Non-cooperative Learning for Robust Spectrum Sharing in Connected Vehicles with Malicious Agents

Haoran Peng*, Hanif Rahbari†, Shanchieh Jay Yang†, and Li-Chun Wang*
*Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
†ESL Global Cybersecurity Institute, Rochester Institute of Technology, Rochester, NY, USA
Emails: peng.ee07@nycu.edu.tw, hanif.rahbari@rit.edu, jay.yang@rit.edu, wang@nycu.edu.tw

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10000791

***Abstract*—Multi-agent reinforcement learning (MARL) has previously been employed for efficient spectrum sharing among cooperative connected vehicles. However, we show in this paper that existing MARL models are not robust against non-cooperative or malicious agents (vehicles) whose spectrum selection strategy may cause congestion and reduce the spectrum utilization. For example, a selfish (non-cooperative) agent aims to only maximize its own spectrum utilization, irrespective of the overall system efficiency and spectrum availability to others. We investigate and analyze the MARL-based spectrum sharing problem in connected vehicles including vehicles (agents) with selfish or sabotage strategies. We then develop a theoretical framework to consider the selfish agent, and study various adversarial scenarios (including attacks with disruptive goals) via simulations. Our robust MARL approach where "robust" agents are trained to be prepared for selfish agents in testing phase achieves more resiliency in the presence of a selfish agent and even a sabotage one; achieving 6.7%∼20% and 50.7%∼138% higher unicast throughput and broadcast delivery success rate over regular benign agents, respectively.**

***Index Terms*—Connected vehicle security, spectrum sharing, multi-agent reinforcement learning, Nash equilibrium**

I. Introduction

Connected vehicle (CV) technologies enable the vehicles to connect to other vehicles and/or nearby devices and services to enhance road safety and intelligent transportation [1]–[3]. CV is currently realized mainly through vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications[1] [4]. V2V communication technology can dramatically mitigate traffic collisions and hence reduce fatalities and injuries by exchanging basic safety information, such as location, speed, and direction, among vehicles on the road at extremely low latency [1], [5]. V2I technology further allows communicating with roadside units (RSUs) or locally relevant servers to support various safety and non-safety applications, such as, infotainment and navigation services for drivers and passengers [6]. However, the limited spectrum resources (5.895–5.925 GHz) allocated to cellular V2V and V2I by FCC in the U.S. combined with their stringent quality-of-service (QoS) constraints pose a fundamental challenge to CV technology [7].

To overcome this challenge, cellular V2V and V2I links need to efficiently share the same 5.9 GHz band to improve spectrum efficiency [7]–[9]. According to LTE-V sidelink Mode 4, the baseline V2V mode, vehicles *independently* monitor and select resources from the spectrum resource pool for their direct V2V communications without involving an eNodeB (eNB) [10]. Besides, recent studies assume that for V2I transmissions an RSU selects idle spectrum resources from the same pool and keeps the selected sub-band until the vehicle is disconnected [9]–[11]. An efficient spectrum sharing scheme is essential to guarantee different QoS for coexisting V2V and V2I links. Specifically, the 3rd Generation Partnership Project (3GPP) Release 16 for Cellular Vehicle-to-Everything (C-V2X) generally requires V2V services to support a maximum latency of 20 ms and message payloads ranging from a few hundred to over 2000 bytes; and V2I links to support a maximum latency of 100 ms and any relative velocity up to 500 km per hour [4]. In such a highly dynamic and potentially non-cooperative environment, sidelink transmissions (for short-range V2V/V2I communications) may interfere with each other over the same band and further affect their future actions in resource selection [10]. It is particularly important as each vehicle must re-select its resources frequently because its surrounding environment is constantly changing. Therefore, each vehicle or RSU is a dynamic factor in the local environment and can cause "state transitions" therein.

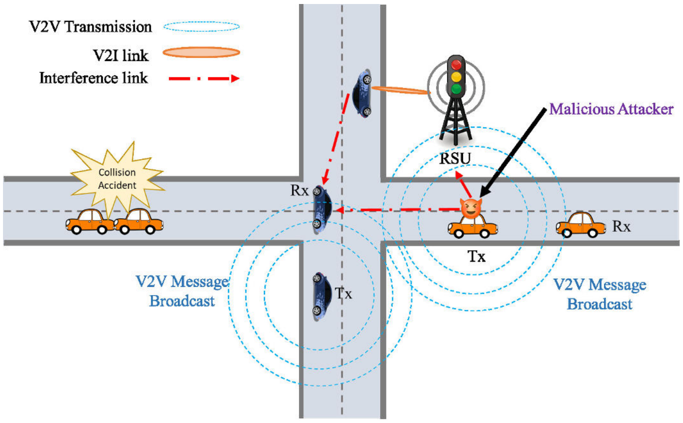

Fig. 1: V2I (unicast) and V2V (broadcast) links coexistence in vehicular environments. All (naive) V2V links share the same observation from the local environment and maintain cooperative goals, whereas malicious/non-cooperative agents employ a selfish or sabotage strategy.

[1] Other CV technologies include vehicle-to-pedestrian (V2P) and vehicle-to-network/cloud (V2N/V2C) that primarily use cellular bands.

978-1-6654-3540-6/22$31.00 © 2022 IEEE

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 11,2023 at 07:14:57 UTC from IEEE Xplore. Restrictions apply.

