IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received January 2, 2020, accepted January 19, 2020, date of publication January 22, 2020, date of current version January 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968595

Deep Deterministic Policy Gradient (DDPG)-Based Resource Allocation Scheme for NOMA Vehicular Communications

YI-HAN XU[1,2], CHENG-CHENG YANG[1], MIN HUA[1], AND WEN ZHOU[1]
[1]College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China
[2]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia

Corresponding author: Yi-Han Xu (xuyihan@njfu.edu.cn)

This work was supported in part by the China Scholarship Council, and in part by the National Natural Science Foundation of China under Grant 61801225.

ABSTRACT This paper investigates the resource allocation problem in vehicular communications based on multi-agent Deep Deterministic Policy Gradient (DDPG), in which each Vehicle-to-Vehicle (V2V) communication acts as agent and adopts Non-Orthogonal Multiple Access (NOMA) technology to share the frequency spectrum that pre-allocated to Vehicle-to-Infrastructure (V2I) communications. Different with conventional D2D communications, the fast varying channel condition due to the high mobility in vehicular environment causes the difficulty of collecting instantaneous Channel State Information (CSI) at base station. Meanwhile, one tremendous challenge faced by vehicular communications is how to maximize the sum-rate of V2I communications simultaneously guaranteeing the latency and reliability requirements for the transmission of safety-critical information in V2V communications. In response, we formulate the resource allocation problem as a decentralized Discrete-time and Finite-state Markov Decision Process (DFMDP), in which allocation decisions are made by multiple agents that do not have complete and global network information. Due to the complexity of the problem, we propose a DDPG algorithm which is capable of handling continuous high dimensional action spaces to find the optimal allocation strategy. Numerical results verify that each agent can effectively learn from the environment by means of the proposed DDPG algorithm to maximize the sum-rate of V2I communications while satisfying the stringent latency and reliability constraints of V2V communications.

INDEX TERMS Vehicular communications, resource allocation, deep deterministic policy gradient (DDPG), non-orthogonal multiple access (NOMA).

I. INTRODUCTION

With the rapid development of sensors and wireless communication technologies, the emergence of vehicular communication aiming to enhance the experience of daily driving and paving the path to intelligent transportation and autonomous driving has attracted great attention from both industry and academia. Vehicular communications also referred to as vehicle-to-everything (V2X) communications, which include Vehicle-to-Infrastructure (V2I), Vehicle-to-Vehicle (V2V) and Vehicle-to-Pedestrian (V2P) connections [1]. Different types of vehicular communication links usually have different service requirements. For example, V2I communications concentrate mainly on sum-rate of data while V2V and V2P communications concern more on latency and reliability. Recently, several existing communication standards have been investigated in literature to realize vehicular communications, such as Dedicated Short Range Communications (DSRC) [2] and Intelligent Transportation System (ITS) [3], both of which are based on IEEE 802.11p standard. Moreover, in order to obtain the accurate Channel State Information (CSI) in vehicular communications, an IEEE 802.11p-based inter-vehicle cooperation channel estimation method was proposed in [4], in which the accurate

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Gao.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

CSI can be obtained by utilizing single or multiple vehicle measurements. Simulation results shown that the estimation performance and the safety-critical data transmission rate were improved. However, the latest studies in [5] and [6] have exposed some inherent defects of IEEE 802.11p-based standards in terms of scalability, mobility management and guaranteed Quality of Service (QoS) as its physical and link layers are originally designed for low mobility communications. To settle these shortcomings, the 3rd Generation Partnership Project (3GPP) has started projects to support diverse QoS requirements of V2X by exploiting the widely investigated Device-to-Device (D2D) communications in Long-Term Evolution (LTE) and upcoming 5G cellular networks [7]–[9]. Without loss of expectation, D2D communications have shown great potential in satisfying the QoS requirements in V2X and simultaneously supporting for high mobility [10]. Therefore, in this paper, we consider resource allocation for D2D-based vehicular communications, where each V2I communication shares frequency spectrum with multiple V2V communications and different vehicular links have diverse QoS requirements.

A. RELATED WORKS

In D2D communications, one major problem faced by both industry and academia is how to mitigate the mutual interference between D2D and cellular links and optimize resource utilization, and this problem will be more severe in the forthcoming heterogeneous cellular networks [11]. Thus, an effective resource allocation scheme is of great importance. In [12], a three-step method which jointly considers spectrum allocation and transmission power control was proposed to maximize the network throughput while guaranteeing the QoS of both D2D and cellular links. Simulation results validate that the method can significantly improve the performance in terms of D2D access rate and network throughput. However, this work did not consider the influence of mobility speed of users. Thus, it may not available to the vehicular environments. In [13], a more challenging research on utilizing D2D networks to deliver mobile video was investigated. The authors proposed to integrate the random mobility of users in mobile opportunistic D2D networks with crowdsourcing to push mobile video. In order to stimulate users, the authors modeled the interaction between depositories and provider as a reverse auction. Furthermore, an online auction algorithm was proposed to maximize the utility of the provider. Unfortunately, resource allocation schemes in conventional D2D communications are extensively studied towards low mobility scenarios [14]–[16], in which instantaneous CSI can be fully obtained. However, due to the high mobility in V2X communications, the fast varying channel condition suppresses the acquisition of high-precision CSI. The resource allocation scheme designed for D2D communications cannot be directly applied in V2X communications. The channel uncertainty issue should be carefully considered in implementing D2D-based V2X communications. Along this line of thought, several previous researches have been done related to resource allocation in V2X communications for different objectives, including network throughput, latency, outage probability and the tradeoff between different objectives. In [17], a location-based resource allocation scheme for D2D-based vehicular communications was proposed. In this scheme, authors considered the spatial resource that could be reused by D2D terminals to improve the sum-rate. However, in vehicular communications, the transmission latency and reliability are more important than sum-rate. Meanwhile, the authors also assume that the full CSI can be obtained, it is not reasonable. In [18], a resource allocation scheme utilizing the slowly varying large-scale fading information of channels to maximize the throughput of V2I links with the constraints of Signal-to-Interference plus Noise Ratio (SINR) of V2V links was proposed. However, the small-scale fading of the channel was not considered in this work. Contrarily, we consider ergodic Rayleigh small-scale fading for all sub-channels in our work. The resource allocation schemes in [19] and [20] can guarantee the reliability and latency requirements of V2V links. However, the authors only considered the transmission latency. The queueing latency which normally cannot be ignored was not taken into account. Although the influence of queueing latency in V2X communications was investigated in [21]–[23], the instantaneous CSI is assumed in most of existing literatures, which may not consistent with realistic situations.

