then switches to the right. Different from that, our Nash double Q controller accelerates firstly and then turns right in time to merge. That is why our overall reward is higher than the MADDPG, also the highest among the used algorithms,

Ramp/ mainline	Wait	Accele.	Decele.	Left	Right
Wait	(0.48, 0.15)	(0.18, 0.05)	(0.43, 0)	(0, 0.29)	(0.02, 0.35)
Accele.	(0, 0.48)	(0.02, 0)	(0, 0)	(0, 0.19)	(0, 0)
Decele.	(0, 0.07)	(0.04, 0)	(0.08, 0)	(0, 0.37)	(0, 0)
Left	(0, 0.15)	(0, 0)	(0.2, 0)	(0.02, 0.12)	(0.35, 0.02)
Right	(0.17, 0)	(0, 0)	(0,0)	(0, 0.21)	(0, 0)

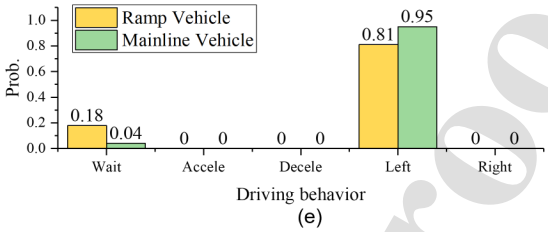
(a)



(b) Ramp vehicle (c) Mainline vehicle



(d)

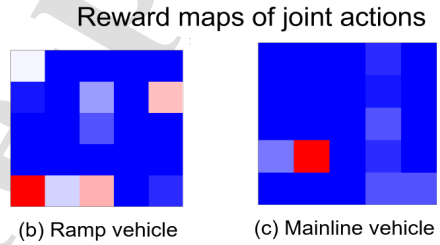


(e)

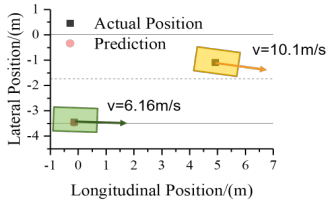
Figure 22: The game analysis of the first scenario at t=0.

Ramp/ mainline	Wait	Accele.	Decele.	Left	Right
Wait	(0.52, 0.08)	(0.22, 0.08)	(0.33, 0)	(0, 0.35)	(0.06, 0.38)
Accele.	(0, 0.51)	(0, 02)	(0, 0)	(0, 0.22)	(0, 0)
Decele.	(0, 0.10)	(0, 0)	(0.09, 0)	(0, 0.37)	(0, 0)
Left	(0.05, 0.11)	(0, 0)	(0.17, 0)	(0, 0.17)	(0.32, 0.01)
Right	(0.25, 0)	(0, 0)	(0,0)	(0, 0.21)	(0, 0)

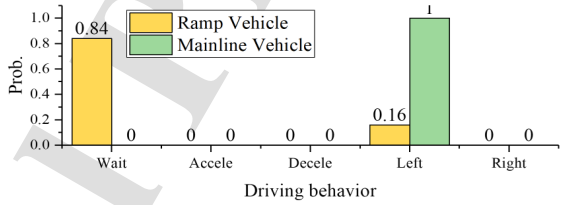
(a)



(b) Ramp vehicle (c) Mainline vehicle



(d)



(e)

Figure 23: The game analysis of the first scenario at t=0.5s.

Table 3: The initial positions and velocity of simulated vehicles

	Longitudinal position/(m)	Lateral position/(m)	Velocity/(m/s)
Ramp vehicle v0	1.27	0	10.1
Competing vehicle v1	-1.97	-3.52	6.25
Surounding vehicle v1f	43	-4.2	4.57
Surounding vehicle v2f	34	-7.6	7.7
Surounding vehicle v2r	1	-7.8	8.9

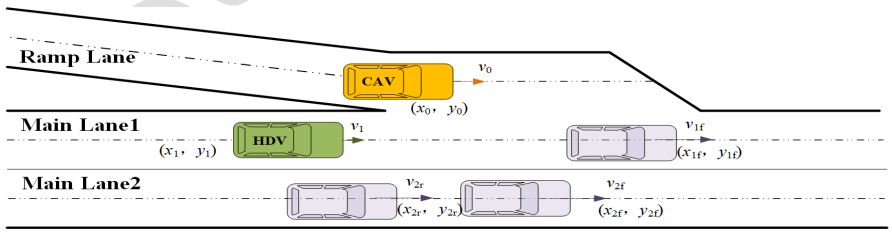


Figure 24: The second on-ramp scenarios: two lanes, multi vehicles.

