JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 1341-1355 (2023)
DOI: 10.6688/JISE.202311_39(6).0006

Rate Fairness Maximization via a DRL Algorithm in Vehicular Networks

BAO GUI, SHENGHUI ZHAO+, GUILIN CHEN
AND BIN YANG
School of Computer and Information Engineering
Chuzhou University
Chuzhou, 239000 P.R. China
E-mail: gui_bao1@163.com; {zsh+; glchen}@chzu.edu.cn;
yangbinchi@gmail.com

Rate fairness is fundamental importance to ensure the quality-of-service (QoS) in vehicular networks. In this paper, we explore the rate fairness maximization (RFM) in a vehicular network including multiple vehicle-to-infrastructure (V2I) pairs and vehicle-to-vehicle(V2V) pairs. Based on this goal, we first formulate the RFM as an optimal problem with the constraints of the resources of spectrum and transmit power, and QoS requirements. To solve this challenging nonlinear and nonconvex optimization problem, we model the spectrum sharing and the transmit powers for V2V and V2I users as a Markov decision process. Then, we propose a deep reinforcement learning (DRL) algorithm to maximize the rate fairness while meeting the QoS requirements by optimally allocating spectrum and transmit power resources. Finally, we conduct simulation study to illustrate the impact of some key parameters on the rate fairness performance.

Keywords: vehicular networks, resource allocation, rate fairness maximization, deep reinforcement learning, quality-of-service

1. INTRODUCTION

Vehicular networks, which are cutting-edge technologies to enable wireless connections among vehicles and road infrastructure, are playing a vital role in the Industry 4.0 [1–3]. In such promising networks, each vehicle can transmit message to infrastructure like base station, *i.e.*, vehicle-to-infrastructure (V2I) communication. Each vehicle can also directly communicate with its nearby vehicle, *i.e.*, vehicle-to-vehicle (V2V) communication [4,5]. To meet high QoS requirements of such networks, it is critical to investigate the fundamental performances (*e.g.*, rate fairness, sum rate).

Some initial works have conducted the studies of sum rate in vehicular networks [6–10]. To improve sum rate performance, the work in [6] uses interference-alignment technique to optimize the spectrum efficiency. By a jointly optimization of spectrum and transmit power resources, the work in [7] explores the maximization of the sum rate for V2V links subject to the constraint of the minimum rate of V2I links. The objective of the work in [8] is to achieve the maximization of the sum rate for V2I links subject to the

Received September 30, 2022; revised October 31, 2022; accepted December 13, 2022.
Communicated by Xiaohong Jiang.
+Corresponding author.

constraints of the latency and reliability requirements of V2I and V2V links. A cluster-based spectrum and power resource management scheme is further proposed in [9] to improve the latency and sum rate performances of V2I links. Recently, the work employs a reconfigurable intelligent surface technique to enhance the weighted sum rate of V2I links [10].

Note that all above works need to obtain instantaneous global network information. However, as the number of vehicles increases, the acquirement of instantaneous channel state information needs huge signaling overhead. Additionally, resource allocation problems are usually formulated as nonlinear and nonconvex optimization problems which are difficult to optimize efficiently using traditional optimization methods. Moreover, these works mainly focus on the study of sum rate maximization, which is to maximize the sum rate of V2V/V2I links aiming at achieving the improvement of the overall system rate performance. But sum rate maximization cannot ensure a good rate performance for each user in the system. Since there may be unfair resource allocation, the rate of some V2V/V2I links may be very low, which leads to unsuccessful data transmission over these links. This is an unfairness to other users in the system. On the other hand, the rate fairness maximization (RFM) is to ensure the rate fairness of each link and thus can avoid the extremely low rate. It can be widely used in the vehicular networks, where each user needs to successful message transmission without occurring outage [11, 12].

To address these issues, based on a DRL algorithm, this paper explores the rate fairness maximization (RFM) problem in vehicular networks by jointly optimizing the resource allocation of spectrum and transmit power. It is notable that the resource allocation is of important issue to enhance system performance, which has received extensive attentions in wireless networks [13–16]. The main contributions of this paper can be summarized as follows.

- We concern on a vehicular network, where there are a base station (BS), multiple V2I and V2V vehicles. In such a network, we explore the RFM of V2I links. This can be modeled as an optimization problem subject to the constraints of the spectrum and transmit power resources, and the basic communication requirements of V2I links and V2V links.

- We formulate the resource allocation of spectrum and transmit power as a Markov decision process. A DRL algorithm is then proposed to maximize the rate fairness satisfying the constraints of the QoS requirements by jointly optimizing the resource allocation of spectrum and transmit power.

- Finally, extensive simulation results are presented to illustrate the impact of some key parameters on the rate fairness/sum rate and also to illustrate our findings.

The rest of the paper is organized as follows. Section 2 presents the system model of this paper. Section 3 gives the problem formulation. Section 4 presents a deep reinforcement learning algorithm. Simulation results are given in Section 5. Finally, Section 6 concludes the paper.