The interactions between CVs, RSUs, and local communication environments for dynamic spectrum sharing can be modeled as a Markov Decision Process (MDP) [12]. Stakeholders who share the same (open) channel are further exposed to the security and safety risk posed by malicious actors, often adversaries or competitors, who may even aim to reduce the spectrum efficiency and induce traffic congestion or collisions. Therefore, spectrum sharing in decentralized vehicular communications is a challenging MDP problems with highly dynamic vehicles, fast channel variations, critical QoS requirements, and security risks [13].

In decentralized settings, multi-agent reinforcement learning (MARL), which effectively solves decentralized MDP problems by maintaining multiple subtasks for different distributed agents, has been widely adopted to optimize spectrum sharing in CVs [9], [14]–[16]. Current MARL-based studies in CVs assume that each agent shares the same observations and maintains cooperative goals with other agents [9]. However, herein, we show that selfish agents can in fact abuse such mechanisms and disrupt the cooperative efficiency of spectrum-sharing systems, e.g., a selfish vehicle may cause more interference to other V2V and V2I transmissions when trying to maximize its own objectives; reducing the main benefits of CVs or even creating chaos. Previous works [17]–[19] have investigated the use of Nash equilibrium (NE) to treat non-cooperative MARL problems in competitive settings. In this study, we adopted NE in the MARL training for CVs using the max-min function to update the action-value (Q-function) of naive agents. Using this approach, no agent, including the non-cooperative ones, can improve its expected payoff by changing actions in the equilibrium state. The resulting NE gives each agent a dynamic policy to follow when deployed in the testing phase.

Specifically, we investigate the risk of non-cooperative selfish agent and even sabotage attacks for existing “cooperative" spectrum sharing frameworks in CVs (e.g., [9]). The selfish agents care only about their transmissions, ignoring the overall spectrum efficiency. We show that naive agents who aim to cooperate with all agents, including the selfish ones, are likely to suffer, and hence the overall spectrum efficiency will deteriorate. Therefore, under our NE-based MARL framework we propose “robust agents” who are aware of the existence of selfish ones and investigate how well our robust agents perform in the presence of selfish agents and combat against them. The proposed robust agent, trained by reaching the NE point with the selfish agent, continuously provides a resilient response to the actions of the selfish agent during the testing period.

Through simulation, we show that our proposed robust agent efficiently improves V2I throughput and V2V delivery success probability by 6.7%–20% and 50.7%–138%, respectively, compared to naive agents (assuming full cooperative MARL) without non-cooperative training. Our simulations further demonstrate that our robust agent can increase the resiliency even against a sabotage agent who only aims to hamper the system performance. To the best of our knowledge, this study is the first to use NE-based MARL to investigate non-cooperative scenarios for secure spectrum sharing in CVs.

The remainder of this paper is organized as follows. In Section II we present the system model. Section III illustrates the proposed robust MARL-based spectrum sharing for mitigating selfish attacks in CVs. In Section IV, we evaluate the effectiveness of the proposed MARL-based spectrum sharing against selfish and sabotage attacks. Concluding remarks and future work are given in Section V.

II. System Model and Objective Formulation

Based on 3GPP Release 16 for advanced C-V2X services, we consider M V2I links and N V2V transmitters that share the same radio-frequency resources [4]. We assume that each vehicular transmitter (Tx) and receiver (Rx) uses a single antenna, whereas RSUs may have multiple antennas [9]. A V2I link is established between one RSU and one vehicle, and each V2V Tx periodically broadcast safety-critical messages. Typically, the set of V2I links is expressed as $\mathcal{M} = \{1, 2, \ldots, M\}$, and that of V2V transmissions is represented by $\mathcal{N} = \{1, 2, \ldots, N\}$. Each V2V transmission n consists of one Tx and L Rx in its coverage. Following the C-V2X physical-layer design for sidelink, we consider orthogonal frequency-division multiple access, which allows concurrent V2V transmission on orthogonal sub-bands.

A. System Model

Suppose the spectrum is divided into J orthogonal sub-bands for M V2I links served by one RSU, where the set of sub-bands is expressed as $\mathcal{J} = \{1, 2, \ldots, J\}$. We assume that the number of sub-bands is equal to that of V2I links [9]. For clarity, we use different symbols, j and m, to express the indices of sub-bands and V2I links, respectively. V2V transmissions share the J sub-bands with V2I links to enhance the utilization of radio resources, and we assume that each Tx only accesses one sub-band each time slot (subframe) to broadcast safety messages. Besides, multiple vehicles may access the same sub-band in the same time slot because decentralized agents in C-V2X do not perform carrier sensing before transmission, despite using the semi-persistent scheduling algorithm [20]. Different agents accessing the same sub-band simultaneously may cause interference at their receivers. The channel between the Tx of the V2V transmission n and the Rx l ($l \in [1, L]$) over the jth sub-band is expressed as

$$g_n^l[j] = \alpha_n^l h_n^l[j] \tag{1}$$

where $h_n^l[j]$ is the small-scale fading component that follows the exponential distribution [9], [21], and α_n^l is the frequency-independent large-scale fading consisting of shadowing and path losses. Assume that the mth V2I link connects the RSU with the mth vehicle by occupying the jth sub-band. Then, the signal-to-interference-plus-noise ratio (SINR) at the RSU is expressed as

$$\begin{aligned} \gamma_m[j] &= \frac{\hat{g}_m^R[j] P_m[j]}{\sigma^2 + \sum_{n=1}^N g_n^R[j] \hat{P}_n[j] \rho_n[j]} \\ s.t. &\sum_{n=1}^{N} \rho_n[j] \leq 1, \quad \forall j = 1, 2, \ldots, J, \\ &\rho_n[j] \in \{0, 1\} \end{aligned} \tag{2}$$

