Received 12 September 2022; revised 28 September 2022; accepted 28 September 2022. Date of publication 3 October 2022; date of current version 13 October 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3211340

Distributed Learning-Based Resource Allocation for Self-Organizing C-V2X Communication in Cellular Networks

NAJMEH BANITALEBI[1], PAEIZ AZMI[1] (Senior Member, IEEE), NADER MOKARI[1] (Senior Member, IEEE), ATEFEH HAJIJAMALI ARANI[2], AND HALIM YANIKOMEROGLU[3] (Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran 14115, Iran

[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[3]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

CORRESPONDING AUTHOR: P. AZMI (e-mail: pazmi@modares.ac.ir)

ABSTRACT In this paper, we investigate a resource allocation problem for a Cellular Vehicle to Everything (C-V2X) network to improve energy efficiency of the system. To address this problem, self-organizing mechanisms are proposed for joint and disjoint subcarrier and power allocation procedures which are performed in a fully distributed manner. A multi-agent Q-learning algorithm is proposed for the joint power and subcarrier allocation. In addition, for the sake of simplicity, it is decoupled into two sub-problems: a subcarrier allocation sub-problem and a power allocation sub-problem. First, to allocate the subcarrier among users, a distributed Q-learning method is proposed. Then, given the optimal subcarriers, a dynamic power allocation mechanism is proposed where the problem is modeled as a non-cooperative game. To solve the problem, a no-regret learning algorithm is utilized. To evaluate the performance of the proposed approaches, other learning mechanisms are used which are presented in Fig. 8. Simulation results show the multi-agent joint Q-learning algorithm yields significant performance gains of up to about 11% and 18% in terms of energy efficiency compared to proposed disjoint mechanism and the third disjoint Q-learning mechanism for allocating the power and subcarrier to each user; however, the multi-agent joint Q-learning algorithm uses more memory than disjoint methods.

INDEX TERMS Cellular vehicle-to-everything (C-V2X) communication, PD-NOMA, resource allocation, learning algorithm.

I. INTRODUCTION

IN RECENT years, the growing demand for local wireless services have created various technical challenges in terms of requiring higher throughput, lower end-to-end latency and power consumption. Cellular vehicle-to-everything (C-V2X) communication has been credited as a key technology in the fifth-generation (5G) networks to improve the performance of the systems. It allows closely located devices to directly communicate with each other and share resources with other devices and cellular users without requiring a centralized controller. By contrast, reuse gain can be achieved by the same radio resource for vehicles and D2D pairs in the C-V2X environment [1]. Accordingly, a self-organizing network (SON) has been considered as a key tool to improve the performance of systems with minimal human intervention. SON allows C-V2X communication to adapt to the changes in the network's conditions and lead their strategies to provide optimal performance in a distributed manner [2], [3]. This can enhance intelligent management while decreasing complexity and operational costs.

Nonorthogonal multiple access (NOMA) techniques have been recognized as a key solution for communication networks of the future by providing spectral efficiency, user fairness, enhanced data rates, and reduced latency. It is therefore expected that future wireless networks will use the new technology power domain of NOMA for improving resource allocation design schemes in ultra-dense topology systems [4].

In this paper, we aim to maximize the energy efficiency of an uplink Power domain non orthogonal multiple access (PD-NOMA) system. To reduce the delay during a vehicular conversation, D2D communication is introduced in the V2X environment. In the proposed system, device-to-device (D2D) pairs share the same uplink resources with other vehicles, and interference produced in the network which impacts on the system performance. Thus, we focus on intra-cell interference and use the successive interference cancellation (SIC) technique to manage the interference among the users in a cellular frequency band [5]. An optimization problem is formulated as a nonlinear integer programming problem. Since users autonomously select their subcarriers based on the environmental information about subcarriers, using machine learning methods seems desirable to reduce both signaling overhead and equipment costs in the system.

Q-learning is a recent form of Reinforcement Learning algorithm that does not need a model of its environment and it is able to compare the expected utility of the available actions without requiring a model of the environment. Q-learning has emerged as a valuable machine learning technique for distributed SONs due to having low complexity and converging to an optimal point. In addition, it is shown that through our distributed Q-learning, D2D users not only are able to learn their resources in a self-organized way, but also achieve better system performance than that using traditional method. Furthermore, SONs can allow systems to configure themselves automatically without manual intervention [6]. Q-learning method is selected for solving the resource allocation problem, which in turn leads to find an optimal policy in the sense of maximizing the expected value of the total reward function for the considered system model [7].

In this paper, we propose two machine learning approaches. In the first, a multi-agent Q-learning algorithm is applied for the joint power and subcarrier allocation. In the second approach, the problem is decoupled into two sub-problems: a power allocation sub-problem and a subcarrier allocation sub-problem. We propose a distributed Q-learning method to allocate subcarriers among users. Given an optimal subcarrier allocation, the power allocation sub-problem modeled as a non-cooperative game. To solve the game, a no-regret algorithm which can be executed in a distributed manner is used. To evaluate the performance of our proposed approaches, we utilize a Q-learning based mechanism presented in [8] for our power allocation problem.

A. RELATED WORKS

Several related works have studied resource allocation for C-V2X communication. In [1], a coalition formation game was proposed to maximize the system sum rate in a statistical average sense for cellular users and multiple C-V2X. An OFDMA-based cellular network with specific frequency bands are considered for each user. As far as, using the fixed frequency band for each user does not seem to be the optimal use in energy, we tried to propose a NOMA-based system and learn the optimal subcarriers for the users. In [9], the authors studied a coalition formation game to address the uplink resource allocation problem for multiple cellular users and C-V2X. In [10], the main contribution was to propose a non-cooperative game and real-time mechanism based on deep reinforcement learning to deal with the energy-efficient power allocation problem in C-V2X networks.. In [11], the authors studied the energy-efficient channel assignment problem for a self-organizing D2D network, and they proposed a distributed game theory-based solution to solve it. In [12], a game theory based learning approach to solve the joint power control and subchannel allocation problem for D2D uplink communications was developed. In [13], the authors studied the behavior of two devices attempting to communicate with a base station from the perspective of non-cooperative game theory, specifying both pure and mixed Nash equilibrium. In [14], to address a resource allocation problem, where C-V2X links use resources common to multiple cells, a new game theory based mechanism was proposed, which indicated that each player had an incentive to conceal their information to improve their profits.

However, the papers mentioned above used game theory based mechanisms, they did not address the energy efficiency issue in C-V2X networks with reinforcement learning mechanisms. We exploit the non-cooperative game to model the power allocation subproblem in a PD-NOMA energy-efficient system, and utilize the no-regret learning method for solving it. C-V2X players tried to learn their resources in a self-organizing manner, independently, which in turn, leads to converge to a Nash equilibrium convergence point more quickly than other methods. Moreover, we utilize a Gibbs sampling scheme to solve the proposed game which is a probabilistic method compared to the approach developed in [15].

In [16], the authors developed a carrier sensing multiple access (CSMA) based algorithm to find the optimal distributed channel allocation of D2D networks. In [17], a multi-agent reinforcement learning-based autonomous mechanism was proposed to achieve optimal channel allocation and effective co-channel interference management for D2D pairs. In [18], to improve the spectral efficiency of a C-V2X network, a spectrum sharing scheme was proposed to provide ad-hoc multi-hop access to a network, however, we proposed the distributed Q-learning method for allocating subcarriers, which in turn leads to reach the optimal resources for the users in terms of maximizing the energy efficiency in the C-V2X network.

In [19], an efficient power control algorithm was proposed to maximize the sum rates. In [20], the authors discussed recent advances in the C-V2X communication system design paradigm from the perspective of a socially aware resource allocation scheme. In [21], the authors first analyze the main streams of the cellular-vehicle-to-everything (C-V2X) technology evolution within the third generation Partnership

Project (3GPP), with focus on the sidelink air interface. Then, they provide a comprehensive survey of the related literature, which is classified and dissected, considering both the Evolution-based solutions and the 5G New Radio-based latest advancements that promise substantial improvements in terms of latency and reliability. In [22], authors addressed the problem of optimizing the energy efficiency of the system by allocating the power and subcarriers in the SC-FDMA wireless networks. The subcarriers are allocated to the users by adopting a multilateral bargaining model. Then, an optimization problem with respect to user's uplink transmission power is formulated and solved. However, we investigate the problem of energy efficiency of the system in the C-V2X communication network in the PD-NOMA system by using the SIC technique to manage the interferences among the users.

Reference [23] presents Open C-V2X, the first publicly available, open-source simulation model of the third generation partnership project (3GPP) release 14 Cellular Vehicle to everything (C-V2X) sidelink, which forms the basis for 5G NR mode 2 under later releases. In [15], the authors proposed an energy-efficient self-organized cross-layer optimization scheme in an OFDMA-based cellular network to maximize the energy-efficiency of a D2D communication system, without jeopardizing the quality-of-service (QoS) requirements of other tiers. In [24], the authors studied interference management in hybrid networks consisting of D2D pairs and cellular links, and they proposed a distributed approach that required minimal coordination yet achieved a significant gain in throughput. In [25], a two-phase resource sharing algorithm was proposed for a D2D communication system whose computational complexity could be adapted according to the network condition. In [26], the authors used the concept of convolution to derive a two-parameter distribution that represented the sum of two independent exponential distributions to enhance the performance of the system. In [27], the authors investigated a power-efficient mode selection and power allocation scheme based on an exhaustive search of all possible mode combinations of devices in a D2D communication system. Note that we utilize an exhaustive search method for joint power and subcarrier allocation problem to compare the results of the proposed methods with the optimal results for resource allocation problem. In [28], the use of self-organized D2D clustering was advocated over the physical random access channel (PRACH), and two D2D clustering schemes were proposed to solve the problem. In [8], the authors employed a Q-learning method to jointly address the channel assignment and power allocation problem to improve the system capacity. In [29], the authors have pointed out D2D based vehicular communication in the V2X environment. In this, device discovery was established using two different techniques that are direct discovery and direct communication.

Most of the technologies have been employed in Table 1.

However, the aforementioned works did not address the energy efficiency issue in C-V2X networks through optimizing power and subcarrier allocations in a distributed manner. In addition, they did not consider a PD-NOMA system with SIC techniques for interference management with QoS constraints. Moreover, using the fifth-generation (5G) technology leads to increase the accuracy and speed of achieving the optimal results compared with previous works. Compared to other Q-learning based approaches, our proposed model uses an novel reward function to maximize the overall sum rate of cells and guarantee minimum interference among users. Moreover, simulation results show the better performances compared with the Q-learning method adopted from the [8], GABS-Dinkelbach algorithm adopted from the [30], VD-RL algorithm and Meta training mechanism with VD-RL algorithm in [31], which are shown in Fig. 8.

B. CONTRIBUTION

The main contribution of this paper is that it introduces a framework for an energy efficiency optimization problem in a C-V2X networks to allocate subcarrier and power among users [32]. Furthermore, SIC technique is performed in the PD-NOMA system to reduce interference among users [33], [34]. To develop this framework, we present two approaches [35].

- In the first approach, a distributed joint Q-learning mechanism for power and subcarrier allocation is proposed. Vehicles and D2D pairs select their transmit power level based on a Gibbs probability distribution. Optimal actions are determined according to the optimal current policy of the proposed multi-agent Q-learning method.
- In the second approach, the optimization problem is divided into two sub-problems: a subcarrier allocation sub-problem and a power allocation sub-problem, due to both binary and continuous optimization variables.
  - In the subcarrier allocation sub-problem, a distributed Q-learning algorithm to allocate the subcarriers is proposed. The value of this method is shown in designing the reward function which contemplates the SIC technique, probability of each subcarrier and energy efficiency of the system. All of the users in the coverage area of the BS choose the subcarriers as the actions, and in each iteration the maximum reward function would be selected for each user, and whenever the agents select the new subcarrier as an action, the current state would be changed. Accordingly, the optimal subcarriers are determined according to the optimal current policy of the Q-learning method.
  - In the power allocation sub-problem, we use a distributed no-regret learning algorithm. In each iteration, each user selects its strategy independently. Furthermore, this distributed approach does not require a control channel for information sharing, and thereby the signaling overhead would be decreased. This approach is suitable when the

TABLE 1. Comparison of existing works with proposed algorithms.

