2021 European Control Conference (ECC)
June 29 - July 2, 2021. Rotterdam, Netherlands

A Hysteretic Q-learning Coordination Framework for Emerging Mobility Systems in Smart Cities

Behdad Chalaki, *IEEE Student Member*, Andreas A. Malikopoulos, *IEEE Senior Member*

***Abstract*— Connected and automated vehicles (CAVs) can alleviate traffic congestion, air pollution, and improve safety. In this paper, we provide a decentralized coordination framework for CAVs at a signal-free intersection to minimize travel time and improve fuel efficiency. We employ a simple yet powerful reinforcement learning approach, an off-policy temporal difference learning called Q-learning, enhanced with a coordination mechanism to address this problem. Then, we integrate a first-in-first-out queuing policy to improve the performance of our system. We demonstrate the efficacy of our proposed approach through simulation and comparison with the classical optimal control method based on Pontryagin's minimum principle.**

I. Introduction

OVER the last decade, the growing population in urban areas without increasing the road capacities has led to traffic congestion, increasing delays, and environmental concerns [1]. The introduction of communication technologies along with computational capabilities into *connected and automated vehicles* (CAVs) has the potential to revolutionize the transportation system to an *emerging mobility system*, in which CAVs can make better operational decisions leading to significant reductions of energy consumption, travel delays, and improvements to passengers' safety [2].

After the seminal work of Levine and Athans [3] on safely coordinating vehicles at merging roadways, several research efforts have explored the benefits of coordinating CAVs in traffic scenarios, such as urban intersections [4]–[8], merging roadways [9]–[11], and speed reduction zones [12] to eliminate congestion in a transportation network while preserving safety by using classical control techniques. A compendious survey of the research efforts reported in the literature to date in control and coordination of CAVs using classical control approaches is provided in [13] and [14].

The evolution of processing power and generation of a massive amount of data have paved the way for reinforcement learning (RL) techniques to emerge as an alternative method for traffic control. RL approaches are used when an agent learns from interaction with an environment without requiring the complete models of environment. Kiumarsi *et al.* [15] surveyed various RL-based techniques to solve optimal control problems in real-time using data measurement along the system trajectories. Q-learning is one of the simplest and most promising RL methods introduced by Watkins [16] in 1989. Since then, numerous studies have been reported in the literature to employ Q-learning in various transportation applications such as traffic signal control [17], ramp-metering control [18]–[20], smart lane-changing maneuvers [21], overtaking [22], and autonomous intersection management [23]. Wu *et al.* [23] modeled CAVs crossing the intersection as a *multi-agent Markov decision process* (MAMDP), in which CAVs cooperate to minimize the intersection delay, and solved it through Q-learning. To mitigate the "curse of dimensionality" and environment non-stationarity, they decomposed the state space of the system for each agent into independent and coordinated parts. The authors updated the corresponding Q-values for those parts separately for each CAV.

Although there have been several research efforts reporting on Q-learning-based frameworks for different transportation applications, to the best of our knowledge, no paper has reported work on a decentralized RL-based coordination framework for CAVs at an intersection intending to minimize energy consumption and improve traffic throughput. In this paper, we establish a decentralized coordination framework for CAVs at a signal-free intersection to minimize travel time and improve fuel efficiency. We formulate the problem by employing a well-known RL approach enhanced with a coordination mechanism called a hysteretic Q-learning, in which two learning rates are considered. Additionally, we integrate a first-in-first-out (FIFO) queuing policy in our hysteretic Q-learning framework to improve the performance of our system. We show our proposed approach's effectiveness through simulation and comparison with the classical optimal control method based on Pontryagin's minimum principle. The contributions of this paper are: (1) the development of a hysteretic Q-learning optimal framework to coordinate CAVs at a signal-free intersection aimed at decreasing both travel time and fuel consumption of each CAV; (2) integrating FIFO queuing policy into our hysteretic Q-learning optimal framework; and (3) comparison of the proposed framework with the benchmark solution from the classical control method based on Pontryagin's minimum principle.

The proposed framework advances the state of the art in the following ways. First, rather than considering a single agent in the RL framework [24]–[26], we propose a decentralized multi-agent framework with 100% penetration rate of CAVs. Second, in contrast to [23], [24], [26], we incorporate energy consumption minimization into our framework in addition to traffic throughput improvement while ensuring both lateral and rear-end safety through our combined hysteretic Q-learning with FIFO framework. Third, in contrast to the

This research was supported in part by ARPAE's NEXTCAR program under the award number DE- AR0000796 and by the Delaware Energy Institute (DEI). This support is gratefully acknowledged.

