2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring) | 979-8-3503-1114-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/VTC2023-Spring57618.2023.10200293

Deep Reinforcement Learning-Based Resource Allocation for Cellular V2X Communications

Yi-Ching Chung[1], Hsin-Yuan Chang[1], Ronald Y. Chang[2], and Wei-Ho Chung[1, 2]
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
Email: {yiching.chung, hyuan.chang}@outlook.com, rchang@citi.sinica.edu.tw, whchung@ee.nthu.edu.tw

***Abstract*—Vehicle-to-everything (V2X) communication is an essential technology for future vehicular applications. It is challenging to simultaneously achieve vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications, given the shared spectrum. Deep reinforcement learning (DRL)-based algorithms have been proposed for resource allocation in V2I and V2V designs. Existing DRL designs focus on the objectives of high-capacity V2I and high-reliability V2V links. In this study, a multi-agent DRL algorithm is proposed to maximize the sum capacity of V2I links while ensuring capacity fairness among the V2V links. The simulation results demonstrate the balance between the V2I–V2V objectives achieved by the proposed algorithm.**

***Index Terms*—Vehicle-to-everything (V2X) communication, deep reinforcement learning (DRL), resource allocation.**

I. Introduction

Vehicle-to-everything (V2X) communications, including vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications [1], [2], are promising techniques for future vehicular applications [3] providing high-transmission capacity experiences and enhancing the safety of road users. The development of autonomous vehicles (AVs) [4], [5] and cooperative localization networks [6]–[9] incorporates V2I and V2V communications. Specifically, V2I communications are exploited to provide services with high capacity, whereas V2V communications play an important role in delivering safety messages. To develop multiple applications using shared spectrum resources, addressing the interference issue while improving service performance is a critical challenge. Resource allocation approaches have been widely discussed in the field of cellular V2X communications for simultaneously improving the performance of V2I and V2V communications, while suppressing interference. In particular, deep reinforcement learning (DRL)-based algorithms [10]–[17] have been introduced owing to their ability to make adaptive decisions through frequent interactions with the environment. Moreover, DRL-based algorithms employ strategies based on the exploration of unknown environments and employment of acquired knowledge. Therefore, DRL is a promising approach that addresses the highly challenging V2X dynamic problem by learning from interactions with the environment and making sequential decisions.

In [10], a DRL algorithm was proposed that adopts the sum capacity and payload delivery rate as the performance metrics of the V2I and V2V links, respectively, to provide high-capacity and high-reliability services. In [11], the developed DRL algorithm for V2V and V2I links both sought high sum capacity and link reliability for V2V links. In [12], a weighted optimization problem of the sum capacity and probability of successful payload transmission was solved using a DRL algorithm for high-capacity V2I and reliable V2V links. In [13], the reliability of V2V links was modeled as a successful transmission probability and simultaneously maximized with the sum capacity of the V2I links. Alternative algorithms [14]–[17] have been proposed to guarantee vehicular safety by focusing on the delay requirements of the V2V links. In [14], a DRL algorithm for maximizing the sum capacity of V2I links was proposed, along with constraints on the sum capacity and outage probability of V2V links, to achieve the latency and reliability requirements. In [15], a DRL algorithm aimed at maximizing both the sum capacities of the V2I and V2V links, with constraints of latency and reliability of the V2V links, was proposed. In [16], the delay requirements of V2V links were investigated and divided into sensitive and insensitive cases, and a DRL algorithm was proposed for providing the services of high-capacity V2I and high-reliability V2V links. In [17], a federated algorithm based on DRL for maximizing the V2I throughput while maintaining the reliability and delay requirements of V2V links was developed.

In contrast to the above studies on the throughput and reliability optimization, our study places a greater emphasis on the fairness of V2V links because all vehicles are of equal importance in acquiring safety-critical information. In this study, we developed a DRL-based algorithm to allocate resources with the objective of maximizing the sum capacity of V2I links and the max–min fairness of V2V links. First, a DRL problem of resource allocation for multiple agents is formulated, and the demands of the V2I and V2V links are emphasized. The proposed algorithm provides a resource allocation strategy to guarantee fairness among V2V links, while maintaining an acceptable sum capacity performance of V2I links. The numerical results show that the proposed algorithm achieves a satisfactory tradeoff between the V2I and V2V objectives. It was also verified by extensive simulations

This work was supported by the Visible Project at the Research Center for Information Technology Innovation, Academia Sinica, and the National Science and Technology Council (formerly, Ministry of Science and Technology), Taiwan, under Grants MOST 109-2221-E-001-013-MY3, 109-2918-I-001-003, and 110-2221-E-007-042-MY3.

