Received 14 January 2024; revised 19 February 2024; accepted 12 March 2024. Date of publication 21 March 2024; date of current version 9 April 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3380509

Resource Allocation in V2X Networks: From Classical Optimization to Machine Learning-Based Solutions

MOHAMMAD PARVINI (Member, IEEE), PHILIPP SCHULZ, AND GERHARD FETTWEIS (Fellow, IEEE)

Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, 01062 Dresden, Germany

CORRESPONDING AUTHOR: M. PARVINI (e-mail: mohammad.parvini@tu-dresden.de)

This work was supported in part by the Federal Ministry of Education and Research (BMBF) as part of the Project 6G-ANNA under Grant 16KISK103.

ABSTRACT As one of the promising intelligent transportation frameworks, vehicular platooning has the potential to bring about sustainable and efficient mobility solutions. One of the challenges in the development of platooning is maintaining the string stability, which ensures that there is no amplification of the signal of interest along the platoon chain. String stability is dependent on reliable inter-vehicle communications and proper controller design. Therefore, in this paper, we formulate radio resource management (RRM) problem with the purpose of satisfying the reliability of the vehicle-to-vehicle (V2V) links and string stability of the platoon. We tackle the optimization problem from different angles. First, we devise centralized classical approaches based on difference of two convex functions (d.c.) programming, in which we assume the base station (BS) has full knowledge over the V2V channel gains. In the second strategy, we develop decentralized resource allocation approaches based on multi-agent reinforcement learning (MARL). In essence, we model each transmitter vehicle in the platoon as an autonomous agent that tries to find an optimal policy according to its local estimated information to maximize the total expected reward. We also investigate whether the integration of federated learning (FL) with decentralized MARL algorithms can bring any potential benefits. This comparison between classical and machine learning (ML)-based RRM strategies helps us make crucial observations in terms of robustness, sensitivity, and efficacy of the policies that are learned by reinforcement learning (RL)-based resource allocation algorithms.

INDEX TERMS Difference of two convex functions (d.c.) programming, optimization, multi-agent reinforcement learning (MARL), platooning, radio resource management (RRM), federated learning (FL).

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

THE emergence of cellular vehicle-to-everything (V2X) communication technologies can be considered as one of the crucial leaps to facilitate intelligent transportation systems and autonomous driving [1]. Essentially, V2X includes vehicle-to-infrastructure (V2I) communications, which aim to connect vehicles to base stations (BSs), and vehicle-to-vehicle (V2V) communications, which enable direct data transmission between vehicles. Entertainment-related applications, which typically require high throughput, rely primarily on V2I links, while low-latency services that require regular dissemination of safety-critical messages usually rely on V2V links. Long Term Evolution (LTE)-V2X was the earliest standardization of the 3rd Generation Partnership Project (3GPP) in Release 14 regarding V2X communications which was followed by further improvements in the subsequent releases. Starting from Release 16, 3GPP has also introduced New Radio (NR)-V2X with the aim of bringing more flexibility and supporting advanced vehicular applications, e.g., platooning, extended sensors, and remote driving [2]. Among these applications, platooning has gained substantial interest from industry and academia as it has the potential to increase road capacity and reduce traffic congestion.

In short, a vehicle platoon is a chain of interconnected vehicles that share a typical moving pattern. One of the principal challenges in platooning is to preserve the string stability of the chain, i.e., a kinematic signal of interest (e.g., velocity, acceleration, etc.) does not amplify as it propagates among the vehicles [3]. To address this issue, various control strategies have been proposed. One such longitudinal control strategy that has been shown to produce satisfactory results in maintaining the string stability of the platoon, is cooperative adaptive cruise-control (CACC) which extends its earlier counterpart, adaptive cruise-control (ACC), by incorporating (control and kinematic) information transmission between the vehicles of the platoon via V2V communications [4], [5]. However, the challenge that faces the integration of CACC is its dependency on reliable V2V connections between the vehicles. Accordingly, developing efficient radio resource management (RRM) algorithms that can facilitate reliable communication in vehicular networks has attracted significant interest in recent years [6].

3GPP has specified two RRM strategies for cellular V2X communications [7], [8]. The first class, which is referred to as mode 1 in NR-V2X and mode 3 in LTE-V2X, is a centralized scheme and is only available if all the vehicles are under cellular coverage. In this case, the vehicles can either request new subchannels anytime they have a new packet for transmission, or the BS can reserve the necessary subchannels for the vehicles. The second class, known as mode 2 in NR-V2X and mode 4 in LTE-V2X, offers a distributed RRM solution. In this case, vehicles access the channel through sensing-based semi-persistent scheduling (SPS) in a distributed manner. The general blueprint in centralized approaches is to first acquire the channel state information (CSI) of all the V2I and V2V links and then solve the problem of interest. This approach becomes problematic when there is mobility in the network, as frequent CSI acquisition is needed from all the vehicles due to the fast channel alternations. This can subsequently increase the signaling overhead on the BS side. However, the second alternative is more error-prone due to its distributed nature and there is a higher probability of packet collisions. Nonetheless, when properly designed, distributed RRM can offer more flexibility and lower latency by eliminating the need for constantly involving the BS in the RRM process.

In conformity with 3GPP specifications on efficient RRM design, various distributed and centralized RRM algorithms have been proposed in the current literature. It is often the case that the RRM problem is modeled as a combinatorial optimization problem. The complexity and non-linearity of the problem usually leave no choice but to use some heuristic or sub-optimal algorithms to derive the solution. Consequently, to deal with these complex problems, there has been a growing tendency towards machine learning (ML)-based RRM algorithms, among which reinforcement learning (RL) maintains a higher ground compared to the other ML algorithms [9]. What makes these RL-based RRM approaches appealing is the level of flexibility they offer to solve the problem of interest, especially when distributed RRM approaches are preferred. The general procedure for treating RRM problems with RL is to model the problem as a Markov decision process (MDP) and then map the equilibrium point to the RRM solution. In most cases, however, the converged equilibrium point does not necessarily indicate the optimal solution. Furthermore, the same line of controversy regarding centralized and distributed approaches also lingers here. Distributed RL algorithms, also termed as multi-agent reinforcement learning (MARL) in the literature, exhibit poor performance and less robustness compared to the case when centralized RL is used [10]. Some studies have proposed federated learning (FL) combined with MARL to bridge this performance gap [11]. The common notion is that FL can help reduce the stationarity problem of the distributed RL algorithms. Indeed, FL reduces the estimation variance due to the periodic averaging performed over the agents' models which can improve the overall performance; however, it is still a matter of debate how close RL algorithms can get to the optimal solution of the designed problem or how much gains can be attained with FL-based RL? One such solution would be to compare the performance of RL with other classical model-based optimization methods to validate its true potential, which has hardly been addressed in the existing literature.

B. RELATED WORK

We have separated the related work into two parts. The first part mainly covers the RL-based RRM algorithms that have been used in more general V2X networks. In the second part, we give an overview of the works that focus more on communication control co-design in platooning and mainly use RL to derive optimal platoon control strategies.

1) RL-BASED RRM IN V2X NETWORKS

A great body of literature has already investigated different MARL-based RRM algorithms for vehicular networks. In [12], a decentralized MARL-based spectrum access scheme for vehicular networks is studied. In [13], the authors investigate the joint problem of mode selection, resource block assignment, and transmit power control for the Internet of vehicles network to maximize the overall network capacity while ensuring strict ultra reliable and low latency communications (URLLC) requirements of V2V links. Furthermore, RL-based resource allocation algorithm for a mobile edge computing (MEC) and unmanned aerial vehicles (UAV) supported vehicular network is investigated in [14]. The authors of [15] propose a RL-based decentralized resource allocation scheme in a vehicular network for both unicast and broadcast scenarios to address the latency constraints on the V2V links. In [16], the authors propose FL empowered computation offloading and resource management algorithm to reduce the task offloading delay and resource cost over heterogeneous vehicular networks. Similarly in [17], the authors propose a FL-based MARL algorithm to jointly optimize the channel selection and power

control for vehicular networks considering the reliability and delay requirements of V2V communication links. Other comparable works, e.g., [18], [19], [20], [21] tackle identical resource allocation problems with RL. In [22], the authors propose a RL based subchannel assignment and power control algorithm aiming at maximizing the data rate of V2I links while satisfying the latency requirement of V2V links. In the same vein, spectrum allocation for device-to-device (D2D) underlay communications has been studied in [23]. They also adopt meta learning to facilitate the fast adaptability of the resource allocation policy in the dynamic environment. The authors of [24] study the spectrum efficiency problem for connected vehicles. They employ long short term memory (LSTM) to predict the mobility pattern of the vehicles and then design a RL algorithm for proper channel allocation. In [25], a MARL-based resource allocation problem is investigated to jointly optimize the channel allocation and power control to satisfy the heterogeneous quality of service (QoS) requirements of V2V communications, including delay-sensitive and non-safety-related applications. The authors of [26] propose radio access network (RAN) slicing to support the heterogeneous requirements of a cellular V2X network and adopt RL to minimize the long-term cost, including, slicing configuration and QoS violation. With the aim of improving the conventional 3GPP Mode 4 resource allocation scheme which is based on SPS, the authors in [27] propose RL-SPS that allows the vehicles to independently select the proper radio resources. Their simulation shows that the proposed improvement, not only reduces the packet collisions but it also outperforms the conventional sensing-based SPS procedure. In [28], the authors propose a hybrid centralized-distributed RRM scheme based on graph matching and RL to maximize the system capacity while guaranteeing reliability requirements. The authors of [29] propose a resource allocation strategy based on Hungarian method and RL for a UAV-assisted vehicular network aiming at maximizing the UAV‘s long term energy efficiency. Finally in [30], the authors investigate the problem of minimizing the age of information (AoI) for a platooning vehicular network. They propose improvements based on objective function decomposition to enhance the performance of MARL-based resource allocation. All previous works study the RRM problem for vehicular networks, mostly focusing on maximizing the overall throughput of the network while considering the reliability of the V2V links. However, as mentioned earlier, the lack of valid comparisons and analysis makes it difficult to make a concrete statement about their proposed RL-based RRM performance.

2) COMMUNICATION CONTROL CO-DESIGN

The impact of V2X communications on platoon control performance has also been a matter of research, e.g., [31], where the authors model platoon control as a sequential stochastic decision problem and then use RL to find the optimal control policy. In [32], the authors propose a two-tiered strategy for resource allocation that considers platoon formation control, focusing on maximizing platoon size and minimizing total power consumption. The authors of [33] formulate the CACC control strategy as an MARL problem and introduce a quantization-based communication protocol to enhance the communication between the platoon's vehicles. In the same vein, [34] and [35] study the joint communication and control co-design problem for autonomous vehicular platoon networks and [36] proposes a RL-based control strategy for automated truck platooning.

The authors of [37] adopt RL to simultaneously learn the platoon control policy as well as V2X communication protocol. Their results demonstrate a notable communication overhead reduction. In [38], the authors investigate the platoon control problem using RL and dynamic programming methods. A parameterized batch actor-critic algorithm is proposed in [39] for longitudinal control of autonomous land vehicles. In the same vein, the authors of [40] study an ACC-based car-following control problem using an RL algorithm. As an extension to their work, they compare the performance of their proposed algorithm with model predictive control methods in [41]. In [42], a hybrid car-following strategy based on RL and CACC is proposed, in which CACC is used to improve the performance of the controller when the performance of RL is poor. Moreover, the proposed strategy can fully utilize the exploration capability of RL to deal with the car-following cases. To illustrate the necessity of the interplay between platoon control and resource allocation, a multi-timescale control and communication problem is designed in [43] and [44]. The authors decompose the problem into two sub-problems, namely, the communication-aware RL-based platoon control sub-problem and a control-aware RL-based RRM sub-problem. In the first part of their work [43], the authors focus on learning an optimal platoon control policy given a particular RRM policy, and in the second part [44], the authors focus on deriving an optimal RRM policy under the condition that the platoon control policy is available.

C. CONTRIBUTIONS OF THIS ARTICLE

There are mainly two angles that make this work stand out from the current literature, which we have discussed in the following subsections.

1) COMPARISON BETWEEN AI- AND NON-AI-BASED ALGORITHMS

Although RL for RRM has been of focal interest in recent literature (as mentioned in Section I-B1), there is still a lack of knowledge on the performance and behavior of the learned policies, which makes it challenging to draw a conclusive statement on the efficacy of the RL-based resource allocation algorithms. In this way forward, classical and RL-based RRM algorithms are designed for a vehicular platooning network to meet the control systems requirements and guarantee the radio links quality. With proper resource

allocation strategies, the string stability of the entire platoon is ensured.