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 11,2023 at 07:14:57 UTC from IEEE Xplore. Restrictions apply.

where σ^2 is the noise power, $\hat{g}_m^R[j]$ represents the V2I channel between the mth vehicle and the RSU over the jth sub-band, and $g_n^R[j]$ the interfering V2V channel from the nth V2V transmitter to the RSU over the jth sub-band. $P_m[j]$ and $\hat{P}_n[j]$ are the transmit powers of the mth V2I and nth V2V transmitters over the same sub-band, respectively. $\rho_n[j]$ indicates a binary parameter that illustrates the usage of the spectrum resources; $\rho_n[j] = 1$ indicates the nth V2V transmitters uses the jth sub-band, and $\rho_n[j] = 0$ indicates otherwise.

The SINR at each Rx l of the nth V2V transmission over the jth sub-band can be expressed as

$$\hat{\gamma}_n^l[j] = \frac{g_n^l[j]\hat{P}_n[j]}{\sigma^2 + \hat{g}_m^l[j]P_m[j] + \sum\limits_{n' \neq n} g_{n'}^l[j]\hat{P}_{n'}[j]\rho_{n'}[j]} \tag{3}$$

where $g_{n'}^l[j]$ and $\hat{g}_m^l[j]$ denote the interfering channels from the n'th ($n' \neq n$) V2V and mth V2I transmitters to each Rx l of the V2V transmission n, respectively, over the jth sub-band. Hence, we can obtain the capacity of the mth V2I link over the jth orthogonal spectrum sub-band by

$$C_m[j] = W_j \log_2(1 + \gamma_m[j]) \tag{4}$$

where W_j represents the bandwidth of the $j$the sub-band. The channel capacity between the Tx and the lth Rx in the V2V transmission n can be obtained by

$$\hat{C}_n^l[j] = W_j \log_2(1 + \hat{\gamma}_n^l[j]). \tag{5}$$

Therefore, the sum-rate of the broadcast V2V transmission n over the jth orthogonal spectrum sub-band can be expressed as

$$\hat{C}_n[j] = \sum_{l=1}^{L} \hat{C}_n^l[j]. \tag{6}$$

B. Objective Formulations and Threat Model

Under C-V2X, each vehicle periodically broadcasts safety messages over a sub-band selected for a short term (up to 1.5 s). Furthermore, the packet payloads may be elastic to acclimate varying amounts of data [4]. As the success probability of delivering V2V packets is essential to guarantee transportation safety, this study aims to increase the average sum-rate of V2I links while guaranteeing the V2V packet delivery rate. We define the probability of V2V packet delivery as follows [21]:

$$\Pr\left\{\sum_{t=1}^{T}\sum_{j=1}^{J} \rho_n[j](t)\hat{C}_n[j](t) \geq B/\Delta_T\right\}, \quad n \in \mathcal{N}, \tag{7}$$

where Δ_T is the channel coherence time, B is the size of the V2V payload during Δ_T, T is the total payload period, and $\hat{C}_n[j](t)$ is the capacity $\hat{C}_n[j]$, defined in (6), in the tth payload generation slot.

The allocation problem of available resources for naive agents herein is to cooperatively allocate the J spectrum resources and the V2V transmission power $\hat{P}_n[j]$ to increase the average capacity of all naive V2I links and guarantee the probability of successful V2V packet delivery. In each time slot t, the objective of each naive agent can be expressed as

$$\begin{aligned} &\max \sum_{m=1}^{M}\sum_{j=1}^{J} C_m[j](t), \\ &s.t.\ C1\ :\ \sum_{j=1}^{J} \rho_n[j]\hat{C}_n[j](t) \geq B/\Delta_T, \forall n \in \mathcal{N}. \end{aligned} \tag{8}$$

However, a malicious actor can easily change the above objective function and act as a selfish agent that intends to maximize its chances to utilize resources. Specifically, a selfish agent will maximize its V2I capacity while guaranteeing its V2V transmissions to other vehicles without concerning the V2V transmissions of other vehicles. Note that the selfish vehicle assumes that its own safety/benefit is ensured by maintaining its own V2V transmissions. The resulting selfish agent m^*, with its V2V links denoted as n^*, will have its objective as:

$$\begin{aligned} &\max \sum_{j=1}^{J} C_{m^*}[j](t), \\ &s.t.\ C1\ :\ \sum_{j=1}^{J} \rho_{n^*}[j]\hat{C}_{n^*}[j](t) \geq B/\Delta_T. \end{aligned} \tag{9}$$

We note that this simple and local change in the selfish or compromised vehicle can cause catastrophic problems for other vehicles when they lose spectrum resources; potentially leading to vehicle collisions.

