Mean-Field-Aided Multiagent Reinforcement Learning for Resource Allocation in Vehicular Networks

Hengxi Zhang, *Graduate Student Member, IEEE*, Chengyue Lu, *Member, IEEE*, Huaze Tang, *Graduate Student Member, IEEE*, Xiaoli Wei, Le Liang, *Member, IEEE*, Ling Cheng, *Senior Member, IEEE*, Wenbo Ding, *Member, IEEE*, and Zhu Han, *Fellow, IEEE*

***Abstract*—As one technique for autonomous driving, vehicular networks can achieve high efficiency with vehicle-and-infrastructure cooperation, bringing high safety and many value-added services. To achieve higher communication efficiency, much effort has been done to cope with the resource allocation issues for vehicular networks. Nevertheless, due to the strong nonconvexity and nonlinearity, the classical joint resource allocation problem in vehicular networks is typically NP-hard. The multiagent reinforcement learning (MARL) has emerged as a promising solution to tackle this challenge but its stability and scalability are not satisfactory when the amount of vehicles gets increased. In this article, we mainly investigate the issue of joint spectrum and power allocation in vehicular communication networks, and carefully consider the interactions between the vehicles and environment by incorporating the cooperative stochastic game theory with MARL, named complete-game MARL (CG-MARL), to achieve a better convergence and stability with the theoretical computational complexity $\mathcal{O}(n^N)$ with n denoting the dimension of action space and N denoting the number of V2X Vehicular. Furthermore, the mean-field game (MFG) theory is employed to further enhance the MARL for decreasing the horrible computing resource consumption caused by the CG-MARL to $\mathcal{O}(n^2)$ while maintaining an approximate performance. The simulation results demonstrate that the proposed mean-field-aided MARL (MF-MARL) for vehicular network resource allocation can achieve 95% near-optimal performance with much lower complexity, which indicates its significant potentials in the scenarios with massive and dense vehicles.**

***Index Terms*—Joint resource allocation, mean-field game (MFG) theory, multiagent reinforcement learning (MARL), vehicular networks.**

I. Introduction

VEHICLE-TO-EVERYTHING (V2X) communications have attracted tremendous interests over the past few years since the cooperation among vehicles and access points can bring a series of advanced services, such as high traffic safety and fuel efficiency, improved infrastructure utilization, and autonomous driving [1], [2]. Nevertheless, the realization of these services requires an ultrahighly reliable transmission of data among communication devices in the whole V2X system [3]. To adequately utilize the communication resources referred in the third partnership project (3GPP), such as power [4], spectrum [5], time slots [6], beams [7], etc., the resource allocation scheme needs to be designed in a careful manner [8], [9].

Although those traditional methods, such as integer linear (or nonlinear) programming [10], [11], [12], graph theoretical approach [13], greedy scheme [14], simulated annealing method [15], [16], etc., from the classic operations research (OR) field have tried to solve the resource allocation problems from different perspectives, the drawbacks of these approaches are also obvious. For instance, since the objective functions of these V2X models are either nonconvex or NP-hard, the algorithms for obtaining the solution need to be iteratively designed, which makes the computational complexity very high [11]. In addition, these conventional approaches rely heavily on the V2X communication network model [12], and the global channel state information (CSI) in the environment is difficult to collect precisely [15], [17]. Hence, the implementations of these methods are constrained to a few V2X scenarios. In those environments with complex requirements,

Manuscript received 27 February 2022; revised 22 June 2022, 16 August 2022, and 22 September 2022; accepted 9 October 2022. Date of publication 14 October 2022; date of current version 24 January 2023. The work of Le Liang was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2242022R30003 and Grant 2242022K30005, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20220810. The work of Ling Cheng was supported in part by the National Research Foundation of South Africa under Grant 148765, Grant 132651, and Grant 129311. The work of Zhu Han was supported by NSF Grant CNS-2107216, Grant CNS-2128368, and Grant CMMI-2222810. *(Corresponding author: Wenbo Ding.)*

Hengxi Zhang, Chengyue Lu, Huaze Tang, and Xiaoli Wei are with the Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: zhanghx20@mails.tsinghua.edu.cn; hibluedeer@gmail.com; thz21@mails.tsinghua.edu.cn; xiaoli_wei@sz.tsinghua.edu.cn).

Le Liang is with the National Mobile Communications Research Laboratory, Frontiers Science Center for Mobile Information Communication and Security, Southeast University, Nanjing 210096, China, and also with Purple Mountain Laboratories, Nanjing 211111, China (e-mail: lliang@seu.edu.cn).

Ling Cheng is with the School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2050, South Africa (e-mail: ling.cheng@wits.ac.za).

Wenbo Ding is with the Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, and also with RISC-V International Open Source Laboratory, Shenzhen 518055, China (e-mail: ding.wenbo@sz.tsinghua.edu.cn).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446701, South Korea (e-mail: hanzhan22@gmail.com).

Digital Object Identifier 10.1109/JIOT.2022.3214525

the performances of these approaches are not as good as expected.

Beyond those traditional approaches above, the machine learning methods, especially the reinforcement learning (RL) approaches, have provided a promising way toward addressing these long-standing and troublesome optimization problems due to their advanced capabilities in making decisions, especially over those dynamic scenarios under uncertainty [18]. That is why and when we are inspired by the RL approaches and intend to use such techniques to revisit these "old but classic" resource allocation problems in V2X networks. RL has demonstrated the superiority and potential in addressing the NP-hard and nonconvex problems [19] and, hence, been naturally deployed in the vehicular communication network [20]. Yet, to deploy the RL algorithm, the communication network needs to be established as a centralized system with the global information [21], which might not be feasible in reality anyway [22]. Accordingly, for coping with the physical conditions where the agents can only observe and collect the local information, the idea of utilizing the distributed RL, i.e., multiagent RL (MARL), has been proposed. For instance, Liang et al. [23] treated each V2V link as an agent and modeled a multiagent communication system in a distributed manner. Afterward, for maximizing the vehicle-to-infrastructure (V2I) sum-rate while still meeting the probability requirement, Vu et al. employed the double deep Q-learning method to deal with the assignment of both spectrum and power [24]. Nevertheless, due to the lack of consideration in the strategical interactions among vehicle or link agents in the environment, the learning quality and stability of the MARL system is limited, especially when the number of agents goes up. In this work, by taking into consideration that the interactions among these agents in the networks and the strategies of agents can be affected by each other [25], [26], we first integrate the stochastic game theory into the MARL system and rebuild it as a game process with complete information, namely, complete-game MARL (CG-MARL), to investigate how the MARL with game formulation will perform in the issue of joint V2X resource allocation. Besides, since the interactions among agents in N-player game become exponentially complicated with the increased agents [27], the mean-field game (MFG) theory is then introduced to MARL to reduce the complexity of the interactions as well as the computation while still maintaining the approximate performance to CG-MARL, which indicates its huge potentials in the scenarios with massive and dense vehicles.

The main contributions of this work are summarized as follows.

1) A specifically designed V2X resource allocation scheme in assigning joint spectrum and power is established as an MARL formulation and three different MARL protocols (two enhanced MARL method compared with one classic MARL) are designed to solve the optimization issue of system capacity over the V2X communication network.
2) To enhance the classic MARL, we reconsider the interaction mechanism among agents and combine the cooperative stochastic game theory with MARL to improve the cooperation in the whole system. Herein, the complete information is shared in this N-player game and the Nash Q-learning [26] is introduced to update the strategies of agents based on assuming the Nash Equilibrium (NE) behaviors over the current Q-values, which has been mathematically proven to converge under certain restrictions.
3) In consideration of the exponential increment of the computing resource consumption in terms of the amount of agents in an N-player stochastic game, which makes the large-scale implementation of the vehicular network nearly impossible, we propose an enhanced MARL protocol, i.e., mean-field-aided MARL (MF-MARL) approach, to reduce the horrible computational complexity of CG-MARL from $\mathcal{O}(n^N)$ to $\mathcal{O}(n^2)$ while still maintaining a relatively high approximation on the general performance of the system capacity.

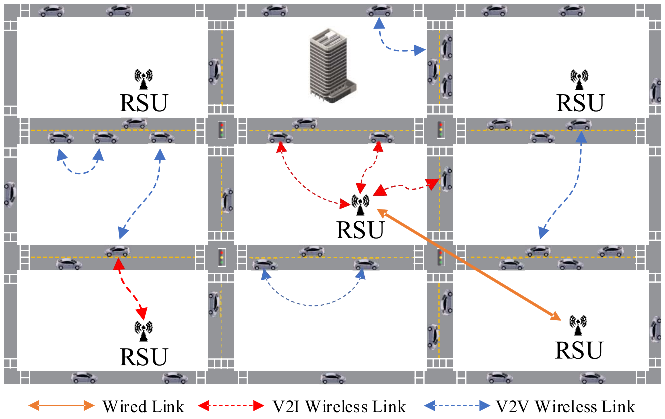


Fig. 1. Schematic of V2X communication networks.

We structure the remainder of this article as follows. In Section II, an evaluation scheme for measuring V2X communication capacity is proposed and the V2V network system is established as a multiagent system with the RL approach. In Section III, we integrate the stochastic game theory into the above original MARL method and afterward keep upgrading it with the MFG theory. Moreover, the comparison of three different MARL algorithms is analyzed. In Section IV, the simulation experiments with all MARL algorithms mentioned above are deployed and the corresponding results are discussed. Finally, conclusion remarks are summarized in Section V.