2. SYSTEM MODEL

2.1 Network Model

As shown in Fig. 1, we focus on an uplink transmission vehicular network composed of a BS, multiple V2I and V2V pairs, where the BS is located at a crossroad and the vehicles travel on a straight highway. In the network, we consider two types of vehicles. One type of vehicle needs to communicate with other vehicles, and another needs to communicate with the BS. Based on the communication requirements, there are M V2I vehicles and K V2V pairs in such network. We use $M' \triangleq \{1,2,\cdots M\}$ and $K' \triangleq \{1,2,\cdots K\}$ denote the sets of V2I and V2V links, respectively. We consider the available network spectrum has a total bandwidth of B MHz. It is divided into M orthogonal spectrum resource blocks of equal size, and thus the these transmission links using different resource blocks do not interfere with each other. To improve the utilization of spectrum resources, V2V links can reuse the uplink spectrum of V2Is. We further consider that each vehicle is equipped with an antenna for receiving or transmitting information.

In our study, one spectrum resource block is assigned to only one V2I link, and thus there is no interference among different V2I links. Each V2V link can reuse the spectrum resource block of one V2I link, and multiple V2V links can also use the same spectrum resource block. Therefore, there exists interference among different V2V links reusing the same resource block. Here, each V2V link can only reuse the resource block of at most one V2I link.

2.2 Channel Model

We consider a time-slotted system and in time slot t, the channel gain of V2I link from the mth V2I vehicle to the BS is expressed as

$$h_{m,B}(t) = g_{m,B}\alpha_{m,B}(t). \tag{1}$$

Here, $\alpha_{m,B}(t)$ denotes the small-scale fading component, which is an exponentially distributed random variable with zero mean. $g_{m,B}$ is the large-scale fading component of the mth V2I link including path loss and shadow fading, which is given by

$$g_{m,B} = G_m\beta_{m,B}{R_{m,B}}^{-\varphi_m}, \tag{2}$$

where G_m is the path loss constant and $\beta_{m,B}$ denotes large-scale fading coefficient. $R_{m,B}$ is the Euclidean distance from the mth V2I vehicle to BS, *i.e.*, $R_{m,B} = \sqrt{(x_{\mathrm{m}} - x_{\mathrm{B}})^2 + (y_{\mathrm{m}} - y_{\mathrm{B}})^2}$. φ_m is the power attenuation constant.

2.3 Link Rate

The signal to interference plus noise ratio (SINR) at the mth V2I receiver can be expressed as

$$\gamma_m^c(t) = \frac{P_m^c h_{m,B}(t)}{\sum_{k\in K'} \rho_k[m]P_k^v h_{k,m}(t) + \sigma^2}, \tag{3}$$

where P_m^c and P_k^v represent the transmission power of the mth V2I vehicle and the transmission power of the kth V2V vehicle, respectively. σ^2 is the variance of the additive

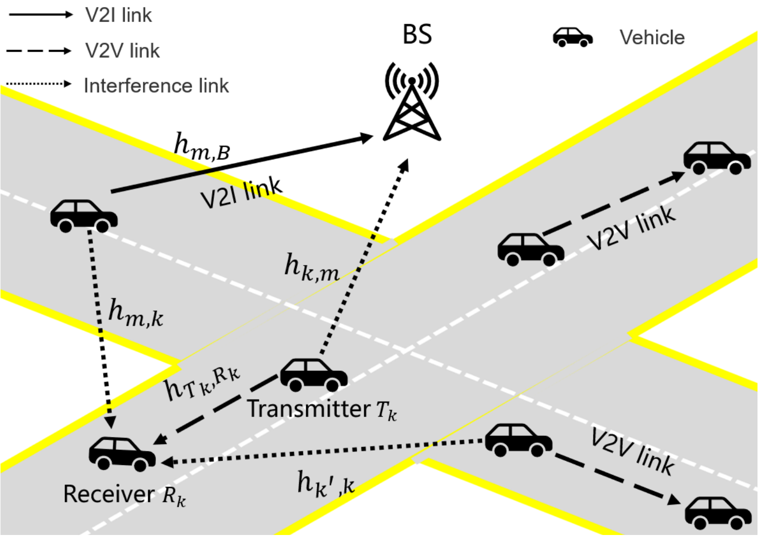

Fig. 1. An illustration of network model.

white Gaussian noise. $\rho_k[m] \in \{0,1\}$ represents an indicator of spectrum resource allocation. When $\rho_k[m]=1$, the kth V2V link reuses the same spectrum resource with the mth V2I link. Otherwise, the kth V2V link does not reuse the spectrum resource.

