Intelligent Resource Management Based on Reinforcement Learning for Ultra-Reliable and Low-Latency IoV Communication Networks

Helin Yang, *Student Member, IEEE*, Xianzhong Xie, *Member, IEEE*, and Michel Kadoch, *Senior Member, IEEE*

***Abstract*—Internet of Vehicles (IoV) has attracted much interest recently due to its ubiquitous message exchange and content sharing among smart vehicles with the development of the mobile communication and computation technologies. In this paper, we investigate the policy for jointly communication mode selection, resource block assignment, and power control in device-to-device-enabled vehicle-to-vehicle (V2V) based IoV communication networks with the purpose of guaranteeing the strict ultra-reliable and low latency communications requirements of V2V links while maximizing the sum capacity of vehicle-to-infrastructure links. Considering the unknown environments dynamics as well as the continuous-valued state and action space in IoV networks, we exploit a decentralized actor-critic reinforcement learning model with a new reward function to learn the policy by interacting with the environment. Moreover, we propose an efficient transfer actor-critic learning (ETAC) approach to effectively enhance the learning efficiency and improve the learning convergence speed, in order to support reliable and delay-sensitive vehicular services in IoV networks. Simulation results show that the proposed ETAC approach can effectively reduce the generated interference in IoV networks and ensure the latency and reliability requirements of V2V link, as well as achieve the fast convergence speed and high convergence stability, compared with other existing approaches.**

***Index Terms*—Internet of Vehicles, ultra-reliable and low latency communications, device-to-device, resource management, transfer reinforcement learning, actor-critic.**

I. Introduction

With the rapid development of smart vehicles, Internet of vehicles (IoV) has been emerging as a promising vision for the road safety, collision avoidance, and information transmission services in intelligent transportation systems [1]–[5]. With the help of IoV, smart vehicles can be interconnected to the internet and intelligently exchange information and contents among smart vehicles with the minimal human interaction [6]. It also applies the mobile communication technology to realize the inter-communication and coordination among vehicle-to-everything (V2X), such as vehicle-to-infrastructure (V2I) communications and vehicle-to-vehicle (V2V) communications, which can greatly improve the traffic safety and efficiency, and reduce the unfortunate incidents and congestion of the road traffic [7]. In IoV networks [8], the information applications and messages sharing require frequent access to the servers and the internet, which involve the considerable amount of transmission data rate provided by V2I links with high-capacity requirements, and the safety-critical information exchange supported by V2V links with the strict ultra-reliable and low latency communications (URLLC) requirements [9], [10]. For example, the METIS reports that a maximum latency of 5 ms, and with the transmission reliability of 99.999% of 1600 bytes message sizes in V2V links [11]. Hence, there is a strong desire for searching efficient resource management approaches to support the above communication requirements in IoV networks.

Recently, various resource allocation approaches have been proposed in IoV communication networks, in order to ensure the quality-of-service (QoS) or quality of experience (QoE) requirements of vehicle users (VUEs) and effectively improve the network capacity. The literatures [12]–[14] presented the optimal resource allocation approaches to improve the network performance (e.g., transmission reliability, resource utilization, road safety, etc.) in cloud-enabled vehicular networks. The authors in [15] and [16] addressed the multiple access problem for vehicular sensor networks with considering the latency- sensitive requirements. In [17] and [18], the metric of efficient resource allocation and interference management was studied in cellular V2X communications, which can achieve a higher link rate and greater link reliability. Moreover, many software -defined networking-based resource allocation schemes were presented to utilize the VUEs states information to guarantee various VUEs QoS or QoE [19], [20].

Device-to-device-enabled V2V (D2D-V2V) communication is identified as a promising candidate for IoV networks with the high efficient and reliable properties [9], [10], [18], [21]–[25], because D2D-V2V communication has the ability to reduce transmission latency and power consumption by directly providing local message dissemination, and enhance the spectrum efficiency due to the proximity and reusing gain [24], [26]. Zhou

Manuscript received November 25, 2018; accepted December 28, 2018. Date of publication January 1, 2019; date of current version May 28, 2019. This work was supported in part by the National Nature Science Foundation of China under Grant 61502067, and in part by the Key Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJZD-K201800603. The review of this paper was coordinated by the Guest Editors of the Special Section on Machine Learning-Based Internet of Vehicles. *(Corresponding author: Xianzhong Xie.)*

H. Yang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: hyang013@e.ntu.edu.sg).

X. Xie is with the Chongqing Key Lab of Computer Network and Communication Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: xiexzh@cqupt.edu.cn).

M. Kadoch is with the École de Technologie Supérieure, University of Quebec, Montreal, QC H3C 1K3, Canada (e-mail: michel.kadoch@etsmtl.ca).

Digital Object Identifier 10.1109/TVT.2018.2890686

et al. [3] investigated the combination of both the physical layer and social layer to achieve the rapid message dissemination in D2D-V2V-based IoV networks. The literature [21] investigated the problem of maximizing the individual data rate within the D2D-assisted IoV for mobile health by considering different levels of emergency. In addition, for D2D-V2V-based IoV networks, the studies [6], [22], [23] jointly considered and optimized the performance of both V2V and V2I communications under a wide variety of intelligent vehicular applications, and the outage probability performance of D2D-V2V links was performed in [24]. In [18], a matrix game theoretical scheme was presented to solve the resource management optimization problem in D2D-based vehicular networks. The above works [18], [21]–[25] mainly pay attention to the physical layer information, and the different kinds of QoS requirements (e.g., strict latency, reliability, etc.) of V2V links has not been well investigated in IoV networks.

There exist many studies on the resource management in D2D-based V2X or IoV networks with considering the stringent latency and reliability constraints [5], [9], [27]–[33]. The resource block scheduling and power control algorithms were proposed to satisfy the latency and reliability requirements of V2V links [9], [27], [28]. In order to enhance the network performance, more researchers have started to joint investigate both the V2V and V2I communications [5], [9], [29]–[33], specifically, to improve the capacity performance of V2I links while satisfying the latency and reliability constraints of V2V links. Silva *et al.* [5] investigated the ethical implications of broader social IoV systems, and analyzed the enhancement of the road safety and traffic efficiency with the autonomous decision making of V2I and V2V links. Liang *et al.* proposed two resource allocation approaches to maximize the sum capacity of V2I links while satisfying the reliability constraints of V2V links in D2D-based V2X communication networks [9], [29]. Furthermore, a joint mode selection and resource allocation scheme was proposed to maximize the sum capacity of Pedestrian users and V2V links [30]. Even the above works [5], [9], [27]–[33] can achieve the considerable performance, they are not intelligent enough to efficiently manage the network resource among smart vehicles with high-level intelligence in IoV networks.

Reinforcement learning (RL) is one of the most powerful machine learning tools for policy control and intelligent decision making [34], which has been widely adopted in wireless communications [35]–[37]. Recently, many works have applied the reinforcement learning tool (e.g., Q-leaning and deep Q-leaning) to solve the intelligent resource management and decision making problem in V2X or IoV networks [38]–[45]. Liu *et al.* [38] proposed a distributed cooperative reinforcement learning approach to manage the traffic with the help of the V2X networks dynamic clustering design, in order to balance the traffic load. In [39], a deep Q-network approach was proposed to improve both the traffic safety and satisfy VUEs QoS requirements in a green IoV networks. The literature [4] combined the RL and software defined network to realize the cognitive capability for IoV networks, in order to search the optimal routing policy under dynamic environments. The Q-learning is adopted in [40]–[42] to realize intelligent learning strategies to improve the V2V communication performance by interacting with the vehicular environments. The literatures [43], [44] presented some case studies that how to apply the RL tool to manage network resources in intelligent V2X networks. Moreover, Ye *et al.* [45] proposed a decentralized resource allocation approach in V2X networks based on a multi-agent deep Q-learning with considering the latency constraints of V2V links. Unfortunately, most popular optimization technologies [5], [9]–[11], [18], [21]–[33] are not suitable for IoV communication networks due to the unknown dynamic vehicular environment and the high intelligent and adaptive scheduling demands of IoV. In addition, the above works [38]–[45] based on Q-learning or deep Q-learning are capable of intelligently manage the network resource in IoV networks, but Q-learning has low convergence speed and the deep Q-learning framework may not always suitable to deal with continuous-valued state and action spaces in IoV networks.