Journal Pre-proof

shown in the Figure 27. In summary, our algorithm is robust when the traffic situation changes. And in dense traffic, the results of MADDPG are advantageous, since the coordination of multi-vehicle is more important.

We also provide the game analysis of the second scenarios in Figure 28, 29, 30. At t=0, the Q value of the two vehicles is given in the Figure 28. Compared with the first scenario, the Q value of each action pair is more obviously different. The competing vehicle is predicted to accelerate, and the ramp vehicle is suggested to accelerate with 2 m/s2 at this time step. From time t=1s, the ramp vehicle begins to turn left based on the prediction that the competing vehicle will wait, as seen from Figure 29. Until time t=1.8s, presented in Figure 30, the ramp vehicle nearly reaches the center line of the mainline, and begins to turn right to follow the center line.

To show the feasibility with application, the computation time during simulation is presented in Figure 31. The simulation is repeated for 50 times. The computation time for every simulation ranges from 0.005 to 0.007s. And planning horizon is usually 0.1s. Therefore, the time complexity could satisfy the requirements of application.

7. Conclusions

Aiming at the uncertain interaction of human-driven vehicles on mainline and ramp vehicles, an interactive merging strategy based on multi-agent deep reinforcement learning (MADRL) is designed, enabling the ramp vehicle (CAV) to consider the dynamic reaction of mainline vehicles. In our method, the longitudinal and lateral motion can be planned at the same time. And the mainline and ramp vehicles can be coordinated by creating two agents. One agent is to predict and analyse the behaviour of mainline vehicles (human-driven vehicles, HDV, or non-connected vehicles), and another agent is created for exploring optimal merging actions of ramp vehicles. To balance the benefits of both vehicles, game theory is used to model the competitive behaviours between ramp and mainline vehicles. And then the Nash equilibrium of joint actions guides the ramp vehicle to learn best response to the mainline vehicle. Also, the Nash double Q algorithm is developed in training process, to ensure the outputs of Q networks efficiently converge to the Nash equilibrium point. And then the trained networks are used for online control.

In our simulations, firstly, the learning performance of single RL (DDPG, RDPG) and existing MADRL (MADDPG) algorithms are compared. It shows that our algorithm could converge to the highest average rewards stably, at around 18.46. Moreover, the different RLs are simulated in real on-ramp scenarios. In the simple scenario, our method shows superior performance in terms of safety, efficiency and comfort. It may be due to a better understanding of the behaviour of the human driver and the driving task, based on the game analysis. In this case, it seems that the historical information (RDPG) is more helpful than the surrounding states (MADDPG). In a denser scenario, our results are still the best in all kinds of driving indicators, including safety, efficiency, and comfort. In summary, our algorithm is robust when the traffic situation changes. And in dense traffic, the results of MADDPG are advantageous, since the coordination of multi-vehicle is more important.

In this study, uncertain interactions of human drivers are focused and the basic game equilibrium is used. Currently, only one ramp vehicle is considered in our model. In future, our work is expected to extend to more agents. To further understand the competition and cooperation among vehicles, more game theory should be studied in our future works. However, it is not difficult to extend to other game theories based on our proposed model.

References

Abdel-Aty, M., Dhindsa, A., Gayah, V., 2007. Considering various ALINEA ramp metering strategies for crash risk mitigation on freeways under congested regime. Transportation Research Part C: Emerging Technologies 15, 113–134. doi:10.1016/j.trc.2007.02.003.

Ackermann, J., Gabler, V., Osa, T., Sugiyama, M., 2019. Reducing overestimation bias in multi-agent domains using double centralized critics. arXiv preprint arXiv:1910.01465 .

Al-Turki, M., Ratrout, N.T., Rahman, S.M., Assi, K.J., 2022. Signalized Intersection Control in Mixed Autonomous and Regular Vehicles Traffic Environment - A Critical Review Focusing on Future Control. IEEE Access 10, 16942–16951. doi:10.1109/ACCESS.2022.3148706.