Reference	Exhaustive	RL	Game	PD-NOMA	SON
[1],[9],[10],[11],[12], [13],[14],[20]	×	×	✓	×	×
[17],[8],[25]	×	✓	×	×	×
[15]	×	×	×	×	✓
[27]	✓	×	×	×	×
[28]	×	×	×	✓	✓
[31]	×	✓	×	✓	×
[34]	×	×	×	✓	×
First joint proposed algorithm	✓	✓	✓	✓	✓
Second disjoint proposed algorithm	✓	✓	×	✓	✓

number of users varies over time, and there is no centralized controller. Furthermore, centralized approaches rely on a single controller. If the controller is compromised, it can lead to failures throughout the network.

The advantage of the first proposed multi-agent joint algorithm is its simplicity and convergence rate relative to the second disjoint Q-learning approach, which requires feedback from UEs. However the proposed multi-agent joint method is about 17% less complex compared with the second disjoint Q-learning method. Increasing the number of subcarriers beyond the 15 cause to increase the complexity of the first multi-agent joint algorithm about 26% than the second disjoint Q-learning method. Moreover, we can show intuitively that the second approach manages the power among UEs more effectively respect to receiving more information from the users during the game. Thus, we can choose the solution that best fits with the priorities of the system.

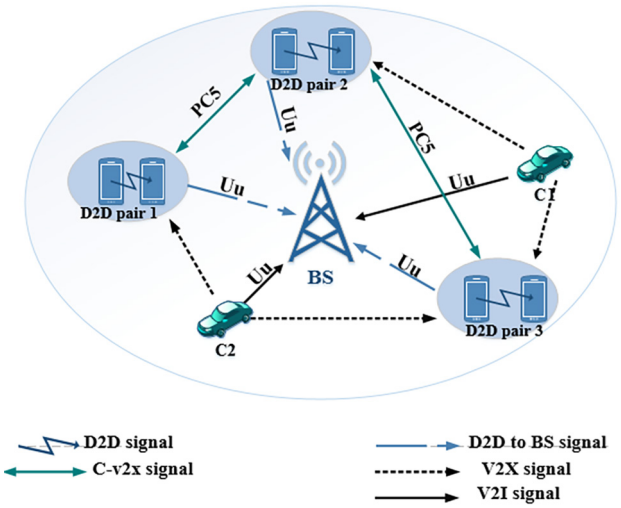


FIGURE 1. Illustration of resource sharing in the system, composed of two vehicles *C*1 and *C*2 and three D2D pairs.

C. ORGANIZATION

The rest of the paper is organized as follows. In Section II, we present the system model and formulate the resource allocation problem. In Section III, we propose a multi-agent joint distributed Q-learning algorithm and a distinct algorithm for allocating the power and subcarrier to each user. We analyze the convergence and complexity of the proposed algorithms in Section IV and V, respectively. In Section VI, we present simulation results. Finally, conclusions are given in Section VII.

II. SOLUTION OF PROPOSED PROBLEM

A. SYSTEM MODEL

We consider a PD-NOMA single-cell system consists of vehicles and D2D pairs shown in Fig. 1, and model the interferences among users in the proposed system model. Considering multi-base stations, just caused to increase in the interferences produced in the system, in which the results are predictable. Thus, to avoid from the complexity of the computation of the interference formula of the system model, we investigate the energy efficiency problem with one base station (BS) located in the center of the area, which is equipped with omni-directional antennas for cellular communications. We assume there are K vehicles labeled as a set of $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ which share their uplink resources with D2D pairs. We denote the set of devices by $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$. We define a binary variable $x_{d_i,n}$ for C-V2X frequency, and thereby if $x_{d_i,n} = 1$, subcarrier n is assigned to the device d_i; otherwise, $x_{d_i,n} = 0$. Similarly for vehicles, $\eta_{c_i,n}$ represents a binary variable that determines the subcarrier assignment for vehicles [36].

The set of all subcarriers is shown by $\mathcal{N}$, and the total available system bandwidth is denoted by B divided into $|N|$ subcarriers with the bandwidth $w = B/N$. In a PD-NOMA system, each subcarrier can be assigned to more than one user, and the corresponding signal is detected by the SIC technique [5]. In this technique, the signal with the highest strength is decoded, subtracted from the combined signal, and a signal with weaker strength is removed. Furthermore, we assume the SIC technique is performed successfully for the user i if

$$\forall n \in \mathcal{N}, \forall i, m \in \mathcal{C} \cup \mathcal{D}, i \neq m, |h_{i,n}|^2 > |h_{m,n}|^2. \quad (1)$$

Since each D2D pair shares the same spectrum with the vehicles or with other D2D pairs, system performance will be reduced; therefore, we focus on the intra-cell interference generated by the users sharing the same frequency band. Three kinds of system interference are described here:

- The vehicle and its corresponding D2D pairs interfere with each other because they share the same uplink spectrum resources.

- The received signals at the BS from the vehicle c_i interfere with the transmitters of the D2D communication system sharing the same spectrum resources in the C-V2X environment.
- The signal at the D2D receiver d_i interferes with the vehicle c_j and the other C-V2X links sharing the same spectrum resources.

The interference power received at vehicle c_i on subcarrier n is defined as (2), shown at the bottom of the page. Parameter $h_{d_i,b,n}$ is a complex Gaussian random variable for the channel coefficient between D2D pair d_i and the BS on subcarrier n, with unit variance and zero mean. Let G_{c_j} denote the transmit antenna gain for vehicle c_j and G_b denote the receive antenna gain for the BS. The signal-to-interference-plus-noise ratio (SINR) of vehicle c_i over subcarrier n is given by

$$\upsilon_{c_i,n} = \frac{|h_{c_i,b,n}|^2 G_{c_i} G_b L^{-\beta}_{c_i,b,n} P_{c_i,n}}{P^{\text{int}}_{c_i,n} + N_{c_i,n} w}. \tag{3}$$

The C-V2X receiver d_i suffers interference from the vehicle c_i and other D2D pairs sharing the same spectrum resources. Therefore, we employ the parameter $P^{\text{int}}_{d_i,n}$ as defined in (4), shown at the bottom of the page, to denote the interference power at D2D $'s$ receiver d_i. Here, $h_{c_i,d_i,n}$ is a complex Gaussian random variable for the channel coefficient gain between D2D pair d_i and vehicle c_i with unit variance and zero mean. Here, G_{d_i} is the transmit antenna gain for D2D pair d_i, and G_{d_j} is the receive antenna gain for D2D pair d_j. The SINR of user d_i over subcarrier n is given by

$$\upsilon_{d_i,n} = \frac{|h_{d_i,d_i,n}|^2 G^2_{d_i} L^{-\beta}_{d_i,d_i,n} P_{d_i,n}}{P^{\text{int}}_{d_i,n} + N_{d_i,n} w}. \tag{5}$$

Accordingly, the problem of allocating resources among D2D users in the C-V2X environment, to maximize the energy efficiency of the system is formulated in the following section.

B. OPTIMIZATION FRAMEWORK

In this section, we formulate an outage-based energy efficiency optimization problem, which is shown in (6), shown at the bottom of the page, and allocates resources effectively to each user, while guaranteeing the QoS requirements for both D2D pairs and vehicles in the C-V2X environment. The system constraints are determined accordingly.

C. SYSTEM CONSTRAINTS

Here, we describe the system constraints, including subcarrier allocation and power allocation constraints, separately.

1) SUBCARRIER ALLOCATION CONSTRAINTS

We define subcarrier allocation constraints in the following form:

$$\eta_{c_i,n}, x_{d_i,n} \in \{0,1\}, \qquad \forall d_i \in \mathcal{D}, c_i \in \mathcal{C}, n \in \mathcal{N}, \tag{7}$$

$$\sum_{n\in\mathcal{N}} x_{d_i,n} \le 1, \qquad \forall d_i \in \mathcal{D}, \tag{8}$$

where (7) indicates the binary variables for cellular and D2D subcarrier assignment, and the constraint defined in (8) indicates that each D2D pair can be assigned to at most one subcarrier.

The SIC technique guarantees that each subcarrier can be reused at most for L_T users. This constraint can be expressed as

$$\sum_{c_i\in\mathcal{C}} \eta_{c_i,n} + \sum_{d_i\in\mathcal{D}} x_{d_i,n} \le L_T, \forall n \in \mathcal{N}, \tag{9}$$

where the system complexity increases as the value of L_T increases. Parameter L_T depends on the signal processing delay in the SIC technique and the receiver's design complexity.

2) POWER ALLOCATION CONSTRAINTS

Parameters $p_{c_i,n}$ and $p_{d_i,n}$ need to satisfy the following constraints:

$$p_{d_i,n} \ge 0, \qquad \forall d_i \in \mathcal{D}, n \in \mathcal{N}, \tag{10}$$

$$p_{c_i,n} \ge 0, \qquad \forall c_i \in \mathcal{C}, n \in \mathcal{N}, \tag{11}$$

$$\sum_{n\in\mathcal{N}} x_{d_i,n} p_{d_i,n} \le P^{\max}_{d_i,n}, \qquad \forall d_i \in \mathcal{D}, \tag{12}$$

$$P^{\text{int}}_{c_i,n} = \sum_{\substack{d_i\in\mathcal{D},\\ |h_{c_i,b,n}|^2 < |h_{d_i,b,n}|^2}} x_{d_i,n} |h_{d_i,b,n}|^2 G_{c_i} G_b L^{-\beta}_{d_i,b,n} P_{d_i,n} + \sum_{\substack{c_i,c_j\in\mathcal{C}, c_i\neq c_j,\\ |h_{c_i,b,n}|^2 < |h_{c_j,b,n}|^2}} \eta_{c_j,n} |h_{c_j,b,n}|^2 G_{c_j} G_b L^{-\beta}_{c_j,b,n} P_{c_j,n}. \tag{2}$$

$$P^{\text{int}}_{d_i,n} = \sum_{\substack{c_i\in\mathcal{C},\\ |h_{d_i,d_i,n}|^2 < |h_{d_i,c_i,n}|^2}} \eta_{c_i,n} |h_{d_i,c_i,n}|^2 G^2_{d_i} L^{-\beta}_{d_i,c_i,n} P_{c_i,n} + \sum_{\substack{d_i,d_j\in\mathcal{D}, i\neq j\\ |h_{d_i,d_i,n}|^2 < |h_{d_i,d_j,n}|^2}} x_{d_i,n} |h_{d_i,d_j,n}|^2 G_{d_i} G_{d_j} L^{-\beta}_{d_i,d_j,n} P_{d_j,n}. \tag{4}$$

$$\text{EE} = \frac{w\left(\sum_{d_i\in\mathcal{D}} \log\left(1+\upsilon_{d_i,n}\right) + \sum_{c_j\in\mathcal{C}} \log\left(1+\upsilon_{c_j,n}\right)\right)}{\sum_{d_i\in\mathcal{D}} P_{d_i,n} + \sum_{c_j\in\mathcal{C}} P_{c_j,n}}. \tag{6}$$

TABLE 2. The main parameters for the proposed resource allocation schemes.

Parameter	Description	Parameter	Description
K	Number of vehicles	M	Number of D2D pairs
$\mathcal{N}$	Set of all subcarriers	w	Subcarrier bandwidth
$\eta_{c_i,n}$	Subcarrier assignment for vehicles	$x_{d_i,n}$	Subcarrier assignment for D2D ppairs
L_T	Maximum users of each subcarrier	β	Pathloss outdoor exponent
$L^{-\beta}_{c_i,b,n}$	Distance between cellular transmitter and BS	$L^{-\beta}_{d_i,c_i,n}$	Distance between D2D pairs transmitter i and vehicle i
$p_{c_i,n}$	Available transmit power for vehicle c_i on subcarrier n	$p_{d_i,n}$	Available transmit power for D2D pair d_i on subcarrier n
$p^{\max}_{c_i,n}$	Transmit power threshold of vehicle i	$p^{\max}_{d_i,n}$	Transmit power threshold of C-V2X links
$N_{c_i,n}$	Cellular spectral density of white Gaussian noise	$N_{d_i,n}$	C-V2X spectral density of white Gaussian noise
$\upsilon_{c_i,n}$	SINR of user c_i over subcarrier n	$\upsilon_{d_i,n}$	SINR of user d_i over subcarrier n
γ_c	Minimum amount for SINR of vehicles	γ_d	Minimum amount for SINR of D2D pairs users
$\pi_{c_i,n(t)}$	Probability of each strategy of cellular player	$\pi_{d_i,n(t)}$	Probability of each strategy of D2D player

$$\sum_{n\in\mathcal{N}} \eta_{c_i,n} p_{c_i,n} \le P^{\max}_{c_i,n}, \qquad \forall c_i \in \mathcal{C}, \tag{13}$$

where (12) and (13) indicate the maximum requirement for the transmit power threshold $P^{\max}_{d_i,n}$ and $P^{\max}_{c_i,n}$ of each D2D pair and cellular user, respectively.