2) COMMUNICATION CONTROL CO-DESIGN

Different from the majority of the works mentioned earlier in Section I-B2, we design an optimization problem that not only takes the radio link parameters, i.e., V2V links reliability, and minimum capacity requirements into account, but it also incorporates the platoon control parameters and its string stability. This joint modeling allows for evaluating the impact of the designed RRM policy on the string stability of the platoon, and, conversely, to assess the impact of the control policy on the RRM.

3) KEY CONTRIBUTIONS

In summary, the contributions of our work can be summarized as follows:

- With the focus on communication and control co-design, in this work, two RRM problems for a platoon of connected vehicles are formulated. More specifically, the first optimization problem aims to increase the total capacity of the V2V links while satisfying the link reliability, minimum capacity requirement, and the string stability of the platoon. The second optimization problem focuses on user fairness, with the goal of improving the minimum rate of all the V2V links in the platoon with similar constraints as in the first optimization problem.
- We extend our analysis in [35] to any arbitrary number of interfering V2V links utilizing the same frequency band and find the feasibility region for their simultaneous transmission. Then, we employ iterative centralized algorithms from [45] to locate the solutions of the sum rate and max-min rate optimization by adhering to the difference of two convex functions (d.c.) representation of the objective functions.
- We also model the proposed RRM problems as MDP and use distributed and FL-based MARL algorithms to solve the optimization problem. The comparison of the results from RL- and model-based optimization methods provides valuable insights that finally reveal the effectiveness and behavior of the policies learned by RL-based RRM methods. This understanding also paves the way for generalizing the behavior of RL-based algorithms when applied to intractable problems.

D. PAPER ORGANIZATION AND NOTATION

The remainder of this article is organized as follows. Section II provides the general system model of the platooning vehicular network where we model the control system and the communication links between the successive vehicles in the platoon along with formulating the optimization problems. In Section III, we design model- and RL-based solutions to tackle the optimization problems. The simulation results are provided in Section IV, and finally, Section V concludes the paper. To ease readability, all the primary notations of the paper are listed in Table 1.

TABLE 1. Primary Notations used in the paper.

Notation	Definition
$N/\mathcal{N}/i$	number/set/index of PMs in platoon
$q_i/v_i/a_i$	position/velocity/acceleration of vehicle i
η_i	internal actuator dynamics of vehicle i
$u_i/u_{fb_i}/u_{ff_i}$	total/feedback/feedforward control input of vehicle i
$\mathbf{x}_i$	dynamical state vector of vehicle i
$d_i^0/d_{r,i}$	desired standstill/inter-vehicle distance
$\mathcal{T}_{d,i}$	headway time
Λ_i	string stability
$\delta_{\mathrm{Tx},i}$	transmission delay
$D_i/F_i/G_i$	delay/feedforward filter/plant model of vehicle i
K_i/H_i	PD controller/spacing strategy of vehicle i
W	bandwidth
$\alpha_{i,i}/g_{i,i}^t$	pathloss/fading component of link i
γ_i^t	SINR of link i
γ_0	SINR threshold
σ^2	noise power spectral density
$\mathfrak{p}_0$	outage probability
c_0	minimum capacity requirement
ς	packet size
$\beta/\beta'/\beta''$	outage-related coefficients
κ_1/κ_2	reward-related coefficients
ζ	discount factor
ϖ	learning rate

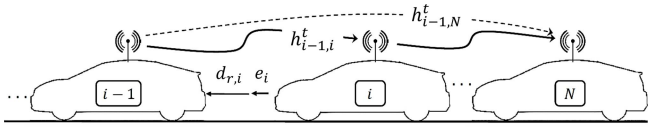


FIGURE 1. Vehicular platooning system model; the desired and interference links are shown with solid and dashed lines, respectively.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, the system model of the vehicular platooning network is presented. First, we analyze the necessary conditions for having a string stable platooning, and then we model the V2V communication links between the vehicles. Finally, two optimization problems are formulated with the purpose of improving the reliability and achievable data rate of the communication links which directly impact the stability of the platoon.

A. CACC DESIGN IN PLATOONING

Fig. 1 indicates a platoon of vehicles of size $N+1$, equipped with the CACC functionality. The first vehicle is assumed to be the platoon leader (PL), and the other vehicles are referred to as platoon members (PMs). We hereby define $\mathcal{N} = \{1, 2, \ldots, N\}, N \in \mathbb{N}$, as the set indicating the PMs of the platoon, indexed by $i \in \mathcal{N}$, and use index $i = 0$

to indicate the PL.[1] We denote the position, velocity and acceleration of each vehicle as q_i, v_i, and a_i, respectively. The dynamics of the vehicles in the platoon can be described by the following set of differential equations [4]

$$\dot{q}_i(t) = v_i(t), \tag{1a}$$
$$\dot{v}_i(t) = a_i(t), \tag{1b}$$
$$\dot{a}_i(t) = -\frac{1}{\eta_i}a_i(t) + \frac{1}{\eta_i}u_i(t), \tag{1c}$$

where η_i is related to the time constant of lag in responding to any commanded deceleration or acceleration and u_i is the control input for vehicle i defined as

$$u_i(t) = u_{fb_i}(t) + u_{ff_i}(t),$$

where u_i, u_{fb_i} and u_{ff_i} denote the total, feedback and feedforward control commands. The distance error between the vehicles can be modelled as

$$e_i(t) = q_{i-1}(t) - q_i(t) - L_i - d_{r,i}(t),$$

where $d_{r,i}(t) = d_i^0 + \mathcal{T}_{d,i}v_i(t)$ is the desired inter-vehicle spacing and L_i is the length of vehicle i. d_i^0 is referred to as the standstill target inter-vehicle distance, required to prevent a near collision at standstill, and $\mathcal{T}_{d,i}$ is the headway time. Following the design guidelines in [4], one can define the string stability transfer function relating the control signals of the vehicle $i-1$ with vehicle i by using the Laplace transform of their respective control signals as

$$\Lambda_i(s) = \left\| \frac{U_i(s)}{U_{i-1}(s)} \right\|_\infty \tag{2a}$$
$$= \left\| \frac{Z_i(s)D_i(s)F_i(s) + K_i(s)G_{i-1}(s)}{1 + K_i(s)G_i(s)H_i(s)} \right\|_\infty, \tag{2b}$$

where $U_i(s)$ refers to Laplace transform of the control signal of vehicle i, with s referring to the complex frequency domain parameter. Furthermore, $D_i(s)$ refers to the communication delay modeled in the s-domain as

$$D_i(s) = e^{-\delta_{\mathrm{Tx},i}s},$$

where $\delta_{\mathrm{Tx},i}$ is the transmission delay in seconds. Considering a packet-based communication model between the vehicles, we assume a zero-order hold (ZOH) signal reconstruction at the receiver side, in which its s-domain transfer function is represented as

$$Z_i(s) = \frac{1 - e^{-sT_i}}{sT_i},$$

with T_i denoting the transmission interval between the vehicles of the platoon.

$F_i(s)$ is the feedforward filter defined as

$$F_i(s) = \frac{G_{i-1}(s)}{H_i(s)G_i(s)},$$

where $G_i(s)$ refers to the plant model formulated as

$$G_i(s) = \frac{1}{s^2(\eta_i s + 1)}.$$

Finally, $K_i(s) = k_{p_i} + k_{d_i}s$ refers to the proportional-derivative (PD) controller and $H_i(s) = 1 + s\mathcal{T}_{d,i}$ models the spacing strategy. The dynamical state vector of vehicle i is comprised of four parts, i.e.,

$$x_i(t) = \left[e_i(t), v_i(t), a_i(t), u_{ff_i}(t)\right]^T \in \mathbb{R}^4.$$

Accordingly, the whole dynamical model of a vehicle (i) in the platoon can be defined as

$$\dot{\mathbf{x}}_i(t) = \mathbf{A}_{i,i}\mathbf{x}_i(t) + \mathbf{A}_{i,i-1}\mathbf{x}_{i-1}(t) + \mathbf{B}_{\mathrm{s},i}u_i(t) + \mathbf{B}_{\mathrm{c},i}u_{i-1}(t), \tag{3}$$

The state matrices $\mathbf{A}_{i,i}$ and $\mathbf{A}_{i,i-1}$ and control matrices $\mathbf{B}_{s,i}$ and $\mathbf{B}_{c,i}$ are obtained as

$$\mathbf{A}_{i,i} = \begin{bmatrix} 0 & -1 & -\mathcal{T}_{d,i} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\eta_i} & 0 \\ 0 & 0 & 0 & -\frac{1}{\mathcal{T}_{d,i}} \end{bmatrix},$$
$$\mathbf{A}_{i,i-1} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \left(1 - \frac{\eta_i}{\eta_{i-1}}\right)\frac{1}{\mathcal{T}_{d,i}} & 0 \end{bmatrix},$$
$$\mathbf{B}_{\mathrm{s},i} = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{\eta_i} \\ 0 \end{bmatrix}, \mathbf{B}_{\mathrm{c},i} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{\eta_i}{\eta_{i-1}}\frac{1}{\mathcal{T}_{d,i}} \end{bmatrix}, \forall i \geq 1. \tag{4}$$

Note that for the homogeneous cases, where we assume similar vehicle characteristics, we have $\frac{\eta_i}{\eta_{i-1}} \approx 1$. Subindices of $\mathbf{B}_{\mathrm{s}}$ and $\mathbf{B}_{\mathrm{c}}$ are used for emphasizing that the signal comes from sensor measurements or through communications, respectively.[2]

B. COMMUNICATION LINK MODELING

We assume that at the beginning of each scheduling slot, a bandwidth of W [Hz] is shared among the vehicles of a platoon [8]. Therefore, spectrum sharing between the vehicles becomes indispensable and interference must also be handled by proper power control. This facilitates reliable communications among the vehicles and subsequently results in string stability of the platoon. The time dimension is partitioned into time slots of equal length.[3] Δ and labeled as $t \in \mathbb{N}$. With some abuse of notation and to ease readibility, we refer to the pair of vehicle $i-1$ and i simply as (i, i). Accordingly, the V2V link between this vehicle pair during one time slot (we assume a coherence time $t_{\mathrm{coh}} \gg \Delta$ due to

[1] In a platoon of size $N+1$, with the exception of the first and last vehicle, all other vehicles are both transmitters and receivers. We therefore have N transmitters and N receivers.

[2] Interested readers can refer to [5] and the references therein, for a more comprehensive review over the CACC model.

[3] One should not confuse the definitions of t presented in the previous and current sections as they refer to a similar concept. Hereinafter, by t we refer to the scheduling time slot.

the small relative velocity of vehicles) is defined as $h_{i,i}^t = \alpha_{i,i}^t g_{i,i}^t$, where $\alpha_{i,i}^t$ denotes the path loss and $g_{i,i}^t$ models the small-scale component which we assume has an exponential power distribution with unit variance. Further, we define the allocated power for this link as p_i^t. Accordingly, the signal-to-interference-plus-noise ratio (SINR) at the i-th receiver can be defined as

$$\gamma_i^t(\mathbf{p}) = \frac{p_i^t h_{i,i}^t}{\sum_{j\in\mathcal{N}\setminus\{i\}} p_j^t h_{j,i}^t + \sigma^2}, \tag{5}$$

where $\mathbf{p} = [p_1, p_2, \ldots, p_N] \in \mathbb{R}^N$ is the vector of the allocated powers to V2V links, and σ^2 is the noise power spectral density. One common way to model the reliability of a communication link is to define the outage probability as

$$\mathrm{O}_i^t = \mathbb{P}\{\gamma_i^t \le \gamma_0\} \tag{6a}$$

$$= \mathbb{P}\left\{p_i^t h_{i,i}^t \le \sum_{j\in\mathcal{N}\setminus\{i\}} \gamma_0 p_j^t h_{j,i}^t + \gamma_0\sigma^2\right\}, \tag{6b}$$

where γ_0 denotes the SINR threshold. One can argue that (6) models the QoS requirements of the V2V links such that whenever (6) is satisfied, reliable communication is provided. Finally, the achievable capacity can also be defined as

$$\mathcal{C}_i^t(\mathbf{p}) = \log_2\left(1 + \frac{p_i^t h_{i,i}^t}{\sum_{j\in\mathcal{N}\setminus\{i\}} p_j^t h_{j,i}^t + \sigma^2}\right).$$

C. OPTIMIZATION PROBLEMS

In this section, we present two optimization problems, one of which is focused on maximizing the total data rate of the V2V links, and the other one is designed to maximize the minimum V2V link data rate. For both optimization problems, string stability, and V2V links reliability have been considered as the constraints, suggesting a similar feasible set for both optimization problems.

