Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 4565203, 18 pages
http://dx.doi.org/10.1155/2016/4565203

Research Article

Dynamic Resource Allocation with Integrated Reinforcement Learning for a D2D-Enabled LTE-A Network with Access to Unlicensed Band

Alia Asheralieva and Yoshikazu Miyanaga

Laboratory of Information Communication Networks, School of Information Science and Technology, Hokkaido University, Sapporo, Japan

Correspondence should be addressed to Alia Asheralieva; aasheralieva@gmail.com

Received 30 May 2016; Revised 8 September 2016; Accepted 16 October 2016

Academic Editor: Juan C. Cano

We propose a dynamic resource allocation algorithm for device-to-device (D2D) communication underlying a Long Term Evolution Advanced (LTE-A) network with reinforcement learning (RL) applied for unlicensed channel allocation. In a considered system, the inband and outband resources are assigned by the LTE evolved NodeB (eNB) to different device pairs to maximize the network utility subject to the target signal-to-interference-and-noise ratio (SINR) constraints. Because of the absence of an established control link between the unlicensed and cellular radio interfaces, the eNB cannot acquire any information about the quality and availability of unlicensed channels. As a result, a considered problem becomes a stochastic optimization problem that can be dealt with by deploying a learning theory (to estimate the random unlicensed channel environment). Consequently, we formulate the outband D2D access as a dynamic single-player game in which the player (eNB) estimates its possible strategy and expected utility for all of its actions based only on its own local observations using a joint utility and strategy estimation based reinforcement learning (JUSTE-RL) with regret algorithm. A proposed approach for resource allocation demonstrates near-optimal performance after a small number of RL iterations and surpasses the other comparable methods in terms of energy efficiency and throughput maximization.

1. Introduction

D2D communication is a direct communication between the users transmitting over the cellular spectrum (inband) or operating on an unlicensed band (i.e., outband). The main advantages of inband D2D communication are the increased spectrum efficiency and possibility of quality of service (QoS) provisioning for different cellular/D2D users. The chief obstacles to the implementation of inband D2D access are (i) interference mitigation (between the users transmitting over the same frequency bands) and (ii) resource allocation [1]. Effective resource allocation and interference management strategies can significantly improve the performance of cellular networks. The objectives here could be different (such as improvement of spectrum efficiency, cellular coverage, network throughput, or user experience) but to achieve the optimal system performance, the problems of cellular/D2D mode selection, spectrum assignment, power allocation, and interference mitigation should be considered jointly in the algorithm design. Related contributions in this area are [2–10] studying the problem of interference mitigation for underlying D2D communication. It should be noted, however, that the majority of proposed formulations (except [2, 3]) does not deal with the issues of mode selection, spectrum assignment, and interference management in a joint fashion but rather by splitting the original problem into smaller subproblems (see e.g., [10]) or by separating the time scales of these subproblems (e.g., [9]). Hence, although the complexity of such methods is less than the complexity of a joint resource allocation, their efficiency in maximizing some certain optimality criterion is clearly downgraded. Outband D2D communication (carried over Wi-Fi Direct [11], ZigBee [12], or Bluetooth [13]) eliminates the need for interference mitigation but can be distorted by

the randomness of unlicensed channels. Existing works on outband D2D access focus on such issues as power consumption (e.g., [14–17]) and coordination between cellular and wireless interfaces ([18–21]). Some of these works ([14, 15, 21]) suggest control of unlicensed band by the cellular network (which requires a certain amount of cooperation and information exchange between different radio interfaces). Other works (e.g., [17, 18, 20]) imply autonomous operation of D2D devices (based on stochastic modeling of unlicensed channels).

The main contributions of this work are as follows. We consider a network-controlled D2D communication in which the licensed and unlicensed spectrum resources, user modes, and transmission power levels are allocated to different device pairs by the LTE eNB to maximize the overall network utility. We consider a general network deployment scenario where the unlicensed band is assumed to be provided by one or more radio access technologies (RATs) based on the orthogonal frequency division multiple access (OFDMA), carrier sense multiple access with collision avoidance (CSMA/CA), frequency-hopping code division multiple access (FH-CDMA), or any other multiple access method. It is assumed that all device pairs are equipped with different wireless interfaces allowing them to connect to the appropriate RAT and use a CSMA/CA to avoid collisions when operating on the unlicensed band. Hence, each unlicensed channel becomes available to a D2D pair only when it is idle. Unlike many previous works, we jointly solve the problems of inband/outband access, mode selection, and spectrum/power assignment by combining these problems into one optimization problem which allows to allocate the inband network resources and offload the D2D traffic in a most effective way (in terms of maximizing the overall network utility). Note that the formulated problem can be solved to optimality only if the global channel and network knowledge (including the precise information on the operating conditions of the licensed and unlicensed channels) is available to the eNB. However, because of the absence of an established control link between the unlicensed and cellular radio interfaces, the eNB cannot get any information about the quality and availability of the unlicensed channels. As a result, a considered resource allocation problem becomes a stochastic optimization problem that can be dealt with by deploying a learning theory [22] (to estimate the random unlicensed channel environment).

Consequently, we formulate the outband D2D access as a dynamic single-player game in which the player (eNB) estimates its possible strategy and expected utility for all of its actions based only on its own local observations using a JUSTE-RL with regret (originally proposed in [23]). The main idea behind RL is that the actions leading to the higher network utility at the current stage should be granted with higher probabilities at the next stage [22]. In the simplest form of RL (described, e.g., in [24]), a learning agent estimates its best strategy based on its observed utility without any prior information about its operating environment. This form of RL requires only algebraic operations but its convergence to the equilibrium state is not guaranteed [25]. In Q-learning [22], a utility is estimated using some value-action function. This RL method converges to a Nash equilibrium (NE) state. However, it requires maximization of the action-value at every stage which can be computationally demanding [22]. In JUSTE-RL algorithm (described, in detail, in [23]), a learning agent estimates not only its own strategy but also the expected utility for all of its actions. Unlike Q-learning, JUSTE-RL does not need to perform optimization of the action-value (since only algebraic operations are required to update the strategies) and, hence, it has a lower computational complexity. On the other hand, compared to a basic RL algorithm, JUSTE-RL converges to a ε-NE [23, 25].

It is worth mentioning that, in wireless communications, RL has been studied in the context of various spectrum access problems. In [26, 27], the learning has been employed to minimize the interference (created by adjacent nodes) in partially overlapping channels. This problem has been formulated as the exact potential graphical game admitting a pure-strategy NE and, therefore, the proposed approach is not realizable in a broader range of problems. A cognitive network with multiple players has been analyzed in [28]. In this work, the learning and channel selection have been separated into two different procedures which increased the complexity of a proposed resource allocation approach. Besides, the stability of a final solution was not verified. A multi-player game for inband D2D access, where the players (D2D users) learn their optimal strategies based on the throughput performance in a stochastic environment, has been studied in [29]. It was assumed that each D2D user can transmit over the vacant cellular channels using a CSMA/CA implying that there are no channels with interfering users (i.e., each orthogonal channel can be occupied by at most one cellular/D2D user). Although the authors consider a scenario with two D2D users operating on the same channel, it is not clear how a D2D user can sense whether the user operating on the channel is cellular or D2D. An autonomous D2D access in heterogeneous cellular networks comprising multiple low-power and high-power BSs with (possibly) overlapping spectrum bands has been investigated in [30]. This problem has been modeled as a stochastic noncooperative game with multiple players (D2D pairs) admitting a mixed-strategy NE. The goal of each player was to jointly select the wireless channel and power level to maximize its reward, defined as the difference between the achieved throughput and the cost of power consumption constrained by the minimum tolerable SINR requirements of this D2D pair. To solve this problem, a fully autonomous multiagent Q-learning algorithm (which does not require any information exchange and/or cooperation among different users) is developed and implemented in an LTE-A network.

The rest of the paper is organized as follows. A general network model for inband and outband network operation is described in Section 2. A general problem and the algorithms for unlicensed and licensed resource allocation are formulated in Section 3. The algorithm implementation, including the proposed resource allocation procedure in an LTE-A networks and performance evaluation, is presented in Section 4. The paper is finalized in Conclusion.

2. Network Model

In this paper, the problem of resource allocation for D2D communication is investigated for both the uplink (UL) and downlink (DL) directions. Similarly, the discussion through the rest of the paper is applicable (if not stated otherwise) to either direction. Consider a basic LTE-A network consisting of one eNB and N user pairs, denoted $\mathrm{PU}_1, \ldots, \mathrm{PU}_N$, with $\mathbf{N} = \{1, \ldots, N\}$ being the set of user pairs' indices. It is assumed that a fixed licensed spectrum band of the eNB spans K resource blocks (RBs), numbered $\mathrm{RB}_1, \ldots, \mathrm{RB}_K$, with $\mathbf{K} = \{1, \ldots, K\}$ denoting the set of RBs' indices comprising the bandwidth. The network runs on a slotted-time basis with the time axis partitioned into equal nonoverlapping time intervals (slots) of the length T_s, with t denoting an integer-valued slot index. Each pair of users can communicate with each other either by the traditional cellular mode (CM) via the eNB or in a D2D mode (DM) without traversing the eNB. Let $\mathbf{C} \subseteq \mathbf{N}$ be the set of the indices of device pairs that can operate only in CM and let $\mathbf{D} = \mathbf{N} \setminus \mathbf{C}$ denote the set of the indices of potential D2D pairs (The indices in $\mathbf{C}$ and $\mathbf{D}$ can be determined based on, e.g., user application (such as video sharing, gaming, and proximity-aware social networking) in which the pair of devices could potentially be in range for the direct communication. Such information can be acquired from a standard session initiation protocol (SIP) procedure (which handles the session setups and users arrivals in LTE networks). Interested readers are referred to [31] for a comprehensive description of an SIP procedure and its use in the D2D access.).

In our network, any potential D2D pair can be allocated with cellular or D2D mode (based on the results of resource allocation procedure). Consequently, we define a binary mode allocation variable $c_n(t)$, $n \in \mathbf{N}$, equaling 1, if PU_n is allocated CM at slot t, and 0, otherwise. Note that $c_n(t) = 1$, for all $n \in \mathbf{C}$. Further, we consider the following models of D2D access.

(i) Inband D2D: a D2D pair operates within the licensed LTE spectrum in an underlay to cellular communication.

(ii) Outband D2D: a D2D pair transmits over the unlicensed band by exploiting other RATs, such as Wi-Fi Direct [11], ZigBee [12], or Bluetooth [13] (It is assumed that all user devices are equipped with the corresponding wireless interfaces to be able to communicate using a suitable RAT.). We assume that there is no coordination and/or information exchange between different wireless interfaces.

To differentiate the pairs according to their D2D access, we define a binary channel access variable $b_n(t)$, $n \in \mathbf{N}$, equaling 1, if PU_n operates inband at slot t, and 0, otherwise. Note that all cellular users can access only the LTE bands. Hence, $b_n(t) = 1$, for all $n \in \mathbf{C}$.

