Multi-Agent Deep Reinforcement Learning for Longitudinal Control of CAVs in Mixed Traffic Environments with Unreliable Communication

Shuncheng Cai[1], Melanie Bouroche[1]

[1]School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

Abstract

Longitudinal control is a crucial aspect of autonomous vehicle operation, which aims to ensure safe spacing between vehicles by controlling their speed. This task is particularly challenging due to the stringent timeless requirements. Connected and Autonomous Vehicles (CAVs) share information about their position and speed to improve control decisions, but inter-vehicle communication cannot be assumed to be reliable in realistic conditions. Multi-agent Deep Reinforcement Learning (MADRL) has been proposed to address this issue and ensure the traffic flow's safety and efficiency. However, existing MADRL approaches face significant challenges in maintaining system stability and adaptability especially under unreliable communication conditions in mixed traffic environments. This paper proposes a longitudinal control strategy for CAVs in mixed traffic environments based on MADRL, aimed at stabilizing platoon control under unreliable communication conditions. Specifically, vehicle-to-vehicle (V2V) communication with varying levels of packet loss is incorporated into the DRL training environment to replicate communication scenarios that may be encountered in the real world. A distribution-based mapping mechanism is designed to smooth the action selection space of the agents. The proposed algorithm is compared with baseline models through simulation experiments, and its control performance and adaptability are further validated under different packet loss levels.

Keywords

Connected and autonomous vehicle (CAV), Longitudinal control, Unreliable communication, Multi-Agent deep reinforcement learning (MADRL)

1. Introduction

In recent years, considerable exploration has been undertaken regarding longitudinal control strategies for Connected and Autonomous Vehicles (CAVs), primarily focusing on Model Predictive Control (MPC)-based strategies and Deep Learning (DL) approaches, particularly Deep Reinforcement Learning (DRL). Both methods have unique advantages under varying communication reliability conditions, and they also encounter specific challenges.

MPC-based strategies excel in optimizing multiple objectives within a flexible framework by predicting future vehicle states and dynamically adjusting driving behavior [1, 2]. As demonstrated by Liu et al., MPC-based strategies effectively automate path planning in structured driving environments, highlighting the precision of these methods in real-world applications [3]. However, MPC typically necessitates the optimization problem to be well-defined and solvable within a short time frame, which can impose significant computational demands depending on the complexity of the formulation, making real-time implementation challenging [4]. Additionally, MPC's performance relies heavily on the system model's accuracy, and any discrepancies can significantly degrade control effectiveness. Finally, computational delays can adversely affect control performance in fast-changing dynamic environments, potentially leading to instability [5, 6].

DRL-based strategies have emerged as a robust alternative, particularly advantageous in environments characterized by stochastic factors and partial observability. These strategies shift the computational burden to the offline phase, allowing for rapid online implementation as turnkey solutions[7, 8, 9, 10].

ATT'24: Workshop Agents in Traffic and Transportation, October 19, Santiago de Compostela, Spain
cais@tcd.ie (S. Cai); melanie.bouroche@tcd.ie (M. Bouroche)
https://www.scss.tcd.ie/Melanie.Bouroche/ (M. Bouroche)
0000-0003-3341-9599 (S. Cai); 0000-0002-5039-0815 (M. Bouroche)

The flexibility of DRL enables it to adapt to dynamic environments without the need for deterministic system dynamics, thus enhancing the controller's responsiveness to real-time changes. Kiran et al. survey various implementations of DRL in autonomous driving, confirming its effectiveness in learning complex policies for dynamic and unpredictable environments[11]. He et al. demonstrate a novel application of DRL for managing multiple vehicle subsystems, showcasing the method's capacity to adapt seamlessly to real-time network conditions without requiring deterministic inputs [12].

However, most existing works focus on controlling and studying CAV platoons and do not address mixed environments. Additionally, in mixed traffic environments, the mechanisms designed to smooth the action selection space of agents are frequently inadequate, leading to abrupt or suboptimal control actions. Finally, the communication between CAVs is typically assumed to be perfect, which is unrealistic.

In mixed traffic, longitudinal control presents unique challenges regarding the interaction between CAVs and different types of vehicles, such as human-driven vehicles (HDVs). Two major challenges are the heterogeneity of vehicle platoons and the random and uncertain behavior of HDVs [13, 14]. An adaptive leader-following control method for heterogeneous platoons, proposed by Li and Sun, performs well under ideal conditions but lacks responsiveness and adaptability in complex, dynamic environments [15]. A robust adaptive cruise control method for heterogeneous platoons with uncertain dynamics, developed by Wang and Zhang, can result in abrupt and discontinuous control actions when handling sudden events and dynamic changes [16]. Although these studies have made progress in addressing HDV uncertainties and mixed traffic heterogeneity, they still fall short in mitigating abrupt or suboptimal control actions in decision-making. This inadequacy can cause disruptions to surrounding HDVs, increasing the risk of traffic accidents.

