A Distributed Approach to Improving Spectral Efficiency in Uplink Device-to-Device-Enabled Cloud Radio Access Networks

Yaohua Sun, *Student Member, IEEE*, Mugen Peng, *Senior Member, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*— Device-to-device (D2D)-enabled cloud radio access networks (C-RANs) are potential solutions for further improving spectral efficiency (SE) and decreasing latency by allowing direct communication between two users. However, due to the need to acquire global channel state information (CSI) and to execute centralized algorithms, heavy burdens are placed on the fronthaul and the baseband unit (BBU) pool. To alleviate these burdens, a distributed approach to mode selection and resource allocation for potential D2D pairs under pre-determined resource allocation of C-RAN users is proposed, in which pairs of users are endowed with decision-making capabilities. The proposed procedure is divided into three stages: communication mode and subchannel selection, utility value determination, and reinforcement-learning-based strategy update. The core idea is that the D2D pairs self-optimize the mode selection and resource allocation without global CSI under several practical constraints. Simulation results show that enabling D2D can significantly improve SE for C-RANs. Furthermore, the impacts of the fronthaul capacity, the centralized signal processing capability of the BBU pool, and the distance between the D2D transmitter and the remote radio head are demonstrated and analyzed.**

***Index Terms*— Cloud radio access networks (C-RANs), device-to-device (D2D), mode selection, resource allocation, game theory.**

I. Introduction

With the explosive increase in mobile data traffic, operators have to continuously improve network performance [1]. As a promising solution, the cloud radio access network (C-RAN) has been proposed to enhance spectral efficiency (SE) and energy efficiency (EE) as well as lower operating and capital expenditures [2], [3]. In C-RANs, the baseband processing and upper-layer functionalities are migrated to the baseband unit (BBU) pool, which connects with distributed remote radio heads (RRHs) via fronthaul. Many studies have been conducted on C-RANs, in terms of computing resource optimization [4], performance analysis [5], system cost minimization [6], and so on.

Since the practical fronthaul is often capacity constrained or time-delay constrained, the C-RAN is far from satisfying the SE and EE requirements of the fifth generation mobile communication network [7]. Hence, advanced techniques can be introduced to further improve the performance of C-RANs. In this paper, device-to-device (D2D) enabled C-RANs are examined, whose core idea is inspired by D2D enabled cellular networks with decreased latency and significant SE/EE gains [8]. In addition, alleviating the burden of fronthaul is another remarkable benefit brought by D2D, since the traffic generated by user equipment (UE) communicating directly does not need to go through fronthaul anymore. In [9], the considerable SE improvement by integrating D2D with C-RANs has been confirmed via stochastic geometry.

To fully boost SE and EE gains, mode selection and resource allocation schemes play key roles. In D2D enabled C-RANs, mode selection refers the situation in which a pair of UEs intending to communicate with each other, termed a potential D2D pair, can select either D2D mode where direct communication is performed or C-RAN mode where the C-RAN helps to transmit the traffic, while resource allocation refers to subchannel allocation among potential D2D pairs. Moreover, appropriate RRH association and D2D power control schemes are needed to meet fronthaul capacity constraints, centralized signal processing capability constraints at the BBU pool, as well as inter-tier interference constraints. In this paper, different from [9] which focuses on performance analysis, the objective of studying the mode selection and resource allocation problem in uplink D2D enabled C-RANs is to optimize the system SE and meanwhile alleviate the burdens of the fronthaul and the BBU pool by making the potential D2D pairs autonomously optimize the mode selection and resource allocation under several practical constraints.

A. Related Work and Challenges

Much attention has been paid to resource allocation in D2D enabled cellular networks recently. In [10], a subchannel sharing protocol is proposed to ensure that the mutual

Manuscript received September 15, 2017; revised February 3, 2018 and June 8, 2018; accepted June 30, 2018. Date of publication July 12, 2018; date of current version December 14, 2018. This work was supported in part by the Chinese State Major Science and Technology Special Project under Grant 2016ZX03001020-006 and 2017ZX03001025-006, in part by the Chinese National Program for Special Support of Eminent Professionals, and in part by the U.S. National Science Foundation under Grants CNS-1702808 and ECCS-1647198. This paper was presented in part at the 2016 IEEE Global Communications Conference. The associate editor coordinating the review of this paper and approving it for publication was J. Choi. *(Corresponding author: Mugen Peng.)*

Y. Sun and M. Peng are with the Key Laboratory of Universal Wireless Communications (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: sunyaohua@bupt.edu.cn; pmg@bupt.edu.cn).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2018.2855212

0090-6778 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

interference of two D2D pairs is negligible when they use the same subchannel. Then, the primal problem aiming at maximizing the sum rate of D2D pairs is further divided into a subchannel assignment problem and a power distribution problem. A minimum weighted EE maximization problem is studied in [11], where the proposed problem is transformed into the joint subchannel and power allocation problem for D2D links.

Furthermore, in [12]–[14], various game models are used to develop resource allocation schemes with low complexity. In [12], focusing on a downlink D2D underlaid cellular network, the spectrum resource allocation for D2D pairs is modeled as a reverse iterative combinatorial auction. The spectrum resources occupied by a cellular user are seen as a bidder who competes for the packages of D2D pairs to maximize the channel rate. In [13], a coalition formation based approach is proposed to address the uplink spectrum sharing for multiple D2D and cellular users, in which a possible coalition structure stands for a possible spectrum sharing relationship. To limit the inter-tier interference, a pricing based D2D power control scheme is designed in [14], and the competition among D2D pairs is modeled as a non-cooperative game.

The aforementioned works typically study resource allocation with all D2D pairs operating in D2D mode. This fixed setting is not beneficial to the overall system performance, and hence mode selection should be taken into account [8]. In [15], a potential D2D transmitter selects the communication mode according to the comparison between the biased D2D link quality and the cellular uplink quality, and a similar method is used in [16]. In [17], a guard zone based mode selection scheme is proposed, where potential D2D transmitters located within the guard zones are required to operate in cellular mode to avoid severe interference to the base stations (BSs).

Actually, resource allocation and mode selection are coupled [18], and thus they need to be optimized jointly. In [19], a joint mode selection and resource allocation problem to maximize the system throughput with a minimum rate guarantee is formulated, and a particle swarm optimization based scheme is proposed, where solutions are mapped onto particles. The joint optimization problem of D2D mode selection, modulation and coding scheme assignment, radio resource allocation, and power allocation is decoupled into two subproblems in [20], and a Tabu Search metaheuristic based approach is used to search for the best mode selection. In [21], three communication modes are considered for D2D users: cellular mode, dedicated mode, and reuse mode, and the system throughput maximization problem is decomposed into a power control problem, and joint mode selection and channel assignment problem which can be solved by the branch-and-bound method. An outer approximation based linearization technique is utilized in [22], to solve the joint admission control, mode selection, and power allocation problem, which is shown to be NP-complete.

Although the centralized methods in [18]–[22] have shown significant performance gains, centralized implementation in D2D enabled C-RANs can put heavy burdens on the fronthaul and the BBU pool due to the need to acquire global channel state information (CSI) and to execute centralized algorithms. Fortunately, the computing capabilities of UEs are increasingly more powerful, and thus it is possible to let potential D2D pairs optimize their communication modes and utilized subchannels autonomously to alleviate the burdens of the C-RAN infrastructures. Note that this idea corresponds with the newly emerging notion of fog computing based radio access networks (F-RANs) [23], and hence our scenario can be seen as a special case of F-RANs. To this end, a distributed approach with low computing burdens as well as low CSI requirement for potential D2D pairs should be developed. Meanwhile, during the self-optimization of D2D pairs, the impacts of limited fronthaul capacity, limited centralized signal processing capability of the BBU pool, and inter-tier interference constraints should be considered as well.

B. Contributions and Organization

In this paper, a distributed approach to mode selection and subchannel allocation for potential D2D pairs under several practical constraints in an uplink D2D enabled C-RAN is proposed, which alleviates the burdens on the fronthaul and the BBU pool. In the proposed approach, D2D pairs are assumed to have decision-making capabilities, which select their communication modes and subchannels according to individual strategies. The strategy profile of all the D2D pairs is ensured to converge to a pure strategy profile by an iterative reinforcement learning process, and the feedback value of utilities is determined by a distributed RRH association process and a distributed power control process where the fronthaul capacity constraints, centralized signal processing capability constraints at the BBU pool, and inter-tier interference constraints are considered. The main contributions of the paper are:

- To improve the SE of C-RANs, the uplink D2D enabled C-RAN is investigated, in which a potential D2D pair can operate either in D2D mode or C-RAN mode. In D2D mode, the two UEs in the same pair communicate directly, while they communicate with the help of the C-RAN in C-RAN mode. Unlike traditional D2D enabled cellular scenarios, the proposed model has unique features such as capacity-constrained fronthaul and uplink coordinated multipoint transmission. Under the pre-determined resource allocation of C-RAN UEs, an optimization problem aiming at maximizing the overall SE is formulated. Specifically, the overall SE is not only related to the communication mode selection and resource allocation of D2D pairs, but also depends on the RRH association process and power control process, where the fronthaul capacity constraints, centralized signal processing capability constraints at the BBU pool, and inter-tier interference constraints are taken into account.
- The formulated problem is handled by a distributed approach with three stages. In the action selection stage, each D2D pair first selects a communication mode and a subchannel according to its individual strategy. Next, in the utility value determination stage, the RRH association of D2D pairs in C-RAN mode is modeled as a many-to-many matching game, and the power control of D2D pairs in D2D mode is modeled as a Stackelberg game. In the strategy update stage, each pair updates its strategy

TABLE I
SUMMARY OF NOTATION

$\mathcal{N}$	The set of potential D2D pairs
$Tx(n)$	The transmitter of the potential D2D pair n
$Rx(n)$	The receiver of the potential D2D pair n
$\mathcal{K}$	The set of RRHs
$\mathcal{C}$	The set of CUEs that are incapable of conducting D2D communication
$x_n^{(D)}$, $x_n^{(C)}$	Mode selection indicators for the potential D2D pair n
$\mathcal{M}$	The set of all transmitting UEs including both transmitters of potential D2D pairs and CUEs
$\mathcal{D}$	The set of available subchannels
$p_{n,d}$	The transmit power of $Tx(n)$ on subchannel d
$h_{n,d}^{(D)}$	The channel coefficient between $Tx(n)$ and $Rx(n)$ on subchannel d
$h_{m,n,d}^{(D)}$	The channel coefficient between transmitting UE m and $Rx(n)$ on subchannel d
$h_{n,k,d}^{(C)}$	The channel coefficient between $Tx(n)$ and RRH k on subchannel d
$\mathbf{h}_{n,d}^{(C)}$	The channel vector from $Tx(n)$ to its associated RRHs on subchannel d
V	The maximum number of RRHs with which a UE can associate
F_k	The maximum number of UEs that the fronthaul between RRH k and the BBU pool can still serve given the RRH association of CUEs
p_{max}	The maximum transmit power of a transmitting UE
$a_{n,i}$	The i-th available action of decision-maker n
$\pi_{n,a_{n,i}}$	The probability that decision-maker n selects action $a_{n,i}$

using a reinforcement learning process. These three stages are repeated until a pure strategy profile of D2D pairs is reached, which is desired in practical implementation. The nature of the proposal is that D2D pairs self-optimize the mode selection and resource allocation under several practical constraints such as fronthaul capacity constraints.

- The properties of the proposed approach are discussed, in terms of optimality, stability, etc. Simulation results show that the proposed method converges and the benefits of enabling D2D in C-RANs are significant. Furthermore, the reasons why D2D can improve SE in C-RANs are analyzed.