The authors are with the Department of Mechanical Engineering, University of Delaware, Newark, DE 19716 USA (emails: {bchalaki;andreas}@udel.edu).

research efforts reported in the literature to date, we compare the results of our proposed framework with the classical control method based on Pontryagin's minimum principle.

The rest of the paper is structured as follows. In Section II, we introduce the modeling framework. We present the simulation framework in Section III, and the corresponding results in Section IV. We present concluding remarks and some discussion for a future research direction in Section V.

II. Problem Formulation

We consider a signal-free intersection, which includes a *coordinator* that stores information about the intersection's geometric parameters and CAVs' information. The coordinator does not make any decision, and it only acts as a *database* among the CAVs. The intersection includes a *contol zone* in which the coordinator can communicate with the CAVs. We assume, there are no errors or delays in the vehicle-to-vehicle and vehicle-to-infrastructure communication. Although this is a strong assumption, it is relatively straightforward to relax this assumption as long as the noise or delays are bounded [8], [27]. We call the area inside the control zone where lateral collisions may occur *merging zone*. The distance from the entry of the control zone to the entry of the merging zone is $L \in \mathbb{R}_{>0}$, and it is assumed to be the same for all entry points. The length of the merging zone is denoted by $D \in \mathbb{R}_{>0}$. We limit our analysis to the cases where left/right turns and lane-changing maneuvers are not allowed.

The problem is formulated as a MAMDP $\langle n,\ \mathcal{S},\ \mathcal{U},\ P,\ R,\ \gamma\rangle$, where $n \in \mathbb{N}$ is total number of CAVs, $\mathcal{S} \coloneqq \times_{i=1}^{n}\mathcal{S}^i$ is a finite set of states of all CAVs, $\mathcal{U} \coloneqq \times_{i=1}^{n}\mathcal{U}^i$ is the joint action space, where $\mathcal{U}^i,\ i \in \{1, 2, \dots, n\}$ is the finite set of actions of CAV i, $P \coloneqq \mathcal{S} \times \mathcal{U} \times \mathcal{S} \to [0, 1]$ is the state transition probability which defines the transition probability between states, $R \coloneqq \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ is the reward function for all CAVs, R^i is the reward function for CAV i, and $\gamma \in [0, 1]$ is a discount factor.

Next, we briefly explain different approaches to formulate the Q-learning updates along with advantages and disadvantages of each approach. For CAV i, we use s_k^i and u_k^i to denote the state and action that CAV i takes at time step $k \in \mathbb{N}$, respectively. Taking action $u_k^i \in \mathcal{U}^i$, CAV i transitions from $s_k^i \in \mathcal{S}^i$ to $s_{k+1}^i \in \mathcal{S}^i$ and receives the reward $r_k^i = R^i(s_k^i, u_k^i)$. In the centralized update rule, the multi-agent system is viewed as a whole and is solved as a single-agent learning task, in which there is only a single Q-function. The update rule is

$$Q(s_k, u_k^1, \dots, u_k^n) \leftarrow (1-\alpha)Q(s_k, u_k^1, \dots, u_k^n) + \alpha\left[r_k + \gamma \max_{u_{k+1}^1, \cdots, u_{k+1}^n} Q(s_{k+1}, u_{k+1}^1, \dots, u_{k+1}^n)\right], \quad (1)$$

where $s_k \in \mathcal{S}$ is the state of the system (collection of states of all CAVs), $r_k = R(s_k, u_k^1, \dots, u_k^n)$ is the total cost incurred on the system at time step $k \in \mathbb{N}$, and $\alpha \in (0, 1]$ is the learning rate. Although, theoretically, this approach converges with probability 1 to the optimal action-value function, it does not scale well when the number of agents is increasing as the size of Q-table grows exponentially.

In the decentralized framework, each CAV is an *independent learner* (IL) with a corresponding Q-function. The update rule for CAV i is

$$Q^i(s_k^i, u_k^i) \leftarrow Q^i(s_k^i, u_k^i) + \alpha\left[r_k^i + \gamma \max_{u_{k+1}^i} Q^i(s_{k+1}^i, u_{k+1}^i) - Q^i(s_k^i, u_k^i)\right], \quad (2)$$

where $s_{k+1}^i \in \mathcal{S}^i$ is the state of CAV i at time step $k+1$. This approach has a smaller size Q-table than the centralized approach, and by increasing the number of agents, the Q-table's size does not grow exponentially. However, one of the drawbacks of this method is the lack of any coordination mechanism.