Packet loss poses a major obstacle in realistic V2V communication simulations. It arises when transmitted packets fail to reach their intended destination due to various factors, including signal degradation, network congestion, or errors in data transmission. Packet loss in a CAV environment can result in outdated or incorrect information regarding vehicle states, which may have an impact on the overall traffic flow and safety [17, 18]. While DRL has demonstrated significant potential in optimizing multi-agent systems, it has scarcely been applied to packet loss scenarios. Shi et al. investigated unreliable communication by incorporating Signal-to-Interference-plus-Noise Ratio(SINR) into agents. However, this study did not specifically address packet loss scenarios, focusing instead on signal quality only [19].

Our goal is to create an integrated and robust DRL distributed control technique that focuses on stabilizing longitudinal control in mixed traffic scenarios under unreliable communication conditions. To achieve this purpose, our DRL framework introduces two novel features:

- Optimizing the Multi-Agent Proximal Policy Optimization (MAPPO) algorithm to enhance the coordination and efficiency of CAVs in mixed environments and under imperfect communication conditions. Specifically, we designed a reward function by weighted aggregation of CAV information and factors affecting CAV driving efficiency and safety, thereby balancing the impact of individual vehicle states in mixed environments. This method not only optimizes the decision-making process of individual vehicles but also significantly enhances the robustness of the overall control strategy, particularly in addressing packet loss.
- Developing a distribution-based mapping mechanism to smooth the action selection space of the agents. This mechanism mitigates abrupt or suboptimal control actions by ensuring more continuous and adaptive decision-making processes. Integrating this mechanism aims to improve the coordination and responsiveness of the CAVs under dynamic traffic conditions and communication disruptions.

The remainder of the paper is structured as follows. A review of the state of the art based on longitudinal control modules is presented in Section 2. The CAV longitudinal control framework and the proposed MADRL algorithm are described in Section 3. The numerical experiments validating the proposed CAV longitudinal control strategy are presented in Section 4. Finally, Section 5 gives the conclusion of our work.

2. State of the Art

Progress in reinforcement learning has resulted in the creation of Proximal Policy Optimisation (PPO), which has made substantial enhancements over previous methods by employing a policy gradient approach to handle continuous domains better [19]. While PPO has shown promise in improving the performance of individual vehicles, its efficacy diminishes in situations involving multiple agents or in mixed traffic environments that consist of both CAVs and HDVs [20]. One of the challenges is the tendency of PPO to converge to suboptimal policies in the presence of multiple interacting agents, leading to non-stationarity [11].

Expanding upon the capabilities of PPO, the introduction of MAPPO aimed to tackle the intricacies of environments that involve multiple interacting agents by incorporating an understanding of agent interdependencies. This understanding is crucial for effective platoon control in CAV systems [21]. Despite the progress made, MAPPO and other RL strategies still face challenges in dealing with communication inconsistencies and non-stationarities commonly encountered in real-world deployments.Zhang et al. investigated the use of MAPPO for coordinated longitudinal control of CAVs in platooning scenarios, demonstrating its effectiveness in maintaining stable inter-vehicle distances and improving traffic flow efficiency [22]. Another study by Chen et al. applied MAPPO to enhance the decision-making process of CAVs during highway merging, showing that the approach can significantly reduce merging times and improve overall traffic safety [21]. These studies highlight the advancements in applying MAPPO to longitudinal control in CAVs, contributing to improved traffic management and safety in autonomous driving scenarios. However, further research is necessary to enhance the robustness and adaptability of these strategies in diverse and dynamic real-world environments.

The research conducted by Shi et al. introduces a methodology that incorporates realistic data and signal-interference-plus noise ratio (SINR) into a DRL framework under mixed traffic environments[19]. However, further investigation into the effects of packet loss and the use of MADRL techniques would be beneficial. These additions would enhance adaptability and robustness in decentralized environments where reliable communication cannot be assumed.

3. Methodology

This paper presents a control framework that incorporates a CAV control strategy using MAPPO and a distribution-based mapping action selection mechanism to enhance CAV driving behavior and mitigate traffic oscillations in real-world scenarios. Specifically, the advantages of the distribution-based mapping action selection mechanism in continuous control tasks have been substantiated in recent RL literature[23, 24]. This framework is designed to adapt to the ever-changing conditions of the communication environment, allowing CAVs to effectively maintain stable and efficient driving behaviors even in the face of communication uncertainties.

The section begins with an overview of the environment settings, before detailing the CAV longitudinal control framework and the mapping-based action selection mechanism.

3.1. Environment Model

The simulation is centered around a car-following scenario on an infinitely long straight highway without lateral movement. The vehicle platoon consists of CAVs controlled by MAPPO agents and HDVs that follow the Krauss model, which is a widely used microscopic traffic model that simulates the longitudinal driving behavior of human drivers, including acceleration, deceleration, and maintaining safe distances. The model is based on the following assumptions:

- Every CAV can transmit the latest state information, including location and speed, to other CAVs in the platoon through V2V communication. However, due to packet loss, this information may not always be successfully received.

- The communication time is considered to have a negligible impact on the simulation process in our study.

To illustrate how the proposed control framework handles communication interruptions, consider that each CAV in the simulation maintains a state vector representing the latest valid data received from upstream vehicles. For instance, if the state vector for vehicle i is represented as $\mathbf{v}_i = [v_{i,1}, v_{i,2}, \ldots, v_{i,N}]$, where each $v_{i,j}$ is the state information from vehicle $i-j$, the vector updates based on the last successfully received data before a packet loss occurs.

