AoI-Aware Resource Allocation for Platoon-Based C-V2X Networks via Multi-Agent Multi-Task Reinforcement Learning

Mohammad Parvini, Graduate Student Member, IEEE, Mohammad Reza Javan, Senior Member, IEEE, Nader Mokari, Senior Member, IEEE, Bijan Abbasi, Senior Member, IEEE, and Eduard A. Jorswieck, Fellow, IEEE

Abstract—This paper investigates the problem of age of information (AoI) aware radio resource management for a platooning system. Multiple autonomous platoons exploit the cellular wireless vehicle-to-everything (C-V2X) communication technology to disseminate the cooperative awareness messages (CAMs) to their followers while ensuring timely delivery of safety-critical messages to the Road-Side Unit (RSU). To lower the computational load at the RSU and cope with the challenges of dynamic channel conditions, we exploit a distributed resource allocation framework based on multi-agent reinforcement learning (MARL), where each platoon leader (PL) acts as an agent and interacts with the environment to learn its optimal policy. Motivated by the existing literature in RL, we propose two novel MARL frameworks based on the multi-agent deep deterministic policy gradient (MADDPG), named Modified MADDPG, and Modified MADDPG with task decomposition. Both algorithms train two critics with the following goals: A global critic which estimates the global expected reward and motivates the agents toward a cooperating behavior and an exclusive local critic for each agent that estimates the local individual reward. Furthermore, based on the tasks each agent has to accomplish, in the second algorithm, the holistic individual reward of each agent is decomposed into multiple sub-reward functions where task-wise value functions are learned separately. Numerical results indicate our proposed algorithms’ effectiveness compared with other contemporary RL frameworks, e.g., federated reinforcement learning (FRL) in terms of AoI performance and CAM message transmission probability.

Index Terms—Resource management, V2X, AoI, Platoon cooperation, MARL.

I. INTRODUCTION

A. Backgrounds

INTelligent transportation systems (ITSs) will become a compulsory component of the future’s smart cities. In essence, ITSs will address the issue of dense traffic networks and transportation bottlenecks by exploiting efficient traffic management approaches [1]. One of the foreseen services of ITS is the so-called autonomous vehicular platoon system [2]. Platooning is the first step toward fully autonomous driving, which is deemed one of the most representative potentials for overcoming the transport costs. Furthermore, platooning improves the intersection’s operational efficiency compared to the case where cars cross the intersection one after another [3]. In summary, a vehicle platoon is a convoy of interconnected vehicles that continuously coordinate their kinetics and share a typical moving pattern. In each platoon formation, the head-of-line vehicle is known as the Platoon Leader (PL), which is responsible for maintaining communication with other Platoon Members (PMs) [4]. In order to reap the benefits of the platooning system properly, several critical issues must be tackled. First, every vehicle in the platoon must have enough information about its relative distance and velocity with other vehicles in the platoon, especially the PL. This perception is needed to allow the vehicles in a platoon to regulate their decisions and to guarantee that any perturbation in the position or velocity of PL does not lead to amplified fluctuations in the behavior of PMs. This balance, known as the string stability, is ensured through the timely exchange of cooperative awareness messages (CAMs) among the vehicles of the platoon, and it is regularly initiated by the PL that manages the group [5]. Then, every platoon must have sufficient information about the other existing platoons and vehicles in the network, especially in the case of intersections or road curves. These points reflect the importance of investigating an efficient resource allocation algorithm that meets the requirements of both inter-platoon connectivity, i.e., communication between different platoons, and intra-platoon connectivity, i.e., communication between the vehicles of a platoon [6].

The advent of vehicle-to-everything (V2X) communication technology has addressed the aforementioned challenges. Platoons communicate with the Road-Side Unit (RSU) through...
vehicle-to-infrastructure (V2I) communications in order to exchange the intersection safety messages, while vehicles in the same platoon exchange CAM messages by either broadcasting or cellular vehicle-to-vehicle (V2V) communications. The more frequently information is exchanged in the network, the sooner each platoon member can react and avoid prospective obstacles [7]. The theoretical potential of Long Term Evolution (LTE) for V2X communications has been appraised in the Third Generation Partnership Project (3GPP) studies [8]. In LTE systems, eNodeBs centrally perform radio resource management (RRM). However, the conventional LTE architecture does not natively support direct V2V communications. Since LTE Release 12, 3GPP has provided several technical specifications to mitigate this problem through device-to-device (D2D) sidelink communications (known as Proximity Services) [9], [10]. Besides, new demands and use cases have been proposed for 5 G V2X enhancements in Release 15 [11].

**B. Motivations**

This paper considers the resource allocation problem in a platooning vehicular network, that incorporates both V2I and V2V connectivity. The V2I links connect the platoons to the RSU, while the V2V links pave the way for the PL and its PMs’ communication. Technically, two modes of resource scheduling have been perceived for V2X services [12], i.e. Mode 3 and Mode 4. In Mode 3 the selection of subchannels and interference control is handled by the RSU and is only available when vehicles are under cellular coverage. On the other hand, in Mode 4 the vehicles select their resources autonomously, and the radio resource allocation and interference management are based on distributed algorithms implemented between the vehicles.

In practice, when it comes to supporting a vast number of vehicles, Mode 3 of resource allocation imposes large overheads on the RSU side. Unlike the conventional vehicular networks, the emergence of platooning systems has introduced new services and requirements, and in most cases, multiple objectives have to be taken into account simultaneously. Accordingly, it is not reasonable to put all the communication and computing burden on the RSU. To fend off this ossification, Mode 4 of resource allocation is a proper solution, because, as long as all or part of the computational load is carried out by the platoons themselves, the network will subsist. However, as Mode 4 is implemented based on distributed algorithms, determining a globally effective method to perfectly match its necessities, indeed seems an insurmountable challenge. While there exists a rich body of literature that applies different conventional optimization methods to tackle these issues; however, their performance and efficiency diminish due to the high network dynamics caused by the mobility of vehicles. Furthermore, owing to the diverse services brought by platooning vehicular networks, the optimization problems are often multi-objective or sometimes hard to be modeled mathematically. Fortunately, recent progress of Reinforcement Learning (RL) has yielded prominent results, and it offers a principled solution towards handling the environment’s dynamics. It also provides a more straightforward way to handle the complex objectives by only translating them into a suitable reward function. Also, another intriguing characteristic of RL is its conformity with the multi-agent environments, which is perfectly matched with the Mode 4 of resource allocation that requires distributed algorithms. One such algorithm is multi-agent deep deterministic policy gradient (MADDPG), which has been widely applied in contemporary papers and has shown outstanding results. However, this algorithm cannot be applied or implemented directly in real-world scenarios due to several reasons. First, the MADDPG requires that each agent shares its information with other agents in the environment. In practice, this aspect of MADDPG can lead to controversial issues in terms of security. Second, the Achilles heel of this method is that the critic’s input grows linearly with the number of agents, and this will result in slow convergence as well as weak performance. Finally, the MADDPG algorithm is trained solely based on a single reward function which is common among the agents. Nevertheless, a careful contemplation can bring us to the conclusion that a single reward function cannot always cater to the needs of a communications system. In applications like vehicular networks, we are always dealing with various objectives, which are most of the time correlated; however, in order to fit them into the RL framework, we inevitably consolidate these objective functions to form a holistic reward function, which often fails to obtain an optimal solution for each objective separately. Therefore, the current structure of MADDPG has to be modified so that it can cope with the aforementioned concerns.

**C. Related Works**

Recently, the platooning system has been considered in various studies. The authors of [13] analyze the capability of the LTE system in establishing intra-platoon communication. In [14], the authors study the reliability and efficiency of the platoon-based V2V communication, investigate the string stability requirements for the platooning systems and design a CAM dissemination mechanism in the LTE-V2V network. The authors of [15] investigate the platoon cooperation in a multi-lane scenario and consider a two-step resource allocation along with developing a dynamic programming based subchannel allocation and power control algorithm to maximize the platoon size as well as to minimize the power consumption. In [16], string stability of the platoons and the maximum wireless system delay that guarantees the stability are analyzed. The resource allocation based on the evolved multimedia broadcast multicast services (eMBMS) capability and D2D communications is examined in [17] to enhance the reliability and reduce the transmission latency in a scenario with a chain of platoons. A two-stage platoon formation algorithm and a time division based intra-platoon resource allocation mechanism are introduced to develop stable platoons in [18]. Most of the issues that have been addressed in the articles mentioned above are related to the platoon’s communications and interactions with each other or controlling algorithms employed to ensure the platoon’s string stability. Nonetheless, an essential common concern that has not yet been elucidated is the fast-changing channel condition in vehicular environments that provoke uncertainty and inaccuracy in estimating the channel state information (CSI). On the other
hand, the gradual increase in users’ number leads to more complicated optimization problems with often nonlinear constraints, making them challenging to optimize by traditional optimization methods. The aforementioned hurdles call for investigating novel methods that can deal with more complex situations efficiently.

As one of the robust machine learning tools, RL has recently attracted substantial attention. In [19], the authors analyze the spectrum allocation scheme by devising a distributed Q-learning approach, where autonomous D2D users try to maximize their throughput and minimize their interference to cellular users. Furthermore, an intelligent resource management problem in the Internet of Vehicles (IoV) networks is analyzed in [20] using an actor-critic RL method. However, the RL methods applied in the above works are suitable in low-dimensional state and action spaces. RL in combination with deep learning has led to the emergence of deep reinforcement learning (DRL) [21]. DRL has sparked a flurry of interest and has found its way into vehicular network literatures [22], [23]. The authors of [24] propose a decentralized resource allocation method in a vehicular network for both unicast and broadcast scenarios employing DRL. In [25], a mobile edge computing-based platooning system has been proposed in which the platoons locate their optimal path through RL. The authors of [26] investigate the problem of channel assignment and power allocation in a platooning vehicular network using the DRL approach. In a similar framework, the spectrum and energy efficiency of the vehicular platooning network is examined in [27]. In addition, in [28], the joint problem of sub-channel selection and transmission power control is investigated. The authors aim to maximize the sum throughput of V2I links while satisfying the reliability and latency requirements of the V2V links, by applying the double dueling deep recurrent Q-network (D3RQ). In [29], the authors investigate the delay-aware user-centric content delivery problem in cache-enabled IoV. In order to address the uncertain cache state and time-varying wireless channels, the authors propose a double deep Q network (DDQN)-based algorithm. In [30], the authors investigate the spectrum sharing in a vehicular network by implementing a multi-agent DRL method. In order to tackle the problem of the environment’s non-stationarity, the authors propose a fingerprint method that incorporates agents’ policies in the observation space. The aforementioned literature focuses primarily on DRL or its multi-agent extension, multi-agent DRL. [24], [25], [27] and [29] model the policy search as a Markov decision process (MDP), which means that all the agents update their policies independently. However, although these algorithms are capable of handling many complex problems, they cannot be applied to multi-agent systems (MASs). In MASs, all the agents act simultaneously and affect the environment, leading to a non-stationary environment [31]. On the other hand, [26], [28], and [30] are based on multi-agent DRL. DRL algorithms employ discrete action spaces which is not appropriate in power control scenarios leading to poor results.