979-8-3503-1114-3/23/$31.00 ©2023 IEEE

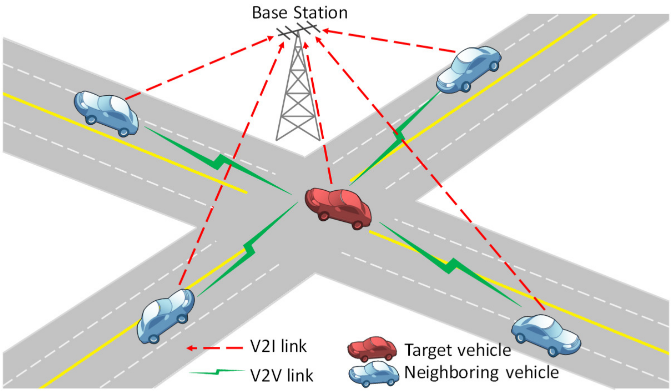


Fig. 1. An illustrative example of the considered vehicular network infrastructure. The red vehicle at the intersection represents the target vehicle, which receives information from the neighboring blue vehicles via V2V links. All vehicles perform uplink transmission towards the BS via V2I links.

that the adopted reward design facilitates the cooperation of multiple agents to achieve better performance.

The remainder of this paper is organized as follows. Sec. II describes the system model, and Sec. III introduces the proposed DRL algorithm for resource allocation. Sec. IV presents the simulation results, followed by conclusions in Sec. V.

II. System Model and Problem Statement

A. Vehicular Network Infrastructure

We consider a cellular-based V2X crossroad infrastructure which has been discussed in the 3GPP Release 15. As shown in Fig. 1, there are multiple neighboring vehicles around a target vehicle, and a base station (BS) beside the crossroad. In particular, N vehicles, M V2I links, and K V2V links are considered. We consider uplink transmissions for V2I links and all vehicles transmit data towards the BS, i.e., $M = N$. Each V2V link comprises a V2V transmitter and a V2V receiver. The $N-1$ neighboring vehicles transmit data towards the target vehicle and the target vehicle transmits data towards the nearest neighbor; therefore, $K = N$. Moreover, each vehicle is characterized by a driving direction and speed and is equipped with a single antenna. The sets of the V2I links and V2V links in the vehicular network are expressed as $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$, respectively. Furthermore, we consider M orthogonal spectrum sub-bands, and the bandwidth of each sub-band is denoted as W.

B. Communication Model

Each V2I link uses a distinct, pre-allocated sub-band in our system. Specifically, the mth V2I link uses the mth sub-band, with a fixed transmit power $P[m] = P$. The K V2V links share the same M sub-bands with the V2I links. We adopt the urban channel model [18], [19] for V2I and V2V links, as described below.

V2I Links: In one coherence time period, the channel power gain $g^{(t)}[m]$ from the mth vehicle to the BS over the mth sub-band is modeled in the dB scale by

$$g^{(t)}[m] = \alpha^{(t)} - 10\log_{10} f^{(t)}[m], \quad (1)$$

where $\alpha^{(t)} = \mathrm{PL}_{\mathrm{V2I}}(d_t) + \mathrm{S}_{\mathrm{V2I}}(\Delta d_t)$ represents the frequency-independent large-scale fading effect at time t, which is the sum of path-loss and shadowing effects in dB, and $f^{(t)}[m]$ represents the frequency-dependent small-scale fading effect, which follows an exponential distribution with a unit mean. The path-loss model of the V2I links is [18]

$$\mathrm{PL}_{\mathrm{V2I}}(d_t) = 128.1 + 37.6\log_{10}(d_t) \quad (d_t \text{ in km}), \quad (2)$$

where d_t represents the distance between the vehicle and the BS at time t. The shadowing value at time t is modeled by [18]

$$\mathrm{S}_{\mathrm{V2I},t}(\Delta d_t) = \begin{cases} s_{\mathrm{V2I}}, \text{ if } t = 0 \\ e^{\frac{-\Delta d_t}{d_{\mathrm{corr}}}} \times \mathrm{S}_{\mathrm{V2I},t-1}(\Delta d_{t-1}) + \sqrt{1 - e^{\frac{-2\Delta d_t}{d_{\mathrm{corr}}}}} \times s_{\mathrm{V2I}}, \text{ else} \end{cases} \quad (3)$$

where $s_{\mathrm{V2I}} \sim \mathcal{N}(0, 8)$ denotes a Gaussian random value, Δd_t is the change in distance of the vehicle to the BS from time $t-1$ to time t, and $d_{\mathrm{corr}} = 50$ m.