3) QUALITY OF SERVICE CONSTRAINTS

The QoS constraints of all users are expressed on the basis of the minimum SINR demands for D2D pair and cellular users according to (3) and (5) as follows:

$$\upsilon_{c_j,n} \ge \gamma_c, \qquad \forall c_j \in \mathcal{C}, n \in \mathcal{N}, \tag{14}$$

$$\upsilon_{d_i,n} \ge \gamma_d, \qquad \forall d_i \in \mathcal{D}, n \in \mathcal{N}. \tag{15}$$

D. OPTIMIZATION PROBLEM

A network can be configured dynamically on the basis of the transmission power vectors P_c, P_d, and subcarrier allocation vectors η_{c_j}, X_{d_i}, for cellular and D2D pairs, respectively. Accordingly, the following optimization problem can be expressed

$$\max_{\mathbf{P_c},\mathbf{P_d},\eta_\mathbf{c},\mathbf{X_d}} \ \mathsf{EE}, \quad \text{s.t.} \ (7), (8), (9), (10), (11), (12), (13), (14), (15). \tag{16}$$

where EE is the energy efficiency of the system, $\eta_c = [\eta^1_{11}, \ldots, \eta^{\mathrm{N}}_{1N}, \eta^1_{21}, \ldots, \eta^{\mathrm{N}}_{2N}, \ldots, \eta^1_{K1}, \ldots, \eta^{\mathrm{N}}_{KN}]$, $X_d = [x^1_{11}, \ldots, x^{\mathrm{N}}_{1N}, x^1_{21}, \ldots, x^{\mathrm{N}}_{2N}, \ldots, x^1_{M1}, \ldots, x^{\mathrm{N}}_{MN}]$, $P_c = [P^n_1, P^n_2, \ldots, P^n_K]$ and $P_d = [P^n_1, P^n_2, \ldots, P^n_M]$. "The optimization problem (16) consists of non-convex objective functions and both integer and continuous variables. Therefore, we have an NP-hard problem, and the available methods to solve the convex optimization problem can not be applied directly. Furthermore, the formulated problem in its original form is not easy to address in a distributed manner [37]. For simplicity, we break problem (16) down into two sub-problems: a subcarrier allocation sub-problem and a power allocation sub-problem.

1) SUBCARRIER ALLOCATION

The subcarrier allocation sub-problem for vehicles and D2D pairs in the C-V2X environment can be written as

$$\max_{\eta_\mathbf{c},\mathbf{X_d}} \ \mathsf{EE}, \quad \text{s.t.} \ (7), (8), (9), (14), (15). \tag{17}$$

For the power allocation sub-problem, we replace constraints (7), (8) and (9) by constraints (10), (11), (12), and (13).

2) POWER ALLOCATION

Here, by assuming the optimally allocated subcarrier from subcarrier allocation sub-problem, an optimization problem can be simplified and derived in the following way:

$$\max_{\mathbf{P_c},\mathbf{P_d}} \ \mathsf{EE}, \quad \text{s.t.} \ (10), (11), (12), (13), (14), (15). \tag{18}$$

First, we investigate the joint subcarrier and power allocation problem in Section IV. We then investigate sub-problems (17) and (18) and propose distributed learning algorithms for solving them in Sections V and VI, respectively.

III. MULTI-AGENT JOINT POWER AND SUBCARRIER ALLOCATION

A. MULTI-AGENT JOINT POWER AND SUBCARRIER ALLOCATION

In this section, we apply a distributed Q-learning mechanism for joint power and subcarrier allocation based on reinforcement learning. Reinforcement learning is an area of machine learning where agents interact with the environment to reach an optimal solution in an autonomous manner [38].

We use a multi-agent extension of the Markov decision process (MDP) to model multi-agent reinforcement learning. An N-agent Markov game is determined by $(\mathcal{S}, \mathcal{A}, r^1_t, \ldots, r^{(M+K)}_t, p)$, where p is the transition probability $p(s^i_{t+1}|s^i_t, a^t_1, \ldots, a^t_{(M+K)})$ where all the agents take actions a^i_t based on the policy π_i.

We define a set of transmit power levels for vehicles and D2D pairs as $\mathcal{P}_\mathcal{L} = \{P_{\min}, aP_{\min}, a^2P_{\min}, a^3P_{\min}, \ldots, P_{\max}\}$ where $P_{\max}$ and $P_{\min}$ represent the maximum and minimum transmit power for all vehicles and D2D pairs, respectively.

Parameter $(a > 1)$ indicates the number of increasing from one level to another fixed in the dBm domain. At first, each agent selects one power level with uniform probability $\pi_{p_l}(t)$ for the vehicles and D2D pairs. Then, in each iteration, the probability function of each power level would be updated. Since the proposed method uses the Boltzmann-Gibbs distribution and probability law for power-levels, it estimates power levels with specific probability distribution and causes a noticeable change in the system. However, the training process for the vehicles occurs at the BS, and the D2D pairs obtain the trained weights for the actions from the BS in the C-V2X environment. Following the actions lead to transits to a new state s^i_{t+1} by agent i and get a reward r^i_t. The accumulated reward R^i_t over time t is expressed as

$$R^i_{t+T} = \sum_{n=0}^{T} \beta_n r^i_{t+n}, \tag{19}$$

where parameter $0 < \beta < 1$ is a discounted factor. Since no user has enough information about the optimal performance of the network, the learner tries to learn the optimal strategy π^* to maximize the accumulated expected returned reward over time t [8], [39], [40]. When the states are selected, the expected return value can be obtained, and the policy for the state action of agent i can be defined as follows:

$$Q^i_t(s, a) = R^i_t(s, a) = r^i_t(s, a) + \beta R^{i*}_t(s_{t+1}, b). \tag{20}$$

As a matter of fact, we developed non-cooperative mechanisms in a distributed manner to reduce the signaling overhead in the system. In this regard, the reward function needs to be improved to make each agent learn independently from other agents, therefore it only captures the local observations so that it yields sub-optimal solutions.

According to the optimal policy π^*, we can define the $R^{i*}_t(s, a) = \arg\max Q^i_t(s, a)$. Therefore, the Q-function for the expected state-action is updated with the learning rate α shown in equation (21), shown at the bottom of the page.

The optimal value of the action for state s is defined as [41], [42]

$$\max_{\mathbf{a_t}} a^* = \max Q(s, a). \tag{22}$$

Here, we define the agents, states, actions, and reward function.

- *Agents.* All the vehicles and D2D pairs.
- *Actions.* At each step, each agent i takes an action, $a_t \in \mathcal{A}$, which selects a subcarrier with a decision policy π_i. The set of all actions is expressed as $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ where a_i represents the subcarrier of the agent i at time slot t. Moreover, a second case study is also studied where the combined power level and subcarriers are selected as an action. The subcarriers distribution statically depends on the BS decisions, however, power levels depend only on a probability model. Therefore, this action result could not maximize the energy efficiency.
- *States.* The key to affect the state of the network environment is the channel and the transmit power of the players. The QoS of users is restricted by the network environment. We can consider a set of states $\mathcal{S}(\mathcal{U}, \mathcal{A}, \mathcal{P}_\mathcal{L}) = \{s_0, s_1, s_2, \ldots, s_t, \ldots, s_T\}$, where $\mathcal{U} = \{u_{d_1}, u_{d_2}, \ldots, u_{d_M}, u_{c_1}, u_{c_2}, \ldots, u_{c_K}\}$ represents the set of all users, $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ represents the set of actions, and $\mathcal{P}_\mathcal{L} = \{p_{l1}, p_{l2}, \ldots, p_{lL}\}$ represents the set of power levels for the vehicle and D2D users. Here, s_t is the system state at time t and defined as $s_t = (u_i, a^j, p_l{}^q)$ where $1 \le i \le M + K$, $1 \le j \le N$ and $1 \le q \le |\mathcal{P}_\mathcal{L}|$. It indicates that the j^{th} subcarrier and q^{th} transmit power level are assigned to the i^{th} player at time t. As a matter of fact, allocating the power and subcarrier to the user u_i is defined as a current state. Hence, the state space contains $NL(M+K)$ states as $S = \{(u_1, a^1, p_l{}^1), \ldots, (u_1, a^N, p_l{}^L), \ldots, (u_{M+K}, a^N, p_l{}^L)\}$.
- *Reward function.* To maximize the energy efficiency of the system, we define a distributed local reward function related to the energy efficiency of the system as

$$r^i_t = p(u_i|a_j)\mathsf{EE}, \tag{23}$$

where $p(u_i|a_j)$ indicates the probability of the presence of u_i in the subcarrier j. To evaluate the system performance at the end of each epoch, we define ε as the threshold of a new state:

$$\varphi = E_i\Big[Q^i_t\big(s^i_t, a^i_t\big) - Q^i_t\big(s^i_{t-1}, a^i_{t-1}\big)\Big]. \tag{24}$$

Whenever the network satisfies this threshold $\varphi > \varepsilon$, it will start a new round of training based on the current state of the system [38].

B. Q-LEARNING SUBCARRIER ALLOCATION

In this section, we apply distributed learning methods to solve the primary problem by simplifying it into subproblems. Some subcarrier parameters are optimized at each step, while others remain fixed. We propose an iterative Q-learning mechanism for the subcarrier allocation and we describe the action, state, and reward functions here.

- *Agents.* D2D and cellular transmitting nodes.
- *Actions.* The set of all actions is expressed as $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ where a_t represents the subcarrier of the agent i at time slot t.
- *States.* We can consider a set of states $\mathcal{S}(\mathcal{U}, \mathcal{A}) = \{s_0, s_1, s_2, \ldots, s_t, \ldots, s_T\}$, where $\mathcal{U}$ represents the set of all players and $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ represents

$$Q^i_{t+1}(s, a) = (1 - \alpha)Q^i_{t+1}(s, a) + \alpha\Big[r^i_{t+1}(s, a) + \beta \max Q^i_t\big(s', a'\big)\Big] \tag{21}$$

the set of actions (subcarriers) [38]. Here, s_t is the system state at time t, and is defined as $s_t = (u_i, a^j)$ where $1 \leq i \leq M + K$ and $1 \leq j \leq N$. It indicates j^{th} subcarrier assigned to the i^{th} user at time t. Hence, the state space contains $N(M + K)$ states as $S = \{(u_1, a^1), \ldots, (u_1, a^N), \ldots, (u_{M+K}, a^N)\}$.

- *Reward function.* To maximize energy efficiency and guarantee the QoS of the system, we define a reward function related to the SINR constraints of all users. If the SINR constraints are satisfied, the reward function is positive; otherwise it is negative. Accordingly, the following reward function for D2D pairs in the C-V2X environment at time t is defined:

$$r_i^t = \lambda p\big(u_i|a_j\big)\sigma\big(u_i|a_j\big)\upsilon_{u_i}, \tag{25}$$

where λ indicates the SINR coefficient for the reward function and is defined as follows:

$$\lambda = \begin{cases} 1, & \text{if constraints (14) and (15) are satisfied,} \\ -1, & \text{otherwise.} \end{cases} \tag{26}$$

$p(u_i|a_j)$ indicates the probability of presence of u_i in the subcarrier j and $\sigma(u_i|a_j)$ is a binary parameter to satisfy the SIC constraint. It is described below

$$\sigma\big(u_i|a_j\big) = \begin{cases} 1, & \text{if constraint (9) is satisfied,} \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

1) 5G NR INTERFERFACE DECISION

Note that vehicles use the NR V2X PC5-interface for selecting the subcarriers. C-V2X employs two complementary transmission modes, and vehicles autonomously select their sub-channels in C-V2X mode 4. Therefore, C-V2X users would be allocated resources according to the environment information in Q-learning method [43]. In each iteration, the feedback report includes information of the transmission and retransmissions of the subcarriers, and cellular users report an ACK to the base station. After receiving feedback report, the BS evaluates if it has to allocate new subcarrier resources to that C-V2X user or not [7], [44], [45]. After each transmission, new resources or sub-channels must be selected and reserved. New resources must also be selected if selected resources do not fit in the resources previously reserved or do not maximize the energy efficiency of the system. As a result, all the C-V2X users are allocated subcarriers according to decisions of the BS.