One of the widely applied MARL frameworks is MADDPG. Spectrum allocation for D2D communication is investigated in [32] in which the authors propose a multi-agent actor-critic method. In [33], the authors study the joint optimization of the channel allocation and power control in a vehicular network by considering both the safety and non-safety-related applications for V2V communications. Furthermore, in [34], a multi-dimensional resource management problem for a vehicular network is proposed. In order to provide on-demand resource access, the authors assume that both the macro eNodeB and UAV are equipped with multi-access edge computing (MEC) servers. All the surveyed literature that apply the MADDPG algorithm, share the same deficiency, that we have already mentioned in Section I-B. Not to mention that, although the proposed multi-agent algorithms reach an optimal solution, there is no explicit notion of coordination between the agents in these works.

In vehicular networks, the traffic and intersection safety information is time-critical, and hence acquiring timely, and fresh traffic updates are of significant importance. Recently, an emerging new metric has been employed for capturing the timeliness of the information, namely the age of information (AoI) [35]. By definition, AoI is the time elapsed since the most recent received information update (from RSU point of view) was generated (at the corresponding platoon). Unlike traditional metrics such as delay, AoI only takes the information that delivers fresh updates to the RSU into account [36]. One of the recent works in this area is [37] where the authors formulate an AoI-aware radio resource management problem in a Manhattan grid V2V network.

D. Contribution

This work considers the AoI minimization problem in a high mobility vehicular platooning system, consisting of multiple connected and autonomous vehicles where PLs attempt to access the frequency spectrum to disseminate the CAM messages between their followers through V2V communications while keeping an updated connection with the RSU over the V2I links. Following the existing literature, this work is based on Mode 4, defined in the 3GPP cellular V2X architecture [38]. The resource scheduling and interference management between the platoons are established based on distributed algorithms implemented between the vehicles [38], [39]. In addition, novel MARL frameworks have been designed that are perfectly consistent with the vehicular network’s requirements and allow for better flexibility in designing the reward function. The novelty of this work lies in the following key contributions:

- We formulate a multi-objective optimization problem for each platoon to jointly minimize the AoI and maximize the CAM message transmission probability.
- We model the spectrum access of the multiple PLs as a multi-agent problem and exploit the recent progress of MARL structures in [40] to build two novel MARL frameworks on top of deterministic policy gradients architectures, named Modified MADDPG, and Modified MADDPG with task decomposition. Both algorithms train two critics: A global critic which estimates the global expected reward and motivates collaboration between multiple agents, and an exclusive local critic for each agent that estimates the local expected reward.
- Unlike the conventional RL frameworks that employ a single reward function, based on the suggested architecture,
we propose two different reward functions: a global reward function which is based on the cumulative interference levels between the platoons, and a local reward function in which its elements are taken root from the objective functions of the optimization problem we are trying to solve. Furthermore, by treating each sub-objective as a separate task, the holistic reward of each agent is decomposed into multiple sub-reward functions for the Modified MADDPG with task decomposition algorithm, where task-wise value functions are learned separately.

- In order to tackle the problem of the overestimation bias in Q-functions, we exploit the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [41] for the global critic.
- Numerical experiments indicate that the proposed framework converges 3 times faster than the conventional RL frameworks and maintains the average AoI quantity within 5-10 milliseconds range, and guarantees a CAM message transmission probability of over 99.9% for various platoon sizes.

### E. paper Organization and Notations

The remainder of the paper is arranged as follows. In Section II, we discuss the proposed system model. Section III describes the proposed multi-agent reinforcement learning algorithms. In Section IV we analyze the complexity of the proposed algorithms. In Section V, we present the simulation results and analyses, and finally, Section VI concludes the paper. Most of the notations applied in this paper are standard. To ease readability, all the primary notations of the paper are listed in Table I.

### II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a cellular V2X based vehicular communication network which consists of one RSU and multiple Platoons, as shown in Fig. 1. The RSU is located at the center of the crossroad and is equipped with single antenna. We assume $\mathcal{P} = \{1, 2, \ldots, P\}$, $P \in \mathbb{N}$, indicates the set of platoons. Each platoon itself is comprised of some connected and automated vehicles. Let $\mathcal{N}_j = \{1, 2, \ldots, N_j\}$, $N_j \in \mathbb{N}$, be the number of vehicles in each platoon $j \in \mathcal{P}$ which are numbered sequentially from one to $N_j$, starting from PL. We discretize the time horizon into equal scheduling slots of length $\Delta t$, indexed by a positive integer $t \in \mathbb{N}$. The system bandwidth is divided into orthogonal subchannels of size $W$. They are indexed by $k \in \mathcal{K} = \{1, 2, \ldots, K\}$. In essence, there are two types of communication modes in a platooning system, namely the intra-platoon and inter-platoon communication. In intra-platoon communication, vehicles within the same platoon, exchange the CAM information periodically through V2I links. According to the 3GPP specifications, [11], CAMs dissemination frequency must be between 10 to 100 Hz. In other words, the CAMs distribution period must be kept in the range of 100 ms or fewer. In inter-platoon communication, the RSU exchanges the intersection safety and platoon control information with every platoon via the V2I links. The first one is crucial in terms of guaranteeing the platoon string stability which lets the vehicles keep a close distance with each other and ensuring that all the platoon members are aware of the kinematics and the decisions of the other platoon members, especially the platoon leader. The latter is essential to inform all the platoons to become aware of the other platoons’ status and traffic condition of the intersection. We exploit the orthogonal frequency division multiplexing (OFDM) to cope with the frequency selective wireless channels.\footnote{It is necessary to mention that in this work, we only consider the channel gains related to the platoon leader interactions with the RSU and its followers.}

Furthermore, we assume that the channel fading is independent across different subchannels and remains constant within one sub channel. We model the channel gain of PL $j \in \mathcal{P}$ in subchannel $k$ during one coherence time period $t$ as

$$h^j_k[k] = \alpha^j_k g^j_k[k],$$

where $\alpha^j_k$ and $g^j_k[k]$ denote the large-scale fading effect comprised of path loss and shadowing, and small-scale fading, respectively. Moreover, we define the binary variable $\beta^j_k \in \{0, 1\}$ that indicates whether subchannel $k$ is allocated to platoon $j$ at time slot $t$. Then PL $j$ will decide whether to use the allocated subchannel for inter-platoon (i.e., to communicate with the RSU) or intra-platoon (i.e., to broadcast the CAM to its followers) communication. For this reason, we define another binary decision

### TABLE I

| Notation | Definition |
|----------|------------|
| $\mathbb{N}$ | the set natural numbers |
| $P/P/j$ | number/index of platoons |
| $N_j/N_f$ | number/index of vehicles in platoon $j$ |
| $K/K/k$ | number/index of subchannels |
| $\alpha_j$ | frequency independent large-scale fading |
| $g_j[k]$ | frequency dependent small-scale fading |
| $R$ | RSU location |
| $\beta^j_k$ | subchannel allocation indicator |
| $\delta^j_{pl}$ | intra/platoon mode selection indicator |
| $c^j_{PL}[k]$ | data rate between PL $j$ and the RSU in subchannel $k$ |
| $h^j_{PL}[k]$ | channel gain from PL $j$ to RSU in subchannel $k$ |
| $c^j_{f}[k]$ | data rate between PL $j$ and its follower $i \in N_j$ |
| $h^j_{f}[k]$ | channel gain from PL $j$ to its PMs in subchannel $k$ |
| $p^j_{s}[k]$ | power usage of PL $j$ |
| $A^j_{\text{cam}}$ | CAM messages size of PL $j$ |
| $c^j_{\text{min}}$ | minimum capacity requirement of PL |

**Fig. 1.** The multi-lane platoon scenario.
variable \( \theta_j^t \in \{0, 1\} \) that indicates the platoon leader’s decision. When \( \theta_j^t = 1 \), that means that the PL will utilize the allocated subchannel for broadcasting (intra-platoon) and \( \theta_j^t = 0 \) indicates that the subchannel will be used for V2I (inter-platoon) communication. We can express the instantaneous rates achieved in V2I communications between PL \( j \) and the RSU according to the Shannon capacity formula as follows:

\[
C_{j,R}^t[k] = \log_2 \left( 1 + \frac{(1 - \theta_j^t)\beta_{j,k}^t p_j^t[k] h_{j,R}^t[k]}{I_{j}^t[k] + \sigma^2} \right),
\]

\[
I_{j}^t[k] = \sum_{j'} \beta_{j',k}^t p_{j'}^t[k] h_{j',R}^t[k], \quad j \neq j', \tag{2}
\]

where the interference from other platoons is treated as noise, \( p_j^t[k] \) is the transmit power level used by PL \( j \) on subchannel \( k \), \( h_{j,R}^t[k] \) is the channel gain from PL \( j \) to RSU in subchannel \( k \), \( \sigma^2 \) is the noise power, \( R \) indicates the RSU location, \( h_{j,R}^t[k] \) is the interfering channel to the RSU from PL \( j' \) in \( P \) functioning in whether inter (\( \theta_j^t = 0 \)) or intra-platoon (\( \theta_j^t = 1 \)) communication mode, and \( I_{j}^t[k] \) represents the total interference power. Furthermore, we can calculate the instantaneous rates between PL \( j \) and its follower \( i \) as

\[
C_{j,i}^t[k] = \log_2 \left( 1 + \frac{\theta_j^t\beta_{j,i,k}^t p_j^t[k] h_{i,j}^t[k]}{I_{j}^{t+1}[k] + \sigma^2} \right),
\]

\[
I_{j}^{t+1}[k] = \sum_{j'} \beta_{j',k}^t p_{j'}^t[k] h_{j',i}^t[k], \quad j \neq j', \quad i \in N_j \setminus \{1\}, \tag{3}
\]