Despite there are various existing investigations in literature intended to exploit traditional optimization methods to solve the resource allocation problems in V2X communications, they found difficulty in fully address the problem, in which diverse QoS requirements should be guaranteed. This type of optimization problem is normally NP-hard, and the optimal solutions are usually difficult to obtain. Fortunately, machine learning has shown impressive potential in addressing decision-making problem under uncertainty and it has been applied in communications and networks fields [24], [25]. In [26], a reinforcement learning framework was proposed to address the resource allocation problem in vehicular clouds to fulfill the dynamic and diverse requirements of different entities. However, this method only can handle low-dimensional state-action space, which is not realistic in vehicular network. Meanwhile, the computational complexity and the convergence speed of this method was not evaluated. A Markov Decision Process (MDP)-based pilot placement optimization problem in vehicular communications was investigated in [27], the authors developed an enhanced pilot placement scheme to jointly evaluate the dynamics of the channel state in time and frequency domains. Simulation results validated the effectiveness of the proposed pilot optimization policy. In [28], an online reinforcement learning based distributed vehicle association scheme in V2X communications was investigated to make load balancing among heterogeneous BSs. However, the spectrum allocation and transmission power control of each vehicle and BS were not described. A multi-agent reinforcement learning based spectrum sharing method was proposed in [29] to jointly optimize the sum capacity of V2I and payload delivery rate of V2V links in vehicular networks. The method leveraged on a proposed fingerprint-based Q-network algorithm to obtain the optimal solutions. Similarly, the authors of [30] propose a decentralized deep reinforcement learning based resource allocation scheme for V2V communications. This scheme enables to find the optimal sub-band and power level without the global network information. Meanwhile, authors demonstrated that the proposed scheme could be applied to both unicast and broadcast scenarios. However, both [30] and [31] assumed that the selectable transmission powers are discrete and limited to four power levels to control the action space with a low dimension. In practice, many applications in vehicular communications have continuous action space, in which deep reinforcement learning based scheme cannot be straightforwardly applied.

B. CONTRIBUTIONS

In this paper, we propose a resource allocation scheme to support concurrent V2I and V2V communications based on Mode 4 specified in 3GPP cellular V2X framework, in which vehicles have a pool of orthogonal frequency spectrum resources that can be selected for V2X communications [32]. In order to improve the spectrum efficiency, we suppose that the sidelink V2I spectrum can be reused by V2V communications with Uu interface under proper interference management. The contributions of this work are summarized in the followings.

- We consider a resource allocation problem for V2X communications to maximize the sum-rate of V2I communications and guarantee the latency and reliability of V2V communications with the joint consideration for both frequency spectrum allocation and transmission power control.
- We formulate the resource allocation problem to be a decentralized Discrete-time and Finite-state Markov Decision Process (DFMDP), in which the state space, action space and the reward function are tactfully designed. In addition, to keep the realistic, we consider a continuous action space, and thus a Deep Deterministic Policy Gradient (DDPG) framework is proposed to solve the problem.
- Through numerical analysis, we validate that the V2V communications enable to learn from the interaction with environment and figure out the optimal strategy to enhance the network performance.

The remainder of this paper is organized as follows. Our system model is presented in Section II. After that, the corresponding resource allocation problem is formulated and the proposed DDPG framework is elaborated in Section III. In Section IV, the simulation setting and results are discussed. Finally, we give the conclusions in Section V

II. SYSTEM MODEL DESCRIPTION

In this section, we first present the system model of the V2X communications. After that, we propose a resource allocation scheme which jointly considers frequency spectrum allocation and transmission power control to maximize the sum-rate of V2I communications meanwhile guaranteeing the latency and reliability of V2V communications. We formulate the resource allocation problem to be a decentralized DFMDP by designing state space, action space and the reward function, respectively.

A. SYSTEM MODEL

In this treatise, the scenario we considered includes a single cellular based vehicular network which is illustrated by Figure 1. The multi-cells based vehicular network scenario which may involve mobility management mechanism is out of scope of this work. Base Station (BS) is located at the center of the cell with radius R, while M V2I communications are denoted as $V2IC = \{1, 2, \ldots, M\}$ and N pairs of V2V communications are denoted as $V2VC = \{1, 2, \ldots, N\}$. Each V2V pair has a transmitter (V2V_TX) and a receiver (V2V_RX). We assume that each V2I communication uses pre-assigned orthogonal cellular uplink spectrum resource. The total available bandwidth is denoted as B_{total}, and B_{total} is divided into M sub-band with the bandwidth of B_{total}/M allocated to each V2I communication. To improve the spectrum efficiency, we suppose each V2V_TX with the NOMA technology can reuse the uplink spectrum that pre-occupied by V2I communications. This assumption is reasonable because of the less intensively use of uplink resource and the stronger anti-interference ability of BS. For simplicity, we suppose each V2V pair cannot multiplex more than one uplink spectrum assigned to V2I communications at a time and the spectrum of each V2I communication can be shared by maximum number of z V2V pairs simultaneously. Therefore, we define a binary parameter $\rho_{V2IC_i,V2VC_j} \in \{0,1\}\,(\forall i \in M, \forall j \in N)$ indicates whether the j-th V2V pair reuses the spectrum resource that pre-allocated to i-th V2I communication. Thus, the assumption can be expressed as Equations 1 and 2:

$$\sum_{V2IC=1}^{M} \rho_{V2VC_j,} \leq 1, \quad (\forall j \in N) \tag{1}$$

$$\sum_{V2VC=1}^{N} \rho_{V2IC_i,} \leq z, \quad (\forall i \in M) \tag{2}$$

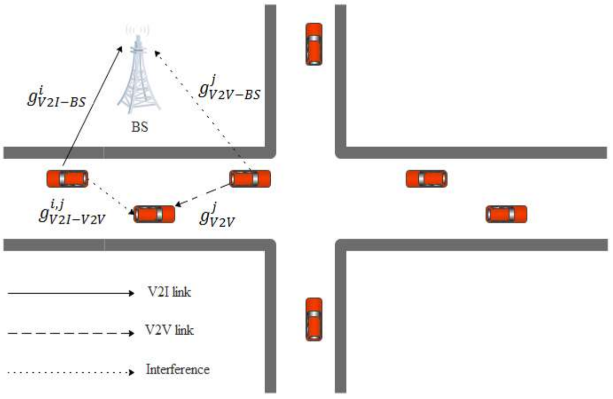


FIGURE 1. System model.

Furthermore, we take the mutual interference between V2I and V2V communications into consideration and derive the SINR of i-th V2I communication and j-th V2V communication over the m-th sub-band ($\forall m \in M$) as Equations 3 and 4, respectively:

$$SINR^i_{V2I}[m] = \frac{p^i_{V2I} \cdot g^i_{V2I-BS}[m]}{\sum_{j\in N} \rho_{V2IC_i} \cdot p^j_{V2V} \cdot g^j_{V2V-BS}[m] + n_0} \tag{3}$$

$$SINR^j_{V2V}[m] = \frac{p^j_{V2V} \cdot g^j_{V2V}[m]}{I^j_{V2V}[m] + n_0} \tag{4}$$

where, p^i_{V2I} and p^j_{V2V} are the transmission powers of i-th V2I transmitter and j-th V2V_TX over m-th sub-band. The set of p^i_{V2I} and p^j_{V2V} is defined to transmission power control strategy. $g^i_{V2I-BS}[m]$ is the channel gain of i-th V2I communication over m-th sub-band. $g^j_{V2V-BS}[m]$ is the channel gain between j-th V2V_TX and BS over m-th sub-band. $g^j_{V2V}[m]$ is the channel gain of j-th V2V communication over m-th sub-band. n_0 is the noise power, which equals to $\frac{B_{total}}{M} \cdot \rho_n$, where ρ_n is the density of noise. $I^j_{V2V}[m]$ is the total interference to the j-th V2V communication on m-th sub-band which can be expressed as Equation (5):

$$\begin{aligned} I^j_{V2V}[m] &= p^i_{V2I} \cdot g^{i,j}_{V2I-V2V}[m] \\ &\quad + \sum_{\substack{k\in N \\ V2V_k \neq V2V_j}} \rho_{V2VC_k} \cdot p^k_{V2V} \cdot g^{k,j}_{V2V-V2V}[m] \end{aligned} \tag{5}$$

where, $g^{i,j}_{V2I-V2V}[m]$ is the channel gain between i-th V2I communication and j-th V2V communication over m-th sub-band. $g^{k,j}_{V2V-V2V}[m]$ is the channel gain between k-th V2V communication and j-th V2V communication over m-th sub-band.