V2V Links: In one coherence time period, the channel power gain $h^{(t)}_{k,k'}[m]$ from the kth V2V transmitter to the k'th V2V receiver over the mth sub-band is modeled in the dB scale by

$$h^{(t)}_{k,k'}[m] = \alpha^{(t)}_{k,k'} - 10\log_{10} f^{(t)}_{k,k'}[m], \text{ for } k \neq k', \quad (4)$$

where $\alpha^{(t)}_{k,k'} = \mathrm{PL}_{\mathrm{V2V}}(d_t) + \mathrm{S}_{\mathrm{V2V}}(\Delta d_t)$ is the frequency-independent large-scale fading effect at time t, which is the sum of path-loss and shadowing effects in dB, and $f^{(t)}_{k,k'}[m]$ is the frequency-dependent small-scale fading effect between the kth and k'th vehicles, which follows an exponential distribution with a unit mean. The path-loss models of the V2V links for line-of-sight (LoS) and non-line-of-sight (NLoS) conditions are given respectively by [19]

$$\mathrm{PL}_{\mathrm{V2V,LoS}}(d_t) = \begin{cases} 22.7\log_{10}(3) + 41 + 20\log_{10}\frac{f_c}{5}, \text{ if } d_t \leq 3 \\ 22.7\log_{10}(d_t) + 41 + 20\log_{10}\frac{f_c}{5}, \text{ if } d_t < d_{\mathrm{BP}} \\ 40\log_{10}(d_t) + 9.45 - 34.6\log_{10}(h_{\mathrm{VE}}) + 2.7\log_{10}\frac{f_c}{5}, \text{ else} \end{cases} \quad (5)$$

where d_t is the distance between two vehicles at time t, and

$$\mathrm{PL}_{\mathrm{V2V,NLoS}}(d_{t,\mathrm{x}}, d_{t,\mathrm{y}}) = \mathrm{PL}_{\mathrm{V2V,LoS}}(d_{t,\mathrm{x}}) + 20 - 12.5n + 10n\log_{10}(d_{t,\mathrm{y}}) + 3\log_{10}(f_c/5). \quad (6)$$

Then, the overall path-loss model is given by

$$\mathrm{PL}_{\mathrm{V2V}}(d_t) =$$

$$\begin{cases} \mathrm{PL}_{\mathrm{V2V,LoS}}(d_t), & \text{if } \min(d_{t,\mathrm{x}}, d_{t,\mathrm{y}}) < 7 \\ \min(\mathrm{PL}_{\mathrm{V2V,NLoS}}(d_{t,\mathrm{x}}, d_{t,\mathrm{y}}), \mathrm{PL}_{\mathrm{V2V,NLoS}}(d_{t,\mathrm{y}}, d_{t,\mathrm{x}})), & \text{else} \end{cases} \tag{7}$$

In (5)–(7), $d_{t,\mathrm{x}}$ ($d_{t,\mathrm{y}}$) is the distance between two vehicles in the x-axis (y-axis) direction at time t, f_c is the carrier frequency, $d_{\mathrm{BP}} = 4(h_{\mathrm{VE}} - 1)^2 f_c/c$ is the breakpoint distance, h_{VE} is the antenna height of vehicles and the effective environment height is assumed to be 1 m, $c = 3 \times 10^8$ m/s is the propagation velocity in free space, and $n = \max(2.8 - 0.0024 d_{t,\mathrm{x}}, 1.84)$. The shadowing value at time t is modeled by [18]

$$\mathrm{S}_{\mathrm{V2V},t}(\Delta d_t) = \begin{cases} s_{\mathrm{V2V}}, & \text{if } t = 0 \\ e^{\frac{-\Delta d_t}{d_{\mathrm{corr}}}} \times \mathrm{S}_{\mathrm{V2V},t-1}(\Delta d_{t-1}) + \sqrt{1 - e^{\frac{-2\Delta d_t}{d_{\mathrm{corr}}}}} \times s_{\mathrm{V2V}}, & \text{else} \end{cases} \tag{8}$$

where $s_{\mathrm{V2V}} \sim \mathcal{N}(0, 3)$ denotes a Gaussian random value, Δd_t is the change in distance of two vehicles from time $t-1$ to time t, and $d_{\mathrm{corr}} = 10$ m. The directions and speeds of the vehicles are uniformly determined from $\{\frac{\pi}{2}, \pi, \frac{3\pi}{2}, 2\pi\}$ and $[10, 15]$ m/s, respectively.