1) SUM-RATE OPTIMIZATION

The optimization problem can be written as

$$\mathcal{P}_1 : \max_{\mathbf{p}} \quad \sum_{i\in\mathcal{N}} \mathcal{C}_i^t(\mathbf{p}), \tag{7a}$$

$$\text{s.t. } \mathbb{P}\{\gamma_i^t \le \gamma_0\} \le \mathfrak{p}_0, \quad \forall i \in \mathcal{N} \tag{7b}$$

$$\mathcal{C}_i^t(\mathbf{p}) \ge c_0, \quad \forall i \in \mathcal{N} \tag{7c}$$

$$\Lambda_i(s) \le 1, \quad \forall i \in \mathcal{N} \tag{7d}$$

$$0 \le p_i^t \le p_{\max}, \quad \forall i \in \mathcal{N} \tag{7e}$$

where $\mathfrak{p}_0$ is the allowed outage probability and c_0 denotes the minimum capacity requirement. Further, (7d) stipulates the string stable platooning criterion [5], and (7e) forces a limit on the maximum power consumption.

2) MAX-MIN OPTIMIZATION

The second optimization problem with the focus on maximizing the minimum V2V link rate can be written as

$$\mathcal{P}_2 : \max_{\mathbf{p}} \min_{i\in\mathcal{N}} \quad \mathcal{C}_i^t(\mathbf{p}), \tag{8a}$$

$$\text{s.t. (7b)} - \text{(7e)}$$

The designed optimization problems aim at different V2X services. Specifically, the sum rate optimization deals with V2X services that require very high data rates, and the max-min optimization deals with safety-critical applications. Both the optimization problems have similar constraints and are formulated in the time domain, except for the string stability constraint (7d) related to CACC, which is defined in the frequency domain and is independent from the other constraints. Both optimization problems (7) and (8) contain non-convex functions. Further manipulations are required to transform them to a convex and tractable problem. In the following section we will provide different centralized and distributed solutions by which we can solve these problems.

III. SOLUTION DESIGN

This section provides different centralized and distributed RRM algorithms one can apply to solve the aforementioned optimization problems. The centralized solutions are assumed to be handled by the BS and it requires the global channel knowledge from all the vehicles. The distributed approaches are handled by the vehicles with partial knowledge of the channel information. For the centralized case, classical model-based solutions based on d.c. programming are exploited and for the distributed case, MARL is applied. This comparison is noteworthy as it reflects the viable gains may or may not be achieved through distributed RL-based approaches compared to the model-based solutions that require global knowledge of the channel to function properly.

A. CENTRALIZED MODEL-BASED APPROACH

In this section, we transform the optimization problems (7) and (8) into convex and tractable problems. Starting with the constraints, as mentioned (7d) is independent from the other constraints. One can replace the inequality with equality as $\Lambda_i(s) = 1$, and find the maximum allowable value for the string stability. The solution satisfying this constraint which yields the maximum allowable transmission interval (MATI) is quite well-known and has been addressed in various literature, e.g., [5], and [46]. More specifically, MATI provides an upper bound on the maximum time allowed for the packets to be transmitted per V2V link. One can translate this constraint to the minimum required rate c_0 in (7c) as

$$t_{\Lambda_i}^* W c_0 \ge \varsigma_i, \tag{9}$$

where $t_{\Lambda_i}^*$ denotes the MATI resulted from (7d) (see Appendix A) and ς_i is the packet size containing the control information. One should also notice the difference of (7b) and (7c) as in the first glance they might appear to be similar. Equation (7b) translates into the reliability of the

V2V links, whereas (7c) denotes the minimum capacity requirement. For example, if we denote the bit error rate (BER) of 10^{-3} as the reliable communication threshold, then for the case of 4-QAM modulation, γ_0 should be at least 10 dB for an additive white Gaussian noise (AWGN) channel. For the case of Rayleigh fading with no diversity and an outage probability of $\mathfrak{p}_0$, γ_0 is shifted to $\bar{\gamma}_0$ as

$$\bar{\gamma}_0 = \frac{\gamma_0}{-\ln(1-\mathfrak{p}_0)}. \tag{10}$$

In other words, it is possible to have a system that satisfies (7c) with $\gamma_i^t < \gamma_0$, however this will lead to a very high BER, meaning that most of the transmitted packets are error prone and the need for re-transmissions becomes necessary. In summary, joint consideration of both (7b) and (7c) is essential to facilitate a reliable communication with adequate data rate. With that being said, using (9), the string stability impact (7d) is translated into constraint (7c). Further, the outage constraint (7b) can be written as

$$\mathbb{P}\left\{ g_{i,i}^t \le \frac{\gamma_0 \sigma^2}{p_i^t \alpha_{i,i}^t} + \sum_{j \in \mathcal{N}\setminus\{i\}} \frac{\gamma_0 p_j^t h_{j,i}^t}{p_i^t \alpha_{i,i}^t} \right\} = \tag{11a}$$

$$= 1 - e^{\frac{-\gamma_0 \sigma^2}{p_i^t \alpha_{i,i}^t}} \prod_{j \in \mathcal{N}\setminus\{i\}} \frac{p_i^t \alpha_{i,i}^t}{p_i^t \alpha_{i,i}^t + \gamma_0 p_j^t \alpha_{j,i}^t}, \tag{11b}$$

where (11b) has been calculated by taking the integral over the fading exponents. We also assume due to the close distance of vehicles in the platoon, noise power is negligible compared to the direct V2V channel gain, i.e., $\sigma^2 \ll \alpha_{i,i}^t$, resulting in

$$\mathrm{O}_i^t : \approx 1 - \prod_{j \in \mathcal{N}\setminus\{i\}} \frac{p_i^t \alpha_{i,i}^t}{p_i^t \alpha_{i,i}^t + \gamma_0 p_j^t \alpha_{j,i}^t} \le \mathfrak{p}_0, \tag{12}$$

with O_i^t denoting the outage constraint for V2V link i. Further simplification of (12) is provided owing to the following Lemma from [47].

Lemma 1: For positive variables $z_1, \ldots, z_n \ge 0$, the following inequality holds

$$1 + \sum_{k=1}^{n} z_k \overset{(a)}{\le} \prod_{k=1}^{n} (1 + z_k) \overset{(b)}{\le} e^{\sum_{k=1}^{n} z_k}, \tag{13}$$

where (a) refers to the Weierstrass inequality [48], and (b) is followed by taking the logarithm and using the inequality $\log(1+z) \le z$.

One should notice, by using either side of the inequality from Lemma 1, the feasible set of the optimization problems (7) and (8) is altered. More specifically, using (a) from Lemma 1 results in a feasible set which is larger than the actual feasible set from (12) and using (b) will result in a feasible set which is tighter, leading to an upper and lower bound respectively. Therefore, one should consider both cases and compare the results to see the gap between the two bounds. With this in mind, by using (a), (12) can be simplified into

$$\mathrm{O}_i^t : \approx \prod_{j \in \mathcal{N}\setminus\{i\}} \left(1 + \frac{\gamma_0 p_j^t \alpha_{j,i}^t}{p_i^t \alpha_{i,i}^t}\right) \le \frac{1}{1-\mathfrak{p}_0}, \tag{14a}$$

$$\overset{(13a)}{\Longrightarrow} \sum_{j \in \mathcal{N}\setminus\{i\}} \left(\frac{\gamma_0 p_j^t \alpha_{j,i}^t}{p_i^t \alpha_{i,i}^t}\right) \le \frac{\mathfrak{p}_0}{1-\mathfrak{p}_0}, \tag{14b}$$

$$\Longleftrightarrow (\beta + \gamma_0) p_i^t \alpha_{i,i}^t \ge \sum_{j \in \mathcal{N}} \gamma_0 p_j^t \alpha_{j,i}^t, \tag{14c}$$

where $\beta = \mathfrak{p}_0/(1-\mathfrak{p}_0)$. One can represent (14) in the matrix form as

$$\gamma_0 \mathbb{E}[\mathbf{H}^t]\mathbf{p} \le (\beta + \gamma_0)\mathrm{diag}(\alpha_{i,i}^t)\mathbf{p}, \tag{15}$$

where $\mathbb{E}[\mathbf{H}^t] = [\alpha_{j,i}^t]_{j,i \in \mathcal{N}}$ denotes the channel matrix comprised of the pathloss components only, and $\mathrm{diag}(\alpha_{i,i}^t)$ is the diagonal matrix whose entries are equal to $\alpha_{i,i}^t$. Similarly, by using (b) and the assumption that the approximation is smaller than $(1-\mathfrak{p}_0)^{-1}$, (12) becomes

$$e^{\sum_{j \in \mathcal{N}\setminus\{i\}} \left(\frac{\gamma_0 p_j^t \alpha_{j,i}^t}{p_i^t \alpha_{i,i}^t}\right)} \le \frac{1}{1-\mathfrak{p}_0}, \tag{16a}$$

$$\Longleftrightarrow (\beta' + \gamma_0) p_i^t \alpha_{i,i}^t - \sum_{j \in \mathcal{N}} \gamma_0 p_j^t \alpha_{j,i}^t \ge 0, \tag{16b}$$

where $\beta' = \ln(1-\mathfrak{p}_0)^{-1}$. Similar to (15), we can define

$$\gamma_0 \mathbb{E}[\mathbf{H}^t]\mathbf{p} \le (\beta' + \gamma_0)\mathrm{diag}(\alpha_{i,i}^t)\mathbf{p}. \tag{17}$$

Also note that constraint (7c) can be rewritten as follows

$$p_i^t h_{i,i}^t + \beta'' \left(\sum_{j \in \mathcal{N}\setminus\{i\}} p_j^t h_{j,i}^t + \sigma^2 \right) \ge 0, \ \ \forall i \in \mathcal{N}, \tag{18}$$

where $\beta'' = (1 - 2^{c_0})$. In matrix form (18) can be written as

$$\mathrm{diag}(h_{i,i}^t)\mathbf{p} + \beta'' \left(\boldsymbol{\sigma}^2 + \left(\mathbf{H}^t - \mathrm{diag}(h_{i,i}^t)\right)\mathbf{p} \right) \ge 0. \tag{19}$$

Here, $\boldsymbol{\sigma}^2 = \sigma^2 \mathbf{1} \in \mathbb{R}^N$ where $\mathbf{1}$ denotes the vector with all entries one. Now it only remains to simplify the objectives in (7) and (8). First, let us define $\mathfrak{F}_i^t(\mathbf{p})$ and $\mathfrak{R}_i^t(\mathbf{p})$ as

$$\mathcal{C}_i^t(\mathbf{p}) = \underbrace{\log_2 \left(\sigma^2 + \sum_{j \in \mathcal{N}} p_j^t h_{j,i}^t \right)}_{\mathfrak{F}_i^t(\mathbf{p})} \tag{20a}$$

$$\underbrace{- \log_2 \left(\sigma^2 + \sum_{j \in \mathcal{N}\setminus\{i\}} p_j^t h_{j,i}^t \right)}_{\mathfrak{R}_i^t(\mathbf{p})}. \tag{20b}$$