Specifically, due to the instantaneous CSI cannot be acquired accurately in vehicular communications, especially while the vehicle mobility speed is high, in this model, the channel gain takes into account of both large-scale fading and small-scale fading. The channel gain on m-th sub-band $g[m]$ can be represented as Equation (6):

$$g[m] = \alpha \cdot h[m] \tag{6}$$

where, α denotes the large-scale fading which includes path loss and shadowing. It can be modeled as $\alpha = PL \cdot \beta (d_{TX-TR})^{-\varphi}$. PL is the path loss constant, β denotes the log-normal shadowing random component, d_{TX-TR} is the distance between transmitter and receiver in a communication link. φ is the power decay exponent. $h[m]$ denotes the small-scale fading which is assumed as ergodic Rayleigh small-scale fading for all sub-channels in this model.

Let $TR^i_{V2I}[m]$ and $TR^j_{V2V}[m]$ are the sum-rates of i-th V2I communication and j-th V2V communication over m-th sub-band. According to Shannon's theorem, $TR^i_{V2I}[m]$ and $TR^j_{V2V}[m]$ can be expressed as

$$TR^i_{V2I}[m] = \frac{B_{total}}{M} \cdot log_2\left(1 + SINR^i_{V2I}[m]\right) \tag{7}$$

and

$$TR^j_{V2V}[m] = \frac{B_{total}}{M} \cdot log_2\left(1 + SINR^j_{V2V}[m]\right) \tag{8}$$

As mentioned in Section I, the goal of this paper is to design a resource allocation scheme to maximize the sum-rate of V2I communications which can be defined as $\sum_{i=1}^{M} \sum_{m=1}^{M} TR^i_{V2I}[m] (\forall i \in M, \forall m \in M)$ to support the high transmission rate entertainment applications. Meanwhile, due to the fact that the V2V communications concentrate mainly on stringent latency and reliability requirements for the safety-critical information while the sum-rate is not significant important. It is difficult to formulate the latency and reliability constraints directly into the optimization problem as it may involve more assumptions and increase the computation complexity. For the sake of convenience, we convert the latency and reliability constraints of V2V communications into a metric of delivery probability which takes both latency and reliability into account. More specifically, we define the delivery probability as the probability of successfully deliver the payload of V2V communications with the size of B bits within an acceptable time T. The delivery probability can be expressed as Equation (9):

$$Prb\left\{\sum_{t=1}^{T} \sum_{m=1}^{M} TR^j_{V2V}[m, t] \geq \frac{B}{T_{cct}}\right\}, \quad \forall j \in N \tag{9}$$

where, B denotes the size of the payload generated by V2V_TX which depends on the vehicle mobility and the advancement of driving services supported by the vehicle. T_{cct} is the channel coherence time in Orthogonal Frequency Division Multiplexing (OFDM) system. Practically, we assume that the buffer of each vehicle has the finite size and served in first in first out fashion. Moreover, we can obtain the update function for the buffer queue length of j-th V2V transmitter as Equation 10:

$$\begin{aligned} TL^t_j = \min\Bigg\{ & TL^{max}_j, TL^{t-1}_j - min\left\{\sum_{m=1}^{M} TR^j_{V2V}[m, t], TL^{t-1}_j\right\} \\ & + B^{t-1}\Bigg\} \end{aligned} \tag{10}$$

where, TL^t_j is the packet queue length of *j-th* V2V_TX in t-th coherence time slot, TL^{max}_j is the maximum queue length of *j-th* V2V_TX

To this end, the problem of maximizing sum-rate of V2I communications simultaneously guaranteeing the stringent latency and reliability of V2V communications can be formulated as:

$$\begin{aligned} \max_{\substack{\{\rho V2IC_i, V2IC_j\} \\ \{p^i_{V2I}\}\{p^j_{V2V}\}}} & \sum_{i\in M} \sum_{m=1}^{M} TR^i_{V2I}[m] \\ &= \sum_{i\in M} \sum_{m=1}^{M} \frac{B_{total}}{M} \cdot log_2\left(1 + SINR^i_{V2I}[m]\right) \end{aligned} \tag{11a}$$

$$\text{Subjects to : } \sum_{V2IC=1}^{M} \rho_{V2VC_j,} \le 1,$$

$$\times \sum_{V2VC=1}^{N} \rho_{V2IC_i,} \le z, \forall j \in N, \quad \forall i \in M \tag{11b}$$

$$\times \sum_{m=1}^{M} \frac{B_{total}}{M} \cdot log_2 \left(1 + SINR^{i}_{V2I}\left[m\right]\right) \ge R_{th} \quad \forall i \in M \tag{11c}$$

$$Prb \left\{ \sum_{t=1}^{T} \sum_{m=1}^{M} TR^{j}_{V2V}\left[m, t\right] \ge \frac{B}{T_{cct}} \right\}, \quad \forall j \in N \tag{11d}$$

$$TL_j^t = \min \left\{ TL_j^{max}, TL_j^{t-1} - min \left\{ \sum_{m=1}^{M} TR^{j}_{V2V}\left[m, t\right], TL_j^{t-1} \right\} + B^{t-1} \right\} \tag{11e}$$

$$0 \le p^{i}_{V2I} \le p^{max}_{V2I} \quad \forall i \in M \tag{11f}$$

$$0 \le p^{j}_{V2V} \le p^{max}_{V2V} \quad \forall j \in N \tag{11g}$$

where, (11b) means that each V2V communication cannot reuse more than one spectrum resource assigned to V2I communications and the spectrum resource that pre-allocated to each V2I communication can be shared by maximum number of z V2V pairs at a time. (11c) depicts the sum-rate of each V2I communication should achieve the minimum required sum-rate threshold R_{th}. (11d) gives the constraint on delivery probability which indicates that the probability of the V2V payload with size of B bits be delivered within a certain time should be larger a delivery threshold. (11e) denotes the update process of the instantaneous buffer queue length of V2V_TX, the status of the queueing packets makes a significant influence on the latency, thus we should take packets queueing latency into account for the V2V communications. (11f) and (11g) presents the transmission powers of both V2I and V2V transmitters cannot exceed the maximum power level to avoid the serious interference.