II. System Model for MARL V2X Networks

In this section, we consider a V2X communication network model consisting of V2I and V2V links, and then incorporate it into the MARL framework where the joint spectrum and power allocation task is performed. The summary of notation of the whole article is represented in Table II.

A. Capacity Evaluation of the Networks

As illustrated in Fig. 1, we suppose there exists a single-cell V2X communication network with K V2I and N V2V links, where V2I links denote the links between K vehicles and the

TABLE I
KEY NOTATIONS

Symbol	Definition
$\gamma_k^c[k]$	SINR of k-th V2I link over k-th spectrum
$\gamma_i^d[k]$	SINR of i-th V2V link over k-th spectrum
ϕ_k^c	Transmit power of k-th V2I link
ϕ_i^d	Transmit power of i-th V2V link
$g_{i,B}$	Channel gain between the i-th V2I transmitter and the BS
g_i	Channel gain of the i-th V2V link
$\hat{g}_{k,B}$	Interference channel gain between the k-th V2I transmitter and the BS
σ^2	Noise power
W	Bandwidth of spectrum
C_k^c	Capacity of the k-th V2I link
C_i^d	Capacity of the i-th V2V link
B_i	Size of the payload of i-th agent
T_i	Time spent by i-th V2V agent for transmitting payload
$\bar{R}$	Average V2V transmission rate
$O(s_t, i)$	Observation of i-th agent at t-th time step
$\kappa_i(t)$	Transmission success rate for i-th agent at time step t
r_t	Reward at time step t
$\mathcal{S}$	State space of V2X environment
$\mathcal{A}^i$	Action space of i-th V2V agent
Ξ	V2V-based stochastic game of V2X network
a_i	Action of i-th V2V agent at time step t
π^i	Strategy of i-th V2V agent
π_*^i	Optimal strategy of i-th V2V agent
$v_i(s, \pi)$	Value of i-th agent with strategy π under state s
$Q_t^i(s, a)$	Q-value of i-th V2V agent with action a and state s at time t
μ_t	State distribution of V2V agents at time t
$m_t(a)$	Mean field of action a at time t
$\bar{\tau}$	Average spectrum efficiency of this V2X environment

base station (BS), while V2V links are employed to support the communications between two individual vehicles.

Both V2I links and V2V links share the same orthogonal spectrum with K sub-bands for the sake of spectral efficiency. In addition, we assume the sub-bands of V2I links are preassigned and there is no interference between V2I links. According to the definition of the signal-to-interference-plus-noise ratios (SINRs) [23], for the kth spectrum, the SINR indices of the kth V2I link and the ith V2V link can be described as

$$\gamma_k^c[k] = \frac{\phi_k^c \hat{g}_{k,B}[k]}{\sum_i \mathbb{1}(a_i = k)\phi_i^d[k] g_{i,B}[k] + \sigma^2} \tag{1}$$

and

$$\gamma_i^d[k] = \frac{\phi_i^d[k] g_i[k]}{P_i[k] + \sigma^2} \tag{2}$$

where the indicator function $\mathbb{1}(a_i = k)$ means the ith V2V link selects the kth spectrum for the payload transmission. In particular, the access for all V2V links is supposed to be less than one band, i.e., $\sum_k \mathbb{1}(a_i = k) \leq 1$, for all $i = 1, 2, \ldots, N$. ϕ_k^c and $\phi_i^d[k]$ correspondingly represent the transmit powers of the kth V2I link and the ith V2V link over the kth spectrum, and σ^2 is the noise power. The interfering channel power gain between the kth V2I transmitter and the BS over the kth spectrum $\hat{g}_{k,B}[k]$ is defined as $\hat{g}_{k,B}[k] = \alpha_{k,B}\hat{h}_{k,B}[k]$, where $\alpha_{k,B}$ describes the large-scale fading effect and $\hat{h}_{k,B}[k]$ denotes the fast-fading depending on the frequency of the spectrum. $g_{i,B}[k]$ and $g_i[k]$ are the channel gain between the ith V2I transmitter and the BS and the channel gain of the ith V2V link over the kth band, respectively, and similarly defined with $\hat{g}_{k,B}[k]$. The interference power of the ith V2V link over the kth spectrum $P_i[k]$ is presented as $P_i[k] = \phi_k^c \hat{g}_{k,i}[k] + \sum_{j \neq i} \mathbb{1}(a_j = k)\phi_j^d[k] g_{j,i}[k]$, where $\hat{g}_{k,i}[k]$ and $g_{j,i}[k]$ denote the interfering channel power gain between the kth V2I transmitter and the ith V2V receiver and the interfering channel power gain between the jth V2V transmitter and the ith V2V receiver over the kth spectrum, respectively.

We furthermore describe the capacity of the kth V2I link over the kth spectrum as $C_k^c[k] = W \log(1 + \gamma_k^c[k])$ and, similarly, the capacity of the ith V2V link can be presented as $C_i^d[k] = W \log(1 + \gamma_i^d[k])$, where W denotes the bandwidth of each spectrum. Both sum capacities of the V2I and V2V links are utilized to evaluate the efficiency of the multiagent communication network η, i.e.,

$$\eta = \omega_c \sum_k C_k^c[k] + \omega_d \sum_i C_i^d[k] \tag{3}$$

where ω_c and ω_d are two hyperparameters to equilibrate the total capacities of V2I and V2I links in the network.

Subsequently, in consideration of the requirement for the instant message delivery, the average V2V transmission rate for the message delivery among the communication network was established to measure the whole performance of the multiagent system given by

$$\bar{R} = \frac{1}{N} \sum_i \frac{B_i}{T_i} \tag{4}$$

where T_i is the time spent by V2V agent i for transmitting the message payload and B_i represents the size of the payload that agent i needs to deliver.

B. MARL for Joint Spectrum and Power Allocation

For obtaining the maximum total capacities of the communication networks mentioned above, we first establish the scenario of joint spectrum and power allocation as an MARL system with the Markov decision process (MDP), which is depicted as Fig. 2, in which each V2V link is treated as an individual agent with its own decision-making policy. Through continually interacting with the V2X environment, each agent will receive a series of rewards from the environment based on the actions that they take, which represent the qualities of their policies. Meanwhile, the environment evolves with the actions from the agents. Since the other agents are also observed as a portion of the environment, each agent needs to ceaselessly update and optimize their strategies against the environment according to the so-called reward-oriented experiences. Considering each V2V agent may perform to compete against other agents for higher capacity in the multiagent system, we set up the same reward in this MARL scenario to encourage cooperative behaviors.

1) Environment State: As mentioned above, the evolution of the communication environment is established as a discrete MDP, where each agent i, representing each V2V link,

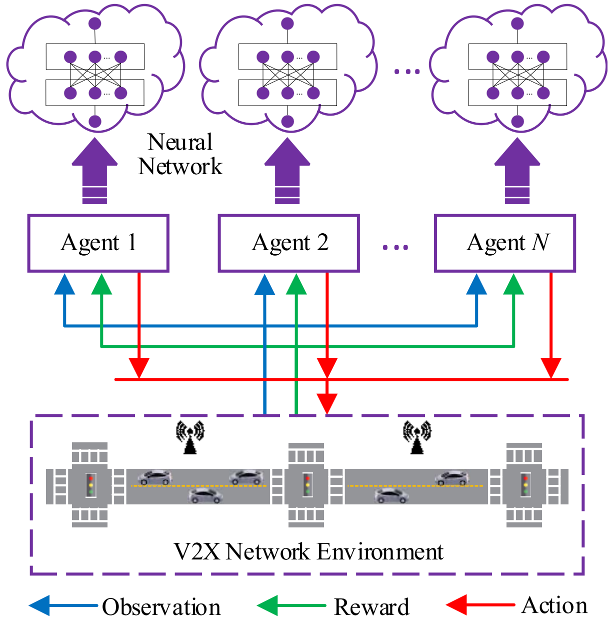


Fig. 2. MDP in RL for the V2X network.

collectively explores and interacts with the unknown V2X communication environment. For handling the issue in which joint spectrum and power need to be appropriately allocated, we include all conditions of the spectrum occupied by V2I and V2V links, all interference caused by all behaviors of agents in the environment state s_t. At each time step t, each V2V agent i needs to perform an observation onto the current environment state s_t, and then selects an action a_t^i from its own action space $\mathcal{A}^i$ based on what it has observed. After all agents executing so, the environment will immediately react to all actions and hand out the rewards, which then leads the environment to evolve from the current state s_t to the next state s_{t+1} with probability $p(s_{t+1}|s_t, \boldsymbol{a}_t)$.

2) Observation Space: In this decentralized multiagent system, each agent can only observe the local conditions of the whole environment. The observation of the environment for agent i at the tth time step O_t^i is hence determined by the observation function $O(s_t, i) = \{B_i^t, T_i^t, \{P_i^t[k]\}_{k=1}^K, \{X_i^t[k]\}_{k=1}^K\}$, where $X_i[k] = \{g_i[k], g_{i,i'}[k], g_{i,B}[k], g_{k,i}[k]\}$. Furthermore, to efficiently help the agents explore the environment with a relatively high exploration rate while maintaining the convergence [28], we adopt the ϵ-greedy policy and update the observation of the environment as

$$Z_t^i = \{O(s_t, i), \epsilon, t\} \tag{5}$$

where $\epsilon \in [0, 1)$ represents the probability of choosing a joint action randomly.