where \( p_j^t[k] \) is the power used by PL \( j \), \( h_{i,j}^t[k] \) is the channel gain from PL \( j \) to its PMs in subchannel \( k \), \( h_{j',i}^t[k] \) is the interfering channel to PL \( j \)’s members from PL \( j' \) in \( P \) functioning in whether inter (\( \theta_j^t = 0 \)) or intra-platoon (\( \theta_j^t = 1 \)) communication mode, and \( I_{j}^{t+1}[k] \) represents the total interference power. As described earlier, the PL has to maintain timely communication with the RSU to exchange the intersection safety messages. In this regard, we note \( A_{j}^t \) as the AoI of platoon \( j \in P \) up to the beginning of scheduling slot \( t \), that is, the time elapsed since the most recently successful V2I communication [35]. The AoI of platoon \( j \in P \) evolves according to

\[
A_{j}^{t+1} = \begin{cases} \Delta t, & \text{if } (1 - \theta_j^t)\beta_{j,k}^t C_{j,R}^t[k] \geq C_{\text{min}}^{j,R}, \\ A_{j}^t + \Delta t, & \text{otherwise}, \end{cases} \tag{4}
\]

where \( C_{\text{min}}^{j,R} \) is the minimum capacity requirement of V2I communication. Furthermore, \( C_{\text{min}}^{j,R} \) also denotes the minimum packet size that must be transmitted from the PL to the RSU during every time slot. It is worth mentioning that the transmission time of the packets has been taken into the AoI formulation, and whenever the transmission time exceeds the scheduling slot, i.e. \( \Delta t \), the transmission is ceased and the AoI increases; however, as (4) suggests, within every successful transmission between the RSU and PL \( j \in P \), the AoI will reset to \( \Delta t \). Accordingly, we can express the multi-objective optimization problem (MoP) for platoon \( j \) as

\[
\min_{\beta, \theta, P} \left\{ \frac{1}{T} \sum_{t=1}^T A_{j}^t, - \Pr \left\{ \sum_{t=1}^T \sum_{k \in K} \min \{ C_{j,R}^t[k] \Delta t \geq \zeta_j \} \right\} \right\},
\]

\[
\text{s.t.} \quad C_1 : C_{j,R}^t[k] \geq C_{\text{min}}^{j,R}, \quad \forall j \in P, \forall k \in K,
\]

\[
C_2 : \beta_{j,k}^t \theta_j^t \in \{0, 1\}, \quad \forall j \in P, \forall k \in K,
\]

\[
C_3 : \sum_{k \in K} \beta_{j,k}^t \leq 1, \quad \forall j \in P, \forall t \in N,
\]

\[
C_4 : p_j^t[k] \leq p_{\text{max}}^j, \quad \forall j \in P, \forall k \in K, \tag{5}
\]

where \( \zeta_j \) is the CAM message size. The objective is to minimize the expected AoI and power consumption for every platoon while maximizing the probability of CAM messages delivery rate among the PMs within every \( T \) seconds.\footnote{As stated in Section II, \( T \) must be below 100 ms according to [11].} Constraint C3 shows that each platoon can access only one subchannel in every time slot and constraint C4 is to satisfy that the transmit power of PL \( j \) remains below its maximum value \( p_{\text{max}}^j \). The optimization problem (5) contains both discrete variables, i.e. the mode selection indicator \( \theta_j^t \) and subchannel selection indicator \( \beta_j \), and continuous variable, i.e. \( p_j^t[k] \). Furthermore, the objectives are non-convex. Therefore (5) is a mixed-integer nonlinear programming (MINLP) problem. Since we have as many MINLP optimization problems as the number of platoons; it is difficult to rapidly solve them using the conventional optimization algorithms. It is critical to obtain an optimal resource allocation decision for a given dynamic environment as fast as possible. In this regard, we will investigate the state-of-the-art RL methods to handle the complexities of the proposed optimization problem.

III. MULTI-AGENT RL BASED RESOURCE ALLOCATION

In this section, we will elaborate on the multi-agent environment and its associated states, actions, and rewards, and finally, we will discuss the proposed MARL algorithm and its relevant formulations.

A. Modeling of Multi-Agent Environment

For a MARL with \( P \) agents (platoons), the optimization problems can be expressed as

\[
\max_{\pi_j} J_j(\pi_j), \quad j \in P, \quad \pi_j \in \Pi_j, \tag{6}
\]

where \( J_j(\pi_j) = \mathbb{E} [\sum_{t=0}^{\infty} \gamma^t R_{j}^{t+1}[\pi_j]^{\theta_j^t}] \), \( \pi_j \) is the policy of agent \( j \), and \( \Pi_j \) is the set of all feasible policies for agent \( j \). Each PL as an agent interacts with the vehicular network environment and takes action according to its policy, aiming at solving the optimization problem (5), or in other words, maximizing its total expected reward (6). At each time \( t \), the PL observes a state, \( s^t \),
and accordingly takes action, \( a^t \). The environment transitions to a new state \( s^{t+1} \) and PL receives a reward based on its selected action. In our proposed system model the state space \( S \), action space \( A \), and the reward function \( r^t \), are defined as follows:

- **State space**: The state observed by the PL \( j \) (agent) at time slot \( t \) consists of several parts: the instant channel information between PL \( j \) and the RSU, \( h^t_{j,\text{RSU}}[k] \), for all \( k \in K \), the channel information between PL \( j \) and its followers, \( h^t_{j,i}[k], i \in N_j \setminus \{1\} \), the previous interference from other platoons to PL \( j \), \( I^{t-1}_{j}[k] \), the AoI of PL \( j \), \( A^t_j \), the remaining intra-platoon payload (CAM message) designated to be transferred by \( T_j \), \( \zeta^t_j \), and the remaining time budget, \( T^t_j \). Hence, the state space of PL \( j \) is

\[
s^t_j = [h^t_{j,\text{RSU}}[k], h^t_{j,i}[k], I^{t-1}_{j}[k], A^t_j, \zeta^t_j, T^t_j], \quad j \in \mathcal{P}.
\]

- **Action space**: The action of each PL \( j \in \mathcal{P} \) is defined as \( a^t_j = \{\beta^t_j, \theta^t_j, p^t_j\} \). As mentioned earlier, \( \beta^t_j \) indicates which subchannel the PL \( j \in \mathcal{P} \) has selected, \( \theta^t_j \) represents the mode selection, and \( p^t_j \) represents the power control. It is noteworthy to mention that because we have applied the deep deterministic policy gradient method, the agent can select any power ranging from 0 to \( p^\text{max}_j \). This is the advantage of policy gradient methods that apply continuous actions spaces and can converge to more accurate results than conventional DQNs in which the power has to be discretized.

- **Reward function**: What makes the reinforcement learning framework fascinating is the flexibility we have in designing the reward function that drives the learning process. In our proposed learning problem, the agents receive two reward signals, a global team reward, which evaluates the agents’ cooperation, and an individual reward, which measures each agent’s performance. Accordingly, we first discuss the proposed learning algorithm and then return to the reward function’s design.

The MARL frameworks’ architecture is shown in Fig. 2, which is based on top of the MADDPG structure. In particular, we have designed two MARL frameworks, namely the Modified MADDPG, which is shown to outperform the MADDPG and other conventional RL frameworks in [40], and Modified MADDPG with task decomposition framework, which is the extension of the first algorithm, where the holistic local reward function of each agent is further decomposed into sub-reward functions and learned separately.

Unlike MADDPG, which uses a single critic to train multiple agents, the proposed framework trains two critics with the following functionalities: The centralized global critic, which is implemented at the RSU and shared between all the agents, takes the observations and actions of all the agents as input and estimates the global team reward for them. The local critic, which is specific for each agent, receives the agent’s local observation and action and estimates the local expected reward. In a sense, the goal is to simultaneously move the policy toward maximizing both global and local rewards and solve the optimization problem (5) for each agent. Furthermore, the agents do not necessarily need to know each other’s policies and take actions based on their own observations. The agents’ performance will be considered as “decent” only when they act in a way that results in a proper global team reward as well as a satisfactory individual reward for each agent.

### B. Modified MADDPG

Let \( \Theta_\pi = (W^{(1)}_\pi, \ldots, W^{(L_\pi)}_\pi) \) and \( \Phi_\psi = (W^{(1)}_\psi, \ldots, W^{(L_\psi)}_\psi) \), be the parameter space of agents’ actor and critic networks and \( \Psi_\psi = (W^{(1)}_\psi, \ldots, W^{(L_\psi)}_\psi) \) be the parameter space of the global critic, where \( L_\pi, L_\psi, L_\psi \) are the number of hidden layers in agents’ actor and critic networks and the global critic, respectively. \( W_\pi \) are the neural networks’ weight matrices and their dimensions are related to the number of nodes in the hidden layers. We consider a vehicular environment consisting of \( P \) platoons (agents) with policies \( \pi = \{\pi_1, \ldots, \pi_P\} \). The agents’ policies \( \pi_j \) and Q-functions \( Q_{\psi_j} \), and the global critic’s Q-function \( Q_{\psi_0} \) are parameterized by \( \theta_\pi \), \( \phi_\psi \), and \( \psi \), respectively, where \( \theta_\pi \in \Theta_\pi, \phi_\psi \in \Phi_\psi \), and \( \psi \in \Psi_\psi \). The MADDPG for platoon \( j \) can be written as

\[
\nabla_{\theta_j} J = \mathbb{E} \left[ \nabla_{\theta_j} \pi_j(a_j | s_j) \nabla_{a_j} Q^\pi_j(s, a_j) \bigg| a_j = \pi_j(s_j) \right].
\]
where \( s = (s_1, \ldots, s_P) \) and \( a = (a_1, \ldots, a_P) \) are the total state and action spaces. \( Q^g_j(s, a) \) is the centralized action-value function that takes the actions and states of the agents as its input to estimate Q-value for platoon \( j \). Based on the framework depicted in Fig. 2, the modified policy gradient for each agent \( j \) can be written as

\[
\nabla_\theta_j J_j = \mathbb{E}_{s, a \sim D} \left[ \nabla_\theta_j \pi_j(a_j | s_j) \nabla_a Q^g_j(s, a) \right] + \mathbb{E}_{s, a_j \sim D} \left[ \nabla_\theta_j \pi_j(a_j | s_j) \nabla_a Q^I_j(s_j, a_j) \right].
\]

(7)

where \( a_j^* = \pi_j(s_j^*) \) is the action the agent \( j \) chooses following its policy \( \pi_j \). The first term in (7) refers to the global critic which takes as input the agents' states and actions and estimates the team reward. The second term in (7) refers to each agent's local critic that unlike the global critic, only takes each agent's local state and action to estimate the agent's individual performance.