As can be seen, $\mathfrak{F}_i^t(\mathbf{p})$ and $\mathfrak{R}_i^t(\mathbf{p})$ are both concave in $\mathbf{p}$, while, their difference is not, but $\mathcal{C}_i^t(\mathbf{p})$ is indeed the difference of two concave functions. One can approximate $\mathfrak{R}_i^t(\mathbf{p})$ by its first order approximation near $\mathbf{p}^{(k)}$. As $\mathfrak{R}_i^t(\mathbf{p})$ is a concave

function, this approximation is tangent to the function from above [49], therefore

$$\mathfrak{R}_i^t(\mathbf{p}) \leq \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right) + \left\langle \nabla \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right), \mathbf{p} - \mathbf{p}^{(k)} \right\rangle, \tag{21}$$

where $\nabla \mathfrak{R}_i^t(\mathbf{p}^{(k)})$ is defined as

$$\nabla \mathfrak{R}_i^t(\mathbf{p}) = \frac{1}{\sigma^2 + \sum_{j \in \mathcal{N} \setminus \{i\}} p_j^t h_{j,i}^t} \boldsymbol{\epsilon}_i, \tag{22}$$

with $\boldsymbol{\epsilon}_i(j) = \frac{h_{j,i}^t}{\ln 2}$ and $\boldsymbol{\epsilon}_i(i) = 0$. Replacing (21) in (20) yields

$$\tilde{\mathcal{C}}_i^t(\mathbf{p}) = \mathfrak{F}_i^t(\mathbf{p}) - \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right) - \left\langle \nabla \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right), \mathbf{p} - \mathbf{p}^{(k)} \right\rangle, \tag{23}$$

where $\tilde{\mathcal{C}}_i^t(\mathbf{p})$ refers to the approximated value for $\mathcal{C}_i^t(\mathbf{p})$. (23) acts as a well approximated lower bound for the non-convex function $\mathcal{C}_i^t(\mathbf{p})$ in (20) as

$$\begin{aligned} &\mathfrak{F}_i^t\left(\mathbf{p}^{(k)}\right) - \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right) - \left\langle \nabla \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right), \mathbf{p}^{(k+1)} - \mathbf{p}^{(k)} \right\rangle \\ &\leq \mathfrak{F}_i^t\left(\mathbf{p}^{(k+1)}\right) - \mathfrak{R}_i^t\left(\mathbf{p}^{(k+1)}\right). \end{aligned}$$

Therefore, maximizing $\mathcal{C}_i^t(\mathbf{p})$ is directly translated into lower bound maximization. With these modifications, now we are ready to redefine the proposed optimization problems (7) and (8). As mentioned, depending on using (15) or (17), we will have a larger or smaller feasible set which will result in two possible solutions. More explanation in this regard will be given in the following.

1) MODIFIED SUM-RATE OPTIMIZATION

The Modified optimization problem (7) reads as follows

$$\begin{aligned} \mathcal{P}_1^{(ub)}: \max_{\mathbf{p}} \quad & \sum_{i \in \mathcal{N}} \mathfrak{F}_i^t(\mathbf{p}) - \sum_{i \in \mathcal{N}} \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right) \\ & - \sum_{i \in \mathcal{N}} \left\langle \nabla \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right), \mathbf{p} - \mathbf{p}^{(k)} \right\rangle, && \text{(24a)} \\ \text{s.t. } & (15),\ (18) && \text{(24b)} \\ & \mathbf{0} \leq \mathbf{p} \leq p_{\max}\mathbf{1}, && \text{(24c)} \end{aligned}$$

where as mentioned, $\mathcal{P}_1^{(ub)}$ refers to the optimization problem with a larger feasible set denoted as $\Omega_1^{(ub)}$. Similarly, we define $\mathcal{P}_1^{(lb)}$ with feasible set $\Omega_1^{(lb)}$, that refers to the same optimization problem as (24) where we replace (15) with (17). As mentioned, the following relation holds true

$$\begin{aligned} \mathcal{P}_1^{(ub)*} &\geq \mathcal{P}_1^* \geq \mathcal{P}_1^{(lb)*} && \text{(25a)} \\ \Omega_1^{(lb)} &\subseteq \Omega_1 \subseteq \Omega_1^{(ub)}, && \text{(25b)} \end{aligned}$$

where Ω_1 refers to the primary optimization problem (7) feasible set. Algorithm 1 presents the sum rate optimization algorithm.

Algorithm 1: Sum Rate Optimization

Input: $\mathbf{H}$, γ_0, c_0, $\mathfrak{p}_0$, $p_{\max}$, ε
Output: $\mathbf{p}^*$

Set initial value for $\mathbf{p}^{(0)}$, and find $\sum_{i \in \mathcal{N}} \mathcal{C}_i^t\left(\mathbf{p}^{(0)}\right)$
Set iteration $k \leftarrow 0$
while *error* $> \varepsilon$ **do**
  Solve optimization problem (24) and find $\mathbf{p}^*$
  $k \leftarrow k + 1$
  $\mathbf{p}^{(k)} \leftarrow \mathbf{p}^*$
  $error = \left| \sum_{i \in \mathcal{N}} \mathcal{C}_i^t\left(\mathbf{p}^{(k)}\right) - \sum_{i \in \mathcal{N}} \mathcal{C}_i^t\left(\mathbf{p}^{(k-1)}\right) \right|$

Algorithm 2: Max-Min Optimization

Input: $\mathbf{H}$, γ_0, c_0, $\mathfrak{p}_0$, $p_{\max}$, ε
Output: $\mathbf{p}^*$

Set initial value for $\mathbf{p}^{(0)}$, and find $\min_{i \in \mathcal{N}} \mathcal{C}_i^t\left(\mathbf{p}^{(0)}\right)$
Set iteration $k \leftarrow 0$
while *error* $> \varepsilon$ **do**
  Solve optimization problem (26) and find $\mathbf{p}^*$
  $k \leftarrow k + 1$
  $\mathbf{p}^{(k)} \leftarrow \mathbf{p}^*$
  $error = \left| \min_{i \in \mathcal{N}} \mathcal{C}_i^t\left(\mathbf{p}^{(k)}\right) - \min_{i \in \mathcal{N}} \mathcal{C}_i^t\left(\mathbf{p}^{(k-1)}\right) \right|$

2) MODIFIED MAX-MIN OPTIMIZATION

With the same procedure Modified (8) reads as follows

$$\begin{aligned} \mathcal{P}_2^{(ub)}: \max_{\mathbf{p}, \vartheta} \quad & \vartheta && \text{(26a)} \\ \text{s.t. } & (15),\ (18), && \text{(26b)} \\ & \forall i \in \mathcal{N}: \ \mathfrak{F}_i^t(\mathbf{p}) - \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right) \\ & - \left\langle \nabla \mathfrak{R}_i^t\left(\mathbf{p}^{(k)}\right), \mathbf{p} - \mathbf{p}^{(k)} \right\rangle \geq \vartheta, && \text{(26c)} \\ & \mathbf{0} \leq \mathbf{p} \leq p_{\max}\mathbf{1}, && \text{(26d)} \end{aligned}$$

where ϑ is another variable that refers to the minimum rate among the links to be maximized. Similarly, one can construct $\mathcal{P}_2^{(lb)}$ by replacing (15) with (17). A similar relation to (25) also holds here as

$$\begin{aligned} \mathcal{P}_2^{(ub)*} &\geq \mathcal{P}_2^* \geq \mathcal{P}_2^{(lb)*} && \text{(27a)} \\ \Omega_2^{(lb)} &\subseteq \Omega_2 \subseteq \Omega_2^{(ub)}, && \text{(27b)} \end{aligned}$$

Algorithm 2 presents the max-min optimization algorithm.

B. DISTRIBUTED MARL APPROACH

In this section, we introduce the proposed MARL based resource allocation algorithm to solve the introduced optimization problems. First, the original problem is transformed into a Markov game and then the respective states, actions and reward function are defined. Finally, we derive the learning procedure based on the twin delayed deep deterministic policy gradient (TD3) [50] and FL algorithm.

1) ENVIRONMENT DESIGN

We define the multi-agent environment as a tuple $(M, S, (\mathcal{A}_i)_{i\in\mathcal{N}}, (r_i)_{i\in\mathcal{N}}, \zeta, (\mathcal{O}_i)_{i\in\mathcal{N}})$, where $M > 1$ is the number of agents, S is the total state space of the environment in which every agent has an imperfect information, $\mathcal{A}_i$ is the action space of agent i, ζ is the discount factor, r_i is the reward agent i gets following its action, and $\mathcal{O}_i$ is the observation space of i-th agent. Note that the designed multi agent system renders a fully distributed solution and no information sharing is assumed between the agents. In what follows, we map the defined elements to the platooning environment.

- *Agents:* We assume each transmitter vehicle as an independent agent, therefore, the total number of agents is $M = N$.
- *Observation:* The observation of each agent i at time instant t contains its estimated channel gain, i.e., $h^t_{i,i}$ and the total received interference as,

$$I_i^{t-1} = \sum_{j\in\mathcal{N}\setminus\{i\}} p_j^{t-1} h_{j,i}^{t-1} + \sigma^2.$$

  In summary, the observation of agent i is denoted as

$$o_i^t = \left(h_{i,i}^t, I_i^{t-1}\right) \in \mathcal{O}_i^t. \tag{28}$$

- *Action:* Each agent decides for its transmission power p_i^t independently. Therefore, the action space for agent i is defined as follows:

$$\mathcal{A}_i^t = \left\{a_i^t := p_i^t \mid p_i^t \in \left[0, p_{\max}\right]\right\}. \tag{29}$$

- *Reward:* Rewards drive the learning process in RL. The reward is defined as follows:

$$r_i^t = \kappa_1\left(\mathcal{C}_i^t - c_0\right) + \kappa_2\left(f(\gamma_i^t) - f(\bar{\gamma}_0)\right), \tag{30}$$

  where $\bar{\gamma}_0$ is the fade margin defined in (10), and κ_1 and κ_2 are the coefficients for balancing the reward. Furthermore, $f(x) = 10\log_{10}(x)$ converts the SINR to be in similar range as the capacity $\mathcal{C}^t$. As can be seen the reward is designed in such way to force the agent towards selecting power values that contribute the most to its data rate and SINR maximization.

2) LEARNING PROCEDURE

In RL, each agent i, given its observation $o_i^t \in \mathcal{O}_i^t$, selects an action $a_i^t \in \mathcal{A}_i^t$ according to its policy $\pi_i{:}\mathcal{O}_i^t \to \mathcal{A}_i^t$, which is parameterized by $\boldsymbol{\theta}_{\psi_i}$ and receives a reward r_i^t. The objective is to find the optimal policy π_i^* in a way that the discounted cumulative reward defined as

$$\mathcal{J}_i\left(\boldsymbol{\theta}_{\psi_i}\right) = \mathbb{E}_{o_i^t, a_i^t}\left[\sum_{\tau=0}^{\infty} \zeta^\tau r_i\left(o_i^t, a_i^t\right)\right] \tag{31}$$

is maximized for each agent. Therefore, the optimization problem that has to be considered for each agent i is defined as

$$\mathcal{P}_3 : \max_{\boldsymbol{\theta}_{\psi_i}} \quad \mathcal{J}_i\left(\boldsymbol{\theta}_{\psi_i}\right), \tag{32a}$$

$$\text{s.t. } a_i^t \sim \pi_i\left(a_i^t \mid o_i^t\right), \tag{32b}$$

$$o_i^{t+1} \sim \mathbb{P}\left(o_i^{t+1} \mid o_i^t, a_i^t\right). \tag{32c}$$

where $\mathbb{P}(o_i^{t+1} \mid o_i^t, a_i^t)$ denotes the Markov transition probability. As (32) is an unconstrained optimization problem, the policy can be updated by taking the gradient of $\mathcal{J}_i(\boldsymbol{\theta}_{\psi_i})$. In actor critic algorithms, the policy which is also referred to as the actor network can be updated through the deterministic policy gradient algorithm [52] as

$$\nabla_{\boldsymbol{\theta}_{\psi_i}} \mathcal{J}_i = \mathbb{E}_{o_i'}\left[\nabla_a Q_i^\pi\left(o_i^t, a_i^t; \boldsymbol{\theta}_{q_i}\right) \nabla_{\boldsymbol{\theta}_{\psi_i}} \pi_i\left(o_i^t; \boldsymbol{\theta}_{\psi_i}\right)\right], \tag{33}$$

where Q_i^π refers to agent i's action-value function (Q-function) parameterized by $\boldsymbol{\theta}_{q_i}$. Following the TD3 algorithm [50], we can accordingly define the loss function for the critic network, as

$$\mathcal{L}\left(\boldsymbol{\theta}_{q_i}\right) = \mathop{\mathbb{E}}_{(o_i, a_i, r, o_i')\sim\mathcal{D}}\left[\left(y - Q_i^\pi\left(o_i^t, a_i^t; \boldsymbol{\theta}_{q_i}\right)\right)^2\right], \tag{34}$$

where $\mathcal{D}$ refers to the replay buffer and

$$y = r_i + \zeta \min_{j=1,2} Q_{i_j}^\pi\left(o_i', a_i'\right), \tag{35}$$

with $a_i' \sim \pi_i(o_i'; \boldsymbol{\theta}'_{\psi_i})$ denoting the action from the target network. In TD3, there are two critic networks for each agent, denoted as $Q_{i_1}^\pi$, and $Q_{i_2}^\pi$, where the loss is calculated by taking the minimum of the two networks' outputs. Similarly, the loss function for the actor network are defined as

$$\mathcal{L}\left(\boldsymbol{\theta}_{\psi_i}\right) = -\mathop{\mathbb{E}}_{o_i\sim\mathcal{D}}\left[Q_i^\pi\left(o_i^t, \pi_i\left(o_i; \boldsymbol{\theta}_{\psi_i}\right); \boldsymbol{\theta}_{q_i}\right)\right]. \tag{36}$$