3) Action Space: Action space stipulates the boundary of a series of actions that the agent or player can execute to interact with the environment [29]. In the vehicular networks, we consider the specific spectrum and power as the action space for each individual agent. In most existing literature of spectrum selection and power control, each V2V link preoccupies one disjoint spectrum and takes continuous value. However, to satisfy the practical circuit restriction and facilitate the learning phase, we stipulate the power selection to four options, i.e., [23, 10, 5, −100] dBm similar to the previous works and standards [23]. Notably, the V2V link with no power transmitted is set as −100 dBm. Consequently, as illustrated in Fig. 3, each joint action consists of the spectrum and power and the dimension of the action space is 4 × K. Herein, the one-hot coding technique [30] is employed to encode the spectrum-power pair for controlling the decision making of each agent using the deep neural network.

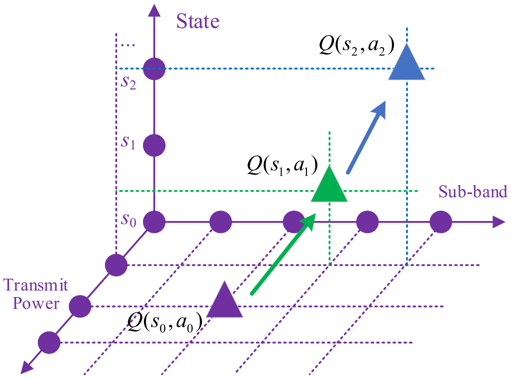


Fig. 3. Action space, where each agent selects a specific transmit power and spectrum at each specific environment state s_t.

4) Design of the Reward: In consideration that all V2I links have already orthogonally preassigned over the specific subbands, and the optimization objective of the proposed MARL is the average V2V transmission rate, we, hence, design the total reward r_{t+1} of the whole multiagent system at time step t as follows:

$$r_{t+1} = \sum_i \kappa_t^i \tag{6}$$

where κ_t^i represents the transmission success rate metric for each agent i at time step t,

$$\kappa_t^i = \begin{cases} \sum_k \mathbb{1}(a_i = k) C_i^d[k, t], & \text{if } B_i \geq 0 \\ c, & \text{otherwise} \end{cases} \tag{7}$$

where c is a constant. Before the total payload is delivered, κ_t^i is set as the effective V2V transmission rate and after the delivery, we employ the constant c as the reward, which is larger than the maximum possible V2V transmission rate and, hence, utilized to encourage the MARL to transmit all the payloads.

III. Mean-Field MARL for V2X Networks

Considering there are multiple agents in the whole V2X communication network, we integrate the stochastic game theory into original MARL and investigate the interaction pattern of this multiagent system. In this section, we first design a multiagent system where the knowledge about other V2V agents is available to all agents, i.e., the stochastic game with complete information, regardless of the exponential expansion of the computational complexity caused by the intricate interaction. Then, the MFG theory is introduced to greatly reduce the computational complexity while

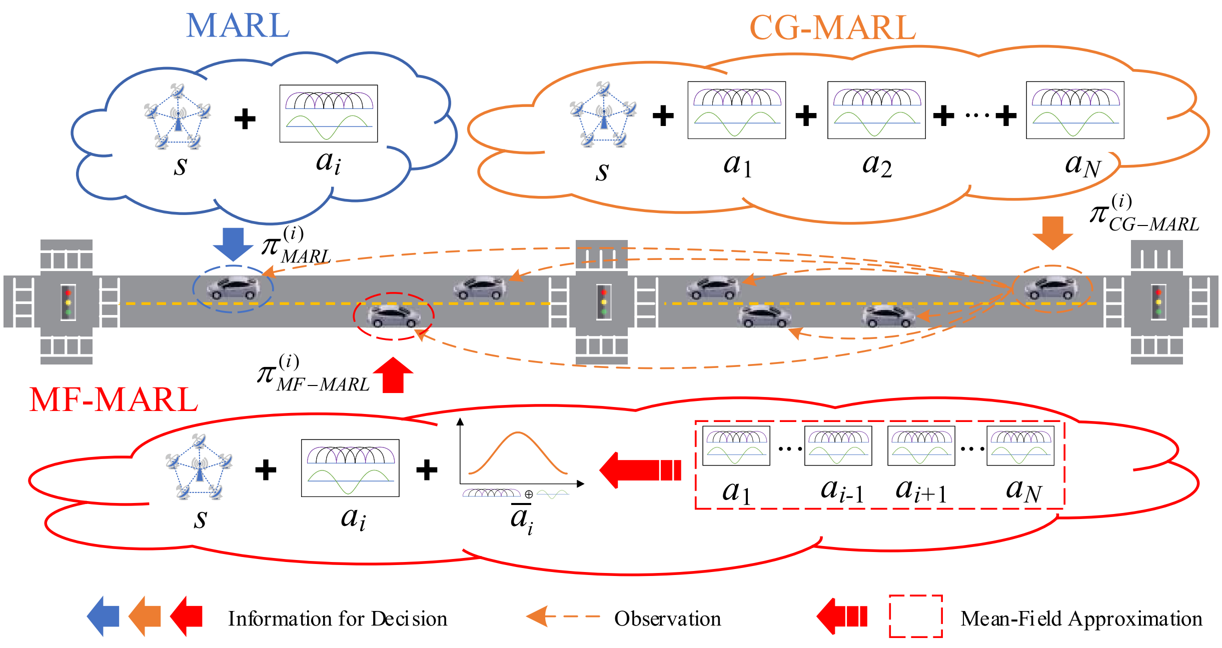


Fig. 4. Schematic illustration and comparison of MARL, CG-MARL, and MF-MARL approaches for V2X resource allocation.

still maintaining a comparable performance. Specifically, the illustrative comparison among the three MARL mechanisms for V2X resource allocation is provided in Fig. 4.

A. CG-MARL

To further investigate the interactions among the agents, we define an N-player stochastic game to describe the dynamic game process [25], [26], [31], [32], which can improve the performance of the multiagent system compared with the MARL one.

1) N-Player V2V-Based Stochastic Games: Suppose there are N homogeneous players, i.e., N V2V agents with same action space, in the game, then the V2V-based stochastic game of this communication network Ξ can be defined as a tuple $\Xi = (\mathcal{S}, \{\mathcal{A}^i\}_{i=1}^N, \{r^i\}_{i=1}^N, p, \beta)$, where $\mathcal{S}$ and $\mathcal{A}^i$ represent the state space of the game, i.e., the V2X environment, and the action space of V2V agent i, respectively. The reward function for player i in the V2X network is denoted as r^i. Given the current environment state $s_t \in \mathcal{S}$, the agent i obtains the observation $z_t^i \in Z_t^i$ and takes an action a_t^i and correspondingly obtains a reward r_t^i from the environment. With all players doing so, the environment state s_t will be updated to the next state s_{t+1} with the transition probability $p : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \to \mathcal{P}(\mathcal{S})$, in which $\mathcal{P}(\mathcal{S})$ denotes the set of probability distribution over state space $\mathcal{S}$. $\beta \in [0, 1)$ is a discount parameter in terms of the reward across time. In addition, as discussed in Section II, considering all agents need to cooperate together to maximize the whole capacity of the same communication network, the reward for each agent is designed as the same, i.e., $r_t = r_t^1 = \cdots = r_t^N$.

In gaming, each V2V agent selects the action following its own strategy $\pi^i : S \to \mathcal{P}(\mathcal{A}^i)$. Since the information in this network is complete, the strategies of other agents can be involved as one of the decision-making references. We use $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^N)$ to denote the joint strategy of all agents. Then, given the specific initial state $z_0 = z$ and the joint strategy of all agents, the value function for each agent i can be expressed to address the following problem:

$$\max_{\pi^i} \; v^i(z;\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi},p}\left[\sum_{t=0}^{\infty}\beta^t r_t^i \mid z_0 = z, \boldsymbol{\pi}\right]$$
$$\text{s.t.} \quad s_{t+1} \sim p(s_{t+1} \mid s_t, \boldsymbol{a}_t), \quad a_t^i \sim \pi_t^i(s_t). \tag{8}$$

2) Equilibrium Strategies: An NE in the communication network can be treated as a joint strategy where each V2V agent does not expect to change its own strategy and prefers to utilize the current one as the best response (BR) against the others. In this V2V-based game, the NE point can be defined as a tuple of N strategies $(\pi_*^1, \ldots, \pi_*^N)$ for all state $z \in Z$ and all agents $i \in \mathcal{N}$ [25] such that for all π^i

$$v^i\left(z; \pi_*^i, \boldsymbol{\pi}_*^{-i}\right) \geq v^i\left(z; \pi^i, \boldsymbol{\pi}_*^{-i}\right) \tag{9}$$

where $\boldsymbol{\pi}_*^{-i} = (\pi_*^1, \ldots, \pi_*^{i-1}, \pi_*^{i+1}, \ldots \pi_*^N)$ denotes the strategies of all V2V agents except for agent i.

3) Nash Cooperative Q-learning: Compared with independent Q-learning, where agents do not communicate with each other, we allow the agents to share information to promote the cooperative behaviors [33]. At each time t, the action selected by agent i is based on its observation of the current observation z_t. Afterward, the agent observes the reward shared by all agents and actions taken by all other agents, as well as the next state s' and observation z'. Then, each agent will immediately calculate the NE $\boldsymbol{\pi}(z')$ for updating its Q-value as follows:

$$Q_{t+1}^i(z_t, \boldsymbol{a}) = (1-\alpha)Q_t^i(z_t, \boldsymbol{a}) + \alpha\left[r_t^i + \beta\mathcal{N}^{Nash}v_t^i\left(z_t'\right)\right] \tag{10}$$

where $\mathcal{N}^{\text{Nash}}$ is an NE operator $\mathcal{N}^{\text{Nash}} = \prod_{i=1}^N \pi_t^i(z')$.