According to the above analysis, in this paper, an efficient transfer actor-critic (AC) RL approach is proposed to address the intelligent resource management problem in D2D-V2V-based IoV networks, in order to maximize the network capacity while guaranteeing the strict URLLC requirements of V2V links. The AC approach can efficiently deal with the continuous-valued state and action spaces (e.g., channel information, transmit power, etc.), where the actor is used to exploit the stochastic actions and the critic is applied to estimate the state-action value function. The main contributions of our work can be summarized as follows:

- This paper firstly presents a joint mode selection, resource block (RB) assignment and transmit power control scheme to maximize the overall network capacity in D2D-V2V-based IoV networks while satisfying the QoS requirements of V2I links, and ensuring the strict URLLC requirements of V2V links. In addition, the decision making problem is formulated as a decentralized RL framework, thus V2I and V2V links are capable of intelligently making their adaptive decisions to improve their performance based on the instantaneous observations under high dynamic vehicular environments.
- In order to support reliable and delay-sensitive vehicular applications in D2D-V2V-based IoV networks, an efficient transfer AC (ETAC) learning approach is proposed to learn the optimal policy for the intelligent resource management under the continuous-valued state and action variables. In detail, in IoV networks, the new VUE or the existing VUE with poor leaning performance can utilize the transferred information from the expert VUE to model its own learning framework with a small transmission overhead, which can effectively enhance the learning efficiency and improve the learning convergence speed. Moreover, the eligibility trace mechanism is adopted to achieve the better convergence property and the advantage function is applied to avoid the high policy gradient variance.
- We evaluate and analyze the performance of our proposed ETAC approach from the perspectives of the convergence, optimality and stability. In addition, we compare it with the classical AC learning, the Q-learning and the random search approaches from perspectives of sum rate, the probability of satisfied V2I and V2V links under different scenarios. Simulation results prove that our proposed ETAC

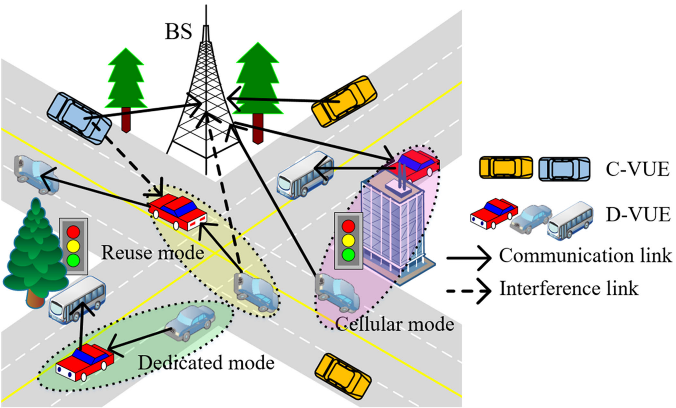


Fig. 1. The D2D-V2V-based IoV communication network.

approach can efficiently deploy intelligent resource management for dynamic IoV networks.

- Non-coordinated resource allocation (denoted as non-coordinated RA): Each cell selfishly optimize its joint subcarrier and power allocation strategy to maximize its data rate while guaranteeing the practical constraints in each cell without coordinating with other cells.

The rest of this paper can be organized as follows. Section II shows the system model and formulates the optimization problem. In Section III, we model the resource management problem as a Markov decision making process and the AC learning approach is adopted to address it. The ETAC learning approach is presented in this section, too. The simulation results and analysis are presented in Section IV. Finally, Section V concludes the paper.

II. System Model and Problem Formulation

A. System Model

We consider a D2D-V2V-based IoV network, as shown in Fig. 1, the IoV communications can be mainly divided into two parts: V2I communication between VUEs and the base station (BS), and V2V communication among VUEs on the road which is a form of D2D communication by directly communicating with each other when two VUEs are close to each other, called D2D-V2V communication. All smart VUEs have the ability to choose both V2I and V2V communication connections based on their requirements.

The D2D-V2V-based IoV network consists K cellular VUEs (denoted C-VUEs) in the V2I communication form and M D2D-V2V pairs (The corresponding VUEs are denoted by D-VUEs) in the D2D-V2V communication form, where each VUE is equipped with one single antenna. The C-VUE set and D2D-V2V pairs set are $\mathcal{K} = \{1, 2, \ldots, K\}$ and $\mathcal{M} = \{1, 2, \ldots, M\}$, respectively. The network adopts the orthogonal frequency division multiple access (OFDMA) technique to serve VUEs in the uplink scenario, and there are a set $\mathcal{N} = \{1, 2, \ldots, N\}$ of $N = K + N_U$ orthogonal RBs being allocated to the VUEs, where K and N_U are the number of used RBs for C-VUEs and the number of current unused RBs. We assume that each C-VUE is only allocated with one RB, and its RB can be reused at most one D2D-V2V pair. Each D2D-V2V pair can only reuse one RB of the C-VUE.

Let h_k and h_m denote the channel gains of the desired transmission for C-VUE k and D2D-V2V pair m, respectively. Let $g_{k,m}$ denotes the interference channel gain between C-VUE k and the receiver of the m-th D2D-V2V pair. Let $g_{k,m,B}$ denotes the interference channel gain from the transmitter of the m-th D2D-V2V pair to the BS on the k-th C-VUE' RB. We assume that all the above channel gains include the path loss, shadowing fading and small-scale fading [9].

B. D2D-V2V Communication Modes

In the D2D-V2V-based IoV network, each D2D-V2V pair can select one of the following three communication modes, as shown in Fig. 1.

1) Reuse Mode: In this mode, two D-VUEs communicate directly by utilizing the uplink RBs resource of C-VUEs when the two D-VUEs are close with each other. In this case, even the spectrum efficiency can be enhanced, the interference between the D2D-V2V pair and the C-VUE is incurred. In the reuse mode, the C-VUE will suffer the interference from the D2D-V2V pair when it shares its RB resource with the D2D- V2V pair. Then, the received instantaneous uplink signal-to interference-plus-noise ratio (SINR) at the BS for the k-th C-VUE is given by

$$\xi^{\mathrm{c}\ (1)}_{k,m} = \frac{P^{\mathrm{c}\ (1)}_{k,m} h_k}{P^{\mathrm{d}\ (1)}_{k,m} g_{k,m,\mathrm{B}} + \sigma^2} \tag{1}$$

where $P^{\mathrm{c}\ (1)}_{k\ m}$ and $P^{\mathrm{d}\ (1)}_{k\ m}$ indicate the transmit power levels of C-VUE k and the m-th D2D-V2V' transmitter on the k-th C-VUE' RB in the reuse mode, respectively. σ^2 is the noise power and we assume all VUEs have the equal noise power.

On the other hand, RB resource reusing will also cause the interference to the D2D-V2V link, where the interference is generated from the co-channel C-VUE on the reused RB. Then, the received uplink SINR at the receiver of D2D-V2V pair m when reusing the RB of C-VUE k can be given by

$$\xi^{\mathrm{d}\ (1)}_{k,m} = \frac{P^{\mathrm{d}\ (1)}_{k,m} h_m}{P^{\mathrm{c}\ (1)}_{k,m} g_{k,m} + \sigma^2} \tag{2}$$

2) Dedicated Mode: In the dedicated mode, a RB resource should be allocated to only one D2D-V2V pair when two D-VUEs communicate directly and they are nearby, and this RB is empty that it is not currently utilized by C-VUEs or other D-VUEs. The uplink SNR of D2D-V2V pair m on the unused RB is expressed as

$$\xi^{\mathrm{d}\ (2)}_{m} = P^{\mathrm{d}\ (2)}_{m} h_m / \sigma^2 \tag{3}$$

where $P^{\mathrm{d}\ (2)}_{m}$ denotes the transmit power of D2D-V2V pair m on the unused RB in the dedicated mode.

3) Cellular Mode: In this mode, two D-VUEs cannot directly communicate with each other if they are far away from each other or the channel gain between them is poor. Under this case, two D-VUEs can communicate through the BS (as a relay) as conventional C-VUEs, and two RBs (one uplink and one uplink) are allocated to this D2D-V2V pair. In this case, we assume that the RB is not allocated to any other D2D-V2V pair. The uplink SNR of D2D-V2V pair m on the unused RB can be written as

$$\xi_m^{\mathrm{d}\,(3)} = P_m^{\mathrm{d}\,(3)} h_m / \sigma^2 \tag{4}$$

where $P_m^{\mathrm{d}\,(3)}$ denotes the transmit power of the m-th D2D-V2V pair on the unused uplink RB in the cellular mode. In this paper, we focus on the resource management in the uplink and assume that the BS can allocate its resource to guarantee that the downlink SNR is not less than $\xi_m^{\mathrm{d}\,(3)}$ [26].

In general, when the C-VUE' RB is not currently reused by any other D2D pairs, it will not suffer the interference from D2D-V2V pairs. In this case, the received uplink SNR at the BS for the k-th C-VUE can be given by

$$\xi_k^{\mathrm{c}\,(2)} = P_k^{\mathrm{c}\,(2)} h_k / \sigma^2 \tag{5}$$

where $P_k^{\mathrm{c}\,(2)}$ denotes the transmit power of C-VUE k in this case.

Then, the sum data rate of C-VUE k and D2D- V2V pair m over their allocated RBs can be given by

$$\begin{aligned} R_k^{\mathrm{c}} = {} & x_{k,m}^{\mathrm{d}\,(1)} \log_2\left(1 + \xi_{k,m}^{\mathrm{c}\,(1)}\right) \\ & + \left(1 - \sum_{m \in \mathcal{M}} x_{k,m}^{\mathrm{d}\,(1)}\right) \log_2\left(1 + \xi_k^{\mathrm{c}\,(2)}\right) \end{aligned} \tag{6}$$

and

$$\begin{aligned} R_m^{\mathrm{d}} = {} & \sum\nolimits_{k \in \mathcal{K}} \left(x_{k,m}^{\mathrm{d}\,(1)} \log_2(1 + \xi_{k,m}^{\mathrm{d}\,(1)})\right) \\ & + x_m^{\mathrm{d}\,(2)} \log_2\left(1 + \xi_m^{\mathrm{d}\,(2)}\right) + x_m^{\mathrm{d}\,(3)} \log_2\left(1 + \xi_m^{\mathrm{d}\,(3)}\right) \end{aligned} \tag{7}$$

respectively, where $x_{k,m}^{\mathrm{d}\,(1)}, x_m^{\mathrm{d}\,(2)}$ and $x_m^{\mathrm{d}\,(3)}$ denote the mode selection indicators, with representing the reuse mode, the dedicated mode and the cellular mode, respectively.