The remainder of this paper is organized as follows. Section II describes the uplink D2D enabled C-RAN model. Section III proposes the optimization problem of interest, and presents a distributed approach to solve it. The utility value determination stage is specified in Section IV, and Section V discusses the properties of the proposed approach. Simulation results are presented in Section VI, followed by the conclusion in Section VII. For convenience, some important notation is listed in Table I.

II. System Model

The considered system model shown in Fig. 1 consists of one BBU pool, one high power node (HPN), multiple RRHs, multiple potential D2D pairs, and multiple C-RAN UEs (CUEs) that are incapable of conducting D2D communication. The HPN is responsible for delivering system information to UEs, and can communicate with the BBU pool through the backhaul [24]. Assume that the system operates in Time-Division Duplex mode, and RRHs connect with the BBU pool via capacity-constrained fronthaul. The set of potential

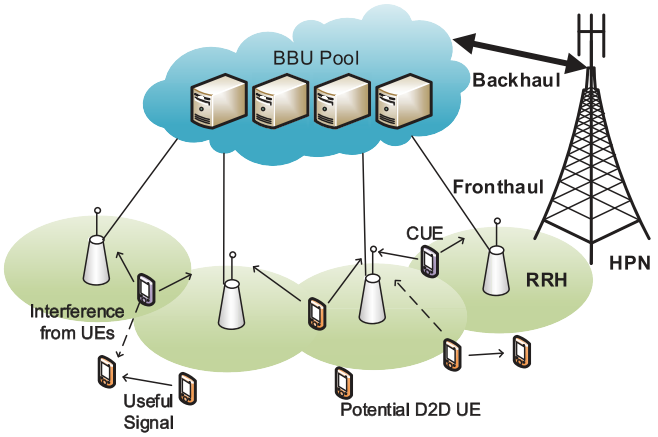


Fig. 1. An uplink D2D enabled C-RAN.

D2D pairs is denoted by $\mathcal{N} = \{1, 2, \ldots, |\mathcal{N}|\}$ with the transmitter and the receiver of the pair n denoted by $Tx(n)$ and $Rx(n)$, respectively, and each pair can operate either in D2D mode or C-RAN mode. Specifically, if the pair n selects the D2D mode, $Tx(n)$ transmits to $Rx(n)$ directly in the uplink time slot. While if the pair n selects the C-RAN mode, $Tx(n)$ is served by multiple RRHs via joint reception with $Rx(n)$ keeping quiet in the uplink time slot, and then the RRHs transmit the message of $Tx(n)$ to $Rx(n)$ via joint beamforming as in [25] with $Tx(n)$ keeping silent in the downlink time slot. Compared to the traditional cellular mode in [18]–[22], the C-RAN mode can be beneficial from the viewpoint of the centralized signal processing capability of the BBU pool. Define the set of RRHs and the set of CUEs as $\mathcal{K} = \{1, 2, \ldots, |\mathcal{K}|\}$ and $\mathcal{C} = \{1, 2, \ldots, |\mathcal{C}|\}$, respectively. The set of all transmitting UEs including both CUEs and the transmitters of potential D2D pairs is denoted by $\mathcal{M} = \{1, 2, \ldots, |\mathcal{C}|, |\mathcal{C}|+1, \ldots, |\mathcal{C}|+|\mathcal{N}|\}$, where $m \in \mathcal{M}$ represents CUE m if $1 \le m \le |\mathcal{C}|$, and $m \in \mathcal{M}$ represents $Tx(m-|\mathcal{C}|)$ if $|\mathcal{C}|+1 \le m \le |\mathcal{C}|+|\mathcal{N}|$. In addition, each RRH and transmitting UE is equipped with a single antenna. The set of available subchannels is denoted by $\mathcal{D} = \{1, 2, \ldots, |\mathcal{D}|\}$ with $|\mathcal{D}| = |\mathcal{C}|$, and the bandwidth of each subchannel is B.

For the pair n in D2D mode that transmits over subchannel d, the received symbol at $Rx(n)$ is given by

$$y_{n,d}^{(D)} = \sqrt{p_{n,d}} h_{n,d}^{(D)} s_n + \sum_{m \in \mathcal{M}, m \neq |\mathcal{C}|+n} \sqrt{p_{m,d}} h_{m,n,d}^{(D)} s_m + w_n, \tag{1}$$

where $s_n \sim \mathcal{CN}(0,1)$ is the symbol of $Tx(n)$, $p_{n,d}$ is the transmit power of $Tx(n)$, $h_{n,d}^{(D)}$ is the channel coefficient between $Tx(n)$ and $Rx(n)$, $h_{m,n,d}^{(D)}$ is the channel coefficient between transmitting UE m and $Rx(n)$, and $w_n \sim \mathcal{CN}(0, \sigma^2)$ is the noise at $Rx(n)$.

Then the data rate achieved by the pair n is given by

$$R_{n,d}^{(D)} = \log\left(1 + \frac{p_{n,d}|h_{n,d}^{(D)}|^2}{\sum_{m \in \mathcal{M}, m \neq |\mathcal{C}|+n} p_{m,d}|h_{m,n,d}^{(D)}|^2 + \sigma^2}\right). \tag{2}$$

When the pair n selects C-RAN mode, $Tx(n)$ would transmit to RRHs. Denote the set of RRHs that $Tx(n)$ associates with by $\mathcal{K}_n$, and the received symbol vector of these RRHs

over subchannel d is given by

$$\mathbf{y}_{n,d}^{(C)} = \sqrt{p_{n,d}}\mathbf{h}_{n,d}^{(C)} s_n + \sum_{m\in\mathcal{M}, m\neq|\mathcal{C}|+n} \sqrt{p_{m,d}}\mathbf{h}_{m,n,d}^{(C)} s_m + \mathbf{z}_n, \tag{3}$$

where $\mathbf{h}_{n,d}^{(C)} \in \mathbb{C}^{|\mathcal{K}_n|\times 1}$ is the channel vector from $Tx(n)$ to its associated RRHs, $\mathbf{h}_{m,n,d}^{(C)} \in \mathbb{C}^{|\mathcal{K}_n|\times 1}$ is the channel vector from transmitting UE m to the associated RRHs of $Tx(n)$, and $\mathbf{z}_n \in \mathbb{C}^{|\mathcal{K}_n|\times 1}$ is the noise vector which is distributed as $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_{|\mathcal{K}_n|})$.

Suppose that minimum mean square error (MMSE) detection is used at the BBU pool; then the estimated symbol for $Tx(n)$ is given by

$$\hat{s}_{n,d} = \mathbf{g}_{n,d}^H \mathbf{y}_{n,d}^{(C)}, \tag{4}$$

where $\mathbf{g}_{n,d} \in \mathbb{C}^{|\mathcal{K}_n|\times 1}$ is the MMSE receiver for $Tx(n)$.

The mean square error (MSE) of $Tx(n)$ can be expressed as

$$\mathbb{E}\left[\left(\hat{s}_{n,d} - s_n\right)\left(\hat{s}_{n,d} - s_n\right)^H\right] = \mathbf{g}_{n,d}^H \mathbf{B}_{n,d}\mathbf{g}_{n,d} - \mathbf{g}_{n,d}^H \mathbf{b}_{n,d} - \mathbf{b}_{n,d}^H \mathbf{g}_{n,d} + 1, \tag{5}$$

where $\mathbf{B}_{n,d} = \mathbf{b}_{n,d}\mathbf{b}_{n,d}^H + \sum_{m\in\mathcal{M}, m\neq|\mathcal{C}|+n} \mathbf{b}_{m,n,d}\mathbf{b}_{m,n,d}^H + \sigma^2\mathbf{I}$, $\mathbf{b}_{n,d} = \sqrt{p_{n,d}}\mathbf{h}_{n,d}^{(C)}$, and $\mathbf{b}_{m,n,d} = \sqrt{p_{m,d}}\mathbf{h}_{m,n,d}^{(C)}$.

Hence, the MMSE receiver for $Tx(n)$ is given by

$$\mathbf{g}_{n,d} = {\mathbf{B}_{n,d}}^{-1}\mathbf{b}_{n,d}, \tag{6}$$

where it can be seen that the main computing burden comes from the calculation of the inverse of a $|\mathcal{K}_n| \times |\mathcal{K}_n|$ matrix $\mathbf{B}_{n,d}$ with computational complexity $\mathcal{O}\left(|\mathcal{K}_n|^3\right)$. In this paper, the centralized signal processing capability of the BBU pool is assumed to be constrained, and hence $|\mathcal{K}_n|$, the number of RRHs that $Tx(n)$ associates with, should be limited. Therefore, define V as the maximum number of RRHs that each UE can associate with; the reasonable value of V can be 3 or 4 according to the result in [26].

When the pair n operates in the C-RAN mode, the transmit rate achieved in uplink can be written as

$$R_{n,d}^{(C)} = \log\left(1 + \frac{p_{n,d}\left|\mathbf{g}_{n,d}^H \mathbf{h}_{n,d}^{(C)}\right|^2}{I_{n,d} + \sigma^2\mathbf{g}_{n,d}^H\mathbf{g}_{n,d}}\right), \tag{7}$$

where $I_{n,d} = \sum_{m\in\mathcal{M}, m\neq|\mathcal{C}|+n} p_{m,d}\left|\mathbf{g}_{n,d}^H \mathbf{h}_{m,n,d}^{(C)}\right|^2$. For CUE $c \in \mathcal{C}$, its achieved data rate in uplink R_c has a similar form to equation (7).

III. Problem Formulation and A Distributed Approach

In this section, an optimization problem aiming at maximizing the overall SE under multiple practical constraints is formulated, and a distributed approach is proposed.

A. Problem Formulation

For CUEs, their RRH association, subchannel allocation, and transmit power can be assumed to be pre-determined. For example, each CUE associates with the V nearest RRHs, and orthogonal subchannels are assigned to CUEs in a random manner with one subchannel for each CUE. Moreover, each CUE and each D2D transmitter in the C-RAN mode transmit at the maximal power level p_{max}. Hence, the system throughput mainly depends on the mode selection, subchannel allocation, RRH association, and transmit power of D2D pairs. However, the joint optimization considering all these items would be quite difficult to solve directly since the corresponding problem contains both continuous variables and integer variables that are tightly coupled. In this case, to reduce the number of optimization variables and keep our work focused, two mappings f_1 and f_2 are presented, which map the mode selection and subchannel allocation of D2D pairs to their RRH association and transmit power, respectively.

Further, the mapping f_1 should meet the fronthaul capacity constraint and the centralized signal processing constraint as follows:

$$\sum_{n\in\mathcal{N}} f_{k,n} \leq F_k, \quad \forall k \in \mathcal{K}, \tag{8}$$

$$\sum_{k\in\mathcal{K}} f_{k,n} \leq V, \quad \forall n \in \mathcal{N}, \tag{9}$$

where $f_{k,n} \in \{0, 1\}$ with $k \in \mathcal{K}$ and $n \in \mathcal{N}$ indicates whether RRH k associates with $Tx(n)$ or not, and F_k is the maximum number of UEs that the fronthaul between RRH k and the BBU pool can still serve under the RRH association of CUEs. Note that the reason for not imposing data rate fronthaul constraints is because they implicitly assume that each fronthaul link can serve an unlimited number of UEs, which is unrealistic in practical systems [6]. In addition, since the pairs in D2D mode reuse the subchannels of CUEs, the mapping f_2 should meet the inter-tier interference constraint and transmit power constraint as follows:

$$I_c \leq \theta_c, \quad \forall c \in \mathcal{C}, \tag{10}$$

$$0 \leq p_n^d \leq p_{max}, \quad \forall n \in \mathcal{N},\ d \in \mathcal{D}, \tag{11}$$

where $I_c = \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}_c} p_n^d \left|h_{n,k,d}^{(C)}\right|^2$ is the total inter-tier interference suffered by CUE c with the subchannel d, $\mathcal{K}_c$ is the set of RRHs that associate with CUE c, p_n^d is the transmit power of the D2D pair n on the subchannel d, $h_{n,k,d}^{(C)}$ is the channel coefficient between $Tx(n)$ and the RRH k, θ_c is the interference threshold, and p_{max} is the maximal transmit power of each pair. In this paper, the mappings f_1 and f_2 actually represent Algorithms 2 and 3 (described below), respectively, which guarantee that the constraints (8), (9), (10), and (11) hold for any mode selection and subchannel allocation of D2D pairs, and the corresponding mathematical problems are a combinatorial utility maximization problem and a mathematical program with equilibrium constraints given by (15) and (18)-(20) in Section IV, respectively. Although f_1 and f_2 are not optimal mappings, they possess some advantages such as distributed

implementation, and further details about their design are discussed in Section IV.