III. Robust MARL

The MARL framework consists of a game environment and multiple agents, where each agent has a similar architecture. The concept of the MARL-based learning problem can be formulated as an MDP of the interaction between agents and the external environment. We consider the transmitters of N V2V transmissions as distributed agents, and the MDP of cooperative awareness MARL can be expressed as

$$\mathcal{G} := \left\langle \mathcal{I}, S, \{a^i\}_{i\in\mathcal{I}}, r, \mathcal{P}, \beta \right\rangle \tag{10}$$

where $\mathcal{I} = [I]$ is the set of agents I, which is equal to the number N of V2V transmitters. At each time step t, the ith agent perceives the current state of the environment $s_t \in S$ and takes an action $a_t^i \in A$, according to its policy $\pi_*^i, i \in \mathcal{I}$, from the given set of action space A. Then, the agent receives a shared reward $r_t(s_t, a_t^1, a_t^2, \ldots, a_t^I)$ and an evolved state s_{t+1} from the environment. $\mathcal{P} : s_t \times a_t^1 \times a_t^2 \times \cdots \times a_t^I \rightarrow s_{t+1}$ denotes the state transition probability that maps the probability distribution from the current state of the environment and the interacting action of all agents with the state of the environment of time step $t+1$. Based on these interactions, each agent develops its independent policy, $\pi_*^i : s_t \rightarrow a_t^i, i \in \mathcal{I}$, to maximize long-term reward $\mathcal{R} = \sum_t r_t\beta^t$. $\beta \in [0, 1]$ is a discounting factor. Similarly, the Q-function of each agent i following the optimal policy can be described using the Bellman equation:

$$\begin{aligned} &Q_*^i(s_t, a_t^i, \{a_t^{-i}\}_{-i\in\mathcal{I}}) = r_t + \\ &\beta \sum_{s_{t+1}\in S} \mathcal{P}(s_{t+1} \mid s_t, a_t^i, \{a_t^{-i}\}_{-i\in\mathcal{I}}) V_*^i(s_{t+1}, \pi_*^i, \{\pi_*^{-i}\}_{-i\in I}) \end{aligned} \tag{11}$$

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 11,2023 at 07:14:57 UTC from IEEE Xplore. Restrictions apply.

where $\pi_*^{-i}(i \neq -i)$ and a^{-1} represent the policies and actions of all other agents, respectively. V_*^i is the state-value function, which indicates the sum of discounted expected rewards following the optimal policy and is given by

$$V_*^i(s_{t+1}, \pi_*^i, \{\pi_*^{-i}\}_{-i \in I}) = \\ \max_{a^i} \left\{ r_t + \beta \sum_{s_{t+1} \in S} \mathcal{P}(s_{t+1} \mid s_t, a^i, \{a_t^{-i}\}_{-i \in \mathcal{I}}) V_*^i(s_{t+1}, \pi_*^i, \{\pi_*^{-i}\}_{-i \in I}) \right\} \tag{12}$$

Therefore, the Q-learning update function can be expressed as

$$Q_{t+1}^i(s_t, a_t^i, \{a_t^{-i}\}_{-i \in \mathcal{I}}) \leftarrow (1 - \epsilon_t)\, Q_t^i(s_t, a_t^i, \{a_t^{-i}\}_{-i \in \mathcal{I}}) \\ + \epsilon_t \left[r_t + \beta \max_{a^i} Q_t^i(s_{t+1}, a^i, \{a_t^{-i}\}_{-i \in \mathcal{I}}) \right] \tag{13}$$

where $\epsilon_t \in [0, 1]$ is the learning rate at time step t.

A. Observation and Action Space

The observation Z_t at the tth time step is constructed for the environment state s_t and shared with all naive and selfish agents. Herein, we suppose that the nth V2V transmitter is the ith distributed agent. Hence, the observation and action spaces are given as follows:

- *Observation*: At each time step t, we assume that agent i maintains the nth V2V transmission over the jth sub-band. The observation, Z_t^i, of agent i, for all $i \in \mathcal{I}$, consists of its V2V channel information $\{g_n^l[j]\}_{l=1}^L$, V2V interference channels, $\{g_{n'}^l[j]\}_{l=1}^L$, for all $n' \neq n$, from other V2V transmitters over the same sub-band, V2I interference channels, $\{\hat{g}_m^l[j]\}_{l=1}^L$, from all V2I transmitters over the same sub-band, and its interference channel produced, $g_n^R[j]$, to RSU. Thus, we can express the observation of the ith agent, for all $i \in \mathcal{I}$ as
$$Z_t^i = \{O(s_t, i)\} \tag{14}$$
where O is the observation function, which determines the observation of the current environment state s_t. The observation of each agent i includes its remaining time budget T_i and V2V payload B_i to acquire the queuing state of the V2V links. For each time step t, observation function $O(s_t, i)$ of the ith agent over the sub-band j can be expressed as
$$O(s_t, i) = \left\{ \{\mathcal{L}_i^j\}_{j \in J}, \{G_i^j\}_{j \in J}, B_i, T_i \right\} \tag{15}$$
where $G_i^j = \{\{g_n^l[j]\}_{l=1}^L, \{g_{n'}^l[j]\}_{l=1}^L, \{\hat{g}_m^l[j]\}_{l=1}^L, g_n^R[j]\}$ for all $m \in \mathcal{M}$. $\mathcal{L}_i^j$ represents the received interference power over the sub-band j and is given by
$$\mathcal{L}_i^j = \sum_{l=1}^L P_m[j] \hat{g}_m^l[j] + \sum_{l=1}^L \sum_{n' \neq n} \hat{P}_{n'}^l[j] g_{n'}^l[j] \rho_{n'}[j]. \tag{16}$$
- *Action Space*: For each time step t, the action $a^i(t)$ of the ith agent includes two predominant components, the sub-band and the power level for transmitting V2V messages. Based on a previous report [9], we suppose four sub-bands and four power levels, $[-100, 5, 10, 23]$ dBm, are provided to the V2V transmitters for selection.