B. DFMDP MODEL

Typically, a DFMDP can be defined by a tuple (S, A, p, r), in which S is a finite set of states, A is a finite set of actions, p is a transition probability from state s *to* state $s'(\forall s \in S, \forall s' \in S)$ after action $a(\forall a \in A)$ is performed and r is the immediate reward obtained after $a(\forall a \in A)$ is executed. We denote π as a policy which is a mapping from a state to an action. The goal of the proposed DDPG framework is to find the optimal policy as denoted π^* to maximize the reward function over a finite time horizon in the DFMDP. In the meantime, it is worth mentioning that the basic framework of reinforcement learning algorithm consists of two components which are agent and environment. In this model, each V2V communication is considered as an agent and interacts with the unknown environment to obtain experiences, which are then iteratively learned to get its optimal policy. Since, each agent explores environment in a distributed fashion and makes strategies decision based on their own observations. Therefore, we formulate the resource allocation problem into a decentralized DFMDP. The detailed tuple in our proposed model are defined as follows:

1)The state of each individual V2V communication can be denoted as $s_j \in S$, which can only be acquired through their own observation from the environment. In this model, we define the state space is a set of states can be observed by V2V communications, which includes the local channel information such as $g^{j}_{V2V}[m]$, interference channels from other V2V communication $g^{k,j}_{V2V-V2V}[m]$, and interference channel from own transmitter to the BS $g^{j}_{V2V-BS}[m]$, and interference channel from all V2I transmitters $\sum_{i=1}^{M} g^{i,j}_{V2I-V2V}[m]$ and the state of queue length in the buffer of each V2V_TX TL^t_j. To ensure the completeness of the exploration of state space, TL^t_j is specified to be an integer and take the values of $[0, 1, \ldots, TL^{max}_j]$.

2) The action $a(\forall a \in A)$ in this scenario should be the resource allocation variables. In this model, we define the action space is a set of actions can be taken by V2I and V2V communications that including the transmission power (p^i_{V2I} and p^j_{V2V}) and spectrum multiplexing factor ($\rho_{V2IC_i,V2VC_j}$). In this work, we assume that the total spectrum is divided into M orthogonal sub-bands, each of which is pre-allocated to one V2I communication. By means of $\rho_{V2IC_i,V2VC_j}$, we can obtain the spectrum allocation strategy. However, the transmission power practically takes continuous value even if some existing works in literature made assumption that the transmission power can be discrete [29], [30]. More realistic, the transmission power is continuous ranges from $[-100, p^{max}]$ dBm in this work. It should be noted that the choice of -100 dBm effectively represents the transmission power is 0. Consequently, the dimension of the action space is continuous which results the conventional Deep Q-Network (DQN) in deep reinforcement learning without the capability to handle with the action space. To solve this obstacle, a DDPG framework is proposed in this paper.

3)Obviously, the purpose of making resource allocation for V2X communications is to select the proper spectrum bands and transmission powers that optimize the different QoS requirements of both V2I and V2V communications. Hence, the reward r should consider two aspects: the sum-rate of V2I communications as expressed in Equation (11a) and the delivery probability of V2V communications which is given in Equation (11d).

However, the reward value-based algorithms such as Monte Carlo [32] and Temporal Difference (TD) [33] algorithms have some shortcomings in practical applications, for instance they cannot process the tasks in continuous action space efficiently and the final solution may not be global optimal. From above analysis, we intend to adopt a policy-based algorithm in this paper. The goal of our proposed DDPG framework is to find out the optimal policy $\pi^*(s_j) \rightarrow A$ for each state in each V2V communication complete state-action space.

III. DDPG FRAMEWORK FOR RESOURCE ALLOCATION

To address the decentralized DFMDP problem, classical Q-learning and deep Q-network algorithms are effective tools. The core idea behind Q-learning algorithm is define value function $V^{\pi}(s_j) \rightarrow r$ to represent the expected value can be obtained by policy π from each state $s_j \in S$. The value function V^{π} quantifies the goodness of the policy π via an infinite horizon and discounted MDP that can be represented as Equation 12:

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}_{\pi}[\sum_{k=0}^{\infty} \gamma \cdot r^k\left(s^k, a^k\right)|s^0 = s] \\ &= \mathbb{E}_{\pi}[r^k(s^k, a^k) + \gamma \cdot V^{\pi}(s^{k+1})|s^0 = s] \end{aligned} \tag{12}$$

Because our goal is to find out the optimal policy π^*, the optimal action at each state can be found by means of the optimal value function as Equation 13:

$$V^*(s) = \max_{a^k}\left\{\mathbb{E}_{\pi}\left[r^k\left(s^k, a^k\right) + \gamma \cdot V^{\pi}\left(s^{k+1}\right)\right]\right\} \tag{13}$$

If we denoted $Q^*(s,a) \triangleq r^k\left(s^k, a^k\right) + \gamma \cdot E_{\pi}[V^{\pi}\left(s^{k+1}\right)]$ as the optimal Q-function, the optimal value function can be rewritten by $V^*(s) = \max_a\{Q^*(s,a)\}$. The $Q^*(s,a)$ can be obtain through iterative process according to the Equation 14:

$$\begin{aligned} &Q^{k+1}\left(s^k, a^k\right) \\ &\quad = Q^k\left(s^k, a^k\right) \\ &\quad +\alpha\left[r^k\left(s^k, a^k\right) + \gamma max Q^k\left(s^k, a^{k+1}\right) - Q^k\left(s^k, a^k\right)\right] \end{aligned} \tag{14}$$

where, α is the learning rate to determine the impact of new information to the existing Q-value, $\gamma \in [0, 1]$ is the discount factor.

However, the Q-learning algorithm can get the optimal policy when the state-action spaces are discrete and the dimension is small. As a result, Q-learning algorithm may insufficient to find the optimal policy within the acceptable time practically. Benefits from the considerably investigation of deep learning techniques, reinforcement learning has shown a significant enhancement by adopting DQN instead of the Q-table in original Q-learning algorithm to derive the approximate value of $Q(s^k, a^k)$. Therefore, the Q-value of DQN in time slot k can be rewritten as $Q(s^k, a^k, \theta)$, where θ is the weight of Deep Neural Network (DNN). After the approximation, the optimal policy π^*(s) will be presented by Equation 15:

$$\pi^*(s) = arg\ \max_{a^k} Q^*\left(s^k, a^{k+1}, \theta\right) \tag{15}$$

where, $Q^*(s,a)$ is the optimal Q-value via DNN approximation. DQN will choose the approximated action $a^{k+1} = \pi^*\left(s^{k+1}\right)$. Then the approximated $\tilde{Q}(s^k, a^k)$ can be given as Equation 16:

$$\begin{aligned} &\tilde{Q}\left(s^k, a^k, \omega\right) \\ &\quad = r\left(s^k, a^k, \omega\right) + \gamma max_{a^{k+1}}\left[\mathrm{Q}\left(s^{k+1}, a^{k+1}, \theta\right)\right] \end{aligned} \tag{16}$$

The value of θ is updated by minimizing the loss as expressed in Equation 17.

$$L = E\left[\left(\tilde{Q}\left(s^k, a^k, \omega\right) - Q\left(s^{k+1}, a^{k+1}, \omega\right)\right)^2\right] \tag{17}$$