In our problem, CAVs need to coordinate to cross the intersection safely. Without a coordination mechanism, a CAV may select an optimal action, but it gets penalized due to the sub-optimal actions of other CAVs. It has been shown in [23], [28], [29], that using decentralized learning in a multi-agent framework with interacting agents leads to more oscillation in the learned policy and poorer performance compared to the centralized approach. In addition, since all CAVs are learning synchronously, the environment is not stationary anymore from the perspective of any CAV. Since past actions of some CAVs may affect the current behavior of other CAVs, the system is not Markovian. The latter implies that convergence is not guaranteed for every single CAV [29].

Matignon *et al.* [29] first presented the hysteretic Q-learning approach to incorporate coordination among ILs in a decentralized RL framework by including two learning rates. They showed that by incorporating two learning parameters α and β, without affecting the Q-table size, the coordination among IL agents could be achieved. In addition, the performance of the system is as good as the centralized approach of multi-agent RL. The update rule is given by

$$\delta \leftarrow r_k^i + \gamma \max_{u_{k+1}^i} Q^i(s_{k+1}^i, u_{k+1}^i) - Q^i(s_k^i, u_k^i), \quad (3)$$

where

$$Q^i(s_k^i, u_k^i) = \begin{cases} Q^i(s_k^i, u_k^i) + \alpha\delta, & \text{if } \delta \geq 0, \\ Q^i(s_k^i, u_k^i) + \beta\delta, & \text{otherwise,} \end{cases} \quad (4)$$

where δ is a temporal difference (TD) error and $\beta < \alpha \in (0, 1]$. By using smaller learning rate when TD error is negative, the update results in a slower degradation of Q-value (hysteresis) associated with positive past experience. For instance, due to the sub-optimal actions of other CAVs in the environment, CAV i may get penalized by doing action u^i at state s^i, for which it received a positive reward in the past. In this case, the effects of this penalty on Q-value of agent i should be less important.

A. Main Elements of Proposed Framework

We adopt the hysteretic Q-learning formulation to update the Q-tables for each CAV which is an IL agent with a unique

assigned index. Next, we present the main elements of our approach, including states, actions, and rewards.

1) State space: At time step k, we consider that CAV i partially observes the system, and its state is $s_k^i := \left\langle p_k^i,\ v_k^i,\ \mathcal{X}_k^{\text{i,rear}},\ \mathcal{P}_k^{\text{i,lat}} \right\rangle$, where p_k^i and v_k^i are its position and speed, respectively; $\mathcal{X}_k^{\text{i,rear}} := \left\langle p_k^j,\ v_k^j \right\rangle$ consists of the position and speed of CAV j, which is immediately ahead of CAV i

2) Action space: CAV i has to choose action u_k^i at time step k which is acceleration/deceleration from a discrete bounded set $\mathcal{U}^i$ with lower bound $u^{i,\min}$ and upper bound $u^{i,\max}$ which correspond to the minimum and maximum allowable control input of CAV i, respectively. Without loss of generality, we do not consider variation among CAVs' maximum and minimum control input. To this end, we set $u^{i,\min} = u^{\min}$ and $u^{i,\max} = u^{\max}$. In order to choose all actions in all states with nonzero probability and balance between exploration and exploitation, we employ the epsilon-greedy algorithm with a linear decay as follows

$$\rho = \max\left\{\frac{\text{total episodes} - \text{current episode}}{\text{total episodes}}, 0\right\}, \tag{5}$$

$$\epsilon = (\epsilon_i - \epsilon_f)\rho + \epsilon_f, \tag{6}$$

where ρ, ϵ_i, and ϵ_f are decay rate, initial and final ratio of exploration, respectively. In a RL framework, each episode represent a simulation, in which there is a corresponding epsilon found from (6). The corresponding epsilon determines the probability that an agent takes a random action at each episode. The epsilon found from (6) is bounded between initial and final ratio of exploration.

$$u_k^i = \begin{cases} \arg\max_{u_k^i} Q^i(s_k^i, u_k^i), & \text{with probability } 1-\epsilon, \\ \text{random action}, & \text{with probability } \epsilon, \end{cases} \tag{7}$$

where ϵ is a small positive number. Employing epsilon-greedy with a linear decay results in more exploration at the earlier episodes and less at the final episodes which can improve the performance of the framework.