Ali, Y., Zheng, Z., Haque, M.M., Yildirimoglu, M., Washington, S., 2021. Clacd: A complete lane-changing decision modeling framework for the connected and traditional environments. Transportation Research Part C: Emerging Technologies 128, 103162. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001807, doi:https://doi.org/10.1016/j.trc.2021.103162.

An, L., Yang, X., Hu, J., 2022. Modeling System Dynamics of Mixed Traffic With Partial Connected and Automated Vehicles. IEEE Transactions on Intelligent Transportation Systems 23, 15755–15764. doi:10.1109/TITS.2022.3145395.

Baltru, T., 2019. Multimodal Machine Learning : A Survey and Taxonomy 41, 423–443.

Bautista-Montesano, R., Galluzzi, R., Ruan, K., Fu, Y., Di, X., 2022. Autonomous navigation at unsignalized intersections: A coupled reinforcement learning and model predictive control approach. Transportation Research Part C: Emerging Technologies 139, 103662. URL: https://doi.org/10.1016/j.trc.2022.103662, doi:10.1016/j.trc.2022.103662.

Bellman, R., 1977. Markovian decision processes. doi:10.1016/S0076-5392(08)61190-X.

Chen, L., Englund, C., 2016. Cooperative intersection management: A survey. IEEE Transactions on Intelligent Transportation Systems 17, 570–586. doi:10.1109/TITS.2015.2471812.

Chen, N., Van Arem, B., Wang, M., 2022. Hierarchical Optimal Maneuver Planning and Trajectory Control at On-Ramps With Multiple Mainstream Lanes. IEEE Transactions on Intelligent Transportation Systems 23, 18889–18902. doi:10.1109/TITS.2022.3167727.

Dahl, J., De Campos, G.R., Olsson, C., Fredriksson, J., 2019. Collision Avoidance: A Literature Review on Threat-Assessment Techniques. IEEE Transactions on Intelligent Vehicles 4, 101–113. doi:10.1109/TIV.2018.2886682.

Deng, F., Jin, J., Shen, Y., Du, Y., 2019. Advanced Self-Improving Ramp Metering Algorithm based on Multi-Agent Deep Reinforcement Learning. 2019 IEEE Intelli-

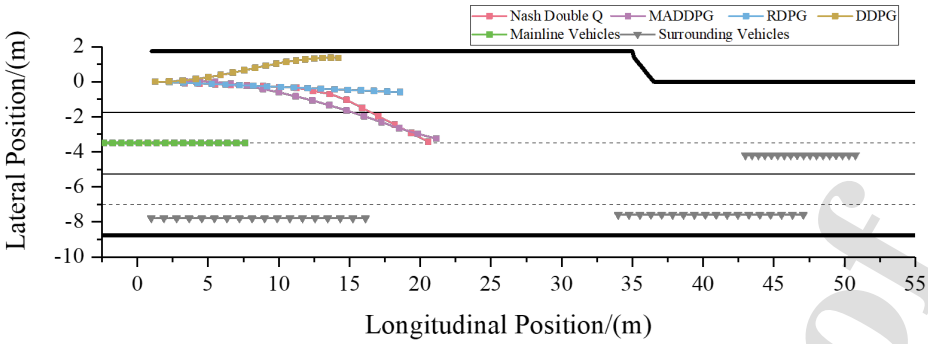


Figure 25: The planned trajectories of different algorithms in the second scenario.

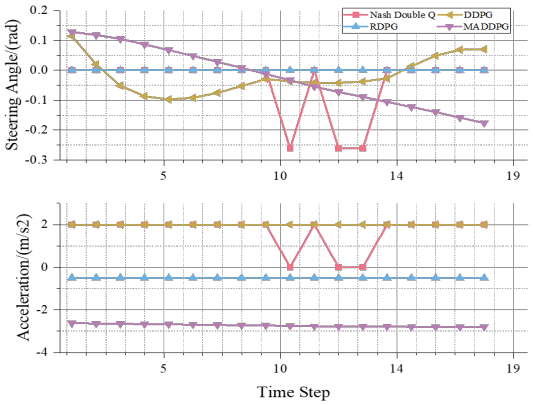


Figure 26: The planned actions of different algorithms in the second scenario..