3) FEDERATED LEARNING

Since the agents only have access to their local information, and thus act independently, the policies that they take might be in contradiction with the other agents. Therefore, federated learning along with MARL can be considered to relieve this issue, and accelerate the training, as proposed in the recent literature [53], [54]. For this, the agents periodically send their network models to a central server, in which the gradients are aggregated iteratively and then shared among the agents. The aggregation procedure can be written as follows

$$\boldsymbol{\theta}_{m+1}^{(i)} = \begin{cases} \frac{1}{M}\sum_{i=1}^{M}\left[\boldsymbol{\theta}_m^{(i)} - \varpi\nabla\mathcal{L}\left(\boldsymbol{\theta}_m^{(i)}\right)\right], & m \bmod \iota = 0 \\ \boldsymbol{\theta}_m^{(i)} - \varpi\nabla\mathcal{L}\left(\boldsymbol{\theta}_m^{(i)}\right), & \text{otherwise} \end{cases} \tag{37}$$

where $\boldsymbol{\theta}_m^{(i)}$ can be either $\boldsymbol{\theta}_{\psi_i}$ or $\boldsymbol{\theta}_{q_i}$, m is the iteration, ι is the aggregation period, and ϖ is the learning rate. A summary of federated TD3-based MARL is given in Algorithm 3.

IV. SIMULATION RESULTS

In this section, we evaluate the proposed d.c. programming- and RL-based RRM algorithms to validate their performance for the platooning vehicular network.[4] All simulation parameters can be found in Table 2. It is noteworthy to mention

[4]Complete source code is available at: https://github.com/M-Parvini/V2X-RRM-IEEE-OJ-COMS-2023.

Algorithm 3: Federated MARL Algorithm

Initialize the Policy and Q-function parameters $\left(\boldsymbol{\theta}_{\psi_i}, \boldsymbol{\theta}^1_{q_i}, \boldsymbol{\theta}^2_{q_i}\right)$ and their corresponding target networks.
for *each episode* **do**
  Update platoon location and channel gains
  for *each time step t* **do**
    for *each agent i* **do**
      Observe state o_i^t
      Select action $a_i^t \sim \pi_i(a_i^t \mid o_i^t)$
    Receive the rewards from (30)
  for *each agent i* **do**
    Update the Q network from (34)
    Update the policy network from (36)
    Update target network parameters:
    $\boldsymbol{\theta}'_{\psi_i} \leftarrow \varrho\boldsymbol{\theta}_{\psi_i} + (1-\varrho)\boldsymbol{\theta}'_{\psi_i}$
    $\boldsymbol{\theta}'_{q_i} \leftarrow \varrho\boldsymbol{\theta}_{q_i} + (1-\varrho)\boldsymbol{\theta}'_{q_i}$
  Perform federated averaging (37) [51].

TABLE 2. Simulation parameters.

Communication parameters	**Value**	**Ref.**
Carrier frequency (f_c)	5.9 GHz	[55]
Bandwidth (W)	1 MHz	[55]
Size of platoon (N)	{7, 9}	
Platoon initial speed	140 km/h	[56]
Vehicles' antenna heights	1.5 m	[55]
Vehicles' antenna gains	3 dBi	[55]
Vehicles' receiver noise figure (N_F)	9 dB	[55]
Vehicles' maximum transmit power	30 dBm	
Packet size (ς)	6000 bytes	[56]
V2V links path loss model	LoS/NLoS from	[55]
Minimum capacity requirement (c_0)	5 bps/Hz	
SINR threshold (γ_0)	{5, 10, 15, 20} dB	
V2V links reliability ($\mathfrak{p}_0$)	1%	
Fast fading update V2V links	Every 1 ms	
Fast fading model	Rayleigh fading	
Vehicular control parameters	**Value**	**Ref.**
MATI ($t^*_{\Lambda_i}$) from (2)	149 ms	[5]
Desired standstill vehicle spacing (d_i^0)	100 m	[56]
Headway time-constant ($\mathcal{T}_{d,i}$)	0.5	[5]
Proportional-derivative controller gains	[k_{p_i}=0.25, k_{d_i}=0.5]	[5]
Plant model gain (k_{g_i})	1	[5]
Actuator dynamics (η)	0.1	[5]
Neural networks parameters	**Value**	
Experience replay buffer size	50000	
Mini batch size	32	
Actor network hidden layers	64, 32, 16	
Critic network hidden layers	64, 32, 16	
Critic/Actor networks learning rate	0.001/0.0001	
Discount factor	0.99	
Actor/Critic soft update parameter	0.01, 0.001	
Number of training episodes	30	
Number of test episodes	10	
Simulation time per episode	30 s	
Federated learning update interval (ϱ)	per episode	
Reward coefficients [κ_1, κ_2]	0.5, 0.5	

that federated aggregation and all the critic networks are only involved during the training and for evaluation only the trained actor models are utilized. All results are obtained by averaging over 10 test episodes, each with a duration of 30 seconds. Our analysis consists of the following algorithms

- *Sum rate optimization*, which we explained in Algorithm 1. In simulations we refer to this algorithm as "*Sum Rate*".
- *Max-Min*, which we explained in Algorithm 2. In simulations we refer to this algorithm as "*Max-Min*".
- *Federated MARL*, which we explained in Algorithm 3. In simulations we refer to this algorithm as "*Fed. MARL*".
- *Decentralized MARL*, which follows the exact procedure as Algorithm 3 excluding the federated averaging in (37). The aim is to indicate whether or not there is any benefit from federated learning compared to fully decentralized MARL. In simulations we refer to this algorithm as "*Dec. MARL*".
- *Centralized RL*, for which we consider the BS as an intelligent RL-based resource scheduler. Here, similar to centralized model-based approaches, we assume that the BS has complete knowledge over all the V2V links of the platoon and it has to decide on the transmitted power of all vehicles jointly. Therefore the observation, action and the reward function for the centralized RL are redefined as

$$o^t = \left(\left[h^t_{ji}\right]_{j\in\mathcal{N}, i\in\mathcal{N}}\right) \in \mathcal{O}^t := \bigcup_{i\in\mathcal{N}} \mathcal{O}^t_i, \tag{38a}$$

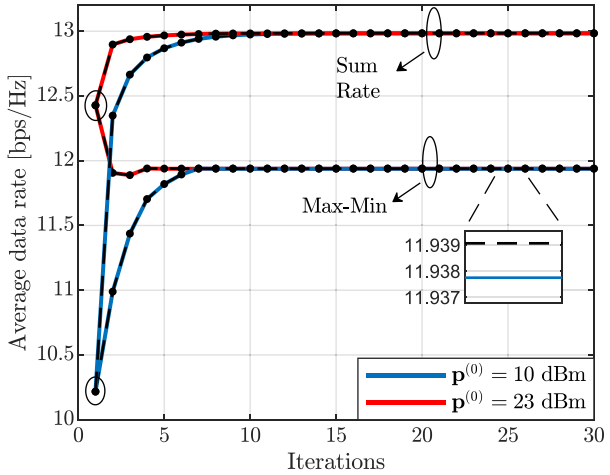
$$a^t = \mathbf{p} \in \mathcal{A}^t := \bigcup_{i\in\mathcal{N}} \mathcal{A}^t_i, \tag{38b}$$

$$r^t = \frac{1}{N}\sum_{i\in\mathcal{N}} r^t_i. \tag{38c}$$

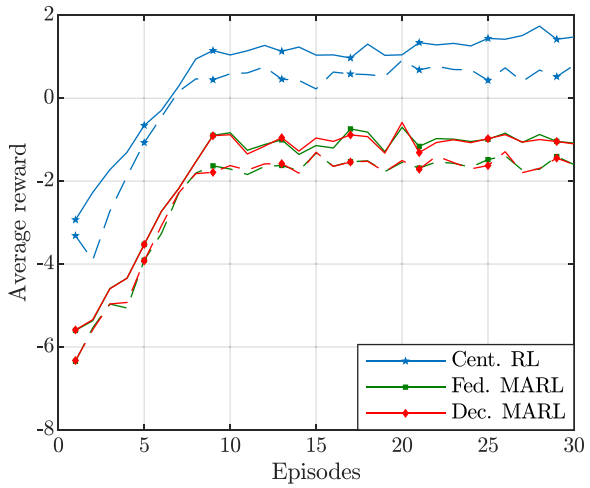
In simulations we refer to this algorithm as "*Cent. RL*".

A. CONVERGENCE EVALUATION

Fig. 2 compares the convergence behavior of the proposed model-based classical approaches with RL-based algorithms. Starting with model-based algorithms, Fig. 2(a), demonstrates fast convergence of both *Sum Rate* and *Max-Min* algorithms. As expected, (25) and (27) both hold true (see the figure). This plot also suggests that the optimal point is achievable irrespective of the initial points. Indeed, the *Sum Rate* reaches higher average data rates compared to *Max-Min* as it targets a different objective function. In comparison we have also shown the convergence behavior of RL-based approaches in Fig. 2(b). As it is shown in the figure, *Cent. RL* achieves higher reward compared to *Fed. MARL* and *Dec. MARL*. The reason comes from the fact that *Cent. RL* has full knowledge over all the V2V channel gains, similar to model-based algorithms, which enables it to converge to better policies. Interestingly, both *Fed. MARL* and *Dec. MARL* have yielded similar convergence behavior, indicating the ineffectiveness of federated averaging. Fig. 2(b) also reveals that better rewards and performance are attainable for the case of fewer vehicles as the level of interference between the vehicles is less. This very behavior is similar for all the RL-based algorithms.

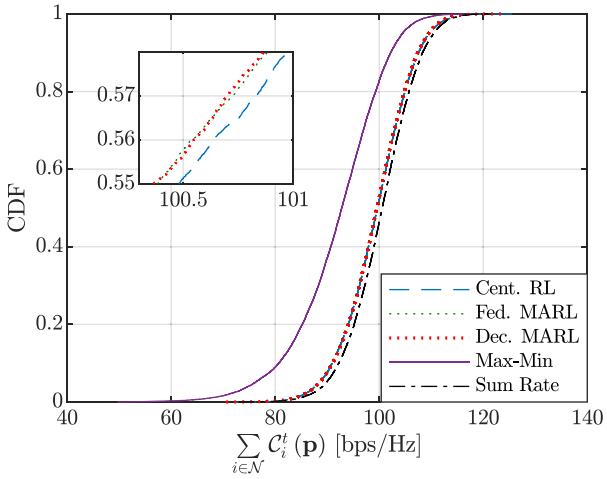


(a) Convergence of the sum data rate (Algorithm 1) and max-min data rate (Algorithm 2). Solutions from $\mathcal{P}_1^{(lb)}$ and $\mathcal{P}_2^{(lb)}$ are shown in solid lines and solutions from $\mathcal{P}_1^{(ub)}$ and $\mathcal{P}_2^{(ub)}$ are shown in dashed lines with markers.

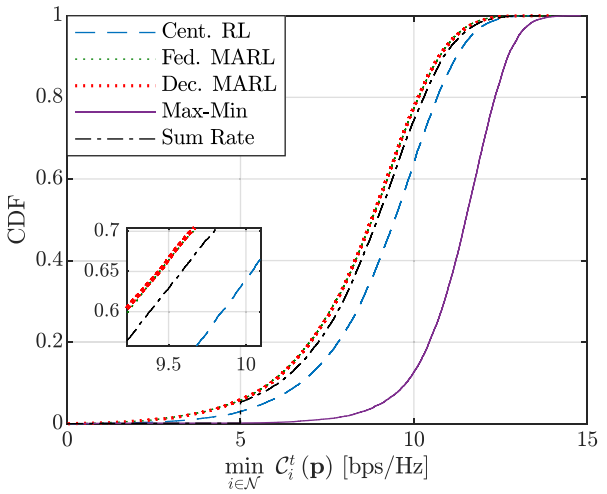


(b) RL-based algorithms convergence. Solid lines indicate $N = 7$ and dashed lines indicate $N = 9$.