Therefore, the achievable rate of the mth V2I link can be written as

$$R_m^c = \frac{B}{M} \cdot \log_2\left(1+\gamma_m^c(t)\right). \tag{4}$$

Similarly, the SINR at the kth V2V receiver can be written as

$$\gamma_k^v(t) = \frac{P_k^v h_{T_k,R_k}(t)}{I_k^v(t)+\sigma^2}, \tag{5}$$

where $I_k^v(t)=\sum_{k'\neq k}\rho_{k'}[m]P_{k'}^v h_{k',k}(t)+P_m^c h_{m,k}(t)$ denotes the sum of the interference from V2I transmitters and other V2V transmitters reusing the same spectrum with the kth V2V transmistter.

Therefore, the achievable rate of the kth V2V link can be written as

$$R_k^v = \frac{B}{M} \cdot \log_2\left(1+\gamma_k^v(t)\right). \tag{6}$$

2.4 QoS Requirements

1) The rate of the mth V2I link:

To meet the QoS requirement of such network, the achievable rate of each V2I link is not less than a given threshold. It can be expressed as

$$R_m^c \geqslant R_{min}^c, \tag{7}$$

where R_{min}^c represents the given threshold and $m \in M'$.

2) The reliability requirements of the kth V2V link: We use the outage probability to measure the reliability of the V2V link transmission. In other words, the outage probability is no more than a given threshold. Here, the outage probability represents the probability that the SINR $\gamma_k^v(t)$ of the kth V2V link is no more than a given value $\gamma_{min}^v(t)$. Thus, we have

$$\mathbb{P}\{\gamma_k^v(t) \leq \gamma_{min}^v(t)\} \leq p_0, \tag{8}$$

where $\mathbb{P}\{\cdot\}$ is the outage probability and p_0 is the threshold. Regarding the Rayleigh fading [17], Eq. (8) can be simplified as

$$\gamma_k^v(t) \geqslant \gamma_{eff}(t) = \frac{\gamma_{min}^v(t)}{\ln \frac{1}{1-p_0}}, \tag{9}$$

where $\gamma_{eff}(t)$ represents the effective outage threshold, $k \in K'$.

3) The transmission latency requirement of the kth V2V link: Then, the transmission latency of the kth V2V link is given by

$$\frac{S_k}{R_k^v(t)} \leq t_{\max}, \tag{10}$$

where S_k represents the size of the message, $t_{\max}$ represents the maximum transmission latency under the condition of ensuring V2V safe communication and $k \in K'$.

3. PROBLEM FORMULATION

Our objective is to achieve the RFM of V2I links by jointly optimizing the resource allocation of spectrum and transmit power, and satisfy the QoS requirements (*i.e.*, V2I rate requirements, V2V latency and reliability) . We use $f_c\left(R_m^c\right)$ to denote the rate fairness metric of V2I links, and then

$$\begin{aligned}
\mathcal{P}_1 : \ & \max_{\rho_k[m], P_k^v, P_m^c} \quad f_c\left(R_m^c\right) \\
\text{s.t. } & \text{C1} - \text{C3} : (7), (9), (10) \\
& \text{C4} : \sum_{m \in M'} \rho_k[m] \leqslant 1, \rho_k[m] \in \{0,1\}, \forall k \in K' \\
& \text{C5} : 0 < P_k^v \leq P_{\max}, \forall k \in K' \\
& \text{C6} : 0 < P_m^c \leq P_{\max}, \forall m \in M'
\end{aligned} \tag{11}$$

where P_{max} represent the maximum transmit powers of V2V and V2I transmitters. The constraints from C1 to C3 represent the minimum rate, reliability and latency requirements, respectively. Constraint C4 indicates that each V2V link can share spectrum resource with only one V2I link. Constraints C5 and C6 represent the ranges of transmit powers of V2V and V2I transmitters, respectively. The objective function $f_c\left(R_m^c\right)$ in Eq. (11) is defined in the following equation,

$$f_c\left(R_m^c\right) = \frac{\left(\sum_{m\in M'} R_m^c\right)^2}{M\sum_{m\in M'}\left(R_m^c\right)^2}. \tag{12}$$

where $f_c\left(R_m^c\right)$ is used to measure the data rate fairness of different links in the network. We can see from Eq. (12) that $f_c\left(R_m^c\right)$ tends to one as the data rate of each link goes to be equal. This means that the data rate fairness can be guaranteed through the maximization of Eq. (12).

This is a non-linear and non-convex optimization problem with non-linear objective function and non-convex constrained conditions. Thus, it is challenging to solve this problem. We will propose a DRL algorithm to tackle with the challenging problem in the following section.

4. DEEP REINFORCEMENT LEARNING ALGORITHM

In this section, we first transform the formulated optimization problem $\mathcal{P}_1$ into a MDP and then propose a DRL algorithm to solve the optimization problem.