C. URLLC Requirements of IoV Communications

In IoV networks, the D2D-V2V links may have different communication services, but these communication services can be mainly divided into two scenarios: the unicast scenario and the normal scenario, where the unicast scenario aims to guarantee the vehicular traffic safety by exchanging the safety information among VUEs, thus it has strict URLLC requirements but it has much looser constraint on the high data rate, while the normal scenario usually provides VUEs' general communication services with the high data rate requirements but they have no strict latency and reliability demands. The above requirements should be considered into the resource management in practical IoV networks and they can be formulated as strict constraints in the mathematical way.

Considering the requirements differentiation for different kinds of communication links, such as, the large data rate requirements for the V2I links and D2D-V2V links under the normal scenario, and the high strict latency and reliability for D2D-V2V links under the unicast scenario.

As illustrated above, the minimum data rate requirement constraints of C-VUE k and D2D-V2V pair m under the normal scenario can be expressed as

$$R_k^{\mathrm{c}} \geq R_k^{\mathrm{c,tar}},\ \forall k,\ \text{and},\ R_m^{\mathrm{d,nor}} \geq R_m^{\mathrm{d,nor,tar}},\ \forall m \tag{8}$$

respectively, where $R_m^{\mathrm{d,nor}}$ is the data rate of D2D-V2V pair m under the normal scenario. $R_k^{\mathrm{c,tar}}$ and $R_m^{\mathrm{d,nor,tar}}$ denote the minimum data rate requirements of C-VUE k and D2D-V2V pair m in the normal scenario, respectively.

In contrast to the above the normal scenario with less strict constraints on the latency and reliability requirements, the D2D-V2V links under the unicast scenario have both the strict latency and reliability requirements even they have much looser constraint on the high data rate.

According to the LTE standard [46], the transmission delay refers to the time interval from the application of the channel access to the successful packet transmission. The packet transmission delay is calculated as follows $T_{\mathrm{tx}} = \tau_{\text{mac}} + \tau_{\text{data}}$ [37], where τ_{mac} is the time needs to finish the three-way handshake process in the media access control (MAC) protocol, and τ_{data} is the time needs to successfully complete one packet transmission $\tau_{\text{data}} = L^{\text{packet}} / R^{\mathrm{d}}$, where L^{packet} and R^{d} are the amount of the data packet size in bits and the achievable link data rate, respectively.

Then, the latency constraint for the m-th D2D-V2V pair under the unicast scenario is guaranteed by controlling the probability of exceeding the threshold value, where the current transmission delay T_{tx} is beyond the predetermined threshold $T_{\max}$. And the probability should be less than the tolerable threshold $p_{\max}^{\text{delay}}$, which can be expressed as

$$p_m^{\text{delay}} = \Pr\left\{T_{\mathrm{tx}} \geq T_{\max}\right\} \leq p_{\max}^{\text{delay}} \tag{9}$$

The reliability of the m-th D2D-V2V pair under the unicast scenario is satisfied by controlling the outage probability, where its transmission data rate $R_m^{\mathrm{d,uni}}$ is below the target threshold $R_m^{\mathrm{d,uni,tar}}$. And the outage probability should be less than the tolerable outage probability $p_{\max}^{\text{outage}}$, which is given by

$$p_m^{\text{outage}} = \Pr\left\{R_m^{\mathrm{d,uni}} \leq R_m^{\mathrm{d,uni,tar}}\right\} \leq p_{\max}^{\text{outage}} \tag{10}$$

D. Problem Formulation

In this paper, our objective is to maximize the overall network capacity while guarantee the above mentioned QoS requirements of VUEs and satisfy the resource constraints of the network. Then, the resource management (joint mode selection, RB assignment, and power control) problem can be mathematically formulated as

$$\max_{\mathbf{x},\mathbf{P}} \left\{ \sum_{k \in \mathcal{K}} R_k^{\mathrm{c}} + \sum_{m \in \mathcal{M}_{\mathrm{nor}}} R_m^{\mathrm{d,nor}} \right\} \tag{11a}$$

$$\text{s.t. } (8), (9), (10);\ x_{k,m}^{\mathrm{d}\,(1)}, x_m^{\mathrm{d}\,(2)}, x_m^{\mathrm{d}\,(3)} \in \{0,1\}, \forall m; \tag{11b}$$

$$\sum_{m\in\mathcal{M}} x_{k,m}^{\mathrm{d}\,(1)} \le 1, \forall k; \tag{11c}$$

$$\sum_{k\in\mathcal{K}} x_{k,m}^{\mathrm{d}\,(1)} + x_{m}^{\mathrm{d}\,(2)} + x_{m}^{\mathrm{d}\,(3)} \le 1, \forall k; \tag{11d}$$

$$\sum_{m\in\mathcal{M}} x_{m}^{\mathrm{d}\,(2)} + \sum_{m\in\mathcal{M}} x_{m}^{\mathrm{d}\,(3)} \le N_{\mathrm{U}}; \tag{11e}$$

$$\begin{aligned}\sum_{k\in\mathcal{K}} \left(x_{k,m}^{\mathrm{d}\,(1)} P_{k,m}^{\mathrm{d}\,(1)}\right) &+ x_{m}^{\mathrm{d}\,(2)} P_{m}^{\mathrm{d}\,(2)} \\ &+ x_{m}^{\mathrm{d}\,(3)} P_{m}^{\mathrm{d}\,(3)} \le P_{\max}^{\mathrm{d}}, \ \forall m;\end{aligned} \tag{11f}$$

$$x_{k,m}^{\mathrm{d}\,(1)} P_{k,m}^{\mathrm{c}\,(1)} + \left(1 - \sum_{m\in\mathcal{M}} x_{k,m}^{\mathrm{d}\,(1)}\right) P_{k}^{\mathrm{c}\,(2)} \le P_{\max}^{\mathrm{c}}, \ \forall k. \tag{11g}$$

where $\mathbf{x} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}$ denote the mode selection and RB assignment matrix. $\mathbf{P}$ is the transmit power matrix of D-VUEs and C-VUEs. M_{nor} denotes the D2D-V2V pairs set in the normal scenario.

In the above, constraint (11c) indicates that one C-VUE' RB is reused by one D2D-V2V pair at most. Constraint (11d) is to ensure that any D2D-V2V pair will choose at most one of the three modes. Constraint (11e) shows that the RBs used by all D2D-V2V pairs in the cellular and dedicated modes should not exceed the total number of unused RBs. Constraints (11f) and (11g) are to satisfy that the transmit power of each D2D-V2V pair and each C-VUE cannot exceed their maximum values, $P_{\max}^{\mathrm{d}}$ and $P_{\max}^{\mathrm{c}}$, respectively.

III. ETAC Learning for Intelligent Resource Management

The optimization problem in (11) is difficult to solve as it is a non-convex combination and NP-hard problem. In addition, the IoV communication network needs an intelligent resource management framework to enable V2X communication links to intelligently make decision with high-level intelligence. Hence, in this section, an ETAC learning approach is proposed to address the intelligent resource management problem in IoV networks, where the different QoS requirements and constraints (e.g., latency and reliability requirements of V2V links or minimum data rate requirements of V2I links) can be directly addressed. Before presenting the proposed approach, the main parts of the RL based Markov Decision Process (MDP) are shown with a new proposed reward function, and an AC learning framework is adopted to address the intelligent resource management problem. Finally, the online solution based on ETAC is presented to learn the optimized policy for intelligent resource management in IoV networks.

A. Markov Decision Process for Resource Management

RL is an important part of machine learning [34], which has the ability to help the IoV communication network to exploit the optimal policy to maximize its profit. Similar to most of existing works [18], [28], we apply MDP to model the policy searching process in the RL formwork.

In MDP model, a sequence of resource management decisions of a learning agent by interacting with the IoV environment at some discrete time scale can be defined as a tuple $(S, A, \mathcal{P}, r_t, \gamma)$, where S denotes the network state space set, A means the action space set, $\mathcal{P}$ is the transition probability: $\mathcal{P}(s_{t+1}|s_t, a_t)$ when the agent takes the action $a_t \in A$ from the current network state $s_t \in S$ to a new state $s_{t+1} \in S$; r_t denotes the immediate reward at the current time slot. $\gamma \in [0, 1)$ is a discount factor.

Agent: Each active V2I link and each active D2D-V2V link.

State space: For each communication link (agent), the network state can be depicted as $s = \{\chi_{\mathrm{rb}}, \chi_{\mathrm{ch}}, \chi_{\mathrm{cq}}, \chi_{\mathrm{tl}}, \chi_{\mathrm{re}}\}$, where χ_{rb} indicates all RBs occupy status among VUEs, χ_{ch} shows the observed channel information, χ_{cq} represents the selected channel quality (SINR or SNR), χ_{tl} indicates the remaining transmission load of D2D-V2V links in the unicast scenario, and χ_{re} indicates the QoS requirements, such as the minimum data rate, the latency and the reliability requirements.

Action space: Three actions are considered for the intelligent resource management in the reinforcement learning framework. We denote $a = \{\beta_{\mathrm{ms}}, \beta_{\mathrm{rb}}, \beta_{\mathrm{po}}\} \in A$ as the candidate of actions of one agent at the state s after making decision in terms of the communication mode selection (β_{ms}), the RB assignment (β_{rb}), and the transmit power level (β_{po}). Note: the action selection of each agent should satisfy the constraints (11c)-(11g).

Reward function: The learning process is driven by the reward function in the RL framework, and each agent makes decision by maximizing its reward with the interactions of the environment. Considering the requirements differentiation for different types of IoV communication links, a new reward function for the intelligent resource management is proposed by following analysis.

In the IoV network, the goal of RL is to learn the optimal policy to intelligently make decision based on the current environment states, through maximizing the expected reward. Hence, how to design an efficient reward function is very important, which directly determines that the optimal policy that the IoV network makes, and the considerable actions it takes.