C. GAME THEORY BASED FRAMEWORK FOR POWER ALLOCATION

In this section, we aim to solve (18) by assuming optimal subcarriers assigned to the users according to the proposed Q-learning subcarrier allocation method. In the proposed approach, we model the competition among vehicles and D2D pairs as a non-cooperative game, where the vehicles and D2D pairs are players and their transmit power levels are selected independently. Then, we apply a no-regret learning approach to solve the sub-problem. We model sub-problem (18) as a non-cooperative game $g = (\mathcal{U}, \{S_u\}_{u\in\mathcal{U}}, \{u_b\}_{u\in\mathcal{U}})$ where $\mathcal{U} = \mathcal{C} \cup \mathcal{D}$ represents the set of cellular and D2D players. Here we assume $\mathcal{S}_u = \mathcal{S}_c \cup \mathcal{S}_d$, where $\mathcal{S}_c = \{s_{c,1}, \ldots, s_{c,|s_c|}\}$ represents the strategy set of vehicles, and $\mathcal{S}_d = \{s_{d,1}, \ldots, s_{d,|s_d|}\}$ represents the strategy set of D2D players. Therefore, $\mathcal{S}_u = \{s_{u,1}, \ldots, s_{u,|s_u|}\}$ is the strategy set of player u, and $s_{u,i}$ denotes the i^{th} pure strategy of player u. The players, strategy sets, and payoff functions are defined as follows:

- *Players:* These include D2D pairs and vehicles.
- *Strategy sets:* The transmit power threshold of the players is defined as a strategy set of the players. We have $\mathcal{S}_c = \{P_c^{\min}, \frac{1}{|S_c|}P_c^{\max}, \ldots, \frac{|S_c|-1}{|S_c|}P_c^{\max}\}$ where $|S_c| > 1$ for cellular users, and s_{c_i} covers the space between $P_c^{\min}$ and $P_c^{\max}$ with uniform probability. For the C-V2X, the strategy set is $\mathcal{S}_d = \{P_d^{\min}, \frac{1}{|S_d|}P_d^{\max}, \ldots, \frac{|S_d|-1}{|S_d|}P_d^{\max}\}$ where $|S_d| > 1$ for C-V2X with uniform probability.
- *Payoff function:* The energy efficiency of the system is defined as a payoff function (6).

A common method for updating the probability distribution assigned to each player u_{d_i} and u_{c_i} at time t is a Boltzmann-Gibbs probability distribution [46], [47]. It is proportional with the energy of each state and system's temperature. The probability for all players can be expressed as follows:

$$P_{b_i}^{Gibs}(t) = \frac{\exp\Big(\frac{1}{k\tau}\mathsf{EE}\Big)}{\sum_{b\in\mathcal{B}}\mathsf{EE}}, \tag{28}$$

where EE is the energy of the system in state s_t, and a constant $k\tau$ is the product of Boltzmann's constant k and thermodynamic temperature τ. In this regard, if $k\tau \longrightarrow \infty$ there will be a uniform distribution over the strategy set of player b, and if $k\tau \longrightarrow 0$, it causes to select the strategy which is mostly reported by the users [48].

1) NO-REGRET BASED LEARNING ALGORITHM

In a no-regret learning algorithm, players learn their environment to choose transmission power levels along with maximizing the energy efficiency of the system. The regret function is defined as the difference between the average payoff function achieved by strategies of the given algorithm until time t and the payoff function obtained by other fixed sequence of decisions due to a change in strategy [49]:

$$D_{s_{b,i}}(t) = \frac{1}{t}\sum_{\tau<t} u\big(s_{b,i}, s_{-b}(\tau)\big) - \hat{u}_b(\tau), \tag{29}$$

where s_{-b} is the strategy of other players. Given a non-cooperative game $G = (\mathcal{B}, \mathcal{S}_{b,i}, u_b \forall b \in \mathcal{B})$, we can define the correlated strategy $p(s)$ as a probability distribution over the strategy profile $s_i \in \mathcal{S}_b$. Given these basic notions, the concept of a ϵ-coarse correlated equilibrium can be defined as the next theorem.

Theorem 1: Given a game G, a distribution $p(s) = p(s_{b,i}, s_{b,-i})$ is defined as a ϵ-coarse correlated equilibrium if no player can ever expect to unilaterally gain by deviating from their recommendation, assuming the other players follow their recommendations [50], [51]. If $(\forall b_i \in \mathcal{B})$, $(s'_{b,i}, s_{b,i}) \in \mathcal{S}_b$ and $(s_{b,-i}) \in \mathcal{S}_{-b}$ we have

$$\sum p(s_{c,i}, s_{c,-i})\big[u(s'_{c,i}, s_{c,-i}) - u(s_{c,i}, s_{c,-i})\big] < \epsilon, \quad (30)$$

and for D2D pairs,

$$\sum p(s_{d,i}, s_{d,-i})\big[u(s'_{d,i}, s_{d,-i}) - u(s_{d,i}, s_{d,-i})\big] < \epsilon. \quad (31)$$

Players estimate the payoff function concerning the balance between minimizing their regret and the average payoff function for all their strategies. Therefore, for each D2D player and $s_{d,i} \in \mathcal{S}_d$, the payoff estimation function can be calculated by [49], [52]

$$\hat{u}_{d,s_{d,i}}(t+1) = \hat{u}_{d,s_{d,i}}(t) + \left(\frac{1}{t+1}\right)^{\gamma}\big(u_d(t+1) - \hat{u}_{d,s_{d,i}}(t)\big). \quad (32)$$

Similarly for each vehicle and $s_{c,i} \in \mathcal{S}_c$, the payoff estimation function can be calculated by [49], [52]

$$\hat{u}_{c,s_{c,j}}(t+1) = \hat{u}_{c,s_{c,j}}(t) + \left(\frac{1}{t+1}\right)^{\gamma}\big(u_c(t+1) - \hat{u}_{c,s_{c,j}}(t)\big), \quad (33)$$

where $\hat{u}_{d,s_{d,i}}(t+1)$ and $\hat{u}_{c,s_{c,j}}(t+1)$ denote the estimated D2D and cellular payoff function at time t. The strategy played at the last iteration sees the corresponding estimated payoff updated, independently. To calculate the regret, each player needs the learning tool to update the estimated regret [53]. Each D2D player estimates its regret function for each $s_{d,i} \in \mathcal{S}_d$ as follows:

$$\hat{R}_{s_{d,i}}(t+1) = \hat{R}_{s_{d,i}}(t) + \left(\frac{1}{t+1}\right)^{\zeta}\Big(\hat{u}_{d,s_{d,i}}(t+1) - u_d(t+1) - \hat{R}_{s_{d,i}}(t)\Big). \quad (34)$$

Similarly each vehicle estimates its regret for each $s_{c,i} \in \mathcal{S}_c$ as follows:

$$\hat{R}_{s_{c,i}}(t+1) = \hat{R}_{s_{c,i}}(t) + \left(\frac{1}{t+1}\right)^{\zeta}\Big(\hat{u}_{c,s_{c,i}}(t+1) - u_c(t+1) - \hat{R}_{s_{c,i}}(t)\Big). \quad (35)$$

The update probability function assigned to each strategy $s_{d,i} \in \mathcal{S}_d$ of D2D players is described next [49]

$$\pi_{d,s_{d,i}}(t+1) = \pi_{d,s_{d,i}}(t) + \left(\frac{1}{t+1}\right)^{\nu}\big(\pi_d(t+1) - \hat{\pi}_{d,s_{d,i}}(t)\big). \quad (36)$$

Similarly the probability assigned to each strategy $s_{c,i} \in \mathcal{S}_c$ of cellular users is updated as

Algorithm 1 Training Joint Q-Learning Algorithm

Input : $\mathcal{N}$, $u(t)$, $p(u_i|a_j)$, $Q_t^i(s,a)$, r_t^i, $\forall u_i \in \mathcal{U}$, $\pi_{p_l}(t)$, $p_l \in \mathcal{P}_{\mathcal{L}}$
Output : $u(t)$, P_c, P_d, X_d, η_c
Initialiation: $t = 1$, T, $\mathcal{D} = \{1, ..., |\mathcal{D}|\}$, $\mathcal{C} = \{1, ..., |\mathcal{C}|\}$

All agents receive initial observation states $\mathcal{S}_0 = \{s_0^1, ..., s_0^N\}$
while $t \leq T^{\max}$ **do**
  for $\forall d_i \in \mathcal{D} \ \vee \forall c_i \in \mathcal{C}$ **do**
    Select: $p_{d_i}(t), p_{c_j}(t)$ using $\pi_{p_l}(t)$
  end for
  All agents select actions a_t^i according to the current policy
  for $\forall d_i \in \mathcal{D} \ \vee \forall c_j \in \mathcal{C}$ **do**
    Calculate: $\upsilon_{c_j,n}(t), \upsilon_{d_i,n}(t)$ according to (3), (5)
  end for
  Obtain λ
  Calculate: $u(t)$ according to (6)
  if $\lambda > 0$ **then**
    All agents Observe immediate reward r_t^i and next state s_{t+1}
    Update the Q table according to (21)
  end if
  All agents choose actions with maximum Q-value (22)
  if $\sigma(u_i|a_t^i)$ is satisfied according to (27) **then**
    Adjust X_d, η_c according to the optimal action $x_{d_i}^{\mathrm{n}} = 1$, $\eta_{c_j}^{\mathrm{n}} = 1$
    Save $(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$
  end if
  $t = t + 1$,
end while

$$\pi_{c,s_{c,i}}(t+1) = \pi_{c,s_{d,i}}(t) + \left(\frac{1}{t+1}\right)^{\nu}\big(\pi_c(t+1) - \hat{\pi}_{c,s_{c,i}}(t)\big). \quad (37)$$

IV. CONVERGENCE ANALYSIS

In this section, we investigate the convergence of learning algorithms.

A. Q-LEARNING ALGORITHM

For the Q-learning algorithms, $Q_t(s,a)$ converges to an optimal value if the following two conditions are satisfied: (1) the learning rate is suitably reduced to 0; (2) each state-action pair is visited infinitely [8], [54], [55].

Theorem 2: Given a finite MDP model, the Q-learning algorithm, given by the update rule (21), converges to the optimal Q-function if

$$\sum_t \alpha_t(s_t, a_t) \to +\infty, \quad (38)$$

$$\sum_t \alpha_t^2(s_t, a_t) < +\infty. \quad (39)$$

Theorem 3: In the proposed Q-learning methods, each agent i takes an action, $a_i \in \mathcal{A}$ with a decision policy

π_i. Since the learning rate, $0 < \alpha_t(s_t, a_t) < 1$, and all state-actions of the users could be visited infinitely in (21), Algorithms 1 and 2 converge to a fixed point.

B. NO-REGRET LEARNING ALGORITHM

The no-regret learning algorithm is based on stochastic approximation theory and uses a Boltzmann-Gibbs distribution to allocate the initial transmit power. For the convergence of the mechanism, the set of $\iota = \{\gamma, \zeta, \nu\}$ should satisfy the following conditions [56], [57]:

$$\lim_{t\to+\infty} \sum_{n=1}^{t} \frac{1}{n^{\iota}} \to +\infty, \tag{40}$$

$$\lim_{t\to+\infty} \sum_{n=1}^{t} \left(\frac{1}{n^{\iota}}\right)^2 < +\infty. \tag{41}$$

Accordingly, the learning rates should be large enough to overcome any undesirable conditions and small enough to guarantee the convergence of no-regret algorithm. We should choose all $\iota = \{\gamma, \zeta, \nu\} \in (0.5, 1)$ and follow $\zeta > \gamma$, $\nu > \zeta$. To this end, the strategies converge if the learning rate exponents satisfy the following criteria

$$\lim_{t\to+\infty} \frac{\left(\frac{1}{t^{\zeta}}\right)}{\left(\frac{1}{t^{\gamma}}\right)} = 0, \tag{42}$$

$$\lim_{t\to+\infty} \frac{\left(\frac{1}{t^{\nu}}\right)}{\left(\frac{1}{t^{\zeta}}\right)} = 0. \tag{43}$$

To obtain an optimal result, the convergence of the utility function and stopping criteria should be verified.