The global critic is updated as

\[
\mathcal{L}(\psi) = \mathbb{E}_{s, a, r, s' \sim \mathcal{D}} \left[ (Q^g_j(s, a) - y_g)^2 \right],
\]

(8)

where \( y_g \) is the target value and is defined as follows:

\[
y_g = r_g + \gamma \mathbb{E}_{s', a' \sim \mathcal{D}} \left[ (Q^g_j(s', a') - y_g') \right].
\]

(9)

where \( \pi' = \{\pi'_1, \ldots, \pi'_P\} \) refers to the target policies which are parameterized by \( \theta' = \{\theta'_1, \ldots, \theta'_P\} \). Similarly the local critic of agent \( j \), \( Q^I_j \), is updated by

\[
\mathcal{L}(\phi_j) = \mathbb{E}_{s, a_j, r_j, s^j} \left[ (Q^I_{\phi_j}(s_j, a_j) - y_I^j)^2 \right],
\]

(10)

and \( y_I^j \) is defined as

\[
y_I^j = r^j_I + \gamma \mathbb{E}_{s', a' \sim \mathcal{D}} \left[ (Q^I_{\phi_j}(s', a') - y_I^j) \right].
\]

(11)

Although the proposed framework can lead to decent results, there is still the problem of overestimation and suboptimal policies in Q-functions due to the function approximation errors. Motivated from the results in [41], the global critic is replaced with the Twin delayed Deterministic Policy Gradient in (7). The resulting policy gradient is

\[
\nabla_\theta_j J_j = \mathbb{E}_{s, a \sim D} \left[ \nabla_\theta_j \pi_j(a_j | s_j) \nabla_a Q^g_j(s, a) \right] + \mathbb{E}_{s, a_j \sim D} \left[ \nabla_\theta_j \pi_j(a_j | s_j) \nabla_a Q^{I_j}(s_j, a_j) \right].
\]

(12)

In (12), the twin global critics are updated as

\[
\mathcal{L}(\psi_1) = \mathbb{E}_{s, a, r, s' \sim \mathcal{D}} \left[ (Q_{\psi_1}^g(s, a) - y_g)^2 \right],
\]

(13)

where \( y_g \) is defined as follows:

\[
y_g = r_g + \gamma \min_{i=1,2} Q_{\psi_i}^g(s, a) \bigg|_{a_i = \pi_j^*(s_j^*)},
\]

(14)

and similarly, the agents' local critics are updated by (10) and (11). The modified MADDPG framework depicted in Fig. 2 is described in Algorithm 1. The core idea in TD3 is to delay the policy updates for \( d \) iterations until the convergence of value estimates. Now, we can return to the issue of designing the reward function. The Reward function must judiciously be adjusted so that the multi-agent system steps on the path of solving the optimization problem (5). In essence, each PL as an agent, tries to access the available subchannels for two reasons: i) maintain an updated connection with the RSU and keep the AoI level at its minimum, ii) disseminate the CAM information \( \zeta \) to its followers. Accordingly, we design the local reward of

---

**Algorithm 1: Modified MADDPG.**

1. Start environment simulator and generate platoons
2. Initialize main global critic networks \( Q^g_{\psi_1} \) and \( Q^g_{\psi_2} \)
3. Initialize target global critic networks \( Q^g_{\phi_1} \) and \( Q^g_{\phi_2} \)
4. Initialize each agent’s policy and critic networks
5. for each episode do
   6. Update platoons locations and respective channel gains
   7. Reset the Intra-platoon payload \( \zeta \) and maximum delivery time \( T \) to 100 ms
   8. for each timestep \( t \) do
      9. for each agent \( k \) do
         10. Observe \( s_k \) and select action \( a_k^* = \pi_k(s_k) \)
         11. \( s' = [s_1, \ldots, s_P], \ a' = [a_1, \ldots, a_P] \)
         12. Receive global and local rewards, \( r_k \) and \( r_k' \)
         13. Store \( (s', a', r_k', r_k^{\text{t}}) \) in replay buffer \( \mathcal{D} \)
      14. Sample minibatch of size \( S \), \( (s', a', r_k', r_k^{\text{t}}) \), from replay buffer \( \mathcal{D} \)
      15. Set \( y_2 = r_k + \gamma \min_{i=1,2} Q^g_{\psi_i}(s', a') \)
      16. Update global critics by minimizing the loss:
         \[ \mathcal{L}(\psi_i) = \frac{1}{S} \sum_j \left( Q^g_{\psi_i}(s', a') - y_2 \right)^2 \]
      17. Update target parameters: \( \psi_i' \leftarrow \tau \psi_i + (1 - \tau) \psi_i' \)
      18. if episode mod \( d \) then
         19. Train local critics and actors
            20. for each agent \( i \) do
               21. Update local critics by minimizing the loss:
                  \[ \mathcal{L}(\phi_i) = \frac{1}{S} \sum_j \left( Q^{I_i}(s_j, a_j) - y_I^i \right)^2 \]
               22. Update local actors:
                  \[ \nabla J_{\phi_i} = \frac{1}{S} \sum_j \left[ \nabla_{\theta_j} \pi_j(a_j | s_j) \nabla_a Q^{I_i}(s_j, a_j) + \nabla_{\theta_j} \pi_j(a_j | s_j) \nabla_a Q^g_{\psi_1}(s', a') \right] \]
               23. Update target networks parameters:
                  \[ \theta_i' \leftarrow \tau \theta_i + (1 - \tau) \theta_i' \]
                  \[ \phi_i' \leftarrow \tau \phi_i + (1 - \tau) \phi_i' \]
            24. end
         25. end
      26. end
   8. end
5. end

---
Algorithm 2: Modified MADDPG with TDec.

1. Start environment simulator and generate platoons
2. Initialize main global critic networks \( Q_{\theta_{G}} \) and \( Q_{\theta_{G}}' \)
3. Initialize target global critic networks \( Q_{\theta_{G}}^T \) and \( Q_{\theta_{G}}'^T \)
4. Initialize each agent's policy networks
5. Initialize each agent's task specific critic networks
6. for each episode do
   7. Update platoons locations and respective channel gains
   8. Reset the Intra-platoon payload \( \zeta \) and maximum delivery time \( T \) to 100 ms
   9. for each timestep \( t \) do
      10. for each agent \( k \) do
          11. Observe \( s_k \) and select action \( a_k^* = \pi_{\theta_k}(s_k) \)
          12. \( s^* = [s_1, \ldots, s_p] \), \( a^* = [a_1, \ldots, a_p] \)
          13. Receive global and local rewards, \( r_j \) and \( r_i^j \)
          14. Store \( (s_i^*, a_i^*, r_i^*, r_j^*, s_i^{t+1}) \) in replay buffer \( D \)
      15. Sample minibatch of size \( S \), \( (s^j, a^j, r_i^j, r_j^j, s^j_{t+1}) \), from replay buffer \( D \)
      16. Set \( y_i^j = r_i^j - \gamma \min \{Q_{\theta_{G}}(s^j, a^j)\} \)
      17. Update global critics by minimizing the loss:
         \[ \mathcal{L}(\phi_i) = \frac{1}{2} \sum_j \left\{ \left( Q_{\theta_{G}}(s^j, a^j) - y_i^j \right)^2 \right\} \]
      18. Update target parameters: \( \psi_i^j \leftarrow \tau \psi_i^j + (1 - \tau) \psi_i^j \)
      19. if episode mod \( d \) then
         20. Train local critics and actors
         21. for each task \( k \) do
             22. Set \( y_{i,k} = r_{i,k} + \gamma \min \{Q_{\theta_{G}}(s^j, a^j)\} \)
             23. Update local critics by minimizing the loss:
                \[ \mathcal{L}(\phi_k) = \frac{1}{2} \sum_j \left\{ \left( Q_{\theta_k}(s^j, a^j) - y_{i,k} \right)^2 \right\} \]
             24. Update local actors:
                \[ \nabla J_{\theta_k} = \frac{1}{S} \sum_j \left( \nabla_{\theta_k} \pi_i(s^j) \nabla_a Q_{\theta_k}(s^j, a^j) \right) + \frac{M}{T} \sum_k \left( \nabla_{\theta_k} \pi_i(s^j) \nabla_a Q_{\theta_k}(s^j, a^j) \right) \]
             25. Update target networks parameters:
                26. for each task \( k \) do
                   27. \( \theta_k^j \leftarrow \tau \theta_k^j + (1 - \tau) \theta_k^j \)
         28. end for
      29. end for
      30. end if
   31. end for
   32. end for

Every platoon \( j \) as

\[
 r_j^i = -\left( \kappa_1 \zeta_j^i / \zeta_j \right) - \kappa_2 A_j^i + \kappa_3 G \left( C_j^i - C_{\text{min}}^i \right) \]

\[
 - \kappa_4 \mathcal{F}(\psi_j^i), \quad (15)
\]

where \( \kappa_1 - \kappa_4 \) are weighting factors used for balancing the reward, and \( \mathcal{F}(\cdot) \) is a function that restricts the power quantity to the same range as the other components in the reward function. Furthermore, \( G(x) \) is a stepwise function given by

\[
 G(x) = \begin{cases} 
 A, & x \geq 0, \\
 0, & x < 0, 
\end{cases}
\]

where \( A > 0 \) is tuned to be a positive constant to indicate the revenue. The reward function in (15) consists of three parts that are matched with the objective function of the optimization problem (5): the first part is related to the reward the agent receives when the intra-platoon communication is chosen, the second part refers to the reward for the agent in the inter-platoon communication mode and the third part is related to the negative reward for the agent due to the power consumption. Correspondingly, we define the global reward function as

\[
 r_g^i = -\frac{1}{P} \sum_{j \in P} \sum_{k \in K} \log_{10} \{ I_j[k] \}. \quad (16)
\]

The inspiration behind choosing the global reward function to be equal to the average interference is that the platoons are derived toward choosing subchannels and power levels that impose less interference on other platoons. It is observed from Algorithm 1 that the global critic is trained more than the local actor and critic networks since we have applied the TD3 algorithm. The introduced delay, which is related to the hyperparameter \( d \), can lead to faster convergence of the system by addressing the overestimation bias of global Q-function.

The following section will discuss the multi-task MARL, its corresponding formulations, and the intuition behind devising such an algorithm.

C. Modified MADDPG With Task Decomposition

In practice, the RL agents have to perform multiple tasks, and in order to drive the policy toward maximizing these tasks simultaneously, we have to integrate these tasks into a single holistic task and design a single reward signal, as stated in (15). However, the drawback of applying such a method is that it cannot guarantee each sub-objective optimality, even though the holistic reward function may exhibit encouraging signs of convergence. In the following, we investigate the decomposition of the holistic local reward of the agents, which was earlier introduced in (15).