In this scenario, the control strategy adjusts the vehicle's actions based on the most recent and reliable data. For example, suppose vehicle i experiences packet loss, and the last known states before the loss were $speed = 30$ m/s and $position = 100$ m. The MAPPO algorithm will use these last valid values to calculate the next best action until updated, reliable data is available. This approach simulates real-world V2V dynamics, ensuring control decisions are based on accurate information, enhancing CAV platoon safety and efficiency.

3.2. Longitudinal Control Scheme

Based on the environment setting, this part describes the control scheme of the proposed strategy, focusing on the regulation of CAVs' longitudinal control using the MAPPO-based approach. We begin with outlining the design of the Deep DRL framework, which includes the definitions of the four fundamental elements of DRL: state, action, policy, and reward. Following this, we detail the implementation of the MAPPO algorithm, focusing on centralized training and decentralized execution. Finally, the training process for the MAPPO model is described. The related notations are defined in Table 1.

Table 1
Notations of the Longitudinal Control Scheme

Symbol	Definition
n	Total number of CAVs and therefore of agents.
a_i^t	Acceleration or deceleration action of CAV i at time t.
v_i^t	Velocity of CAV i at t.
x_i^t	Position of CAV i at t.
d_i^t	Distance between CAV i and the preceding vehicle at time t.
μ_i^t, σ_i^t	NN-predicted mean and standard deviation for CAV i at time t.
$\mathcal{N}(\mu_i^t, \sigma_i^{t^2})$	Normal distribution for action a_i^t.
$mapping\_factor$	Scaling factor for action values.
$L^{CLIP}(\theta)$	PPO's clipped objective function.
$\hat{\mathbb{E}}_t$	Expectation over time.
$r_i^t(\theta)$	Probability ratio under policies for CAV i at time t.
ϵ	Clipping range hyperparameter.
w_1, w_2	Weights for distance metrics.
$d^t_{\text{desired},i}, d^t_{\text{safe},i}$	Desired and safe distances for CAV i at time t.

3.2.1. DRL Design

DRL can be recognized as a Markov decision process, consisting of four elements: state, action, policy, and reward.

The state representation $\mathbf{s}_i^t$ in the DRL framework captures each CAV's velocity and position. The definition is as follows:

$$\mathbf{s}_i^t = [v_1^t, x_1^t, v_2^t, x_2^t, \ldots, v_n^t, x_n^t] \tag{1}$$

where, the velocity and position of the i-th CAV at time t are represented by v_i^t and x_i^t respectively.

The action a_i^t represents the acceleration or deceleration for CAV i at each timestep. It is defined as:

$$a_i^t \in \mathbb{R} \tag{2}$$

Specifically, we suggest implementing a method for mapping actions. In contrast to previous writers, such as Shi et. al,[19], we put forward a strategy for action selection that is based on distribution. This is because, in practical scenarios, it is challenging to alter the acceleration within a brief time frame significantly. In contrast to the value-based selection mechanism commonly proposed in research publications, the distribution-based mapping approach offers a more seamless acceleration and is more adept at managing uncertainties in real-world scenarios. This results in more resilient decision-making in the face of uncertainty. Specifically, the actions are sampled from a normal distribution, which is parameterized by neural network outputs that predicts the mean and standard deviation based on the current state[25, 26]:

$$\mu_i^t, \sigma_i^t = \mathrm{NN}(s_i^t), \quad a_i^t \sim \mathcal{N}(\mu_i^t, {\sigma_i^t}^2) \tag{3}$$

Here, μ_i^t and σ_i^t represent the mean and standard deviation of the action distribution for CAV i at time t, respectively. To ensure that the action values are within a feasible range, a hyperbolic tangent function ($tanh$) is applied, followed by scaling with a predefined mapping factor:

$$a_i^t = mapping\_factor \cdot \tanh(\mathcal{N}(\mu_i^t, {\sigma_i^t}^2)) \tag{4}$$

This scaling transformation allows the action values to be specifically tailored to the dynamic range required for optimal vehicle control, enhancing the flexibility and precision of the system.

The policy $\pi(s_i^t)$ is updated within the PPO framework using the clipped surrogate objective:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_i^t \left[\min(r_i^t(\theta)\hat{A}_i^t, \mathrm{clip}(r_i^t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i^t) \right] \tag{5}$$

where $r_t(\theta)$ is the ratio of the probabilities under the new and old policies for the taken action, and ϵ is a hyperparameter defining the clipping range, which helps control the magnitude of policy updates and ensure training stability.

The reward function $r(s_i^t, a_i^t)$ is designed to incentivize driving behaviors that improve safety and traffic flow efficiency:

$$r(s_i^t) = \sum_{i=0}^{n} \left(w_1 \cdot |d_i^t - d_{\mathrm{desired},i}^t| - w_2 \cdot \max(0, d_{\mathrm{safe},i}^t - d_i^t) \right) \tag{6}$$

This feature integrates multiple elements to guarantee optimal vehicle spacing. Simultaneously, it will impose penalties for the upkeep of hazardous gaps. By employing the control program, CAVs are able to prioritize safety while also maintaining optimal efficiency.