2.1. Inband Network Operation. In LTE/LTE-A, RBs are allocated to cellular users by the eNBs using a standard packet scheduling procedure [32]. The use of packet scheduling in a D2D-enabled LTE-A network is described, in detail, in [33]. In short, a packet scheduling process can be explained as follows. In the UL direction, at the beginning of any slot t, each user is required to collect and transmit its buffer status information. After collecting this data, a user sends the scheduling request (SR) with its buffer status information to the eNB via a dedicated physical uplink control channel (PUCCH). After receiving all the SRs, the eNB allocates the RBs to the users (according to a certain scheduling algorithm) and responds to all the SRs by sending the scheduling grants (SGs) together with the allocation information to the corresponding users via dedicated physical downlink control channels (PDCCHs) [33]. In the DL, the eNB readily finds out the DL buffer status for each user, allocates the RBs, and sends the SGs with allocation information via PDCCHs [33]. In the framework used in this paper, the above scheduling process is applied for both the cellular and D2D communication with some modifications (the corresponding resource allocation procedure will be described in Section 4).

Let us further define a binary RB allocation variable $a_n^k(t)$, $n \in \mathbf{N}$, $k \in \mathbf{K}$, equaling 1, if PU_n is allocated with RB_k at slot t, and 0, otherwise. Each RB can be allocated to at most one cellular user. Hence,

$$\sum_{n \in \mathbf{N}} a_n^k(t)\, b_n(t)\, c_n(t) \leq 1, \quad \forall k \in \mathbf{K}. \tag{1a}$$

The number of D2D users operating on the same RBs is unlimited. Additionally, to maximize the network utilization, we enforce each RB to be allocated to at least one user. That is,

$$\sum_{n \in \mathbf{N}} a_n^k(t)\, b_n(t) \geq 1, \quad \forall k \in \mathbf{K}. \tag{1b}$$

Note that both the OFDMA used for DL transmissions and single carrier frequency division multiple access (SC-FDMA) applied in the UL direction provide orthogonality of resource allocation to cellular communications. This allows achieving a minimal level of cochannel interference between the transmitter-receiver pairs located within one cell [34]. Thus, when information is transmitted by cellular/D2D user, it will be distorted only by the users operating on the same RB(s).

Let G_{nm}^k, $n \in \mathbf{N}$, $m \in \mathbf{N}$, and $k \in \mathbf{K}$, denote the channel gain coefficient between the transmitter and receiver of PU_n and PU_m operating on RB_k (for $n \in \mathbf{C}$, G_{nn}^k indicates the channel gain coefficient between PU_n operating on RB_k and the eNB). In LTE system, the instantaneous values of G_{nm}^k can be obtained from the channel state information (CSI) through the use of special reference signals (RSs) [35] and, hence, they are known to the eNB and the users. Then, for any PU_n operating on RB_k, the SINR at slot t in the UL direction is described by

$$\mathrm{SINR}_n^k(t) = \frac{a_n^k(t)\, p_n(t)\, G_{nn}^k}{\sum_{j \in \mathbf{N} \setminus \{n\}} a_n^k(t)\, a_j^k(t)\, p_n(t)\, G_{jn}^k + N_0}, \quad \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}, \tag{2}$$

where N_0 is the variance of zero-mean additive white Gaussian noise (AWGN) power and $p_n(t)$ is the transmission

power allocated to PU_n at slot t. Clearly, $p_n(t)$ is nonnegative and cannot exceed some predefined maximal level $P_n^{\max}$; that is

$$0 \le p_n(t) \le P_n^{\max}, \quad \forall n \in \mathbf{N}. \tag{3}$$

At any t, the inband service rate of PU_n depends on the number of RBs allocated to this device pair and the SINR in each RB. That is,

$$r_n^L(t) = \omega^L \sum_{k \in \mathbf{K}} a_n^k(t) \log\left(1 + \text{SINR}_n^k(t)\right), \quad \forall n \in \mathbf{N}, \tag{4}$$

where $r_n^L(t)$ is the service rate of PU_n (in bits per slot or bps) over licensed (inband) spectrum and ω^L is the bandwidth of one LTE RB ($\omega = 180$ kHz).

2.2. Outband Network Operation. We consider M separate outband wireless channels numbered, for notation consistency, as $C_{K+1}, \ldots, C_{K+M}$ (In this paper, we consider a general scenario when the unlicensed outband access can be based on OFDMA, CSMA/CA (in case of Wi-Fi Direct), FH-CDMA (in case of Bluetooth), or any other multiple access method.). We denote by $\mathbf{M} = \{K+1, \ldots, K+M\}$ the set of channel indices within the unlicensed band and use a binary channel allocation variable $a_n^m(t)$, $n \in \mathbf{N}$, $m \in \mathbf{M}$, to indicate if PU_n is allocated with the unlicensed channel C_m (in which case, $a_n^m(t) = 1$) or not ($a_n^m(t) = 0$). Note that $a_n^m(t) = 0$, for all $n \in \mathbf{C}$ and $m \in \mathbf{M}$ (since cellular users can access only the LTE bands). For $n \in \mathbf{D}$, $b_n(t)$ equals 0, if $\sum_{m \in \mathbf{M}} a_n^m(t) \ge 1$, and 1, otherwise (i.e., if $\sum_{m \in \mathbf{M}} a_n^m(t) = 0$). Hence,

$$b_n(t) = \max\left\{0,\ 1 - \sum_{m \in \mathbf{M}} a_n^m(t)\right\}, \quad \forall n \in \mathbf{D}. \tag{5a}$$

To avoid collisions, the D2D pairs use a CSMA/CA method when operating outband. As a result, each unlicensed channel C_m is available to D2D communication only when it is idle. Additionally, to reduce the possibility of collisions between D2D users, we assume that, at any slot t, at most one device pair can transmit over each unlicensed channel C_m. That is,

$$\begin{aligned} &\sum_{n \in \mathbf{D}} a_n^m(t) \le 1, \\ &\sum_{n \in \mathbf{D}} \left(1 - b_n(t)\right) \le M, \\ &\qquad\qquad \forall m \in \mathbf{M}. \end{aligned} \tag{5b}$$

The transmission procedure for the pair of D2D users operating outband is described as follows. At the beginning of slot t, one of the users starts sensing the allocated unlicensed channel C_m (for simplicity, we assume perfect sensing). If the channel is free, the transmission phase (of the length T_r^m, such that $0 \le T_r^m \le T_s$) begins. Note that the duration of T_r^m is random. It depends on the availability of the channel C_m and the applied CSMA/CA scheme. The probability density function (p.d.f.) of T_r^m is not calculated here (since it has no impact on the further analysis in this paper). An example of such calculations can be found in [36].

Let G_{nn}^m, for all $m \in \mathbf{M}$, denote the channel gain coefficient between the transmitter and receiver of PU_n operating on unlicensed channel C_m. Then, the SINR of PU_n transmitting over the channel C_m at slot t can be expressed by

$$\text{SINR}_n^m(t) = \frac{a_n^m(t)\, p_n(t)\, G_{nn}^m}{N_0}, \quad \forall n \in \mathbf{D},\ \forall m \in \mathbf{M} \tag{6a}$$

and the service rate of PU_n over unlicensed (outband) spectrum is described by

$$\begin{aligned} r_n^U(t) = \sum_{m \in \mathbf{M}} \frac{T_r^m \omega_m^U}{T_s} a_n^m(t) \log\left(1 + \text{SINR}_n^m(t)\right), \\ \forall n \in \mathbf{D}, \end{aligned} \tag{6b}$$

where ω_m^U is the bandwidth (in Hz) of unlicensed channel C_m. Note that neither the eNB nor D2D users have prior information about quality and availability of unlicensed channels. Therefore, the exact values of G_{nn}^m and T_r^m are unknown to the eNB and the D2D users.

3. Resource Allocation Problem

3.1. Problem Statement. We define a binary $N \times K$-dimensional RB allocation matrix $\mathbf{a}^L$ and a binary $N \times M$-dimensional unlicensed channel allocation matrix $\mathbf{a}^U$ as

$$\begin{aligned} \mathbf{a}_t^L &= \begin{bmatrix} a_1^1(t) & \cdots & a_1^K(t) \\ \vdots & \vdots & \vdots \\ a_N^1(t) & \cdots & a_N^K(t) \end{bmatrix}, \\ \mathbf{a}_t^U &= \begin{bmatrix} a_1^{K+1}(t) & \cdots & a_1^{K+M}(t) \\ \vdots & \vdots & \vdots \\ a_N^{K+1}(t) & \cdots & a_N^{K+M}(t) \end{bmatrix}, \end{aligned} \tag{7}$$

respectively. We also define a binary N-dimensional D2D access allocation vector $\mathbf{b}_t = (b^1(t), \ldots, b^N(t))$, a binary N-dimensional mode allocation vector $\mathbf{c}_t = (c^1(t), \ldots, c^N(t))$, and a real-valued N-dimensional power allocation vector $\mathbf{p}_t = (p^1(t), \ldots, p^M(t))$. Then, the sets of all admissible values for $\mathbf{a}_t^L$, $\mathbf{a}_t^U$, $\mathbf{b}_t$, $\mathbf{c}_t$, and $\mathbf{p}_t$ are described by

$$\mathbf{A}^L = \left\{\mathbf{a}_t^L \mid a_n^k(t) \in \{0,1\},\ \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}\right\}; \tag{8a}$$

$$\begin{aligned} \mathbf{A}^U = \Bigg\{ &\mathbf{a}_t^U \mid a_n^m(t) \in \{0,1\},\ a_i^m(t) = 0,\ \sum_{j \in \mathbf{D}} a_j^m(t) \\ &\le 1\ \forall n \in \mathbf{D},\ \forall i \in \mathbf{C},\ \forall m \in \mathbf{M} \Bigg\}; \end{aligned} \tag{8b}$$

$$\begin{aligned} \mathbf{B} = \Bigg\{ &\mathbf{b}_t \mid b_n(t) \in \{0,1\},\ b_i(t) = 1,\ \sum_{j \in \mathbf{D}} \left(1 - b_j(t)\right) \\ &\le M\ \forall n \in \mathbf{D},\ \forall i \in \mathbf{C} \Bigg\}; \end{aligned} \tag{8c}$$