Taking a closer look to (15) reveals that the term related to the AoI, i.e. \( \frac{T}{T} \sum_{t=1}^T A_j^i \), and the term associated with the CAM message transmission, i.e. \( -\Pr \left\{ \sum_{t=1}^T \sum_{k \in K} \min_i \{C_j^i[k]\} \Delta t \geq \zeta_j \right\} \), are connected with \( (\theta) \). By changing this variable, the platoon will focus on minimizing the AoI or maximizing the CAM transmission probability. Consequently, it is possible to evaluate the platoon’s performance on these performance metrics separately. Care must be taken, the adopted procedure does not necessarily infer that these objectives are totally independent but rather means that the evaluation of the decomposed parts can be separate. The same reasoning cannot be applied for the objective related to power consumption (third objective in (15)), since the power quantity is influencing both the AoI and CAM message. Therefore, the local reward function which was derived for

\[
 F^i_j \]

This could have been possible if we had adopted different power allocation variables for AoI and CAM dissemination.
Modified MADDPG in (15) is decomposed into the following sub-reward functions:

- **Task. 1 reward (CAM message transmission)**
  \[
  r_{\ell}^{1,1} = - \left\{ \kappa_1 C_j^f / C_j^g \right\} - \theta_j^k \kappa_4 F\{ p_j^f \},
  \]

- **Task. 2 reward (AoI minimization)**
  \[
  r_{\ell}^{1,2} = - \kappa_2 A_j^2 + \kappa_3 G \left( C_j^f - \ell_m \right)
  - \left( 1 - \theta_j^k \right) \kappa_4 F\{ p_j^f \},
  \]

where \( \kappa_4 = \kappa_4 \) in (15). In other words we have

\[
  r_{\ell}^j = r_{\ell}^{1,1} + r_{\ell}^{1,2}, \quad \forall j \in \mathcal{P}.
\]

Therefore, for a MARL system consisting of \( M \) tasks and \( P \) agents, we change the optimization problem (6) as follows:

\[
  \max_{\pi_j} \mathcal{J}_j(\pi_j), \quad j \in \mathcal{P}, \quad \pi_j \in \Pi_j
\]

\[
  \mathcal{J}_j(\pi_j) = [\mathcal{J}_j^1(\pi_j), \ldots, \mathcal{J}_j^M(\pi_j)],
\]

where \( \mathcal{J}_j^M(\pi_j) \) is related to the agent \( j \)’s objective function for the \( M \)th task. The following Theorem provides the condition for task decomposition, which results from decomposing the holistic reward function into sub-reward functions.

**Theorem 1:** If the reward function \( R \) can be decomposed into \( M \) sub-reward functions, i.e., \( R_j(s, a, s') = \sum_{k=1}^{M} r_{j,k}^k(s, a, s') \), then the holistic objective function \( \mathcal{J}_j(\pi_j) \) can be written as \( \mathcal{J}_j(\pi_j) = \sum_{k=1}^{M} \mathcal{J}_j^k(\pi_j) \),

\[
  \mathcal{J}_j^k(\pi_j) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_{j,k}^{t+1}(s_t = s) \right], \quad k = 1, \ldots, M.
\]

**Proof:** Following the procedure applied in [42], [43], the objective function which is written as

\[
  \mathcal{J}_j(\pi_j) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_{j,k}^{t+1}(s_t = s) \right]
\]

by replacing \( R_j(s, a, s') = \sum_{k=1}^{M} r_{j,k}^k(s, a, s') \), changes as follows:

\[
  \mathcal{J}_j(\pi_j) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_{j,k}^{t+1,k}(s_t = s) \right]
\]

\[
  = \mathbb{E} \left[ \sum_{k=1}^{M} \sum_{t=0}^{\infty} \gamma^t r_{j,k}^{t+1,k}(s_t = s) \right],
\]

\[
  = \sum_{k=1}^{M} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_{j,k}^{t+1,k}(s_t = s) \right],
\]

\[
  = \sum_{k=1}^{M} \mathcal{J}_j^k(\pi_j).
\]

Based on Theorem 1, we can decompose the agents’ local critics in (12) based on the sub-tasks, and the resulting policy gradient considering the functionality of the global critic would be,

\[
  \nabla_{\pi_j} \mathcal{J}_j = \mathbb{E}_{a, a' \sim D} \left[ \nabla_{\pi_j} \pi_j(s, a) \nabla_{\pi_j} Q^{\theta_j}(s, a) + \nabla_{\pi_j} \pi_j(s, a) \right] + \nabla_{\pi_j} \pi_j(s, a).
\]

where the parameters of sub-critics for agent \( j \) are updated as

\[
  \mathcal{L}_{j,k}^{\pi_j}(\phi_{j,k}) = \mathbb{E}_{a, a', s} \left[ \left( Q_{\phi_{j,k}}^j(s, a_j) - y_{j,k} \right)^2 \right],
\]

\[
  y_{j,k} = r_{j,k} + \gamma Q_{\phi_{j,k}}^j(s', a'_j) | a'_j = \pi_j(s'_j), k = 1, \ldots, M.
\]

According to (26), the performance evaluation of the actor network for each platoon, comes from different critic networks. The global critic estimates the efficiency of the selected subchannel and power level of each agent based on (16), and the local critics estimate the performance of the platoon based on the sub-reward functions. In other words, each platoon is equipped with one actor network, and we are employing different critic networks to evaluate the overall performance of it. Again, this procedure does not necessarily mean that these subtasks are independent; since if it was the case, then we should have used different actor networks for each sub-task. However, as it is obvious from (26), each platoon operates based on a single actor network. Comparing (26) with (12) reveals that

\[
  Q^j(s_j, a_j) = \sum_{k=1}^{M} Q_{j,k}^j(s_j, a_j),
\]

which can be easily derived from Theorem 1. In other words, the decomposition of the holistic reward function leads to the decomposition of the value functions. The corresponding algorithm of modified MADDPG with task-decomposition is shown in Algorithm 2.

**IV. ALGORITHM ANALYSIS**

**A. Complexity Analysis**

The complexity analysis is crucial to the utility of the algorithms. Therefore, we analyze the computational complexity of the proposed RL methods and compare them with the conventional MADDPG framework, upon which our proposed algorithms are built. Furthermore, to give a comprehensive insight regarding the applied methods, we have also provided the complexity analysis of the learning methods we have used as baselines in Section V. However, it is of great importance to first have a profound introduction of these RL algorithms that we have investigated their numerical performance in the simulation part. The complete description of the baseline frameworks are as follows:

- **Modified MADDPG:** In this algorithm, the global critic, which is implemented at the RSU, motivates cooperation.
between the platoons by periodically reporting the effectiveness of platoons’ chosen action. The local critics and actor networks are implemented in each platoon and trained with each platoon’s local training dataset without the need for other platoons’ information.

- Federated Reinforcement Learning: In this algorithm, each platoon is equipped with a segregated actor and critic network and is trained based on its local information. Unlike the proposed algorithms that the RSU trains a separate neural network based on the local states and actions of platoons to motivate cooperation between the platoons, in FRL the RSU’s (central server) role does not involve any training process. In FRL the agents transmit the weights of their actor and critic networks to the RSU instead of their local information, i.e., states and actions. The RSU accumulates these weights, and based on a pre-set algorithm, aggregates these weights and then again sends them back to the agents. The aggregation rule which is adopted at the RSU is formulated by [44]

$$\Theta^{t+1} = \Theta^t \cdot \Omega,$$

(29)

where $\Theta^t = [\Theta_1^t, \ldots, \Theta_p^t]$ denotes the vector of all the agents’ parameters at the $t$-th learning epoch, and the $\Omega$ is defined as follows:

$$\Omega = \begin{bmatrix}
\omega & \frac{1 - \omega}{P - 1} & \ldots & \frac{1 - \omega}{P - 1} \\
\frac{1 - \omega}{P - 1} & \omega & \ldots & \frac{1 - \omega}{P - 1} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{1 - \omega}{P - 1} & \frac{1 - \omega}{P - 1} & \ldots & \omega
\end{bmatrix}.$$  

(30)

Under the proposed aggregation scheme, each agent preserves its parameters with weight $\omega$ and mixes the other agents parameters with weight ($\frac{1 - \omega}{P - 1}$).

- Fully decentralized MADDPG: To illustrate the global critic’s impact on the network performance, in this algorithm, the RSU’s role is not taken into account, and the platoons choose their actions in a fully decentralized way, based on their observations.

- DDPG: In this algorithm, the RSU has to acquire all the platoons’ observations and actions and is considered a fully centralized algorithm in which all the computations and decision-making have to be performed in the RSU [45].

In the following, we have provided the complexity analysis of the introduced algorithms. In essence, this analysis depends on four parameters, i) the number of trainable parameters, ii) the total number of neural networks used in the algorithms, iii) the computational complexity, and iv) the communication overhead between the agents and the central server or the RSU. This inclusive overview can give us an authentic insight into the applicability and the scalability of these algorithms. Furthermore, since the proposed algorithms are based on MADDPG, we have also included its complexity analysis in the following evaluations.

i) The number of trainable parameters: In MADDPG, the number of trainable parameters for MADDPG would be $O(P^2(\omega + \alpha))$, where $P$ indicates the number of agents. Conversely, the two proposed RL methods incorporate two types of critic networks: the global and local critics. Both the algorithms share a global centralized Q-function whose parametric space increases linearly and is represented as $O(P(\omega + \alpha))$. On the other hand, the local critics in the two RL methods only take the respective agent’s observation and action as their input. Consequently, the parameter space of local critics can be expressed as $O(\omega + \alpha)$, and this is similar for both the algorithms. Similarly, in FRL, and Fully decentralized MADDPG, the parameter space of the agents’ local critics is denoted by $O(\omega + \alpha)$ as the agents are solely operate based on their own observations. Finally, in DDPG, since there is only one actor and critic network, both of them have to take all the states and actions of the agents into account; therefore, the parameter space of this method would be $O(P(\omega + \alpha))$.