For notation brevity, we drop time t hereafter. The received signal-to-interference-plus-noise ratio (SINR) of the mth V2I link over the mth sub-band can be formulated as

$$\Gamma_{\mathrm{V2I}}[m] = \frac{P g[m]}{\sigma^2 + \sum\limits_{k \neq 1} \rho_k[m] P_k[m] h_{k,1}[m] + \rho_1[m] P_1[m] h_{1,\widehat{k}}[m]}, \tag{9}$$

where the first term of the denominator denotes the received noise power, the second and third terms of the denominator denote the interference due to transmissions from the neighboring vehicle (denoted by the kth vehicle) to the target vehicle (denoted by the 1st vehicle), and the transmission from the target vehicle to the nearest neighboring vehicle (denoted by the $\widehat{k}$th vehicle), $P_k[m]$ denotes the transmit power of the kth V2V transmitter over the mth sub-band, and $\rho_k[m]$ denotes the binary spectrum allocation indicator where $\rho_k[m] = 1$ (0) when the kth V2V link employs (not employs) the mth sub-band. Likewise, the received SINR of the kth V2V link over the mth sub-band is represented as

$$\Gamma_{\mathrm{V2V},k}[m] = \begin{cases} \frac{\rho_k[m] P_k[m] h_{k,\widehat{k}}[m]}{\sigma^2 + P g[m] + \sum\limits_{k' \neq k} \rho_{k'}[m] P_{k'}[m] h_{k',k}[m]}, & \text{if } k = 1 \\ \frac{\rho_k[m] P_k[m] h_{k,1}[m]}{\sigma^2 + P g[m] + \sum\limits_{\substack{k' \neq 1 \\ k' \neq k}} \rho_{k'}[m] P_{k'}[m] h_{k',1}[m] + \rho_1[m] P_1[m] h_{1,\widehat{k}}[m]}, & \text{else} \end{cases} \tag{10}$$

Then, the capacities of the mth V2I link and the kth V2V link over the mth sub-band are given respectively by

$$C_{\mathrm{V2I}}[m] = W \log(1 + \Gamma_{\mathrm{V2I}}[m]), \text{ and} \tag{11}$$

$$C_{\mathrm{V2V},k}[m] = W \log(1 + \Gamma_{\mathrm{V2V},k}[m]). \tag{12}$$

C. Problem Statement

Our objective is to determine power allocation $P_k[m]$ and spectrum allocation $\rho_k[m]$ for V2V links so that fairness (in terms of max-min fairness) among V2V communications is achieved while satisfactory sum capacity performance for V2I links is maintained. Specifically, we consider the following problem:

$$\max_{P_k[m], \rho_k[m]} \left\{ c_1 \sum_{m \in \mathcal{M}} C_{\mathrm{V2I}}[m] + c_2 \min_{k \in \mathcal{K}} C_{\mathrm{V2V},k}[m] \right\}, \tag{13}$$

where $c_1 \in [0, 1]$ and $c_2 \in [0, 1]$ are weighting factors that determine the tradeoffs between V2I and V2V objectives and that $c_1 + c_2 = 1$. Problem (13) embodies a multi-objective optimization (MOO) problem which contains multiple potentially conflicting goals. We reformulate the MOO problem into a tractable single-objective optimization (SOO) problem by the weighted sum approach as in (13). This paper proposes a DRL-based method to solve (13) for its superior dynamic programming capabilities. The proposed method is described next.

III. Multi-Agent DRL-Based Resource Allocation Algorithm

In this section, we first define the state, action, and reward of the proposed multi-agent DRL algorithm and then introduce the learning algorithm in training and testing stages. We use deep Q-learning with experience replay to train multiple V2V agents, where agents are trained based on the action-value function under the policy π. To emphasize the dynamic evolution of DRL, we associate the related notations with an additional time index t.

A. State, Action, and Reward

1) State: Let the state $S_{k,t}$ contain the observed channel information of the kth V2V agent at the tth time slot. Since further analysis shows that decision-making policies of agents are highly related to the number of training iterations and the rate of exploring the environment, we factor them into the state, which is defined as

$$S_{k,t} = \left\{ \{ H_t^k[m], I_t^k[m] \}_{m \in \mathcal{M}}, \omega, \epsilon \right\}, \tag{14}$$

where ω is the ωth training iteration, ϵ is the probability of random action selection in the ϵ-greedy policy,

$$H_t^k[m] = \begin{cases} \left\{ h_{k,\widehat{k}}[m], h_{k',k}[m], g[m] \right\}, & \text{if } k = 1 \\ \left\{ h_{k,1}[m], h_{k',1}[m], h_{1,\widehat{k}}[m], g[m] \right\}, & \text{else} \end{cases} \tag{15}$$

indicates the observed channel by the kth agent, and

$$I_t^k[m] = \begin{cases} P g[m] + \sum\limits_{k' \neq k} \rho_{k'}[m] P_{k'}[m] h_{k',k}[m], & \text{if } k = 1 \\ P g[m] + \sum\limits_{\substack{k' \neq 1 \\ k' \neq k}} \rho_{k'}[m] P_{k'}[m] h_{k',1}[m] + \rho_1[m] P_1[m] h_{1,\widehat{k}}[m], & \text{else} \end{cases} \tag{16}$$

indicates the received interference power in (10).