At this point, the throughput can be expressed with a function of mode selection and subchannel allocation of D2D pairs, and the overall SE maximization for the uplink can be formulated as

$$\max_{\left\{x_n^{(D)},x_n^{(C)}\right\},\mathbf{Q}} \frac{U_{system}\left(\left\{x_n^{(D)},x_n^{(C)}\right\},\mathbf{Q},f_1,f_2\right)}{|\mathcal{D}|\,B}$$
$$s.t.\ (a)\ x_n^{(D)},\quad x_n^{(C)}\in\{0,1\},\ \forall n\in\mathcal{N},$$
$$(b)\ x_n^{(D)}+x_n^{(C)}=1,\quad \forall n\in\mathcal{N},$$
$$(c)\ \sum_{d=1}^{|\mathcal{D}|} q_{d,n}=1,\quad \forall n\in\mathcal{N}, \tag{12}$$

where $x_n^{(D)}$ and $x_n^{(C)}$ are the mode selection indicators for the D2D pair n, $\mathbf{Q}$ is the matrix with elements $q_{d,n}\in\{0,1\}$ with $d\in\mathcal{D},n\in\mathcal{N}$, which indicates whether the subchannel d is allocated to D2D pair n or not, $U_{system}=\sum_n\left[x_n^{(D)}\sum_{d,q_{d,n}=1}BR_{n,d}^{(D)}+x_n^{(C)}\sum_{d,q_{d,n}=1}BR_{n,d}^{(C)}\right]+\sum_c BR_c$ is the system throughput in the uplink, $f_1\left(\left\{x_n^{(D)},x_n^{(C)}\right\},\mathbf{Q}\right)$ represents the RRH association scheme, and $f_2\left(\left\{x_n^{(D)},x_n^{(C)}\right\},\mathbf{Q}\right)$ represents the power control scheme. The constraints (a) and (b) jointly require that each D2D pair can select only one communication mode, while the constraint (c) states that each pair can use only one subchannel. Note that the same constraints as constraint (c) can be found in [27] and [28].

B. Distributed Mode Selection and Resource Allocation

The problem (12) is hard to solve directly due to its NP-hardness [21], and the corresponding centralized approaches lead to high complexity and require global CSI, putting heavy burdens on the BBU pool and fronthaul. Fortunately, with the ever increasing computing capabilities of UEs, it is possible to let them help alleviate the burden on the C-RAN infrastructure. Hence, in this paper, UEs are endowed with decision-making capabilities, and the problem (12) can be handled in a way such that D2D pairs autonomously optimize the mode selection and resource allocation.

To achieve this goal, inspired by the works [29] and [30], distributed reinforcement learning with joint utility and strategy estimation is adopted whose process is given as follows:

$$\begin{cases}\hat{r}_{n,a_{n,i}}(t)=\hat{r}_{n,a_{n,i}}(t-1)\\ \quad+\theta(t)\,\mathbb{1}_{\{a_n(t)=a_{n,i}\}}\left(\tilde{r}_n(t)-\hat{r}_{n,a_{n,i}}(t-1)\right),\\ \pi_{n,a_{n,i}}(t)=\pi_{n,a_{n,i}}(t-1)\\ \quad+\gamma(t)\left(\tilde{\beta}_{n,a_{n,i}}^{(\zeta)}\left(\hat{\boldsymbol{r}}_n(t)\right)-\pi_{n,a_{n,i}}(t-1)\right),\end{cases} \tag{13}$$
$$(i)\ \lim_{T\to\infty}\sum_{t=1}^{T}\theta(t)=+\infty,\quad \lim_{T\to\infty}\sum_{t=1}^{T}\theta(t)^2<+\infty,$$
$$(ii)\ \lim_{T\to\infty}\sum_{t=1}^{T}\gamma(t)=+\infty,\quad \lim_{T\to\infty}\sum_{t=1}^{T}\gamma(t)^2<+\infty,$$
$$(iii)\ \lim_{t\to\infty}\frac{\gamma(t)}{\theta(t)}=0, \tag{14}$$

where $a_{n,i}$ is the i-th available action of decision-maker n, $\hat{r}_{n,a_{n,i}}$ is the utility estimation of action $a_{n,i}$, $\hat{\boldsymbol{r}}_n(t)=\left(\hat{r}_{n,a_{n,1}}(t),\ldots,\hat{r}_{n,a_{n,A}}(t)\right)$ where A is the number of all possible actions, $\tilde{r}_n(t)=r_n(t)+\xi_n$ is the feedback value of decision-maker n's utility at iteration t with ξ_n being the feedback noise, $\pi_{n,a_{n,i}}$ is the probability that decision-maker n selects action $a_{n,i}$, the strategy of decision-maker n is a vector of $\pi_{n,a_{n,i}}$ denoted by $\boldsymbol{\pi}_n$, $\tilde{\beta}_{n,a_{n,i}}^{(\zeta)}\left(\hat{\boldsymbol{r}}_n(t)\right)=\frac{\exp\left(\zeta\hat{r}_{n,a_{n,i}}(t)\right)}{\sum_{j=1}^{A}\exp\left(\zeta\hat{r}_{n,a_{n,j}}(t)\right)}$, ζ is a parameter to balance exploitation and exploration [29], and $\theta(t)$ and $\gamma(t)$ are learning rates.

The first equation in (13) for utility estimation originates from the standard reinforcement learning process to learn the expected utility [31], while the second equation in (13) is a revised version of the strategy update in [32]. Meanwhile, conditions (i) and (ii) in (14) are essential to guarantee the convergence of the learning procedure, whose roles are explained in Section V. The reinforcement learning process is totally distributed and composed of two coupled phases. First, using the feedback value of the utility, each decision-maker estimates the utility of each action. Second, decision-makers update their strategies based on the estimated values. The benefits of the process lie in the low computational burdens and low information requirements for the decision-makers, because only simple mathematical operations are executed, and each decision-maker needs only to acquire its own utility value.

To apply the learning process, an action of the D2D pair n is taken as a combination of a communication mode and a subchannel, and the utility of the D2D pair n is given by $r_n=\frac{U_{system}}{|\mathcal{D}|B}$. It has been shown that better system performance can be achieved by taking the utility of each decision-maker as the system objective, and meanwhile this setting can guarantee the convergence of the learning process [30]. The following algorithm is proposed to solve problem (12):

The iterative approach proposed in Algorithm 1 contains three stages. At each iteration, each D2D pair first selects one action randomly according to its current strategy, then utility value determination is performed, including distributed RRH association and distributed power control. Based on the noisy utility feedback, each pair updates its strategy individually. Note that pairs can acquire the feedback values from the HPN. Specifically, after RRH association and power control are performed, potential D2D transmitters and CUEs transmit. Then the BBU pool measures the total data rate of all the UEs accessing RRHs whose values are delivered to the HPN via backhaul, while each receiver of D2D pairs in the D2D mode measures its data rate whose value is delivered to the HPN via the control channel. Once the HPN receives all the necessary data rate information, it calculates

Algorithm 1 Distributed Mode Selection and Resource Allocation

1: **Stage 1: Initialization**
$\forall n \in \mathcal{N}$, $\hat{\boldsymbol{r}}_n(t=0) = (0, \ldots, 0)$,
$\boldsymbol{\pi}_n(t=0) = \frac{1}{2|\mathcal{D}|}(1, \ldots, 1)$.
2: **Stage 2: Action Selection**
$\forall n \in \mathcal{N}$, the D2D pair n randomly selects an action according to $\boldsymbol{\pi}_n(t)$.
3: **Stage 3: Utility Value Determination**
The transmitters of D2D pairs selecting C-RAN mode compete for RRH association by the many-to-many matching based process shown in Algorithm 2.
Next, D2D pairs selecting D2D mode participate in the Stackelberg game based power control process shown in Algorithm 3.
4: **Stage 4: Strategy Update**
Each D2D pair updates its strategy using formulas in (13).
5: **Stage 5: Convergence Check**
If the strategy of each player converges, the algorithm terminates. Otherwise, go back to **Stage 2** with the updated strategies.

the system SE and broadcasts the result to the UEs. Since the utility of each pair resulting from selecting each action is determined by the RRH association process and power control process where constraints (8), (9), and (10) are considered, the nature of Algorithm 1 is that D2D pairs self-optimize the communication mode selection and subchannel allocation under fronthaul capacity constraints, centralized signal processing capability constraints, and inter-tier interference constraints.

Actually, the reinforcement learning can handle the communication mode selection, subchannel allocation, transmission power optimization, and RRH association directly by modifying the action of each pair as the combination of a communication mode, a subchannel, a discrete power level, and a set of RRHs it will associate with, and meanwhile setting the utility of each pair to 0 if any of the constraints in our model like the fronthaul capacity constraint is violated. However, involving the selection of power levels and the set of RRHs will cause considerable growth of the number of actions for each pair, and hence the convergence time of reinforcement learning can be unacceptable due to the long time needed for estimating the utilities of actions.

IV. Utility Value Determination

After each pair selects its action, RRH association is determined for D2D transmitters in C-RAN mode based on a matching process, while the pairs selecting D2D mode participate in a Stackelberg game based power control process, where the inter-tier interference constraint (10) would be satisfied when the process terminates.

A. Many-to-Many Matching Based RRH Association

Matching theory can provide a distributed solution for RRH association that is easy for fast implementation [33]. In the matching theory based RRH association process, the RRH set $\mathcal{K}$ and the set of D2D transmitters in C-RAN mode $\mathcal{N}_{C,Tx}$ are two teams of players. To characterize the constraints (8) and (9), the quota of each transmitter is set to V, and the quota of RRH k is set to F_k. In addition, each transmitter would like to access RRHs with better CSI, and each RRH would like to be associated with transmitters with better communication environments. The matching is defined as an assignment of the players in $\mathcal{N}_{C,Tx}$ to the players in $\mathcal{K}$. If player $n \in \mathcal{N}_{C,Tx}$ is assigned to player $k \in \mathcal{K}$, it means that $Tx(n)$ would be associated with RRH k. Note that the BBU pool acts on behalf of RRHs, and the considered matching is a many-to-many matching since each transmitter can be associated with multiple RRHs and each RRH can accommodate multiple transmitters. Formally, a many-to-many matching is defined as follows:

Definition 1: A many-to-many matching η is a mapping from the set $\mathcal{N}_{C,Tx} \cup \mathcal{K}$ into the set of all the subsets of $\mathcal{N}_{C,Tx} \cup \mathcal{K}$ such that for every $n \in \mathcal{N}_{C,Tx}$ and $k \in \mathcal{K}$:

1) $\eta(n)$ is contained in $\mathcal{K}$ and $\eta(k)$ is contained in $\mathcal{N}_{C,Tx}$;
2) $|\eta(n)| \leq V$ for all n in $\mathcal{N}_{C,Tx}$;
3) $|\eta(k)| \leq F_k$ for all k in $\mathcal{K}$;
4) k is in $\eta(n)$ if and only if n is in $\eta(k)$,

where $\eta(n)$ is the set of RRHs that $Tx(n)$ is associated with, and $\eta(k)$ is the set of transmitters communicating with RRH k. The definition states that a matching is a many-to-many relation in the sense that each RRH is matched to a set of transmitters, and vice-versa.