3) Rewards: CAV i takes an action u_k^i at time step k, transitions from state s_k^i to the new state s_{k+1}^i, and receives a reward (or penalty) r_k^i based on the multi-objective cost $r_k^i = w_1 \cdot r_{\text{fuel}}^i + w_2 \cdot r_{\text{delay}}^i + w_3 \cdot r_{\text{speed}}^i + w_4 \cdot r_{\text{rear}}^i + w_5 \cdot r_{\text{lateral}}^i$, where $w_1, \ldots, w_5 \in \mathbb{R}_{\geq 0}$ are the weighting factors corresponding to the following costs.

a) Fuel Efficiency: We use the L^2-norm of the control input at each time step k as a penalty to reduce the control effort, which decreases fuel consumption.

$$r_{\text{fuel}}^i = -\frac{\|u_k^i\|^2}{(\max\{\|u^{\max}\|, \|u^{\min}\|\})^2}. \tag{8}$$

b) Delay: To improve the travel time, we define the time delay at each time step as a difference between the time that it takes for CAV i to reach its current position from the entry of the control zone and the time it would have taken for CAV i to cruise with the initial speed from the entry of the control zone until its current position. Considering CAV i at time step k, the traveled distance measured from the entry of the control zone is p_k^i, and its entry speed is denoted by v_0^i, the normalized penalty corresponding to delay is $r_{\text{delay}}^i = -\dfrac{(k\Delta t - \tau)}{\tau}$, where $\Delta t \in \mathbb{R}_{>0}$ is the time step and $\tau = \frac{p_k^i}{v_0^i}$.

c) Speed Limits Violation: For each CAV i, at each time step k, the speed is bounded by $0 \leq v^{\min} \leq v_k^i \leq v^{\max}$, where $v^{\min}, v^{\max}$ are the minimum and maximum speed limit, respectively. To ensure the speed constraint does not become active, we have

$$r_{\text{speed}}^i = \begin{cases} p_{\text{speed}}, & \text{if speed violates the constraint,} \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $p_{\text{speed}} \in \mathbb{R}_{<0}$ is the penalty for violating the speed constraint, and is decided a priori.

d) Rear-end Safety: To ensure the absence of rear-end collision between CAV i and a preceding CAV j at time step k, we impose $p_k^j - p_k^i \geq d_{\text{safe}}$, where $d_{\text{safe}} \in \mathbb{R}_{>0}$ is a safe constant distance. The associated penalty for violating rear-end safety at each time step k is

$$r_{\text{rear}}^i = \begin{cases} p_{\text{col}}, & \text{if } \ p_k^j - p_k^i < d_{\text{safe}}, \\ 0, & \text{if } \ p_k^j - p_k^i \geq d_{\text{safe}}, \end{cases} \tag{10}$$

where $p_{\text{col}} \in \mathbb{R}_{<0}$ is the penalty for violating the safety, and is decided a priori.

e) Lateral Safety: To guarantee lateral safety as CAVs cross the merging zone, we limit the merging zone occupancy to only one CAV at a time for CAVs with lateral collision potentials. Considering that CAVs i and j might have a lateral collision inside the merging zone, we construct the penalty according to the following two cases: Case 1: CAV i is outside the merging zone at time step k, and by taking an action u_k^i, it enters the merging zone while CAV j, which previously entered the merging zone, is either still inside the merging zone, or it enters the merging zone at the same time as CAV i. In this case, CAV i receives the penalty $r_{\text{lateral}}^i = p_{\text{col}} \in \mathbb{R}_{<0}$. Case 2: CAV i is outside the merging zone at time step k, and by taking an action u_k^i, it enters the merging zone while, at the same time step k, CAV j exits the merging zone. In this case, CAV i receives no penalty.

If by the time CAV i enters the merging zone, there are more than one CAVs inside the merging zone that can cause lateral collision with CAV i, then CAV i gets penalized for each of these CAVs separately equal to p_{col}. To encourage CAVs to have a safe pass through the intersection, each CAV receives a terminal reward equal to $n \cdot r_{\text{suc}}$ (recall that n is the total number of CAVs) when it exits the control zone, if the episode did not have any crashes, where $r_{\text{suc}} \in \mathbb{R}_{>0}$ is the success reward. Additionally, the episodes with a crash are terminated, and a new episode starts.