Note that the Q-values of other agents are the critical parameters for each agent to obtain the NE point $\boldsymbol{\pi}(z')$. However,

Algorithm 1 V2X Joint Spectrum and Power Allocation With CG-MARL

```
Initialize: Environment, all agents' $Q$-networks
for episode $epi = 1, 2, \ldots, E$ do
    Set greedy parameter $\epsilon$
    Update agents' positions and channel fadings
    for time step $t = 1, 2, \ldots, T$ do
        Calculate the distances toward all other agents
        for agent $i = 1, 2, \ldots, N$ do
            Choose action $a^i_t$ according to $Q$-value $Q^i_t(z^i_t, \boldsymbol{a}^i_t)$
            Collect all agents' observations $z_t$ and actions $a^{-i} = [a^1, \ldots, a^{i-1}, a^{i+1}, \ldots, a^N]$
            Update $Q$-values $Q^i_{t+1}(z, \boldsymbol{a}) = (1-\alpha)Q^i_t(z, \boldsymbol{a}) + \alpha\left[\boldsymbol{r}_t + \beta\mathcal{N}^{Nash}v^i_t(z')\right]$
        end for
        All agents take actions simultaneously and act to the environment
        Environment hand outs rewards $\boldsymbol{r}_{t+1}$ to all agents
        Update environment
        for agent $i = 1, 2, \ldots, N$ do
            Observe the environment $z_t$
            Store $\{z_t, a^i_t, \boldsymbol{a}^{-i}_t, r_{t+1}, z_{t+1}\}$ in the experience replay memory $\mathcal{E}_i$
        end for
    end for
    Update $Q$-network using Deep $Q$-Network (DQN) method
end for
```

since this part is not shared in the game, each agent needs to speculate it by itself with the iteration of the game. The Q-function of agent j can be conjectured by agent i using the asynchronous updating rule [26], and vector formulation of this process can be expressed as

$$\boldsymbol{Q}_{t+1}(z, \boldsymbol{a}) = (1-\alpha)\boldsymbol{Q}_t(z, \boldsymbol{a}) + \alpha\left[\boldsymbol{r}_t + \beta\mathcal{N}^{\text{Nash}}\boldsymbol{Q}_t(z')\right] \quad (11)$$

where $\boldsymbol{Q}(z, \boldsymbol{a}) = [Q^1(z, \boldsymbol{a}), \ldots, Q^N(z, \boldsymbol{a})]$, and $\boldsymbol{r} = [r^1, \ldots, r^N]$. With the iterative procedure offered above, we are able to compute the convergent Nash strategy in this V2X network using Nash Q-learning. The CG-MARL algorithm is presented as Algorithm 1.

However, this CG-MARL algorithm proposed above will be extremely complicated to implement due to the computational complexity. As the number of V2V agents increases, we must deal with the exponential expansion of the dimension of joint action $\boldsymbol{a}$.

B. MF-MARL

In consideration of the ability of the MFG in presenting the mass behaviors of the multiagent system as a mean-field formulation [34], we then utilize the MFG theory to decline the computational complexity in the CG-MARL, where the dimension of joint action $\boldsymbol{a}$ grows exponentially with respect to the number of agents N.

1) Mean-Field Formulation: In the MFG theory, the interaction between individual agent and the mass of the whole group of a continuum of players are represented as the Hamilton–Jacobi–Bellman (HJB) and Fokker–Planck–Kolmogorov (FPK) equations, respectively [35]. Since the state space over the environment at time t is described as $s_t = \{\{P_t[k]\}_{k\in\mathcal{K}}, \{X_t[k]\}_{k\in\mathcal{K}}\}$ according to the V2X network model established above, we then utilize the FPK equation to model the evolution of the state distribution in a discrete horizon as

$$\mu_{t+1}(s') = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p(s'|a, s, m(a))\pi(a|s)\mu_t(s) \quad (12)$$

where μ_t is the state distribution at time t and $p(\cdot)$ means the transition probability of the state. Here, s and s' denote the current and next states, respectively.

Correspondingly, the action distribution of the multiagent system will get changed due to the update of state distribution, which can be defined as

$$m_t(a) = \sum_{s\in\mathcal{S}} \pi(a|s)\mu_t(s) \quad (13)$$

where $m_t(a)$ denotes the mean field of action a.

On the other hand, the HJB equation can be used to emphasize the value function [36]. For the CG-MARL mentioned above, the value function of agent i in the HJB equation can be reformulated as

$$v^i(z) = \max_\pi \mathbb{E}_\pi\left[\int_0^\infty \beta^t r_t(z^i_t, a^i_t, m_t(a))dt\right] \quad (14)$$

where $z^i_t \in Z^i_t = \{O(s_t, i), \epsilon, t\}$ denotes the observation of agent i, $r(\cdot)$ denotes the reward function, i.e., payoff function, for player i in this dynamic game. Let the discount factor β be $e^{-\lambda}$, $\lambda > 0$, representing the erosion of the reward over time.

Meanwhile, based on the Bellman optimality theorem, where the optimal strategy for agent i should be constituted by its series of future optimal actions, the optimal action at any time can be obtained from

$$\lambda v^i(z) + \mathcal{L}v^i(z) = 0 \quad (15)$$

where $\mathcal{L}v^i(z)$ represents the Hamiltonian operator associated with the dynamics.

To integrate the MFG theory into the Nash Q-learning in CG-MARL with a discrete-time setup [37], we deploy a finite N-player discrete-time MFG model with a finite amount of agent actions and environment states. Then, the value function for V2V agent i in the HJB equation can be reconsidered as

$$v^i_t(z) = \max_\pi \mathbb{E}_\pi\left[\sum_{t=0}^\infty \beta^t r(z_t, a^i_t, m_t(a))\right]. \quad (16)$$

Furthermore, combined with (8) and (16), (14) can be updated as the Bellman equation

$$v^i(z; \boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi},p}\left[r(z, a^i, m(a)) + \beta v^i(z'; \boldsymbol{\pi})\right]. \quad (17)$$

Algorithm 2 V2X Joint Spectrum and Power Allocation With MF-MARL

```
Initialize: Environment, all agents' $Q$-networks
for episode $epi = 1, 2, \ldots, E$ do
    Set greedy parameter $\epsilon$
    Update agents' positions and channel fadings
    for time step $t = 1, 2, \ldots, T$ do
        Calculate the distances toward all other agents
        for agent $i = 1, 2, \ldots, N$ do
            Choose action $a_t^i$ according to $Q$-value
            Observe the environment $Z_t^i$
            Update the mean-field $Q$-value function $Q(z_t, a_t^i, m_t(a))$ using the HJB equation
            Update the strategy $\pi_t^i$ using softmax strategy
        end for
        Update mean field of state distribution and action distribution using the FPK equation
        All agents take actions simultaneously and act to the environment
        Environment hand outs rewards $\boldsymbol{r}_{t+1}$ to all agents
        Update environment
        for agent $i = 1, 2, \ldots, N$ do
            Observe the environment $z_{t+1}^i$
            Store $\{z_t^i, a_t^i, \bar{a}_t^i, r_{t+1}, z_{t+1}^i\}$ in the experience replay memory $\mathcal{E}_i$
        end for
    end for
    for agent $i = 1, 2, \ldots, N$ do
        Update $Q$-network using DQN method
    end for
end for
```

Hereto, the mean-field formulation can be integrally expressed as follows:

$$\begin{cases} v(z;\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi},p}\big[r(z,a,m(a)) + \beta v(z';\boldsymbol{\pi})\big] \\ \qquad \lambda v(z) + \mathcal{L}v(z) = 0 \\ \mu_{t+1}(s) = \sum_s \sum_a p(s'|s,a,m)\pi(a|s)\mu_t(s) \\ \qquad m_t(a) = \sum_s \pi(a|s)\mu_t(s) \end{cases} \tag{18}$$

where $z_t^i \in Z_t^i = \{O(s_t, i), \epsilon, t\}$.

2) MFG to MF-MARL Formulation: Considering the NE of the game or the optimal actions for agents cannot be acquired by the value function $v(\cdot)$, we further utilize the state–action-value function, i.e., the Q-value function, to help each V2V agent effectively obtain its optimal strategy

$$Q^i(z, \boldsymbol{a}) = \max_{a^i \in \mathcal{A}^i} \mathbb{E}\left[\sum_t \beta^t r_t\big(z_t, a_t^i, m_t\big) \mid (z_0, \boldsymbol{a}_0) = (z, \boldsymbol{a})\right]. \tag{19}$$

The Q-function is then factorized in consideration of the pairwise local interactions [38], i.e., $Q^i(z, \boldsymbol{a}) = 1/N^i \sum_j Q^j(z, a^i, a^j)$, where $j \in \mathcal{N}(i)$, $\boldsymbol{a} = [a^1, \ldots, a^N]$, and $\mathcal{N}(i)$ is the index set of the neighboring agents of agent i with size $N^i = |\mathcal{N}(i)|$. With Taylor's theorem, the pairwise Q-function $Q^j(z, a^i, a^j)$ can be furthermore expended and expressed as $1/N^i \sum_j Q^j(z, a^i, a^j) \xrightarrow{\text{Taylor}} Q^j(z, a^i, m(a))$, where $a^i = [a_1^i, \ldots, a_D^i]$ can be interpreted as the empirical distribution of the actions taken by agent i's neighbors and the mean field of action $m^i(t, a)$ can be treated as $\bar{a}^i = 1/N^i \sum_j a^j$ as a specific condition based on the interactions of neighborhood $\mathcal{N}(i)$ of agent i. The interaction is thus simplified and expressed by the mean-field Q-function below.