Before designing the matching algorithm, the preference list for each player should be defined. For $Tx(n)$ selecting subchannel d, it holds a preference list where the indices of RRHs are ordered according to $l_{n,k,d} = \left|h^{(C)}_{n,k,d}\right|^2$, and the index of the RRH related to the largest $l_{n,k,d}$ is at the top of the preference list. For RRH k, it holds a preference list containing the indices of all the transmitters which are ordered according to $l'_{n,k,d} = \frac{l_{n,k,d}}{\sum_{j \in \mathcal{M}_d, j \neq |\mathcal{C}|+n} l_{j,k,d}}$ with $\mathcal{M}_d$ denoting the set of transmitting UEs operating on subchannel d, and the transmitter related to the largest $l'_{n,k,d}$ is most preferred. Note that $l'_{n,k,d}$ can coarsely characterize the communication environment experienced by $Tx(n)$ from RRH k's point of view.

The many-to-many matching process with transmitters proposing is designed for RRH association as summarized in Algorithm 2, which involves three stages. In the first stage, the transmitters and the BBU pool collect necessary CSI. In the second stage, each transmitter makes its own preference list, while the preference list for each RRH is made by the BBU pool. The third stage contains two steps. In the first step, each transmitter proposes to the V RRHs at the top of the preference list, and then RRH k accepts the best F_k proposals among the received proposals and rejects the rest. In the second step, $Tx(n)$ deletes the indices of RRHs who have rejected its proposal from the preference list and proposes to the V RRHs at the top of the new preference list. Afterwards, RRH k accepts the best F_k proposals among the received proposals and rejects the rest. The second step is executed repeatedly until no rejection occurs. The corresponding mathematical problem that Algorithm 2 solves is a combinatorial utility maximization problem with fronthaul capacity constraints and

computing capability constraints as follows [34]:

$$\begin{aligned}\max_{f_{k,n}} \ & \sum_k \sum_n f_{k,n} {l'}_{n,k,d_n} + \sum_k \sum_n f_{k,n} l_{n,k,d_n} \\ (1) \ & \sum_{n\in\mathcal{N}} f_{k,n} \le F_k, \quad \forall k \in \mathcal{K}, \\ (2) \ & \sum_{k\in\mathcal{K}} f_{k,n} \le V, \quad \forall n \in \mathcal{N}, \\ (3) \ & f_{k,n} \in \{0,1\}, \end{aligned} \tag{15}$$

where d_n represents the subchannel allocated to pair n.

Algorithm 2 CSI Based Many-to-Many Matching Algorithm for RRH Association

1: **Stage 1:**
The transmitters of pairs in C-RAN mode and the BBU pool gather necessary CSI.
2: **Stage 2:**
Each pair and each RRH holds a preference list based on the collected information.
3: **Stage 3:**
Each transmitter proposes to the V RRHs at the top of its preference list.
Each RRH accepts the proposals from the best transmitters with regard to its quota and preference list, and rejects the remaining proposals.
4: **Stage 4:**
Repeat:
Each transmitter deletes the indices of RRHs rejecting its proposal, and proposes to the V RRHs at the top of the new preference list.
Each RRH accepts the proposals from the best transmitters regarding its quota and preference list, and rejects the remaining proposals.
Until: No rejection occurs.

B. Stackelberg Game Based D2D Power Control

To suppress the inter-tier interference to CUEs, the transmit power of the D2D pairs in D2D mode should be properly configured. To this end, the BBU pool can set a price for the inter-tier interference generated by D2D pairs based on which D2D pairs adjust their transmit powers. Since the policy is hierarchical, a Stackelberg game is naturally suited to help develop a distributed power control scheme, where the BBU pool is taken as the leader and D2D pairs act as followers [35]. Focusing on an arbitrary subchannel d, the price of unit interference is denoted by μ_d. Then for the D2D pair n operating on subchannel d, it aims to maximize its utility function through solving the following optimization problem:

$$\begin{aligned}&\max_{p_n^d} \ U_n^d\left(\mu_d, p_n^d, \boldsymbol{p}_{-n}^d\right) = \mathrm{R}_{n,d}^{(D)} - \mu_d \sum_{k\in\mathcal{K}_c} p_n^d l_{n,k,d}, \\ &s.t.: \ 0 \le p_n^d \le p_{max}, \end{aligned} \tag{16}$$

where $\boldsymbol{p}_{-n}^d$ denotes the transmit power profile of all the other pairs in D2D mode using subchannel d, $\mathcal{K}_c$ denotes the set of associated RRHs of CUE c allocated with subchannel d, and the optimization objective of the pair n contains the data rate and the cost of causing inter-tier interference. It should be mentioned that an additional price term $-\eta_d \sum_{m\in\mathcal{N}', m\neq n} p_n^d \left|h_{n,m,d}^{(D)}\right|^2$ can be added to the utility function to account for the impact of interference generated by the D2D pair n on other D2D pairs with η_d the price per unit of co-tier interference on subchannel d and $\mathcal{N}'$ the set of pairs in D2D mode using subchannel d, which can help to achieve a better convergence point at the cost of more overhead to estimate the CSI between the transmitter of the pair n and the receivers of other pairs. Note that this change would not affect the analysis and conclusions below. In the following, the subchannel index d is dropped for simplicity of notation.

By solving (16), the optimal value of p_n under fixed μ and $\boldsymbol{p}_{-n}$ is given by

$$p_n^*\left(\mu, \boldsymbol{p}_{-n}\right) = \left[\frac{1}{\mu l_n \ln 2} - \frac{I_n^D + I_n^C + \sigma^2}{\left|h_n^{(D)}\right|^2}\right]_0^{p_{max}}, \tag{17}$$

where $l_n = \sum_{k\in\mathcal{K}_c} l_{n,k}$, $I_n^D = \sum_{m\in\mathcal{N}', m\neq n} p_m \left|h_{m,n}^{(D)}\right|^2$, I_n^C denotes the total interference from CUE c and D2D transmitters in C-RAN mode, and $[x]_0^{x_{\max}} = \min\{x_{\max}, \max\{x, 0\}\}$.

For the BBU pool, its objective is assumed to make $|I_c - \theta_c|$ less than a very small value, which is enough to make constraint (10) approximately satisfied. Based on the above derivation, a Stackelberg game based D2D power control process is developed which is shown in Algorithm 3. In the proposed process, all the pairs using subchannel d first transmit at maximal powers, and the BBU pool measures I_c for CUE c. If $I_c > \theta_c$, the power control process is triggered, and then the BBU pool initializes a price $\mu = \frac{\mu_{upper}+\mu_{lower}}{2}$, where μ_{upper} and μ_{lower} are the maximal value and the minimal value of the price, respectively. Given μ, pairs in D2D mode update their transmit power to maximize utilities. After the update process converges, the BBU pool again measures I_c. If $I_c > \theta_c$, the BBU pool would set $\mu_{lower} = \mu_d$ and then set $\mu_d = \frac{\mu_{upper}+\mu_{lower}}{2}$. If $I_c < \theta_c$, the BBU pool would set $\mu_{upper} = \mu_d$ and then set $\mu_d = \frac{\mu_{upper}+\mu_{lower}}{2}$. The process repeats until $|I_c - \theta_c|$ is less than a very small value.

Actually, Algorithm 3 solves the following optimization problem, which is known as a mathematical program with equilibrium constraints:

$$\min_{\mu} \left|I_c - \theta_c\right|, \tag{18}$$

$$s.t. \ \mu \ge 0, \tag{19}$$

$$p_n \in \arg\max_{p_n} \left\{U_n : \ 0 \le p_n \le p_{\max}\right\}, \quad \forall n \in \mathcal{N}', \tag{20}$$

where the subproblem composed of (18) and (19) is the upper level problem related to the BBU pool which sets the interference price, while the subproblem (20) is the lower level problem related to each D2D pair, which represents the equilibrium constraints.

Algorithm 3 Stackelberg Game Based D2D Power Control

1: **Stage 1:**
For subchannel d allocated to CUE c, all the pairs using this subchannel transmit at maximal powers, and the BBU pool measures I_c. If $I_c > \theta_c$, the BBU pool sets $\mu_d = \frac{\mu_{upper}+\mu_{lower}}{2}$, and the process goes to **Stage 2**.
2: **Stage 2:**
Pairs in D2D mode update the transmit power by using (17) iteratively:
$p_n(t) = p_n^*(\boldsymbol{p}_{-n}(t-1))$,
until the transmit powers of all the pairs converge.
3: **Stage 3:**
The BBU pool measures I_c. Then set $\mu_{lower} = \mu_d$ if $I_c > \theta_c$, and set $\mu_{upper} = \mu_d$ if $I_c < \theta_c$.
Set $\mu_d = \frac{\mu_{upper}+\mu_{lower}}{2}$.
4: **Stage 4:**
Repeat stage 2 and 3 until $|I_c - \theta_c|$ is less than a very small value.

V. Properties of the Proposed Approach

In this section, the properties of the proposed approach in Algorithm 1 are discussed to make the proposal clear.

A. CSI Requirement

For the D2D pairs, first, the CSI is not needed in the reinforcement learning process. Second, in the RRH association process, each pair only needs to know the CSI between itself and RRHs. Third, in the power control process, each pair only needs to know the CSI between its transmitter and receiver, and the CSI between its transmitter and RRHs, since the interference from other pairs and CUEs plus noise can be directly measured at the receiver. For the BBU pool, the CSI between transmitting UEs and RRHs is needed to make preference lists for RRHs in the RRH association process and calculate the MMSE receivers. Compared to the centralized approaches where global CSI is needed in a central node, D2D pairs and the BBU pool need only partial CSI in the proposed distributed approach.