B. FIFO queuing policy

In this subsection, we provide a brief overview of a common approach in motion planning of CAVs at signal-free intersections called FIFO queuing policy. By imposing

a FIFO queuing policy, each CAV must enter the merging zone in the same order that it entered the control zone. Let $N(t) \in \mathbb{N}$ be the total number of CAVs that have entered the control zone by the time t, $\mathcal{N}(t) = \{1, \ldots, N(t)\}$ be the queue which designates the order in which CAVs enter the merging zone, and $t_i^0,\ t_i^m \in \mathbb{R}_{>0}$ denote the time when CAV $i \in \mathcal{N}(t)$ enters the control zone and merging zone, respectively. The optimal entry time, which satisfies safety and speed constraint, can be found through the following recursive structure [4]. If $i = 1$, $t_i^{m^*} = t_i^0 + \frac{L}{v_0^i}$; if $i-1 \in$ *safe*, $t_i^{m^*} = \max\{t_{i-1}^{m^*},\ t_j^{m^*} + t_h,\ t_i^c\}$;or if $i-1 \in \{$*lateral*, *rear-end*$\}, t_i^{m^*} = \max\{t_{i-1}^{m^*} + t_h,\ t_i^c\}$, where based on the path of CAV $i-1$ and CAV i, CAV $i-1$ belongs to one of the following subsets: (1) *safe*, if there is no potential for collision with CAV i. (2) *lateral*, if there is a potential for lateral collision with CAV i. (3) *rear-end*, if CAV $i-1$ is the CAV immediately positioned in front of CAV i. The earliest feasible time that CAV i can reach the merging zone is denoted by t_i^c, and t_h is the safe time-headway to ensure safety at the entrance of the merging zone. If $i = 1$, CAV i cruises with the constant speed that it entered the control zone. Index j in t_j^m represents CAV j which is physically located in front of CAV i.

C. Combined Hysteretic Q-learning with FIFO Framework

In our hysteretic Q-learning framework, we aimed at achieving the lateral safety through our state and reward architecture. Due to the fact that after each crash the simulation episode is terminated, an increasing number of CAVs may require a greater number of simulation episodes, which might become less applicable in the real systems. In this subsection, we introduce an enhanced framework which is a combination of FIFO and hysteretic Q-learning. In this framework, CAVs first find the optimal arrival time at merging zone recursively through a FIFO queuing policy at the start of each simulation episode. Since the lateral safety, and time-delay minimization are considered in the FIFO queuing policy [4], we need to modify the state and reward function in our Q-learning framework. The revised state of CAV i at time step k is $s_k^i := \left\langle p_k^i,\ v_k^i,\ \mathcal{X}_k^{\text{i,rear}}, \Delta t_i^{m^*} \right\rangle$, where p_k^i and v_k^i are its position and speed, respectively; and $\mathcal{X}_k^{\text{i,rear}} := \left\langle p_k^j,\ v_k^j \right\rangle$ consists of the position and speed of CAV j, which is immediately ahead of CAV i. The difference between the optimal arrival time at merging zone, and arrival time at the control zone is captured in the last element $\Delta t_i^{m^*} = t_i^{m^*} - t_i^0$, which takes value from a bounded set defined by speed limits of the roads. In the revised reward function, the weights regarding the lateral collision and delay terms are set to zero and $r_k^i = w_1' \cdot r_{\text{fuel}}^i + w_2' \cdot r_{\text{speed}}^i + w_3' \cdot r_{\text{rear}}^i + w_4' \cdot r_{\text{FIFO}}^i$, where $w_1', \ldots, w_4' \in \mathbb{R}$ are new weighting factors. To encourage CAV i to reach the merging zone at the planned arrival time $t_i^{m^*}$, we define $r_{\text{FIFO}}^i = -(\text{EAT} - t_i^{m^*})^2$ to be the negative of the normalized squared error of arrival time at the merging zone computed as the difference of the estimated arrival time (EAT) at the merging zone and the optimal arrival time. At time step k, the EAT of CAV i is approximated by assuming that CAV i cruises with a constant speed v_k^i for the rest of the path until the merging zone.

As CAV i enters the merging zone, the crossing time t_i^m is compared to the optimal arrival time $t_i^{m^*}$, and CAV i receives the last FIFO reward as $r_{\text{FIFO}}^i = p_{\text{FIFO}} \times (t_i^m - t_i^{m^*})^2$, where $p_{\text{FIFO}} \in \mathbb{R}_{<0}$ is the penalty for violating the FIFO arrival time, and is decided a priori. After entering the merging zone, w_4' for CAV i is set to zero.