Considering the practical QoS differentiation for different types of communication links, we propose a new reward function for the intelligent resource management problem in IoV networks, which can be given by

$$\begin{aligned} r = {} & \underbrace{c_1 \left(\sum_{k\in\mathcal{K}} R_k^{\mathrm{c}} + \sum_{m\in\mathcal{M}_{\mathrm{nor}}} R_m^{\mathrm{d,nor}} \right)}_{\text{part 1}} \\ & \underbrace{- c_2 \left(\sum_{m\in\mathcal{M}_{\mathrm{uni}}} \left(p_m^{\text{delay}} + p_m^{\text{outage}} \right) \right)}_{\text{part 2}} \\ & \underbrace{- c_3 \left(\sum_{k\in\mathcal{K}} \left(R_k^{\mathrm{c,tar}} - R_k^{\mathrm{c}} \right) + \sum_{m\in\mathcal{M}_{\mathrm{nor}}} \left(R_m^{\mathrm{d,nor,tar}} - R_m^{\mathrm{d,nor}} \right) \right)}_{\text{part 3}} \end{aligned} \tag{12}$$

where the Part 1 is the immediate utility (the sum data rate), the Part 2 and Part 3 are the immediate cost functions in terms of the unsatisfied latency and reliability requirements for unicast V2V links, the unsatisfied minimum capacity requirements for V2I links and normal V2V links, respectively. The coefficient c_i, $i \in \{1, 2, 3\}$ are the weights of the three parts in (12), and they are also used to for balancing the utility and cost. M_{uni} denotes the D2D-V2V pairs set in the unicast scenario.

If the policy is optimal, i.e., the highest transmission data rate is achieved with the lowest outage probability and transmission delay for different types of communication links in the IoV network, the reward value is large. In contrast, if the low transmission date rate, the frequent outage events happen and the high transmission delay in the network, all these negative factors will lead to the low reward value. The low reward value reflects the poor policy of the current learning mechanism, so the learning policy should be improved.

In the IoV network, each agent aims to choose a policy π to maximize its expected reward. Let $Q^\pi(s,a)$ denotes the state-action function, which is a cumulative discounted reward for starting the network state s with a given policy π, which is expressed as

$$Q^\pi(s,a) = E\left\{\sum_{t=1}^{\infty} \gamma^t r_t(s_t, a_t)|s_0 = s, \pi\right\} \tag{13}$$

The function $Q^\pi(s,a)$ in (13) is applied to evaluate how good the policy π is when taking the action at the current state s. It also can be computed recursively using the Bellman equation [34]–[36], which is expressed as

$$Q^\pi(s,a) = E\left\{r_t(s_t, a_t) + \gamma Q^\pi(s_{t+1}, a_{t+1})\right\} \tag{14}$$

The goal of the intelligent resource management is to search the policy π that maximizes the network objective reward value, which is given by [34]

$$J(\pi) = E\left\{Q^\pi(s,a)\right\} = \int_S d(s) \int_A \pi(s,a) Q^\pi(b,a) dads \tag{15}$$

where $d(s)$ denotes the state distribution function, and $\pi(s,a)$ is a stochastic policy with the state s over the current action a, which shows the conditional probability density of the action a at the state s.

From (15), we observe that the policy can be optimized numerically by adopting the value iteration method [34] with the RL tools, such as Q-learning, policy gradient method, actor-critic and so on [34]. In the next subsection, a model-free RL is exploited to solve the intelligent resource management in IoV networks.

B. AC Learning for Intelligent Resource Management

The intelligent resource management problem formulated in Section II-D can be solved by applying the Q-learning, deep Q-learning and policy gradient methods. However, the Q-function or the Q-function approximator in the Q-learning network is slow to be learned [34], [35], leading to the low convergent rate, thus it may fail to search the optimal policy with the given iterations. Moreover, Q-learning is not capable of learning stochastic policies, where the IoV network has continuous action spaces with continuous- valued stochastic variables (e.g., channel condition, transmit power). In addition, the policy gradient method can achieve a good policy by operating the learning strategy in the policy space with the faster convergence rate than Q-learning, but it may converge to a local optimal policy.

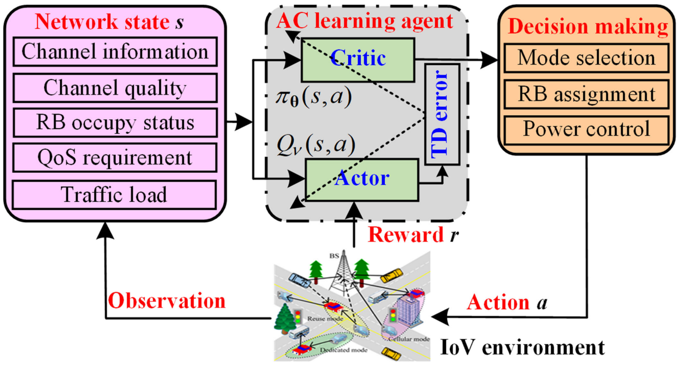


Fig. 2. The AC learning for IoV communication networks.

Hence, in order to learn the optimal policy for the intelligent resource management with the continuous-valued states and actions, an actor-critic learning algorithm is adopted to optimize the policy by combining the process of the policy learning and value learning designs with good convergence properties.

The framework of the AC learning for the intelligent resource management in IoV networks is shown in Fig. 2. In the AC learning framework, there are two parts, namely, the actor and the critic, where the actor is represented through adopting a control policy with action selections based on the observed network state, and the critic evaluates the input policy by a reward function from the environment feedback. In IoV networks, V2I and V2V links can be regraded as agents and the network constitutes the environment. Each link observes the current network state, then it makes decision by itself based on its learned policy strategy without sharing its strategy with other links in a decentralized way. After that, the IoV environment provides a new network state and the immediate reward r in (12) to the agents. According to the feedback, all agents learn a new policy in the next step.

1) The Critic Process: The Critic Process: The goal of the critic part of the AC learning is to evaluate the quality of the policy that the learning framework searches. However, the optimal state-action function $Q^\pi(s,a)$ cannot be computed for the infinite state (observation state) with the Bellman equation [34]. For this problem, the AC learning can adopt the function estimator to approximate the value function [34]. The linear-function estimator is usually adopted to evaluate the function approximation [34], and the value function approximator can be written as

$$Q_\nu(s,a) = \nu^T \varphi(s,a) = \sum_{i \in \mathcal{L}} {\nu_i}^T \varphi_i(s,a) \tag{16}$$

where $\varphi(s,a) = (\varphi_1(s,a), \ldots \varphi_M(s,a))^T$ is the basis function vector when the network chooses the action a at the state s,

$\nu(s,a) = (v_1, \ldots v_E)^T$ is the parameter vector of weights. $\mathcal{L}$ is the set of the elements in the state space.

The temporal-difference learning scheme is also adopted in the AC reinforcement learning to compute the temporal difference (TD) error between the estimated value and the real value, which can be expressed as

$$\delta_t = r_{t+1} + \gamma Q_\nu(s_{t+1}, a_{t+1}) - Q_\nu(s_t, a_t) \tag{17}$$

When adopting the linear function estimator shown in (16), the parameter vector of weights $\nu(s,a)$ can be updated by using the gradient descent method

$$\nu_{t+1} = \nu_t + \beta_c \delta_t \nabla_\nu Q_v(s,a) = \nu_t + \beta_c \delta_t \varphi(s,a) \tag{18}$$

where β_c is the critic learning rate. After optimizing the parameter $\nu(s,a)$ in (18), then the actor updates its state-value function in (16).

2) The Actor Process: The stochastic Policy Gradient (PG) method [34] is usually adopted in the actor part to update the parameterized policies, in order to optimize the policy step by step to improve the objective function in (15). The policy $\pi(s,a)$ can be initially built by using the parameter vector $\theta = \{\theta_1, \theta_2, \ldots, \theta_u\}$, which is denoted by $\pi_\theta(s,a) = \Pr(a|s,\theta)$.

Then, the gradient of the policy with respect to θ in terms of the objective function in (15) is given by

$$\nabla_\theta J(\pi_\theta) = = \int_S d(s) \int_A \nabla \pi_\theta(s,a) Q^{\pi_\theta}(s,a) dads \tag{19}$$

The parameterized stochastic policy $\pi_\theta(s,a)$ is generally represented by the Gibbs distribution [34], i.e.,

$$\pi_\theta(s,a) = \frac{\exp(\theta^T \cdot \mathbf{\Phi}(s,a))}{\sum_{a' \in A} \exp(\theta^T \cdot \mathbf{\Phi}(s,a'))} \tag{20}$$

where $\mathbf{\Phi}(s,a)$ is the feature vector [34].

Then, the policy parameter vector is updated based on the gradient of the objective reward in (15) as follows

$$\theta_{t+1} = \theta_t + \beta_a \nabla_\theta J(\pi_\theta) \tag{21}$$

where β_a denotes the learning rate of the actor.

C. ETAC for Intelligent Resource Management

The previous subsection addresses the intelligent resource management problem by exploiting the classical AC learning approach to help the IoV network to smartly select communication mode, assign RBs and control transmit power under different QoS requirements. The AC learning is capable of optimizing the policy with the low computational cost and fast convergence properties by using the gradient method. However, the actor part using the policy gradient method may converge to a local optimal policy, and directly updating the actor and critic parts with the TD error value causes the AC learning to diverge. Moreover, in the high dynamic IoV environment, the classical AC learning in will lead to poor convergence performance due to the high variance of the gradient estimation.