B. The Convergence of Reinforcement Learning

The convergence proof relies on two-time-scale stochastic approximation. Specifically, (13) can be analyzed by studying the following ordinary differential equations (ODEs) owing to conditions (i) and (ii) in (14) [36]:

$$\begin{aligned}\dot{\hat{r}}_{n,a_{n,i}} &= \pi_{n,a_{n,i}}\left(\bar{r}\left(a_{n,i},\boldsymbol{\pi}_{-n}\right) - \hat{r}_{n,a_{n,i}}\right),\\ \dot{\hat{\pi}}_{n,a_{n,i}} &= \beta_{n,a_{n,i}}\left(\hat{\boldsymbol{r}}_n\right) - \pi_{n,a_{n,i}},\end{aligned} \tag{21}$$

where $\beta_{n,a_{n,i}}\left(\hat{\boldsymbol{r}}_n\right) = \frac{\exp\left(\zeta\hat{r}_{n,a_{n,i}}\right)}{\sum_{a_{n,j}\in\mathcal{A}_n}\exp\left(\zeta\hat{r}_{n,a_{n,j}}\right)}$ and $\bar{r}\left(a_{n,i},\boldsymbol{\pi}_{-n}\right)$ is the expected utility of potential D2D pair n that results from taking the i-th action when the strategies of other pairs are $\boldsymbol{\pi}_{-n}$. Then, under condition (iii) in (14), the first ODE, which corresponds to the fast process $\hat{r}_{n,a_{n,i}}$, can be analyzed by fixing strategy $\pi_{n,a_{n,i}}$, and the second ODE, which corresponds to the slow process $\pi_{n,a_{n,i}}$, can be analyzed as if the utility estimation is accurate, i.e., $\hat{r}_{n,a_{n,i}} = \bar{r}\left(a_{n,i},\boldsymbol{\pi}_{-n}\right)$ [37]. From the first ODE, it can be derived that $\hat{r}_{n,a_{n,i}} = \bar{r}\left(a_{n,i},\boldsymbol{\pi}_{-n}\right)$ when $\pi_{n,a_{n,i}} \neq 0, \forall a_{n,i}$. While for the second ODE, the global convergence of its trajectory is guaranteed when the utility of D2D pairs satisfies

$$\begin{aligned}&r_n\left(a_{n,i},\boldsymbol{a}_{-n}\right) - r_n\left(a'_{n,i},\boldsymbol{a}_{-n}\right)\\ &\qquad = \phi\left(a_{n,i},\boldsymbol{a}_{-n}\right) - \phi\left(a'_{n,i},\boldsymbol{a}_{-n}\right),\end{aligned} \tag{22}$$

which indicates that the interaction between D2D pairs is an exact potential game for which a Lyapunov function must exist [38] and obviously holds for our utility function setting. Since the convergence point must be a zero point of the associated ODE, we have $\pi_{n,a_{n,i}} = \beta_{n,a_{n,i}}\left(\boldsymbol{\pi}_{-n}\right)$, and this stable state is called logit equilibrium.

Note that for practical scenarios with time-varying CSI, $\hat{r}_{n,a_{n,i}}$ will converge to $\mathbb{E}_{\mathrm{h}}\left[\bar{r}\left(a_{n,i},\boldsymbol{\pi}_{-n}\right)\right]$ [36], while $\pi_{n,a_{n,i}}$ will be equal to $\frac{\exp(\zeta\mathbb{E}_{\mathrm{h}}[\bar{r}(a_{n,i},\boldsymbol{\pi}_{-n})])}{\sum_{a_{n,j}\in\mathcal{A}_n}\exp(\zeta\mathbb{E}_{\mathrm{h}}[\bar{r}(a_{n,j},\boldsymbol{\pi}_{-n})])}$ [36]. Then, in the following, each device needs only to select actions based on the learned and fixed strategy, and hence can make a real-time decision on the selected channel and communication mode, implicitly taking into account the dynamics of CSI.

C. The Convergence of the RRH Association Process

The convergence of the matching algorithm is guaranteed by the following theorem.

Theorem 1: The convergence of Algorithm 2 is ensured.

Proof: For each transmitter, the number of RRH indices on its preference list is limited, and meanwhile the indices of RRHs rejecting its proposal would be deleted. Hence, after **Stage 4** of Algorithm 2 repeats finitely many times, no rejections would occur, which means the Algorithm 2 must converge to a final matching. ■

The concept of pair-wise stable matching is given below to make preparations for the proof of stability.

Definition 2 (Pairwise-Stable [39]): The matching η is pairwise-stable if there are no transmitter n and RRH k that are not paired with each other under η such that $Ch_n\left(\eta(n)\cup\{k\}\right) \underset{n}{\succ} \eta(n)$ and $Ch_k\left(\eta(k)\cup\{n\}\right) \underset{k}{\succ} \eta(k)$, where $\forall j \in \mathcal{K}\cup\mathcal{N}_{C,Tx}$, $Ch_j(\mathcal{W})$ represents the subset of $\mathcal{W}$ that player j likes best with $\mathcal{W}$ denoting the set of possible partners, and $\mathcal{W} \underset{j}{\succ} \mathcal{B}$ means j prefers $\mathcal{W}$ to $\mathcal{B}$ with $\succ$ being irreflexive.

Based on the above definition, the stability of the final matching is ensured by the following theorem.

Theorem 2: The final matching reached by Algorithm 2 is pairwise-stable.

Proof: The proof follows by contradiction. Assume that the final matching η is unstable. Then, there must exist a transmitter-RRH pair (n,k) satisfying $n \notin \eta(k)$ and $k \notin \eta(n)$ and meanwhile $Ch_k\left(\eta(k)\cup\{n\}\right) \underset{k}{\succ} \eta(k)$ and $Ch_n\left(\eta(n)\cup\{k\}\right) \underset{n}{\succ} \eta(n)$. Overall, there are two possible cases under the instability assumption. One case is that transmitter n has never proposed to RRH k. In this case, $Ch_n\left(\eta(n)\cup\{k\}\right) = \eta(n)$, and $Ch_n\left(\eta(n)\cup\{k\}\right) \underset{n}{\succ} \eta(n)$ does not hold. The other case is that transmitter n has proposed

to RRH k. Nevertheless, $n \notin \eta(k)$ means the proposal was rejected. Hence, $Ch_k(\eta(k) \cup \{n\}) = \eta(k)$, and then $Ch_k(\eta(k) \cup \{n\}) \underset{k}{\succ} \eta(k)$ does not hold. Since contradictions are found under both cases, the final matching η must be pairwise-stable. ■

Next, the optimality of the proposal is given by the following theorem.

Theorem 3: The stable matching reached by Algorithm 2 is Pareto-optimal for D2D transmitters.

Proof: Assume that the stable matching η is not Pareto-optimal, and then there must exist an RRH and D2D transmitter pair (k, n) satisfying $n \notin \eta(k)$ and $k \notin \eta(n)$ such that $Ch_n(\eta(n) \cup \{k\}) \underset{n}{\succ} \eta(n)$. Nevertheless, since the matching η is stable, $Ch_k(\eta(k) \cup \{n\}) \underset{k}{\succ} \eta(k)$ does not hold, as otherwise the stability of the matching η would be violated. Actually, since the preference of RRHs over D2D transmitters is strict, pairing with D2D transmitter n would hurt RRH k because RRH k must reject a D2D transmitter that is preferred by it to accommodate D2D transmitter n. Hence, even though pairing with RRH k helps the improvement of D2D transmitter n, RRH k would not allow this pairing. Therefore, under stable matching η, the RRH association of each D2D transmitter cannot be improved, and thus the matching η must be Pareto-optimal for D2D transmitters. ■

Finally, we discuss the signaling overhead to show the implementation feasibility. In each round of Algorithm 2, the total number of association requests from D2D transmitters is at most $|\mathcal{N}_C|V$, where $\mathcal{N}_C$ is the set of D2D transmitters in C-RAN mode. Meanwhile, each D2D transmitter needs one bit to send a request to the desired RRH, and each RRH needs one bit to inform the corresponding D2D transmitter of the decision on its association. Therefore, the total signalling overhead in each round requires at most $2|\mathcal{N}_C|V$ bits. Nevertheless, once the request from a D2D transmitter is rejected by the desired RRH, the D2D transmitter would not send the request information to the RRH for the second time. Hence, with the execution of Algorithm 2, the number of available RRHs on the preference list of each D2D transmitter can be reduced, and partial transmitters can send request information to fewer than V RRHs. Thus, the total signalling overhead in each round is less than $2|\mathcal{N}_C|V$ bits with Algorithm 2 running. Moreover, considering that each D2D transmitter can only possibly associate with RRHs within a certain distance, the number of RRHs on the preference list of each D2D transmitter is much less than the total number of RRHs. Thus, Algorithm 2 can reach a stable matching after a small number of rounds, and it can be concluded that the amount of information exchange required by Algorithm 2 is tolerable.

D. The Convergence of the Power Control Process

We first consider the convergence and stability of the power update procedure under a given price. For subchannel d, the power update procedure of the corresponding pairs in Algorithm 3 can be seen as a non-cooperative game. In each update, each pair reacts to the power profiles of other pairs by playing the best response defined in (17). Denote the set of pairs in D2D mode using subchannel d by $\mathcal{N'}_D \subseteq \mathcal{N}$, and re-label these pairs with $i = 1, 2, \ldots, |\mathcal{N'}_D|$. Following the proof in [40], the following proposition can be derived.

Theorem 4: Under a given price, the power update procedure in Algorithm 3 is guaranteed to converge to a pure-strategy Nash equilibrium (NE) if

$$\left\|\bar{\mathbf{H}}\right\|_1 < \frac{1}{\max\limits_{n \in \mathcal{N}} \frac{1}{\left|h_n^{(D)}\right|^2}}, \tag{23}$$

where $\bar{\mathbf{H}}$ is a $|\mathcal{N'}_D| \times |\mathcal{N'}_D|$ matrix, and

$$\bar{\mathbf{H}}_{ij} = \begin{cases} \left|h_{j,i}^{(D)}\right|^2, & if\ i \neq j, \\ 0, & otherwise. \end{cases} \tag{24}$$

The proof follows by writing the best responses of all the pairs into a vector form and then applying the Banach contraction theorem. The physical interpretation for this convergence result is that the power update process converges if no pair generates excessive interference to others. The specific definition of NE is given as follows.

Definition 3: A transmit power profile $\left[p_1^*, \ldots, p_{|\mathcal{N'}_D|}^*\right]$ is a pure-strategy NE under a given price μ if, for each pair i, the following holds:

$$U_i\left(\mu, p_i^*, \boldsymbol{p}_{-i}^*\right) \geq U_i\left(\mu, p_i, \boldsymbol{p}_{-i}^*\right), \quad \forall p_i \in \mathcal{P}_i, \tag{25}$$

where $\boldsymbol{p}_{-i}^*$ is the transmit power profile of all the pairs except pair i, and $\mathcal{P}_i = \{p_i : 0 \leq p_i \leq p_{max}\}$.

Now we discuss the convergence and stability of the overall power control process. First, it is intuitive that I_c is a continuous function of μ. Second, I_c can achieve its minimal value when μ is sufficiently large, while I_c can achieve its maximal value when μ takes a sufficient small value. Hence, a μ making $|I_c - \theta_c|$ less than a given small value can always be found by the process in Algorithm 2 if μ_{lower} is sufficiently small and μ_{upper} is sufficiently large. Once such a μ is found, the BBU pool would not change the price according to the assumption about its objective. Under this converged price, the result of the power update procedure would not change. Therefore, the convergence of Algorithm 2 is ensured. Meanwhile, under the converged outcome $(\mu^*, \{p_n^*\})$, the BBU pool has no incentive to change the price under $\{p_n^*\}$, and each pair has no incentive to change its transmit power unilaterally under μ^*. Hence, $(\mu^*, \{p_n^*\})$ is a Stackelberg equilibrium.

E. Overall Convergence

Once the distributed reinforcement learning converges to the logit equilibrium, each pair can reach its pure strategy by letting the parameter ζ become sufficiently large as discussed in subsection G, and then the mode selection and subchannel allocation of each pair will not change. Therefore, the RRH association and power control results will not change either, and the mode selection, subchannel allocation, RRH association, and transmit power output determined by Algorithm 1 will be invariant over iterations, which implies the convergence of Algorithm 1.