III. Simulation Setup

In our decentralized hysteretic Q-learning framework, at each time step each CAV needs to store the updated Q-value corresponding to the pair of current state and selected action. The discretization level not only directly affects the size of the Q-table, which each CAV stores, but it also influences the performance of our approach. Hence, determining proper discretization levels is a trade-off between the Q-table size and performance of the algorithm. Moreover, selecting improperly large or small values for time discretization results in poor performance, or even oscillating behavior. For instance, selecting a very small time step compared to the state discretization level leads to a situation in which a CAV takes action, but its state does not change. On the other hand, by selecting a very large time step, the system's safety might be jeopardized. The discretization level for position, speed, control input, and time are selected as $\Delta p = 2$ m, $\Delta v = 1$ m/s, $\Delta u = 0.5$ m/s^2, and $\Delta t = 0.5$ s, respectively. The rest of the parameters are set as follows: $\epsilon_i = 0.6$, $\epsilon_f = 0.01$, $\alpha = 0.4$, $\beta = 0.05$, $p_{\text{speed}} = -1$, $p_{\text{col}} = -100$, $p_{\text{FIFO}} = -10$, $r_{\text{suc}} = 10$, $w_1 = w_3 = w_4 = w_5 = 1$, $w_2 = 0.3$, and $w_i' = 1, i = \{1, \ldots, 4\}$.

At the start of each episode, the initial conditions of CAVs are reset. In order to do that, the initial speed is drawn randomly from a uniform feasible speed distribution, and the arrival time of CAV i is computed as $t_i^0 = \sum_{a=0}^{i} Y_a$, where t_i^0 is the sum of i independent and identically distributed random variable Y_a drawn from an exponential distribution with mean 2 s.

IV. Simulation Results

To evaluate the effectiveness of our proposed framework, we investigate the coordination of CAVs at a signal-free intersection in two scenarios. We use the following parameters for the simulation: $d_{\text{safe}} = 4$ m, $v_{\min} = 5$ m/s, $v_{\max} = 15$ m/s, $u_{\max} = 3$ m/s^2, $u_{\min} = -3$ m/s^2.

Scenario 1: For our first scenario, we consider coordination of four CAVs using the hysteretic Q-learning framework in an intersection where the length of each road connected to the intersection is $L = 32$ m, the length of the merging zones is $D = 18$ m, and total episodes of simulation are set to 2,000,000. CAV #1, #2, #3, #4, enter the control zone from southbound (SB), eastbound (EB), northbound (NB), and westbound (WB), respectively. Figure 1 shows the Cartesian norm of Q-table for each CAV and the average norm, respectively; which are computed after each 100 episodes. As it can be seen in Fig. 1, the norm of Q-table for all four CAVs reach to stable values and does not change that much at the

final episodes. The difference in the converged value is due to the fact that during the earlier episodes of training, CAV #3 and #4 are more likely to cause accidents and get penalized compared to CAV #1 and CAV #2, since CAV #3 and #4 enter the control zone later. After the training phase, we test

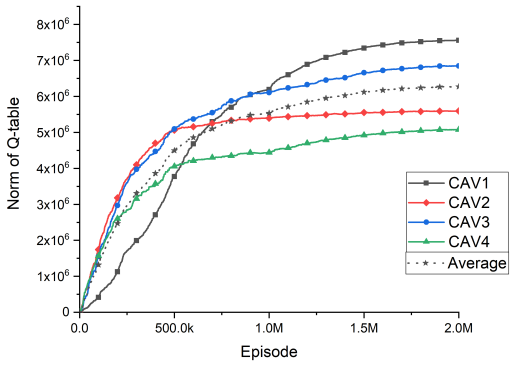

Fig. 1: Norm of Q-table for CAVs in Scenario 1.

the policies for 1,000 randomly generated simulation. The same initial conditions for four CAVs are used to simulate the optimal control framework. The optimal control framework consists of throughput maximization and energy minimization problems. In the throughput maximization problem, each CAV computes its arrival time at the merging zone based on a FIFO queuing policy. By restricting CAVs to have a constant speed after entering the merging zone, each CAV derives its energy-optimal control input from the control zone's entry until it reaches the merging zone considering speed and control constraints. Details of this approach can be found in [4]. The position trajectories of four CAVs for a randomly selected simulation is shown in Fig. 2 (solid lines). The major difference in the trajectories is because in the optimal control framework (dashed lines) the arrival time at merging zones for each CAV is found first, and then for each CAV, the optimal control problem is formulated from the arrival time at the control zone to the arrival time at the merging zone. On the other hand, our hysteretic Q-learning approach determines the policy with respect to the designed reward. Although trajectories resulted from our approach happen to respect the FIFO queuing policy (i.e., CAVs enter the merging zone in the same order they entered the control zone) without being enforced to, they appear to be more aggressive in minimizing the travel time. One can explore tuning w_1 and w_2 to find the trade-off between minimizing fuel consumption or delay.