3.2.2. MAPPO Algorithm

MAPPO offers a decentralized perspective for organized oversight and effective and secure operation of CAV platoon management, particularly in situations where CAVs possess limited local observation capabilities. The MAPPO architecture utilizes the actor and critic components of the PPO algorithm to implement a centralized training and decentralized execution structure [27]. Therefore, every intelligent entity undergoes two distinct stages: centralized training and decentralized execution. It is important to understand that the training process occurs without an active connection, and exploration is employed to discover the most effective policy. During the execution phase, it is necessary to propagate forward without introducing any randomization into the exploration process. In the following sections, we will demonstrate the process of training the MAPPO model in a centralized manner and then implementing the trained model in a decentralized fashion. To show this, we will use one of n agents as an example.

During the centralized training process, the critic assumes the role of the central coordinator and calculates the centralized action-value function $Q(s^t, a^t|\phi)$ using the global state information, which

includes the actions and observations of all agents. The centralized Q function assesses the actor's activities from a comprehensive viewpoint and directs it toward selecting better actions. The critic network then updates the parameters by minimizing the loss:

$$\text{loss}(\theta^C) = E\left[Q(s^t, a^t|\phi) - y^{MAPPO}\right]^2 \tag{7}$$

where,

$$y^{MAPPO} = r_i^t + \gamma Q(s^{t'}, at'|\phi') \tag{8}$$

Here $s = (s_1^t, s_2^t, \ldots, s_n^t)$, $a = (a_1^t, a_2^t, \ldots, a_n^t)$ and ϕ are the parameters of the critic network. $s^{t'}$ denotes the updated state of the target network and ϕ' is the updated parameter of the evaluation network. On the other hand, the actor-network updates the network parameters θ and outputs actions based on the centralized Q function computed by the critic network and its own observations. Specifically, the actor-network adjusts the network parameters θ directly in the direction of $\nabla_\theta J(\theta)$. The global state makes value learning faster and easier by assuming a centralized value function that changes a partially observable MDP into a fully observable MDP:

$$\nabla_\theta J(\theta) \approx E[\nabla_\theta Q(s, a|\phi)\nabla_\theta \pi_\theta(s)]. \tag{9}$$

During the decentralized execution process, the critic network is omitted, and only the network of trained actors operates online. Decentralized executors rely solely on local observations from CAVs to make choices. During the whole execution process, there is a single forward propagation procedure, resulting in significant reductions in time and computational resource consumption compared to the training phase. Within a group of actors educated using parametric methods, each actor has the ability to achieve an action that is very close to the best possible outcome, even without having knowledge of the actions taken by other actors.

3.3. Distributed Action Mapping Mechanism

Our implementation enhances the robustness and adaptability of the action selection process by utilizing a distribution-based approach rather than directly obtaining single action values. The policy network, parameterized by θ, outputs the parameters of a Gaussian distribution: the mean μ and standard deviation σ. The action a_i^t is sampled from this distribution:

$$a \sim \mathcal{N}(\mu(s_i^t|\theta), \sigma(s_i^t|\theta))$$

where s_i^t represents the state input of CAV i at time t. This approach allows the agent to explore a range of possible actions instead of being restricted to deterministic policy outputs. The sampled action is then transformed using a hyperbolic tangent function to map values to the desired range:

$$a_i^{t'} = \tanh(a_i^t)$$

This bounds the action values within $[-1, 1]$, ensuring feasible control task actions. Mathematically, this can be represented as:

$$a_i^{t'} = \tanh(\mu(s_i^t|\theta) + \sigma(s_i^t|\theta) \cdot \epsilon)$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is a noise term to facilitate exploration. This distribution-based approach provides several advantages: balanced exploration and exploitation, robustness against uncertainties and environmental variations, and effective end-to-end training via the reparameterization trick.

The stochastic policy $\pi(a_i^t|s_i^t, \theta)$ is expressed as:

$$\pi(a_i^t|s_i^t, \theta) = \frac{1}{\sigma(s_i^t|\theta)\sqrt{2\pi}} \exp\left(-\frac{(a_i^t - \mu(s_i^t|\theta))^2}{2\sigma(s_i^t|\theta)^2}\right)$$

The transformed action $a_i^{t'}$ ensures a smooth, bounded action space, enhancing stability and control in real-world scenarios.

4. Experiments and Evaluation

In this section, we conduct a series of numerical experiments to evaluate the effectiveness of the proposed distribution-based MAPPO control strategy for CAV platoon management. The experiments are designed to compare the performance of the MAPPO strategy with distribution-based and value-based PPO strategies under varying conditions of communication reliability. We utilize SUMO as the traffic simulation platform to create controlled scenarios and analyze key metrics such as velocity, acceleration, vehicle spacing, and reward values. Additionally, we investigate the impact of different packet loss rates on the relative congestion index (RCI) to assess the robustness and generalizability of the control algorithms.A significant number of studies have examined mixed traffic efficiency utilising travel rate and congestion index as primary performance metrics. The trip rate indicates the duration a vehicle need to traverse a certain segment of the road network, whereas the RCI provides a more precise assessment of traffic flow conditions. An RCI score exceeding 2 signifies extremely heavy traffic congestion.