ii) The total number of neural networks: In MADDPG, the total number of neural networks used during the training process is equal to $2 \times (P(1_\alpha + 1_\ell))$, where the multiplication by 2 is because of the target networks, $1_\alpha$ and $1_\ell$ represent that there is one critic and actor network specific for each agent, and $P$ is the total number of agents. For the modified MADDPG framework, the total number of neural networks is $2 \times (P(1_\alpha + 1_\ell) + 1_\omega)$, where $1_\omega$ indicates the total number of global critics. It is worth mentioning that applying the TD3 algorithm doubles the number of global critics, and in this case the number of neural networks will be $2 \times (P(1_\alpha + 1_\ell) + 2_\omega)$. Finally, for the modified MADDPG method with task decomposition, number of neural networks will be $2 \times (P(1_\alpha + 1_\ell) + 2_\omega)$, where $2_\omega$ indicates that there is a separate Q-function for each agent’s decomposed tasks. Similarly, this number will be $2 \times (P(1_\alpha + 1_\ell) + 2_\omega)$, whenever the TD3 algorithm is further applied. In a similar procedure, for FRL, and the Fully decentralized MADDPG, the total number of neural networks will be $2 \times (1_\omega + 1_\ell)$. Finally, for DDPG this number will be $2 \times (1_\omega + 1_\ell)$.

iii) The computational complexity: In order to put our analytically into mathematical expressions, let us first assume that $\Gamma_a^i$ and $\Gamma_c^i$ denote the number of the neurons in the $i$-th layer of the actor and critic networks, respectively. Since both the actor and critic networks are fully connected, their computational complexity can be written as $O(\sum_{i=1}^{L_a-1}(\Gamma_a^i \cdot \Gamma_a^{i+1}))$ and $O(\sum_{i=2}^{L_c-1}(\Gamma_c^i \cdot \Gamma_c^{i+1} + \Gamma_c^i \cdot \Gamma_c^{i+1}))$, respectively, where $L_a$ and $L_c$ are the total number of layers in the actor and critic networks. Hence, the complexity of the mentioned frameworks will be as follows:

- MADDPG: $O(P \cdot N_b \cdot E \cdot I \cdot (\mathcal{C}_a^g + \mathcal{C}_c^g))$

- Modified MADDPG with TDec. ($k$ tasks): $O \left( N_b \cdot E \cdot I \cdot (\mathcal{C}_a^g + \mathcal{C}_c^g) + O \left( P \cdot N_b \cdot E \cdot I \cdot (\mathcal{C}_a^g + k \cdot \mathcal{C}_c^g) \right) \right)$.

- Modified MADDPG: $^4$ $O \left( N_b \cdot E \cdot I \cdot (\mathcal{C}_a^g + \mathcal{C}_c^g) + O \left( P \cdot N_b \cdot E \cdot I \cdot (\mathcal{C}_a^g + \mathcal{C}_c^g) \right) \right)$

$^4$ Indexes $g$ and $\ell$ are related to the global and local neural networks, and $k$ is the number of tasks.
• Federated Reinforcement Learning:
  \[ O(P \cdot N_h \cdot E \cdot I \cdot (C^a_i + C^c_i)) \]

• Fully Decentralized MADDPG: \[ O(P \cdot N_h \cdot E \cdot I \cdot (C^a_i + C^c_i)) \]

• DDPG: \[ O(N_h \cdot E \cdot I \cdot (C^a_i + C^c_i)) \]
  where \[ C^a_i = \sum_{i=2}^{L_o-1} (\Gamma^a_{i-1} \Gamma^a_i + \Gamma^a_{i+1}), \quad C^c_i = \sum_{i=2}^{L_o-1} (\Gamma^c_{i-1} \Gamma^c_i + \Gamma^c_{i+1}) \]
  and \( P \) is the number of platoons (agents), \( N_h \) is the mini-batch sampling size, \( E \) is the number of episodes, and \( I \) is the max training steps of each episode.

iv) The communication overhead: Communication overhead is one of those performance metrics that is often overlooked; especially in communication systems that are established based on MARL frameworks. The reason is obvious; most of the MARL frameworks are built upon agents’ communications. This exchange of data is necessary to stabilize the learning process and prompt the agents into cooperative behavior. Nonetheless, it is crucial to lowering the overhead as much as possible. In the following, we have provided the total communication overhead for the proposed frameworks. In this analysis, we have focused on the fact that how many times the agents have to interact with other agents and the RSU during the learning process.

In MADDPG, the overhead is equal to \( P(P-1) \), while it reduces to \( P \) for our proposed algorithms, FRL, and DDPG. The Fully decentralized MADDPG does not impose any overhead on the network side and this is due to the fact the agents act independently in this framework.

B. Convergence Analysis

Unlike the Q-learning algorithm that has a straightforward convergence analysis [46], for the policy-based RL algorithms, and especially their multi-agent extension, e.g. our proposed algorithms, the convergence is hard to prove, as multiple agents are interacting with the environment simultaneously. Therefore, we assess the convergence of the proposed algorithms through simulations in Section V.

V. PERFORMANCE EVALUATION

In this section, we assess the simulation results to validate the proposed multi-agent RL based resource allocation for the platooning system. We have built our simulation following the urban case defined in Annex A of [8]. Major simulation parameters, including the channel models for V2I and V2V links, are listed in Table II. In addition, the Gaussian noise \( \epsilon \sim N(0, 0.2) \) is added to the actions chosen by the target actor networks, and then clipped to \((-0.5, 0.5)\) to smooth the target policy, and the policy update delay factor is set to \( \delta = 2 \). Throughout the simulations, the number of available RBs is fixed to three; however, we have varied the number of platoons, the number of PMs, and the intra-platoon spacing to investigate their impact on the system’s overall performance. It is worthwhile to mention that we fix the large-scale fading during each episode and let the small-scale fading alter; therefore, the RL algorithm can better procure the underlying fading dynamics. Due to the sensitivity of RL algorithms to the reward function design, the global reward function in (16) is normalized to be consistent with the local reward’s range.

A. Simulation Results

Fig. 3 indicates the convergence of agents sub-tasks when the intra-platoon gap is 25 m, and the number of platoon members at each platoon is 6 (30 vehicles in total). For each agent, we have plotted its sub-tasks reward function. Two notable trends stand out in the figure; first, it can be seen that all the agents have been able to fulfill their associated tasks and maximize the designated reward functions in (17) and (18) during the \( T \) seconds. Second, the proposed algorithm is quite fast in convergence time. It is observed that for most of the agents, the task-wise reward functions converge in less than 50 episodes. In addition to some fluctuations due to the channel fading that arose by platoons’ movements in the environment, the following observations can also be noticed. Since the number of vehicles is large compared to the available resources, there is high contention between the platoons in terms of accessing the available resources. Therefore, the platoons have to share the resources. However, they

### Table II: Simulation Parameters

| Vehicular environment parameters          | Value |
|------------------------------------------|-------|
| Carrier frequency                        | 2 GHz |
| Number of RBs                            | 3     |
| Bandwidth of each RB                     | 180 kHz|
| Number of Vehicles                       | 16 – 50|
| Size of Platoons                         | 4 – 10 |
| Platoons Speed                           | 36 – 54 km/h|
| Intra-platoon gap                        | 5, 15, 25, 35 m|
| RSU and vehicles antenna heights         | 25 m  |
| RSU and vehicles antenna gains           | 8, 3 dB |
| RSU and receiver noise figure            | 5, 9 dB|
| Vehicles mobility model                  | Urban case of A.1.2 [8]|
| Vehicles maximum power                   | 30 dBm|
| Noise power \( \sigma^2 \)               | -114 dBm|
| Time constraint of CAM dissemination, \( T \) | 100 ms|
| CAM message size                         | 4000 bytes|
| V2I links\(^1\) minimum capacity requirement, \( C^\text{I}_\text{min} \) | 3 bps/Hz [47]|
| V2I links path loss model               | 128.1 + 37.6 log_10 (d) |
| V2V links\(^2\) path loss model         | LOS in WINNER+ B1 Manhattan [48]|
| Shadowing distribution                   | Log-normal|
| Shadowing standard deviation for V2I links| 8 dB |
| Shadowing standard deviation for V2V links| 3 dB |
| Decorrelation distance for V2I/V2V links | 50, 10 m|
| Pathloss/shadowing update for V2I/V2V links | Every 100 ms [8]|
| Fast fading update for V2I/V2V links     | Every 1 ms [8]|
| Fast fading                             | Rayleigh fading\(^3\) |

\(^1\)Link between PL and RSU. \(^2\)Link between PL and its followers. \(^3\)It is worth noting that when there exist line-of-sight communications between vehicles and RSUs, the LoS fading model is preferred, e.g., the Rician fading model; however, we have used Rayleigh fading model for both LoS and NLoS scenarios for simplicity.
have to control their power usage jointly with the mode they choose to operate so as not to impose much interference to the other platoons reusing the same resources. This issue is of paramount importance as the platoons choose their actions based on their own observations. The figure implicitly indicates that different components of the system have somehow reached an equilibrium. In other words, not only the global critic has been able to drive the platoons toward selecting proper resources to impose less interference on each other, but also the local critics have motivated their respective platoons to flexibly alter their decisions between inter and intra-platoon modes and meet the predetermined requirements.

Fig. 3 also reveals that the number of episodes agent three and agent five needed for proper convergence is longer compared to the other agents. Starting with agent three, it is observed that agent five has focused only on one of its tasks leading to an increase in one task’s reward and a substantial decrease in the other one. These fluctuations are demonstrated with red and black arrows for agent five’s task-wise reward functions. In general, the proposed MARL method has robust functionality, and yields compelling results even in complex environments consisting of even more vehicles.

Fig. 4 compares the convergence of the five approaches in terms of the average reward performance when the number of platoons is five and seven, respectively. At first glance, the proposed methods indisputably outperform the other three baselines. The DDPG method has the worst reward performance among the considered RL algorithms in both figures. The reason for this weak execution can be related to the DDPG’s centralized behavior. Since the DDPG has to take all the agents’ observations and actions as input and evaluate how decent the policy has performed for all the agents, it fails to address the agents’ individual performance and acts non-stationarily in multi-agent environments. This improper execution is further intensified with the number of agents in Fig. 4(b).