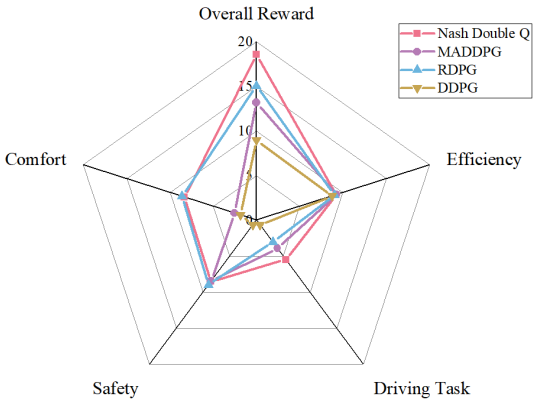


Figure 27: The reward comparisions in second scenario.

gent Transportation Systems Conference, ITSC 2019 , 3804–3809doi:10.1109/ITSC.2019.8916903.

Fan, X., Wang, Q., Liu, J., 2020. Work-in-progress abstract: A reliable wireless smart vehicle highway on-ramp merging protocol with constant time headway safety guarantee. Proceedings - 2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems, ICCPS 2020 , 184–185doi:10.1109/ICCPS48487.2020.00025.

Fukuyama, S., 2020. Dynamic game-based approach for optimizing merging vehicle trajectories using time-expanded decision diagram. Transportation Research Part C: Emerging Technologies 120, 102766. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2030677X, doi:https://doi.org/10.1016/j.trc.2020.102766.

Heess, N., Hunt, J.J., Lillicrap, T.P., Silver, D., 2015. Memory-based control with recurrent neural networks , 1–11URL: http://arxiv.org/abs/1512.04455, arXiv:1512.04455.

Heppeler, G., Sonntag, M., Wohlhaupter, U., Sawodny, O., 2017. Predictive planning of optimal velocity and state of charge trajectories for hybrid electric vehicles. Control Engineering Practice 61, 229–243. URL: http://dx.doi.org/10.1016/j.conengprac.2016.07.003, doi:10.1016/j.conengprac.2016.07.003.

Herings, P.J.J., Peeters, R., 2009. Homotopy methods to compute equilibria in game theory. Economic Theory 42, 119–156. doi:10.1007/s00199-009-0441-5.

Heshami, S., Kattan, L., 2021. Ramp metering control under stochastic capacity in a connected environment: A dynamic bargaining game theory approach. Transportation Research Part C: Emerging Technologies 130, 103282. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002941, doi:https://doi.org/10.1016/j.trc.2021.103282.

Information, I., 2016. Next Generation Simulation (NGSIM) Interstate 80 Freeway Dataset Identification Information Time Period of Content .

Jeon, S., Lee, K., Kum, D., 2022. Overtaking decision and trajectory planning in highway via hierarchical architecture of conditional state machine and chance constrained model predictive control. Robotics and Autonomous Systems 151, 104014. URL: https://doi.org/10.1016/j.robot.2021.104014, doi:10.1016/j.robot.2021.104014.

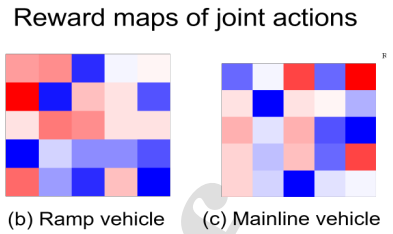
Jing, S., Hui, F., Zhao, X., Rios-Torres, J., Khattak, A.J., 2022. Integrated Longitudinal and Lateral Hierarchical Control of Cooperative Merging of Connected and Automated Vehicles at On-Ramps. IEEE Transactions on Intelligent Transportation Systems 23, 24248–24262. doi:10.1109/TITS.2022.3204033.

Jing, S., Zhao, X., Hui, F., Khattak, A.J., Yang, L., 2021. Cooperative CAVs optimal trajectory planning for collision avoidance and merging in the weaving section. Transportmetrica B 9, 219–236. URL: https://doi.org/10.1080/21680566.2020.1845852, doi:10.1080/21680566.2020.1845852.