Consequently, motivating by the idea of transfer learning [27], we propose an efficient transfer AC learning approach to enhance the convergence speed and improve the learning efficiency toward the optimal policy in IoV networks, where a part of VUEs are encouraged to utilize the learned knowledge from their neighbors by exchanging information with each other, as shown in Fig. 3.

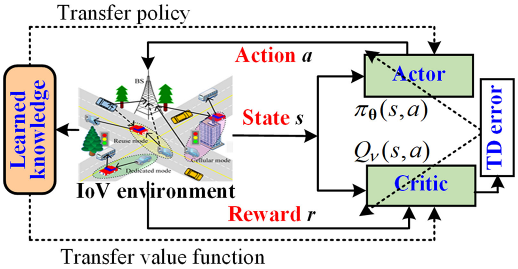


Fig. 3. Transfer AC learning based VUE-to-VUE teaching.

There are two following issues we should investigate when we propose the ETAC approach for the intelligent resource management in IoV networks.

1) The Expert VUE Selection: Let us consider that when a new VUE joins IoV networks, instead of building a MDP model, it searches a neighboring VUE as the expert to utilize the current learned strategy from the expert VUE (such as mode selection, RB assignment and transmit power). In addition, if one VUE has slow convergence performance or its QoS requirements fail to be satisfied based on its current policy, it can request one neighboring VUE as the expert to transfer the learned strategies. In order to search the expert, the VUEs can exchange the following three types of information among their neighbors, such as, the IoV communications types (V2I or V2V), the communication services (unicast scenario or normal scenario), and the QoS requirements (minimum capacity, reliability and latency requirements).

The similarity level between the learning VUE and the expert VUE can be evaluated by adopted the Bregman Ball concept [26], and its calculation contains the above three metrics. The VUE has the highest similarity level with the learning VUE is selected as the expert VUE, and the learning VUE utilizes the transferred knowledge from the expert VUE.

2) Efficient Transfer AC Learning Based Learned Knowledge: The AC learning optimizes both the policy function (or actor parameter) and the value function (or critic parameter) alternately according to the transferred knowledge. We implement the ETAC learning process as follows:

(i) Action Selection: Once the learning VUE finds its expert VUE who has the highest similarity, it utilizes the transferred action information from the expert VUE and its current action space to generate an overall action space range $[u_{\mathrm{tr}}, u_{\mathrm{cu}}]$, where u_{cu} is the agent' current selected action and u_{tr} denotes the transferred action information (Note that we suppose $u_{\mathrm{tr}} \leq u_{\mathrm{cu}}$). After that, the agent greedily selects the actual action from them by interacting with the environment.

Without loss of generality, at the network state , the selected action can be updated by

$$\mu(s) = \omega\ u_{\mathrm{tr}} + (1 - \omega) u_{\mathrm{cu}} \tag{22}$$

where $\omega \in [0, 1]$ can be defined as the transfer rate, which will be decreased after each learning step to gradually remove the effect of the transferred strategy information on the new action selection.

(ii) Critic Parameters Update: In order to avoid the bias generated by the approximated policy gradient, we use the squared value error between the value function approximator $Q_v(s, a)$ and the optimal $Q^{\pi_\theta}(s, a)$ to evaluate the quality of the parameter vector $\nu(s, a)$ in the critic part. Then, the squared value error between $Q^{\pi_\theta}(s, a)$ and $Q_v(s, a)$ is given by

$$\varepsilon_{\nu}^{\pi_\theta}(s, a) = [Q^{\pi_\theta}(s, a) - Q_v(s, a)]^2 \tag{23}$$

From (23), the smaller the mean square error is, the more accurate the value function parameter $\nu(s, a)$ achieves.

Considering the fact that the selected action and policy may affect both the immediate reward and the accumulative reward in the future learning process, only updating the parameter vector $\nu(s, a)$ is not enough. Hence, an eligibility trace mechanism is adopted to improve the convergence speed and the efficient learning way by encouraging each agent to overview all the rewards forward and better integrate them [34]. Let $\mathbf{z}_t$ denotes the eligibility trace vector at the time step t, then, the update equations for the eligibility trace vector $\mathbf{z}$ and the critic parameter vector ν can be expressed as

$$\mathbf{z}_{t+1} = \gamma \lambda_c \mathbf{z}_t + \varphi(s_t, a_t) \tag{24}$$

$$\nu_{t+1} = \nu_t + \beta_c \delta_t \mathbf{z}_{t+1} \tag{25}$$

respectively, where $\lambda_c \in [0, 1]$ is the eligibility trace decay factor. With the help of the eligibility trace mechanism, the value function can be efficiently updated and hence the learning process can be improved.

After optimizing the parameter $\nu(s, a)$ by minimizing the mean square error in (23), the actor updates its state-value function as

$$Q_v(s, a) = \nu^T \varphi(s, a) \tag{26}$$

Finally, the nearest estimator of $Q^{\pi_\theta}(s, a)$ by $Q_v(s, a)$ is achieved iteratively by updating the parameter vector

For the gradient function in (19), the baseline function $b(s)$ is adopted to reduce policy gradient variance without changing the policy gradient value. We denote the best selection of baseline function is the state-value function $V^\pi(s)$. Then, the gradient in (19) is expressed as

$$\nabla_\theta J(\pi_\theta) = \int_S d(s) \int_A \nabla \pi_\theta(s, a)(Q_\nu(b) - b(s)) da ds \tag{27}$$

When the estimated value $Q_\nu(s, a)$ in (26) is sufficiently equal to the optimal value $Q^{\pi_\theta}(s, a)$, the baseline function $b(s) = V^\pi(s)$ is the best baseline that achieves the minimum variance in the action-value function approximator. By submitting $b(s) = V^\pi(s)$ into (27), we can get

$$\nabla_\theta J(\pi_\theta) = \int_S d(s) \int_A \nabla \pi_\theta(s, a) B^\pi(s, a) da ds \tag{28}$$

where $B^\pi(s, a) = Q_\nu(s, a) - V^\pi(s)$ is the advantage function [25].

(iii) Actor (Policy) Parameter Update: The updated is achieved after the optimization of the advantage function in (28), the policy gradient is given by

$$\nabla_\theta J(\pi_\theta) = \int_S d(s) \int_A \nabla \pi_\theta(s, a) Q_\nu(s, a) da ds \tag{29}$$

Then, the policy parameter vector can be updated as

$$\theta_{t+1} = \theta_t + \beta_a \nabla_\theta J(\pi_\theta) \tag{30}$$

If every action is executed under an infinite number of iterations at each network state, that is to say, the learning policy is greedy with the infinite explorations, the state-value function $Q_v(s, a)$ and the policy strategy $\pi_\theta(s, a)$ will iteratively converge to the final points, respectively, with a probability of 1 [35], [36].

D. Online Solution Based Efficient Transfer AC Learning

For the proposed ETAC approach, the value function parameter vector ν is updated in the critic part, while the policy parameter vector θ is updated by the actor, both these two parameter vectors can be updated iteratively and simultaneously. During the learning process, the eligibility trace mechanism is adopted to improve the efficient learning way and the learning process, the advantage function combined with the baseline approximator is used to decrease the variance in the gradient calculation as well as enhance the function estimation accuracy in the critic part, and the action strategy transfer leaning is presented to increase the convergence speed and improve the overall learning quality. The complete proposed ETAC approach is shown in **Algorithm 1** (Note: the IoV environment simulators include the C-VUEs and D-VUEs, and their estimated channel state information (CSI). The positions of vehicles are randomly dropped on the road so that the CSI of V2V and V2I links is generated according to their current positions).

In IoV networks, in each learning step, each communication link (agent) observes the current network state s (the RB occupy status, the channel quality, its QoS requirements, etc.), then it chooses an action a (mode selection, RB assignment, and transmit power level) according to the policy strategy $\pi_\theta(s, a)$. Then, the IoV environment provides a new network state and the immediate r reward in (12) to the agents. In the learning framework, the critic part then estimates the value function $Q_v(s, a)$ and calculates the TD error δ. At the same time, the critic updates the eligibility trace vector $\mathbf{z}$ as well as the critic parameter vector ν, and compute the advantage function $B^\pi(s, a)$. After that, the actor part uses the advantage function to achieve the policy gradient $\nabla_\theta J(\pi_\theta)$ before updating its policy parameter vector θ by applying the policy gradient method. If a new VUE has just joined the network or one VUE has poor learning performance (e.g., fail to guarantee its QoS requirements or has poor convergence speed, etc.), then the transfer learning is operated for these VUEs to improve their learning efficiency. The parameters θ and ν in the ETAC learning framework will be improved with an infinite number of learning steps, when the efficient transfer AC leaning converges to the final value function and policy.

Then, the best actions for the intelligent resource management are achieved for IoV networks.

Generally, for most of communication links, the action of each VUE is chosen independently according to its own local information without sharing its learned strategy with other communication links if the VUE does not need to utilize the transferred strategy from any expert. Only when the new VUE or the VUE with poor learning performance requests the learned experience from its expert VUE by using a small transmission overhead. Hence, compared with the centralized IoV network, the advantage of our proposed efficient transfer AC learning approach is that the agent can learn independently when it has an efficient learning performance, rather than continuously exchange information among VUEs.