FIGURE 2. Comparison of the convergence behavior of the model-based classical approaches and RL-based algorithms.



(a) Sum data rate CDF.



(b) Minimum data rate CDF.

FIGURE 3. CDF plot comparison of the proposed algorithms in terms of sum data rate and min data rate, with $\gamma_0 = 15$ dB, and $N = 9$.

B. ACHIEVABLE RATE DISTRIBUTION

In Fig. 3(a), we compare the cumulative distribution functions (CDFs) of the instantaneous sum capacity achieved by the proposed algorithms. From the figure, it can be seen that the *Sum Rate* algorithm attains a higher sum data rate compared to the other baselines. *Max-Min* however, has the worst performance as the objective is centered around maximizing the minimum rate. MARL-based algorithms perform closely to *Sum Rate*, with *Cent. RL* achieving higher values in comparison with *Fed. MARL* and *Dec. MARL*. In the same vein, Fig. 3(b) indicates the CDFs of the proposed algorithms in terms of the minimum achievable data rate. This comparison is of the highest interest as it demonstrates how the proposed algorithms behave in terms of per-link data rate fairness. From the figure, the *Max-Min* algorithm exhibits the highest fairness and transcends other algorithms by a larger margin. Finally, *Dec. MARL* and *Fed. MARL* perform equally poorly and worse than the other algorithms.

C. RELIABILITY EVALUATION

Fig. 4 compares the proposed algorithms in terms of average V2V links reliability. From the figure, the *Max-Min* algorithm yields the highest reliability compared to the other algorithms with reliability of more than 99% up to $\gamma_0 = 15$ dB. The reason for this can be seen in Fig. 3(b), which shows that *Max-Min* provides better fairness compared to the other baselines and tries to raise the performance of the worst users. The *Sum Rate* algorithm performs worse but is better than RL-based solutions. This algorithm prioritizes maximizing the total data rate of V2V links while paying less attention to the weaker links, which are the main source of unreliable communication. With similar interpretations as before, *Cent. RL* shows higher reliability compared to *Fed. MARL* and *Dec. MARL*. Again, federated averaging had no

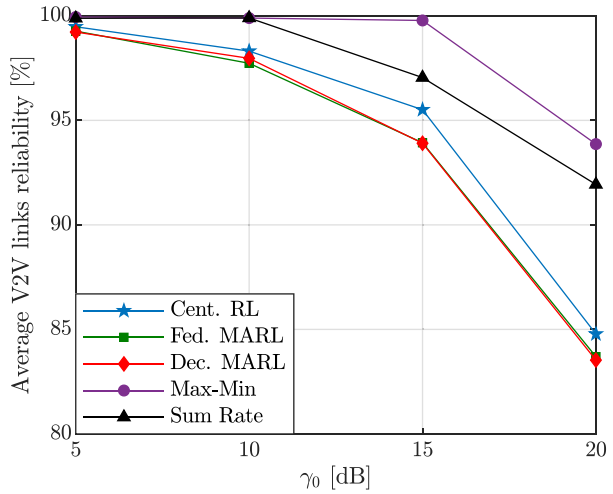


FIGURE 4. Average V2V links reliability with varying γ_0 threshold, assuming $N = 9$.

TABLE 3. Average data rate of the algorithms in bps/Hz.

	Cent. RL	*Fed. RL*	*Dec. RL*	*Max-Min*	*Sum Rate*
$\gamma_0 = 5$ dB	12.56	12.56	12.56	11.75	12.73
$\gamma_0 = 15$ dB	12.58	12.57	12.58	12.04	12.73

impact on the performance of *Dec. MARL*. Finally, as one can expect, by increasing the threshold γ_0, the reliability of the V2V links will be compromised.

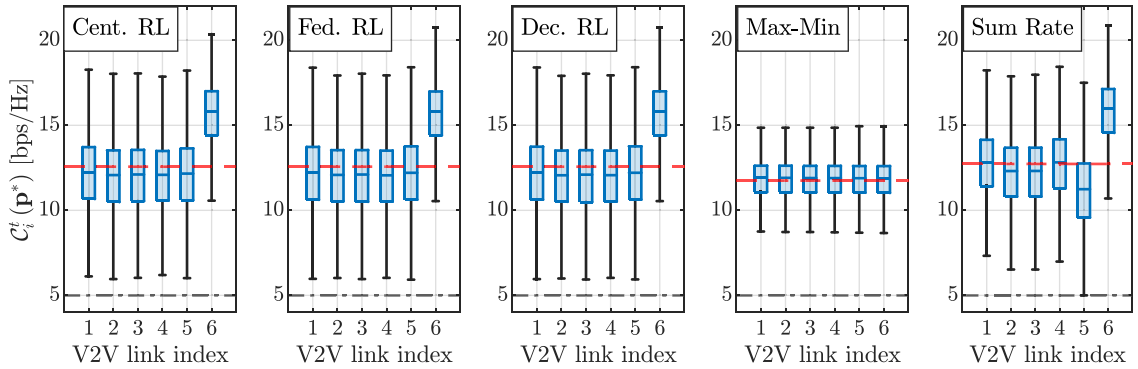
D. FAIRNESS EVALUATION AND POWER CONSUMPTION

In Fig. 5 we have compared each algorithm's performance in terms of per V2V link capacity. This comparison is of the highest interest, as looking only at the total capacity of the V2V links does not provide a comprehensive insight into the behavior of the algorithms in terms of per-link performance. Further, we have shown the average achieved data rate of the proposed algorithms for different values of γ_0 in Table 3. The boxplots shown in Figs. 5(a) and 5(b) represent the distributions of the V2V links' data rates of several simulation runs. Besides, the dashed black horizontal line shows the minimum required data rate (c_0), which all the links must satisfy. Taking a closer look at Fig. 5(a) and Fig. 5(b) reveals the reason why the *Max-Min* had a higher reliability compared to the other algorithms. For both the cases of $\gamma_0 = 5$ dB and $\gamma_0 = 15$ dB, this algorithm upholds all the V2V links rate higher than the minimum rate (c_0) by a large margin, fulfilling the outage constraint more effectively and resulting in higher reliability factors. Furthermore, a slight change in the performance can be seen when $\gamma_0 = 15$ dB. The reason for this is that as γ_0 is increased, the feasible set contracts, so that the same optimal points that resulted in a uniform distribution of rates for all V2V connections when γ_0 was 5 dB are no longer feasible.

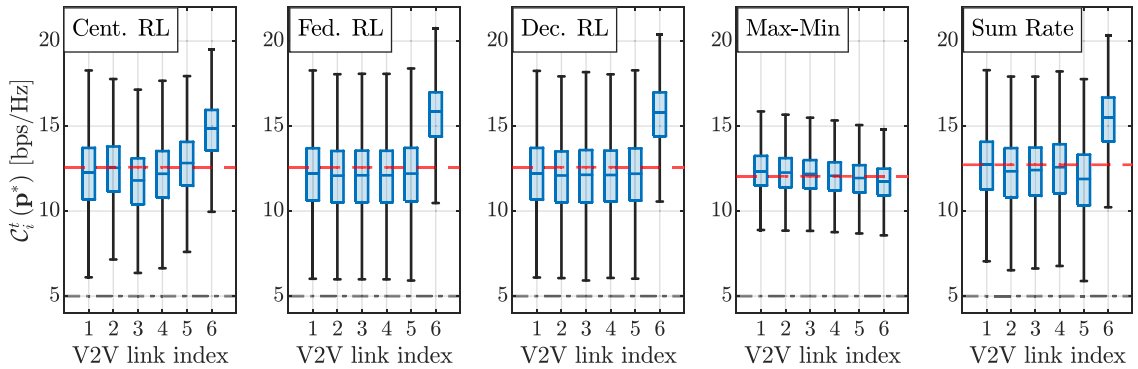
Interesting results can also be seen for the *Sum Rate* algorithm by comparing Figs. 5(a) and 5(b). By looking into the lower whisker of link 5 in Fig. 5(a), one can see that the *Sum Rate* algorithm, unlike *Max-Min*, has increased the link 5 rates up to exactly $c_0 = 5$ bps/Hz. The rationale is that $\gamma_0 = 5$ dB translates into a capacity of approximately 2 bps/Hz which is less than the minimum capacity requirement of $c_0 = 5$ bps/Hz. Therefore, the outage probability constraint has less impact on the optimization problem solution and the *Sum Rate* increases the 5th link's rate up to a level that satisfies the minimum capacity. However, by increasing γ_0 to 15 dB, Fig. 5(b) demonstrates that this algorithm has shifted its focus more to increasing this link's data rate as the outage constraint's (7b) impact is now more pronounced. Further, we can see that link 6 has the highest data rate compared to the other links. The reason is due to the fact it gets the least interference from the other transmitters as there is always a vehicle blockage on the received signal path (see Fig. 1). This very behavior was also followed by the RL-based algorithms as all the approaches yield higher data rates for the last link. Besides, *Fed. MARL* and *Dec. MARL* exhibit identical behavior. Nonetheless, *Fed. MARL* achieves slightly higher data rates for that last link compared to *Dec. MARL* when $\gamma_0 = 15$ dB, which can be directly related to the slim improvement that FL brings on top of *Dec. MARL*.

Another intriguing result is the variance of the V2V links data rates. From the figure, it can be seen that the *Max-Min* algorithm results in the least variance compared to the other algorithms which can be considered as its robustness. A direct impact of this was shown in the reliability factor of these algorithms in Fig. 4.

To complete our analysis, in Fig. 6 we have compared the average allocated power to the V2V links following the proposed algorithms. This comparison has been overlooked by most of the works in the literature. However, it turns out to be an important ingredient in determining the optimality of the RL-based RRM algorithms. From the figure, it can be seen that the *Max-Min* algorithm has a lower average power consumption compared to the other algorithms. *Max-Min* attempts to balance the data rate distribution in the network by increasing the minimum rate between V2V links, as can be seen in Figs. 5(a) and 5(b). Therefore, this algorithm reduces the power allocated to the vehicles to compensate for the interference between vehicles. However, by increasing γ_0 to 15 dB, the average power has slightly increased, as can be seen from the figure. The reason for this is twofold. First, the total interference power distribution in the platoon is uneven, and second, as mentioned earlier, the last vehicle in the platoon always gets less interference from the other vehicles. In order to balance the rates, *Max-Min* allocates less power to the last V2V link compare to the other links. When $\gamma_0 = 5$ dB, the equilibrium point where all V2V links exhibit similar rates is achievable (Fig. 5(a)). However, when $\gamma_0 = 15$dB, the average transmit power has to be higher compared to the previous case because there is a tighter restriction from the outage constraint that forces the links to have higher SINR values. Therefore, the same reduction in the allocated powers to the V2V links, which was viable earlier, becomes infeasible. This ultimately results in a higher average rate as denoted in Table 3 and non



(a) Per-link data rate analysis with $\gamma_0 = 5$ dB and $N = 7$.



(b) Per-link data rate analysis with $\gamma_0 = 15$ dB and $N = 7$.

FIGURE 5. Fairness analysis of the proposed algorithms in terms of per link capacity. Dashed red lines represent average data rate achieved by each of the proposed algorithms (refer to Table 3) and dashed black lines show the minimum capacity requirement $c_0 = 5$ bps/Hz, respectively.