V. COMPUTATIONAL COMPLEXITY ANALYSIS

In each iteration, the computational complexity depends on the number of subcarriers *(N)* and the number of vehicle and D2D pairs *(M+K)*. Furthermore, the overall complexity depends on the number of iterations *(T)* needed for convergence. Here, we calculate the complexity of each proposed algorithm.

A. SUBCARRIER ALLOCATION

The complexity of the exhaustive search algorithm for the subcarrier allocation sub-problem can be calculated as follows:

$$O(\Lambda_1) = O\left(C^{M+K}_{N(M+K)}\right), \tag{44}$$

which denotes all the probable combinations of selecting $(M + K)$ states from $N(M + K)$ existing states.

For the Q-learning algorithm, there are $N(M + K)$ states, and the complexity can be represented in the following way:

$$O(\Lambda_2) = O(TN(M + K)). \tag{45}$$

Algorithm 2 Training Subcarrier Allocation Q-Learning

Input : $\mathcal{N}$, $p(u_i|a_j)$, $Q^i_t(s, a)$, r^i_t, $\forall u_i \in \mathcal{U}$
Output : X_d, η_c
Initialiation: $t = 1$, T, $\mathcal{D} = \{1, ..., |\mathcal{D}|\}$, $\mathcal{C} = \{1, ..., |\mathcal{C}|\}, X_d, \eta_c$

for $\forall d_i \in \mathcal{D} \ \vee \ \forall c_j \in \mathcal{C}$ **do**
  Select a initial state s_0 randomly
  while $t \leq T^{\max}$ **do**
    Select an action a_t based on strategy
    Calculation: $\upsilon_{c_j,n}(t)$, $\upsilon_{d_i,n}(t)$ according to (3), (5)
    Observe λ
    if $\lambda\sigma(u_i|a_t) > 0$ **then**
      Obtain immediate reward r^i_t and next state s_{t+1}
      Update the Q table according to (21)
    end if
    Choose the action for the user u_i with maximum Q-value (22)
    Adjust X_d, η_c according to the optimal action $x^{\mathrm{n}}_{d_i} = 1$, $\eta^{\mathrm{n}}_{c_j} = 1$
    $t = t + 1$,
  end while
end for

B. POWER ALLOCATION

For the no-regret learning algorithm, in each iteration, subcarrier allocation matrixes are updated with the complexity of $O(TN(M + K))$ and then each vehicle and D2D player chooses a power level with the complexity of $O(1)$. Players calculate their SINR and check the QoS with the complexity of $O(N(M+K))$, and the learning functions are updated with the complexity of $O(N)$. To this end, the whole complexity can be represented as follows:

$$O(\Lambda_3) = O(T(M + K)(TN + M + K + 2)). \tag{46}$$

C. MULTI-AGENT JOINT POWER AND SUBCARRIER ALLOCATION

In this mechanism, all the agents take the actions with a maximum Q-value according to the optimal policy. Hence, the corresponding space complexity is reduced, and it can be written as

$$O(\Lambda_4) = O(T(M + K)(N + M + K + 3)). \tag{47}$$

The above analysis provides the computational complexity for the proposed algorithms [58], [59]. We can observe a trade-off between the performance and convergence speed of the proposed algorithms. The results are shown in Table 3 and Table 4.

VI. SIMULATION RESULTS

We consider a single-cell scenario, where D2D pairs and vehicles are uniformly distributed over an area of $500 \times 500m^2$ with the BS located in the center of the C-V2X environment. We consider a fixed number of vehicles and D2D pairs determined according to the closest distance. When

TABLE 3. The computational complexity and performance of the subcarrier allocation schemes.

	Exhaustive	Q-learning	Exhaustive over the Q-learning
Computational Complexity	$O(\Lambda_1)$	$O(\Lambda_2)$	60% ↑
Performance (Energy Efficiency (Mbps/Watt))	2.1	1.55	14.5% ↑

TABLE 4. The computational complexity and performance of the power allocation schemes.

	Multi joint Q-learning	No-regret	Multi joint Q-learning over the no-regret
Computational Complexity	$O(\Lambda_3)$	$O(\Lambda_4)$	83% ↓
Performance (Energy Efficiency (Mbps/Watt))	1.8	1.5	9% ↑

Algorithm 3 No-Regret Power Allocation Algorithm

Input : $\mathcal{N}$, $u_{s_{d_i},n}(t)$, $R_{s_{d_i},n}(t)$, $\pi_{s_{d_i},n}(t)$, $\forall d_i \in \mathcal{D}$ and $s_{d_i} \in \mathcal{S}_d$, $\hat{u}_{s_{c_j},n}(t)$, $R_{s_{c_j},n}(t)$, $\pi_{s_{c_j},n}(t)$, $\forall c_j \in \mathcal{C}$ and $s_{c_j} \in \mathcal{S}_c$

Output : $u(t)$, $u_{s_{d_i},n}(t+1)$, $R_{s_{d_i},n}(t+1)$, $\pi_{s_{d_i},n}(t+1)$, $\forall d_i \in \mathcal{D}$ and $s_{d,i} \in \mathcal{S}_d$, $u_{s_{c_j},n}(t+1)$, $R_{s_{c_j},n}(t+1)$, $\pi_{s_{c_j},n}(t+1)$, $\forall c_j \in \mathcal{C}$ and $s_{c,j} \in \mathcal{S}_c$

Initialiation: $t = 1, \mathcal{D} = \{1, ..., |\mathcal{D}|\}$, $\mathcal{C} = \{1, ..., |\mathcal{C}|\}$

while $t \leq T^{\max}$ **do**
  Update: X_d, η_c
  for $\forall d_i \in \mathcal{D} \ \vee \ \forall c_i \in \mathcal{C}$ **do**
    Select: $p_{d_i,n}(t)$ using $\pi_{s_{d_i},n}(t)$
    Select: $p_{c_j,n}(t)$ using $\pi_{s_{c_j},n}(t)$
  end for
  for $\forall d_i \in \mathcal{D} \ \vee \ \forall c_j \in \mathcal{C}$ **do**
    Calculate: $\upsilon_{c_j,n}(t)$, $\upsilon_{d_i,n}(t)$ according to (3), (5)
  end for
  if $\left(\upsilon_{c,n}(t) > \gamma_c\right) \wedge \left(\upsilon_{d,n} > \gamma_d\right)$ **then**
    Calculate: $u(t)$ according to (6)
  end if
  for $\forall c_j \in \mathcal{C}$ **do**
    Update: $u_{s_{c_j},n}(t+1)$,$R_{s_{c_j},n}(t+1)$,$\pi_{s_{c_j},n}(t+1)$ according to (33), (35), (37)
  end for
  for $\forall d_i \in \mathcal{D}$ **do**
    Update: $u_{s_{d_i},n}(t+1)$,$R_{s_{d_i},n}(t+1)$,$\pi_{s_{d_i},n}(t+1)$ according to (32), (34), (36)
  end for
  $t = t + 1$,
end while

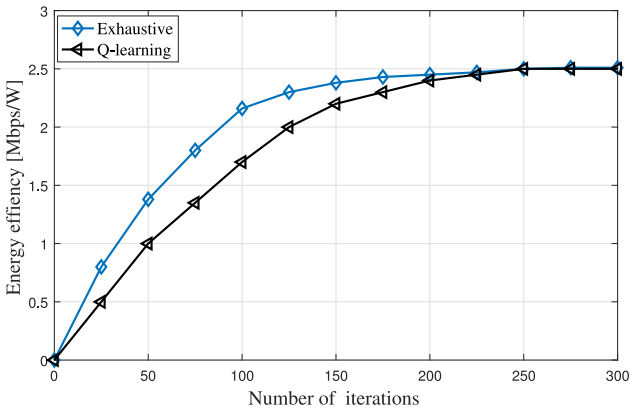


FIGURE 2. Average energy efficiency of the system versus number of iterations.

TABLE 5. Simulation parameters.

Parameter	Value
Physical link type	uplink
Noise power spectral density	-174 dBm/Hz
Rayleigh fading channel(h)	CN(0,1)
Learning rate exponent $\gamma, \zeta, \nu, \alpha$	0.6, 0.7, 0.8, 0.9
Cell radius	500 m
Reward discount β	0.98
The training iteration period	4000
Cellular transmit power	25 dBm
D2D transmit power	20 dBm
Pathloss exponent	2
Antenna gain of device	0.5 dBi
Antenna gain of BS	14 dBi

two D2D users are physically close, a Rayleigh C-V2X communication channel is established. For a fixed number of vehicles and D2D pairs, we ran 500 independent simulations, and we present the average of these results. The pathloss model and shadow fading were considered for C-V2X links, and we set the pathloss exponent in a free space propagation model to be 2. Furthermore, we vary the number of vehicles and D2D pairs, and observe the performance of the system. The simulation parameters are summarized in Table 5.

In Figs. 2–4, we investigate our proposed disjoint approach for allocating the subcarriers to each user by varying the number of subcarriers. However, to evaluate the results of the Q-learning method for allocating the subcarriers, we utilize the Exhaustive search method for finding the optimal subcarriers and comparing the results with each other.

In Fig. 2, the proposed Q-learning algorithm for subcarrier allocation brought about a convergence approximately as fast as the exhaustive search method for subcarrier allocation. We noted only a 14.5% difference between the two algorithms in term of energy efficiency to achieve the same converge point, while Q-learning algorithm implies a much lower complexity than the exhaustive search method.

In Figs. 3 and 4, we varied the number of subcarriers to demonstrate the impact of this on the performance of our proposed Q-learning algorithm. We set the number of D2D pairs and cellular users to be 10 and 5, respectively. As we can see in Fig. 3, varying the number of subcarriers from 5 to 12 brings about a significant performance gain, due to

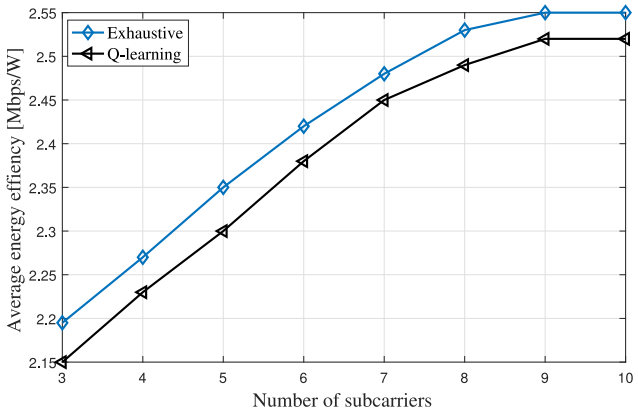


FIGURE 3. Average energy efficiency versus number of subcarriers.

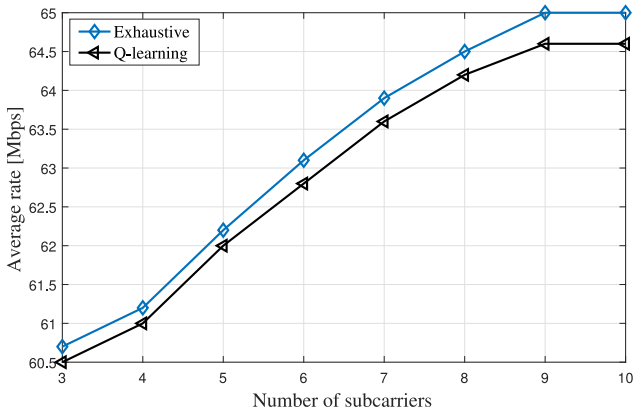


FIGURE 4. Average system rate versus number of subcarriers.