4.1 DRL Framework

As illustrated in Fig. 2, the DRL framework consists of agents and environment. Each vehicle, which serves as an agent, interacts with the environment, and then takes an action according to a policy π. The interaction between the agent and environment is modeled as an MDP. The agents interact with the environment and continuously learn knowledge to adapt to the environment based on the reward or penalty. At each time slot t, the agent observes a state s_t from state space, and correspondingly performs an action a_t (*i.e.*, selecting spectrum and power resources) from action space based on the policy π. Then, the current state of the environment transits to a new state s_{t+1} and the agent obtains a reward $r(t)$.

In the network, the sate space, action space and reward function can be described as follows.

1) State Space $\mathcal{S}$: The state observed by the agent for depicting the environment includes eight elements: the channel gain $h_{k,m}(t)$ from the V2V transmitter to the BS, the channel gain $h_{m,k}(t)$ from the mth V2I transmitter to the V2V receiver, the channel gain $h_{T_k,R_k}(t)$ of V2V link, the channel gain $h_{m,B}(t)$ from the mth V2I transmitter to the BS, the interference power $I_k^v(t-1)$ of V2V link in the previous time slot, the remaining time $T(t)$, the load $L(t)$, and the spectrum resource $F(t-1)$ occupied by the mth V2I pair. Thus, the state space $\mathcal{S}$ can be expressed as

$$\begin{aligned}\mathcal{S} = \{&h_{k,m}(t), h_{m,k}(t), h_{T_k,R_k}(t), h_{m,B}(t), I_k^v(t-1), L(t), T(t), F(t-1)\\ &\mid \forall m \in M', k \in K'\}.\end{aligned} \tag{13}$$

2) Action Space $\mathcal{A}$: Regarding the current state s_t, each agent performs an action a_t based on the policy π. Here, the action a_t corresponds to the selection of spectrum and power resources. The action space $\mathcal{A}$ can be written as

$$\mathcal{A} = \{\rho_k[m] \in \{0,1\}, \{P_k^v = \frac{nP_{max}}{N_p - 1} \mid n \in \{0,1,2,\cdots,N_p - 1\}\} \mid \forall k \in K', m \in M'\}, \tag{14}$$

where N_p is the number of the transmit power levels, so the size of the action space can be denoted as $M \times N_p$.

3) Reward Function r_t: The reward function drives the learning process in the DRL, and each agent attempts to maximize its reward with the interactions of the environment. Here, the reward function is to achieve the maximum rate fairness and also to guarantee the QoS requirements of V2I and V2V links. Then, r_t is given by

$$\begin{aligned} r_t = {} & \lambda_1 f_c\left(R_m^c\right) + \lambda_2 \sum_{m \in M'} H\left(R_m^c - R_{min}^c\right) + \lambda_3 \sum_{k \in K'} H\left(\gamma_k^v(t) - \gamma_{eff}(t)\right) \\ & + \lambda_4 \sum_{m \in M'} H\left(t_{\max} - \frac{S_k}{R_k^v(t)}\right). \end{aligned} \tag{15}$$

Here, λ_1, λ_2, λ_3 and λ_4 are weight coefficient, used to measure the importance of each component in the reward function. The first part of the reward function represents the fairness metric, while the second, third and fourth parts represent V2I rate, V2V reliability and latency, respectively. And the piecewise function $H(x)$ is expressed as

$$H(x) = \begin{cases} A & x \geqslant 0 \\ x & x < 0 \end{cases}, \tag{16}$$

where A is a positive constant, $H(x)$ denotes a reward or a penalty, depending on whether the communication links meet the QoS requirements.

4.2 DRL Algorithm

In DRL, the goal of each agent is to find an optimal policy π^* to maximize the long-term expected cumulative reward denoted by $Q(s_t, a_t)$. Then, it is defined as

$$Q^{\pi^*}(s_t, a_t) = \max \mathrm{E}\Big[\sum_{t'=t}^{T} \beta^{t'-t} r_{t'} \mid (s_t, a_t)\Big], \tag{17}$$

where $Q^{\pi^*}(s_t, a_t)$ represents the maximum value of $Q(s_t, a_t)$. T is the total time step, and $\beta \in (0,1)$ represents the discount factor. At time slot t', $r_{t'}$ is the corresponding reward, and the policy π is a function of state.

According to the Bellman equation [18], we obtain the following recursive relationship as

$$Q_{new}(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \beta \underset{s_t \in \mathcal{S}}{\operatorname{argmax}} Q(s_t, a_t) - Q(s_t, a_t)], \tag{18}$$

where $Q_{new}(s_t, a_t)$ is the updated value, α denotes the learning rate [19].