Algorithm 1 gives the DQN-based resource allocation algorithm

Algorithm 1 The DQN-Based Resource Allocation Algorithm

```
initialize replay memory $D$ to 10000
initialize the Q-network Q with random weights $\omega$
for $episode = 1$ to $M$ do
  Initialize the V2I and V2V communication scenario,
    receive initial observation state $s_1$
  for $k = 1$ to $K$ do
    select a random action $a^k$ (actions are sub-band allocation
    and transmission power control with the probability $\varepsilon$
    Otherwise select $a^k = argmaxQ^*(s^k, a^k, \omega)$
    perform action $a^k$ and observe immediate reward $r^k$
    and next state $s^{k+1}$
    store transition $(s^k, a^k, r^k, s^{k+1})$ in $D$
    select randomly samples c$(s_i, a_i, r_i, s_{i+1})$ from $D$
    the weights of the of DNN are updated by using
    stochastic gradient descent with respect to the $\omega$ to
    minimize the loss as Equation 14
    update the policy $\pi\left(s^k\right) = \arg \max_{a^{k+1}} Q^*\left(s^k, a^{k+1}, \omega\right)$
    after every a fixed number of steps
  end for
end for
```

Despite the deep reinforcement learning algorithm is superior to the classical Q-learning algorithm as it enables to solve problems with high-dimensional state spaces, but it still cannot handle with the situation in which the action space is continuous. This is because the deep reinforcement learning algorithm relies on the selection of the best action to maximize the Q-value. However, it is infeasible to find the optimal action in a continuous space. In the light of this, we propose to use DDPG framework to optimize the resource allocation problem for V2X communications in this work. DDPG algorithm is first proposed in [34], in which the authors introduce a model-free off-policy actor-critic algorithm using DNN to learn policies in continuous action space. In fact, the fundamental concept of DDPG algorithm is to integrate DNN into Deterministic Policy Gradient (DPG) algorithm proposed in [35] to improve the learning efficiency. The key idea of DPG algorithm is the policy function and Q-value function can be approximated by two networks, namely, actor network and critic network, respectively. These two functions maintain a parameterized actor function $\mu\left(s;\theta^{\mu}\right)$ with parameter vector θ which specifies the current policy by deterministically mapping states to a specific action. The critic $Q\left(s, a\right)$ is learned by using the Bellman equation as in Q-learning. The actor is updated by applying the chain rule to the expected return from the start distribution J. The update of actor network can be derived as Equation 18. It is valuable to mention that the exploration is the major challenge for the learning in a continuous action space, which requires the target values should be updated slowly to improve the learning stability. Therefore, the θ^{μ} and θ^{Q} in target networks should be updated slowly with the tracking on the learned networks.

$$\begin{aligned}\nabla_{\theta^{\mu}} J &\approx \mathbb{E}[\nabla_{\theta^{\mu}} Q(s, a; \theta^{Q})|_{s=s^k, a=\mu(s^k|\theta^{\mu})}] \\ &\approx \mathbb{E}\left[\nabla_a Q\left(s, a; \theta^{Q}\right)|_{s=s^k, a=\mu(s^k)} \nabla_{\theta^{\mu}} \mu\left(s; \theta^{\mu}\right)|_{s=s^k}\right]\end{aligned} \tag{18}$$

Based on this update rule, DDPG algorithm was introduced to learn competitive policies by using DNN. In DDPG algorithm, we can rewrite the policy function to $a = \pi(s; \theta^{\mu})$ and Q-value function to $Q(s, a; \theta^{Q})$, in which θ^{μ} and θ^{Q} represent the weight of DNN in actor network and critic network, respectively. Without the loss of generality, the goal of the policy function is to maximize the long-term cumulative reward expectation with discount rate of γ as expressed in Equation 19:

$$\begin{aligned}\mu^* &= argmax_{\mu} J\left(\mu\right) \\ &= argmax_{\mu} \mathbb{E}_{\mu}\left[r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots + \gamma^{k-1} r_k\right]\end{aligned} \tag{19}$$

We present the proposed DDPG-based resource allocation algorithm in-detail in Algorithm 2.

Algorithm 2 The DDPG-Based Resource Allocation Algorithm

```
1. Initialize replay memory $D$ to 10000 and mini-batch to 200
2. Randomly initialize the weights of actor network $\theta^{\mu}$ and critic network $\theta^{Q}$, respectively
3. Initialize the target network with weights $\theta^{\mu'} \leftarrow \theta^{\mu}$ and $\theta^{Q'} \leftarrow \theta^{Q}$, respectively
4. for $episode = 1$ to $M$ do
5.   Initialize the V2I and V2V communication scenario, receive initial observation state $s_1(g^{j}_{V2V}[m], g^{k,j}_{V2V-V2V}[m], g^{j}_{V2V-BS}[m], \sum_{i=1}^{M} g^{i,j}_{V2I-V2V}[m]$, and $TL^{t}_{j})$
6.   for $k = 1$ to $K$ do
7.     Select a random action $a^k = \mu\left(s^k \mid \theta^{\mu}\right) + N^k$ (actions are sub-band allocation and transmission power control, $N^k$ is the exploration noise)
8.     Perform action $a^k$, get the immediate reward $r^k$ and next state $s^{k+1}$ store transition $(s^k, a^k, r^k, s^{k+1})$ in $D$
9.     if the replay memory $D$ is full, do 10. Randomly sample mini-batch of $C$ transitions
         $(s^k, a^k, r^k, s^{k+1})$ from $D$
11.      Set $\tilde{Q}\left(s^k, a^k \mid \theta^{Q}\right) = r^k + \gamma Q(s^{k+1}, \mu(s^k|\theta^{\mu'})|\theta^{Q'})$
12.      Update the $\theta^{Q}$ in critic network by minimizing the loss:
13.      $L = \frac{1}{N}\sum_{k}(\tilde{Q}\left(s^k, a^k \mid \theta^{Q'}\right) - Q\left(s^k, a^k \mid \theta^{Q}\right))^2$
14.      Update the $\theta^{\mu}$ in actor network by using the sampled policy
         gradient: $\nabla_{\theta^{\mu}} J \approx \frac{1}{N}\sum_{k} \nabla_a Q\left(s^k, a^k \mid \theta^{Q}\right) \nabla_{\theta^{\mu}} \mu(s^k|\theta^{\mu})$
15.      Update the target networks:
           $\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}$
           $\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$
16.    end if
17.  end for
end for
```

IV. SIMULATION RESLUTS AND ANALYSIS

In this section, performance of the proposed DDPG-based resource allocation scheme and the features of V2X communications are investigated. Simulation studies are carried out by using MATLAB and Tensorflow 1.0 to evaluate the performance of the proposed scheme with other two comparison schemes: (1) deep reinforcement learning based resource allocation scheme and (2) random resource allocation scheme. To verify the effectiveness of the proposed scheme, we evaluate the performance metric in terms of sum-rate of V2I communications and the delivery probability of the V2V communications in various different network settings.