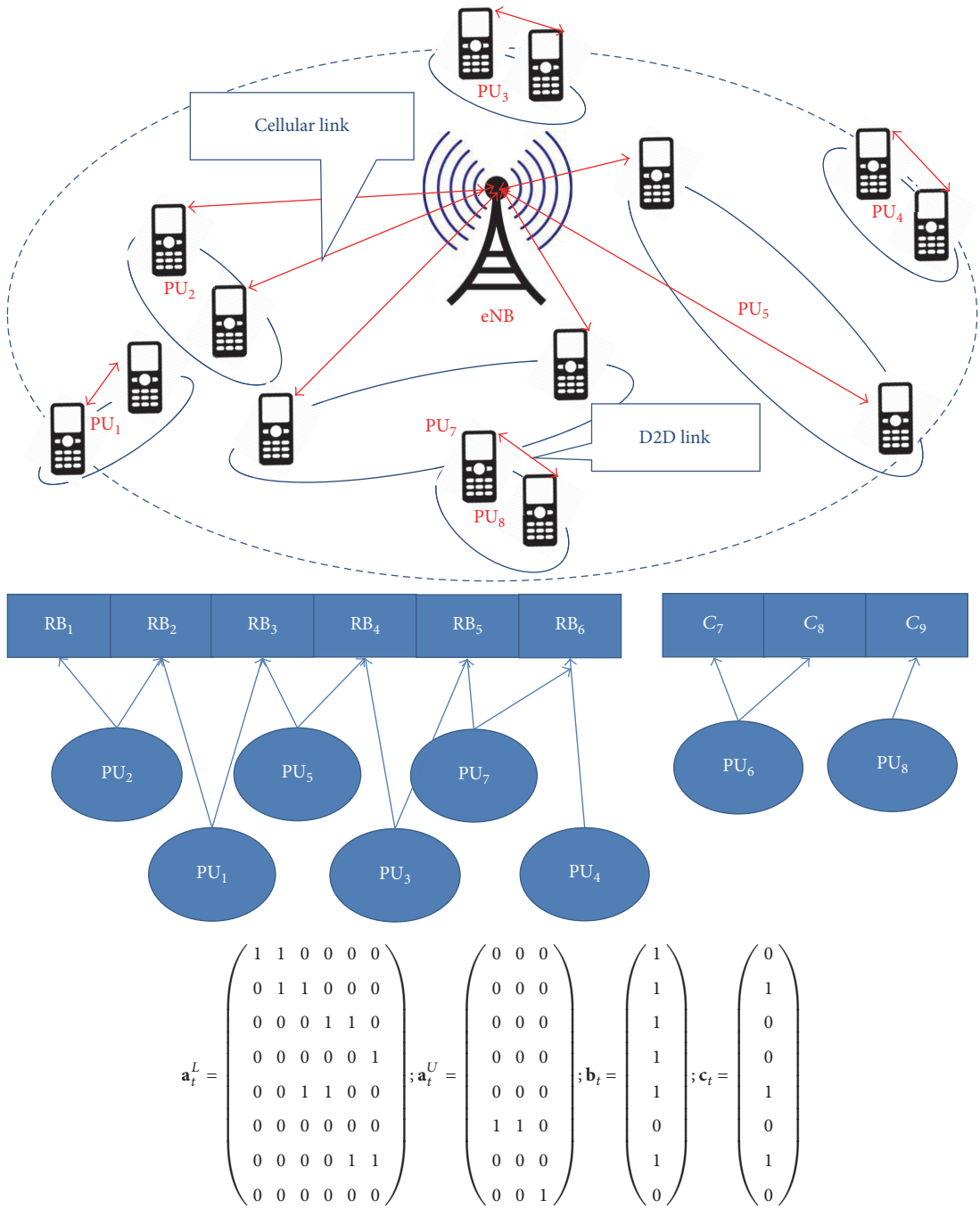

FIGURE 1: A D2D-enabled cellular network with three cellular pairs (PU_2, PU_5, and PU_7) and five D2D pairs (PU_1, PU_3, PU_4, PU_6, and PU_8). Three of the D2D pairs (PU_1, PU_3, and PU_4) use inband access and two D2D pairs (PU_6, PU_8) are allocated with the unlicensed channels. In this example, different cellular and D2D pairs interfere with each other when transmitting over RB_2 (PU_2 interferes with PU_1), RB_3 (PU_1 interferes with PU_5), RB_4 (PU_5 interferes with PU_3), RB_5 (PU_7 interferes with PU_3), and RB_6 (PU_4 interferes with PU_7).

$$\mathbf{C} = \left\{\mathbf{c}_t \mid c_n(t) \in \{0,1\},\ c_i(t) = 1,\ \forall n \in \mathbf{N},\ \forall i \in \mathbf{C}\right\}; \quad (8d)$$

$$\mathbf{P} = \left\{\mathbf{p}_t \mid 0 \le p_n(t) \le P_n^{\max},\ \forall n \in \mathbf{N}\right\}. \quad (8e)$$

Example of a D2D-enabled network with all defined optimization variables is shown in Figure 1.

Ideally, at any slot t, the eNB should distribute the network resources among the users to maximize their aggregated service rate. That is, to maximize the sum:

$$\sum_{n\in\mathbf{N}} r_n(t) = \sum_{n\in\mathbf{N}} \left(b_n(t)\, r_n^L(t) + (1 - b_n(t))\, r_n^U(t)\right), \quad (9)$$

where $r_n(t)$ represents the service rate of PU_n (operating either inband or outband). However, when communicating

over the unlicensed spectrum, each D2D pair should transmit at a maximal power level to achieve the high SINR regime (and, consequently, service rate) which, in turn, results in increased power consumption of mobile terminals. Therefore, when formulating the utility of each device pair, we should also consider the cost of power consumption, to quantify the trade-off between the achieved rate and power level (as in [37]). Accordingly, we can define a utility $u_n(t)$ of PU_n at slot t, as the difference between its instantaneous service rate $r_n(t)$ and the cost of power consumption:

$$\begin{aligned} u_n(t) &= r_n(t) - \upsilon_n p_n(t) \\ &= b_n(t)\, r_n^L(t) + \left(1 - b_n(t)\right) r_n^U(t) - \upsilon_n p_n(t), \end{aligned} \tag{10}$$

where $\upsilon_n \geq 0$ is the cost per unit (W) level of power for PU_n.

Using the above definition, we can express our resource allocation problem as follows:

$$\text{maximize} \quad \sum_{n\in\mathbf{N}} u_n(t) = \sum_{n\in\mathbf{N}} \left[b_n(t)\, r_n^L(t) + \left(1 - b_n(t)\right) r_n^U(t) - \upsilon_n p_n(t) \right], \tag{11a}$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{a}_t^L \in \mathbf{A}^L, \\ & \mathbf{a}_t^U \in \mathbf{A}^U, \\ & \mathbf{b}_t \in \mathbf{B}, \\ & \mathbf{c}_t \in \mathbf{C}, \\ & \mathbf{p}_t \in \mathbf{P}, \end{aligned} \tag{11b}$$

$$\sum_{n\in\mathbf{N}} a_n^k(t)\, b_n(t)\, c_n(t) \leq 1, \quad \forall k \in \mathbf{K}, \tag{11c}$$

$$\sum_{n\in\mathbf{N}} a_n^k(t)\, b_n(t) \geq 1, \quad \forall k \in \mathbf{K}, \tag{11d}$$

$$b_n(t) = \max\left\{0,\ 1 - \sum_{m\in\mathbf{M}} a_n^m(t)\right\}, \quad \forall n \in \mathbf{D}, \tag{11e}$$

$$\mathrm{SINR}_n^k(t) \leq \mathrm{SINR}_n^{\mathrm{tar}}, \quad \forall n \in \mathbf{N},\ \forall k \in \mathbf{K} \cup \mathbf{M}, \tag{11f}$$

where the constraint (11f) is necessary to protect the users from heavy interference (here $\mathrm{SINR}_n^{\mathrm{tar}}$ stands for the minimal SINR level acceptable by PU_n). Note that information on the sets $\mathbf{C}$ and $\mathbf{D}$ is readily available at the eNB. The values of G_{nm}^k for $n \in \mathbf{N}$, $m \in \mathbf{N}$ and $k \in \mathbf{K}$, are obtained by the eNB from the CSI carried by the RSs. The only missing information is related to $r_n^U(t)$ that depends on the parameters T_r^m (representing the availability of the unlicensed channel C_k in our model) and G_{nn}^m (which defines the quality of unlicensed channel C_k), for all $m \in \mathbf{M}$. The latter parameter is determined by the unlicensed channel allocations and, hence, the eNB can adapt to the changes of G_{nn}^m in time and space. Since there is no coordination (and no information exchange) between the LTE and outband RAT interfaces, solving (11a)–(11f) to optimality might be impossible, which is a rather strong argument in favor of applying a well-known reinforcement learning (RL) for resource allocation.

The main idea behind RL is that the actions (unlicensed channel allocations) leading to the higher network utility at slot t should be granted with higher probabilities at slot $t+1$ and vice versa [22]. In the simplest form of RL (presented in [24]), the learning agent estimates its possible strategies based on the locally observed utility without any prior information about the operating environment. This form of RL requires only algebraic operations but does not guarantee the convergence to an equilibrium [25]. In Q-learning [22], the agent's utility is estimated using some value-action function. Given the certain (easy to follow) conditions, this algorithm converges (with probability 1) to an NE state. However, it requires maximization of the action-value at every slot t (which can be computationally demanding depending on the structure of a chosen value-action function) [20]. In JUSTE-RL algorithm [23], the learning agent estimates not only its own strategy but also the expected utility for all of its actions. Unlike Q-learning, JUSTE-RL does not need to perform optimization of the action-value (since only algebraic operations are required to update the strategies) and, hence, it has a lower computational complexity. On the other hand, compared to a basic RL algorithm, JUSTE-RL converges to a ε-NE [23, 25]. We now show how a JUSTE-RL with regret can be applied to our problem.

3.2. Unlicensed Channel Allocation. To apply JUSTE-RL with regret to our problem, we represent it as a game with one player (the eNB) having no information about the operating environment. A finite set of the eNB's actions $\mathbf{A}^U$ represents the set of all admissible unlicensed channel allocation decisions. The objective of the eNB is to select, at any slot t, an

action $A_t = \mathbf{a}_t^U \in \mathbf{A}^U$ to maximize the eNB's utility $u_t = \sum_{n\in\mathbf{N}} u_n(t)$. In the following, we use notation $\overline{a}_n^m(t)$, $\forall m \in \mathbf{M}$, to specify the eNB's decision regarding the allocation of an unlicensed channel C_k to a pair PU_n and $\overline{\mathbf{a}}_t^U$ to describe all unlicensed channel allocations by the eNB when selecting a particular action A_t at slot t. We also use $\overline{\mathbf{b}}_t$ to denote the D2D access allocation vector and $\overline{r}_n^U(t)$ to indicate the outband service rate achieved by playing the action A_t. After taking an action $A_t = \overline{\mathbf{a}}_t^U$ at slot t, the eNB observes the (random) service rate $\overline{r}_n^U(t)$ and estimates the network utility $\overline{u}_t = \overline{u}_t(A_t)$ by solving the following problem:

$$\text{maximize} \quad u_t = \sum_{n\in\mathbf{N}} \left[\overline{b}_n(t)\, r_n^L(t) + \left(1 - \overline{b}_n(t)\right)\overline{r}_n^U(t) - \upsilon_n p_n(t)\right], \tag{12a}$$

$$\text{subject to} \quad \mathbf{a}_t^L \in \mathbf{A}^L,$$

$$\mathbf{c}_t \in \mathbf{C}, \tag{12b}$$

$$\mathbf{p}_t \in \mathbf{P},$$

$$\sum_{n\in\mathbf{N}} a_n^k(t)\,\overline{b}_n(t)\,c_n(t) \le 1, \quad \forall k \in \mathbf{K}, \tag{12c}$$

$$\sum_{n\in\mathbf{N}} a_n^k(t)\,\overline{b}_n(t) \ge 1, \quad \forall k \in \mathbf{K}, \tag{12d}$$

$$\text{SINR}_n(t) \le \text{SINR}_n^{\text{tar}}, \quad \forall n \in \mathbf{N}, \tag{12e}$$

where

$$\overline{b}_n(t) = \max\left\{0, 1 - \sum_{m\in\mathbf{M}} \overline{a}_n^m(t)\right\}, \quad \forall n \in \mathbf{D} \tag{12f}$$

and $\overline{b}_n(t) = 0, \forall n \in \mathbf{C}$. Note that, unlike problem (11a)–(11f), problem (12a)–(12e) can be solved to optimality (since $\overline{r}_n^U(t)$ is known). It has three optimization variables $\mathbf{a}_t^L$, $\mathbf{c}_t$, and $\mathbf{p}_t$ and, hence, its complexity is lower than the complexity of (11a)–(11f) (the method for solving (12a)–(12e) is presented in the next subsection).