2) Action: At the tth time slot, the kth agent's action $A_{k,t}$ consists of the aforementioned spectrum allocation indicator $\rho_{k,t}[m] \in \{0,1\}$ and the transmission power $P_{k,t}[m] \in \{-100, 5, 10, 23\}$. Note that the selection of transmission power is limited to a finite set, which contains four discrete power levels, to facilitate the learning procedure and reflect actual circuit limitations [10]. Overall, the kth agent's action $A_{k,t}$ is defined as $A_{k,t} = \{P_{k,t}, \rho_{k,t}\}$, where the sets $P_{k,t} = \{P_{k,t}[m] : m \in \mathcal{M}\}$ and $\rho_{k,t} = \{\rho_{k,t}[m] : m \in \mathcal{M}\}$ contain full results of transmission power and sub-band selections, respectively.

3) Reward: The reward is designed according to the problem objective in (13) as

$$R_{t+1} = c_1 \sum_{m \in \mathcal{M}} C_{\mathrm{V2I}}\left[m, t\right] + c_2 \left(\min_{k \in \mathcal{K}} C_{\mathrm{V2V},k}\left[m, t\right] - \delta \right), \tag{17}$$

where $C_{\mathrm{V2I}}[m,t]$ and $C_{\mathrm{V2V},k}[m,t]$ are defined in (11) and (12), respectively. The hyperparameter δ is introduced to alter the exploration ability of the algorithm. Specifically, δ is made adaptive in the sense that its value is set to 0 initially and then adjusted to the minimal capacity among all V2V links at half of the episodes to facilitate escaping from local minima [20]. Setting δ to the minimal capacity among all V2V links renders the second term of (17) zero, thereby motivating agents to improve V2V functionalities. Adjusting δ in the middle of the training procedure allows sufficient time to learn the policy before and after the adjustment during the training. The expected return, i.e., the cumulative discounted future reward, is given by

$$U_t = \sum_{i=0}^{\infty} \gamma^i R_{t+i+1}, \tag{18}$$

where $\gamma \in [0,1]$ is the discount factor.

B. Learning Algorithm

The proposed multi-agent DRL algorithm can be divided into training and testing phases. In the training phase, we adopt centralized learning, implying that the reward is known to all agents and therefore agents are capable of adjusting their actions toward the optimal policy by updating the parameters of their own deep Q-networks (DQNs). In the testing phase, distributed learning is implemented. Each agent receives observations within the local scope and then selects appropriate actions to interact with the environment globally according to its own trained DQN.

1) Training Phase: The action-value function of the kth agent is defined as the expected return of making an action in a certain state by following a specific strategy, which can be written as

$$Q_k(S_{k,t}, A_{k,t}) = \mathbb{E}\left[U_t | S_{k,t}, A_{k,t}\right]. \tag{19}$$

Algorithm 1 Multi-Agent DRL-Based Resource Allocation Algorithm for V2X Networks

```
Start environment simulator, producing vehicles and links
Initialize Q-networks for all V2V agents
for each episode do
    Renew vehicle locations and large-scale fading $\alpha$, $\alpha_{k,k'}$
    for each step $t$ do
        for each V2V agent $k$ do
            Observe $S_{k,t}$
            Select action $A_{k,t}$ from $S_{k,t}$ according to $\epsilon$-greedy policy
        end for
        All agents make actions and receive reward $R_{t+1}$
        Renew channel small-scale fading $f[m]$, $f_{k,k'}[m]$
        for each V2V agent $k$ do
            Observe $S_{k,t+1}$
            Store $(S_{k,t}, A_{k,t}, R_{t+1}, S_{k,t+1})$ in the replay memory $\mathcal{D}_k$
        end for
    end for
    for each V2V agent $k$ do
        Uniformly fetch mini-batches from $\mathcal{D}_k$
        Use variant of stochastic gradient descent to optimize the error between Q-network and learning objectives
    end for
end for
```

The data in the form of time sequence may introduce improper training for the model. Thus, a replay memory is used to store the data. We record the complete loop as $(S_{k,t}, A_{k,t}, R_{t+1}, S_{k,t+1})$ and store it in the replay memory $\mathcal{D}_k$. In order to update the trainable parameter θ_k of DQN for the kth agent, a small batch of the experience $\mathcal{D}$ is uniformly fetched from the replay memory in each episode to decrease the temporal correlation in continuous updates and minimize the sum-squared error function:

$$\sum_{\mathcal{D}} \left[R_{t+1} + \gamma \max_{a'} Q'_k(S_{k,t+1}, a'; \theta'_k) - Q_k(S_{k,t}, A_{k,t}; \theta_k) \right]^2, \tag{20}$$

where θ'_k is the parameter set of a target Q-network, which is periodically copied from the training Q-network for updates. With the experience replay design, the correlation of successive updates is effectively broken by repeatedly extracting samples from experience, thus stabilizing learning.