F. Convergence Time

The previous proof has shown that distributed reinforcement learning converges to a logit equilibrium $\boldsymbol{\pi}^* = \left(\boldsymbol{\pi}_1^*, \ldots, \boldsymbol{\pi}_{|\mathcal{N}|}^*\right)$. Following that, we have the inequality for D2D pair $n \in \mathcal{N}$ given by

$$\sum_{\boldsymbol{a}\in\mathcal{A}} r_n(\boldsymbol{a})\,\Pi_{\boldsymbol{a}} - \sum_{\boldsymbol{a}\in\mathcal{A}} r_n(\boldsymbol{a})\,\Pi_{\boldsymbol{a}}^* \leq \varepsilon, \quad \forall \boldsymbol{\pi}_n \neq \boldsymbol{\pi}_n^*, \tag{26}$$

where $\boldsymbol{a}$ is the action profile of all pairs whose set is denoted by $\mathcal{A}$, $\Pi_{\boldsymbol{a}} = \pi_{n,a_n} \prod_{m\in\mathcal{N},m\neq n} \pi^*_{m,a_m}$, $\Pi^*_{\boldsymbol{a}} = \pi^*_{n,a_n} \prod_{m\in\mathcal{N},m\neq n} \pi^*_{m,a_m}$, and π_{n,a_n} is the probability that the pair n plays action a_n. This inequality suggests that if one D2D pair changes its strategy, it cannot achieve an average utility improvement larger than ε. According to Proposition 5 in [36], the convergence time for the adopted reinforcement learning to reach equilibrium is $\mathcal{O}\left(\log\left(\frac{1}{\varepsilon}\right)\right)$. This measure means it will cost more time to achieve a more exact version of Nash equilibrium. Hence, in reality, if a network designer wants to build a more stable network, the learning algorithm should be executed for a longer time. Meanwhile, it should be noted that this measure is for a fixed network size. Concerning the impact of network size on convergence time, when the number of D2D pairs and the number of subchannels increase, the convergence time will be longer, since the interactions between D2D pairs will be more complex and each pair needs to estimate utilities for more actions.

G. Optimality

Note that when all pairs share an identical utility taken as the system performance U_{system}, the interactions among pairs to select communication mode and subchannel can be seen as an exact potential game with potential function $\Phi = U_{system}$ [35]. Since all Nash equilibria of this game maximize Φ either locally or globally, all Nash equilibria of this game maximize U_{system} either locally or globally. Further, by Proposition 1 in [30], once Algorithm 1 reaches the equilibrium, ε can approach 0 by increasing ζ. Meanwhile, when Algorithm 1 converges to the equilibrium, the probability of each D2D pair playing the i-th action is given by

$$\pi_{n,a_{n,i}} = \frac{\exp\left(\zeta \hat{r}_{n,a_{n,i}}(t)\right)}{\sum_{j=1}^{A} \exp\left(\zeta \hat{r}_{n,a_{n,j}}(t)\right)}. \tag{27}$$

Hence, when ζ increases, each pair intends to play the action with maximal estimated utility, and thus as ζ goes to infinity, the following result can be derived from (26):

$$r_n\left(a_n, \boldsymbol{a}^*_{-n}\right) - r_n\left(a^*_n, \boldsymbol{a}^*_{-n}\right) \leq 0, \quad \forall a_n \neq a^*_n, \tag{28}$$

where a^*_n is the action of the pair n with the highest estimated utility when Algorithm 1 converges, and $\boldsymbol{a}^*_{-n}$ is the action profile of all the other pairs. Note that this inequality is the standard definition of Nash equilibrium, and it can be concluded that Algorithm 1 will converge to a global or local optimal solution to problem (12) as ζ goes to infinity. For the practical implementation, ζ should be taken at first as a small value to allow pairs to estimate the utility sufficiently and then taken as a large value to improve the system performance once the algorithm converges [29].

H. Overall Complexity

In the action selection stage, a pair needs to generate a random number to select an action, whose complexity is $O(1)$. Then for the utility determination stage, the complexity of the distributed matching based RRH-D2D association for each D2D pair mainly lies in the sorting algorithm to make a preference list about RRHs. The sorting complexity for each pair is $\mathcal{O}(K \log K)$ with K the number of RRHs in the network. In addition, the power update formula of each pair contains only multiplications and additions, and hence the complexity is $\mathcal{O}(1)$. Therefore, the power control complexity for each pair is $O(I)$ with I the number of power updates until the power control algorithm converges. As for the strategy update stage based on reinforcement learning, since each pair updates the utility estimation of only one action each time, this update complexity is $\mathcal{O}(1)$. For the second equation of reinforcement learning for the strategy update, the complexity is $O(|\mathcal{D}|)$ because each pair needs to update strategies for $2|D|$ actions. Hence, the complexity of all three stages of Algorithm 1 for each pair is $\mathcal{O}((K \log K) + I + |\mathcal{D}|)$.

I. The Setting of ζ and Learning Rates

The network configuration has significant impact on the setting of ζ and learning rates. First, for the parameter ζ, when the number of D2D pairs increases, the interactions between D2D pairs will become more complex. Hence, to estimate the utility of each action accurately, each pair should keep a small value of ζ for a longer time to extend the duration of action exploration. In addition, when the number of subchannels increases, the number of actions for each D2D pair will increase. In this case, each pair should keep a small value of ζ for a longer time as well, since the utilities of more actions need to be estimated. Second, for the learning rate, it affects the size of the step for updates in reinforcement learning. When the number of pairs or subchannels increases, each pair should select learning rates that diminish more slowly with time. This is because the exploration time should be extended in both situations as discussed, and hence longer learning time is needed.

VI. Simulation Results and Analysis

In this section, simulation and numerical results are shown to evaluate the effectiveness of the proposed appoach, and demonstrate the benefits of D2D enabled C-RANs.

A. Scenarios and Parameters

The simulation scenario is shown in Fig. 2, where the axes are in units of meters, and each arrow starts from a D2D transmitter and points to the paired D2D receiver. We assume that two subchannels are available, and CUEs 1 and 2 transmit over subchannels 1 and 2, respectively. The channel coefficient of each link is composed of path loss and fast fading. The

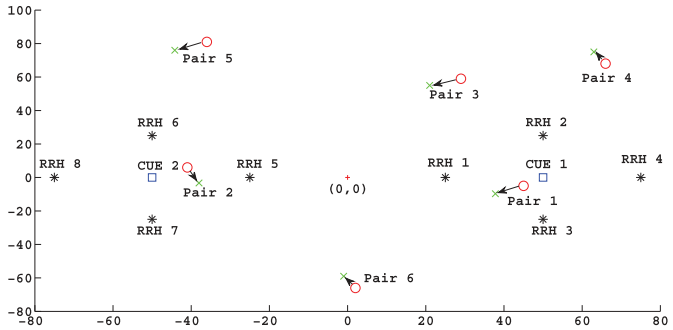


Fig. 2. The simulation scenario.

pathloss model is d^{-2} with d being the distance between two nodes in the same link, and the fast fading is modeled as an independent complex Gaussian random variable distributed as $\mathcal{CN}(0,1)$. Moreover, each UE can associate with at most four RRHs, and the maximal UE transmit power is 20 dBm. The number of UEs that each fronthaul link can serve is fixed as two, and the noise power is set to -104 dBm. The learning parameter is taken as $\zeta = t/1000$, and the learning rates $\theta(t)$ and $\gamma(t)$ are set to $\frac{1}{(t+1)^{0.6}}$ and $\frac{1}{(t+1)^{0.7}}$, respectively. Note that the reason for adopting the topology in Fig. 2 is to facilitate the analysis in Subsection B and the performance comparison with the optimal solution obtained by exhaustive search. When the number of subchannels is quite large, the number of actions of each D2D pair will be large, and hence it can take a long time for our proposed algorithm to converge to the logit equilibrium since each D2D pair has more actions to try. Faced with this problem, we can divide the whole area into multiple districts, and allocate these districts with orthogonal subsets of subchannels. In this way, the number of each pair's actions can be reduced significantly, and the interactions between D2D pairs in different districts are independent. Then, the proposed approach can be applied in each district.

For each pair, four different actions are available:

- *Action 1*: Operating in D2D mode over subchannel 1.
- *Action 2*: Operating in D2D mode over subchannel 2.
- *Action 3*: Operating in C-RAN mode over subchannel 1.
- *Action 4*: Operating in C-RAN mode over subchannel 2.

To show the advantages of enabling D2D in C-RANs, two schemes are considered:

- *Scheme 1*: D2D pairs can choose all the actions listed above.
- *Scheme 2*: D2D pairs can only choose *Actions* 3 and 4, which suggests the D2D communication is not allowed.

B. Convergence

Figs. 3 and 4 depict the convergence behavior of the proposed approach, which are obtained by a single simulation trial. Specifically, the evolution of the strategy of each pair is presented in Fig. 3. Initially, pairs play actions with equal probability, but however, after exploring the action spaces for a period of time, each pair takes the action with the highest estimated utility.

Fig. 4 shows the evolution of the system SE, where the optimal system performances under *Schemes* 1 and 2 are obtained by exhaustive search. In the conducted trial, the SE achieved by the reinforcement learning process first increases with the number of iterations, and then a near-optimal performance is achieved, which demonstrates the effectiveness of the proposed approach. Moreover, it can be seen that the performance of the C-RAN is greatly improved by allowing direct communication between UEs. Note that it is common for the reinforcement learning process to take many iterations to converge, e.g. 2000 iterations, even for a small-scale network as shown in [30, Fig. 2].

TABLE II

RESULTS UNDER *Scheme* 1

D2D Pair	Communication Mode	Selected Subchannel	Transmit Power	Associated RRHs
1	C-RAN mode	2	100 mW	1,2,3,4
2	C-RAN mode	1	100 mW	5,6,7,8
3	D2D mode	1	100 mW	
4	D2D mode	2	100 mW	
5	D2D mode	2	100 mW	
6	D2D mode	1	100 mW	

TABLE III

RESULTS UNDER *Scheme* 2

D2D Pair	Communication Mode	Selected Subchannel	Transmit Power	Associated RRHs
1	C-RAN mode	2	100 mW	1,2,3,4
2	C-RAN mode	2	100 mW	6,8
3	C-RAN mode	2	0 mW	
4	C-RAN mode	1	0 mW	
5	C-RAN mode	2	0 mW	
6	C-RAN mode	1	100 mW	5,7

To give greater insight into the poor performance of *Scheme* 2, we list the mode selection, subchannel allocation, RRH association, and power control results when the reinforcement learning process converges under *Scheme* 1 and the optimal performance is achieved under *Scheme* 2 in Table II and III, respectively. Note that the transmit power of UEs in C-RAN mode that are not associated with any RRH are set to 0. Then, with the help of the two tables and Fig. 2, the following conclusions can be drawn. First, due to the fronthaul capacity constraints, not all the UEs can fully benefit from the centralized signal processing capability of the C-RAN. For example, in Table III, pairs 2 and 6 are only associated with two RRHs, and pairs 3, 4 and 5 are not even served by any RRH. Second, for the UEs that are not near enough to RRHs, D2D mode is more attractive because of the proximity of D2D receivers. For example, in Table II, D2D pair 6 operates in D2D mode instead of C-RAN mode.