A. SIMULATION SETTING

In simulations, we consider a scenario that includes a single cellular network with the radius of 1000 m. BS is located at the center of this topology with the carrier frequency of 2 GHz. The setting of this simulation is based on the Manhattan case specified in 3GPP TR 36.885 [36], in which the vehicle drop models, speeds, densities, and vehicular channels are described in detail. In this simulation, there are total 9 blocks with both the line-of-sight (LOS) and non-line-of-sight (NLOS) channels. The vehicles are dropped on the lane randomly with the mobility speeds range from 30 km/h to 60 km/h. The maximum distance of each V2V pair is set to 200m. M V2I communications are conducted by using M pre-allocated orthogonal cellular uplink sub-band. N V2V communications are formed between two nearby vehicles by reusing the sub-band that pre-occupied to V2I communications. Meanwhile, we assume that the number of N is three times of the number of M. This assumption is to ensure that all the sub-bands are fully reused by V2V communications and it also can verify the robustness of the proposed scheme. In our DDPG framework, we set 200 time instants for each episode and each performance metric will be averaged to reduce the instability. The DNN utilized in both actor and critic networks contain three fully connected hidden layers, in which 500, 250 and 120 neurons are set respectively. The activation function of rectified linear unit (ReLU) is adopted which is defined as Equation 20.

$$f(x) = max(0, x) \tag{20}$$

The learning rate is set to 0.01 initially and decreases exponentially to 0.001. The ϵ-greedy policy is used to balance the exploration and exploitation with the value of ϵ is set linearly annealed from 1 to 0.02 [37]. For each configuration, we generate 100 independent runs and average the performance of sum-rate of V2I communications and delivery probability of V2V communications. All of the detailed simulation variables used in this paper is summarized in Table 1.

B. RESULTS AND ANALYSIS

Figure 2 illustrates the comparison of the optimization process for sum-rate of V2I communications between the proposed DDPG-based algorithm and DQN-based algorithm. In this simulation, we set the number of vehicles to 60. The simulation result gives an observation that DQN-based algorithm performs better than DDPG-based algorithm before 42 episodes. After 42 episodes, the DDPG-based algorithm starts to outperform the DQN-based algorithm owning to the implementing of DNN in both actor and critic networks. It is worth noting that the DDPG-based algorithm is unstable initially. However, as the episodes increase, the performance trends to stable. Meanwhile, the DDPG-based algorithm outperforms the DQN-based algorithm for over approximate 26% after 120 episodes. Moreover, it is clear that DDPG-based algorithm performs quite stable after 71 episodes rather than 120 episodes for the DQN-based algorithm in this scenario, which indicates that DDPG-based algorithm achieves convergence faster than the DQN-based algorithm.

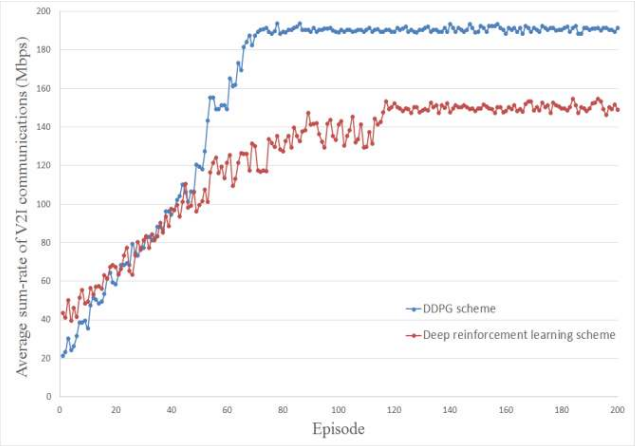


FIGURE 2. **The optimization process for average sum-rate of V2I communications.**

TABLE 1. **Simulation parameters setting [36].**

Parameters	Value
R	1000 m
Distance of each V2V pair	Random distributed in [20, 200] m
Number of V2I communications M	[5:1:30]
Number of V2V communications N	[8:2:46]
z	3
Carrier frequency	2 GHz
Bandwidth per channel	1.5 MHz
BS antenna height	25 m
BS antenna gain	8 dBi
BS receiver noise figure	5 dBi
Vehicle antenna height	1.5 m
Vehicle antenna gain	3 dBi
Vehicle receiver noise figure	9 dBi
Vehicle speeds	Random distributed in [30, 60] km/h
V2I transmission power p_{V2I}^{i}	Range from [-100, 23] dBm
V2V transmission power p_{V2V}^{j}	Range from [-100, 23] dBm
ρ_n	-174 dBm/Hz
V2V payload size B	[0.4:0.2:1.6] Mb
TR_{th}	1 bps/Hz
Latency constraint of V2V payload delivery T_{cct}	100 ms
Path loss model of V2I communications	$128.1 + 37.6log_{10}d$ d in km
Path loss model of V2V communications	LOS in WINNER +B1 Manhattan [38]
Shadowing distribution of V2I communications	Log-normal
Shadowing distribution of V2V communications	Log-normal
Fast fading	Rayleigh fading
Number of neurons in first hidden layer	500
Number of neurons in second hidden layer	250
Number of neurons in third hidden layer	120

Figure 3 presents the sum-rate of V2I communications versus the different number of vehicles. From the results, it can be observed that as the increase of the number of vehicles, the sum-rate of V2I communications decrease for all schemes. This is because the number of vehicles increase results the number of V2V communications increase, thus the interference from V2V communications to V2I communications grow which causes the drop of sum-rate of V2I communications. However, it is clear that the proposed DDPG scheme achieves the highest sum-rate of V2I communications and random resource allocation scheme with the worst sum-rate

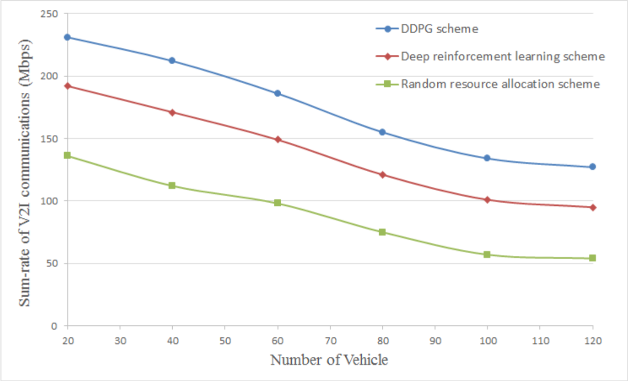


FIGURE 3. Average sum-rate of V2I communication versus different number of vehicle.

of V2I communications. This is due to the fact that the random resource allocation scheme allocates sub-band and transmission power randomly which generates catastrophic mutual interference between V2I and V2V communications. Furthermore, as compared with deep reinforcement learning scheme, even if deep reinforcement learning scheme can find the optimized resource allocation strategy, but the strategy may not be the global optimum as the states of transmission powers are continuous. It is worth noting that the proposed DDPG scheme enables to find the optimal resource allocation strategy with the probability 100%.