TABLE II
Computational Complexity

Algorithm	Q-value	Action Space	Complexity
MARL	$Q^i(s, a^i)$	$\mathcal{A}$	$\mathcal{O}(n)$
CG-MARL	$Q^i(s, a^i, a^{-i})$	$\mathcal{A}^N$	$\mathcal{O}(n^N)$
MF-MARL	$Q^i(s, a^i, m(a))$	$\mathcal{A}^2$	$\mathcal{O}(n^2)$

3) Mean-Field Q-Update: With the MF-MARL formulation, the mean-field value function with respect to Q-value for agent i at time t can be obtained as

$$v_{t+1}^i(z') = \sum_{a^i} \pi_{t+1}^i\big(a^i \mid z', m\big)\mathbb{E}\big[Q_{t+1}^i\big(z', a^i, m\big)\big] \tag{20}$$

where the mean-field Q-function can be updated in an MDP manner as

$$Q_{t+1}^i\big(z, a^i, m\big) = (1-\alpha)Q_t^i\big(z, a^i, m\big) + \alpha\big[r_t^i + \beta v_t^i\big(z'\big)\big] \tag{21}$$

and the softmax strategy is employed as

$$\pi_\varphi^i\big(a^i \mid z', m(a)\big) = \frac{\exp\big(\varphi Q^i\big(s, a^i, m(a)\big)\big)}{\sum_{a_j} \exp\big(\varphi Q^i\big(z, a^j, m(a)\big)\big)} \tag{22}$$

where φ is a hyperparameter that controls the Softmax operator [27]. Then, this MF-MARL problem is to figure out the BR π_t^i for agent i at time t, which can be acquired through the mean-field Q-function $Q_t^i(s, a^i, m(a))$. The procedure of MF-MARL method is displayed in Algorithm 2.

C. Complexity Analysis

As the dimension of the joint action $\boldsymbol{a}$ is affected proportionally by the number of the V2V agents, the computational complexity must be considered for deploying the V2X communication networks more efficiently. In this part, we list the dimension of the action space via taking the example of an individual V2V agent, and analyze and compare the computational complexities for three different MARL algorithms in Table II.

1) MARL: The Q-value for each V2V agent in the MARL algorithm includes the state and its own action, i.e., the action for choosing both sub-band and power to transmit the message. Since the dimension of the action and the action space for each V2V agent are two and $\mathcal{A}$, the computational complexity can be presented as $\mathcal{O}(n)$, where n denotes the dimension of the action space.

2) CG-MARL: For CG-MARL, each V2V agent not only needs to focus on its own action but also pays attention to the actions of all other agents, which results in the exponential expansion of the dimension of the joint action. In such a scenario, the action space for each agent becomes $\mathcal{A}^N$. We would say the computational complexity increases to $\mathcal{O}(n^N)$ from $\mathcal{O}(n)$, where N means the number of the V2V agents.

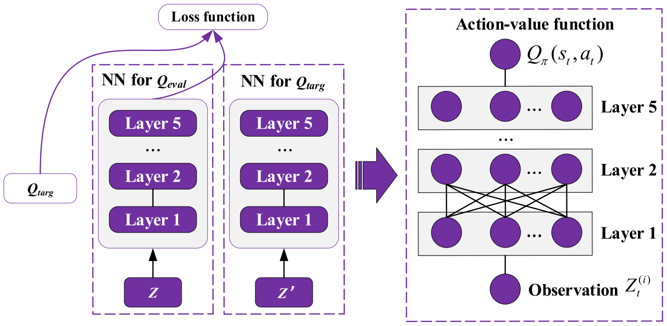


Fig. 5. DQN with experience replay.

3) MF-MARL: In MF-MARL, all actions except for each V2V individual are mathematically generalized as a so-called *mean action* according to the mean-field theory, Thus, in addition to its own action, each agent only needs to observe this mean action that summarizes all action information of the others, where the action space for each V2V agent is immediately reduced to $\mathcal{A}^2$. Meanwhile, the complexity of computation decreases correspondingly to $\mathcal{O}(n^2)$.

D. Deep Q-Network

In the learning phase, considering the high dimension of the observation for the agents, we utilize the DQN method to train the agents' neural networks [19], [39], for effectively supporting the learning process of each agent and obtaining excellent performance in selecting the spectrum and transmit power.

To avoid the waste of training data and break the data correlation in successive training iterations, the deep Q-network method with experience replay is deployed to improve the learning efficiency as well as stabilizing the learning process. The structure of DQN with experience replay is presented in Fig. 5.

At each time step t, the transition experience data $(z_t^i, a_t^i, x_t^i, r_{t+1}^i, z_{t+1}^i)$ is collected by agent i into its memory pool, where x_t^i represents the corresponding action type in CG-MARL and MF-MARL, respectively. Afterward, with a dedicated DQN as the decision-making brain, each V2V agent uniformly samples a minibatch of experiences e from the memory pool at each episode, to update its Q-network weights using the stochastic gradient-descent method. Then, the training of the Q-network can be described as an optimization problem

$$\mathbb{E}_{e\sim\mathcal{E}}\left[r_{t+1} + \beta \max_{a_{t+1}} Q\big(z_{t+1}, a_{t+1}, x_{t+1}; \boldsymbol{\theta}'\big) - Q(z_t, a_t, x_t; \boldsymbol{\theta})\right]^2 \tag{23}$$

where the sum-squared error needs to be minimized. $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ are the parameter sets of the target Q-network and evaluation Q-network, respectively.

IV. Experiments and Results

In the experiment part, we design a V2X communication scenario according to 3GPP Release 15, where the settings are presented as Table III to evaluate the performance of the whole multiagent communication system with the series of implanted MARL approaches. Specifically, the mobility model in this work is based on the Manhattan case defined in Annex A of 3GPP TR 36.885 [2]. To ensure the consistency of experiments, the same amount of V2V agents, from 8 to 32, are deployed in all MARL simulation scenarios.

TABLE III
Simulation Settings [2], [40]

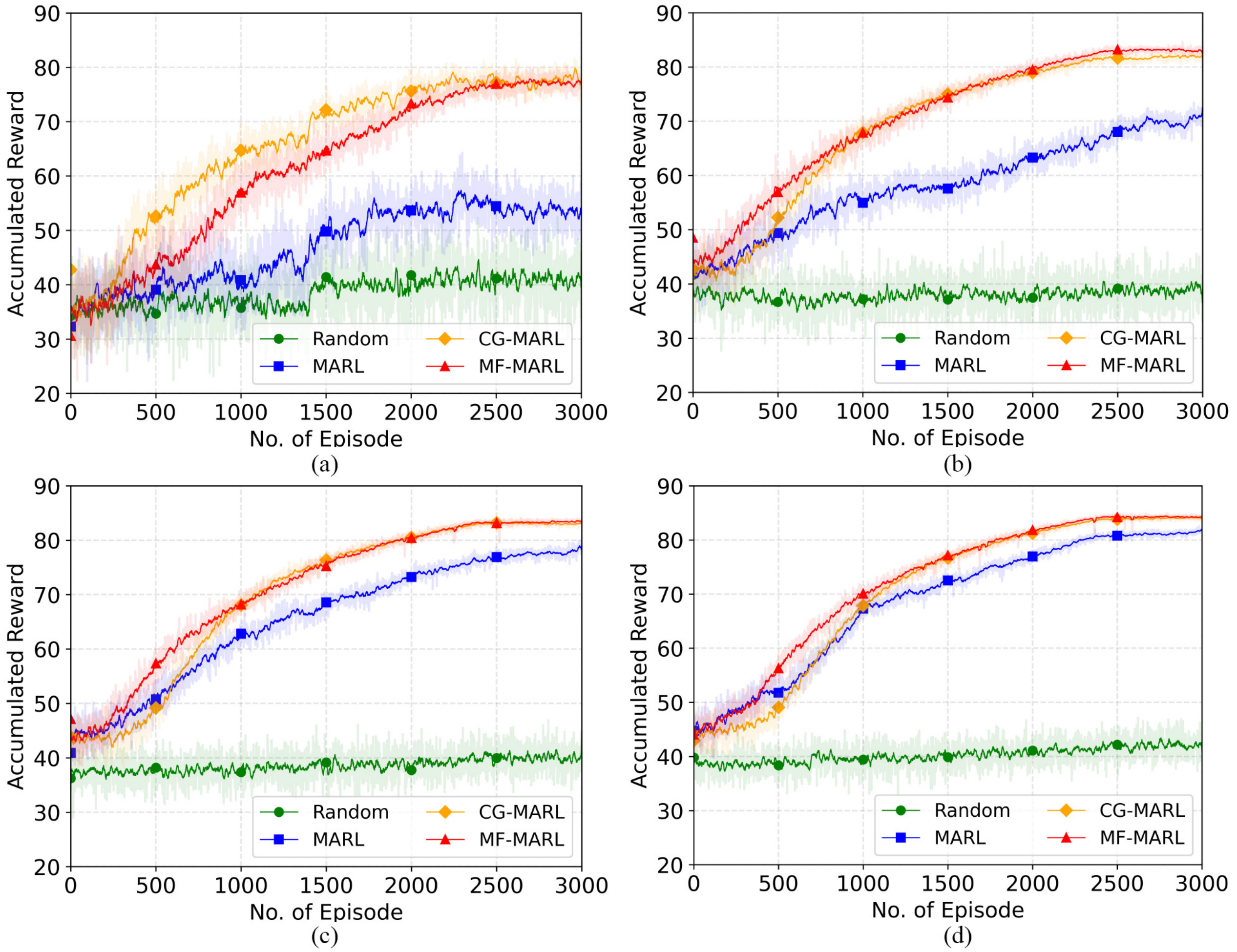
Network Parameter	Detail
Number of V2V Agents	From 8 to 32
Speed of Vehicles	From 36 to 54 km/h
Carrier Frequency	2 GHz
Bandwidth	4 MHz
V2I Transmit Power	23 dBm
V2V Transmit Power	[23,10,5,-100] dBm
Remaining Time for Message Delivery	100 ms
Message Size	2 × 1060 bytes
DQN Parameter	**Detail**
Number of Episode	3000
Time Step	100
ϵ Annealing Length	2400
Learning Rate	0.01
Discount Factor	0.9
Number of DQN Hidden Layers	5
Node Activation DQN	ReLU Function