B. Reward Function

The objective of each naive agent is to improve the sum rate of all V2I links while ensuring the successful delivery of its V2V payload. For each time step t, we suppose the V2I subobjective is the instantaneous sum rate of all V2I links $\sum_m C_m[j](t)$. The V2V subobjective of each agent is the V2V transmission rate until its payload is successfully delivered, and, thereafter, the reward is a constant σ greater than the maximum possible V2V transmission rate. Thus, the partial reward of V2V transmission n can be expressed as

$$\hat{r}^i(t) = \begin{cases} \sum_{j=1}^J \rho_n[j] \hat{C}_n[j](t), & \text{if } B_i \geq 0, \\ \sigma, & \text{otherwise.} \end{cases} \tag{17}$$

For each time step t, the cooperative reward for each naive agent i can be expressed as follows:

$$r^i(t) = \lambda \sum_m C_m[j](t) + (1 - \lambda) \sum_i \hat{r}^i(t) \tag{18}$$

where $\lambda \in [0, 1]$ is the weight to balance the objectives of the V2V and V2I behavior. However, the malicious agent is trained by a selfish reward without considering the system performance. The reward for each selfish agent $i'(i' \neq i)$ that occupies the sub-band $j' \in \mathcal{J}$ can be expressed as

$$r^{i'}(t) = \lambda C_{m*}[j'](t) + (1 - \lambda) \hat{r}^{i'}(t), \forall m* \in \mathcal{M}. \tag{19}$$

The objective function of the sabotage agent is to reduce the total reward of other naive agents.

C. Robust MARL algorithm

NE in MARL is the equilibrium point of a joint policy, $\pi_* := \left(\pi_*^1, \ldots, \pi_*^I\right)$, sisfied by each agent's policy π_*^i as follows:

$$\mathcal{R}^i(\pi_*^i, \pi_*^{-i}) \geq \mathcal{R}^i(\pi^i, \pi_*^{-i}), \forall i \in \mathcal{I} \tag{20}$$

where $\pi^i \neq \pi_*^i$ represents the deviated policy of agent i. This indicates that there is no incentive for agent i to deviate from π_*^i. Consequently, we formulate the problem of non-cooperative MARL as a robust MDP as

$$\bar{\mathcal{G}} := \left\langle \mathcal{I}, S, \{a^i\}_{i \in \mathcal{I}}, \{\bar{r}^i\}_{i \in \mathcal{I}}, \mathcal{P}, \beta \right\rangle. \tag{21}$$

For convenience, $\mathcal{A} := a^i \times \cdots \times a^I$. $\bar{r}_{i \in I}^i \subseteq \mathbb{R}^{|\mathcal{A}|}$ represents the set of mixed objective rewards, which varies for different agents i. For simplicity, we define the naive agent, robust agent, selfish agent, and sabotage agent as follows:

- *Naive agent*: The agent is trained by cooperative strategy, and it shares the same reward function with other agents. In practice, the naive agent only trains with other naive agents and does not know the selfish agent.
- *Selfish agent*: The agent is trained with naive agents and maintains a selfish objective function.
- *Sabotage agent*: The agent is trained with naive agents and maintains a sabotage objective function.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 11,2023 at 07:14:57 UTC from IEEE Xplore. Restrictions apply.

TABLE I: **Simulation parameters.**

Parameters	**Default Value**
Number of vehicles	4
Number of selfish agents	1
Number of neighbors L	1
Bandwidth per sub-band	1 MHz
The height of the RSU antenna	10 meters
The height of vehicles' antenna	1.5 meters
V2I power P_m	23 dBm
V2V power $\hat{P}_n$	[-100, 5, 15, 23] dBm
The absolute velocity of Vehicles	10 ~ 15 meters/s
B	1200 bytes
σ^2	-114 dBm
Fast fading update	1 msec
Slow fading update	100 msec
The objective weight λ	0.3
Steps per episode	100
Vehicle position update	100 msec

- *Robust agent*: The agent is trained with the selfish agent and maintains an NE-based objective.