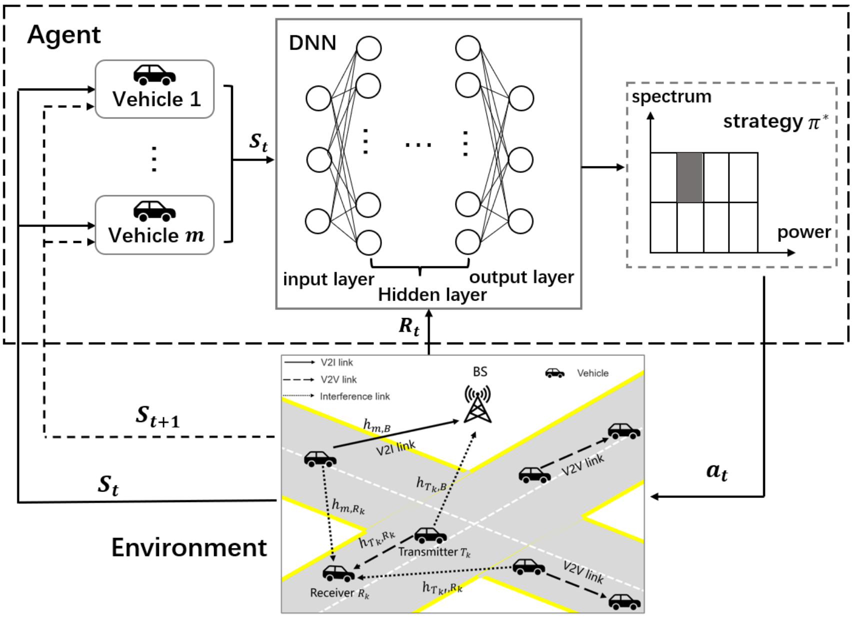

Fig. 2. DRL framework.

Algorithm 1: DRL Algorithm.

```
Initialize the Q-network for all agents.
for each episode do
    Update vehicle locations and channel large-scale fading.
    Reset the value of $L(t)$, $T(t)$ and $F(t-1)$.
    for each step $t$ do
        for each V2V agent $k$ do
            Observe $s_t \in \mathcal{S}$.
            Choose action $a_t \in \mathcal{A}$ according to policy $\pi$.
        end for
        All agents take actions and receive reward $r_t$.
        Update channel small-scale fading.
        for each V2V agent $k$ do
            Observe $s_{t+1}$ from environment.
            Store $(s_t, a_t, r_t, s_{t+1})$ in the reply memory $\mathcal{D}$.
        end for
    end for
    for each V2V agent $k$ do
        Sample mini-batches from $\mathcal{D}$.
        Update the loss function defined in (19).
    end for
end for
```

We consider that $Q(s,a)$ is approximated by the deep neural networks (DNN) $Q(s,a;\theta)$ with weight θ [20]. Here, DNN is also called Q-network. We use experience replay buffer $\mathcal{D}$ to solve the instability of function approximation. We get the latest sample space by removing the sample data without being used for the longest time such that the experience buffer $\mathcal{D}$ is always up-to-date. In each iteration, the Q-network randomly selects a min-batch of experience samples (s_t, a_t, r_t, s_{t+1}). Then we can optimize the approximation by minimizing the following loss:

$$L(\theta) = E[y_t - Q(s_t, a_t \mid \theta))^2], \tag{19}$$

where

$$y_t = r_t(s_t, a_t) + \beta \max_{a_{t+1}} (s_{t+1}, a_{t+1} \mid \theta'). \tag{20}$$

The target network is added to further improve the stability of the algorithm. The estimated values may be runaway if a set of highly dynamic values is used to update the parameters of the target network during training. It can cause instability to the algorithm. To solve the problem, we utilize the target network to update the parameters frequently and slowly. thus reducing the association between the target value and the estimated value, to steady the algorithm. As a result, the relevance between the target value and the estimated value is reduced and the stability of the algorithm is achieved. We can know the process of updating the network parameters, which can be expressed as

$$\theta' = \theta + \tau\theta + (1-\tau)\theta', \tag{21}$$

where $\tau \in (0,1)$ is used to slowly update the target network.

The detailed DRL algorithm for spectrum and power allocation is provided in Algorithm 1.

5. SIMULATION RESULTS

In this section, we present a simulation study to illustrate the impact of some key parameters on the sum rate and rate fairness under our DRL algorithm. In addition, we compare the performance with two other algorithms, the random algorithm and the vehicle grouping algorithm [21]. For the random algorithm, each agent randomly selects spectrum and transmit power resources. As for the vehicle grouping algorithm, the agents in each group are assigned the same spectrum and transmit power resources.

We consider a two-lane highway scenario, where all vehicles are located at the crossroad following Poisson distribution and a BS is located at the center. The concerned DNN network consists of three fully connected hidden layers with $\{500, 250, 120\}$ neurons per layer, and the corresponding learning rate is 0.001. The simulation parameters are set according to the the 3GPP TR.36.885 White Paper. The remaining parameters are listed in Table 1.

We explore how the number of V2V links affects the sum rate of V2I links under these three algorithms. We summarize in Fig. 3 how the sum rate of V2I links varies with

Table 1. Simulation parameters.