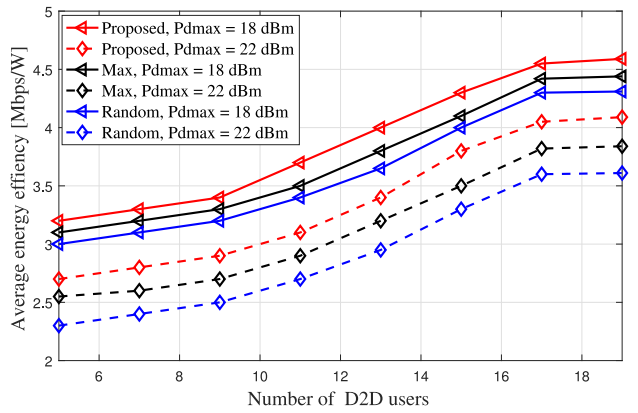


FIGURE 5. Average energy efficiency versus number of D2D pairs.

increasing allocated subcarriers to the users. Adopting the proposed Q-learning algorithm for allocating the subcarriers results in much better performance for the cellular and D2D links. The proposed Q-learning approach can gain the value as well as the exhaustive search method with only a 9% difference in average energy efficiency.

In Fig. 4, we can see that an increase in the number of subcarriers results in an increase in the spectrum available for users and a decrease in the interference among users in the system, which in turn lead to increase in the data rate of

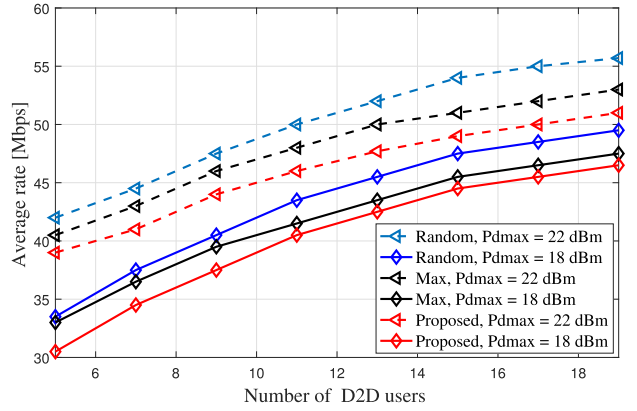


FIGURE 6. Average system rate versus number of D2D pairs.

the system. There is only a difference about 13% compared with the exhaustive search results.

In Figs. 5–7, we show how the performance of the proposed no-regret learning algorithm for power allocation in the non-cooperative game achieves better performance. For the sake of simplicity, we set the number of subcarriers and vehicles to 10 and 5, and we vary the number of D2D pairs from 5 to 19. However, we determine the number of users and subcarriers as variable parameters in the proposed algorithms and they could be assigned a large number. Furthermore, we compare our proposed self-organizing mechanism with three following benchmark references:

- D2D pairs and vehicles choose their transmit power level according to the roulette wheel method; it is labeled as (Proposed, Pdmax = 18 dBm) and (Proposed, Pdmax = 22 dBm).
- D2D pairs and vehicles choose maximum transmit power threshold level; it is labeled as (Max, Pdmax = 18 dBm) and (Max, Pdmax = 22 dBm).
- D2D pairs and vehicles choose random transmit power level; it is labeled as (Random, Pdmax = 18 dBm) and (Random, Pdmax = 22 dBm).

Fig. 5 shows the average utilities achieved by different methods which is increased by varying the number of D2D pairs. However, the proposed method using the Boltzmann-Gibbs distribution to assign the probability to each subcarrier indicates the higher value compared to the algorithm using the two other methods for selecting the power level. Since the proposed method using the Boltzmann-Gibbs distribution, is based on probability law, estimates power level with specific probability distribution and causes a noticeable change in the system. Moreover, by increasing the power threshold level (Pdmax) from 18 dBm to 22 dBm, the average energy efficiency of the system decreased. This was due to the fact that increasing the power level may lead to an increase in energy consumption and result in a decrease in energy efficiency.

As we can see in Fig. 6, the average data rate of the system achieved with these methods increased by varying the number of D2D pairs. Furthermore, by using the roulette wheel

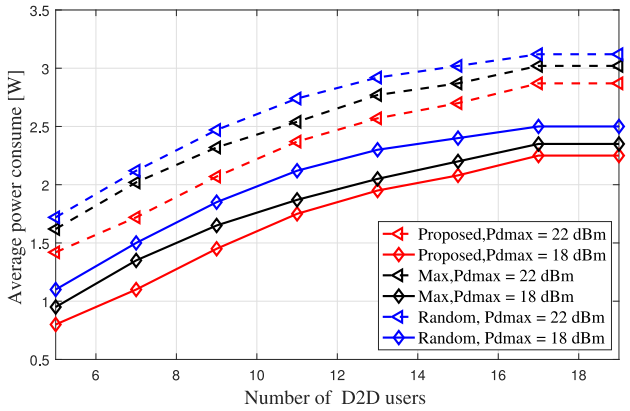


FIGURE 7. Average power consumption versus number of D2D pairs.

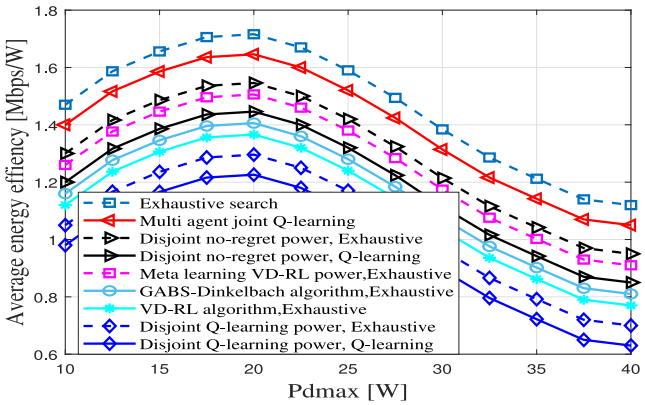


FIGURE 8. Average energy efficiency of the system versus maximum power threshold.

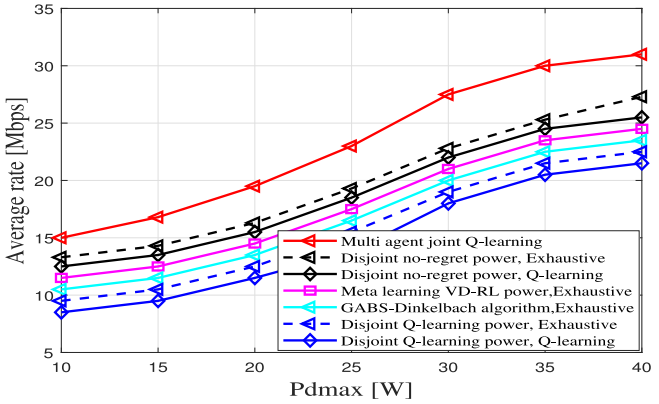


FIGURE 9. Average system rate versus maximum power threshold.

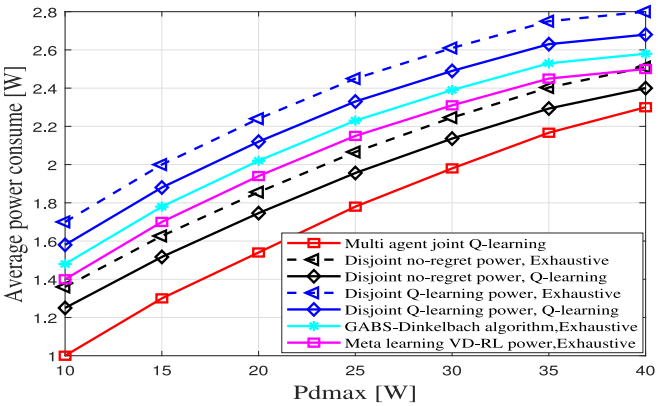


FIGURE 10. Average power consumption versus maximum power threshold.

method, which is based on the probability distribution law, the Nash equilibrium is reached faster than with the other methods in the simulation. Simulation results show that the first algorithm using the roulette wheel method can attain data rates respectively 3% and 5% higher than the maximum and random power levels. Furthermore, it can be observed increasing the power threshold level (Pdmax) results in an increase in the average system sum rate. This is due to strong management of interference among the users. Since the proposed mechanism performs well at a power threshold of 22 dBm, it yields higher average result about 32% compared with the result at a power threshold of 18 dBm.

Fig. 7 shows the power consumption of the system achieved by these three methods. Using the roulette wheel method for selecting the transmit power level results in a faster convergence, and consumes less energy than the other two methods that use the maximum and random power levels. In addition, increasing the power threshold level (Pdmax) from 18 dBm to 22 dBm increases the average energy consumption of the system. This is due to the fact that the number of the strategies in the game increases, which may lead to greater competition among users to achieve an optimal power level, thereby using more energy.

In Figs. 8–10, we show the performance of the our proposed two multi-agent joint Q-learning and disjoint Q-learning algorithm compared with each other. To evaluate the performance of our proposed joint and disjoint algorithms, we use the Q-learning method adopted from the [8], GABS-Dinkelbach algorithm adopted from the [30], VD-RL algorithm and Meta training mechanism with VD-RL algorithm in [31]. Moreover, to evaluate the optimality of the proposed methods, the results would be compared with exhaustive search method for allocating the optimal subcarriers and powers to the users. Results show that increasing the power threshold levels from 10 dBm to 40 dBm brings about a significant performance; however, increasing the power threshold beyond 40 dBm only achieves marginal benefits in the above algorithms. We compare our proposed algorithms with following benchmark references:

- *Multi-agent joint Q-learning*. This algorithm is executed to allocate the joint power and subcarriers.
- *No-regret disjoint algorithm*. This algorithm is proposed for power allocation. If the Q-learning algorithm is implemented for allocating the subcarriers, it is labeled as (Disjoint no-regret power, Q-learning). If an exhaustive search method is implemented for subcarrier allocation, it is labeled as (Disjoint no-regret power, Exhaustive).

- *Q-learning disjoint algorithm*. This algorithm is developed in [8], and used for power allocation. If a Q-learning algorithm is implemented for allocating the subcarriers, it is labeled as (Disjoint Q-learning power, Q-learning). If the exhaustive search method is implemented for subcarrier allocation, it is labeled as (Disjoint Q-learning power, Exhaustive).
- *Disjoint GABS-Dinkelbach algorithm*. This algorithm is developed in [30], and used for power allocation. Moreover, the exhaustive search method is implemented for subcarrier allocation, which is labeled as (Disjoint GABS-Dinkelbach power, Exhaustive).
- *Disjoint VD-RL algorithm*. This algorithm is used in [31] for power allocation, and exhaustive search method is implemented for subcarrier allocation, which is labeled as (Disjoint VD-RL power, Exhaustive).
- *Disjoint meta learning VD-RL power*. This algorithm is developed in [31], and used for power allocation. Moreover, the exhaustive search method is implemented for subcarrier allocation, which is labeled as (Disjoint meta learning VD-RL power, Exhaustive).
- *Exhaustive search algorithm*. This algorithm is executed to allocate the joint power and subcarriers.

We evaluate the performance of our proposed algorithms in terms of different power levels.

Fig. 8 shows the average energy efficiency of the system. Increasing the power threshold puts the system within a maximum value range of 18-20 dBm, while increasing the power threshold beyond the 20dBm, enhance the right to choose the transmit power strategy and lead to consume more energy. Thus, it drops down slowly. The proposed multi-agent joint Q-learning algorithm converges to an optimal point faster than other disjoint algorithms. This is due to the simultaneous allocation of resources and low complexity. Accordingly the second proposed disjoint algorithm which is involved with the no-regret algorithm for power allocation has the faster convergence rate than the disjoint Q-learning method and GABS-Dinkelbach algorithm, which are taken from other papers. It has a greater influence on the energy efficiency of the system, due to the fact that the no-regret algorithm uses the regret function and the probability-based which increases the convergence rate. The multi-agent joint Q-learning algorithm can yield a higher average energy efficiency, of up to 11%, 14% and 18%, than the proposed disjoint mechanism with no-regret learning, GABS-Dinkelbach algorithm and other Q-learning methods for power allocation, respectively.

The results also show that using the proposed meta training mechanism with VD-RL algorithm in [31], can find optimal solution in an unseen environment with faster convergence speed than VD-RL algorithm. However, there is about 14% differences between the proposed joint Q-learning algorithm and the meta-learning with VD-RL methods. This is because that, joint Q-learning proposed method is competitive and users learn their strategies in a distributed manner without the information of others. Moreover, past data from the meta-training method, can be recycled to adapt the policy on a new task in the proposed joint Q-learning method, which in turn lead to reach more efficient results than the meta-learning method. Therefore, the proposed Q-learning method compares favorably with the state of the art in meta-RL.