Regarding the fully decentralized MADDPG, the agents act absolutely oblivious without any knowledge about other existing agents’ policies or actions in the environment. This unawareness can lead to increased levels of interference in the system, which will degrade the agents’ overall performance. This phenomenon is not very severe when the number of platoons is low, as can be seen from Fig. 4(a); however, by increasing the number of platoons, its tendency even to perform worse than DDPG is not inconceivable, as observed from Fig. 4(b). Federated reinforcement learning is built on top of the fully decentralized MADDPG; however, in this framework, the central server periodically collects the agents’ weights and after performing a pre-planned aggregation, like the one defined in (29), sends the weights back to the agents afterward. This aggregation can help the agents towards cooperative behavior, and that is why its performance is superior to the fully decentralized MADDPG. One prominent feature that separates our proposed RL frameworks from the other baselines, aside from their better reward performance and faster convergence, is their stability and minimal fluctuations during the convergence. We can summarize the primary reasons for this performance gap as follows: i) The proposed frameworks can learn to maximize the individual and global rewards for all the agents simultaneously, leading to improved collaboration between the agents, hence driving towards better performance. ii) The global critic, which is based on TD3, considers the correspondence between function approximation error in both policy and value updates. On the other hand, the DDPG method is highly susceptible to inaccuracies provoked by function approximation errors, making it overfit to narrow peaks in the value estimate. iii) Last but not least, unlike the original implementation of DDPG, which leverages the correlated Ornstein-Uhlenbeck noise, the proposed MARL framework applies an uncorrelated Gaussian noise for exploration. Eventually, by analyzing Figs. 4(a) and 4(b), it is unveiled that the proposed RL algorithms tend to converge to the same quantity even though the vehicle density has increased in the environment, while the other baselines’ performance diminishes.
with the increased load. From the evidence provided by the figures, we can infer that neither a fully centralized approach, i.e., DDPG, nor a fully decentralized one, i.e., fully decentralized MADDPG is a suitable approach to be taken for the proposed scenario. In contrast, our proposed methods along with the federated learning yield promising results. Not to mention that compared to our proposed RL frameworks, federated learning converges to a lower value. To understand the logic behind this issue, it is better to take a closer look at the building blocks of each framework. In our proposed methods the central server (RSU) trains a separate neural network from the agents and participates in the learning process by evaluating the total interference levels between the agents. However, in federated learning, the central server’s role is only limited to aggregating the agents’ parameters. This aggregation has its drawbacks. Any untrained agent can aggravate the performance of the other agents since the agents’ parameters are aggregated with each other. Hence, federated learning suffers from slow convergence rates.

Fig. 5 illustrates the mean AoI of platoons as a function of the intra-platoon gap when P = 5 and N = 4. From the figure, it can be observed that the AoI quantity rises for all the considered algorithms as the intra-platoon spacing increases. The intuition behind this observation is straightforward. By increasing the intra-platoon gap, it is perceptible that the channel conditions from PLs to their followers sustain more variations, leading to lower data rates. Accordingly, the PL spends more time transmitting the CAM message to its followers and operating in Mode 1. In the meantime, the PL less frequently interacts with the RSU; therefore, the average AoI increases. Nonetheless, our proposed MARL frameworks perform significantly more reliable than the other baselines, maintaining the average AoI quantity within 5-10 milliseconds range, and guarantees better QoS. Stunningly, the proposed frameworks act close to each other. This behavior is anticipated as both the algorithms leverage the global and local critics simultaneously to learn a global and individual reward. However, there is still a slight performance gap between them due to the task decomposition in our second algorithm. In comparison, the DDPG acts less stable, and its performance degrades by increasing the intra-platoon spacing.

It is also observed that the performance of fully decentralized MADDPG is close to our proposed algorithms up to 25 meters intra-platoon gap; however, there can be seen a sharp jump in the AoI quantity when the intra-platoon gap rises to 35 meters. This is because, with longer distances between the PL and its followers, the PLs tend to use more power to compensate for the reduced levels of channel gains to guarantee the CAM message transmission to their followers, which inevitably results in severe interference to other platoons, and as these platoons are acting in a fully decentralized way, they cannot discern the appropriate resources to select, hence leading to these sharp changes in the performance metrics. However, the former behavior is relieved in federated learning since the platoons can to some extent control their power and interference levels through the knowledge they have gained as a result of aggregation performed by the RSU. The aforementioned analysis is also extendible to results in Fig. 6(a), which demonstrates the average AoI versus the number of platoon followers.

Another compelling result can be observed from Figs. 5(b) and 6(b), which show the CAM message transmission probability. From the figures, the performance metric drops for all the schemes as the intra-platoon gap increases. In conjunction with the observations from Figs. 5(a) and 6(a), the intuition behind this phenomenon is explicit. However, as Figs. 5(b) and 6(b) suggest the proposed framework is robust against alterations in platoon sizes or intra-platoon spacing variations. The proposed framework maintains a transmission probability of over 99 percent for different platoon sizes when the intra-platoon gap is less than 25 meters, whereas this metric drops significantly for DDPG and fully decentralized MADDPG. We finalize the respective analysis with a critical look at all four figures. By comparing Figs. 5(a)–6(b), it is conceivable that the number of vehicles significantly impacts the performance metrics quantity. In Fig. 6(b), by increasing the number of vehicles up to 30, except DDPG, all the algorithms have shown a similar behavior. As we continue increasing the number of vehicles up to 50, the gap between these algorithms starts to grow. One interesting observation from this figure is that the CAM message transmission probability has dropped to 65 percent for fully decentralized MADDPG, even worse than DDPG. From Fig. 6(b), it is inferred...
that when the number of platoon members is 4 or 6, the fully decentralized MADDPG performs almost close to our proposed methods. This is due to the fact that although the platoons in this method act in a fully decentralized way; however, since the number of platoon members is not that large, this method acts close to the proposed methods. In other words, choosing the best resources is not that critical when the number of platoon members is not very large, and the platoons are most likely capable of performing effectively. In the meantime, the DDPG framework, as mentioned earlier, surely lacks other certain standards. The weak performance stems from its weak convergence and as it is apparent from the figure, its performance is the worst. However, by increasing the number of platoon members to 8 and then to 10, the importance of choosing the best resources gets prominent and as it is shown in the figure, the performance of Fully decentralized MADDPG falls even below DDPG, which directly relates to its lack of interference management when the number of vehicles is considerable.

To better underline the validity of our proposed methods, we have also adopted two heuristic algorithms based on greedy and exhaustive search as baselines. For both of these algorithms, we adopt a common framework to determine the resource allocation binary variable (β) which is proposed in [49]. This framework works on the basis of distributing the users on different channels based on their interference levels. This algorithm tries to divide strongly interfering cellular links into different sets in order to reduce mutual interference. For the decision-making (θ) and power allocation (p), we adopt two approaches,

1) The platoons process the tasks separately, i.e. they send the CAM message first and then switch to the AoI minimization part. For the power allocation, we assume that the platoons send at the maximum available power. This analysis is of most interest as it provides a clear understanding of how the proposed algorithms can handle the interference between the platoons by properly allocating an appropriate amount of power.

2) Both the decision-making and power allocation are handled randomly, hoping to find the best order of allocation.
As is shown in Fig. 7, it can be seen that there is a considerable gap between our proposed algorithms and the greedy scheme. In addition, the performance gap between the exhaustive search and our proposed methods (especially the Modified MADDPG with Task Dec.) is very close. Not to mention that the complexity of an exhaustive search is very high.

Another interesting result can be seen in Fig. 8. We have shown the behavior of one platoon during the 100 ms time budget to see how frequently it is jumping between two tasks, and also how much power it consumes during the different time slots. The blue curves show the transmitted power of the platoon leader during Task 1 (CAM dissemination, solid line), and Task 2 (AoI minimization, dashed line). The orange one is related to the remaining CAM message after each transmission. During the initial time slots, the platoon leader only focused on disseminating the CAM messages with full power. Taking a closer look at the remaining CAM messages (orange curve) shows that the PL has managed to disseminate almost all of this packet in 20 ms. To reduce the AoI, the platoon leader has followed unpredictable behavior. Based on the figure, the platoon leader has communicated with RSU only when it is necessary, meaning that it has kept its communication with the RSU to the level that it can guarantee a fair level of AoI.

Overall, what drives the learning procedure in RL is the reward function design, which is done by translating the objective functions into a proper reward function. Extra flexibility can be added to how an RL agent sees an optimization problem by decomposing the reward function. What we can infer from the figures is that with task decomposition, the RL agents can infer the underlying architecture of the learning environment faster. Nevertheless, care must be taken since these observations are based on the particular setting for the simulation, and additional caution is required when generalizing them. We can still conclude that our proposed frameworks, especially the one proposed in Algorithm 2 indicated a very robust behavior against the parameter modifications and outperformed the other baselines.

VI. CONCLUSION

In this paper, novel MADDPG-based resource allocation methods, named Modified MADDPG, and Modified MADDPG with task decomposition were developed for a platooning system, aiming at minimizing the AoI of platoons while guaranteeing the CAM message delivery to PMs. The proposed MARL frameworks consist of a collaborative setting where a group of PLs simultaneously learn to maximize the collective global reward and individual local reward. Furthermore, in the second algorithm, we decomposed the agents’ holistic reward signal into multiple sub-reward functions based on their sub-tasks and evaluated them separately. The following interesting results were observed from the simulations. First, the proposed algorithms were shown to outperform the other benchmark algorithms in terms of AoI and CAM message transmission probability. Second, we demonstrate that the proposed RRM schemes were robust and effective in encouraging platoons to improve system-level performance, although the PLs independently select their transmission mode, RB, and power levels. Finally, decomposing the holistic reward function of the agents led to the faster convergence rate. Future work will carry an in-depth extension of the proposed framework to Non-orthogonal Multiple Access (NOMA) and Multiple-Input and Multiple-Output (MIMO) scenarios for the platooning system. Also, examining the spectrum sharing scenarios in vehicular networks is another encouraging direction worth further investigation.

REFERENCES

[1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, “Data-driven intelligent transportation systems: A survey,” IEEE Trans. Intell. Transp. Syst., vol. 12, no. 4, pp. 1624–1639, Dec. 2011.
[2] R. Hall and C. Chin, “Vehicle sorting for platoon formation: Impacts on highway entry and throughput,” Transp. Res. Part C: Emerg. Technol., vol. 13, no. 5-6, pp. 405–420, Oct. 2005.
[3] J. Lioris, R. Pedarsani, F. Y. Tascikaraoglu, and P. Varaiya, “Platoons of connected vehicles can double throughput in urban roads,” Transp. Res. Part C: Emerg. Technol., vol. 77, pp. 292–305, Apr. 2017.
[4] D. Jia, K. Lu, J. Wang, X. Zhang, and X. Shen, “A survey on platoon-based vehicular cyber-physical systems,” IEEE Commun. Surv. Tut., vol. 18, no. 1, pp. 263–284, Jan.-Mar. 2016.

[5] Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service, European Standard (EN) 302 637-2, European Telecommunications Standards Institute, Sophia Antipolis, France, Sep. 2014, version 1.3.2.