Algorithm 1: ETAC for Intelligent Resource Allocation.

```
Input: learning rate factor $\beta_a$ and $\beta_c$, discount parameter
  $\gamma$, decay factor $\lambda$, all IoV environment simulators.
1:  Initialize: initial state $s_0$, state-value function $Q_v(s,a)$,
      policy function $\pi_\theta(s,a)$, eligibility trace $z_0$;
2:  for each time step $t = 0, 1, 2, \ldots$ do
3:     Observe the environment state $s_t$;
4:     Receive the current reward $r_t$;
5:     if the VUE is new or has poor performance then;
6:        Find the expert VUE with the highest similarity;
7:        Get the transferred strategy from the expert;
8:        Generate the overall action space $[u_{\mathrm{tr}}, u_{\mathrm{cu}}]$;
9:        Update the transfer rate, and get the selected action
            by (22);
10:       Perform efficient AC learning from step 14 to
            step 20;
11:    else
12:       Directly perform efficient AC learning from step
            14 to step 20;
13:    end if
     Critic Process
14:    Compute value function: $Q_v(s,a) = \nu^T \varphi(s,a)$;
15:    Compute TD error: $\delta_t = r_{t+1} + \gamma Q_\nu(s_{t+1}, a_{t+1}) -$
         $Q_\nu(s_t, a_t)$ ;
16:    Update eligibility trace: $\mathbf{z}_{t+1} = \gamma\lambda_c \mathbf{z}_t + \varphi(s_t, a_t)$;
17:    Update critic parameters: $\nu_{t+1} = \nu_t + \beta_c \delta_t \mathbf{z}_{t+1}$ ;
     Actor Process
18:    Compute advantage function: $B^\pi(s,a) =$
         $Q_\nu(s,a) - V^\pi(s)$ ;
19:    Update policy gradient: $\nabla_\theta J(\pi_\theta) = \int_S d(s)$
         $\int_A \nabla \pi_\theta(s,a) Q_\nu(s,a) da ds$
20:    Update policy parameters: $\theta_{t+1} = \theta_t + \beta_a \nabla_\theta J(\pi_\theta)$.
21: End for
```

IV. Numerical Results and Discussion

In this section, simulation results are conducted in Matlab R2016a to evaluate the performance of our proposed intelligent resource management based the ETAC approach in the IoV network, and we compare it with the following approaches: 1. The optimal resource management approach, which is achieved by solving the optimization problem (11) in a centralized way (denoted as Optimal), where each VUE can obtain the full knowledge of environment information; 2. AC learning by using the policy gradient method (denoted as AC-PG); 3. Q-learning approach (denoted as Q-learning); 4. Optimizing the objective function (11a) with the constraints from (11b) to (11g) by decomposing the problem (11) into two subproblems: i). mode selection and resource block, ii). power control, which can be solved iteratively in a centralized way, refers to [9], [10], [26] (denoted as Baseline 1); 5. Random search approach (denoted as random search).

TABLE I
Simulation Parameters

Parameters	Value
Cell radius	500 m
Carrier frequency, each RB bandwidth	2 GHz, 180KHz
Number of lanes (4 in each direction)	16
Max D2D-V2V communication distance	100 m
Number of C-VUEs	20
Number of unused RBs, number of RBs,	10, 30
Number of D-VUEs	20, 40, …, 120
Total number of D2D-V2V links	40, 80, …, 240
Number of D2D-V2V links in normal links	8, 10, …., 18
Tolerable probability threshold of the latency in unicast links	0
Maximum C-VUE or D-VUE transmit power	23 dBm
Background noise power σ^2	-114 dBm
Absolute vehicle speed	20, 30, …., 70 km/h
Learning rates of actor and critic, β_a and β_c	0.02, 0.02

We consider a single cell where all smart vehicles are dropped in a crossroad (two roads cross each other) based on the spatial Poisson process, as shown in Fig. 1, and each road is a multi-lane freeway that passes through the cell with the BS being locating at the cell center. The C-VUEs and D-VUEs are randomly selected among the active smart vehicles, and each D-VUE can simultaneously build two D2D-V2V links (pairs). We use the simulation setup based on the Manhattan case detailed in 3GPP TR 36.885 [47]. The radio resource is organized in a number of uplink RBs with 180 kHz per RB in one time-slot (0.5 ms), and all RBs are with both the Non-Light-of-Sight (NLOS) links as well as the Light-of-Sight (LOS) links in IoV communications. In order to achieve the demand of the METIS project [11], we set the latency threshold $T_{\max} = 5$ ms, the transmission reliability is 0.001 with the minimum capacity being 0.5 bits/s/Hz, the safety-critical message sizes are 1600 bytes and the SINR threshold in unicast links is 3dB. In order to satisfy the general communication services, the minimum capacity of C-VUE and D-VUE in normal links are 3.5 and 3 bits/s/Hz, respectively. We set $c_1 = 1$, $c_2 = 10^3$ and $c_3 = 10$ to balance the utility and the cost in (12) [40], [45]. The RL approaches (ETAC, AC-PG and Q-learning) have the following same parameters' values: the discount factor $\gamma = 0.95$, the trace decay factor $\lambda_c = 0.3$ and the learning rate is 0.02 [34]. Other main simulation parameters are shown in Table I.

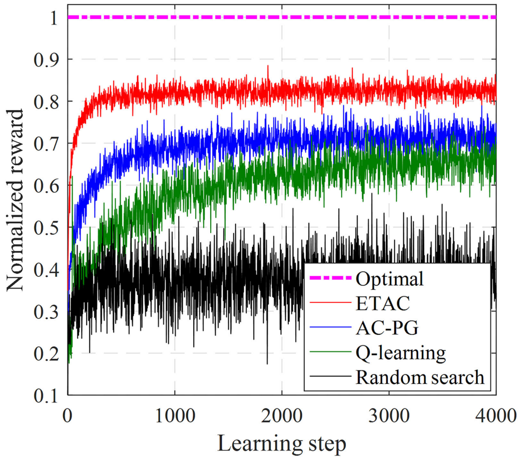
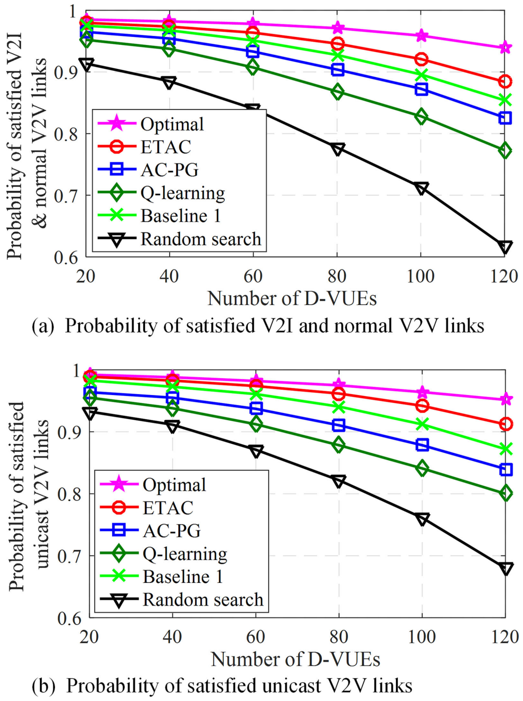


Fig. 4. Learning process comparison.

Fig. 4 shows the learning process of the five approaches in terms of the reward performance when the number of D-VUEs is 60 and the absolute vehicle speed is 40 km/h. We can see that the three RL approaches greatly outperform the random search approach, especially, the proposed ETAC approach achieves the best reward performance with the fastest convergence speed and the most stable learning process (less fluctuations) compared with other three approaches excepts the optimal approach. For the Q-learning approach, it needs more learning steps to optimize the Q-function approximator, so the slow convergence may fail to guarantee the strict latency requirements in IoV networks. In addition, the AC-PG learning approach has a fast convergence speed compared with Q-learning, but it may converge to the local optimal point. For the random search approach, its performance is worst among the five approaches, because it randomly search the policy only based on the current immediate reward, but it has the simple structure. The optimal approach is capable of searching the optimal policy in a centralized way with the best reward performance, while other four approaches search their optimized policy in a decentralized way. However, it is not practical that the optimal approach can simultaneously collects the feedback information from VUEs, and it needs a lot of transmission overhead. Our proposed approach adopts the efficient transfer learning to enhance the learning efficiency and the convergence rate, and the optimized policy will be learned after a finite number of learning steps.

Fig. 5 demonstrates the probability of the satisfied IoV communication links versus the number D-VUEs when the absolute vehicle speed is 40 km/h. From Fig. 5, we can see that the probability of the satisfied V2I and V2V links decreases as the increase number of active D-VUEs, because under the limited radio resource situation, the large number of V2V links need to be connected and different QoS requirements should to be satisfied, all approaches may fail to complete all the communication services with the large number of D-VUEs, leading to bring down the satisfied links. However, the proposed approach achieves a much larger satisfied probability for different IoV communication links to guarantee the minimum capacity, latency and reliability requirements through effectively searching the resource management policy, especially the performance gap between the proposed approach and other approaches (excepts the optimal approach) becomes more obviously when the number of D-VUEs is large. The high satisfied communication links indicates that the ETAC approach can effectively guarantee the strict latency and transmission reliability requirements of unicast V2V links, and ensure the minimum capacity requirements of V2I links and normal V2V links. We would like to mention that Baseline 1 approach has higher performance than AC-PG and Q-learning due to its centralized resource management policy at the BS, but it has lower performance than that of the ETAC approach due to its suboptimal solution by dividing the optimization problem into two subproblems.



(a) Probability of satisfied V2I and normal V2V links

(b) Probability of satisfied unicast V2V links

Fig. 5. Probability of satisfied IoV communication links with varying number of D-VUEs.