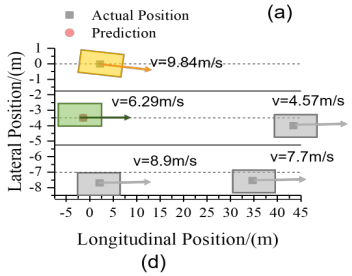
Khattak, Z.H., Smith, B.L., Fontaine, M.D., Ma, J., Khat-

Ramp/ mainline	Wait	Decele.	Accele.	Left	Right
Wait	(3.01, 0.98)	(-0.56, -0.33)	(-1.83, -3.33)	(1.75, -0.04)	(-2.78, 0.42)
Decele.	(-2.44, 1.04)	(0.29, -0.63)	(-0.67, 1.40)	(-0.81, -1.75)	(-1.46, 3.11)
Accele.	(1.24, 1.63)	(2.76, 0.15)	(**2.52**, **1.72**)	(1.21, -2.06)	(1.35, -3.07)
Left	(4.58, 0.92)	(-2.25, -3.57)	(1.71, 0.9)	(1.14, 0.62)	(-1.34, -0.68)
Right	(2.24, -1.96)	(2.54, 0.24)	(-1.82,3.13)	(0.72, -1.89)	(0.91, 4.41)

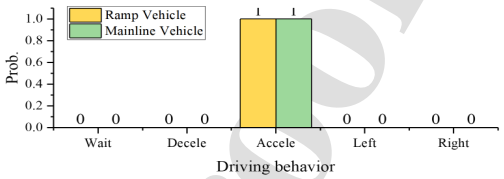
(a)



(b) Ramp vehicle (c) Mainline vehicle



(d)

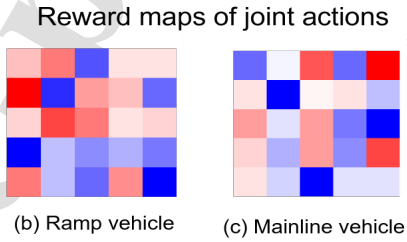


(e)

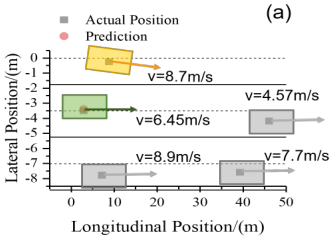
Figure 28: The game analysis of the second scenario at t=0.

Ramp/ mainline	Wait	Decele.	Accele.	Left	Right
Wait	(1.71, 0.65)	(-0.22, -0.10)	(-1.37, -4.15)	(2.24, -0.01)	(-3.49, -0.63)
Decele.	(-3.58, 0.80)	(-0.67, -1.13)	(-1.55, 1.72)	(-0.54, -1.09)	(-1.24, 3.13)
Accele.	(0.96, 1.41)	(3.20, -0)	(2.42, 1.41)	(1.21, -1.56)	(1.21, -3.59)
Left	(**4.18**, **0.45**)	(-2.20, -3.53)	(1.76, -0.08)	(1.45, 0.75)	(-1.64, -0.66)
Right	(0.96, -1.98)	(1.93, 0.27)	(-2.02,2.68)	(0.86, -1.73)	(0.68, 4.33)

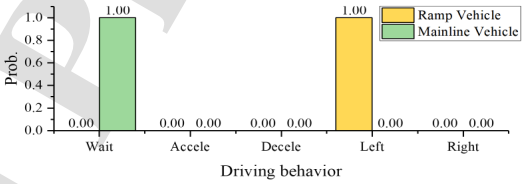
(a)



(b) Ramp vehicle (c) Mainline vehicle



(d)

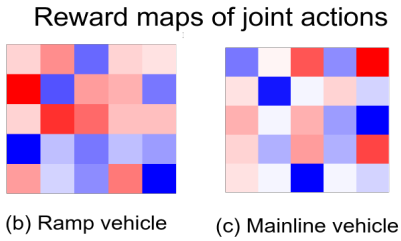


(e)

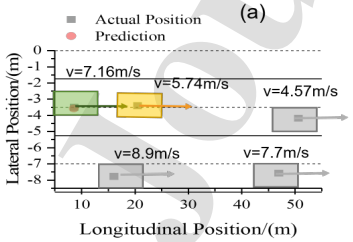
Figure 29: The game analysis of the second scenario at t=1s.