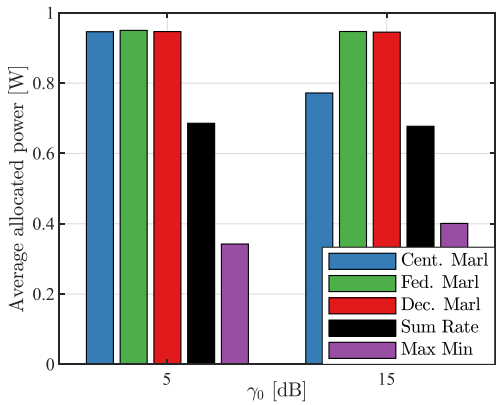


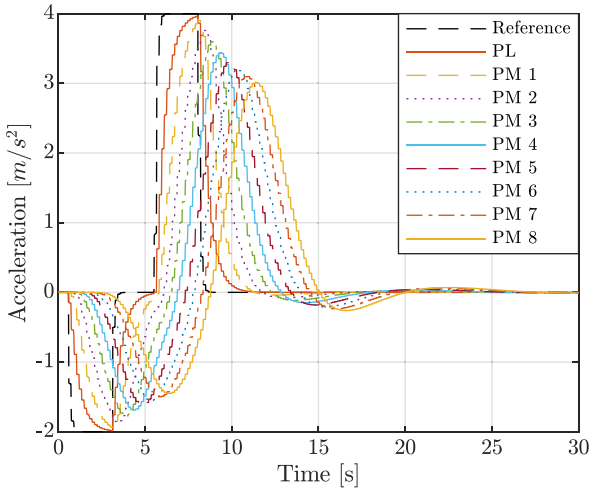
FIGURE 6. Average allocated power to V2V links following the proposed algorithms with varying γ_0 threshold, assuming $N = 7$.

uniform distribution of the rates in the platoon. On the other hand, *Sum Rate* results in higher power consumption compared to *Max-Min* since the purpose is to maximize the total rate of V2V links. When $\gamma_0 = 15$ dB, due to the outage constraint, the last V2V link has slightly lower rates in this case compared to the case when $\gamma_0 = 5$ dB; whereas link 5 receives higher data rates. Overall, *Sum Rate* algorithm shows impervious behavior to the change in the γ_0 levels and keeps the average rate of the V2V links similar, as shown in the Table 3.

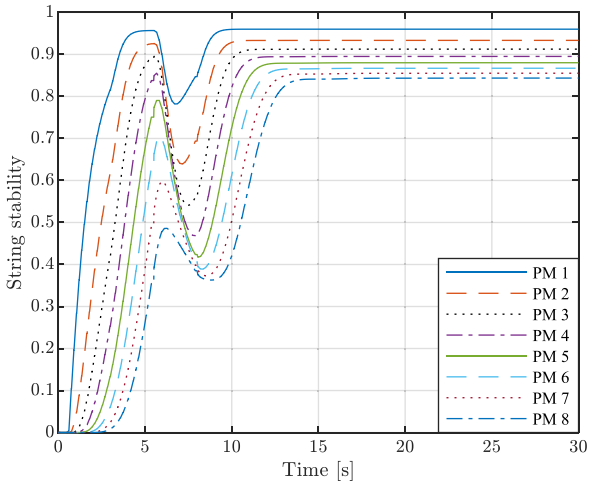
Coming to RL-based algorithms, only *Cent. RL* has shown some sensitivity to the increase in the γ_0 levels. From the figure, *Cent. RL* has decreased the power consumption to reduce the interference levels and satisfy the constraints. However, both *Fed. MARL* and *Dec. MARL* keep the average power consumption more than 0.9 W and no sensitivity is shown when the γ_0 is changed to 15 dB. This behavior is logical to some extent. Since we assume local knowledge of channel gains and interference power in these two algorithms, the strategy is forced to maximize the rate by keeping the power as high as possible to mitigate the effects of interference. As long as there is no information sharing between vehicles, this learned strategy will not change. The very purpose of federated learning was to promote this information exchange; however, the results have shown no impact.

E. STRING STABILITY EVALUATION

In the following, we show the platoon's time domain analysis during 30 seconds of simulation time. Figs. 7(a) and 7(b) plot the acceleration and string stability of the PMs with respect to the preceding PM. These two plots hold true for all the proposed algorithms, i.e., both AI- and non-AI-based RRM solutions. The reference acceleration has been shown with a dashed black line in Fig. 7(a), which is expected for



(a) Acceleration change. $N = 9$ and $\gamma_0 = 15$.



(b) String stability of platoon with $N = 9$ and $\gamma_0 = 15$.

FIGURE 7. Acceleration and string stability change of the platoon's vehicles. The results are similar for all the proposed algorithms.

all the PMs to follow. Fig. 7(b) has been derived from the time domain analysis of string stability, i.e.,

$$\frac{\|u_i(J(t))\|_{\mathcal{L}_2}}{\|u_{i-1}(J(t))\|_{\mathcal{L}_2}} = \frac{\sqrt{\sum_{j=0}^{J(t)/\Delta} \int_{t_j}^{t_{j+1}} u_i^2(t)dt}}{\sqrt{\sum_{j=0}^{J(t)/\Delta} \int_{t_j}^{t_{j+1}} u_{i-1}^2(t)dt}}, \quad (39)$$

where $J(t)$ determines up to which time slot the string stability is calculated. A change in the acceleration profile of the platoon ultimately results in an increase or decrease in the vehicle speed and inter-vehicle spacing and therefore changes the wireless channel behavior between the vehicles. To compensate for these modifications, proper resource allocation strategies must be implemented with quick adaptability to the change in the wireless channel conditions so that the string stability of the platoon is preserved. From the figures it can be seen that the proposed resource allocation schemes have facilitated the timely exchange of control packets between the vehicles. For this reason, the string stability criterion for all vehicles in Fig. 7(b) has remained below one, which satisfies eq. (2). The change in the string stability quantity is due to the increase and decrease in the acceleration profile of the vehicles. One can infer from these plots that AI-based algorithms can bring about string-stable platooning even though they lead to suboptimal solutions in terms of achievable capacity (refer to Fig. 3).

F. DISCUSSION

In this subsection, we will attempt to consolidate our observations from the comparisons between classical and RL-based RRM approaches. For the sake of transparency, we have divided our discussion into the following parts.

- *Complexity:* As we have seen in Fig. 2, both the classical and RL-based algorithms require multiple iterations to converge to the final solution. The key point here is that one needs to run the classical model-based algorithms every time the channel gains or the environment changes, whereas with RL, the trained models can be used after training is complete without the need for retraining. This is indeed a major advantage of ML, which enables an autonomous response to the changes in the environment. The individual complexity of each iteration is also important. Although classical methods may require multiple iterations, the complexity of each iteration can be quite low, so the overall complexity can still be comparable to ML. Therefore, the development of ML models with simpler designs is of utmost importance.
- *Optimality:* Figs. 3 and 4 showed that model-based approaches using d.c. programming, as expected, outperform RL-based algorithms. Overall in this work, RL leads to suboptimal solutions regardless of whether the state information is fully or only partially known. Further, we noticed that *Centralized RL* can outperform distributed architectures in the case study we had in this paper. One of the goals of this paper was to also find out whether integrating FL with RL can bring any benefits in terms of better RRM policies. The results showed that both decentralized and FL-based MARL algorithms deliver similar performance. Although it is not proved mathematically, the results highlight that no extra gains are attainable through FL inclusion in RL algorithms. Overall, we concluded that fully distributed MARL algorithms can bring about string-stable platooning even though they lead to suboptimal solutions in terms of achievable capacity. Caution is needed in interpreting the results here. The MARL design that is mostly studied in the current literature assumes centralized training and distributed execution [57], [58], which may ultimately lead to similar strategies as the centralized case. The distributed MARL we studied assumes no information sharing between the vehicles to mimic the exact situation in real scenarios.

- *Sensitivity:* As we have shown in Figs. 5 and 6, RL is less sensitive to the changes in the constraints, unlike the model-based approaches. This impervious behavior has cast a shadow over RL-based algorithms and yet has rarely been mentioned by the current literature.
- *Robustness:* As we have shown in Fig. 7, all the RL-based solutions, including *Fed. MARL* successfully provide a string-stable platooning vehicle network, although, as mentioned, the learned policies are sub-optimal compared to the conventional model-based approaches (refer to Fig. 3). The correlation between the trustworthiness and robustness of an algorithm appears to be closely related to the particular application for which it is intended [59]. In the context of our study, the prescribed latencies of the platooning control system facilitated the adaptability of RL methods, enabling effective problem solving capabilities.
- *Signaling overhead:* All the mentioned algorithms need full or partial knowledge of the total CSI. *Sum Rate*, *Max-Min*, and *Cent. RL* require all the V2V links CSI. As the BS is mainly in charge of acquiring this information, estimation of the uplink CSI is also necessary. These schemes introduce challenges when there is high mobility in the environment, especially in V2X networks. Due to the fast movement of vehicles which leads to high Doppler effects, the estimated uplink CSI can have high error. To reduce this, frequent uplink channel information is required. Upon accurate estimation, the centralized algorithms can decode the V2V links CSI, and run the specified algorithms to allocate the required powers to fulfill (7) and (8). Then, it is also necessary that the allocated power is sent back to the vehicles via feedback channels. Overall, while centralized algorithms can provide optimal solutions, they result in higher latency and signaling overhead. On the other hand, distributed algorithms like *Dec. MARL*, require no interaction with the BS, and both the training and evaluation are performed in a distributed manner, though the solutions may be inaccurate. Not to mention, unlike the uplink CSI estimation, the V2V links CSI estimation is less error-prone as the vehicle relative velocity in the platoon is quite small hence introducing less Doppler. *Fed. MARL* is considered a semi-distributed algorithm where the training has to be done centrally, introducing the same issues as other centralized algorithms. In conclusion, taking the overhead issue into account, it seems developing robust distributed algorithms is beneficial, especially for V2X networks.

From the preceding analysis, it can be concluded that while RL can lead to suboptimal solutions, it also offers some advantages, especially in terms of complexity and autonomous response to environmental changes. It is yet too early to make definitive conclusions on the performance of RL and further studies need to be conducted to fill the gap mentioned in this article.

V. CONCLUSION

In this work, we formulated RRM problem for a platooning vehicle network under the string-stable CACC design scheme. To shed light on the efficacy of ML-based RRM algorithms, we solved the problem of interest using both classical and MARL-based strategies. First, we developed model-based classical approaches based on the d.c. representation of the objective function. Then, we modeled the RRM problem as MDP and used both centralized and decentralized MARL approaches to solve the problem. The simulation results showed that RL-based resource allocation algorithms are not able to provide similar performance and robustness as the classical approaches. Moreover, the analysis showed that the combination of FL and MARL is not sufficient to achieve a better estimation of the optimal RRM strategies. However, under realistic conditions where a constant estimation of channel gains is required, classical solutions can lead to a large overhead on the BS side. This is more challenging in vehicular networks, where channel gains change rapidly due to the dynamic movement of vehicles. In these cases, depending on the application of interest, MARL, although sub-optimal, can lead to satisfactory results, as was the case in this work.

Further investigations to close the gap between RL-based solutions and classical approaches should be the subject of future studies. In parallel, the possibility of designing ML models with low complexity to further reduce training and inference time (as was the case in the study here and [60]) must also be the subject of future works.

APPENDIX

A. MATI CALCULATION

The methodology used here to derive the MATI, follows the exact same steps stipulated in [5] and [46]. First, we equate the string stability condition in (2) to 1 as we are interested in obtaining the maximum communication delay. Then, after substituting the defined transfer functions, and substituting $s = j\omega$, one can express the magnitude of the string stability transfer function as

$$|\Lambda_i(s)|_\infty = \frac{\sqrt{\left(\mathrm{Re}\left(N_{\Lambda_i}(j\omega)\right)\right)^2 + \left(\mathrm{Im}\left(N_{\Lambda_i}(j\omega)\right)\right)^2}}{\sqrt{\left(\mathrm{Re}\left(D_{\Lambda_i}(j\omega)\right)\right)^2 + \left(\mathrm{Im}\left(D_{\Lambda_i}(j\omega)\right)\right)^2}} = 1, \tag{40}$$

where N_{Λ_i} and D_{Λ_i} refer to the numerator and denominator of the string stability transfer function $\Lambda_i(s)$. By reordering the terms, one can get

$$\left(\mathrm{Re}\left(N_{\Lambda_i}(j\omega)\right)\right)^2 + \left(\mathrm{Im}\left(N_{\Lambda_i}(j\omega)\right)\right)^2. - \left(\mathrm{Re}\left(D_{\Lambda_i}(j\omega)\right)\right)^2 - \left(\mathrm{Im}\left(D_{\Lambda_i}(j\omega)\right)\right)^2 = 0. \tag{41}$$

After some algebraic manipulations, (41) can be rearranged as a polynomial of the transmission interval term T as

$$T_i^p c_p(\omega) + T_i^{p-1} c_{p-1}(\omega) + \cdots c_0(\omega) = 0, \tag{42}$$

where c_p are themselves polynomial functions in ω. Finally, the optimal value for MATI (denoted as $t^*_{\Lambda_i}$ hereinafter) corresponds to the least positive real root obtained from the resultant polynomial (42) with $\omega = \omega_n$, where ω_n is the system natural frequency. To avoid a string unstable system, the following relationship must hold between the communication delay $\delta_{\mathrm{Tx},i}$ and $t^*_{\Lambda_i}$ as

$$\delta_{\mathrm{Tx},i} \leq t^*_{\Lambda_i}.$$

ACKNOWLEDGMENT

The authors alone are responsible for the content of the paper.