Fig. 6 depicts the sum rate performance of V2I links and normal V2V links for the six approaches with different number of active D-VUEs when the absolute vehicle speed is 40 km/h. From Fig. 6, we can see that the performance of all approaches reduces as the increase number of D-VUEs. Because more D-VUEs increases the more V2V links, all links aim to share the fixed radio resource, consequently leading to the growing amount of generated interference to V2I links and normal V2V links, hence the sum rate performance reduces as the increase number of D-VUEs. It is interesting to note that our proposed approach and the Baseline 1 approach have the comparable sum rate performance and they outperform other approaches (except the optimal approach) through adjusting the decision making policy to manage network resource and effectively mitigating the total generated interference to V2I links and normal V2V links. However, even the optimal approach and Baseline 1 achieve the considerable performance, they performed in the BS in a centralized way which needs to continuously exchange

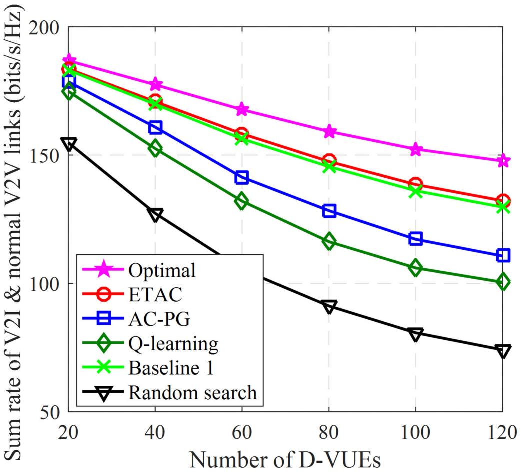
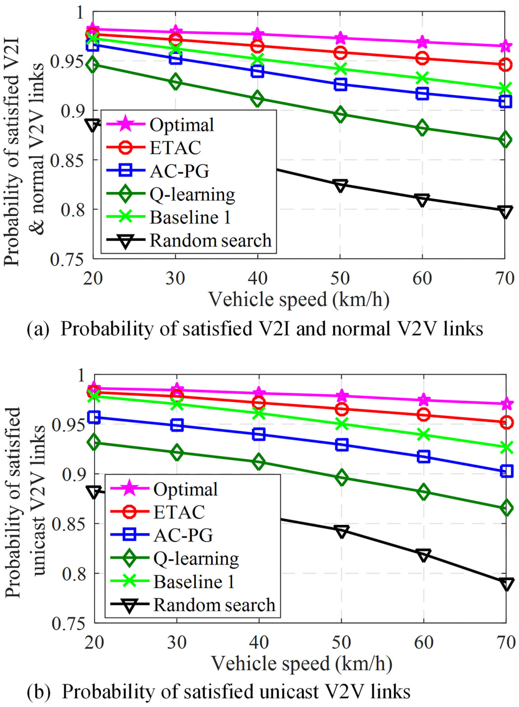


Fig. 6. Sum rate performance of V2I and normal V2V links with varying number of D-VUEs.



(a) Probability of satisfied V2I and normal V2V links

(b) Probability of satisfied unicast V2V links

Fig. 7. Probability of satisfied IoV communication links with varying vehicle speed.

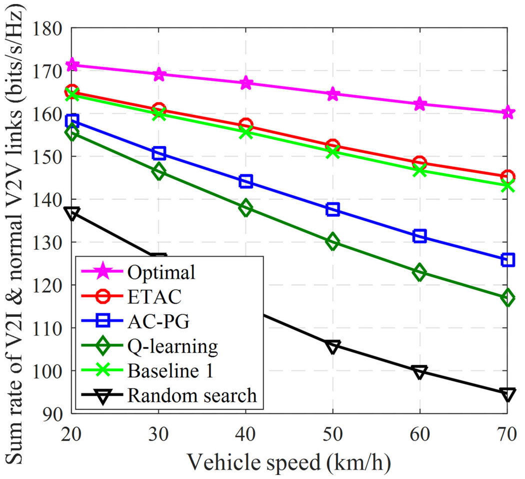


Fig. 8. Sum rate performance of V2I and normal V2V links with varying vehicle speed.

information between VUEs and the BS, leading to the heavy transmission overhead.

Fig. 7 shows the probability of satisfied IoV communication links with an increasing absolute vehicle speed, when the number active D-VUEs is 60. We observe that the probability of the satisfied communication links decreases as the vehicles move faster. Such the performance degradation results from the lower received desired power in IoV communication links due to the average increase inter-vehicle distance when the higher movement speed induces sparser traffic. In addition, the higher speed results in the high dynamic vehicular environment, resulting in the high observation uncertainties (e.g., the channel state information and the received interference), which decreases the learning efficiency, hence the more unsatisfied communication link events happen in the high vehicle speed regions. However, as the vehicle speed increases with high uncertainties, our proposed approach can still maintain the probability of satisfied IoV communication links at a considerable level, and outperforms other approaches with the higher satisfied probability (except the optimal approach), especially in the high vehicle speed regions. This reveals that the proposed ETAC approach are more stable and robust in high dynamic IoV networks.

Fig. 8 presents the impact of the vehicle speed on the sum data rate of V2I and normal V2V links when the number active D-VUEs is 60. From the figure, the sum rate performance reduces as the increase of vehicle speed, because the high observation uncertainties are induced from the high dynamic vehicular network due to the high vehicle speed, so four approaches fail to find the optimal policy, especially the random search approach has the worst performance. However, our proposed approach can still achieves the better performance than other approaches under different vehicle speed values approaches (except the optimal approach).

Finally, we compare the convergence computational time of the RL approaches, where the all approaches are run are run in the Matlab 2016a environment on a PC with Intel(R) Core(TM) i7-6700 CPU @ 3.40 GHz, 16 RAM, and the operating system is Windows 10 Ultimate 64 bits. Fig. 9 shows the convergence computational time of different approaches with varying number of D-VUEs. (Note: The policy used in the IoV communication network for resource management is random at the beginning and then gradually improved by the updated learning strategy.). As the increase number of D-VUEs, all approaches need the more computational time to achieve the convergence. This is because that the vehicular environment becomes more complex and dynamic as the increase number of D-VUEs, all approaches need more iterations and computational time to converge their optimized solutions. However, we can observe that other two RL approaches (AC-PG and Q-learning) need more computational time to converge to their final solutions than our proposed ETAC approach, especially when the number of D-VUEs is large, the

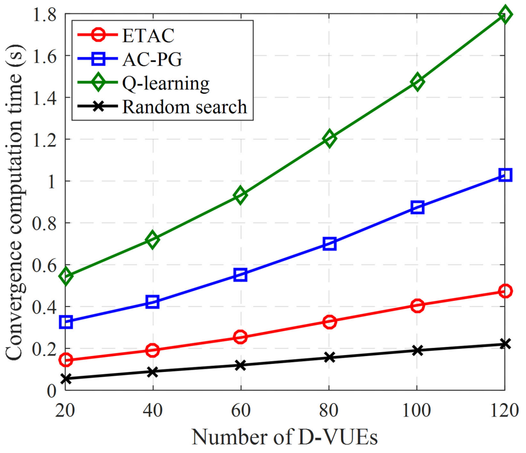


Fig. 9. Convergence computational time with varying number of D-VUEs.

computational time gap between the two RL approaches and the proposed algorithm become more obvious. In addition, the convergence time of proposed ETAC approach is substantially less related to the number of D-VUEs than those of other two RL approaches. It is interesting to note that the random search approach needs the least computational time compared with other approaches, but it achieves the worse performance in terms of the probability of satisfied links and sum rate performance as shown from Fig. 5 to Fig. 8 in the paper.

V. Conclusion

This paper has studied the joint mode selection, RB assignment and power control problem in D2D-V2V-based IoV networks, and formulated the intelligent resource management problem as a decentralized reinforcement learning model to ensure the URLLC requirements of unicast V2V links which maximizing the sum capacity of V2I links and normal V2V links. Since the dynamic environments as well as the continuous-valued state and action space exist in IoV networks, the model-free actor-critic learning framework is adopted to learn the optimal policy. Moreover, the ETAC approach was proposed to effectively enhance the learning efficiency and improve the learning convergence speed, in order to support reliable and delay-sensitive vehicular applications in IoV networks. Simulation results conforms that the proposed ETAC reinforcement learning approach can effectively learn to ensure the latency and reliability requirements of V2V links while minimizing the generated interference in IoV networks. In our future works, we will pay more attention to design efficient and robust reinforcement learning algorithms (e.g., combine with deep learning techniques), which can effectively provide ”real time” analytics in the large-scale IoV communication networks, and it is capable of adapting resource management strategy for different scenarios in dynamic/complex IoV networks.

References

[1] Y. Zhang, F. Tian, B. Song, and X. Du, “Social vehicle swarms: A novel perspective on socially aware vehicular communication architecture,” *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 82–89, Aug. 2016.

[2] O. Kaiwartya *et al.*, “Internet of vehicles: motivation, layered architecture, network model, challenges, and future aspects,” *IEEE Access*, vol. 4, pp. 5356–5373, May 2016.

[3] Z. Zhou, C. Gao, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez, “Social big-data-based content dissemination in internet of vehicles,” *IEEE Trans. Ind. Inform.*, vol. 14, no. 2, pp. 768–777, Feb. 2018.

[4] C. Wang, L. Zhang, Z. Li, and C. Jiang, “SDCoR: Software defined cognitive routing for internet of vehicles,” *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3513–3520, Oct. 2018.

[5] R. Silva and R. Iqbal, “Ethical implications of social internet of vehicles systems,” *IEEE Internet Things J.*, vol. 6, no. 1, pp. 517–531, Feb. 2019.

[6] C. Wang, C. Chou, P. Lin, and M. Guizani, “Performance evaluation of IEEE 802.15.4 nonbeacon-enabled mode for internet of vehicles,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3150–3159, Dec. 2015.