For each time step t, we assume that the policy and action of the selfish agent are π' and ${a'}_t$, respectively. Thus, the Q-function of each NE-based robust agent is updated as

$$\begin{aligned} &Q^i_{t+1}(s_t, a^i_t, \{a^{-i}_t\}_{-i\in\mathcal{I}}) \leftarrow (1-\epsilon_t)\, Q^i_t(s_t, a^i_t, \{a^{-i}_t\}_{-i\in\mathcal{I}}) \\ &+ \epsilon_t \left[r_t + \beta\, \mathbf{Nash}_i(s_{t+1}, Q^1_t, Q^2_t, \ldots, Q^I_t)\right] \end{aligned} \tag{22}$$

where $\mathbf{Nash}_i(\cdot)$ is the state-value function of the robust agent i where the NE point is achieved and is given by

$$\begin{aligned} \mathbf{Nash}_i(s_{t+1}, Q^1_t, Q^2_t, \ldots, Q^I_t) = \max_{\pi^i_*(\cdot|s_{t+1})} \min_{{a'}_t} \sum_{a^1_t, a^2_t, \ldots, a^I_t} \\ \left(\pi^1_*(a^1_t) \ldots \pi^I_*(a^I_t) Q^i_t(s_{t+1}, a^i_t, \{a^{-i}_t\}_{-i\in\mathcal{I}}, {a'}_t)\right). \end{aligned} \tag{23}$$

Robust agents are assumed to cooperate to maximize their common interests and reduce the value of the selfish agent. According to Eqs. (18) and (19), the objective based on the NE of the robust agent i is given by

$$\begin{aligned} r^i_{robust}(t) =& \lambda \sum_{m\neq m*} C_m[j](t) + (1-\lambda) \sum_{i\neq i'} \hat{r}^i(t) - \\ & \lambda_* \left(\lambda C_{m*}[j'](t) + (1-\lambda)\hat{r}^{i'}(t)\right) \end{aligned} \tag{24}$$

where λ_* is the weight to balance the objective of maximizing the common interest and reducing the agent's reward.

IV. Performance Evaluation

For each V2V agent, we adopted three fully connected layers with 256, 128, and 64 neurons, respectively. The initial position and direction of each agent are randomly generated and consistently moved during the training or testing phases. Specifically, each vehicle updates its position and neighbor per episode (100 ms). During each episode, the system updates the communication environment per step (1 ms). The simulation parameters are given in Table I [4], [9], [22]. Following [9], we assume that the mth V2I link connects the RSU with the mth vehicle by preoccupying the jth sub-band in our simulations. We considered six scenarios described as follows.

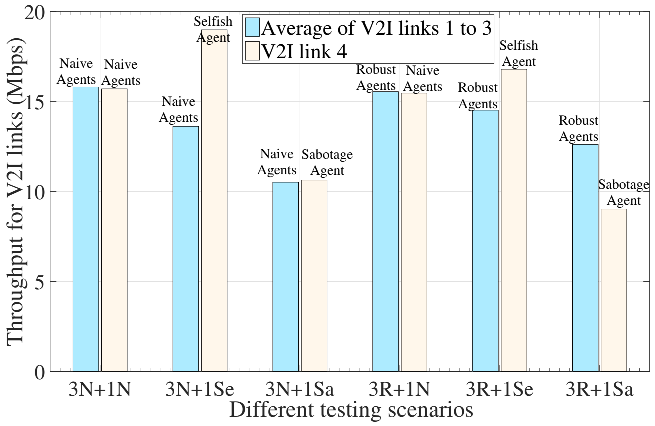

Fig. 2: Average V2I throughput per testing episode.

1) *3 Naive agents + 1 Naive agent (3N+1N)*: Four naive agents are trained by the full cooperative strategy and tested without any malicious agents.
2) *3 Naive agents + 1 Selfish agent (3N+1Se)*: Three naive agents are trained without any malicious agents but tested with one selfish agent.
3) *3 Naive agents + 1 Sabotage agent (3N+1Sa)*: Three naive agents are trained without any malicious agents but tested with one sabotage agent.
4) *3 Robust agents + 1 Naive agent (3R+1N)*: Three robust agents are trained with a selfish agent but tested with a naive agent.
5) *3 Robust agents + 1 Selfish agent (3R+1Se)*: Three robust agents are trained and tested with a selfish agent.
6) *3 Robust agents + 1 Sabotage agent (3R+1Sa)*: Three robust agents are trained with a selfish agent but tested with a sabotage agent.

For each scenario, we conducted 100 independent experiments in the testing phase. To compare the proposed robust agent with the state-of-the-art approaches, this study trains naive agents via the setting in [9]. Therefore, the simulation results of the naive agents represent the performance of the method provided by [9] in our environment.

Fig. 2 shows the V2I performance of the different test scenarios for 100 episodes. Naive agents achieved an average V2I throughput of 15.81 Mbps in an all-cooperative environment, but the throughput dropped to 13.62 Mbps with the selfish agent. Meanwhile, the selfish agent achieved a V2I throughput of nearly 19 Mbps. Robust agents could reach an average V2I throughput of 14.53 Mbps with a selfish agent, a 6.7% improvement over that of the naive agents. The robust agents showed a stable V2I throughput of 14.53 ~ 15.55 Mbps in both cooperative and selfish scenarios. Furthermore, the naive agent only achieves a V2I throughput of 10.52 Mbps in the sabotage scenario, where the robust agent can achieve 12.62 Mbps, a

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 11,2023 at 07:14:57 UTC from IEEE Xplore. Restrictions apply.