4.1. Evaluation Scenario

At the onset of the experiment, we establish a less complex scenario as a preliminary stage. The experiment utilizes SUMO (Simulation of Urban MObility) as the traffic simulation platform to create a scenario consisting of a single-lane, intersection-free, infinitely long straight road. This scenario is specifically created to accurately replicate the behavior of vehicles traveling in formation. The vehicle formation consists of six vehicles, categorized into two types: HDVs and CAVs, as shown in Figure 1. The HDV serves as the lead vehicle and is formally represented by SUMO's Krauss model. It possesses a driving behavior index (sigma) of 0.7, indicating a high level of randomness in its behavior. To differentiate it from the other vehicles, the HDV is highlighted in red. All five following vehicles are CAVs and are visibly designated with the color yellow. The CAVs, listed in order from left to right, are $veh0, veh1, veh2, veh3, veh4$. The lead vehicle is an HDV with the number $vehx$. All vehicles follow an identical path and are arranged in a straight line along the road. All vehicles initially have a velocity of 10 m/s and are positioned sequentially with 20-meter gaps between them. The allowed range of velocities is between -3 m/s and +3 m/s, and the maximum velocity is capped at 30 m/s. The simulation begins at time 0 seconds and continues until 60 seconds. Each simulation step is set to 0.1 seconds to ensure accurate data collection and real-time performance, allowing the algorithm to respond effectively to dynamic changes in the simulation environment. Fluctuating alterations occur as the scenario progresses.

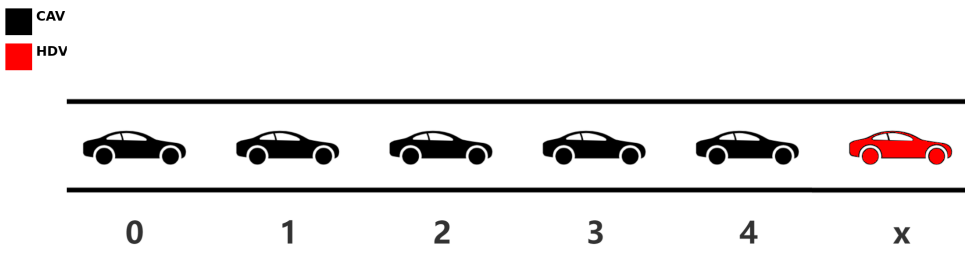
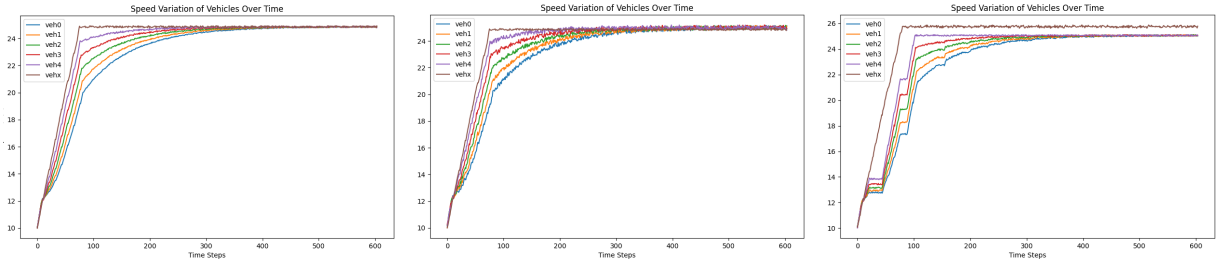


Figure 1: Simulation environment

4.2. Performance Evaluation

During the creation of the experimental situations, we systematically compare the suggested distribution-based MAPPO control strategy, the distribution-based PPO control strategy, and the value-based PPO control strategy. Here are two PPOs, one implementing the distribution-based mapping mechanism presented in this paper and the other directly calculating the values. Fig. 3-6 display the simulated outcomes with a 10% packet loss rate obtained from applying three distinct strategy to CAVs platoon.

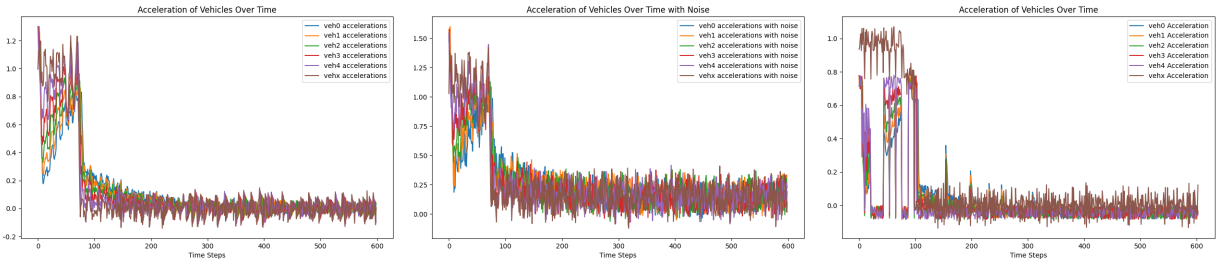
Our primary concern is the velocity variation of the platoon throughout the entire procedure. Figure 2 demonstrates that the distribution-based MAPPO method achieves smoother and more stable speed transitions, aligning well with the expected performance of a distribution-based strategy. In contrast, while the distribution-based PPO employs a similar strategy, its overall speed fluctuations are more pronounced compared to MAPPO, indicating greater difficulty in achieving stability under the same conditions. Finally, the value-based PPO method, although it eventually reached a consistent speed, exhibited notable disparities in speed and intermittent fluctuations in speed as compared to the leading HDV *vehx*. This indicates a possible delay in response or adjustment.