[6] 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 Trans. Intell. Transp. Syst., vol. 18, no. 5, pp. 1066–1077, May 2016.

[7] G. Nardini, A. Virdis, C. Campolo, A. Molinaro, and G. Stea, “Cellular V2X communications for platooning: Design and evaluation,” Sensors, vol. 18, no. 5, May 2018, Art. no. 1527.

[8] 3rd Generation Partnership Project (3GPP), “Study on LTE-based V2X services,” 3GPP, Sophia Antipolis, France, Tech. Specification 36. 885, Jun. 2016, version 14.0.0.

[9] 3rd Generation Partnership Project (3GPP), “Proximity-based services (ProSe),” 3GPP, Sophia Antipolis, France, Tech. Specification 23. 033, Jul. 2020, version 16.0.0.

[10] 3rd Generation Partnership Project (3GPP), “Vehicle to vehicle (V2V) services based on LTE sidelink,” 3GPP, Sophia Antipolis, France, Tech. Specification 36. 785, Oct. 2016, version 14.6.0.

[11] 3rd Generation Partnership Project (3GPP), “Study on enhancement of 3GPP support for 5G V2X services,” 3GPP, Sophia Antipolis, France, Tech. Rep. 22. 886, Dec. 2018, version 16.2.0.

[12] 3rd Generation Partnership Project (3GPP), “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures,” 3GPP, Sophia Antipolis, France, Tech. Specification 36. 213, Oct. 2018, version 14.8.0.

[13] C. Campolo, A. Molinaro, G. Araniti, and A. O. Berthet, “Better platooning control toward autonomous driving: An LTE device-to-device communications strategy that meets ultralow latency requirements,” IEEE Veh. Technol. Mag., vol. 12, no. 1, pp. 30–38, Mar. 2017.

[14] J. Mei, K. Zheng, L. Zhao, L. Lei, and X. Wang, “Joint radio resource allocation and control for vehicle platooning in LTE-V2V network,” IEEE Trans. Veh. Technol., vol. 67, no. 12, pp. 12238–12250, Dec. 2018.

[15] P. Wang, B. Di, H. Zhang, K. Bian, and L. Song, “Platoon cooperation in cellular V2X networks for 5G and beyond,” IEEE Trans. Wireless Commun., vol. 18, no. 8, pp. 3919–3932, Aug. 2019.

[16] T. Zeng, O. Semieri, W. Saad, and M. Bennis, “Joint communication and control for wireless autonomous vehicular platoon systems,” IEEE Trans. Commun., vol. 67, no. 11, pp. 7907–7922, Nov. 2019.

[17] H. Peng et al., “Resource allocation for cellular-based inter-vehicle communications in autonomous platoons,” IEEE Trans. Veh. Technol., vol. 66, no. 12, pp. 11249–11263, Dec. 2017.

[18] R. Wang, J. Wu, and J. Yan, “Resource allocation for D2D-enabled communications in vehicle platooning,” IEEE Access, vol. 6, pp. 50526–50537, 2018.

[19] K. Zia, N. Javed, M. N. Sial, S. Ahmed, A. A. Pirzada, and F. Pervez, “A distributed multi-agent RL-based autonomous platooning scheme for multi-access edge computing,” in 2020 5th Annual IEEE China Council Conference (ICCC), 2020, pp. 1–5.

[20] Z. Li and C. Guo, “Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications,” IEEE Trans. Veh. Technol., vol. 69, no. 2, pp. 1828–1840, Feb. 2020.

[21] J. Tian, Q. Liu, H. Zhang, and D. Wu, “Multi-agent deep reinforcement learning based resource allocation for heterogeneous QoS guarantees for vehicular networks,” IEEE Internet Things J., vol. 9, no. 3, pp. 1683–1695, Feb. 2022.

[22] H. Peng and X. Shen, “Multi-agent reinforcement learning based resource management in MEC-and UAV-assisted vehicular networks,” IEEE J. Sel. Areas Commun., vol. 39, no. 1, pp. 131–141, Jan. 2021.

[23] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, “Minimizing age of information in vehicular networks,” in Proc. IEEE Commun. Soc. Conf. Sensor Mesh Ad Hoc Commun. Netw., 2011, pp. 350–358.

[24] M. K. Abdel-Azziz, C. F. Liu, S. Samarakoona, M. Benni, and W. Saad, “Ultra-reliable low-latency vehicular networks: Taming the age of information tail,” in Proc. IEEE Glob. Commun. Conf., 2018, pp. 1–7.

[25] X. Chen et al., “Age of information aware radio resource management in vehicular networks: A proactive deep reinforcement learning perspective,” IEEE Trans. Wireless Commun., vol. 19, no. 4, pp. 2268–2281, Apr. 2020.

[26] R. Molina-Masegosa and J. Gozalvez, “LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications,” IEEE Veh. Technol. Mag., vol. 12, no. 4, pp. 30–39, Dec. 2017.

[27] S. Chen, J. Hu, Y. Shi, and L. Zhao, “LTE-V: A TD-LTE-based V2X solution for future vehicular network,” IEEE Internet Things J., vol. 3, no. 6, pp. 997–1005, Dec. 2016.

[28] H. U. Sheikh and L. Bölöni, “Multi-agent reinforcement learning for problems with combined individual and team reward,” in Proc. IEEE Int. Joint Conf. Neural Netw., 2020, pp. 1–8.

[29] S. Fujimoto, H. van Hoof, and D. Meer, “Addressing function approximation error in actor-critic methods,” in Proc. 35th Int. Conf. Mach. Learn., 2018, pp. 1588–1597.

[30] H. Van Seijen, M. Fatemi, J. Romoff, R. Laroche, T. Barnes, and J. Tsang, “Hybrid reward architecture for reinforcement learning,” in Proc. Ann. Conf. Neural Inf. Process. Syst., 2017, pp. 5396–5406.

[31] C. Sun, W. Liu, and L. Dong, “Reinforcement learning with task decomposition for cooperative multiagent systems,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 5, pp. 2894–2895, 2020.

[32] Z. Zha, S. Wan, P. Fan, and K. B. Letaief, “Federated multi-agent actor-critic learning for age sensitive mobile edge computing,” IEEE Internet Things J., vol. 9, no. 2, pp. 1053–1067, Jan. 2022.

[33] H. Peng and X. Shen, “Deep reinforcement learning based resource management for multi-access edge computing in vehicular networks,” IEEE Trans. Netw. Sci. Eng., vol. 7, no. 4, pp. 2416–2428, Oct.-Dec. 2020.

[34] C. Watkins and P. Dayan, “Q-learning,” Mach. Learn., vol. 8, no. 3-4, pp. 279–292, 1992.

[35] 3rd Generation Partnership Project (3GPP), “Study on evaluation methodology of new vehicle-to-everything V2X use cases for LTE and NR (Release 16),” 3GPP, Sophia Antipolis, France, Tech. Specification (TS) 37. 985, Jun. 2018, version 15.3.0.

[36] Y. J. Bultitude and T. Rautiainen, “IST-4-027756 WINNER II DI 1.2 V1.2 WINNER II channel models,” EBITG, TUL, UOUUL, CUC/RCC, NOKIA, Tech. Rep, 2007. [Online]. Available: https://www.cept.org/files/8339/winner2%20-%20final%20report.pdf

[37] L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, “Graph-based resource sharing in vehicular communication,” IEEE Trans. Wireless Commun., vol. 17, no. 7, pp. 4579–4592, Jul. 2018.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Mohammad Parvini (Graduate Student Member, IEEE) received the B.Sc. degree in electrical engineering from the Amirkabir University of Technology, Tehran, Iran, in 2019, and the M.Sc. degree in communication systems from Tarbiat Modares University, Tehran, in 2021. He is currently working toward the Ph.D. degree in communication systems with the Technical University of Dresden, Dresden, Germany. His research interests include wireless communication systems, optimization, and reinforcement learning.

Mohammad Reza Javan (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, in 2003, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2006, and the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, in 2013. He is currently with the Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran. His research interests include the design and analysis of wireless communication networks with emphasis on the application of optimization theory and machine learning methods.

Nader Mokari (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 2014. In October 2015, he joined the Department of Electrical and Computer Engineering, Tarbiat Modares University as an Assistant Professor. He has been elected as an IEEE exemplary Reviewer in 2016 by IEEE Communications Society. He is currently an Associated Professor with the Department of Electrical and Computer Engineering, Tarbiat Modares University. His research interests include many aspects of wireless technologies with a special emphasis on wireless networks. Dr. Mokari is on the Editorial board of the IEEE TCOM. In recent years, his research has been funded by Iranian Mobile Telecommunication Companies, Iranian National Science Foundation. He was the recipient of the Best Paper Award at ITU K-2020, and also IEEE outstanding Ph.D. thesis award. He was also involved in a number of large scale network design and consulting projects in the telecom industry.

Bijan Abbasi (Senior Member, IEEE) received the B.Sc. degree from the Shiraz University, Shiraz, Iran, in 1995, and the M.S. and Ph.D. degrees in telecommunications engineering from Tarbiat Modares University, Tehran, Iran, in 1997 and 2003, respectively. From 2003 to 2005, he was a Researcher of electromagnetic propagation Department, Iran, Telecommunication Research Center. In 2005, he joined the Satellite communication laboratory, Tarbiat Modares University as a Postdoctoral Researcher. Since 2010, he has been an Assistant Professor with the Faculty of Electrical and Computer Engineering, Tarbiat Modares University. He has also more than 35 papers and publications in journals and conferences.

Eduard A. Jorswieck (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from TU Berlin, Berlin, Germany, in 2004. He is currently the Managing Director with the Institute of Communications Technology and the Head of the Chair for Communications Systems and a Full Professor with the Technische Universität Braunschweig, Braunschweig, Germany. From 2008 to 2019, he held the Chair of Communication Theory at TU Dresden, Dresden, Germany. He has authored or coauthored more than 160 journal articles, 15 book chapters, one book, three monographs, and some 300 conference papers. His main research interests include the broad area of communications. He was the recipient of the IEEE Signal Processing Society Best Paper Award. He and his colleagues were also recipients of the Best Paper and Best Student Paper Awards at the IEEE CAMSAP 2011, IEEE WCSP 2012, IEEE SPAWC 2012, IEEE ICUPN 2018, PETS 2019, and ISWCS 2019. Since 2017, he has been the Editor-in-Chief of the EURASIP Journal on Wireless Communications and Networking. Since 2022, he has been on the Editorial board of the IEEE TRANSACTIONS ON COMMUNICATIONS. He was on the Editorial boards of the IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.