Scenario 2: For the second scenario, we consider coordination of eight CAVs using the combined hysteretic Q-learning with FIFO framework in an intersection which each road connecting to the intersection to be $L = 100$ m, the length of the merging zones to be $D = 18$ m, and total episodes of simulation are set to 400,000. In this scenario, we employ the state and reward architecture based on FIFO queuing policy. This extension allows us to reduce the state

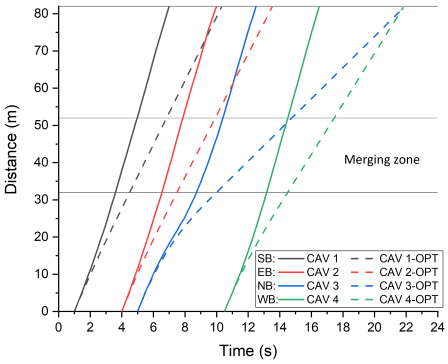

Fig. 2: Position trajectories of CAVs in Scenario 1.

space size significantly. Namely, we are able to increase the control zone length in order to have enough space for CAVs to coordinate with each other. The position trajectories of eight CAVs for a randomly selected simulation (solid lines) along with the corresponding trajectories computed from the optimal control (dashed lines) are shown in Fig. 3. We note that CAVs following our combined hysteretic Q-learning with FIFO framework arrive at the merging zone at the planned arrival time with very small deviation. The trajectories for CAVs in our combined hysteretic Q-learning with FIFO framework do not deviate very much from the energy-optimal trajectories found from the optimal control techniques. Our proposed RL-based approach requires more time in the training phase compared to the classical control techniques, but after that Q-table is converged, it can be implemented in real-time. The videos from our simulation analysis can be found at the supplemental site, https://sites.google.com/view/ud-ids-lab/HQLC

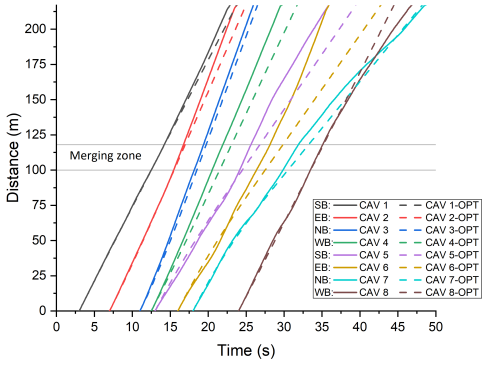

Fig. 3: Position trajectories of CAVs in Scenario 2.

V. Concluding Remarks and Discussion

In this paper, we proposed a learning-based decentralized coordination framework for CAVs at a signal-free intersection to minimize travel delay and improve fuel consumption while ensuring rear-end and lateral safety. We embedded

a coordination mechanism into our decentralized learning framework by using hysteretic Q-learning to update the Q-table of each CAV. We also integrated FIFO queuing policy in our framework to improve the performance of our system. Finally, we showed the effectiveness of our proposed approach through simulation and comparison with the classical optimal control method based on Pontryagin's minimum principle.

Ongoing research considers the presence of noise in the framework originated from the vehicle-level control and also investigates the effects of errors and delays in the communication. Our framework can be further extended to include lane-changing maneuvers and left/right turns by considering a different queuing policy instead of FIFO, such as upper-level motion planning proposed in [30]. Coordination for mixed-traffic scenarios and the interaction of human-driven vehicles and CAVs, is another potential direction for future research. Future studies should also investigate approaches for transferring the policy to real-world scenarios. We have explored zero-shot transfer of an autonomous driving policy inside a roundabout directly from simulation to a scaled testbed under Gaussian noise [31] and multi-agent adversarial noise [32].

REFERENCES

[1] B. Schrank, B. Eisele, and T. Lomax, "2019 Urban Mobility Scorecard," Texas A& M Transportation Institute, Tech. Rep., 2019.

[2] Z. Wadud, D. MacKenzie, and P. Leiby, "Help or hindrance? the travel, energy and carbon impacts of highly automated vehicles," *Transportation Research Part A: Policy and Practice*, vol. 86, pp. 1–18, 2016.

[3] W. Levine and M. Athans, "On the optimal error regulation of a string of moving vehicles," *IEEE Transactions on Automatic Control*, vol. 11, no. 3, pp. 355–361, 1966.

[4] A. A. Malikopoulos, C. G. Cassandras, and Y. Zhang, "A decentralized energy-optimal control framework for connected automated vehicles at signal-free intersections," *Automatica*, vol. 93, pp. 244–256, 2018.

[5] R. Hult, M. Zanon, S. Gros, and P. Falcone, "Optimal coordination of automated vehicles at intersections: Theory and experiments," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 6, pp. 2510–2525, 2018.

[6] Y. Bichiou and H. A. Rakha, "Developing an optimal intersection control system for automated connected vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1908–1916, 2018.