We also define a mixed-strategy probability π_t of playing an action $A_t = \overline{\mathbf{a}}_t^U$ at slot t as

$$\pi_t(A_t) = \{\pi_t(B)\}_{B\in\mathbf{A}^U},$$
$$\pi_t(B) = \Pr\{A_t = B\}, \tag{13a}$$
$$\forall A_t \in \mathbf{A}^U,$$

and a regret $\rho_t(A_t)$ for not playing this action at slot t as

$$\rho_t(A_t) = \max\{0, \overline{u}_{t-1} - \overline{u}_t\}, \quad \forall A_t \in \mathbf{A}^U. \tag{13b}$$

In JUSTE-RL, the probability distribution of a regret over all possible actions becomes the Boltzmann–Gibbs distribution (aka canonical ensemble), given by [22]

$$G\{\rho_t(A_t)\} = \frac{\exp\{-\rho_t(A_t)/kT_B\}}{\sum_{B\in\mathbf{A}^U}\exp\{-\rho_t(B)/kT_B\}}, \tag{14}$$
$$\forall A_t \in \mathbf{A}^U,$$

where $k = 1.38064852 \times 10^{-23}$ J/K is the Boltzmann constant; T_B is the system temperature (in K). High temperatures make all actions almost equiprobable and low temperatures result in greedy action selection [22].

Using the above definitions, the dynamics of a JUSTE-RL with regret can be described as [23]

$$\overline{u}_{t+1}(B) = \overline{u}_t(B) + \alpha_t \mathbf{1}_{A_t=B}\left(\overline{u}_t - \overline{u}_t(B)\right),$$
$$\rho_{t+1}(B) = \rho_t(B) + \beta_t\left(\overline{u}_t(B) - \overline{u}_t - \rho_t(B)\right), \tag{15}$$
$$\pi_{t+1}(B) = \pi_t(B) + \gamma_t\left(G\{\rho_t(B)\} - \pi_n(B)\right),$$

for all $B \in \mathbf{A}^U$, where α_t, β_t, and γ_t are the learning rates, such that [23]

$$\lim_{t\to\infty}\sum_{\tau=1}^{t}\alpha_\tau = +\infty,$$
$$\lim_{t\to\infty}\sum_{\tau=1}^{t}\alpha_\tau^2 < +\infty,$$
$$\lim_{t\to\infty}\sum_{\tau=1}^{t}\beta_\tau = +\infty,$$
$$\lim_{t\to\infty}\sum_{\tau=1}^{t}\beta_\tau^2 < +\infty,$$
$$\lim_{t\to\infty}\sum_{\tau=1}^{t}\gamma_\tau = +\infty,$$
$$\lim_{t\to\infty}\sum_{\tau=1}^{t}\gamma_\tau^2 < +\infty,$$

Initialization:
(1) **Input** $\upsilon_\alpha, \upsilon_\beta, \upsilon_\gamma, T$;
(2) **For all** $A_0 \in \mathbf{A}^U$, **set** $\overline{u}_0(A_0) \leftarrow 0, \rho_0(A_0) \leftarrow 0, \pi_0(A_0) \leftarrow 0$;
(3) **For all** $n \in \mathbf{N}$, **set** $\overline{r}_n^U(0) \leftarrow 0$;
Main Loop:
(4) **While** $(t \leq T)$ **do**
(5) **Select** $A_t \leftarrow \arg\max_{B \in \mathbf{A}^U}(\pi_t(B))$ **and set** $\overline{\mathbf{a}}_t^U \leftarrow A_t$;
(6) **For all** $n \in \mathbf{D}$, **set** $\overline{b}_n(t) = \max\{0,\ 1 - \sum_{m \in \mathbf{M}} \overline{a}_n^m(t)\}$;
(7) **For all** $n \in \mathbf{C}$, **set** $\overline{b}_n(t) = 0$;
(8) **Execute** A_t **and observe** $\overline{r}_n^U(t)$, **for all** $n \in \mathbf{N}$;
(9) **Solve** (12a)–(12e) to find an optimal $\overline{u}_t(A_t) \leftarrow \overline{u}_t$;
(10) **For all** $B \in \mathbf{A}^U$, **update** $\overline{u}_t(B), \rho_t(B), \pi_t(B)$ using (15);
(11) **End**.

Algorithm 1: JUSTE-RL with regret for unlicensed channel allocation. RL algorithm for outband channel allocation.

$$\lim_{t\to\infty} \frac{\gamma_t}{\alpha_t} = 0,$$
$$\lim_{t\to\infty} \frac{\beta_t}{\gamma_t} = 0. \tag{16a}$$

Typically, the learning rates are set equal [22]:

$$\alpha_t = \frac{1}{(t+1)^{\upsilon_\alpha}},$$
$$\beta_t = \frac{1}{(t+1)^{\upsilon_\beta}}, \tag{16b}$$
$$\gamma_t = \frac{1}{(t+1)^{\upsilon_\gamma}},$$

where $\upsilon_\alpha \in (0.5, 1]$, $\upsilon_\beta \in (0.5, 1]$, $\upsilon_\gamma \in (0.5, 1]$, and $\upsilon_\alpha \leq \upsilon_\beta \leq \upsilon_\gamma$. The initializations $\overline{u}_0(A) \geq 0$, $\rho_0(A) \geq 0$, and $\pi_0(A) \geq 0$ should be sufficiently close to zero, for all $A \in \mathbf{A}^U$. The dynamics (15) converge to the ε-Nash equilibrium. Note that a Nash equilibrium point for (15) is given by [23]:

$$\lim_{t\to\infty} \pi_t(A) = \pi^*(A), \quad \forall A \in \mathbf{A}^U. \tag{17}$$

The corresponding learning algorithm for unlicensed channel allocation is presented in Algorithm 1 (where T indicates the total simulation length in slots). Note that this algorithm converges when $\pi_t(B) = \pi_{t-1}(B)$, for all $B \in \mathbf{A}^U$. The complexity of JUSTE-RL with regret is mainly determined by the size of an action set $\mathbf{A}^U$, since, at any slot t, we have to select an action $A_t \in \mathbf{A}^U$ that maximizes $\pi_t(A_t)$ (the dynamics in (15) require only algebraic operations and, thus, its computational complexity is negligible). Consequently, the worst-case time complexity of Algorithm 1 is $O(n)$, where $n = |\mathbf{A}^U| \leq N \times M$ is the size of our action set.

3.3. Inband Resource Allocation. Consider (12a)–(12e) that represents a joint mode, RB, and power level allocation problem. This problem has two binary optimization variables $\mathbf{a}_t^L$ and $\mathbf{c}_t$, one real-valued variable $\mathbf{p}_t$, nonlinear objective (12a), and nonlinear constraints (12c) and (12e). Hence, it belongs to a family of the mixed-integer nonlinear programming (MINLP) problems. It has been well established in the past (see, e.g., [38]) that all MINLP problems involving binary variables (such as (12a)–(12e)) are Nondeterministic Polynomial-time- (NP-) hard. For immediate NP-hardness proof for a considered problem note that, given that $\mathbf{a}_t^L$ can be either 0 or 1, any feasible solution to (12a)–(12e) is a subset of vertices. The constraint (12d) also implies that at least one end point of every edge is included in this subset. Hence, the solution to this problem describes a vertex cover, for which finding a minimum is NP-hard.

Most of the MINLP solution techniques involve the construction of the following relaxations to the considered problem: a nonlinear programming (NLP) relaxation (the original problem without integer restrictions) and a mixed-integer linear-programming (MILP) relaxation (an original problem where the nonlinearities are replaced by supporting hyperplanes). To form the MILP and NLP relaxations to (12a)–(12e), let us first represent in equivalent form the following:

$$\text{minimize} \quad \sum_{n \in \mathbf{N}} \left[\upsilon_n p_n(t) - \omega^L \overline{b}_n(t) \sum_{k \in \mathbf{K}} a_n^k(t)\, d_n^k - \left(1 - \overline{b}_n(t)\right) \overline{r}_n^U(t) \right], \tag{18a}$$

$$\text{subject to} \quad \mathbf{a}_t^L \in \mathbf{A}^L,$$
$$\mathbf{c}_t \in \mathbf{C}, \tag{18b}$$
$$\mathbf{p}_t \in \mathbf{P},$$

$$\sum_{n\in\mathbf{N}} a_n^k(t)\,\bar{b}_n(t) \geq 1, \quad \forall k \in \mathbf{K}, \tag{18c}$$

$$g_k^1\left(\mathbf{a}_t^L, \mathbf{c}_t\right) = \sum_{n\in\mathbf{N}} a_n^k(t)\,\bar{b}_n(t)\,c_n(t) - 1 \leq 0, \quad \forall k \in \mathbf{K}, \tag{18d}$$

$$g_n^2\left(\mathbf{a}_t^L, \mathbf{p}_t\right) = \mathrm{SINR}_n(t) - \mathrm{SINR}_n^{\mathrm{tar}} \leq 0, \quad \forall n \in \mathbf{N}, \tag{18e}$$

$$g_{n,k}^3\left(\mathbf{a}_t^L, \mathbf{p}_t\right) = \mathrm{SINR}_n^k(t) - 2^{d_n^k} + 1 \leq 0, \quad \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}, \tag{18f}$$

where objective (18a) and constraints (18b) and (18c) are linear, while constraints (18d)–(18f) are nonlinear. The MILP relaxation to (18a)–(18f) in a given point $(\mathbf{a}_t^{L0}, \mathbf{c}_t^0, \mathbf{p}_t^0)$ is given by

$$\text{minimize} \quad \sum_{n\in\mathbf{N}} \left[v_n p_n(t) - \omega^L \bar{b}_n(t) \sum_{k\in\mathbf{K}} a_n^k(t)\,d_n^k - \left(1 - \bar{b}_n(t)\right)\bar{r}_n^U(t) \right], \tag{19a}$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{a}_t^L \in \mathbf{A}^L, \\ & \mathbf{c}_t \in \mathbf{C}, \\ & \mathbf{p}_t \in \mathbf{P}, \end{aligned} \tag{19b}$$

$$\sum_{n\in\mathbf{N}} a_n^k(t)\,\bar{b}_n(t) \geq 1, \quad \forall k \in \mathbf{K}, \tag{19c}$$

$$g_k^1\left(\mathbf{a}_t^L, \mathbf{c}_t\right) + \nabla g_k^1\left(\mathbf{a}_t^{L0}, \mathbf{c}_t^0\right)^T \begin{bmatrix} \mathbf{a}_t^L - \mathbf{a}_t^{L0} \\ \mathbf{c}_t - \mathbf{c}_t^0 \end{bmatrix} \leq 0, \quad \forall k \in \mathbf{K}, \tag{19d}$$

$$g_n^2\left(\mathbf{a}_t^L, \mathbf{p}_t\right) + \nabla g_n^2\left(\mathbf{a}_t^{L0}, \mathbf{p}_t^0\right)^T \begin{bmatrix} \mathbf{a}_t^L - \mathbf{a}_t^{L0} \\ \mathbf{p}_t - \mathbf{p}_t^0 \end{bmatrix} \leq 0, \quad \forall n \in \mathbf{N}, \tag{19e}$$

$$g_{n,k}^3\left(\mathbf{a}_t^L, \mathbf{p}_t\right) + \nabla g_{n,k}^3\left(\mathbf{a}_t^{L0}, \mathbf{p}_t^0\right)^T \begin{bmatrix} \mathbf{a}_t^L - \mathbf{a}_t^{L0} \\ \mathbf{p}_t - \mathbf{p}_t^0 \end{bmatrix} \leq 0, \quad \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}. \tag{19f}$$