C. The Effects of RRH Association and Power Control

Fig. 5 shows the RRH association evolution when all the pairs operate in C-RAN mode, where the RRH index 0 means all the pairs have no RRHs to associate with at the beginning. It can be seen that each RRH can only accommodate one UE, and each UE can access four RRHs at most, which shows that the fronthaul capacity and computing capability constraints strictly hold by our matching based RRH association process. Note that each RRH can serve only one UE because it has accommodated one traditional C-RAN UE, and meanwhile the RRH association of only three pairs are demonstrated because the association requests of the other three pairs during the entire matching process are all rejected. As for the power

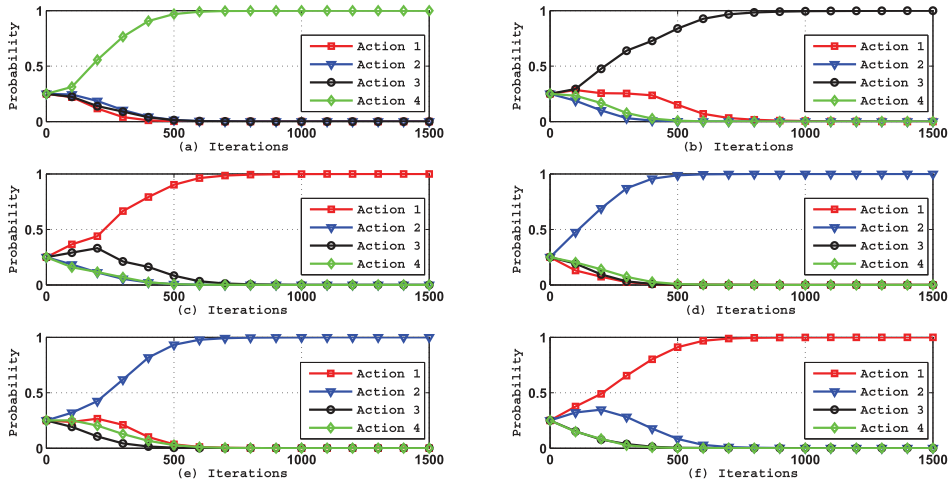


Fig. 3. The evolution of the D2D pairs' strategies with Figs. (a)-(f) corresponding to the D2D pairs 1-6, respectively.

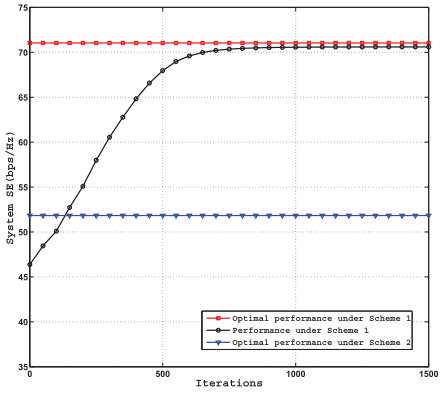


Fig. 4. The evolution of the overall SE.

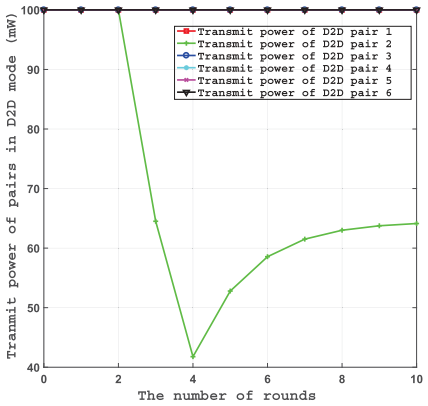


Fig. 6. The evolution of the transmission power.

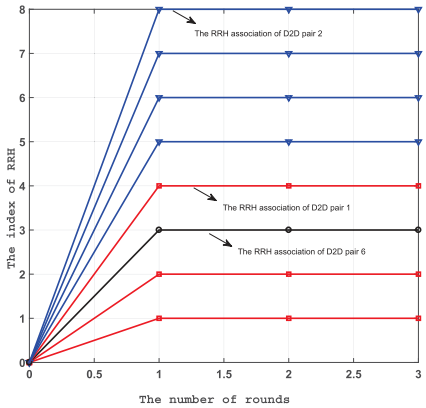


Fig. 5. The evolution of RRH association.

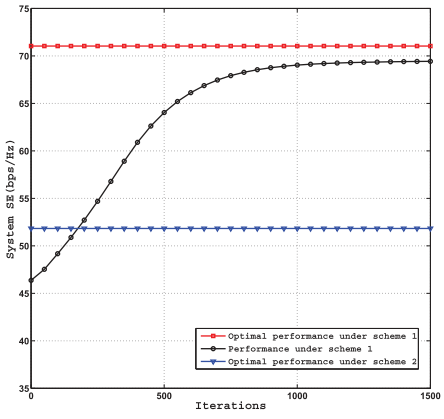


Fig. 7. The benefits of enabling D2D in C-RANs with constrained fronthaul capacity.

control algorithm, we consider the case in which all the D2D pairs select subchannel 2 and operate in D2D mode. Fig. 6 shows the evolution of the transmission power of D2D pairs. It can be seen that the transmission power of pair 2 is greatly suppressed because it causes most of the interference to CUE 2 occupying subchannel 2.

D. Enabling D2D Benefit

Fig. 7 highlights the benefit of enabling D2D in C-RANs by comparing the SE achieved under *Schemes* 1 and 2, and the results are obtained by conducting 500 independent trials and then taking the average value. It is reasonable to find that the SE achieved under *Scheme* 1 is significantly higher than the optimal performance under *Scheme* 2, since D2D mode is enabled in *Scheme* 1 and hence UEs can communicate with the paired receivers directly, which is preferable when the UEs are far from RRHs or RRHs in the neighborhood are not available due to the fronthaul capacity constraints. While in Fig. 8, the benefit of enabling D2D is shown when the

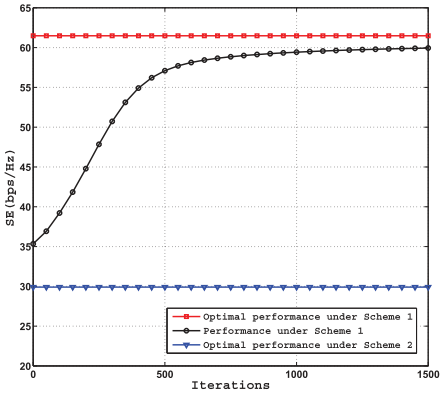


Fig. 8. The benefits of enabling D2D in C-RANs with constrained centralized signal processing capability.

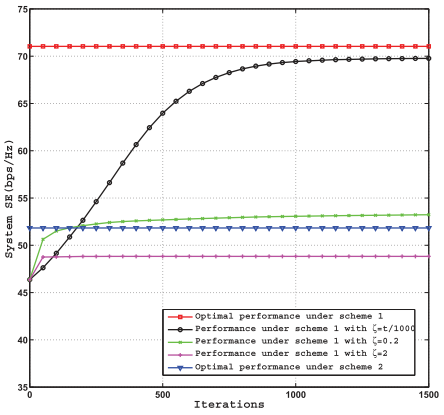


Fig. 9. The impacts of different learning parameters.

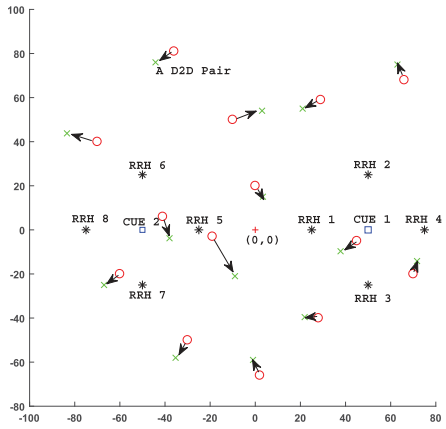


Fig. 10. The simulation scenario with 14 D2D pairs.

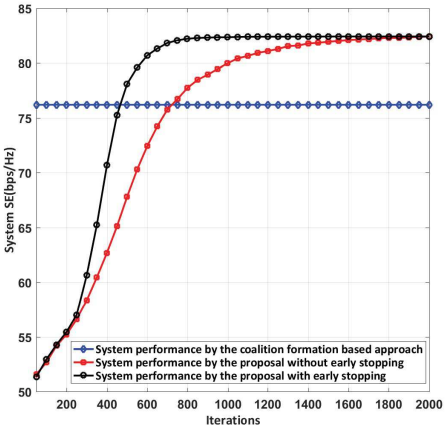


Fig. 11. The performance of the proposed appoach compared to a state-of-the-art approach.

centralized signal processing capability of the BBU pool is not powerful, where the fronthaul capacity is set to 7 and each D2D transmitter is only allowed to associate with one RRH. Not surprisingly, it can be seen that the optimal performance under *Scheme* 2 is much worse than the performance achieved under *Scheme* 1. This is because the advantage of C-RAN mode over D2D mode mainly comes from the centralized signal processing capability of the BBU pool.

E. The Impact of ζ

The impact of the learning parameter ζ on SE is evaluated in Fig. 9. Similar to Figs. 7 and 8, the results are obtained by conducting multiple independent trials and then taking the average value. In particular, it can be seen that either a large learning parameter or a small learning parameter leads to poor performance, while the parameter setting $\zeta = t/1000$ achieves a near-optimal performance. This verifies the discussion about learning parameter selection in Section V.

F. Performance in Large-Scale Networks

To show the performance of our algorithm in large scale networks, the network scenario in Fig. 10 is considered, which contains 14 D2D pairs. Since the computation time of exhaustive search for the optimal solution in this setting becomes prohibitive, we compare our algorithm with a coalition formation based mode selection and subchannel allocation algorithm proposed in [41], and the same RRH association and power control algorithms as in our proposed appoach are adopted for this comparison scheme. In addition, to reduce the number of iterations, early stopping is adopted [42], which sets the selection probability of an action to 1 if the updated probability is larger than 0.55. From Fig. 11, a performance improvement of 8.2% is observed, which demonstrates the effectiveness of our proposed appoach. Moreover, it can be seen that early stopping can effectively reduce the number of iterations for convergence, meanwhile achieving the same performance as that of the case without early stopping.

G. Overall Benefits

In this subsection, the benefits of our proposed distributed algorithm over the centralized scheme adopted in our simulation are discussed. The centralized case means the mode selection, subchannel allocation, user-RRH association, and power control are all optimized at the BBU pool. Specifically, mode selection and subchannel allocation are optimized based on exhaustive search, while user-RRH association and power control schemes are the same as those in our proposed appoach. Denote the computing resource

consumed by the centralized method by Δ. Since the complexity of exhaustion is $O\left((2\left|\mathcal{D}\right|)^{|\mathcal{N}|}\right)$, Δ can be modeled as $\lambda(2\left|\mathcal{D}\right|)^{|\mathcal{N}|}$ [26], where λ is a constant parameter. Denote the number of iterations of Algorithm 1 by ϖ, and then the computing cost saved by the proposed distributed approach is $\rho_1\left(1-\frac{\varpi}{(2|\mathcal{D}|)^{|\mathcal{N}|}}\right)\lambda(2\left|\mathcal{D}\right|)^{|\mathcal{N}|}=\rho_1\lambda\left((2\left|\mathcal{D}\right|)^{|\mathcal{N}|}-\varpi\right)$ with ρ_1 the price per unit computing resource.