Parameter	Value
Number of V2I users M	5
Number of V2V users K	20
Bandwidth *B*	10 MHz
Maximum transmit power P_{max}	23 dBm
Vehicle speed *v*	36 km/h
Noise power σ^2	–114 dBm

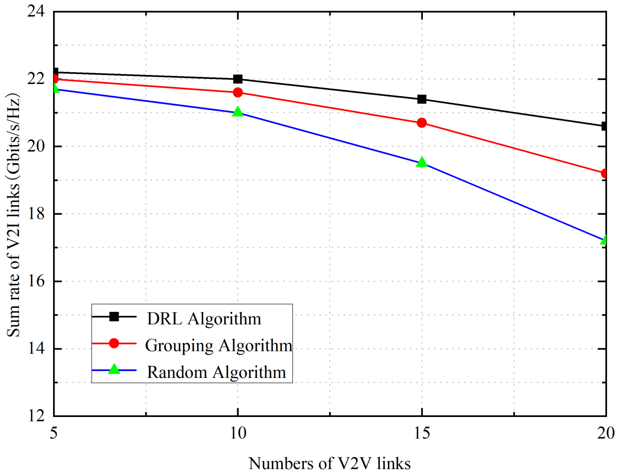

Fig. 3. Sum rate of V2I links versus the number of vehicles.

the number of V2V link for a setting of $P_{max} = 5$ dBm. We can observe from Fig. 3 that the sum rate under these three algorithms decrease as the number of V2V links increases. This can be explained as follows. As the number of V2V links increases, more V2V links reuse the same spectrum resource, which increases the interference among different V2V links. This leads to the decrease of the sum rate under each algorithm. Another observation from Fig. 3 indicates that for each fixed number of V2V links, our DRL algorithm can achieve the highest sum rate performance compared to the other two algorithms.

We explore how the number of V2I links affects the sum rate of V2I links. We summarize in Fig. 4 the relationship between the number of V2I links and the sum rate of V2I links with a setting of $P_{max} = 5$ dBm. In Fig. 4, we can see that as the number of V2I links increases, the sum rate of V2I links gradually increases. This phenomenon is due to the fact that as the number of V2I links increases, the number of resource blocks also increases in the network. Thus, V2V links have more opportunities to reuse the resource blocks that can reduce the mutual interference. Besides, the sum rate of V2I link under the DRL algorithm is also obviously greater than these under the other two algorithms.

We then investigate that how the number of V2V links affects the rate fairness of V2I links. The results are summarized in Fig. 5 with a setting of $P_{max} = 5$ dBm. We can see from Fig. 5 that the rate fairness of V2I links increases as the number of V2V links increases. This is because the increase of the interference results in the decrease

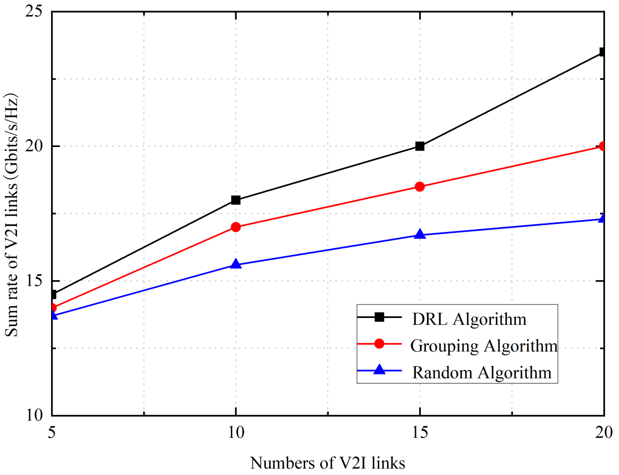

Fig. 4. Sum rate of V2I links versus the number of V2I links.

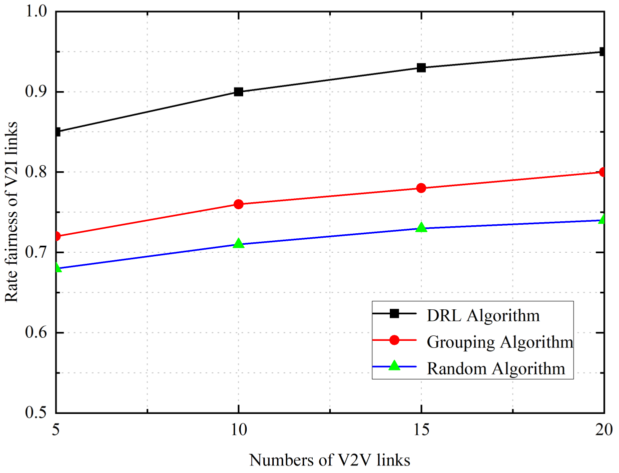

Fig. 5. Rate fairness of V2I links versus the number of vehicles.

of the difference between different links rate, and thus the rate fairness increases. We can also observe from Fig. 5 that the rate fairness under our DRL algorithm is highest in comparison with the other two algorithms.