2) Testing Phase: During the execution phase, each V2V agent selects the action of the largest action value at each time slot based on its own observed local channels and its trained Q-network. All V2V agents make corresponding sub-band selection and transmission power level selection according to the chosen actions.

The complete DRL procedure is summarized in Algorithm 1.

IV. Simulation Results

We simulate the V2X network in an urban scenario as described in Sec. II. Similar to [10], we set the default simulation parameters as listed in Table I. The DQN of each V2V agent is composed of three fully connected hidden layers

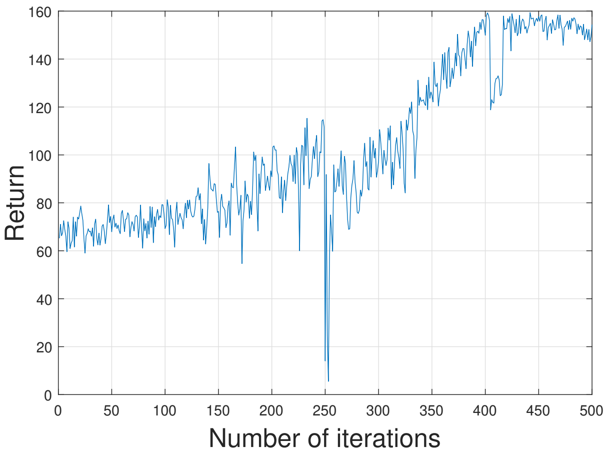


Fig. 2. The expected return vs. the number of iterations for each training episode.

TABLE I
SIMULATION PARAMETERS

Parameter	Symbol	Value
Number of vehicles	N	5
Number of V2I links	M	5
Number of V2V links	K	5
Bandwidth	W	1 MHz
Carrier frequency	f_c	2 GHz
Vehicle antenna height	h_{VE}	1.5 m
V2I transmit power	P	23 dBm
V2V transmit power	P_k	$\{-100, 5, 10, 23\}$ dBm
Noise power	σ^2	-114 dBm
Weights in reward function	c_1, c_2	0.2, 0.8
Discount factor	γ	1
Large-scale fading update	-	Every 100 ms
Small-scale fading update	-	Every 1 ms

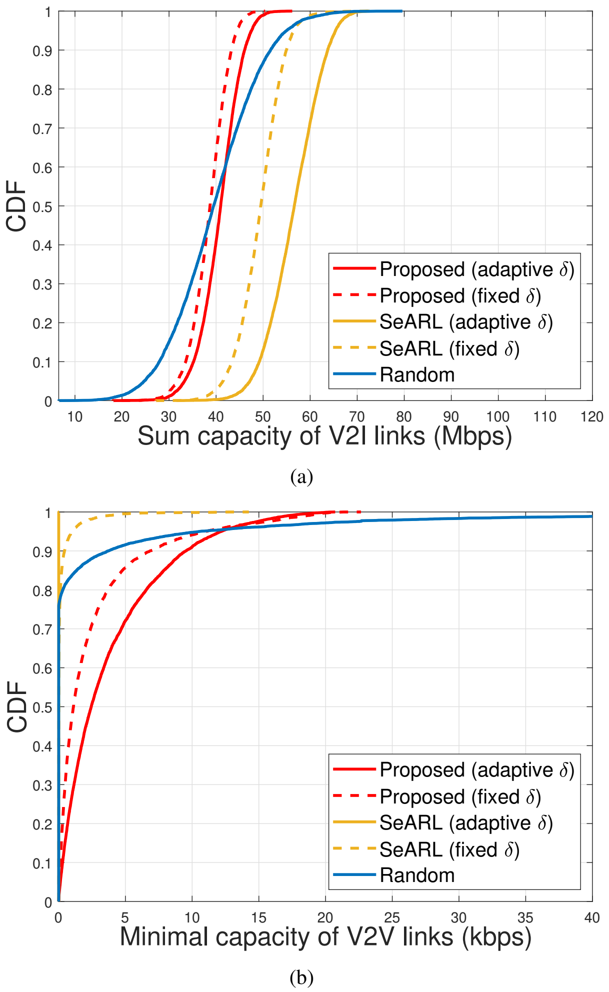
of 500, 250, and 120 neurons, respectively, with rectified linear unit (ReLU) activation. The RMSProp optimizer is used to update the parameters of the neural network with a learning rate of 0.001. The Q-network of each V2V agent is trained for 500 episodes, where the exploration rate ϵ decays linearly from 1 to 0.02 in the first 400 episodes, and remains unchanged thereafter.