Next, we analyze the loss of the distributed scheme incurred by information exchange in Algorithms 2 and 3. Based on the analysis of Algorithm 2, the total number of information exchanges in each round is at most $2|\mathcal{N}_C|V$. However in Algorithm 3, the information exchange is caused by broadcasting the interference prices corresponding to $|\mathcal{D}|$ subchannels via the high power node. Therefore, an upper bound on the information exchange cost can be calculated as $\rho_2\left(\varpi * T_{matching} * 2|\mathcal{N}_C|V+\varpi * |\mathcal{D}| * T_{stackelberg} * 1\right)$ with ρ_2 the price of a single information exchange, $T_{matching}$ the number of rounds needed for Algorithm 2 to reach a stable matching in the worst case, and $T_{stackelberg}$ the number of rounds needed for Algorithm 3 to reach Stackelberg equilibrium in the worst case. Thus, the net gain brought by the proposed distributed approach can be calculated as

$$\Gamma=\rho_1\lambda\left((2\left|\mathcal{D}\right|)^{|\mathcal{N}|}-\varpi\right)-\rho_2\left(\varpi * T_{matching} * 2|\mathcal{N}_C|V+\varpi * |\mathcal{D}| * T_{stackelberg} * 1\right), \quad (29)$$

and to achieve a positive Γ, ρ_1 and ρ_2 should satisfy the following relationship:

$$\rho_1 > \frac{T_{matching} * 2|\mathcal{N}_C|V+|\mathcal{D}| * T_{stackelberg}}{\lambda\left(\frac{(2|D|)^{|N|}}{\varpi}-1\right)}\rho_2. \quad (30)$$

In our simulation setting, $|\mathcal{D}|=2$, $|\mathcal{N}|=6$, $T_{matching}=3$ as shown in Fig. 6, $T_{stackelberg}=10$ as shown in Fig. 7, $|\mathcal{N}_C|=6$, $\lambda=10$, and $V=4$. Then, it can be seen that only if the price per unit computing resource is about 5 times larger than the price for single control information exchange, can the proposed distributed approach bring positive gain for the operator in our simulation scenario. Meanwhile, since Algorithm 1 can be further speeded up by transfer learning and early stopping, the condition (30) for a positive Γ will more possibly hold owing to the reduction of ϖ. In addition, with the network scale becoming large, the first term in (29) will increase exponentially, while the second term increases more slowly. Hence, it is anticipated that the distributed approach can achieve a more significant benefit in larger scale networks.

VII. Future Work

In this paper, considering that subchannel allocation and beamforming design in the downlink time slot will incur extra challenges, only the optimization for the uplink time slot of the proposed TDD based D2D enabled C-RAN has been considered, where D2D transmission and the uplink transmission of D2D pairs in C-RAN mode coexist. In the future, we plan to investigate joint uplink and downlink system performance optimization. Meanwhile, since the convergence of the distributed reinforcement learning does not rely on the specific form of the utility function, it can be used to optimize latency as well by replacing the current utilities of D2D pairs with utilities related to latency.

In addition, another interesting topic is to accelerate the convergence of distributed reinforcement learning to make it feasible for D2D communications and tackle the dynamics of the environment such as the dynamic change of C-RAN UEs' states. A potential solution is to involve transfer learning in the reinforcement learning process. Specifically, the author in [43] improves the actor-critic algorithm by transfer learning, which shows faster convergence by avoiding the agent having to learn from scratch. Inspired by [43], the utility estimates and strategies learned under previous environments can be taken as useful experience for the utility estimation and strategy estimation of the distributed learning process, respectively, considering that there will likely exist some similarity between the states of environments like C-RAN UEs' states in different time periods. In this way, when the state of the environment such as the C-RAN UEs' states change, D2D pairs will not learn from initial points and can achieve good performance with decreased learning time. Meanwhile, early stopping can be adopted as well to speed up the learning process.

VIII. Conclusion

A distributed approach to joint mode selection and resource allocation has been developed for an uplink device-to-device enabled cloud radio access network to optimize the overall spectral efficiency under the fronthaul capacity, centralized signal processing capability of the baseband unit pool, and inter-tier interference constraints. The approach contains three stages: communication mode and subchannel selection, utility value determination, and strategy update. The second stage includes a many-to-many matching based remote radio head association process and a Stackelberg game based power control process. Based on the feedback values of utilities, D2D pairs update their strategies using a reinforcement learning process. The proposed approach imposed only a light computing burden on each D2D pair, and meanwhile each pair and the BBU pool need to acquire only partial channel state information. The performance gain of enabling D2D in C-RANs has been confirmed by numerical simulations, and the gain depends on the distance between D2D transmitters and RRHs, the fronthaul capacity, as well as the centralized signal processing capability of the BBU pool.

References

[1] A. Checko *et al.*, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[2] Y. Zhong, T. Q. S. Quek, and W. Zhang, "Complementary networking for C-RAN: Spectrum efficiency, delay and system cost," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4639–4653, Jul. 2017.

[3] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 29, no. 2, pp. 6–14, Mar. 2015.

[4] Y. Liao, L. Song, Y. Li, and Y. A. Zhang, "How much computing capability is enough to run a cloud radio access network?" *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 104–107, Jan. 2017.

[5] Z. Yang, Z. Ding, and P. Fan, "Performance analysis of cloud radio access networks with uniformly distributed base stations," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 472–477, Jan. 2016.

[6] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, "System cost minimization in cloud RAN with limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371–3384, May 2017.

[7] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[8] P. Mach, Z. Becvar, and T. Vanek, "In-band device-to-device communication in OFDMA cellular networks: A survey and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1885–1922, 4th Quart., 2015.

[9] J. Liu, M. Sheng, T. Q. S. Quek, and J. Li, "D2D enhanced coordinated multipoint in cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4248–4262, Jun. 2016.

[10] W. Zhao and S. Wang, "Resource sharing scheme for device-to-device communication underlaying cellular networks," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4838–4848, Oct. 2015.

[11] T. D. Hoang, L. B. Le, and T. Le-Ngoc, "Energy-efficient resource allocation for D2D communications in cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 6972–6986, Sep. 2016.

[12] C. Xu *et al.*, "Efficiency resource allocation for device-to-device underlay communication systems: A reverse iterative combinatorial auction based approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 348–358, Sep. 2013.

[13] 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.

[14] H.-H. Nguyen *et al.*, "Distributed resource allocation for D2D communications underlay cellular networks," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 942–945, May 2016.

[15] H. ElSawy, E. Hossain, and M. S. Alouini, "Analytical modeling of mode selection and power control for underlay D2D communication in cellular networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4147–4161, Nov. 2014.

[16] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[17] J. Ye and Y. J. Zhang, "A guard zone based scalable mode selection scheme in D2D underlaid cellular networks," in *Proc. ICC*, London, U.K., Jun. 2015, pp. 2110–2116.

[18] C. Gao, J. Tang, X. Sheng, W. Zhang, S. Zou, and M. Guizani, "Enabling green wireless networking with device-to-device links: A joint optimization approach," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2770–2779, Apr. 2016.

[19] L. Su, Y. Ji, P. Wang, and F. Liu, "Resource allocation using particle swarm optimization for D2D communication underlay of cellular networks," in *Proc. WCNC*, Shanghai, China, Apr. 2013, pp. 129–133.

[20] H. Zhou, Y. Ji, J. Li, and B. Zhao, "Joint mode selection, MCS assignment, resource allocation and power control for D2D communication underlaying cellular networks," in *Proc. WCNC*, Apr. 2014, pp. 1667–1672.

[21] G. Yu, L. Xu, D. Feng, R. Yin, G. Y. Li, and Y. Jiang, "Joint mode selection and resource allocation for device-to-device communications," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3814–3824, Nov. 2014.

[22] M. Azam *et al.*, "Joint admission control, mode selection, and power allocation in D2D communication systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7322–7333, Sep. 2016.

[23] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul./Aug. 2016.

[24] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[25] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[26] M. Peng, D. Liang, Y. Wei, J. Li, and H.-H. Chen, "Self-configuration and self-optimization in LTE-advanced heterogeneous networks," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 36–45, May 2013.

[27] H. Zhang, Y. Wang, and H. Ji, "Resource optimization-based interference management for hybrid self-organized small-cell network," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 936–946, Feb. 2016.

[28] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Tech.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[29] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-Aho, "Dynamic clustering and on/off strategies for wireless small cell networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2164–2178, Mar. 2016.

[30] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, Aug. 2016.

[31] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[32] Y. Xing and R. Chandramouli, "Stochastic learning solution for distributed discrete power control game in wireless data networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 932–944, Aug. 2008.

[33] J. Wang *et al.*, "D-FROST: Distributed frequency reuse-based opportunistic spectrum trading via matching with evolving preferences," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3794–3806, 3rd Quart., 2018.

[34] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H. V. Poor, "Context-aware small cell networks: How social metrics improve wireless resource allocation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 5927–5940, Nov. 2015.

[35] M. Peng, X. Xie, Q. Hu, J. Zhang, and H. V. Poor, "Contract-based interference coordination in heterogeneous cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1140–1153, Jun. 2015.

[36] S. M. Perlaza, H. Tembine, and S. Lasaulce, "How can ignorant but patient cognitive terminals learn their strategy and utility?" in *Proc. SPAWC*, Marrakesh, Morocco, Jun. 2010, pp. 1–5.

[37] E. J. Collins and D. S. Leslie, "Convergent multiple-timescales reinforcement learning algorithms in normal form games," *Ann. Appl. Probab.*, vol. 13, pp. 1231–1251, Feb. 2002.

[38] J. Hofbauer and W. H. Sandholm, "Evolution in games with randomly disturbed payoffs," *J. Econ. Theory*, vol. 132, no. 1, pp. 47–69, 2007.

[39] M. Sotomayor, "Three remarks on the many-to-many stable matching problem," *Math. Social Sci.*, vol. 38, no. 1, pp. 55–70, 1999.

[40] K. W. Shum, K. K. Leung, and C. W. Sung, "Convergence of iterative waterfilling algorithm for Gaussian interference channels," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 1091–1100, Aug. 2007.

[41] D. Wu, Y. Cai, R. Hu, and Y. Qian, "Dynamic distributed resource sharing for mobile D2D communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5417–5429, Oct. 2015.

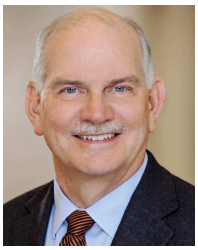
[42] C. Fan, B. Li, C. Zhao, W. Guo, and Y.-C. Liang, "Learning-based spectrum sharing and spatial reuse in mm-wave ultradense networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 4954–4968, Jun. 2018.

[43] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014.

Yaohua Sun (S'18) received the bachelor's degree (Hons.) in telecommunications engineering (with management) from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communications (Ministry of Education). His research interests include game theory, resource management, deep reinforcement learning, network slicing, and fog radio access networks. He was a recipient of the National Scholarship in 2011 and 2017. He has been a reviewer for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and IEEE COMMUNICATIONS LETTERS.

Mugen Peng (M'05–SM'11) received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. He joined BUPT, where he has been a Full Professor with the School of Information and Communication Engineering since 2012. In 2014, he was an Academic Visiting Fellow with Princeton University, Princeton, NJ, USA. He leads a Research Group focusing on wireless transmission and networking technologies with the Key Laboratory of Universal Wireless Communications (Ministry of Education), BUPT. He has authored and co-authored over 60 refereed IEEE journal papers and over 200 conference proceeding papers. His main research areas include wireless communication theory, radio signal processing, and convex optimizations, with a particular interests in cooperative communication, self-organization networking, heterogeneous networking, cloud communication, and Internet of Things.

Dr. Peng is a fellow of the IET. He was a recipient of the 2018 Heinrich Hertz Prize Paper Award, the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the Best Paper Award in the IEEE WCNC 2015. He is on the Editorial/Associate Editorial Board of the *IEEE Communications Magazine*, IEEE ACCESS, IEEE INTERNET OF THINGS JOURNAL, *IET Communications*, and *China Communications*. He has been a Guest Leading Editor for special issues of IEEE WIRELESS COMMUNICATIONS.



H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in electrical engineering and computer science from Princeton University in 1977. From 1977 to 1990, he was a faculty member at the University of Illinois at Urbana–Champaign. Since 1990, he has been a faculty member at Princeton, where he is currently the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he was the Dean of Princeton's School of Engineering and Applied Science. He has held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Information Theoretic Security and Privacy of Information Systems* (Cambridge University Press, 2017).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences. He is also a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the Marconi Award and the Armstrong Award from the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, Honorary Professorships at Peking University and Tsinghua University, both conferred in 2017, and a D.Sc. *honoris causa* from Syracuse University in 2017.