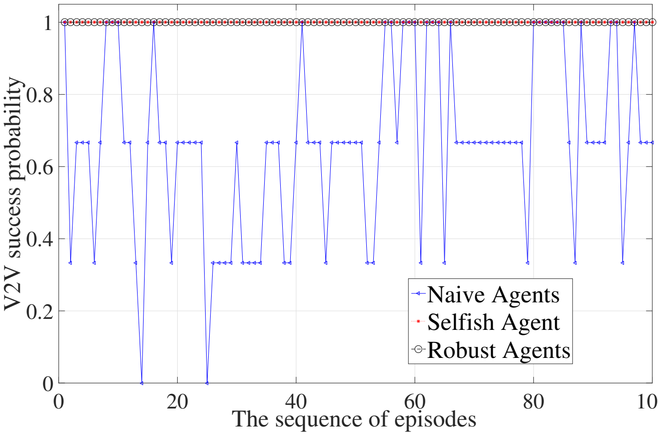

Fig. 3: The V2V delivery success probability per episode.

20% improvement. Notably, the selfish agent achieves higher V2I throughput than the naive agent when the scenario already has robust agents. This is because the current robust agent can improve its situation but cannot penalize the selfish agent.

Fig. 3 shows the probability of V2V payload delivery success for different agents. In the full cooperative scenario, all naive agents could successfully deliver their payloads. However, the average probability of success for naive V2V transmissions is only 66.33% in the selfish scenario. Note that this V2V communication lost can lead to catastrophic problems such as vehicle collisions. The success probability of the V2V transmissions of the robust agent was achieved at 100% in both selfish and cooperative scenarios, which is 50.7% better than that seen by the naive agents. The 100% V2V for the selfish agent helps itself to continue receive critical information from other vehicles. Furthermore, robust agents can provide a success probability of 97.6% V2V transmissions in the sabotage scenario, while naive agents only achieve 41%. Therefore, the robust agent improves 138% over the naive agent concerning V2V performance in the sabotage scenario. This is because the proposed robust agent maximizes its rewards by reaching NE with the selfish agent in dynamic communication environments. With the robust agents recovering the V2V communication, the proposed approach provides a stronger assurance of vehicle safety.

Figs. 4 and 5 show the delivery speed of the V2V payload in the 3N + 1Se and 3R + 1Se scenarios, respectively. As shown in Fig. 4, the delivery speed of the selfish agent was significantly faster than that of the naive agents. The average payload delivery time per naive agent was 52.2 msec, whereas the selfish agent cost only 2.11 msec. Most naive agents could not deliver all payloads in 100 msec. However, the proposed robust agent significantly improved the average delivery time to 11.26 msec, a 78.43% improvement over that of the naive agents, whereas the average delivery time of the selfish agent was 1.75 msec.

In summary, the selfish agent can reduce system performance with full cooperative MARL-based spectrum sharing in vehicular networks. Due to the aggressive and disruptive use of resources by selfish agents, the naive agent cannot deliver all payloads, and its V2I link suffers severe interference. With the proposed NE-based MARL, the robust agents can improve the system performance for both V2I and V2V transmissions. Unfortunately, in our current setup, the selfish agent achieves high performance, even with robust agents. This is because the NE-based policy is to find an equilibrium that reduces the deviate incentive for all agents, not to damage other participants' strategies.

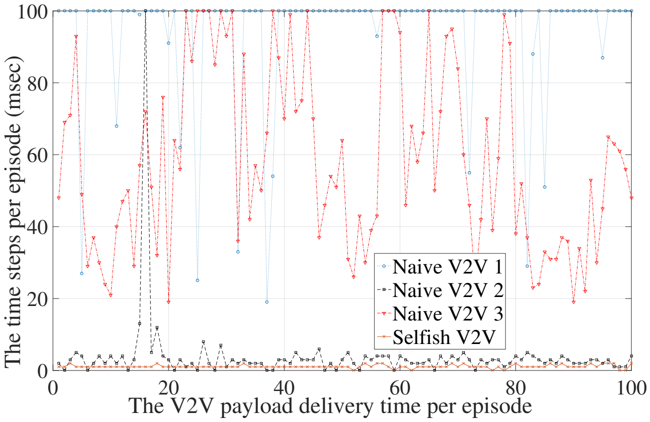

Fig. 4: The V2V payload delivery time per episode for the 3N + 1Se scenario.

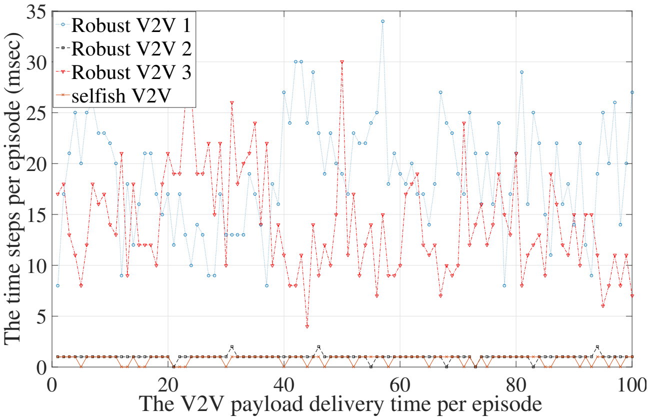

Fig. 5: The V2V payload delivery time per episode for the 3R + 1Se scenario.