[7] W. Saad, Z. Han, A. Hjorungnes, D. Niyato, and E. Hossain, “Coalition formation games for distributed cooperation among roadside units in vehicular networks,” *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 48–60, Jan. 2011.

[8] J. Cheng, J. Cheng, M. Zhou, F. Liu, S. Gao, and C. Liu, “Routing in internet of vehicles: a review,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2339–2352, Oct. 2015.

[9] L. Liang, G. Y. Li, and W. Xu, “Resource allocation for D2D-enabled vehicular communications,” *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3186–3197, Jul. 2017.

[10] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, “A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3850–3860, Jun. 2018.

[11] “Scenarios, requirements and KPIs for 5G mobile and wireless system,” ICT-317669-METIS/D1.1, METIS deliverable D1.1, Apr. 2013. [Online]. Available: https://www.metis2020.com/documents/deliverables/

[12] K. Zheng, H. Meng, P. Chatzimisios, L. Lei, and X. Shen, “An SMDP-based resource allocation in vehicular cloud computing systems,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7920–7928, Dec. 2015.

[13] R. Yu, J. Ding, X. Huang, M. Zhou, S. Gjessing, and Y. Zhang, “Optimal resource sharing in 5G-enabled vehicular networks: a matrix game approach,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 7844–7856, Oct. 2016.

[14] L. Zhang, Z. Zhao, Q. Wu, H. Zhao, H. Xu, and X. Wu, “Energy-aware dynamic resource allocation in UAV assisted mobile edge computing over social internet of vehicles,” *IEEE Access*, vol. 6, pp. 56700–56715, 2018.

[15] N. Kumar, S. Misra, and M. S. Obaidat, “Collaborative learning automata-based routing for rescue operations in dense urban regions using vehicular sensor networks,” *IEEE Syst. J.*, vol. 9, no. 3, pp. 1081–1090, Sep. 2015.

[16] C. Chen, Q. Pei, and X. Li, “A GTS allocation scheme to improve multiple-access performance in vehicular sensor networks,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1549–1563, Mar. 2016.

[17] C. Wei, A. C. Huang, C. Chen, and J. Chen, “QoS-aware hybrid scheduling for geographical zone-based resource allocation in cellular vehicle-to-vehicle communications,” *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 610–613, Mar. 2018.

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

[19] K. Z. Ghafoor, L. Kong, D. B. Rawat, E. Hosseini, and A. S. Sadiq, “Quality of service aware routing protocol in software-defined internet of vehicles,” *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2817–2828, Apr. 2019.

[20] Q. Zheng, K. Zheng, H. Zhang, and V. C. M. Leung, “Delay-optimal virtualized radio resource scheduling in software-defined vehicular networks via stochastic learning,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 7857–7867, Oct. 2016.

[21] D. Lin, Y. Tang, Y. Yao, and A. V. Vasilakos, “User-priority-based power control over the D2D assisted internet of vehicles for mobile health,” *IEEE Internet Things J.*, vol. 4, no. 3, pp. 824–831, Jun. 2017.

[22] X. Cheng, L. Yang, and X. Shen, “D2D for intelligent transportation systems: A feasibility study,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1784–1793, Aug. 2015.

[23] W. Sun, E. G. Ström, F. Brännström, K. C. Sou, and Y. Sui, “Radio resource management for D2D-based V2V communication,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6636–6650, Aug. 2016.

[24] N. Cheng *et al.*, “Performance analysis of vehicular device-to-device underlay communication,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5409–5421, Jun. 2017.

[25] Z. Zhou, F. Xiong, C. Xu, Y. He, and S. Mumtaz, “Energy-efficient vehicular heterogeneous networks for green cities,” *IEEE Trans. Ind. Inform.*, vol. 14, no. 4, pp. 1522–1531, Apr. 2018.

[26] G. Yu, L. Xu, D. Feng, R. Yin, G. Y. Li, and Y. Jiang, “Joint mode selection and resource allocation for device-to-device communications,” *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3814–3824, Nov. 2014.

[27] W. Sun, D. Yuan, E. G. Ström, and F. Brännström, “Cluster-based radio resource management for D2D-supported safety-critical V2X communications,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2756–2769, Apr. 2016.

[28] D. Han, B. Bai, and W. Chen, “Secure V2V communications via relays: Resource allocation and performance analysis,” *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 342–345, Jun. 2017.

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

[30] X. Li, R. Shankaran, M. Orgun, L. Ma, and Y. Xu, “Joint autonomous resource selection and scheduled resource allocation for D2D-based V2X communication,” in *Proc. IEEE 87th Veh. Technol. Conf.*, 2018, pp. 1–5.

[31] F. Abbas and P. Fan, “A hybrid low-latency D2D resource allocation scheme based on cellular V2X networks,” in *Proc. Int. Conf. Commun. Workshops*, Kansas City, MO, USA, 2018, pp. 1–6.

[32] Q. Wei, W. Sun, B. Bai, L. Wang, E. G. Ström, and M. Song, “Resource allocation for V2X communications: A local search based 3D matching approach,” in *Proc. IEEE Int. Conf. Commun.*, Paris, France, 2017, pp. 1–6.

[33] W. Sun, D. Yuan, E. G. Ström, and F. Brännström, “Resource sharing and power allocation for D2D-based safety-critical V2X communications,” in *Proc. IEEE Int. Conf. Commun. Workshop*, London, U.K., 2015, pp. 2399–2405.

[34] R. S. Sutton, A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[35] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, “TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014.

[36] A. M. Kaushik, F. Hu, and S. Kumar, “Intelligent spectrum management based on transfer actor-critic learning for rateless transmissions in cognitive radio networks,” *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1204–1215, May 2018.

[37] Y. Saleem, K. A. Yau, H. Mohamad, N. Ramli, M. H. Rehmani, and Q. Ni, “Clustering and reinforcement-learning-based routing for cognitive radio networks,” *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 146–151, Aug. 2017.

[38] W. Liu, G. Qin, Y. He, and F. Jiang, “Distributed cooperative reinforcement learning-based traffic signal control that integrates V2X networks dynamic clustering,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 8667–8681, Oct. 2017.

[39] R. F. Atallah, C. M. Assi, and M. J. Khabbaz, “Scheduling the operation of a connected vehicular network using deep reinforcement learning,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1669–1682, May 2019.

[40] L. Xiao, T. Chen, C. Xie, H. Dai, and H. V. Poor, “Mobile crowdsensing games in vehicular networks,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1535–1545, Feb. 2018.

[41] C. Wu, T. Yoshinaga, X. Chen, L. Zhang, and Y. Ji, “Cluster-based content distribution integrating LTE and IEEE 802.11p with fuzzy logic and Q-Learning,” *IEEE Comput. Intell. Mag.*, vol. 13, no. 1, pp. 41–50, Feb. 2018.

[42] R. F. Atallah, C. M. Assi, and J. Y. Yu, “A reinforcement learning technique for optimizing downlink scheduling in an energy-limited vehicular network,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 4592–4601, Jun. 2017.

[43] H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu, and M. Wu, “Machine learning for vehicular networks: Recent advances and application examples,” *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 94–101, Jun. 2018.

[44] G. R. de Campos, P. Falcone, R. Hult, H. Wymeersch, and J. Sjöberg, “Traffic coordination at road intersections: autonomous decision-making algorithms using model-based heuristics,” *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 1, pp. 8–21, Spring 2017.

[45] H. Ye, G. Y. Li, and B. F. Juang, “Deep reinforcement learning based resource allocation for V2V communications,” 2018. [Online]. Available: https://arxiv.org/abs/1805.07222.

[46] Y. Wu *et al.*, “A learning-based QoE-driven spectrum handoff scheme for multimedia transmissions over cognitive radio networks,” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 11, pp. 2134–2148, Nov. 2014.

[47] *3rd Generation Partnership Project: Technical Specification Group Radio Access Network: Study LTE-Based V2X Services: (Release 14)*, Standard 3GPP TR 36.885 V2.0.0, Jun. 2016.

Helin Yang (S’15) received the B.S. and M.S. degrees from the School of Telecommunications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China, in 2013, and 2016, respectively. He is currently working toward the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is a Reviewer for IEEE international journals such as the IEEE COMMUNICATIONS MAGAZINE, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, etc. His current research interests include wireless communication, visible light communication, and resource management.

Xianzhong Xie received the Ph.D. degree in communication and information systems from Xidian University, Xi’an, China, in 2000. He is currently a Professor with the School of Optoelectronic Engineering and the Director of Chongqing Key Lab of Computer Network and Communication Technology at Chongqing University of Posts and Telecommunications, Chongqing, China. He is the principal author of five books on cooperative communications, 3G, MIMO, cognitive radio, and TDD technology. He has authored or coauthored more than 120 papers in journals and 30 papers in international conferences. His research interests include MIMO precoding, cognitive radio networks, and cooperative communications.

Michel Kadoch (S’86–M’91–SM’04) received the Ph.D. degree from Concordia University, Montreal, QC, Canada, in 1992. He is currently a Full Professor with Ecole de technologie superieure (ETS), University of Quebec, Montreal, QC, Canada. He is the Director of the research laboratory LAGRIT, ETS. He is also an Adjunct Professor with Concordia University. As the Principal Investigator, he has managed and participated actively in a research program on QoS for Multicast in high speed networks sponsored by Bell Canada and NSERC. He is currently working on reliable multicast in wireless ad hoc networks and 5G heterogeneous networks. He has authored or coauthored many articles and is the author of a book *Protocoles et réseaux locaux* (Edition ETS).