Furthermore, in order to evaluate the optimality of the proposed methods, we utilize the Exhaustive search method for finding the optimal convergence point. There is only 8% differences between the joint proposed method and the Exhaustive search method in term of energy efficiency to achieve the same convergence point, while Q-learning algorithm implies a much lower complexity than the exhaustive search method.

Fig. 9 shows the average throughput when the power threshold level increases. Increasing the power threshold causes an increase in the average throughput. The main reason for this, is that the D2D links use the same radio frequency band used by cellular links in the adjacent zones. Therefore, the throughput of the D2D link is affected by the transmission power of the cellular link and the surrounding D2D links. Thus, if the transmission power of the D2D link becomes greater than that of the cellular link, the throughput of the system increases. For instance, the joint multi-agent mechanism yields up to 18%, 26% and 35% improvement in terms of throughput, relative to the proposed disjoint mechanism with no-regret learning, GABS-Dinkelbach algorithm and other Q-learning methods for power allocation, respectively. Furthermore, by increasing the power threshold, the average throughput in the disjoint mechanisms GABS-Dinkelbach algorithm have the almost same performance as no-regret algorithm for allocating the power.

Fig. 10 shows the average power consumption when the power threshold level increases. As the Pdmax increases, energy consumption increases because the interference becomes stronger, and users require more power to meet QoS constraints. The multi-agent joint Q-learning algorithm consumes less energy, about 5.3%, 10.2% and 15.2% compared with the proposed disjoint mechanism with no-regret learning, GABS-Dinkelbach algorithm and other Q-learning methods for power allocation, respectively. This is due to the fact that, allocating the subcarrier and power simultaneously in a distributed manner causes minimal human interference and complexity. Moreover, for a given Pdmax, the second and third disjoint algorithms with the proposed Q-learning method for subcarrier allocation consume less energy compared to the other approaches which involve exhaustive search methods for subcarrier allocation. However, they have almost the same performance.

In Fig. 11, we show the performance of our two proposed methods in term of power consumption; first, multi-agent joint power and subcarrier allocation algorithm and second, the disjoint distributed learning algorithm. Varying the number of subcarriers from 5 to 20 yields a significant performance gains for the joint algorithm due to the more efficient management of interference among users with the Q-learning method. As a matter of fact, There is a gap about

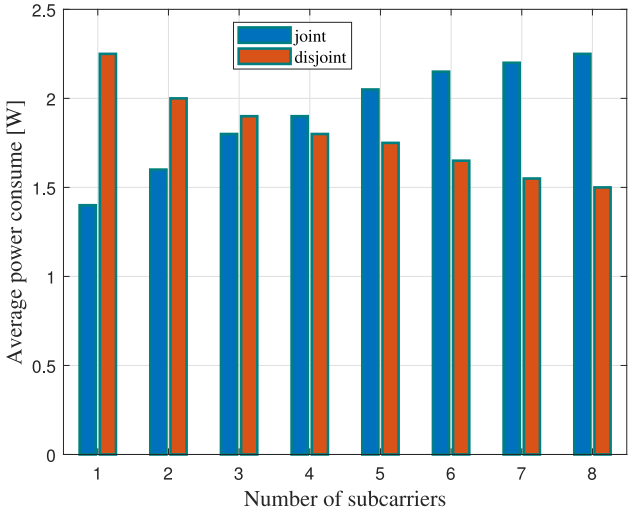


FIGURE 11. Average power consumption versus number of the subcarriers.

12% between the results of the joint and disjoint algorithm in terms of energy efficiency of the system. This is because the feasibility region of finding the optimal value of variables in the joint multi-agent Q-learning method is larger than that of disjoint learning method. Thus, it is reasonable that the joint method gives larger value rather than the disjoint method. Note that the disjoint method searches for the optimal values in the smaller region (because in each subproblem one variable is fixed and the other is optimized), so it gets a lower EE value.

However, the proposed multi-agent joint method has about 16.2% lower complexity compared with the second disjoint Q-learning method, increasing the number of subcarriers beyond 20 caused to increase the memory usage and the complexity of the first joint algorithm about 11% over the second proposed disjoint method.

VII. CONCLUSION AND FUTURE WORK

In this paper, we investigated the resource allocation problem for a C-V2X network to improve the energy efficiency. We proposed two approaches using machine learning. In the first, a multi-agent Q-learning algorithm was applied for the joint power and subcarrier allocation. In the second approach, we broke the problem down into two sub-problems: a power sub-problem and a subcarrier allocation sub-problem. To allocate the subcarrier among users, a distributed Q-learning algorithm was proposed. Then, given optimal subcarrier allocation, we modeled the power allocation sub-problem as a non-cooperative game. To solve the game, an algorithm was used, which could be executed in a distributed manner. Moreover, we compared the results with a third Q-learning algorithm for power allocation. Simulation results showed that the multi-agent joint Q-learning approach yielded significant performance gains of about 36% and 27% in terms of energy efficiency and sum rate over other disjoint learning algorithms. In addition, our no-regret based learning approach for power allocation was shown to provide better performance, of about 14% and 16% compared with a disjoint benchmark algorithm which utilizes a Q-learning algorithm for power allocation, in terms of the average energy efficiency and average throughput. In the future work, it is interesting to consider multi-base stations, which causes to increase the interferences produced in the system, and try to optimize the resource allocation in the system.

APPENDIX

A. PROOF OF THEOREM 1

$$\lim_{\theta\to 0}\sup\Big(\sum p(s_{c,i}, s_{c,-i})\big[u(s'_{c,i}, s_{c,-i}) - u(s_{c,i}, s_{c,-i})\big] = \frac{1}{|\mathcal{S}_b^*|}\sum\big[u(s'_{c,i}, s_{c,-i}) - u(s_{c,i}, s_{c,-i})\big] > 0.$$

Now we prove (30):

$$\begin{aligned}\sum p(s)\big[u(s'_{c,i}, s_{c,-i}) - u(s_{c,i}, s_{c,-i})\big] &= \frac{|\mathcal{S}_c{}^*|\Big[u\Big(s'_{c,i}, s_{c,-i}\Big) - u(s_{c,i}, s_{c,-i})\Big]}{|\mathcal{S}_c{}^*|} \\ &+ \frac{\sum\exp\Big(\frac{1}{\theta}y_s\Big)\Big[u\Big(s'_{c,i}, s_{c,-i}\Big) - u(s_{c,i}, s_{c,-i})\Big]}{|\mathcal{S}_c{}^*|} \\ &\le \sum\exp\left(\frac{1}{\theta}y_s\right) \le \varepsilon.\end{aligned}$$

Similarly (31) is proved for D2D pairs [50]. ■

B. PROOF OF THEOREM 2

We rewrite (21) as

$$\begin{aligned}Q_{t+1}(s_t, a_t) &= (1-\alpha_t(s_t, a_t))Q_t(s_t, a_t) \\ &\quad + \alpha_t(s_t, a_t)[r_t + \beta\max Q_t(s_{t+1}, b)]. \\ \triangle_t(s_t, a_t) &= (1-\alpha_t(s_t, a_t))\ \triangle_t\ (s_t, a_t)) \\ &\quad + \alpha_t(s, a)\big[r_t + \beta\max Q_t(s_{t+1}, b) - Q_t^*(s_t, a_t)\big].\end{aligned}$$

If we define

$$F_t(s, a) = r(x, a, X(s, a)) + \beta\max Q_t(s_{t+1}, b) - Q^*(s_t, a_t),$$

we have

$$E[F_t(s, a)|F_t] = \sum_{n\in\mathcal{N}} P_a(x, X(s, a))\big[r(x, a, X(s, a)) \beta\max Q_t(s_{t+1}, b) - Q^*(s_t, a_t)\big].$$

Since $Q^* = HQ^*$, we can write

$$\begin{aligned}E[F_t(s, a)|F_t] &= HQ_t(s, a) - HQ^*(s, a). \\ \|E[F_t(s, a)|F_t]\|_\infty &< \beta\|Q_t - Q^*\|_\infty = \gamma\|\ \triangle_t\ \|_\infty. \\ \mathrm{var}[\mathrm{F_t(s)|F_t}] &= E\Big[(r(s, a, X(s, a)) + \beta\max Q_t(y, b) \\ &\quad -Q^*(s, a) - (HQ_t)(s, a) + Q^*(s, a))^2\Big] \\ &= \mathrm{var}\big[(\mathrm{r(s,\ a,\ X(s,\ a))} + \beta\max \mathrm{Q_t(y,\ b)|F_t}\big]\end{aligned}$$

Since r is bounded, this indicates

$$\mathrm{var}[\mathrm{F_t(s)|F_t}] \le \mathrm{C}\Big(1 + \|\ \triangle_t\ \|_\infty^2\Big)$$

for a given constant C [54], [60]. ■

C. PROOF OF THEOREM 3

Algorithm (1) solves (16) by alternating maximum Q-value and calculating the energy efficiency of the system. Since maximum reward function maximizes the Q-function, we want to show that reward function in Algorithm (1) does not increase the objective value of (16). According to line (16) of Algorithm (1), computational resource allocation does not increase the objective value of (16). In addition, based on (38) and (39), convergence of Algorithm (1) is guaranteed. In i^{th} iteration of algorithm (1), energy efficiency of the system depends on the numbers of users and their power levels. As a matter of fact, it would be equal to EE_k for cellular and D2D users when the numbers of users are larger than their maximum acceptable value. Therefore, we have EE_k^+, and need to show that EE_k, does not increase after i^{th} iteration. If $\mathsf{EE}_k = \max_{\mathbf{N}} \mathsf{EE}$ after i iterations, varying the number of users more than N_i caused to increase the power consumption and decrease the energy efficiency of the system. Thus, EE_{max} does not increase more than EE_k when increasing the number of users in other iterations.

$$\mathsf{EE}^{i+1} \leq \mathsf{EE}_k^i. \tag{48}$$

Moreover, the learning rate is suitably reduced to 0, which is vital for convergence of the algorithm (1). As a result, the objective value of (16) is non-increasing in each iteration, and since it is lower bounded by zero, Algorithm (1) is convergent. ■

REFERENCES

[1] Y. Chen, B. Ai, Y. Niu, K. Guan, and Z. Han, "Resource allocation for device-to-device communications underlaying heterogeneous cellular networks using coalitional games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4163–4176, Jun. 2018.

[2] J. Moysen and L. Giupponi, "From 4G to 5G: Self-organized network management meets machine learning," *Comput. Commun.*, vol. 129, pp. 248–268, Sep. 2018.

[3] "3rd generation partnership project; technical specification group services and system aspects; study on 5G smart energy and infrastructure, version 18.0.1," 3rd Gener. Partnership Project (3GPP), Sophia Antipolis, France, Rep. 22.867, Jun. 2021.

[4] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[5] S. Alemaishat, O. A. Saraereh, I. Khan, and B. J. Choi, "An efficient resource allocation algorithm for D2D communications based on NOMA," *IEEE Access*, vol. 7, pp. 120238–120247, 2019.

[6] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, "A survey of machine learning techniques applied to self-organizing cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2392–2431, 4th Quart., 2017.

[7] M. H. C. Garcia *et al.*, "A tutorial on 5G NR V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1972–2026, 3rd Quart., 2021.

[8] Y. Luo, Z. Shi, X. Zhou, Q. Liu, and Q. Yi, "Dynamic resource allocations based on Q-learning for D2D communication in cellular networks," in *Proc. IEEE Int. Wavelet Active Media Technol. Inf. Process. (ICCWAMTIP)*, 2014, pp. 385–388.

[9] Y. Li, D. Jin, J. Yuan, and Z. Han, "Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3965–3977, Jul. 2014.

[10] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and M.-N. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 100480–100490, 2019.

[11] N. Ul Hasan, W. Ejaz, A. Shahid, I. Baig, and M. Zghaibeh, "Self-organized energy efficient channel assignment for cognitive D2D communication in 5G networks," in *Proc. IEEE 5th Int. Conf. Electr. Eng. (ICEE)*, 2018, pp. 404–407.

[12] H. Dai, Y. Huang, R. Zhao, J. Wang, and L. Yang, "Resource optimization for device-to-device and small cell uplink communications underlaying cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1187–1201, Feb. 2018.