V. Conclusions and Future Work

Regular (full cooperative) multi-agent reinforcement learning (MARL)-based spectrum sharing is promised to achieve high spectrum efficiency for V2I and V2V coexisting vehicular networks. However, existing MARL-based spectrum sharing techniques have proven to be vulnerable to selfish and sabotage agents that are, respectively, designed to focus only on its interests and create chaos. In this study, we formulated non-cooperative objectives for naive and malicious agents. We then proposed a robust agent that is based on NE theory for spectrum sharing in non-cooperative scenarios. The proposed robust agent can give the best response to attacks by reducing

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 11,2023 at 07:14:57 UTC from IEEE Xplore. Restrictions apply.

non-cooperative incentives of malicious agents. The simulation results showed the effectiveness and efficiency of the proposed approach for the spectrum sharing in connected vehicles in both cooperative and non-cooperative scenarios. Future studies could look at exploring the resilience of the proposed robust agent to diverse and multiple threat models. Furthermore, the practical system model and simulations based on upcoming 3GPP specification releases are worthwhile to be studied in future work.

Acknowledgment

This work has been partially funded by the Ministry of Science and Technology under the Grants MOST 110-2634-F-A49-006- and MOST 110-2221-E-A49-039-MY3, Taiwan. This work was also financially supported by the Center for Open Intelligent Connectivity from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and was supported by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and MOE, Taiwan.

References

[1] U.S. DOT, “What are connected vehicles and why do we need them?” https://www.its.dot.gov/cv_basics/cv_basics_what.htm, retrieved: 2022-02-23.

[2] 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, May 2014.

[3] W. Yuan, S. Li, L. Xiang, and D. W. K. Ng, “Distributed estimation framework for beyond 5G intelligent vehicular networks,” *IEEE Open J. Veh. Technol.*, vol. 1, no. 1, pp. 190–214, Apr. 2020.

[4] “3GPP Release 16,” https://www.3gpp.org/release-16, retrieved: 2020-07-03.

[5] E. Uhlemann, “Introducing connected vehicles,” *IEEE Veh. Technol. Mag.*, vol. 10, no. 1, pp. 23–31, Feb. 2015.

[6] K. Abboud, H. A. Omar, and W. Zhuang, “Interworking of DSRC and cellular network technologies for V2X communications: A survey,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9457–9470, Dec. 2016.

[7] U.S. Federal Communications Commission, “FCC modernizes 5.9 GHz band for Wi-Fi and auto safety,” https://www.fcc.gov/document/fcc-modernizes-59-ghz-band-improve-wi-fi-and-automotive-safety, retrieved: 2020-11-18.

[8] S. Chen, J. Hu, Y. Shi, L. Zhao, and W. Li, “A vision of C-V2X: Technologies, field testing, and challenges with Chinese development,” *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3872–3881, May 2020.

[9] 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, Aug. 2019.

[10] 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.

[11] Y. Chen, Y. Wang, J. Zhang, and M. D. Renzo, “QoS-driven spectrum sharing for reconfigurable intelligent surfaces (RISs) aided vehicular networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5969–5985, Sep. 2021.

[12] R. Bellman, “A markovian decision process,” *J. Appl. Math. Mech.*, vol. 6, no. 5, pp. 679–684, May 1957.

[13] H. Zhou, W. Xu, Y. Bi, J. Chen, Q. Yu, and X. S. Shen, “Toward 5G spectrum sharing for immersive-experience-driven vehicular communications,” *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 30–37, Dec. 2017.

[14] 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 Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8243–8256, May 2020.

[15] X. Zhu, Y. Luo, A. Liu, M. Z. A. Bhuiyan, and S. Zhang, “Multiagent deep reinforcement learning for vehicular computation offloading in IoT,” *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9763–9773, Nov. 2021.

[16] H. V. Vu, Z. Liu, D. H. N. Nguyen, R. Morawski, and T. Le-Ngoc, “Multi-agent reinforcement learning for joint channel assignment and power allocation in platoon-based C-V2X systems,” *arXiv preprint arXiv:2011.04555*, Nov. 2020.

[17] K. Zhang, T. Sun, Y. Tao, S. Genc, S. Mallya, and T. Basar, “Robust multi-agent reinforcement learning with model uncertainty,” in *Proc. NeurIPS*, vol. 33, no. 1, Dec. 2020, pp. 10 571–10 583.

[18] J. Hu and M. P. Wellman, “Nash Q-learning for general-sum stochastic games,” *J. Mach. Learn. Res.*, vol. 4, no. 1, pp. 1039–1069, Dec. 2003.

[19] M. L. Littman, “Friend-or-foe Q-learning in general-sum games,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 1, Jun. 2001, pp. 322–328.

[20] G. Twardokus and H. Rahbari, “Vehicle-to-nothing? securing C-V2X against protocol-aware DoS attacks,” in *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, May 2022.

[21] L. Liang, H. Ye, and G. Y. Li, “Multi-agent reinforcement learning for spectrum sharing in vehicular networks,” in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[22] Y. d. J. Bultitude and T. Rautiainen, “Ist-4-027756 winner ii d1. 1.2 v1. 2 winner ii channel models,” *EBITG, TUI, UOULU, CU/CRC, NOKIA, Tech. Rep*, 2007.

Authorized licensed use limited to: Beijing Information Science & Tech Univ. Downloaded on February 11,2023 at 07:14:57 UTC from IEEE Xplore. Restrictions apply.