The proposed multi-agent DRL method is compared with the following benchmarks:

- *SeARL [11]:* Multiple V2V agents decide actions in sequence with the reward function in (17), using a common DQN with three hidden layers of 500, 250, and 120 neurons, respectively, and the ReLU activation; and
- *Random:* Multiple V2V agents take random actions at each time slot, using distinct DQNs.

Furthermore, the SeARL and proposed schemes using the reward function in (17) (denoted by "(adaptive δ)") are compared with their respective counterparts trained using the same reward function but with fixed parameter $\delta = 0$ (denoted by "(fixed δ)") to study the reward design.

Fig. 2 depicts the expected return of the proposed multi-agent DRL algorithm during the learning process to examine convergence. As can be seen, a general trend is that the return increases continuously until it reaches saturation after training, showing that the proposed algorithm converges to optimal weightings. The fluctuations during the training process are due to the dynamic channels in the vehicular environment. The noticeable drop at training episode $= 250$ can be attributed to channel fading and to the increased value of parameter δ in the reward function. As previously mentioned, parameter δ, initially set to zero, is adjusted to the minimal capacity among V2V links at half of the episodes to help V2V agents escape from the local minima.

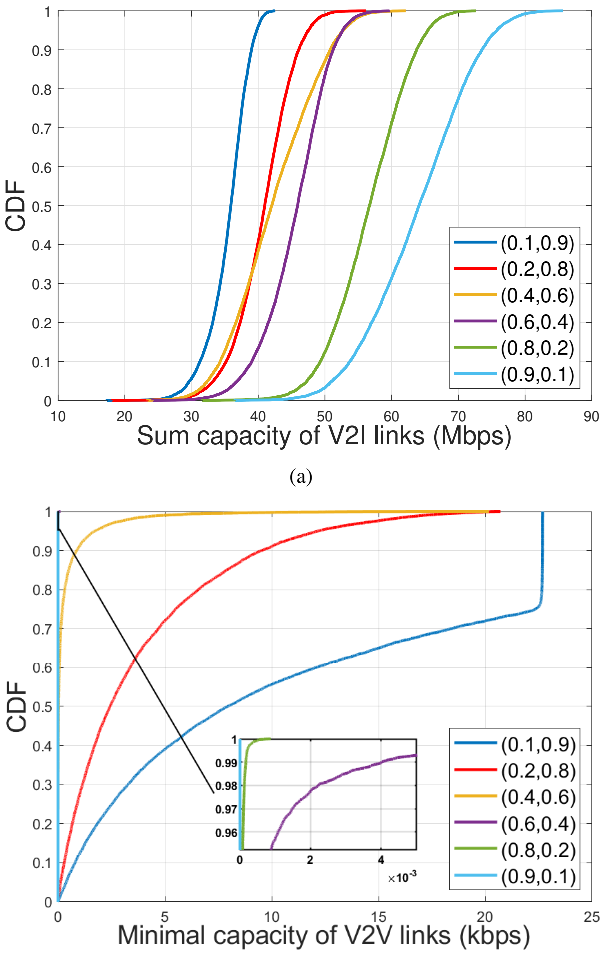


(a)

(b)

Fig. 3. CDF of (a) the sum capacity of the V2I links and (b) the minimal capacity of the V2V links.

Fig. 3 shows the performance of V2I and V2V links. First, comparing the proposed scheme and SeARL, it is seen that the proposed scheme achieves superior V2V fairness but inferior V2I performance. This can be explained by examining the learning mechanism. SeARL employs a single DQN shared by all V2V agents, and therefore it is difficult for SeARL to learn a strategy that increases the capacity of the present minimal-capacity V2V agent while in the meantime decreases the capacity of other V2V agents by redistributing resources



(a)

(b)

Fig. 4. CDF of (a) the sum capacity of the V2I links and (b) the minimal capacity of the V2V links, for the proposed algorithm with various weights (c_1, c_2) in the reward function.