(a) distribution-based MAPPO (b) distribution-based PPO (c) value-based PPO

Figure 2: Velocity results of the mixed vehicle platoon

Subsequently, we examine variations in the acceleration of the three control strategies. As shown in Figure 3, The acceleration of the distribution-based MAPPO strategy indicates that the vehicle experiences rapid acceleration and soon stabilizes within the first 100 time steps. This demonstrates a high level of smoothness and stability, efficiently managing the instability caused by unreliable communication and HDV. Furthermore, despite the distribution-based PPO strategy exhibiting faster acceleration during the initial stages, it is evident that it cannot still quickly adjust to the instability of HDV and is marginally less adaptable. Overall, the platoon managed by the distribution-based PPO method and the MAPPO strategy demonstrates a superior capacity to adjust to the unpredictability and volatility of the HDV. This is evident from the fact that the CAVs align more closely with the frequency of acceleration changes in the HDV. The value-based PPO strategy exhibits the greatest variation in acceleration during the entire experiment, particularly in the initial phase. The estimation of acceleration by CAVs is consistently lower than the actual value of HDV, indicating a clear deficiency in dynamic adaptability.

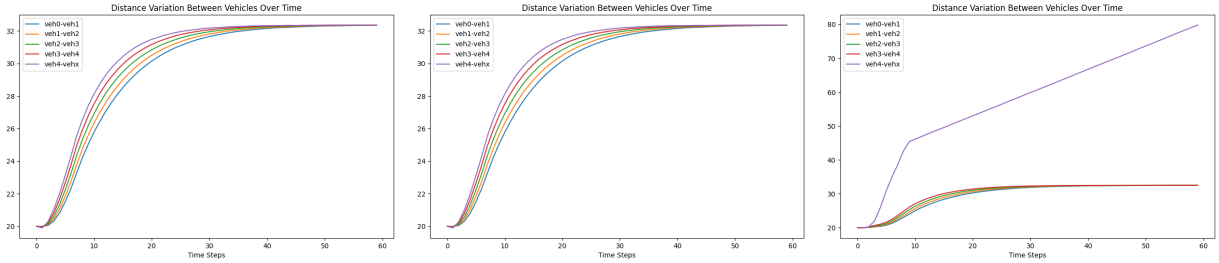


(a) distribution-based MAPPO (b) distribution-based PPO (c) value-based PPO

Figure 3: Acceleration results of the mixed vehicle platoon

Figure 4 demonstrates the strong consistency and stability of the distribution-based MAPPO and distribution-based PPO techniques in regulating the distance between vehicles. These two tactics efficiently achieve a rapid stabilization of vehicle spacing, minimizing fluctuations between vehicles and ensuring the overall formation and safe distance of the platoon. Nevertheless, the value-based PPO strategy demonstrated a substantial disparity from the leading vehicle, particularly throughout

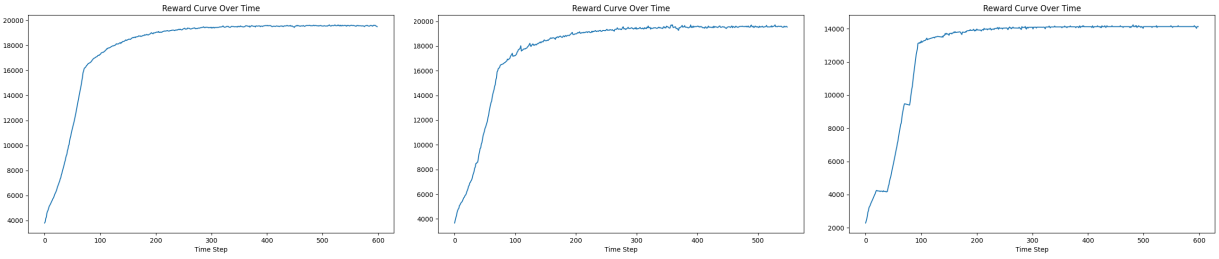
the middle and late phases of the experiment. The reason for this could be that value-based PPO may lack the flexibility of distribution-based methods when it comes to handling rapidly changing network conditions. As a result, it becomes challenging to maintain constant distances between cars.