Figure 4 gives the average delivery probability of V2V communication while different numbers of vehicles are deployed. In this simulation, we set the payload *B* for each V2V communication with the constant size of 1Mb. From the results, we can find that as the number of vehicle increases, the average delivery probabilities decrease for all schemes, including the proposed DDPG scheme. This is because that more vehicles are deployed will cause more mutual interference between V2I and V2V communications. However, the proposed DDPG scheme still gives the better performance as compared to other two schemes throughout the tested cases. Remarkably, the proposed DDPG scheme is capable of guaranteeing the average V2V delivery probability above 90% even in the worst case in which 120 vehicles are deployed. Moreover, in conjunction with the results from Figure 4, we can validate the robustness of the proposed DDPG scheme against the variation of the number of vehicles meanwhile satisfying the requirements of both V2I and V2V communications.

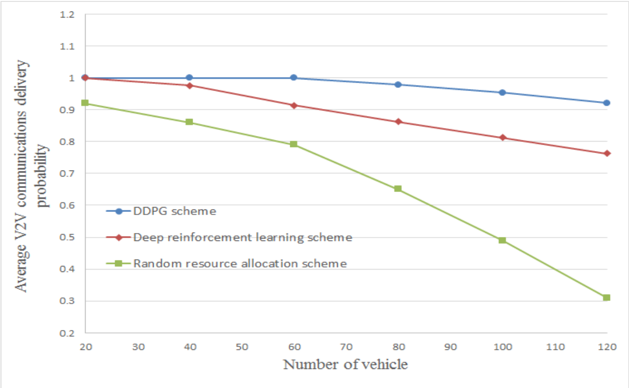


FIGURE 4. Average delivery probability of V2V communication versus different number of vehicle.

Figure 5 depicts the average delivery probability of V2V communication versus different payload size *B* of each V2V communication while the number of vehicle is set to 60. We can observe from the results that the proposed DDPG scheme gets highly desirable performance throughout all the cases. This is due to fact that the proposed DDPG scheme always enables to find the optimal sub-band spectrum allocation and transmission power control strategy to maintain the delivery probability of V2V communications. The worst case is that when the payload size of each V2V communication increases to 1.6Mb, the average delivery probability still achieves 95.6%. In other words, the proposed DDPG scheme has the capability to guarantee the latency and reliability requirements for V2V communications in a high mobility vehicular network. Looking at other two schemes, deep reinforcement learning scheme has the second performance with the delivery probability above 80%. The worst case appear while the payload size increases to 1.6Mb, as the payload size increases, due to the deep reinforcement learning scheme cannot fully guarantee the latency of each communication, and thus the delivery probability will drop. Similarly, the random resource allocation scheme has the worst performance as it allocates sub-band spectrum and transmission power randomly.

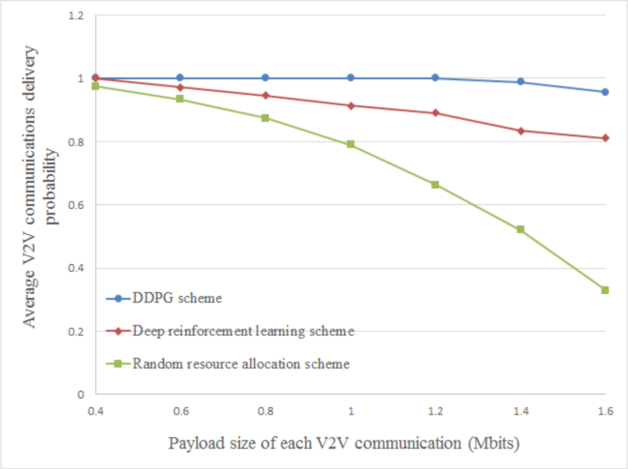


FIGURE 5. Average delivery probability of V2V communication versus different number of vehicle.

Figure 6 illustrates the comparison of the average sum-rate of V2I communications among the proposed DDPG scheme, deep reinforcement learning scheme and random resource allocation scheme while different mobility speeds are set to the vehicles in V2I communications. The velocity of vehicles in V2V communications are randomly distributed between 30 km/h and 60 km/h. From the results, we can observe

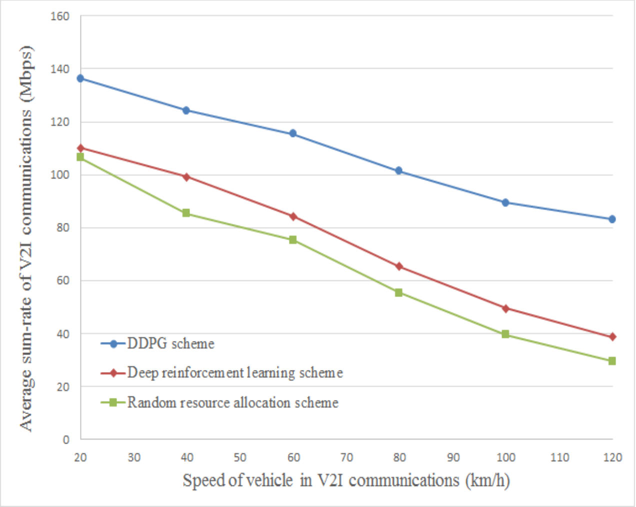


FIGURE 6. **Average sum-rate of V2I communication versus different speeds of vehicle.**

that the average sum-rate of V2I communications decreases as the vehicle speed increases. This is due to the growing vehicle speed causes the fast varying of channel condition which results in more packet loss meanwhile the growing speed generates more interference between V2I and V2V communications which also reduces the sum-rate. In spite of this, the proposed DDPG scheme still achieves the best performance among three schemes as the proposed DDPG scheme enables to adjust the resource allocation to the optimal strategy dynamically.

Figure 7 plots the average delivery probability of V2V communications while different mobility speeds are set to the vehicles in V2V communications. The velocity of vehicles in V2I communications are randomly distributed between 30 km/h and 60 km/h. Similarly, as the vehicle speed increases, the average delivery probability of V2V communications decreases. This is due to the reason that higher mobility speed in V2V communications causes the more stringent latency constraints which further reduces the delivery probability. However, from the results, it is clear that the proposed DDPG scheme still enables to obtain the good performance even when the mobility speed is set to the maximum value of 120 km/h.

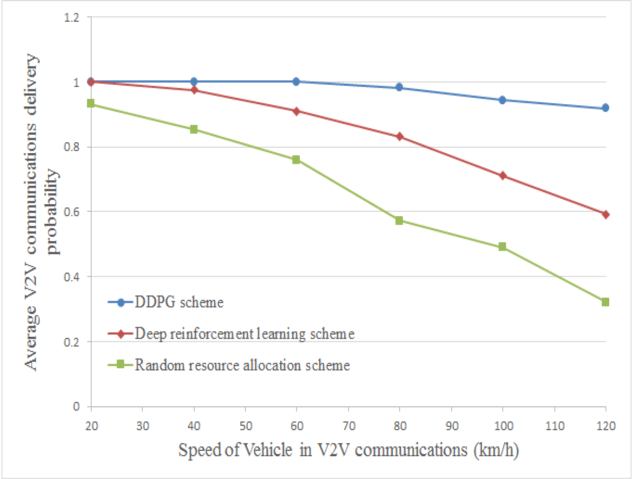


FIGURE 7. **Average delivery probability of V2V communication versus different speeds of vehicle.**

V. CONCLUSION

The main motivation of this paper is to study the resource allocation scheme for V2X communications underlaying cellular networks. Unlike the traditional D2D communication, V2X communication is characterized by stringent diverse service requirements. Specifically, the sum-rate of V2I communications and the latency and reliability of V2V communications should be guaranteed simultaneously. In this paper, we jointly consider the frequency spectrum allocation and transmission power control, formulated the resource allocation problem into a decentralized DFMDP with the goal of maximizing the sum-rate of V2I communications meanwhile guaranteeing the delivery probability of V2V communications. Due to the high complexity of the problem, we also propose a DDPG framework to solve the problem. Through extensive simulation, it is shown that the proposed scheme enables each agent adaptively learn from environment to satisfy the V2V communications constraint while maximizing the sum-rate of V2I communications. However, in this paper, we did not give an in-depth analysis of the robustness of the proposed DDPG algorithm. Future work will investigate the robustness of DDPG algorithm to obtain better understanding on when the trained actor network and critic network need to be updated and how to perform such update efficiently.