The NLP relaxation to (18a)–(18f) is given by

$$\text{minimize} \quad \sum_{n\in\mathbf{N}} \left[v_n p_n(t) - \omega^L \bar{b}_n(t) \sum_{k\in\mathbf{K}} a_n^k(t)\,d_n^k - \left(1 - \bar{b}_n(t)\right)\bar{r}_n^U(t) \right], \tag{20a}$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{a}_t^L \in \widetilde{\mathbf{A}}^L, \\ & \mathbf{c}_t \in \widetilde{\mathbf{C}}, \\ & \mathbf{p}_t \in \mathbf{P}, \end{aligned} \tag{20b}$$

$$\sum_{n\in\mathbf{N}} a_n^k(t)\,\bar{b}_n(t) \geq 1, \quad \forall k \in \mathbf{K}, \tag{20c}$$

$$g_k^1\left(\mathbf{a}_t^L, \mathbf{c}_t\right) \leq 0, \quad \forall k \in \mathbf{K}, \tag{20d}$$

$$g_n^2\left(\mathbf{a}_t^L, \mathbf{p}_t\right) \leq 0, \quad \forall n \in \mathbf{N}, \tag{20e}$$

$$g_{n,k}^3\left(\mathbf{a}_t^L, \mathbf{p}_t\right) \leq 0, \quad \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}, \tag{20f}$$

where

$$\widetilde{\mathbf{A}}^{L} = \left\{\mathbf{a}_t^L \mid 0 \le a_n^k(t) \le 1,\ \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}\right\}; \tag{20g}$$

$$\widetilde{\mathbf{C}} = \left\{\mathbf{c}_t \mid 0 \le c_n(t) \le 1,\ c_i(t) = 1,\ \forall n \in \mathbf{N},\ \forall i \in \mathbf{C}\right\}. \tag{20h}$$

In general, all MINLP problems can be solved using either exact techniques (e.g., branch-and-bound [39]) or heuristic methods (such as local branching [40], large neighborhood search [41], and feasibility pump [42]). Since we are interested in a reasonably simple and fast algorithm, it is more convenient to use heuristics to solve (18a)–(18f). Among numerous heuristic techniques, feasibility pump (FP) [43] is perhaps the most simple and effective method for producing more and better solutions in a shorter average running time (the local convergence properties of FP for nonconvex problems have been proved in [44]). The fundamental idea of an FP heuristic is to decompose the MINLP problem into two parts: integer feasibility and constraint feasibility. The former is achieved by rounding (solving the MILP relaxation to an original problem), the latter by projection (solving the NLP relaxation). The algorithm generates two sequences of integral and rounding points. The first sequence of integral points, $\{(\overline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i})\}_{i=1}^I$, $I = 1, 2, \ldots$, contains the solutions that may violate the nonlinear constraints; the second sequence, $\{(\underline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i})\}_{i=1}^I$, comprises the rounding points that are feasible for the MILP relaxation but might not be integral.

Particularly, with the input $(\underline{\mathbf{a}_t^{L1}, \mathbf{c}_t^1, \mathbf{p}_t^1})$ being a solution to an NLP relaxation (20a)–(20f), FP generates two sequences by solving the following problems, for $i = 1, \ldots, I$,

$$\left(\overline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i}\right) = \arg\min \quad \left\|\left(\mathbf{a}_t^L, \mathbf{c}_t, \mathbf{p}_t\right) - \left(\underline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i}\right)\right\|_1, \tag{21a}$$

$$\text{subject to} \quad \mathbf{a}_t^L \in \mathbf{A}^L,$$

$$\mathbf{c}_t \in \mathbf{C}, \tag{21b}$$

$$\mathbf{p}_t \in \mathbf{P},$$

$$\sum_{n \in \mathbf{N}} a_n^k(t)\, \overline{b}_n(t) \ge 1, \quad \forall k \in \mathbf{K}, \tag{21c}$$

$$g_k^1\left(\mathbf{a}_t^L, \mathbf{c}_t\right) + \nabla g_k^1\left(\mathbf{a}_t^{L0}, \mathbf{c}_t^0\right)^T \begin{bmatrix} \mathbf{a}_t^L - \mathbf{a}_t^{L0} \\ \mathbf{c}_t - \mathbf{c}_t^0 \end{bmatrix} \le 0, \quad \forall k \in \mathbf{K}, \tag{21d}$$

$$g_n^2\left(\mathbf{a}_t^L, \mathbf{p}_t\right) + \nabla g_n^2\left(\mathbf{a}_t^{L0}, \mathbf{p}_t^0\right)^T \begin{bmatrix} \mathbf{a}_t^L - \mathbf{a}_t^{L0} \\ \mathbf{p}_t - \mathbf{p}_t^0 \end{bmatrix} \le 0, \quad \forall n \in \mathbf{N}, \tag{21e}$$

$$g_{n,k}^3\left(\mathbf{a}_t^L, \mathbf{p}_t\right) + \nabla g_{n,k}^3\left(\mathbf{a}_t^{L0}, \mathbf{p}_t^0\right)^T \begin{bmatrix} \mathbf{a}_t^L - \mathbf{a}_t^{L0} \\ \mathbf{p}_t - \mathbf{p}_t^0 \end{bmatrix} \le 0, \quad \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}; \tag{21f}$$

$$\left(\underline{\mathbf{a}_t^{Li+1}, \mathbf{c}_t^{i+1}, \mathbf{p}_t^{i+1}}\right) = \arg\min \quad \left\|\left(\mathbf{a}_t^L, \mathbf{c}_t, \mathbf{p}_t\right) - \left(\overline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i}\right)\right\|_2, \tag{22a}$$

$$\text{subject to} \quad \mathbf{a}_t^L \in \widetilde{\mathbf{A}}^L,$$

$$\mathbf{c}_t \in \widetilde{\mathbf{C}}, \tag{22b}$$

$$\mathbf{p}_t \in \mathbf{P},$$

$$\sum_{n \in \mathbf{N}} a_n^k(t)\, \overline{b}_n(t) \ge 1, \quad \forall k \in \mathbf{K}, \tag{22c}$$

$$g_k^1\left(\mathbf{a}_t^L, \mathbf{c}_t\right) \le 0, \quad \forall k \in \mathbf{K}, \tag{22d}$$

$$g_n^2\left(\mathbf{a}_t^L, \mathbf{p}_t\right) \le 0, \quad \forall n \in \mathbf{N}, \tag{22e}$$

$$g_{n,k}^3\left(\mathbf{a}_t^L, \mathbf{p}_t\right) \le 0, \quad \forall n \in \mathbf{N},\ \forall k \in \mathbf{K}, \tag{22f}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are l_1-norm and l_2-norm, respectively. The rounding is carried out by solving the problem (21a)–(21f) and the projection is the solution to (22a)–(22f). Consequently, an FP algorithm alternates between the rounding and

projection steps until $(\overline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i}) = (\underline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i})$ (which implies feasibility) or until the number of iterations i has reached its predefined limit I. The workflow of the algorithm is presented in Algorithm 2. Note that to retain the local convergence, the problems (21a)–(21f) and (22a)–(22f) have to be solved exactly in the FP algorithm. The problem (22a)–(22f) (and, hence, (20a)–(20f)) can be solved using any standard NLP method. In this paper, an interior point algorithm (described, e.g., in [45]) which has a polynomial $O(n^2)$ time complexity is applied to solve (20a)–(20f) and (22a)–(22f). The MILP problem (21a)–(21f) is relatively simple and, therefore, it can be solved to optimality by any technique from the family of the branch-and-bound methods (e.g., [46]).

Note that, in general, finding an optimal solution to any joint resource allocation problem with integrality constraints is NP-hard (which has been shown in [47]). Consequently, most of the recent approaches to deal with such kind of problems focus on finding the high-quality suboptimal solutions using, for example, relaxation (by removing all the integer restrictions, as it has been done in [47, 48]) or iterative two-stage algorithms for determining the optimal integral solutions given fixed power levels and, then, finding the optimal power allocation with fixed integral points (e.g., [49]). In this paper, instead of relaxation or iteration, we directly apply a heuristic FP algorithm that has a polynomial $O(n^c)$ time complexity in the size n of the problem (with c being some real constant) [43] (Note that in our case, the size n of the problem (18a)–(18f) is proportional to $|\mathbf{A}^L| \times |\mathbf{C}| \times |\mathbf{P}| = N^3 \times K$. The numerical results showing the complexity of a proposed algorithm will be presented in Section 4.). Hence, the presented heuristic approach has moderate complexity compared to the previously proposed algorithms for resource allocation with integrality constraints whose complexity ranges from linear [21, 30, 47, 48, 50] to polynomial [2, 3, 49, 51–53].

4. Algorithm Implementation

4.1. Resource Allocation Procedure. We now discuss the implementation of the proposed algorithms (presented in Section 3) in an LTE-A network. The following scheduling procedure is repeated at the beginning of each slot t as follows.

(i) All users send their SRs to the eNB via dedicated PUCCHs. Note that the SRs may contain some useful control information, such as updated target SINR level $\text{SINR}_n^{\text{tar}}$ or observed throughput on the unlicensed channel $\bar{r}_n^U(t)$.

(ii) After receiving the SRs from all of the users, the eNB performs resource allocation (by assigning the modes, RBs and unlicensed channels, and power levels to user pairs according to Algorithms 1 and 2) and sends the SGs with optimal allocations to the corresponding users via PDCCHs.

(iii) After receiving the SGs, the users start their data transmissions over allocated RBs/unlicensed channels with assigned mode and power levels.

As it was already been mentioned, we deploy a CSMA/CA for outband D2D access using a procedure described in IEEE 802.11 [54]. As dictated by [54], if a certain D2D pair PU_n, $n \in \mathbf{D}$, is allocated with one or more unlicensed channels then, prior to transmission, one of the users must first sense the channel (to determine whether it is idle) for the duration of a distributed coordination function interframe space (DIFS). DIFS (which is 34 μs long) consists of a short interframe space (SIFS) equaling 16 μs and 2 Wi-Fi slots (each equals 9 μs). After DIFS, a user must typically defer its transmission for a random number of slots, generated from 0 to CW-1 (contention window size), to allow the other devices to share a channel in a fair manner. Given that the minimum CW value is $\text{CW}_{\min} = 16$, the device will, on an average, wait for about 7.5 Wi-Fi slots before transmission. Thus, the average channel access delay is 16 μs + 9.5 × 9 μs = 101.5 μs (independent of service rate). Since the slot duration in LTE system (T_s = 1 ms) is much longer than the average channel access delay (101.5 μs), it is expected that (in average) a D2D pair will be able to exchange the data within the scheduled period. In this case, each of the users in a D2D pair should observe achieved throughput $\bar{r}_n^U(t)$ and report this value to the eNB when sending its SR. Otherwise (if a D2D pair is not able to exchange the data within one slot), the D2D users send the value $\bar{r}_n^U(t) = 0$ to the eNB. Note that a CSMA/CA does not allow two-way data transmission. Hence, the second device in a D2D pair can start the data transmission only after the first user has finished its transmission.

It is worth mentioning that, at some point in time, a JUSTE-RL will reach its equilibrium state. However, even after the equilibrium has been reached, the eNB continues the learning process, because the network environment (channel quality, network traffic, and the number of active users) is likely to change over time resulting in different optimal mode, RB/unlicensed channel, and power allocations.