(a) distribution-based MAPPO (b) distribution-based PPO (c) value-based PPO

Figure 4: Spacing results of the mixed vehicle platoon

Finally, we documented the reward curves as shown in Figure 5. From the reward curves, it is evident that both the distribution-based MAPPO and PPO strategies exhibit a significant increase in reward values during the initial phase of the experiment. However, after around 200-time steps, the reward values reach a plateau, with a final value close to 20,000. The swift convergence seen indicates that these two strategies, which rely on distribution-based methods, are highly effective in addressing unpredictable and dynamic settings. Furthermore, they demonstrate a remarkable ability to learn and optimize their performance. On the other hand, the PPO strategy is based on values, and while it also shows an initial increase in rewards during the experiment, it ultimately reaches a stable reward value of around 14,000, which is considerably lower than the other two strategies. The results indicate that distribution-based methods may offer better performance in a reinforcement learning system, particularly in communication-limited contexts that need complexity and robustness.



(a) distribution-based MAPPO (b) distribution-based PPO (c) value-based PPO

Figure 5: Reward results of the mixed vehicle platoon

4.3. Packet Loss Rate Effects

In addition to performance comparisons and tests, we also conducted experiments on the impact of packet loss rate on platoon control in mixed traffic scenarios. These experiments served to assess the algorithms' robustness and generalizability. We have selected the relative congestion index (RCI) as our evaluation criterion since it precisely represents the current condition of traffic flow operation. RCI values exceeding 2 indicate a significant level of traffic congestion. Analysis has demonstrated that trip rates are not a reliable indicator of traffic flow conditions in crowded traffic situations[28]. In situations of heavy traffic congestion, vehicles either move at slow rates with few or no changes in speed, or at somewhat higher speeds with frequent speed changes.

The RCI of the three algorithms was measured at packet loss rates of 10%, 30%, and 50%.

Table 2
RCI Values for the Different Algorithms Under Various Packet Loss Rates

Packet Loss Rate	Distribution-based MAPPO	Distribution-based PPO	Value-based PPO
10%	1.247	1.253	1.392
30%	1.326	1.387	1.517
50%	1.449	1.597	1.684

Table 2 displays the RCI values of the three algorithms at various packet loss rates for comparison. The data clearly demonstrates that the distribution-based MAPPO method consistently achieves outstanding performance across all test conditions. The RCI value only exhibits a modest increase from 1.247 to 1.449, indicating exceptional control even in the presence of a high packet loss rate of 50%. This demonstrates the exceptional flexibility and resilience of the strategy in ever-changing and demanding conditions.

The distribution-based PPO strategy has commendable performance, as seen by its RCI value remaining within the lower range of up to 1.597. This suggests that it possesses effective traffic flow control capabilities, however slightly less superior compared to MAPPO.

At 50% packet loss, the RCI value of the value-based PPO method exhibits substantial oscillations, with a peak value of 1.684, which is notably greater than the RCI values of the other two techniques. This tendency indicates that the value-based method may have more difficulties in sustaining platoon stability and smoothness in environments with substantial packet loss, particularly in situations that need quick adaptation to changes in the mixed environment.

5. Conclusion

This paper introduces a MADRL strategy for CAVs in contexts with inconsistent communication. Our methodology enhances classic DRL by incorporating uncertainty in the platoon's HDVs and accounting for the effects of varying packet loss rates on the control algorithm. The suggested technique aims to alleviate the effects of unreliable communication links on control signals and improve the vehicle's reaction by using a novel distribution-based action mapping approach and a weighted aggregate reward function. During our research, we perform simulations implementing a group of six vehicles consisting of one HDV at the front and five CAVs following behind. The purpose is to assess the effectiveness of our proposed algorithm under different levels of communication reliability. The experiments involve doing performance analysis using various algorithms and evaluating the impact under varied packet loss rates. The results demonstrate that the suggested distribution-based control method surpasses the two analyzed DRL algorithms in terms of platoon control performance. The results of our study show that using MADRL is both possible and beneficial for improving the control of CAVs in situations under mixed traffic environments. This strategy significantly outperforms current approaches in dealing with uncertainty in communication. To expand on this study, future research could investigate the integration of lateral control into the CAV framework to more effectively handle the intricacies of merge, diverge, or reroute operations. Furthermore, the development of more precise modeling and prediction techniques for HDVs could lead to the implementation of more resilient and effective control mechanisms.

Acknowledgments

This work was supported by the Science Foundation Ireland Centre for Research Training in Artificial Intelligence (CRT AI) under Grant No. 18/CRT/6223, and the Science Foundation Ireland CONNECT Research centre Phase 2, Grant $13/RC/2077_p2$. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

[1] P. Falcone, F. Borrelli, J. Asgari, H. Tseng, D. Hrovat, Predictive active steering control for autonomous vehicle systems, IEEE Transactions on Control Systems Technology 15 (2007) 566–580.

[2] E. F. Camacho, C. Bordons, Model Predictive Control, Springer Science & Business Media, 2004.

[3] C. Liu, S. Lee, S. Varnhagen, et al., Path planning for autonomous vehicles using model predictive control, in: IEEE Intelligent Vehicles Symposium, IEEE, 2017, pp. 797–802.

[4] J. B. Rawlings, D. Q. Mayne, Model Predictive Control: Theory and Design, Nob Hill Publishing, 2009. URL: https://nobhillpublishing.com/MPC/.