REFERENCES

[1] M. H. C. Garcia et al., "A tutorial on 5G NR V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1972–2026, 3rd Quart., 2021.

[2] M. Harounabadi, D. M. Soleymani, S. Bhadauria, M. Leyh, and E. Roth-Mandutz, "V2X in 3GPP Standardization: NR sidelink in release-16 and beyond," *IEEE Commun. Stand. Mag.*, vol. 5, no. 1, pp. 12–21, Mar. 2021.

[3] D. Jia, K. Lu, J. Wang, X. Zhang, and X. Shen, "A survey on platoon-based vehicular cyber-physical systems," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 263–284, 1st Quart., 2016.

[4] G. J. L. Naus, R. P. A. Vugts, J. Ploeg, M. J. G. van de Molengraft, and M. Steinbuch, "String-stable CACC design and experimental validation: A frequency-domain approach," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4268–4279, Nov. 2010.

[5] A. González, "Analytic approaches for obtaining the communications requirements of string stable Platooning," Ph.D. dissertation, Dept. Elect. Comput. Eng., Tech. Univ. Dresden, Dresden, Germany, Mar. 2022.

[6] K. Sehla et al., "Resource allocation modes in C-V2X: From LTE-V2X to 5G-V2X," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8291–8314, Jun. 2022.

[7] "Study on NR vehicle-to-everything (V2X); (Release 16)," 3GPP, Sophia Antipolis, France, Rep. TR 38.885, Mar. 2019.

[8] "Overall description of radio access network (RAN) aspects for vehicle-to-everything (V2X) based on LTE and NR; (Release 17)," 3GPP, Sophia Antipolis, France, Rep. TR 37.985, 2022.

[9] F. Tang et al., "Comprehensive survey on machine learning in vehicular network: Technology, applications and challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 2027–2057, 3rd Quart., 2021.

[10] X. Lyu et al., "Contrasting centralized and decentralized critics in multi-agent reinforcement learning," 2021, *arXiv:2102.04402*.

[11] J. Posner, L. Tseng, M. Aloqaily, and Y. Jararweh, "Federated learning in vehicular networks: Opportunities and solutions," *IEEE Netw.*, vol. 35, no. 2, pp. 152–159, Mar./Apr. 2021.

[12] P. Xiang et al., "Multi-agent RL enables decentralized spectrum access in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10750–10762, Oct. 2021.

[13] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, May 2019.

[14] 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.

[15] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[16] S. B. Prathiba, G. Raja, S. Anbalagan, K. Dev, S. Gurumoorthy, and A. P. Sankaran, "Federated learning empowered computation offloading and resource management in 6G-V2X," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 5, pp. 3234–3243, Sep./Oct. 2022.

[17] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang, and H. Du, "Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 8, pp. 8810–8824, Aug. 2022.

[18] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.

[19] X. Zhang, M. Peng, S. Yan, and Y. Sun, "Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6380–6391, Dec. 2020.

[20] H. Zhang et al., "Mean-field-aided multiagent reinforcement learning for resource allocation in vehicular networks," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2667–2679, Feb. 2023.

[21] Z. Guo, Z. Chen, P. Liu, J. Luo, X. Yang, and X. Sun, "Multi-agent reinforcement learning-based distributed channel access for next generation wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1587–1599, May 2022.

[22] Y. Yuan, G. Zheng, K.-K. Wong, and K. B. Letaief, "Meta-reinforcement learning based resource allocation for dynamic V2X communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 8964–8977, Jul. 2021.

[23] 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.

[24] A. S. Kumar, L. Zhao, and X. Fernando, "Multi-agent deep reinforcement learning-empowered channel allocation in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1726–1736, Feb. 2022.

[25] J. Tian, Q. Liu, H. Zhang, and D. Wu, "Multiagent 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.

[26] Y. Cui, H. Shi, R. Wang, P. He, D. Wu, and X. Huang, "Multi-agent reinforcement learning for slicing resource allocation in vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 2005–2016, Feb. 2024.

[27] B. Gu, W. Chen, M. Alazab, X. Tan, and M. Guizani, "Multiagent reinforcement learning-based semi-persistent scheduling scheme in C-V2X mode 4," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 12044–12056, Nov. 2022.

[28] C. Guo, C. Wang, L. Cui, Q. Zhou, and J. Li, "Radio resource management for C-V2X: From a hybrid Centralized-distributed scheme to a distributed scheme," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1023–1034, Apr. 2023.

[29] W. Qi, Q. Song, L. Guo, and A. Jamalipour, "Energy-efficient resource allocation for UAV-assisted vehicular networks with spectrum sharing," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7691–7702, Jul. 2022.

[30] M. Parvini, M. R. Javan, N. Mokari, B. Abbasi, and E. A. Jorswieck, "AoI-aware resource allocation for platoon-based C-V2X networks via multi-agent multi-task reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 9880–9896, Aug. 2023.

[31] L. Lei, T. Liu, K. Zheng, and L. Hanzo, "Deep reinforcement learning aided platoon control relying on V2X information," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 5811–5826, Jun. 2022.

[32] 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.

[33] D. Chen, K. Zhang, Y. Wang, X. Yin, Z. Li, and D. Filev, "Communication-efficient decentralized multi-agent reinforcement learning for cooperative adaptive cruise control," 2023, *arXiv:2308.02345*.

[34] T. Zeng, O. Semiari, 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.

[35] M. Parvini, A. Gonzalez, A. Villamil, P. Schulz, and G. Fettweis, "Joint resource allocation and string-stable CACC design with multi-agent reinforcement learning," in *Proc. IEEE Int. Conf. Commun.*, Rome, Italy, 2023, pp. 1232–1237.

[36] R. Lian, Z. Li, B. Wen, J. Wei, J. Zhang, and L. Li, "Multiagent deep reinforcement learning for automated truck Platooning control," *IEEE Trans. Intell. Transp. Syst. Mag.*, vol. 16, no. 1, pp. 116–131, Jan./Feb. 2024.

[37] T. Liu, L. Lei, Z. Liu, and K. Zheng, "Jointly learning V2X communication and platoon control with deep reinforcement learning," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Toronto, ON, Canada, 2023, pp. 1–6.

[38] T. Liu, L. Lei, K. Zheng, and K. Zhang, "Autonomous platoon control with integrated deep reinforcement learning and dynamic programming," *IEEE Internet Things J.*, vol. 10, no. 6, pp. 5476–5489, Mar. 2023.
[39] Z. Huang, X. Xu, H. He, J. Tan, and Z. Sun, "Parameterized batch reinforcement learning for longitudinal control of autonomous land vehicles," *IEEE Trans. Syst., Man, Cybern, Syst.*, vol. 49, no. 4, pp. 730–741, Apr. 2019.
[40] Y. Lin, J. McPhee, and N. L. Azad, "Longitudinal dynamic versus kinematic models for car-following control using deep reinforcement learning," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Auckland, New Zealand, 2019, pp. 1504–1510.
[41] Y. Lin, J. McPhee, and N. L. Azad, "Comparison of deep reinforcement learning and model predictive control for adaptive cruise control," *IEEE Trans. Intell. Veh.*, vol. 6, no. 2, pp. 221–231, Jun. 2021.
[42] R. Yan, R. Jiang, B. Jia, J. Huang, and D. Yang, "Hybrid car-following strategy based on deep deterministic policy gradient and cooperative adaptive cruise control," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 2816–2824, Aug. 2022.
[43] T. Liu, L. Lei, K. Zheng, and X. Shen, "Multi-timescale control and communications with deep reinforcement learning—Part-I: Communication-aware vehicle control," *IEEE Internet Things J.*, early access, Dec. 29, 2023, doi: 10.1109/JIOT.2023.3348590.
[44] L. Lei et al., "Multi-timescale control and communications with deep reinforcement learning—Part-II: Control-aware radio resource allocation," *IEEE Internet Things J.*, early access, Dec. 29, 2023, doi: 10.1109/JIOT.2023.3348594.
[45] H. Kha, H. D. Tuan, and H. H. Nguyen, "Fast global optimal power allocation in wireless networks by local D.C. programming," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 510–515, Feb. 2012.
[46] S. Öncü, "String stability of interconnected vehicles: Network-aware modelling, analysis and experiments," Ph.D. dissertation, Dept. Mech. Eng., Eindhoven Univ. Technol., Eindhoven, The Netherlands, Jan. 2014.
[47] S. Kandukuri and S. Boyd, "Optimal power control in interference-limited fading wireless channels with outage-probability specifications," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 46–55, Jan. 2002.
[48] S. Wu, "Some results on extending and sharpening the weierstrass product inequalities," *J. Math. Anal. Appl.*, vol. 308, no. 2, pp. 689–702, 2005.
[49] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. press, 2004.
[50] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[51] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
[52] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn.*, Bejing, China, 2014, pp. 387–395.
[53] X. Xu, R. Li, Z. Zhao, and H. Zhang, "The gradient convergence bound of federated multi-agent reinforcement learning with efficient communication," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 507–528, Jan. 2024.
[54] M. Krouka, A. Elgabli, C. B. Issaid, and M. Bennis, "Communication-efficient and federated multi-agent reinforcement learning," vol. 8, no. 1, pp. 311–320, Mar. 2022.
[55] "Study on evaluation methodology of new vehicle-to-everything V2X use cases for LTE and NR; (Release 15)," 3GPP, Sophia Antipolis, France, Rep. TR 37.885.
[56] "Enhancement of 3GPP support for V2X scenarios; stage 1; (Release 17)," 3GPP, Sophia Antipolis, France, Rep. TS 22.186, Mar. 2022.
[57] R. Lowe, Y. Wu, A. Tamar, J. H. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2017, pp. 6382–6393.
[58] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning Control*. Cham, Switzerland: Springer, 2021, pp. 321–384.
[59] H. Boche, A. Fono, and G. Kutyniok, "Mathematical algorithm design for deep learning under societal and judicial constraints: The algorithmic transparency requirement," 2024, *arXiv:2401.10310*.
[60] M. Q. Khan, A. Gaber, M. Parvini, P. Schulz, and G. Fettweis, "A low-complexity machine learning design for mmWave beam prediction," 2023, *arXiv:2310.19323*.

MOHAMMAD PARVINI (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 pursuing the Ph.D. degree in communication systems with the Technical University of Dresden. His research includes wireless communication systems, signal processing, optimization, and machine learning.

PHILIPP SCHULZ received the M.Sc. degree in mathematics and the Ph.D. (Dr.-Ing.) degree in electrical engineering from Technische Universität Dresden, Germany, in 2014 and 2020, respectively, where he was a Research Assistant in the field of numerical mathematics, modeling, and simulation. In 2015, he joined TU Dresden's Vodafone Chair for Mobile Communications Systems and became a member of the System-Level Group. After more than one year at the Barkhausen Institut, Dresden, Germany, where he studied rateless codes in the context of multi-connectivity, he is currently a Research Group Leader with the Vodafone Chair and focuses on resilience of wireless communications systems. His research interests include flow-level modeling and the application of queuing theory on communications systems with respect to ultrareliable low-latency communications.

GERHARD FETTWEIS (Fellow, IEEE) received the Ph.D. degree under the supervision of H. Meyr from RWTH Aachen University in 1990. After postdoctoral work with IBM Research, San Jose, CA, USA, he joined TCSI, Berkeley, USA. Since 1994, he has been the Vodafone Chair Professor with Technische Universität Dresden. Since 2018, he has also been headed the new Barkhausen Institute. In 2019, he was elected into the DFG Senate (German Research Foundation). He researches wireless transmission and chip design, coordinates 5G Lab Germany, has spun out 17 tech and three non-tech startups. He is a member of the German Academy of Sciences (Leopoldina) and the German Academy of Engineering (acatech).