4.2. Simulation Model. A simulation model of the network has been implemented upon a standard LTE-A platform using the OPNET simulation and development package [55]. The model consists of one eNB and N user pairs randomly positioned inside a three-sector hexagonal cell (with the antenna pattern specified in [56]). It is assumed that the users operate outdoors in a typical urban environment and are stationary throughout all simulation runs. Each user device has its own traffic generator, enabling a variety of traffic patterns. For simplicity, in the examples below, the user traffic is modeled as a full buffer with load of 10 packets per second and packet size of 1500 bytes. In all simulations, $|\mathbf{C}| = 10$ cellular pairs, $\upsilon_\beta = 1.1\upsilon_\alpha$, $\upsilon_\gamma = 1.1\upsilon_\beta$, $T = 1000$ slots, $T_B = 10^6$ K, and $I = 1000$ iterations, the target SINR levels for each device pair are set as $\text{SINR}_n^{\text{tar}} = \text{SINR}^{\text{tar}} = 0$ dB, for all $n \in \mathbf{N}$. The licensed band of the eNB comprises $K = 100$ RBs (equivalent to 20 MHz). The unlicensed band comprises $M = 4$ nonoverlapping OFDM channels with $\omega_m^U = 10$ MHz, for $m \in \mathbf{M}$. The main simulation parameters of our model are listed in Table 1. Other parameters are set in accordance with 3GPP specifications [56].

Initialization:
(1) **Input** I, T;
(2) **While** $(t \leq T)$ **do**
  (3) **Input** $\bar{r}_n^U(t), \bar{\mathbf{a}}_t^U, \bar{\mathbf{b}}_t$;
  (4) **Solve** (20a)–(20f) to find the optimal $(\underline{\mathbf{a}_t^{L1}, \mathbf{c}_t^1, \mathbf{p}_t^1})$;
  Main Loop:
  (5) **While** $(i \leq I)$ **do**
    Rounding:
    (6) **Solve** (21a)–(21f) to find the optimal $(\overline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i})$;
    (7) **If** $((\overline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i}) = (\underline{\mathbf{a}_t^{Li}, \mathbf{c}_t^i, \mathbf{p}_t^i}))$ **then break**;
    Projection:
    (8) **Solve** (22a)–(22f) to find the optimal $(\underline{\mathbf{a}_t^{Li+1}, \mathbf{c}_t^{i+1}, \mathbf{p}_t^{i+1}})$;
    (9) **Set** $i \leftarrow i + 1$;
(10) **End**.

Algorithm 2: FP algorithm for inband resource allocation.

Table 1: Simulation parameters of LTE-A Model.

Parameter	Value
Cell radius	500 m
Frame structure	Type 2 (time division duplex)
Slot duration	1 ms
TDD configuration	0
eNodeB Tx power	46 dBm
UE active/idle Tx power	23/2 dBm
Noise power	−174 dBm/Hz
Path loss and cellular link	$128.1 + 37.6 \log(d)$, d [km]
NLOS path loss and D2D link	$40 \log(d) + 30 \log(f) + 49$, d [km], f [Hz]
LOS path loss and D2D link	$16.9 \log(d) + 20 \log(f/5) + 46.8$, d [m], f [GHz]
Shadowing st. dev.	10 B (cell mode); 12 dB (D2D mode)

In this paper, the evaluation of a proposed approach for inband and outband resource allocation, referred to as JRA (JUSTE-RL based resource allocation), is divided into two parts. In the first part, we analyze the performance of JUSTE-RL with regret for unlicensed channel allocation (Algorithm 1). In the second part, we examine the efficiency of a proposed joint inband/outband resource allocation (Algorithms 1 and 2). In the following, the performance of JRA is compared with the performance of the following resource allocation techniques.

(i) First is joint inband/outband resource allocation with ε-greedy Q-learning (GQL) [57] based on formulations (12a)–(12e) and (18a)–(18f), where the unlicensed channels are allocated to the users by the LTE eNB. In GQL, at any slot t, an action with the largest Q-value is selected with probability $1 - \varepsilon$ and the other actions are selected uniformly at random with probability ε. In all simulation experiments, the value of ε is set in accordance with the most common suggestions (provided, e.g., in [22]), as $\varepsilon = 0.1$.

(ii) Second is centralized optimal strategy (COS), where the inband and outband network resources are allocated to the users by solving (11a)–(11f) directly based on global channel and network knowledge. Note that COS corresponds to the most efficient (in terms of network utility maximization) strategy although it is not practically realizable (since in the real network deployment scenarios, the precise information about quality and availability of unlicensed channels is not available) (In this paper, we use an FP algorithm to find the optimal solution to (11a)–(11f) or (18a)–(18f) in GQL and COS.).

(iii) Third is social heuristic for multimode D2D communication (SMD) in an LTE-A network proposed in [21] to reduce the complexity of an original optimization problem for joint inband/outband resource allocation. This algorithm assigns user modes and resources to maximize the social welfare based on the global channel and network knowledge. The eNB creates a randomly ordered list of the D2D pairs. Then, it computes the aggregated network utility for each mode of the first user in the list and assigns this user with a mode that provides the highest aggregated utility. This process is repeated for all D2D pairs.

(iv) Fourth is greedy heuristic for multimode D2D communication (GMD) in LTE-A networks [21], where the modes and inband/outband network resources are allocated to maximize the individual users' welfare based on the global channel and network knowledge. Similar to SMD, the eNB creates a randomly ordered list of the D2D pairs. After this, it computes the utility for each mode of the first user in the list and assigns this user a mode assuring the highest individual utility. This process is repeated for all D2D pairs.

(v) Fifth is ranked heuristic for multimode D2D communication (RMD) in LTE-A networks [21]. Here the eNB evaluates the utility of each user in each mode (based on the global channel and network knowledge)

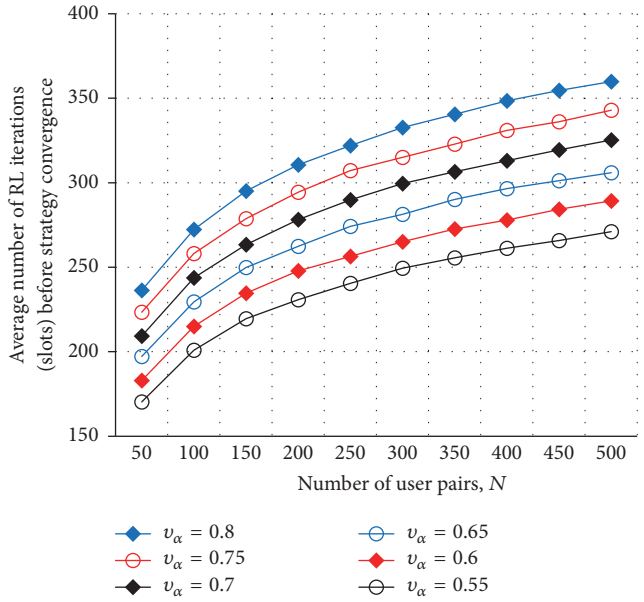

Figure 2: The average number of RL iterations (slots) necessary for convergence of strategies in JRA with different values of υ_α and fixed $|\mathbf{C}| = 10$.

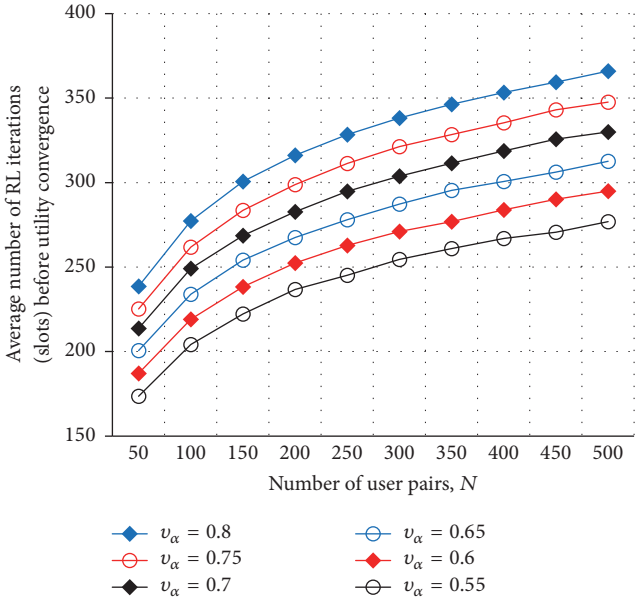

Figure 3: The average number of RL iterations (slots) necessary for convergence of utilities in JRA with different values of υ_α and fixed $|\mathbf{C}| = 10$.

and sorts the D2D pairs according to their utilities in a descending order. Next, the eNB allocates the first user in the list a mode that guarantees the highest aggregated network utility. This process is repeated for all D2D pairs.

Note that all algorithms used in our performance evaluation are simulated with identical system parameters.

4.3. Performance of a Learning Algorithm. We start with the performance evaluation of JUSTE-RL with regret for unlicensed channel allocation (outlined in Algorithm 1). Figures 2 and 3 demonstrate the learning speed of JRA. Figure 2 shows the average number of RL iterations (slots) necessary for convergence of strategies in JRA (at the point where $\pi_t(B) = \pi_{t-1}(B)$, for all $B \in \mathbf{A}^U$) with different values of $\upsilon_\alpha \in [0.55, 0.8]$, $\upsilon_\beta = 1.1\upsilon_\alpha \in [0.61, 0.88]$, and $\upsilon_\gamma = 1.1\upsilon_\beta \in [0.67, 0.97]$ and a varying number of user pairs, $N \in [50, 500]$. The average number of RL iterations (slots) necessary for convergence of utilities in JRA (at the point where $u_t(B) = u_{t-1}(B)$, for all $B \in \mathbf{A}^U$) with $\upsilon_\alpha \in [0.55, 0.8]$, $\upsilon_\beta \in [0.61, 0.88]$, and $\upsilon_\gamma \in [0.67, 0.97]$ and $N \in [50, 500]$) is plotted in Figure 3. The accuracy of estimation in JRA is presented in Figures 4 and 5. Figure 4 shows the absolute error of strategy estimation in JRA, denoted as δ_π, defined as a sum of the absolute differences between the actual optimal strategies and the estimated strategies upon the algorithm termination. That is,

$$\delta_\pi = \sum_{B \in \mathbf{A}^U} \left| \pi_T(B) - \pi^*(B) \right|, \tag{23}$$

where $\pi_T(B)$ is an optimal strategy estimated in JRA upon the algorithm termination (at slot T) and $\pi^*(B)$ is the actual

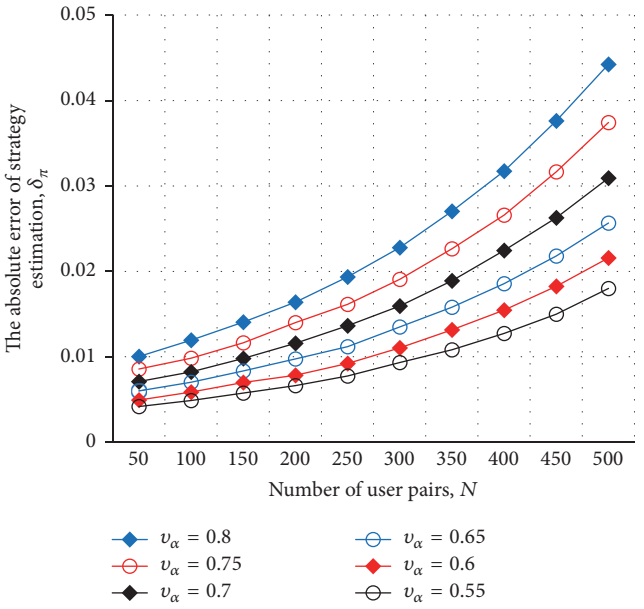

Figure 4: The absolute error of strategy estimation δ_π in JRA calculated upon the algorithm termination with different values of υ_α and fixed $|\mathbf{C}| = 10$.