to achieve max-min fairness of V2V performance. However, the same mechanism (a single DQN shared by all V2V agents) makes it relatively easy to train a strategy to achieve the objective of sum V2I capacity. In fact, as the numerical data suggests, SeARL learns the best strategy for achieving the highest reward in (17) by focusing on the V2I objective, which results in the better V2I performance but compromised V2V fairness performance. In contrast, V2V agents in the proposed scheme use multiple DQNs of distinct weights to achieve both V2V and V2I objectives. Second, comparing the proposed scheme and Random, it is seen that Random results in a wide range of performance distributions by making decisions with equal probabilities. In contrast, the proposed scheme yields a more robust performance by adopting a more reliable strategy-making process. Third, the adoption of an adaptive δ poses advantages in general over fixed δ, except for the V2V performance for SeARL. This is because, as previously mentioned, SeARL achieves the desired reward by predominantly focusing on the objective of sum V2I capacity and sacrificing the objective of minimal V2V capacity. After the parameter δ is introduced and subtracted from the reward, the performance of sum V2I capacity is further boosted and the performance of minimal V2V capacity is thus degraded. Overall, considering the performances of the V2I and V2V links, the proposed method achieves a favorable balance between V2I–V2V tradeoffs by maintaining the sum V2I capacity and fair V2V communications.

Fig. 4 examines the performance of the proposed scheme with various (c_1, c_2) settings in the reward function. Increasing c_1, which corresponds to the weight for the sum V2I capacity in the reward, generally leads to increased sum V2I capacity performance. Likewise, increasing weight c_2 increases the minimal V2V capacity performance. As can be seen, a judicious configuration of (c_1, c_2) is critical to nontrivial V2V performance. Overall, any desired V2I–V2V tradeoff points can be achieved by properly configuring the reward function in the proposed scheme.

V. Conclusion

In this paper, we propose a resource allocation algorithm based on the DRL technique to accomplish the high-capacity and capacity fairness services of V2I and V2V links by exploiting suitable parameter settings in the reward design. Moreover, a detailed investigation is conducted to determine the weighting factors of the proposed algorithm, thus completing the algorithm design. The simulation results confirm the superiority of the proposed algorithm in terms of V2I–V2V tradeoffs. Specifically, based on the results, the achieved capacity of V2I links is found to be acceptable, and the fairness of V2V links is significantly improved.

References

[1] 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. 10 647–10 659, 2017.

[2] 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, 2018.

[3] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 289–299, Aug. 2014.

[4] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transp. Res. part C: Emerg. Technol.*, vol. 89, pp. 384–406, Apr. 2018.

[5] L. C. Bento, R. Parafita, H. A. Rakha, and U. J. Nunes, "A study of the environmental impacts of intelligent automated vehicle control at intersections via V2V and V2I communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 41–59, 2019.

[6] C.-H. Lin, Y.-H. Fang, H.-Y. Chang, Y.-C. Lin, W.-H. Chung, S.-C. Lin, and T.-S. Lee, "GCN-CNVPS: Novel method for cooperative neighboring vehicle positioning system based on graph convolution network," *IEEE Access*, vol. 9, pp. 153 429–153 441, Nov. 2021.

[7] W.-Y. Chen, H.-Y. Chang, C.-Y. Wang, and W.-H. Chung, "Cooperative neighboring vehicle positioning systems based on graph convolutional network: A multi-scenario transfer learning approach," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 3226–3231.

[8] A. J. Alami, K. El-Sayed, A. Al-Horr, H. Artail, and J. Guo, "Improving the car GPS accuracy using V2V and V2I communications," in *IEEE Int. Multidiscip Conf. on Eng. Tech. (IMCET)*, 2018, pp. 1–6.

[9] A. Amini, R. M. Vaghefi, J. M. de la Garza, and R. M. Buehrer, "Improving GPS-based vehicle positioning for intelligent transportation systems," in *IEEE Intell. Veh. Symp. Proc.*, 2014, pp. 1023–1029.

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

[11] 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.

[12] J. Li, J. Zhao, and X. Sun, “Deep reinforcement learning based wireless resource allocation for V2X communications,” in *IEEE Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2021, pp. 1–5.

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

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

[15] D. Zhao, H. Qin, B. Song, Y. Zhang, X. Du, and M. Guizani, “A reinforcement learning method for joint mode selection and power adaptation in the V2V communication network in 5G,” *IEEE Trans. on Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 452–463, Jun. 2020.

[16] J. Tian, Q. Liu, H. Zhang, and D. Wu, “Multiagent deep-reinforcement-learning-based resource allocation for heterogeneous QoS guarantees for vehicular networks,” *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1683–1695, 2021.

[17] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang, and H. Du, “Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications,” *IEEE Trans. Veh. Technol.*, 2022.

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

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

[20] L. Luo, J. Zhang, S. Chen, X. Zhang, B. Ai, and D. W. K. Ng, “Downlink power control for cell-free massive MIMO with deep reinforcement learning,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6772–6777, Jun. 2022.