A. Training Phase

The unicast scheme is implemented in the experimental simulation where the V2V agent of each vehicle chooses the nearest neighbor to transmit the message. When the neighbor is chosen, the V2V agent selects a specific sub-band and power to transmit. We deployed different amount of vehicles (from 8 to 32), respectively, to investigate how the V2V agents will perform in such unicast scenarios. To observe the performance of the proposed MF-MARL proposed above, we set up MARL, CG-MARL as the experimental comparisons, respectively.

As shown in Fig. 6, the performance of MARL is not as excellent as CG-MARL and MF-MARL due to the limited observation information of other agents existing in the same scenario in support of the next action for each agent. Meanwhile, CG-MARL and MF-MARL performed much better than MARL as the Q-value inside for action selecting has action parameters to operate. With both updated MARL algorithms, each V2V agent has a complete view of the action of others, in particular, the agent in the CG-MARL protocol is able to observe each action selected by all individuals while the agent in MF-MARL can only observe the general information, i.e., mean-field of action, of the whole multiagent system.

In Fig. 6, the training performances of MF-MARL and CG-MARL are almost the same for four scenarios in the convergence phase (after 2500 episodes). As the amounts of agents are relatively small, like eight agents in Fig. 6(a), CG-MARL performs better since the agents only need to spend little time on finding the behavioral coordination. However, the coordination strategy for the agent group in the CG-MARL protocol becomes more complex due to the exponentially increasing observation with the number of agents, and more



(a)

(b)

(c)

(d)

Fig. 6. Training performances of agents. (a) 8 Agents. (b) 16 Agents. (c) 24 Agents. (d) 32 Agents.

time is required for agents to form their strategies. Therefore, MF-MARL is able to perform better in the first beginning during the training phase since it costs much lower computing resources and the cooperative mode can be found much quicker in contrast to the CG-MARL, shown in Fig. 6(b)–(d). In addition, the classical MARL (the blue line in Fig. 6) can also be treated as a critical baseline compared with the CG-MARL and MF-MARL methods.

To further investigate the performance of MF-MARL, we take 200 episodes as an observation batch and obtain the approximation gap between CG-MARL and MF-MARL using the mean of the absolute value of both accumulated rewards

$$\delta_i = \frac{1}{U_i} \sum_{u=1}^{U_i} |G^u_{\text{MF-MARL}} - G^u_{\text{CG-MARL}}| \tag{24}$$

where $G_{\text{MF-MARL}}$ and $G_{\text{CG-MARL}}$ denote the accumulated rewards of the multiagent system with CG-MARL and MF-MARL, respectively, and i and U are the number and size of the observation batch, respectively. In Fig. 7, before 500 episodes, the agent groups in both CG-MARL and MF-MARL protocols start exploring. Since the way that they find the optimal strategies are quite distinguishable, the differences of the performances for both modes come out. Yet, the performance gap between CG-MARL and MF-MARL gradually becomes smaller and smaller as training proceeds, and after 500 episodes, the MF-MARL approximates more

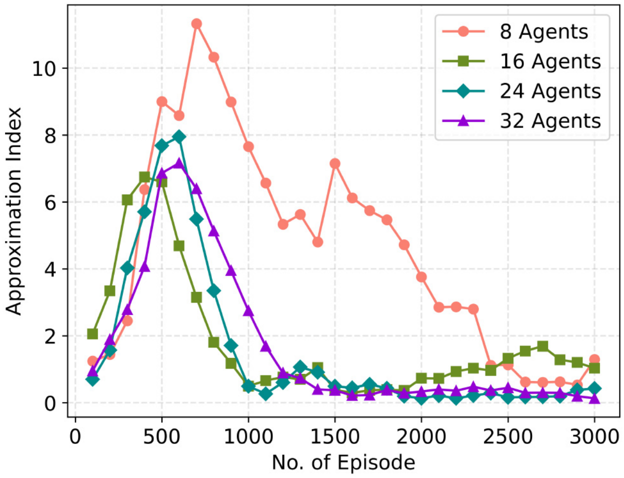
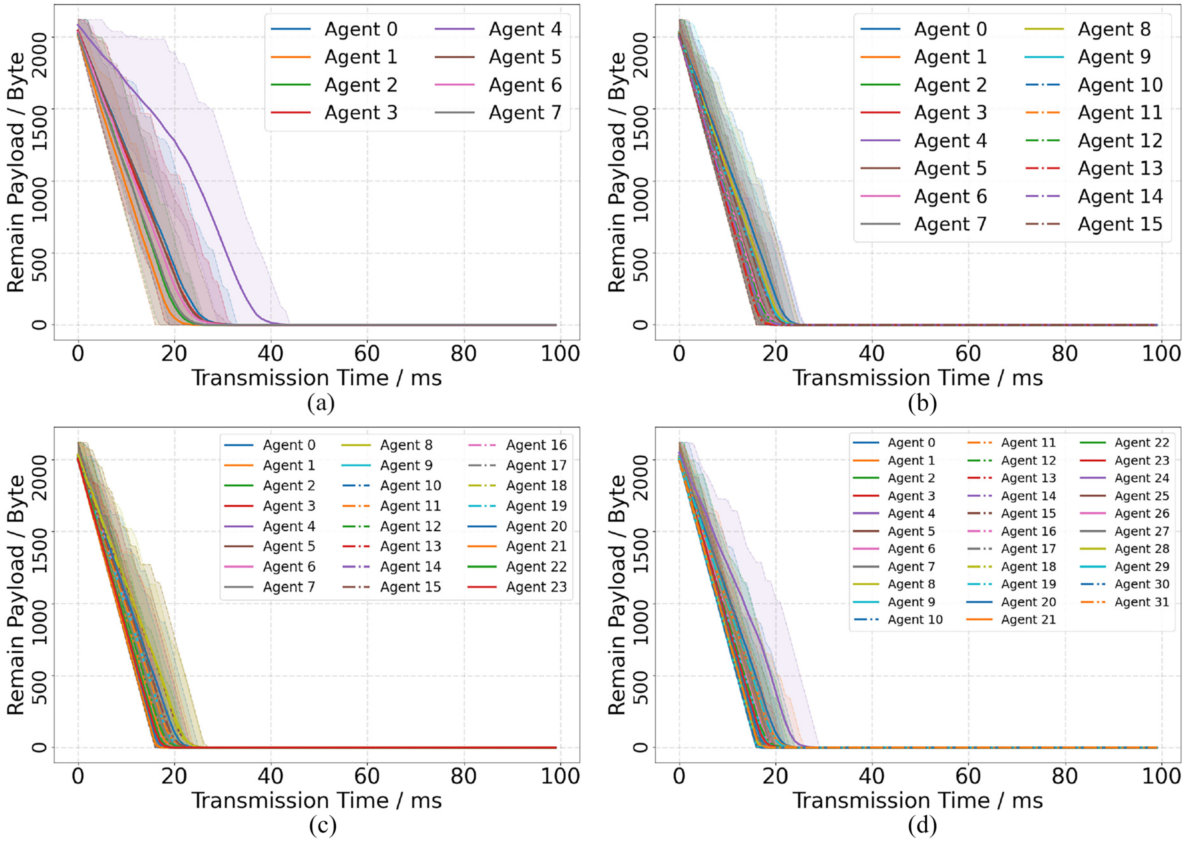


Fig. 7. Reward gap between CG-MARL and MF-MARL.



(a)

(b)

(c)

(d)

Fig. 8. Payload delivery for each MF-MARL agent in the test phase.

to the CG-MARL while still maintaining a much lower computational complexity.

B. Testing Phase

To test the real performance of the MF-MARL, we have built a testing environment to observe how the agents will act after training. The first evaluation index that we take is the payload transmission. Fig. 8 presents that the payload in MF-MARL unicast scenario can be transmitted efficiently by all V2V agents (around 20 ms for each). Note that the transmission rates for agents in communication implementation with a relatively smaller amount of agents, the one with eight agents, are slightly unstable and there exists one V2V agent spending around 40 ms to finish the message delivery. While with the increment of agents, the whole multiagent communication system gradually performs better and agents inside seem to update their strategies appropriately to cooperate with each other due to the mean-field approximation.