[13] S. Driouech, E. Sabir, and H. Tembine, "Self-organized device-to-device communications as a non-cooperative quitting game," in *Proc. IEEE Int. Wireless. Netw. Mobile Commun. Conf. (WINCOM)*, 2017, pp. 1–8.

[14] J. Huang, C.-C. Xing, Y. Qian, and Z. J. Haas, "Resource allocation for multicell device-to-device communications underlaying 5G networks: A game-theoretic mechanism with incomplete information," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2557–2570, Mar. 2018.

[15] A. Shahid, K. S. Kim, E. De Poorter, and I. Moerman, "Self-organized energy-efficient cross-layer optimization for device to device communication in heterogeneous cellular networks," *IEEE Access*, vol. 5, pp. 1117–1128, 2017.

[16] S. M. Zafaruddin, I. Bistritz, A. Leshem, and D. Niyato, "Distributed learning for channel allocation over a shared spectrum," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2337–2349, Oct. 2019.

[17] S. Sharma and B. Singh, "Weighted cooperative reinforcement learning-based energy-efficient autonomous resource selection strategy for underlay D2D communication," *IET Commun.*, vol. 13, no. 14, pp. 2078–2087, Aug. 2019.

[18] B. Kaufman and B. Aazhang, "Cellular networks with an overlaid device to device network," in *Proc. IEEE 42nd Int. Asilomar Conf. Signals Syst. Comput. (ASILOMAR)*, Pacific Grove, CA, USA, 2008, pp. 1537–1541.

[19] A. Ramezani-Kebrya, M. Dong, B. Liang, G. Boudreau, and S. H. Seyedmehdi, "Joint power optimization for device-to-device communication in cellular networks with interference control," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5131–5146, Aug. 2017.

[20] Y. Zhao, Y. Li, Y. Cao, T. Jiang, and N. Ge, "Social-aware resource allocation for device-to-device communications underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6621–6634, Dec. 2015.

[21] A. Bazzi, A. O. Berthet, C. Campolo, B. M. Masini, A. Molinaro, and A. Zanella, "On the design of sidelink for cellular V2X: A literature review and outlook for future," *IEEE Access*, vol. 9, pp. 97953–97980, 2021.

[22] E. E. Tsiropoulou, A. Kapoukakis, and S. Papavassiliou, "Energy-efficient subcarrier allocation in SC-FDMA wireless networks based on multilateral model of bargaining," in *Proc. IFIP Netw. Conf.*, 2013, pp. 1–9.

[23] B. McCarthy, A. Burbano-Abril, V. R. Licea, and A. O'Driscoll, "OpenCV2X: Modelling of the V2X cellular sidelink and performance evaluation for aperiodic traffic," 2021, *arXiv:2103.13212*.

[24] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 441–454, Feb. 2015.

[25] H. H. Esmat, M. M. Elmesalawy, and I. I. Ibrahim, "Adaptive resource sharing algorithm for device-to-device communications underlaying cellular networks," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 530–533, Mar. 2016.

[26] P. Oguntunde, O. Odetunmibi, and A. Adejumo, "On the sum of exponentially distributed random variables: A convolution approach," *Eur. J. Stat. Probab.*, vol. 2, no. 1, pp. 1–8, Mar. 2014.

[27] M. Jung, K. Hwang, and S. Choi, "Joint mode selection and power allocation scheme for power-efficient device-to-device (D2D) communication," in *Proc. IEEE 75th Int. Veh. Technol. Conf. (VTC Spring)*, 2012, pp. 1–5.

[28] M. Gharbieh, A. Bader, H. ElSawy, H.-C. Yang, M.-S. Alouini, and A. Adinoyi, "Self-organized scheduling request for uplink 5G networks: A D2D clustering approach," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1197–1209, Feb. 2019.

[29] K. P. Sharmila and C. Ramesh, "Analyzing performance and QoS parameter estimation for VANET using D2D," in *Innovations in Electronics and Communication Engineering*. Singapore: Springer, 2019, pp. 249–259.

[30] A. Ihsan, W. Chen, S. Zhang, and S. Xu, "Energy-efficient NOMA multicasting system for beyond 5G cellular V2X communications with imperfect CSI," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10721–10735, Aug. 2022.

[31] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3177–3192, Oct. 2021.

[32] "3rd generation partnership project; technical specification group services and system aspects; release 16 description; summary of rel-16 work items," 3rd Gener. Partnership Project (3GPP), 3GPP Rep. TR 21.916, Dec. 2021.

[33] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[34] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[35] K. Sehla, T. M. T. Nguyen, G. Pujolle, and P. B. Velloso, "Resource allocation modes in C-V2X: From LTE-V2X to 5G-V2X," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8291–8314, Jun. 2022.

[36] L. Lai, H. El Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *IEEE Trans. Mobile Comput.*, vol. 10, no. 2, pp. 239–253, Feb. 2011.

[37] M. Simsek, M. Bennis, and I. Güvenç, "Learning based frequency- and time-domain inter-cell interference coordination in HetNets," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4589–4602, Oct. 2015.

[38] G. Zhao, Y. Li, C. Xu, Z. Han, Y. Xing, and S. Yu, "Joint power control and channel allocation for interference mitigation based on reinforcement learning," *IEEE Access*, vol. 7, pp. 177254–177265, 2019.

[39] L. R. Faganello, R. Kunst, C. B. Both, L. Z. Granville, and J. Rochol, "Improving reinforcement learning algorithms for dynamic spectrum allocation in cognitive sensor networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2013, pp. 35–40.

[40] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1828–1840, Feb. 2020.

[41] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[42] U. Challita, L. Dong, and W. Saad, "Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4674–4689, Jul. 2018.

[43] D. P. M. Osorio *et al.*, "Towards 6G-enabled Internet of Vehicles: Security and privacy," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 82–105, 2022.

[44] M. Harounabadi, D. M. Soleymani, S. Bhadauria, M. Leyh, and E. Roth-Mandutz, "V2X in 3GPP standardization: NR sidelink in release-16 and beyond," *IEEE Commun. Stand. Mag.*, vol. 5, no. 1, pp. 12–21, Mar. 2021.

[45] M. Segata, P. Arvani, and R. L. Cigno, "A critical assessment of C-V2X resource allocation scheme for platooning applications," in *Proc. IEEE 16th Annu. Conf. Wireless On-Demand Netw. Syst. Serv. Conf. (WONS)*, 2021, pp. 1–8.

[46] O. Morgenstern and J. Von Neumann, *Theory of Games and Economic Behavior*. Princeton, NJ, USA: Princeton Univ. Press, 1953.

[47] J. R. Marden and J. S. Shamma, "Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation," *Games Econ. Behav.*, vol. 75, no. 2, pp. 788–808, Jul. 2012.

[48] M. Derakhshani and T. Le-Ngoc, "Distributed learning-based spectrum allocation with noisy observations in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3715–3725, Oct. 2014.

[49] A. H. Arani, A. Mehbodniya, M. J. Omidi, F. Adachi, W. Saad, and I. Güvenç, "Distributed learning for energy-efficient resource management in self-organizing heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9287–9303, Oct. 2017.

[50] C. A. Marks, "No-regret learning and game-theoretic equilibria," Ph.D. dissertation, Dept. Comput. Sci., Brown Univ., Providence, RI, USA, 2008.

[51] N. D. Stein, P. A. Parrilo, and A. Ozdaglar, "Correlated equilibria in continuous games: Characterization and computation," *Games Econ. Behav.*, vol. 71, no. 2, pp. 436–455, Mar. 2011.

[52] A. H. Arani, A. Mehbodniya, M. J. Omidi, and F. Adachi, "Learning-based joint power and channel assignment for hyper dense 5G networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–7.

[53] M. Bennis, S. M. Perlaza, and M. Debbah, "Learning coarse correlated equilibria in two-tier wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 1592–1596.

[54] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1994, pp. 703–710.

[55] F. S. Melo and M. I. Ribeiro, "Convergence of Q-learning with linear function approximation," in *Proc. IEEE Eur. Control Conf. (ECC)*, 2007, pp. 2671–2678.

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

[57] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming: An overview," in *Proc. IEEE 34th Conf. Decis. Control*, vol. 1, 1995, pp. 560–564.

[58] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

[59] J. Zheng, Y. Cai, Y. Liu, Y. Xu, B. Duan, and X. Shen, "Optimal power allocation and user scheduling in multicell networks: Base station cooperation using a game-theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6928–6942, Dec. 2014.

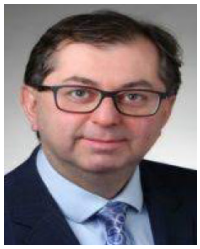
[60] C. Ribeiro and C. Szepesvári, "Q-learning combined with spreading: Convergence and results," in *Proc. ISRF-IEE Int. Conf. Intell. Cogn. Syst.*, 1996, pp. 32–36.

NAJMEH BANITALEBI received the B.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 2016, and the M.Sc. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 2019. She has been currently involved in designing and programming of digital communication systems. Her research interests include design, programming, and optimization of communication networks, wireless communications, resource management, and machine learning.

PAEIZ AZMI (Senior Member, IEEE) was born in Tehran, Iran, in April 17, 1974. He received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Sharif University of Technology, Tehran, in 1996, 1998, and 2002, respectively. Since September 2002, he has been with the Electrical and Computer Engineering Department, Tarbiat Modares University, Tehran, where he became an Associate Professor on January 2006, and he is currently a Full Professor. From 1999 to 2001, he was with the Advanced Communication Science Research Laboratory, Iran Telecommunication Research Center (ITRC), Tehran. From 2002 to 2005, he was with the Signal Processing Research Group, ITRC. His current research interests include modulation and coding techniques, digital signal processing, wireless communications, resource allocation, molecular communications, and estimation and detection theories.

NADER MOKARI (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 2014, where he joined as an Assistant Professor with the Department of Electrical and Computer Engineering in 2015. He has been involved in a number of large scale network design and consulting projects in the telecom industry. His research interests include design, analysis, and optimization of communication networks

ATEFEH HAJIJAMALI ARANI received the Ph.D. degree from the Isfahan University of Technology in 2018. During the Ph.D., she was a Research Fellow with Tohoku University. She is currently a Postdoctoral Fellow with the University of Waterloo and a member of the National Research Council of Canada/UW Collaboration Centre. Her scientific research interests are in the areas of wireless communications, resource management, machine learning, and self-organizing heterogeneous networks.



HALIM YANIKOMEROGLU (Fellow, IEEE) was born in Giresun, Turkey, in 1968. He received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 1990, and the M.A.Sc. degree in electrical engineering and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 1992 and 1998, respectively. From 1993 to 1994, he was with the Research and Development Group of Marconi Kominikasyon A. S., Ankara. Since 1998, he has been with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, where he is currently a Full Professor. He was a Visiting Professor with the TOBB University of Economics and Technology, Ankara, from 2011 to 2012. In recent years, his research has been funded by Huawei, Telus, Allen Vanguard, Blackberry, Samsung, Communications Research Centre of Canada, and DragonWave. This collaborative research resulted in over 25 patents (granted and applied). His research interests include wireless technologies with a special emphasis on wireless networks. He was a recipient of the IEEE Ottawa Section Outstanding Educator Award in 2014, the Carleton University Faculty Graduate Mentoring Award in 2010, the Carleton University Graduate Students Association Excellence Award in Graduate Teaching in 2010, and the Carleton University Research Achievement Award in 2009. He is a Registered Professional Engineer in Ontario, Canada. He has been involved in the organization of the IEEE Wireless Communications and Networking Conference (WCNC) from its inception, including serving as a Steering Committee Member and the Technical Program Chair or the Co-Chair of the WCNC 2004, Atlanta, GA, USA; the WCNC 2008, Las Vegas, NV, USA; and the WCNC 2014, Istanbul, Turkey. He was the General Co-Chair of the IEEE Vehicular Technology Conference (VTC) 2010-Fall held in Ottawa and serving as the General Chair of the IEEE VTC 2017-Fall which will be held in Toronto. He has served on the Editorial Boards of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He was the Chair of the IEEE Wireless Technical Committee. He is a Distinguished Lecturer of the IEEE Communications Society and a Distinguished Speaker of the IEEE Vehicular Technology Society.