Ramp/ mainline	Wait	Decele.	Accele.	Left	Right
Wait	(2.31, 0.83)	(-0.47, -0.3)	(-1.66, -3.68)	(2.07, -0.12)	(-3.27, -0.16)
Decele.	(-3.13, 0.99)	(-0.43, -0.81)	(-1.08, 1.66)	(-0.68, -1.35)	(-1.32, 3.11)
Accele.	(1.07, 1.67)	(3.18, -0.17)	(2.44, 1.66)	(0.95, -1.78)	(1.13, -3.39)
Left	(4.27, 0.74)	(-2.27, -3.65)	(1.79, 0.35)	(1.40, 0.62)	(-1.65, -0.69)
Right	(1.46, -1.98)	(2.28, 0.16)	(-1.79,2.82)	(0.93, -1.92)	(0.95, 4.33)

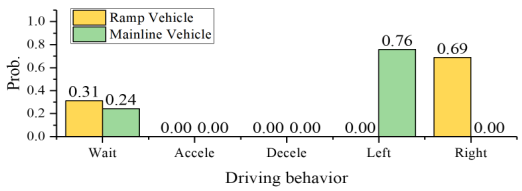
(a)



(b) Ramp vehicle (c) Mainline vehicle



(d)



(e)

Figure 30: The game analysis of the second scenario at t=1.8s.

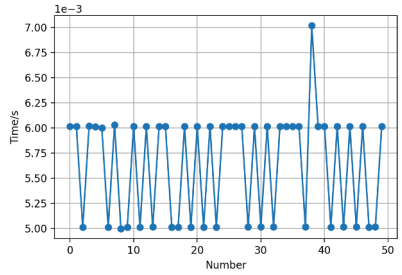


Figure 31: The computation time of our method.

tak, A.J., 2022. Active lane management and control using connected and automated vehicles in a mixed traffic environment. Transportation Research Part C 139, 103648. URL: https://doi.org/10.1016/j.trc.2022.103648, doi:10.1016/j.trc.2022.103648.

Koller, D., Pfeffer, A., 1997. Representations and solutions for game-theoretic problems. Artificial Intelligence 94, 167–215. doi:10.1016/s0004-3702(97)00023-4.

Kotsialos, A., 2021. A varying parameter multi-class second-order macroscopic traffic flow model for coordinated ramp metering with global and local environmental objectives. Transportation Research Part C: Emerging Technologies 128, 103106. URL: https://doi.org/10.1016/j.trc.2021.103106, doi:10.1016/j.trc.2021.103106.

LEMKE, C.E., HOWSON, J.T., JR., 1964. EQUILIBRIUM POINTS OF BIMATRIX GAMES .

Li, P., Pei, X., Chen, Z., Zhou, X., Xu, J., 2022. Human-like motion planning of autonomous vehicle based on probabilistic trajectory prediction. Applied Soft Computing 118, 108499. URL: https://doi.org/10.1016/j.asoc.2022.108499, doi:10.1016/j.asoc.2022.108499.

Li, Z., Yu, H., Zhang, G., Dong, S., Xu, C.Z., 2021. Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. Transportation Research Part C: Emerging Technologies 125, 103059. URL: https://doi.org/10.1016/j.trc.2021.103059, doi:10.1016/j.trc.2021.103059, arXiv:2104.09936.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2016. Continuous control with deep reinforcement learning. 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings arXiv:1509.02971.

Liu, B., Shi, Q., Song, Z., El Kamel, A., 2019. Trajectory planning for autonomous intersection management of connected vehicles. Simulation Modelling Practice and Theory 90, 16–30. URL: https://doi.org/10.1016/j.simpat.2018.10.002, doi:10.1016/j.simpat.2018.10.002.

Madridano, Á., Al-Kaff, A., Martín, D., de la Escalera, A., 2021. Trajectory planning for multi-robot systems: Methods and applications. Expert Systems with Applications 173. doi:10.1016/j.eswa.2021.114660.

Nash, J.F., 1950. Equilibrium points in n person games. Proceedings of the National Academy of Sciences 36, 48–49. doi:10.1073/pnas.36.1.48.