[5] S. J. Qin, T. A. Badgwell, A survey of industrial model predictive control technology, Control Engineering Practice 11 (2003) 733–764. URL: https://doi.org/10.1016/S0967-0661(02)00186-7.

[6] D. Q. Mayne, J. B. Rawlings, C. Rao, P. Scokaert, Constrained model predictive control: Stability and optimality, Automatica 36 (2000) 789–814. URL: https://doi.org/10.1016/S0005-1098(99)00214-9.

[7] T. Campisi, A. Severino, M. Al-Rashid, G. Pau, The development of the smart cities in the connected and autonomous vehicles (cavs) era: From mobility patterns to scaling in cities, Infrastructures 6 (2021) 100.

[8] L. Dormehl, The Formula: How Algorithms Solve All Our Problems... and Create More, Penguin, 2014.

[9] E. Guizzo, How google's self-driving car works, IEEE Spectrum 18 (2011) 1–4.

[10] C. Yang, K. Ozbay, X. Ban, Developments in connected and automated vehicles, Journal of Intelligent Transportation Systems 31 (2017) 154–164.

[11] B. Kiran, I. Sobh, V. Talpaert, P. Mannion, et al., Deep reinforcement learning for autonomous driving: A survey, IEEE Transactions on Intelligent Transportation Systems 22 (2021) 712–733.

[12] Y. He, N. Zhao, H. Yin, Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach, IEEE Transactions on Vehicular Technology 66 (2017) 10660–10675.

[13] J. M. Anderson, N. Kalra, K. D. Stanley, P. Sorensen, C. Samaras, O. A. Oluwatola, Autonomous vehicle technology: A guide for policymakers, Rand Corporation (2014).

[14] R. Garcia, M. Tawadrous, D. Martin, A survey of deep learning techniques for autonomous driving, Journal of Robotics and Automation 2015 (2015) 111–122.

[15] H. J. Li, S., D. Sun, Cooperative control of heterogeneous connected vehicle platoons: An adaptive leader-following approach, IEEE Transactions on Intelligent Transportation Systems 20 (2019) 761–772.

[16] J. R. Wang, H., W. Zhang, Robust cooperative adaptive cruise control of heterogeneous vehicle platoons with uncertain dynamics, IEEE Transactions on Intelligent Transportation Systems 21 (2020) 1177–1187.

[17] T. L. Willke, P. Tientrakool, N. F. Maxemchuk, A survey of inter-vehicle communication protocols and their applications, IEEE Communications Surveys & Tutorials 11 (2009) 3–20.

[18] H. Hartenstein, K. P. Laberteaux, A tutorial survey on vehicular ad hoc networks, IEEE Communications Magazine 46 (2008) 164–171.

[19] H. Shi, Y. Zhou, X. Wang, S. Fu, S. Gong, B. Ran, A deep reinforcement learning-based distributed connected automated vehicle control under communication failure, Computer-Aided Civil and Infrastructure Engineering 37 (2022) 2033–2051.

[20] H. Shi, Y. Zhou, K. Wu, S. Chen, B. Ran, Q. Nie, Connected automated vehicle cooperative control with a deep reinforcement learning approach in a mixed traffic environment, Transportation Research Part C: Emerging Technologies 133 (2021) 103421.

[21] D. Chen, Z. Li, Y. Wang, L. Jiang, Y. Wang, Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic, IEEE Transactions on Intelligent Transportation Systems 23 (2022) 113–125.

[22] Y. Zhang, J. Zhao, G. Cao, Data dissemination in vehicular ad hoc networks, IEEE Signal Processing Magazine 28 (2011) 84–94.

[23] S. Ju, P. van Vliet, O. Arenz, J. Peters, Digital twin of a driver-in-the-loop race car simulation

with contextual reinforcement learning, IEEE Robotics and Automation Letters (2023). URL: https://www.ias.informatik.tu-darmstadt.de/uploads/Site/EditPublication/RAL_Siwei_Ju.pdf.
[24] A. Pinosky, I. Abraham, A. Broad, B. Argall, T. Murphey, Hybrid control for combining model-based and model-free reinforcement learning, The International Journal of Robotics Research 42 (2023) 337–355. doi:10.1177/02783649221083331.
[25] e. a. Zhou, Rl-based car-following model for cavs, Journal of Advanced Transportation 45 (2020) 123–134.
[26] S. Qu, L. Song, Z. Zhang, J. Ren, A physics-informed generative car-following model for connected autonomous vehicles, Entropy 25 (2023) 1050. doi:10.3390/e25071050.
[27] Y. Savid, R. Mahmoudi, R. Maskeliunas, R. Damaševičius, Simulated autonomous driving using reinforcement learning: A comparative study on unity's ml-agents framework, Information 14 (2023) 290. doi:10.3390/info14050290.
[28] K. Hamad, S. Kikuchi, Developing a measure of traffic congestion: Fuzzy inference approach, Transportation Research Record 1802 (2002) 77–85.

A. Online Resources

The sources for the ceur-art style are available via

- GitHub,
- Overleaf template.