REFERENCES

[1] C. Guo, L. Liang, and G. Y. Li, “Resource allocation for low-latency vehicular communications: An effective capacity perspective,” *IEEE J. Select. Areas Commun.*, vol. 37, no. 4, pp. 905–917, Apr. 2019.
[2] J. B. Kenney, “Dedicated short-range communications (DSRC) standards in the United States,” *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, Jul. 2011.
[3] *Intelligent Transport Systems (ITS); Cooperative ITS (C-ITS); Release 1*, document ETSI TR 101 607 V1.1.1, May 2013.
[4] Y. Yang, D. Fei, and S. Dang, “Inter-vehicle cooperation channel estimation for IEEE 802.11p V2I communications,” *J. Commun. Netw.*, vol. 19, no. 3, pp. 227–238, Jun. 2017.
[5] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, “LTE for vehicular networking: A Survey,” *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 148–157, May 2013.
[6] L. Liang, H. Peng, G. Y. Li, and X. Shen, “Vehicular communications: A physical layer perspective,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10647–10659, Dec. 2017.
[7] *Study on Enhancement of 3GPP Support for 5G V2X Services*, document 3GPP TR 22.886, v16.1.1, Sep. 2018.
[8] *Enhancement of 3GPP Support for V2X Scenarios*, document 3GPP TS 22.186, v16.0.0, Sep. 2018.
[9] H. Peng, L. Liang, X. Shen, and G. Y. Li, “Vehicular communications: A network layer perspective,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1064–1078, Feb. 2019.
[10] 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.
[11] Y. Liu, L. Hao, Z. Liu, K. Sharif, Y. Wang, and S. K. Das, “Mitigating interference via power control for two-tier femtocell networks: A hierarchical game approach,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 7194–7198, Jul. 2019.

[12] D. Feng, L. Lu, Y. W. Yi, and G. Y. Li, "Device-to-device communications underlaying cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3541–3551, Aug. 2013.

[13] Y. Liu, W. Quan, T. Wang, and Y. Wang, "Delay-constrained utility maximization for video ADS push in mobile opportunistic D2D networks," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 4088–4099, Oct. 2018.

[14] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[15] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Turányi, "Design aspects of network assisted device-to-device communications," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.

[16] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Survey Tuts.*, vol. 16, no. 4, pp. 1801–1819, 4th Quart., 2014.

[17] Y. Ren, F. Liu, Z. Liu, C. Wang, and Y. Ji, "Power control in D2D-based vehicular communication networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5547–5562, Dec. 2015.

[18] 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.

[19] W. Sun, D. Yuan, E. G. Strom, and F. Brannstrom, "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.

[20] W. Sun, E. G. Strom, F. Brannstrom, 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.

[21] M. Patra, R. Thakur, and C. S. R. Murthy, "Improving delay and energy efficiency of vehicular networks using mobile femto access points," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1496–1505, Feb. 2017.

[22] 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.

[23] C.-F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1292–1295, Jun. 2018.

[24] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Survey Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[25] Y. Liu, H. Wang, M. Peng, J. Guan, J. Xu, and Y. Wang, "DeePGA: A privacy-preserving data aggregation game in crowdsensing via deep reinforcement learning," *IEEE Internet Things J.*, to be published, doi: 10.1109/jiot.2019.2957400.

[26] M. A. Salahuddin, A. Al-Fuqaha, and M. Guizani, "Reinforcement learning for resource provisioning in the vehicular cloud," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 128–135, Aug. 2016.

[27] Y. Yang, S. Dang, Y. He, and M. Guizani, "Markov decision-based pilot optimization for 5G V2X vehicular communications," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 1090–1103, Feb. 2019.

[28] Z. Li, C. Wang, and C.-J. Jiang, "User association for load balancing in vehicular networks: An online reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 2217–2228, Aug. 2017.

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

[30] 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.

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

[32] K. H. Quah and C. Quek, "MCES: A novel monte carlo evaluative selection approach for objective feature selections," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 431–448, Mar. 2007.

[33] W. Caarls and E. Schuitema, "Parallel online temporal difference learning for motor control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1457–1468, Jul. 2016.

[34] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, and Y. Tassa, "Continuous control with deep reinforcement learning," *Comput. Sci.*, vol. 8, no. 6, pp. 1–14, Jul. 2019.

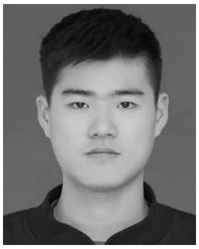
[35] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 1–9.

[36] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on LTE-based V2X Services; (Release 14)*, document 3GPP TR 36.885 V14.0.0, Jun. 2016.

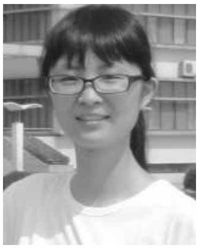
[37] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44–55, Jan. 2018.

[38] *WINNER II Channel Models*, document IST-4-027756 WINNER II D1.1.2 V1.2, Sep. 2007.

YI-HAN XU received the Ph.D. degree in telecommunications engineering from the University of Malaya, Malaysia, in 2014. He is currently a Conjoint Associate Professor with the College of Information Science and Technology, Nanjing Forestry University, China, and the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. His general research interests include statistical signal processing, the Internet of Things, and machine learning for various wireless communications.



CHENG-CHENG YANG received the bachelor's degree in electrical information engineering from the Taizhou Institute of Science and Technology, Nanjing University of Science and Technology, China, in 2019. He is currently pursuing the master's degree with the College of Information Science and Technology, Nanjing Forestry University. His research fields include the Internet of Things, V2V communications, and wireless and mobile communications.



MIN HUA received the Ph.D. degree from the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2018. Since 2018, she has been with the College of Information Science and Technology, Nanjing Forestry University, where she is currently a Lecturer. She has over ten IEEE journal publications. Her research interests are in the areas of wireless communications, signal processing, and the Internet of Things.

WEN ZHOU received the Ph.D. degree in engineering from The Hong Kong University, in 2010. He is currently with the College of Information Science and Technology, Nanjing Forestry University, where he is also an Associate Professor. His research interests include optimization designs in MIMO systems, information geometry in engineering, and the forestry Internet of Things.

• • •