optimal strategy obtained by playing an action $B \in \mathbf{A}^U$. Figure 5 demonstrates the absolute error of utility estimation in JRA, denoted δ_u, defined as a sum of the absolute differences between the actual and the estimated optimal network utilities upon the algorithm termination; that is,

$$\delta_u = \sum_{B \in \mathbf{A}^U} \left| u_T(B) - u^*(B) \right|, \tag{24}$$

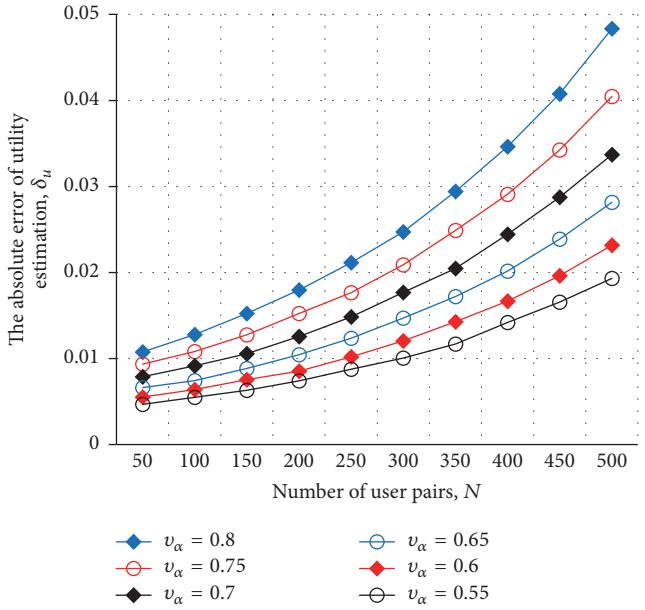

FIGURE 5: The absolute error of utility estimation δ_u in JRA calculated upon the algorithm termination with different values of υ_α and fixed $|\mathbf{C}| = 10$.

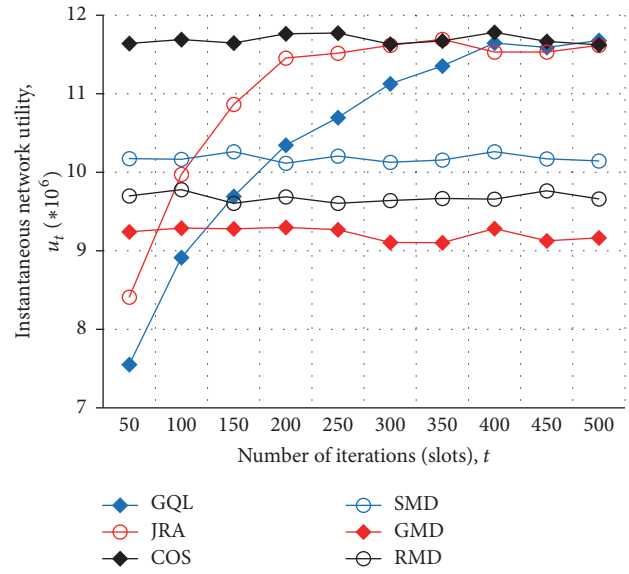

FIGURE 6: The instantaneous network utility u_t in different algorithms with fixed $N = 100$ and $|\mathbf{C}| = 10$.

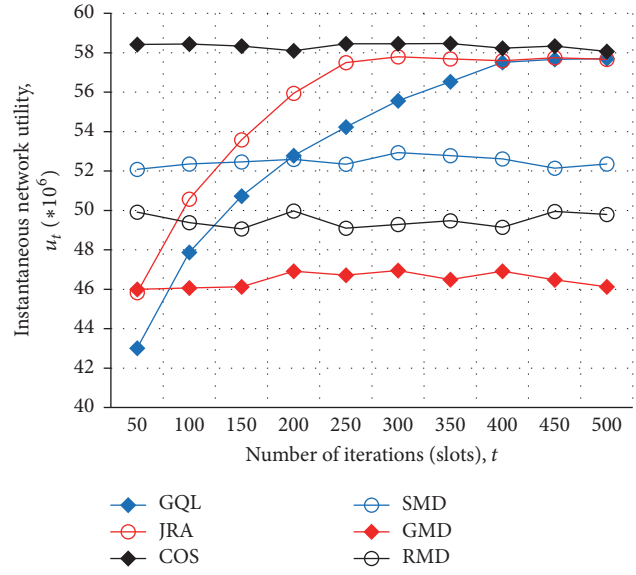

FIGURE 7: The instantaneous network utility u_t in different algorithms with fixed $N = 500$ and $|\mathbf{C}| = 10$.

where $u_T(B)$ is an optimal network utility estimated in JRA upon the algorithm termination (at slot T) and $u^*(B)$ is the actual optimal network utility obtained by playing an action $B \in \mathbf{A}^U$. The observations in Figures 2–5 show that the rates of convergence of strategies and utilities and the accuracy of strategy and utility estimation are almost the same. Furthermore, we find that the number of iterations necessary for the algorithm convergence and absolute estimation error strongly depend on the setting of the parameters υ_α, υ_β, and υ_γ: the worst performance is attained with $\upsilon_\alpha = 0.8$, $\upsilon_\beta = 0.61$, and $\upsilon_\gamma = 0.67$ and the best with $\upsilon_\alpha = 0.55$, $\upsilon_\beta = 0.88$, and $\upsilon_\gamma = 0.97$. Such results are rather predictable since the parameters υ_α, υ_β, and υ_γ are related to the parameters α_t, β_t, and γ_t (see (16b)) which have a direct influence on the learning rate of JRA [22, 23].

In Figures 6 and 7, the instantaneous network utility $u_t = \sum_{n \in \mathbf{N}} u_n(t)$ is presented as a function of time in scenarios with low network load ($N = 100$) and high network load ($N = 500$) and fixed. Here a proposed JRA technique is simulated with the settings $\upsilon_\alpha = 0.5$, $\upsilon_\beta = 0.61$, and $\upsilon_\gamma = 0.67$. The graphs in these figures show that the efficiency of JRA and GQL improves gradually over time. After about 300 slots (which is the average time necessary for the convergence of strategies and utilities in Algorithm 1), JRA demonstrates near-optimal results. GQL needs a little longer time (≈400 slots) to converge, after which its performance also becomes very close to the performance of COS. Unlike JRA and GQL, the performance of COS, SMD, GMD, and RMD is consistent over time (since these algorithms do not involve any learning process). We also observe that the network utility attained in SMD, GMD, and RMD is much smaller than that in COS. To understand such poor performance of SMD, GMD, and RMD, note that, in these algorithms, the original resource allocation problem is divided into two separate problems: (i) mode selection and (ii) packet scheduling. After that, the mode selection problem is solved using very plain heuristics (social, greedy, or ranked) which reduces the complexity of an original optimization problem (from exponential to linear) but has a negative impact on the performance of these techniques in terms of network utility maximization [21].

4.4. Performance of Joint Inband/Outband Allocation. We now evaluate the efficiency of a proposed inband/outband resource allocation (Algorithms 1 and 2). The graphs in Figures 8–10 demonstrate the computational complexity, solution time, and solution accuracy of different resource

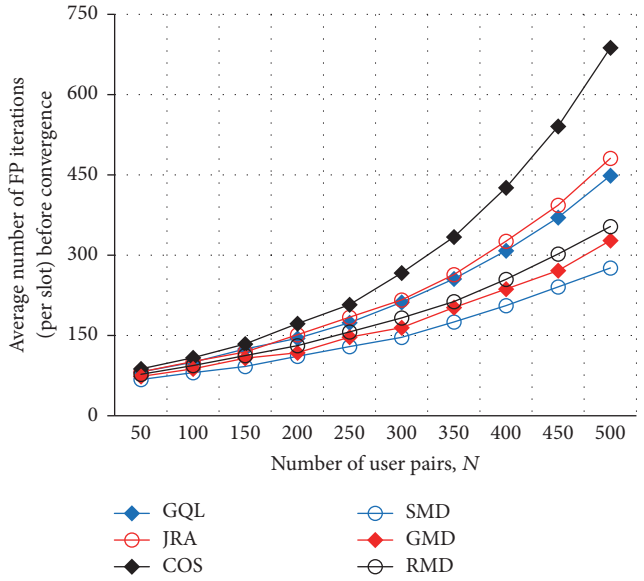

FIGURE 8: The average number of FP iterations (per slot) necessary for the convergence of the algorithms with fixed $|\mathbf{C}| = 10$ collected during T slots.

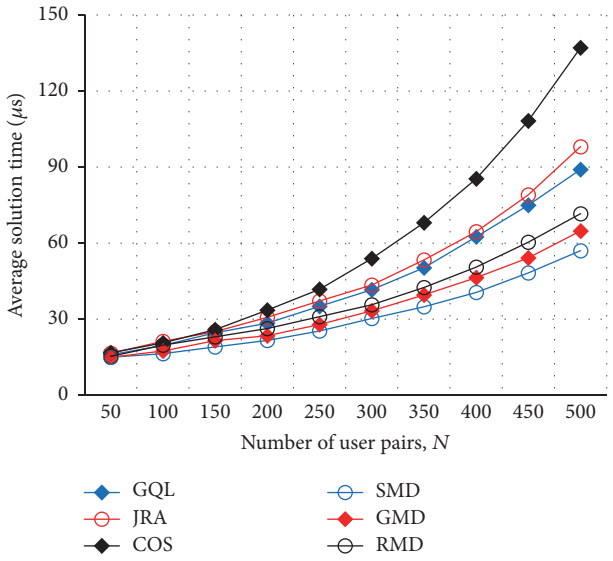

FIGURE 9: The average solution time (in μs) of different algorithms with fixed $|\mathbf{C}| = 10$ collected during T slots.

allocation techniques in the experiments with $\upsilon_\alpha = 0.55$, $\upsilon_\beta = 0.61$, and $\upsilon_\gamma = 0.67$ (in JRA) collected during the entire simulation period T. Particularly, in Figures 8 and 9, the average number of algorithm iterations (per slot) and solution time (in μs) are presented as a function of the number of user pairs N. Figure 10 shows the average relative deviation from the optimal solution, denoted as Δ and calculated according to

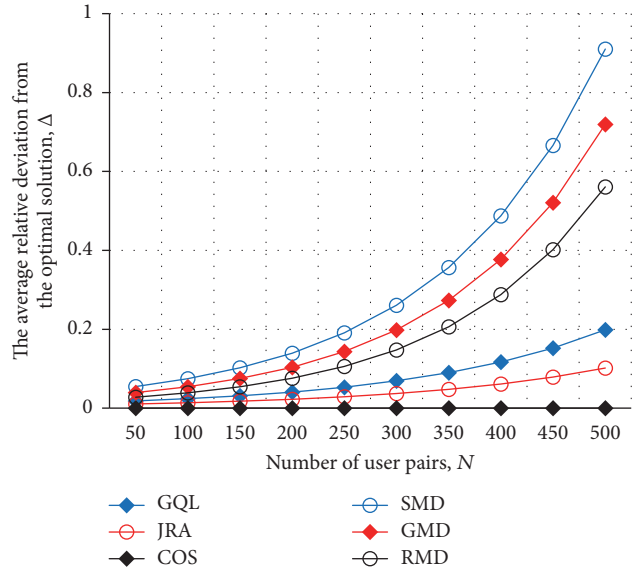

FIGURE 10: The average relative deviation from the optimal solution Δ in different algorithms with fixed $|\mathbf{C}| = 10$ collected during T slots.