Ntousakis, I.A., Nikolos, I.K., Papageorgiou, M., 2016. Optimal vehicle trajectory planning in the context of cooperative merging on highways. Transportation Research Part C: Emerging Technologies 71, 464–488. URL: http://dx.doi.org/10.1016/j.trc.2016.08.007, doi:10.1016/j.trc.2016.08.007.

Peng, T., Su, L., Zhang, R., Guan, Z., Zhao, H., Qiu, Z., Zong, C., Xu, H., 2020. A new safe lane-change trajectory model and collision avoidance control method for automatic driving vehicles. Expert Systems with Applications 141. doi:10.1016/j.eswa.2019.112953.

Rios-Torres, J., Malikopoulos, A.A., 2017. A Survey on the Coordination of Connected and Automated Vehicles at Intersections and Merging at Highway On-Ramps. IEEE Transactions on Intelligent Transportation Systems 18, 1066–1077. doi:10.1109/TITS.2016.2600504.

Schaul, T., Quan, J., Antonoglou, I., Silver, D., 2016. Prioritized experience replay. 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings , 1–21arXiv:1511.05952.

Schester, L., Ortiz, L.E., 2021. Automated driving highway traffic merging using deep multi-agent reinforcement learning in continuous state-action spaces. IEEE Intelligent Vehicles Symposium, Proceedings 2021-July, 280–287. doi:10.1109/IV48863.2021.9575676.

Scholte, W.J., Zegelaar, P.W., Nijmeijer, H., 2022. A control strategy for merging a single vehicle into a platoon at highway on-ramps. Transportation Research Part C: Emerging Technologies 136, 103511. URL: https://doi.org/10.1016/j.trc.2021.103511, doi:10.1016/j.trc.2021.103511.

Shou, Z., Chen, X., Fu, Y., Di, X., 2022. Multi-agent reinforcement learning for Markov routing games: A new modeling paradigm for dynamic traffic assignment. Transportation Research Part C: Emerging Technologies 137, 103560. URL: https://doi.org/10.1016/j.trc.2022.103560, doi:10.1016/j.trc.2022.103560, arXiv:2011.10915.

Su, H., Zhong, Y.D., Chow, J.Y., Dey, B., Jin, L., 2022. Emvlight: A Multi-Agent Reinforcement Learning Framework for an Emergency Vehicle Decentralized Routing and Traffic Signal Control System. SSRN Electronic Journal 146, 1–25. doi:10.2139/ssrn.4152865, arXiv:2206.13441.

Talebpour, A., Mahmassani, H.S., Hamdar, S.H., 2015. Modeling lane-changing behavior in a connected environment: A game theory approach. Transportation Research Procedia 7, 420–440. URL: https://www.sciencedirect.com/science/article/pii/S2352146515000903, doi:https://doi.org/10.1016/j.trpro.2015.06.022. 21st International Symposium on Transportation and Traffic Theory Kobe, Japan, 5-7 August, 2015.

Uno, A., Sakaguchi, T., Tsugawa, S., 1999. Merging control algorithm based on inter-vehicle communication. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC , 783–787doi:10.1109/itsc.1999.821160.

Wang, J., Zheng, Y., Xu, Q., Wang, J., Li, K., 2021. Controllability Analysis and Optimal Control of Mixed Traffic Flow with Human-Driven and Autonomous Vehicles. IEEE Transactions on Intelligent Transportation Systems 22, 7445–7459. doi:10.1109/TITS.2020.3002965, arXiv:2002.02099.

Wang, Y., Wenjuan, E., Tang, W., Tian, D., Lu, G., Yu, G., 2013. Automated on-ramp merging control algorithm based on internet-connected vehicles. IET Intelligent Transport Systems 7, 371–379. doi:10.1049/iet-its.2011.0228.

Wu, J., Chen, X., Bie, Y., Zhou, W., 2023. A co-evolutionary lane-changing trajectory planning method for automated vehicles based on the instantaneous risk identification. Accident; analysis and prevention 180, 106907. URL: https://doi.org/10.1016/j.aap.2022.106907, doi:10.1016/j.aap.2022.106907.