Meanwhile, we dig deeper to investigate how much time each V2V agent exactly needs to transmit the message in MF-MARL scenario. Therefore, the transmission time of each agent is studied. The definition of transmission time T_i is given as the time consumed to transmit the payload of 2120 bytes (2 × 1060 bytes), namely

$$T_i = \frac{B_i}{\bar{R}_i} \tag{25}$$

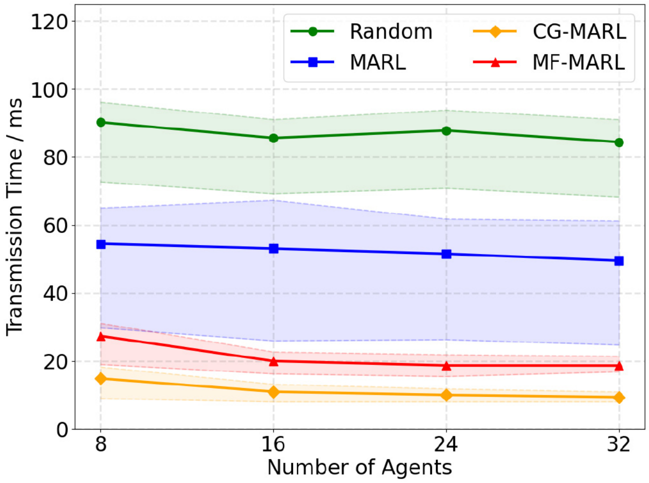


Fig. 9. Average transmission rates for MARL, CG-MARL, and MF-MARL in the test phase.

where B_i denotes the payload of V2V agent i and $\bar{R}_i$ is the average transmission rate for agent i before the transmission finishing. The results of transmission time for all agents corresponding to three different MARL protocols and random mechanism are visualized in Fig. 9. The real line in the figure denotes the average transmission time over all agents, while the light-color area is the distribution of all agents. Due to the instability of the independent MARL, where each agent in the system is not able to consider the information

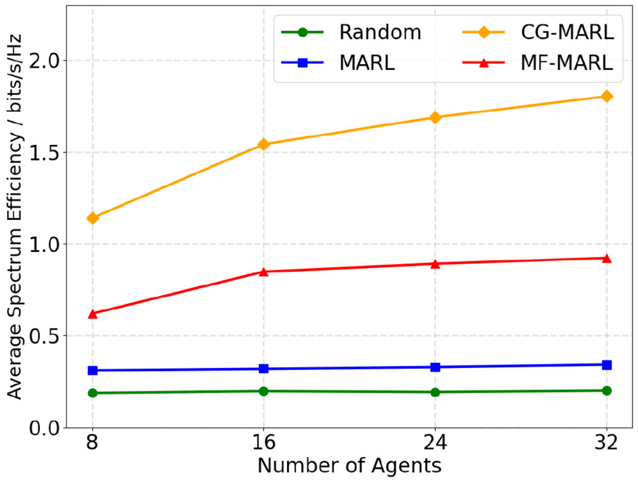


Fig. 10. Average spectrum efficiencies for MARL, CG-MARL, and MF-MARL in the test phase.

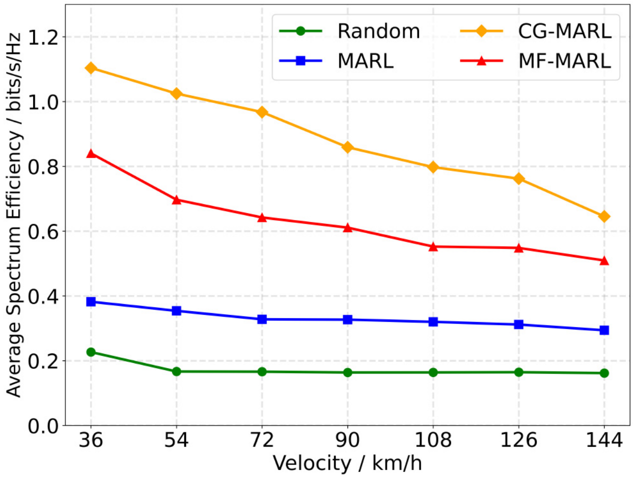


Fig. 11. Average spectrum efficiencies with different velocities.

and strategies from others, there may exist some randomness for this kind of MARL approach, which can be observed from the light blue region. Therefore, we concentrate more on the average tendency of transmission rate for this baseline method drawn in the blue line. However, the transmission time spent by agents in CG-MARL and MF-MARL is still much lower than the one in baseline MARL even with such a condition. With the increment of the number of agents, the performances of CG-MARL and MF-MARL become closer due to the approximation of the MFG. To be specific, the average V2V rate of MF-MARL in transmitting the payload converges to CG-MARL, in particular, while maintaining a much lower computational complexity.

To investigate how the spectrum is leveraged in the multiagent system, we measure the average spectrum efficiency of this V2X environment $\bar{\tau}$ referenced in [41], given as

$$\bar{\tau} = \frac{R_{\text{total}}}{B_{\text{total}}} = \frac{\sum_{k\in[K]}\sum_{i=1}^{N} \mathbb{1}(a_i = k)\bar{R}_i T_i}{WN\sum_{i=1}^{N} T_i} \tag{26}$$

where R_{total} is total transmission rate, B_{total} is total used bandwidth, $\bar{R}_i$ is the average transmission rate for agent i and T_i is the time spent for transmitting the payload. W is the bandwidth. In Fig. 10, the average spectrum efficiencies of three MARL protocols are represented, respectively. Since the agents in the MARL protocol is not able to find the optimal cooperative strategies, which affects how they make use of the spectrum, the average spectrum efficiency is not as good as expected. While the agent groups in both CG-MARL and MF-MARL have already learned behavioral coordination, the spectrum can be utilized effectively by all agents in the V2X environment even with the increment of the number of agents.

We have further analyzed the relationship between the average spectrum efficiency $\bar{\tau}$ and the velocity of the vehicle. The efficiency is evaluated by setting four different velocities from 36 to 144 km/h (or from 10 to 40 m/s) with 18 km/h (or 5 m/s) as the velocity increment in the same scenario. As presented in Fig. 11, for all algorithms, the average spectrum efficiency for the whole multiagent system decreases with the increment of the velocity of the vehicle. However, the performance of each algorithm keeps the same, i.e., CG-MARL $>$ MF-MARL $>$ MARL $>$ Random approach. Since the transmission rate is highly relevant to the fading and path loss, which will be basically affected by the velocity of the vehicle and the distance between vehicles, the faster the vehicles are, the lower the spectrum efficiency is.

V. Conclusion

In this article, we have proposed an enhanced MARL, named the MF-MARL approach, which fuses the MFG theory to MARL in the vehicular network to implement the joint spectrum and power allocation with massive V2V links. According to the experimental simulation, MF-MARL is able to achieve about 95% performance of the CG-MARL while greatly reducing the computational complexity from $\mathcal{O}(n^N)$ to $\mathcal{O}(n^2)$. We believe that the proposed method may overcome the constraints of the agent number in traditional MARL and makes it possible to the deployment of massive agents in V2X communication networks. Furthermore, we also find that the excellent performance of MF-MARL in the spectrum and power allocation reveals the potential in allocating other communication resources as well, and we are also willing to cope with such problems in the short period of future to promote the practical value of our work, which is discussed in Section V of the revised manuscript as well.

References

[1] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, “An open approach to autonomous vehicles,” *IEEE Micro*, vol. 35, no. 6, pp. 60–68, Nov./Dec. 2015.

[2] “Technical specification group radio access network; study LTE-based V2X services; (Release 14),” 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 36.885 V14.0.0, Jun. 2016.

[3] S. S. Husain, A. Kunz, A. Prasad, E. Pateromichelakis, and K. Samdanis, “Ultra-high reliable 5G V2X communications,” *IEEE Commun. Standards Mag.*, vol. 3, no. 2, pp. 46–52, Jun. 2019.

[4] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[5] M. Morelli, C.-C. J. Kuo, and M.-O. Pun, “Synchronization techniques for orthogonal frequency division multiple access (OFDMA): A tutorial review,” *Proc. IEEE*, vol. 95, no. 7, pp. 1394–1427, Jul. 2007.

[6] D. D. Falconer, F. Adachi, and B. Gudmundson, “Time division multiple access methods for wireless personal communications,” *IEEE Commun. Mag.*, vol. 33, no. 1, pp. 50–57, Jan. 1995.

[7] D. J. Love, R. W. Heath, and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[8] Y. Qi, Y. Zhou, Y.-F. Liu, L. Liu, and Z. Pan, "Traffic-aware task offloading based on convergence of communication and sensing in vehicular edge computing," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17762–17777, Dec. 2021.

[9] Y. Qi, L. Tian, Y. Zhou, and J. Yuan, "Mobile edge computing-assisted admission control in vehicular networks: The convergence of communication and computation," *IEEE Veh. Technol. Mag.*, vol. 14, no. 1, pp. 37–44, Mar. 2019.

[10] A. Moubayed, A. Shami, P. Heidari, A. Larabi, and R. Brunner, "Edge-enabled V2X service placement for intelligent transportation systems," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1380–1392, Apr. 2021.