As shown in Fig. 6, we examine how the maximum transmit power P_{max} affects the rate fairness of V2I links. In Fig. 6, we can observe that as the maximum transmission power P_{max} of V2V vehicles increases, the rate fairness of V2I links decreases. This is because a big P_{max} can more interference to the V2I links using the same spectrum resource, which leads to the decrease of the V2I link rates. Thus, this also increases the unfairness.

Finally, we conduct a convergence comparison between our DRL algorithm and vehicle grouping algorithm. We summarize in Fig. 7 how the rate fairness varies with the

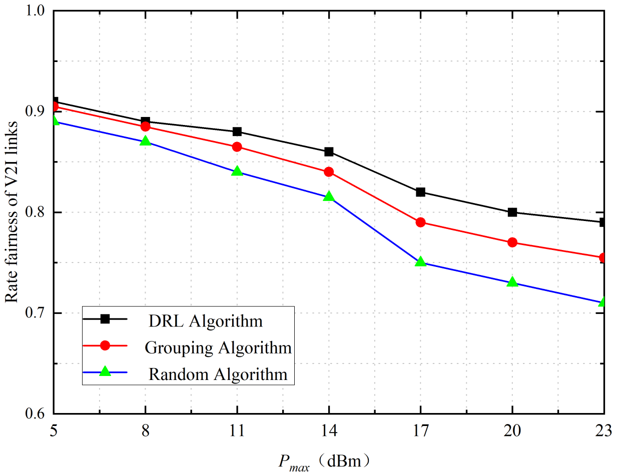

Fig. 6. Rate fairness of V2I links versus the maximum transmit power P_{max}.

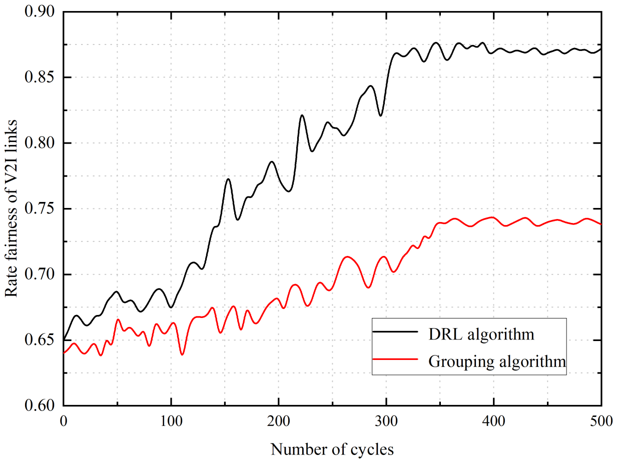

Fig. 7. Rate fairness of V2I links versus the number of cycles.

number of cycles for a setting of $P_{max} = 5$ dBm and 5 V2V links. We can see from Fig. 7 that the convergence of our DRL algorithm is faster than that of the vehicle grouping algorithm, and our DRL algorithm can achieve stable performance while the vehicle grouping algorithm is more fluctuant.

6. CONCLUSION

This paper investigated the RFM by a joint optimization of spectrum and transmission power resources. We first formulated the RFM as an optimization problem subject to the QoS requirements, spectrum and transmission power resources. We further proposed a DRL algorithm to solve this optimization problem. Finally, the simulation results showed that the rate fairness under our algorithm outperforms these under the random and vehicle grouping algorithms. Besides, our algorithm could also achieve a better convergence performance than vehicle grouping algorithm.

ACKNOWLEDGMENT

This work was supported by the Research and Innovation Team of Anhui Provincial Smart Elderly Care and Health Project under Grant 2022AH010067; by the Anhui Provincial Major Scientific and Technological Special Project under Grant 201903A06020026; by the National Natural Science Foundation of China under Grant 61962033; and by the Natural Science Project of Anhui/ChuZhou University under Grant KJ2021ZD0128, and Grant 2022XJZD12.