Xiao, W., Cassandras, C.G., Belta, C.A., 2021. Bridging the gap between optimal trajectory planning and safety-critical control with applications to autonomous vehicles. Automatica 129. doi:10.1016/j.automatica.2021.109592, arXiv:2008.07632.

Yao, Z., Jiang, H., Cheng, Y., Jiang, Y., Ran, B., 2020. Integrated Schedule and Trajectory Optimization for Connected Automated Vehicles in a Conflict Zone. IEEE Transactions on Intelligent Transportation Systems 23, 1841–1851. doi:10.1109/tits.2020.3027731.

Yao, Z., Wu, Y., Wang, Y., Zhao, B., Jiang, Y., 2023. Analysis of the impact of maximum platoon size of CAVs on mixed traffic flow : An analytical and simulation method. Transportation Research Part C 147, 103989. URL: https://doi.org/10.1016/j.trc.2022.103989,

Journal Pre-proof

doi:10.1016/j.trc.2022.103989.

You, F., Zhang, R., Lie, G., Wang, H., Wen, H., Xu, J., 2015. Trajectory planning and tracking control for autonomous lane change maneuver based on the cooperative vehicle infrastructure system. Expert Systems with Applications 42, 5932–5946. doi:10.1016/j.eswa.2015.03.022.

Zhang, R., Mao, S., Kang, Y., 2023. A novel traffic flow prediction model: Variable order fractional grey model based on an improved grey evolution algorithm. Expert Systems with Applications 224. doi:10.1016/j.eswa.2023.119943.

Zhang, X., Zhao, C., Liao, F., Li, X., Du, Y., 2022. Online parking assignment in an environment of partially connected vehicles: A multi-agent deep reinforcement learning approach. Transportation Research Part C: Emerging Technologies 138, 103624. URL: https://doi.org/10.1016/j.trc.2022.103624, doi:10.1016/j.trc.2022.103624.

Zhongcheng, L., Dihua, S., Min, Z., Hang, Z., 2022. CPS-based ? ∞ pinning approach for mixed traffic considering human driver differences. Transportation Research Part C 140, 103674. URL: https://doi.org/10.1016/j.trc.2022.103674, doi:10.1016/j.trc.2022.103674.

Zhou, S., Zhuang, W., Yin, G., Liu, H., Qiu, C., 2022. Cooperative On-Ramp Merging Control of Connected and Automated Vehicles: Distributed Multi-Agent Deep Reinforcement Learning Approach. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2022-Octob, 402–408. doi:10.1109/ITSC55140.2022.9922173.

Zhou, Y., Cholette, M.E., Bhaskar, A., Chung, E., 2019. Optimal Vehicle Trajectory Planning with Control Constraints and Recursive Implementation for Automated On-Ramp Merging. IEEE Transactions on Intelligent Transportation Systems 20, 3409–3420. doi:10.1109/TITS.2018.2874234.

Zhu, L., Lu, L., Wang, X., Jiang, C., Ye, N., 2022. Operational Characteristics of Mixed-Autonomy Traffic Flow on the Freeway With On-and Off-Ramps and Weaving Sections: An RL-Based Approach. IEEE Transactions on Intelligent Transportation Systems 23, 13512–13525. doi:10.1109/TITS.2021.3125241.

Journal Pre-proof

Journal Pre-proof

Contributor Roles Taxonomy

Lin Li: Conceptualization, Writing - Original Draft, review&editing, Methodology, Software, Validation

Wanzhong Zhao: corresponding author, Supervision, Funding acquisition

Chunyan Wang: Supervision

Abbas Fotouhi: Supervision

Xuze Liu: Writing - Review

Journal Pre-proof

Journal Pre-proof

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof

https://dspace.org/

Nash double Q-based multi-agent deep reinforcement learning for interactive merging strategy in mixed traffic

Li, Lin; Zhao, Wanzhong; Wang, Chunyan; Fotouhi, Abbas; Liu, Xuze

2023-03-01
Elsevier
https://doi.org/10.1016/j.eswa.2023.121458
Li L, Zhao W, Wang C, et al., (2024) Nash double Q-based multi-agent deep reinforcement learning for interactive merging strategy in mixed traffic. Expert Systems with Applications, Volume 237, Part B, March 2024, Article number 121458

Downloaded from a DSpace repository. https://dspace.org/