[11] F. Jameel, W. U. Khan, N. Kumar, and R. Jäntti, "Efficient power-splitting and resource allocation for cellular V2X communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3547–3556, Jun. 2021.

[12] X. Li, L. Ma, R. Shankaran, Y. Xu, and M. A. Orgun, "Joint power control and resource allocation mode selection for safety-related V2X communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7970–7986, Aug. 2019.

[13] L. F. Abanto-Leon, A. Koppelaar, C. B. Math, and S. H. de Groot, "Impact of quantized side information on subchannel scheduling for cellular V2X," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, 2018, pp. 1–5.

[14] F. Abbas, P. Fan, and Z. Khan, "A novel low-latency V2V resource allocation scheme based on cellular V2X communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2185–2197, Jun. 2019.

[15] X. Li, L. Ma, Y. Xu, and R. Shankaran, "Resource allocation for D2D-based V2X communication with imperfect CSI," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3545–3558, Apr. 2020.

[16] W.-C. Chiang and R. A. Russell, "Simulated annealing metaheuristics for the vehicle routing problem with time windows," *Ann. Operat. Res.*, vol. 63, no. 1, pp. 3–27, 1996.

[17] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[18] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[19] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[20] T. Şahin, R. Khalili, M. Boban, and A. Wolisz, "Reinforcement learning scheduler for vehicle-to-vehicle communications outside coverage," in *Proc. IEEE Veh. Netw. Conf.*, Taipei, Taiwan, 2018, pp. 1–8.

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

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

[23] 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.

[24] H. V. Vu, M. Farzanullah, Z. Liu, D. H. Nguyen, R. Morawski, and T. Le-Ngoc, "Multi-agent reinforcement learning for joint channel assignment and power allocation in platoon-based C-V2X systems," 2020, *arXiv:2011.04555*.

[25] A. M. Fink, "Equilibrium in a stochastic *n*-person game," *J. Sci. Hiroshima Univ. Ser. AI*, vol. 28, no. 1, pp. 89–93, Jan. 1964.

[26] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, pp. 1039–1069, Nov. 2003.

[27] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," Oct. 2019, *arXiv:1901.09585*.

[28] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proc. Int. Conf. Mach. Learn.*, Vienna, Austria, Aug. 2017, pp. 2681–2690.

[29] A. Kanervisto, C. Scheller, and V. Hautamäki, "Action space shaping in deep reinforcement learning," in *Proc. IEEE Conf. Games (CoG)*, Osaka, Japan, Aug. 2020, pp. 479–486.

[30] S. Okada, M. Ohzeki, and S. Taguchi, "Efficient partition of integer optimization problems with one-hot encoding," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, 2019.

[31] F. Fu and U. C. Kozat, "Stochastic game for wireless network Virtualization," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 84–97, Feb. 2013.

[32] J. Zheng, Y. Cai, Y. Wu, and X. Shen, "Dynamic computation offloading for mobile cloud computing: A stochastic game-theoretic approach," *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 771–786, Apr. 2019.

[33] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, Sao Paulo, Brazil, May 2017, pp. 66–83.

[34] L. Li et al., "Delay optimization in multi-UAV edge caching networks: A robust mean field game," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 808–819, Jan. 2021.

[35] R. A. Banez, L. Li, C. Yang, and Z. Han, *Mean Field Game and Its Applications in Wireless Networks*. Cham, Switzerland: Springer, 2021.

[36] T. Li et al., "A mean field game-theoretic cross-layer optimization for multi-hop swarm UAV communications," *J. Commun. Netw.*, vol. 24, no. 1, pp. 68–82, Feb. 2022.

[37] A. Angiuli, J.-P. Fouque, and M. Lauriere, "Reinforcement learning for mean field games, with applications to economics," Jun. 2021, *arXiv:2106.13755*.

[38] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2018, pp. 5567–5576.

[39] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep Q-learning," in *Proc. 2nd Conf. Learn. Dyn. Control*, vol. 120, Aug. 2020, pp. 486–489.

[40] "WF on SLS evaluation assumptions for EV2X," 3GPP, Sophia Antipolis, France, 3GPP Rep. R1–165704, May 2016.

[41] X. Chen, X. Wu, S. Han, and Z. Xie, "Joint optimization of EE and SE considering interference threshold in ultra-dense networks," in *Proc. Int. Wireless Commun. Mobile Comput. Conf.*, Tangier, Morocco, Apr. 2019, pp. 1305–1310.

Hengxi Zhang (Graduate Student Member, IEEE) is currently pursuing the M.S. degree in data science and information technology with Tsinghua–Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China.

His research interests include multiagent reinforcement learning, game theory, and multirobot system.

Chengyue Lu (Member, IEEE) received the B.S. degree from the School of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou, China, in 2020.

He is currently a Research Assistant of Data Science and Information Technology with Smart Sensing and Robotics Group, Tsinghua University, Shenzhen, China. His research interests include machine learning, Vehicle to Everything, Internet of Things, and robotics.

Huaze Tang (Graduate Student Member, IEEE) received the B.S. degree (Hons.) from Chien-Shiung Wu College, Southeast University, Nanjing, China, in 2021. He is currently pursuing the M.S. degree in data science and information technology with Smart Sensing and Robotics Group, Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

His research interests include reinforcement learning, Internet of Things, and robotics.

Xiaoli Wei received the B.Sc. degree from the University of Science and Technology of China, Hefei, China, in 2014, the M.Sc. degree from the Université Paris Dauphine, Paris, France, in 2015, and the Ph.D. degree in applied mathematics from the Université Paris Diderot, Paris, in 2018.

She is an Assistant Professor with Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. Prior to that, she was a Postdoctoral Fellow with the Department of Industrial Engineering and Operations Research, University of California at Berkeley, Berkeley, CA, USA. Her research interests include stochastic controls, stochastic differential games, and financial mathematics.

Le Liang (Member, IEEE) received the B.E. degree in information engineering from Southeast University, Nanjing, China, in 2012, the M.A.Sc. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2015, and the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2018.

He is with the National Mobile Communications Research Laboratory, Frontiers Science Center for Mobile Information Communication and Security, Southeast University, and also with the Purple Mountain Laboratories, Nanjing. He was a Research Scientist with Intel Labs, Hillsboro, OR, USA, from 2019 to 2021. He has been with the National Mobile Communications Research Laboratory, Southeast University, since 2021. His main research interests are in wireless communications, signal processing, and machine learning.

Dr. Liang received the Best Paper Award of IEEE/CIC ICCC in 2014 and was named an Exemplary Reviewer of the IEEE Wireless Communications Letters in 2018. He has been serving as an Editor for the IEEE COMMUNICATIONS LETTERS since 2019. He served as an Associate Editor for the IEEE JSAC Series on Machine Learning in Communications and Networks from 2020 to 2022. He has been a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society since 2021. He was a Technical Program Committee Co-Chair of the 18th International Symposium on Wireless Communication Systems in 2022.

Ling Cheng (Senior Member, IEEE) received the B.Eng. degree *(cum laude)* in electronics and information from Huazhong University of Science and Technology, Wuhan, China, in 1995, and the M.Ing. *(cum laude)* and D.Ing. degrees in electrical and electronics from the University of Johannesburg, Johannesburg, South Africa, in 2005 and 2011, respectively.

He joined the University of the Witwatersrand, Johannesburg, in 2010, where he was promoted to a Full Professor in 2019. He has been a Visiting Professor with five universities and the Principal Advisor for over 40 full research postgraduate students. He has published more than 100 research papers in journals and conference proceedings. His research interests are in telecommunications and artificial intelligence.

Prof. Cheng was awarded the Chancellor's Medals in 2005 and 2019 and the National Research Foundation ratings in 2014 and 2020. The IEEE ISPLC 2015 Best Student Paper Award was made to his Ph.D. student in Austin. He serves as the associate editor for three journals. He is the Vice-Chair of the IEEE South African Information Theory Chapter.

Wenbo Ding (Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) from Tsinghua University, Beijing, China, in 2011 and 2016, respectively.

He worked as a Postdoctoral Research Fellow with Georgia Tech, Atlanta, GA, USA, from 2016 to 2019. He is currently a Tenure-Track Assistant Professor and a Ph.D. Supervisor with Tsinghua–Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, where he leads the Smart Sensing and Robotics Group. His research interests are diverse and interdisciplinary, which include self-powered sensors, energy harvesting, and wearable devices for health and soft robotics with the help of signal processing, machine learning, and mobile computing.

Dr. Ding has received many prestigious awards, including the Gold Medal of the 47th International Exhibition of Inventions Geneva and the IEEE Scott Helt Memorial Award. He has been serving as the Editorial Board Member for the *Digital Signal Processing* since 2021.

Zhu Han (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 1999 and 2003, respectively.

He was a Research and Development Engineer with JDSU, Germantown, MD, USA, from 2000 to 2002. He was a Research Associate with the University of Maryland from 2003 to 2006. He was an Assistant Professor with Boise State University, Boise, ID, USA, from 2006 to 2008. He is currently a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department as well as with the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid.

Dr. Han received the NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, the IEEE Leonard G. Abraham Prize in the field of Communications Systems (Best Paper Award in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS) in 2016, and several best paper awards in IEEE conferences. He is also the Winner of the 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks." He has been a 1% Highly Cited Researcher since 2017 according to the Web of Science. He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018, and has been an AAAS Fellow since 2019 and an ACM Distinguished Member since 2019.