REFERENCES

1. Y. Zhou, F. Tang, Y. Kawamoto, and N. Kato, "Reinforcement learning-based radio resource control in 5G vehicular network," *IEEE Wireless Communications Letters*, Vol. 9, 2019, pp. 611-614.
2. T. Wu, P. Zhou, K. Liu, Y. Yuan, X. Wang, H. Huang, and D. O. Wu, "Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks," *IEEE Transactions on Vehicular Technology*, Vol. 69, 2020, pp. 8243-8256.
3. M. Noor-A-Rahim, Z. Liu, H. Lee, G. M. N. Ali, D. Pesch, and P. Xiao, "A survey on resource allocation in vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, 2020, pp. 701-721.
4. H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Transactions on Vehicular Technology*, Vol. 68, 2019, pp. 3163-3173.
5. Y. Hou, X. Wu, X. Tang, X. Qin, and M. Zhou, "Radio resource allocation and power control scheme in V2V communications network," *IEEE Access*, Vol. 9, 2021, pp. 34529-34540.
6. H. E. Elkotby, K. M. F. Elsayed, and M. H. Ismail, "Exploiting interference alignment for sum rate enhancement in D2D-enabled cellular networks," in *Proceedings of IEEE Wireless Communications and Networking Conference*, 2012, pp. 1624-1629.
7. C. Kai, H. Li, L. Xu, Y. Li, and T. Jiang, "Joint subcarrier assignment with power allocation for sum rate maximization of D2D communications in wireless cellular networks," *IEEE Transactions on Vehicular Technology*, Vol. 68, 2019, pp. 4748-4759.
8. W. Sun, E. G. Ström, F. Brännström, Y. Sui, and K. C. Sou, "D2D-based V2V communications with latency and reliability constraints," in *Proceedings of IEEE Globecom Workshops*, 2014, pp. 1414-1419.
9. F. Abbas, G. Liu, Z. Khan, K. Zheng, and P. Fan, "Clustering based resource management scheme for latency and sum rate optimization in V2X networks," in *Proceedings of IEEE 89th Vehicular Technology Conference*, 2019, pp. 1-6.
10. D. L. Dampahalage, K. B. S. Manosha, N. Rajatheva, and M. Latva-Aho, "Weighted-sum-rate maximization for an reconfigurable intelligent surface aided vehicular network," *IEEE Open Journal of the Communications Society*, Vol. 2, 2021, pp. 687-703.

11. W. Ahsan, W. Yi, Z. Qin, Y. Liu, and A. Nallanathan, "Resource allocation in uplink NOMA-IoT networks: A reinforcement-learning approach," *IEEE Transactions on Wireless Communications*, Vol. 20, 2021, pp. 5083-5098.
12. I. Budhiraja, N. Kumar, and S. Tyagi, "Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay D2D communication," *IEEE Internet of Things Journal*, Vol. 8, 2020, pp. 3143-3156.
13. S. Zhang, H. Gu, K. Chi, L. Huang, K. Yu, and S. Mumtaz, "DRL-based partial offloading for maximizing sum computation rate of wireless powered mobile edge computing network," *IEEE Transactions on Wireless Communications*, Vol. 21, 2022, pp. 10934-10948.
14. B. Zhu, K. Chi, J. Liu, K. Yu, and S. Mumtaz, "Efficient offloading for minimizing task computation delay of NOMA-based multiaccess edge computing," *IEEE Transactions on Communications*, Vol. 70, 2022, pp. 3186-3203.
15. L. Huang, R. Nan, K. Chi, Q. Hua, K. Yu, N. Kumar, and M. Guizani, "Throughput guarantees for multi-cell wireless powered communication networks with non-orthogonal multiple access," *IEEE Transactions on Vehicular Technology*, Vol. 71, 2022, pp. 12104-12116.
16. J. Niu, S. Zhang, K. Chi, G. Shen, and W. Gao, "Deep learning for online computation offloading and resource allocation in NOMA," *Computer Networks*, Vol. 216, 2022, p. 109238.
17. L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, "Graph-based resource sharing in vehicular communication," *IEEE Transactions on Wireless Communications*, Vol. 17, 2018, pp. 4579-4592.
18. Y. Feng, L. Li, and Q. Liu, "A kernel loss for solving the bellman equation," *Advances in Neural Information Processing Systems*, Vol. 32, 2019, pp. 1-12.
19. K. Wang, Y. Dou, T. Sun, P. Qiao, and D. Wen, "An automatic learning rate decay strategy for stochastic gradient descent optimization methods in neural networks," *International Journal of Intelligent Systems*, Vol. 37, 2022, pp. 7334-7355.
20. W. Lee, O. Jo, and M. Kim, "Intelligent resource allocation in wireless communications systems," *IEEE Communications Magazine*, Vol. 58, 2020, pp. 100-105.
21. M. I. Ashraf, M. Bennis, C. Perfecto, and W. Saad, "Dynamic proximity-aware resource allocation in vehicle-to-vehicle (V2V) communications," in *Proceedings of IEEE Globecom Workshops*, 2016, pp. 1-6.

Bao Gui received the MS degree from Anhui University of Science and Technology, China, in 2022. His research interests include vehicular networks and wireless communication.

Shenghui Zhao received the MS degree from the Hefei University of Technology, China, in 2003, and the Ph.D. degree from Southeast University, China, in 2013. She is currently a Professor with the School of Computer and Information Engineering, Chuzhou University, Anhui, China. Her current research interests include trusted computing, wireless networks, healthcare, and Internet of Things.

Guilin Chen received the BS degree from Anhui Normal University, China, in 1985, and the MS degree from the Hefei University of Technology, in 2007. He is currently a Professor with the School of Computer and Information Engineering, Chuzhou University, Anhui, China. His current research interests include cloud computing, wireless networks, healthcare, and Internet of Things.

Bin Yang received his Ph.D. degree in Systems Information Science from Future University Hakodate, Japan in 2015. He is a Professor with the School of Computer and Information Engineering, Chuzhou University, China. His research interests include unmanned aerial vehicle networks, cyber security and Internet of Things.