[7] J. Lee and B. Park, "Development and evaluation of a cooperative vehicle intersection control algorithm under the connected vehicles environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 81–90, 2012.

[8] B. Chalaki and A. A. Malikopoulos, "Optimal control of connected and automated vehicles at multiple adjacent intersections," *arXiv preprint arXiv:2008.02379*, 2020.

[9] I. A. Ntousakis, I. K. Nikolos, and M. Papageorgiou, "Optimal vehicle trajectory planning in the context of cooperative merging on highways," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 464–488, 2016.

[10] W. Xiao, C. Belta, and C. G. Cassandras, "Decentralized merging control in traffic networks: A control barrier function approach," in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019, pp. 270–279.

[11] J. Rios-Torres and A. A. Malikopoulos, "Automated and Cooperative Vehicle Merging at Highway On-Ramps," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 780–789, 2017.

[12] A. A. Malikopoulos, S. Hong, B. B. Park, J. Lee, and S. Ryu, "Optimal control for speed harmonization of automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 7, pp. 2405–2417, 2019.

[13] J. Guanetti, Y. Kim, and F. Borrelli, "Control of connected and automated vehicles: State of the art and future challenges," *Annual Reviews in Control*, vol. 45, pp. 18–40, 2018.

[14] J. Rios-Torres and A. A. Malikopoulos, "A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1066–1077, 2016.

[15] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2042–2062, 2017.

[16] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[17] S. El-Tantawy and B. Abdulhai, "An agent-based learning towards decentralized and coordinated traffic signal control," in *13th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2010, pp. 665–670.

[18] C. Jacob and B. Abdulhai, "Integrated traffic corridor control using machine learning," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4. IEEE, 2005, pp. 3460–3465.

[19] M. Davarynejad, A. Hegyi, J. Vrancken, and J. van den Berg, "Motorway ramp-metering control with queuing consideration using q-learning," in *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2011, pp. 1652–1658.

[20] E. Ivanjko, D. K. Nečoska, M. Gregurić, M. Vujić, G. Jurković, and S. Mandžuka, "Ramp metering control based on the q-learning algorithm," *Cybernetics and Information Technologies*, vol. 15, no. 5, pp. 88–97, 2015.

[21] L. Wang, F. Ye, Y. Wang, J. Guo, I. Papamichail, M. Papageorgiou, S. Hu, and L. Zhang, "A q-learning foresighted approach to ego-efficient lane changes of connected and automated vehicles on freeways," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1385–1392.

[22] D. C. K. Ngai and N. H. C. Yung, "A multiple-goal reinforcement learning method for complex vehicle overtaking maneuvers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 509–522, 2011.

[23] Y. Wu, H. Chen, and F. Zhu, "Dcl-aim: Decentralized coordination learning of autonomous intersection management for connected and automated vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 103, pp. 246–260, 2019.

[24] D. Isele and A. Cosgun, "To go or not to go: a case for q-learning at unsignalized intersections," 2017.

[25] S. M. Seliman, A. W. Sadek, and Q. He, "Automated vehicle control at freeway lane-drops: a deep reinforcement learning approach," *Journal of Big Data Analytics in Transportation*, pp. 1–20, 2020.

[26] T. Tram, A. Jansson, R. Grönberg, M. Ali, and J. Sjöberg, "Learning negotiating behavior between cars in intersections using deep q-learning," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3169–3174.

[27] B. Chalaki and A. A. Malikopoulos, "Robust learning-based trajectory planning for emerging mobility systems," *arXiv preprint arXiv:2103.03313*, 2021.

[28] L. Meng-Lin, C. Shao-Fei, and C. Jing, "Adaptive learning: A new decentralized reinforcement learning approach for cooperative multi-agent systems," *IEEE Access*, 2020.

[29] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 64–69.

[30] A. A. Malikopoulos, L. E. Beaver, and I. V. Chremos, "Optimal time trajectory and coordination for connected and automated vehicles," *Automatica*, vol. 125, p. 109469, 2021.

[31] K. Jang, E. Vinitsky, B. Chalaki, B. Remer, L. Beaver, A. A. Malikopoulos, and A. Bayen, "Simulation to scaled city: zero-shot policy transfer for traffic control via autonomous vehicles," in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019, pp. 291–300.

[32] B. Chalaki, L. E. Beaver, B. Remer, K. Jang, E. Vinitsky, A. Bayen, and A. A. Malikopoulos, "Zero-shot autonomous vehicle policy transfer: From simulation to real-world via adversarial learning," in *IEEE 16th International Conference on Control & Automation (ICCA)*, 2020, pp. 35–40.