$$\Delta = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\left|\mathbf{a}_t^{Li} - \mathbf{a}_t^{L*}\right|}{\mathbf{a}_t^{L*}} + \frac{\left|\mathbf{a}_t^{Ui} - \mathbf{a}_t^{U*}\right|}{\mathbf{a}_t^{U*}} + \frac{\left|\mathbf{c}_t^{i} - \mathbf{c}_t{}^{*}\right|}{\mathbf{c}_t{}^{*}} + \frac{\left|\mathbf{p}_t^{i} - \mathbf{p}_t{}^{*}\right|}{\mathbf{p}_t{}^{*}}\right) \tag{25}$$

in GQL, JRA, and COS and

$$\Delta = \frac{1}{T}\sum_{t=1}^{T}\frac{\left|x_t^{i} - x_t{}^{*}\right|}{x_t{}^{*}} \tag{26}$$

in SMD, FMD, and RMD. In the above equations, $(\mathbf{a}_t^{Li}, \mathbf{a}_t^{Ui}, \mathbf{c}_t^{i}, \mathbf{p}_t^{i})$ is the optimal solution found in GQL, JRA, and COS, $(\mathbf{a}_t^{L*}, \mathbf{a}_t^{U*}, \mathbf{c}_t{}^{*}, \mathbf{p}_t{}^{*})$ is the actual optimal solution to the original resource allocation problem (11a)–(11f), x_t^{i} is the mode allocation in SMD, FDM, and RMD, and $x_t{}^{*}$ is the actual optimal allocation (a solution to the optimization problem originally stated in [21]). It follows from these figures that all simulated strategies have moderate computational complexity. Predictably, COS has the highest complexity because the number of optimization variables $(\mathbf{a}_t^{L}, \mathbf{a}_t^{U}, \mathbf{c}_t, \mathbf{p}_t)$ in this algorithm is bigger than that in JRA, GQL, SMD, FMD, and RMD. The lowest complexity and solution accuracy are achieved in SMD, FMD, and RMD (which have a linear time complexity but are based on very raw approximations and plain heuristic assumptions).

Figures 11–13 present the observations collected at slot $t = 500$ with $\upsilon_\alpha = 0.55$, $\upsilon_\beta = 0.61$, and $\upsilon_\gamma = 0.67$ (in JRA). The average user throughput r_t (in kbits/s) and the average

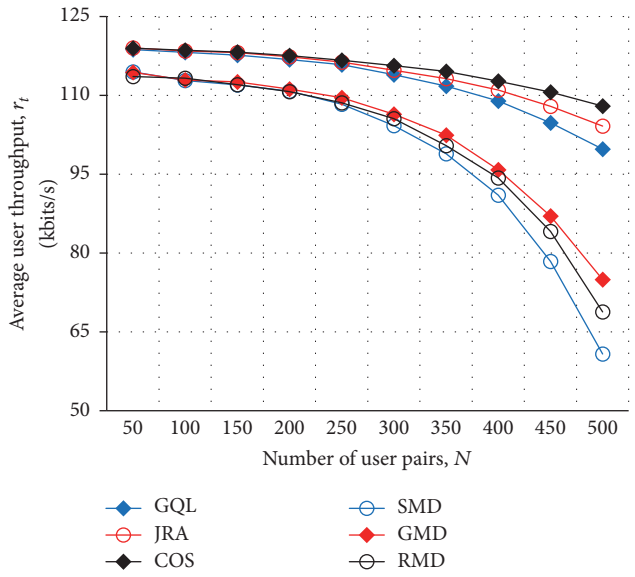

Figure 11: The average user throughput r_t (in kbits/s) in different algorithms with fixed $|\mathbf{C}| = 10$ observed at slot $t = 500$.

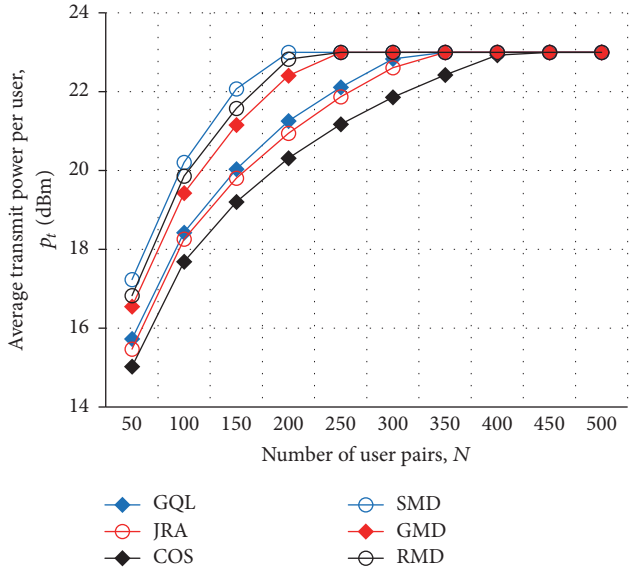

Figure 12: The average user transmit power (in dBm) in different algorithms with fixed $|\mathbf{C}| = 10$ observed at slot $t = 500$.

transmission power (per user) p_t (in dBm) in different algorithms estimated according to

$$r_t = \frac{1}{N} \sum_{n \in \mathbf{N}} \left(r_n^L(t) + r_n^U(t) \right), \quad p_t = \frac{1}{N} \sum_{n \in \mathbf{N}} p_n(t), \tag{27}$$

are shown in Figures 11 and 12, respectively. The instantaneous network utility u_t in different algorithms depending on the target SINR level, SINR^{tar}, with fixed number of user pairs, $N = 100$, is plotted in Figure 13. The obtained results

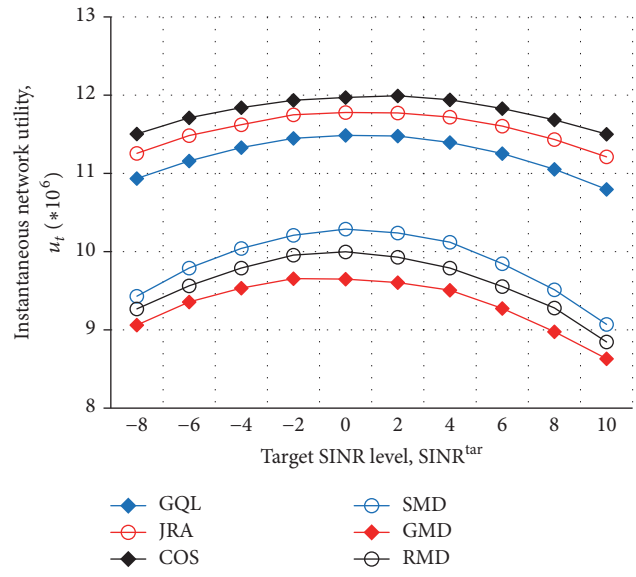

Figure 13: The instantaneous network utility u_t in different algorithms with fixed $N = 100$ and fixed $|\mathbf{C}| = 10$ observed at slot $t = 500$.

demonstrate that the average user throughput decreases with the number of user pairs N (Figure 11). This is rather predictable because when the network load increases, the number of RBs or unlicensed channels available for each user decreases resulting in a reduced throughput. Besides, to achieve the desired SINR levels, the users tend to transmit at a higher power level (see Figure 12) when the total number of user pairs in the network increases. The graphs in Figure 13 show that the network utilities in different resource allocation schemes are described by some concave functions of SINR^{tar}. To understand such results, note that with too low settings of SINR^{tar} ($\text{SINR}^{\text{tar}} < 0$ dB), the total user throughput reduces because of the bad channel conditions leading to the decreased network utility. On the other hand, when SINR^{tar} is too high ($\text{SINR}^{\text{tar}} > 4$ dB), the throughput (and, consequently, network utility) degrades due to the shortage of available bandwidth since the number of channels with suitable data transmission conditions becomes very small (because not all of them satisfy the SINR requirements of the users). We also observe that, in all simulated scenarios, the performance of JRA is very close to optimal (i.e., the one achieved in COS). GQL performs a little worse than JRA but still better than heuristic algorithms (SMD, GMD, and RMD).

5. Conclusion

This paper introduces a JRA algorithm for a D2D-enabled LTE-A network with access to unlicensed band provided by one or more RATs based on different channel access methods (OFDMA, CSMA/CA, FH-CDMA, etc.). In the presented framework, the inband/outband network resources (cellular/D2D modes, spectrum, and power) are allocated jointly by the LTE eNB to maximize the total network utility. Unlike most of the previously proposed techniques for

outband D2D communication (which presume a certain level of coordination and information exchange between licensed and unlicensed systems), our JUSTE-RL based approach for unlicensed channel assignment is fully autonomous and has demonstrated relatively fast (≈300 RL iterations) convergence to ε-Nash equilibrium (given the appropriate settings of learning rates). Simulations results also show that the proposed joint inband/outband resource allocation strategy outperforms other relevant spectrum and power management schemes in terms of energy efficiency and throughput maximization.

Competing Interests

The authors declare that they have no competing interests.

References

[1] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[2] A. Asheralieva and Y. Miyanaga, "Dynamic buffer status-based control for LTE-a network with underlay D2D communication," *IEEE Transactions on Communications*, vol. 64, no. 3, pp. 1342–1355, 2016.

[3] A. Asheralieva and Y. Miyanaga, "QoS oriented mode, spectrum and power allocation for D2D communication underlaying LTE-A network," *IEEE Transactions on Vehicular Technology*, 2016.

[4] J. Kim, S. Kim, J. Bang, and D. Hong, "Adaptive mode selection in D2D communications considering the bursty traffic model," *IEEE Communications Letters*, vol. 20, no. 4, pp. 712–715, 2016.

[5] D. Penda, L. Fu, and M. Johansson, "Energy efficient D2D communications in dynamic TDD systems," https://arxiv.org/abs/1506.00412.

[6] K. Yang, S. Martin, L. Boukhatem, J. Wu, and X. Bu, "Energy-efficient resource allocation for device-to-device communications overlaying LTE networks," in *Proceedings of the IEEE 82nd Vehicular Technology Conference (VTC Fall '15)*, pp. 1–6, Boston, Mass, USA, September 2015.

[7] A. Asadi, P. Jacko, and V. Mancuso, "Modeling multi-mode D2D communications in LTE," *Acm Sigmetrics*, vol. 42, no. 2, pp. 55–57, 2014.

[8] F. Malandrino, C. Casetti, C. F. Chiasserini, and Z. Limani, "Uplink and downlink resource allocation in D2D-enabled heterogeneous networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference Workshops (WCNCW '14)*, pp. 87–92, April 2014.

[9] D. Feng, L. Lu, Y.-W. Yi, G. Y. Li, G. Feng, and S. Li, "Device-to-device communications underlaying cellular networks," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3541–3551, 2013.

[10] L. Su, Y. Ji, P. Wang, and F. Liu, "Resource allocation using particle swarm optimization for D2D communication underlay of cellular networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '13)*, pp. 129–133, IEEE, Shanghai, China, April 2013.

[11] Wi-Fi Alliance, Wi-Fi Peer-to-Peer (P2P) Specification version 1.1, Wi-Fi Alliance Specification, 1, 2010.

[12] Z. Alliance, Zigbee Specification, Document 053474r06 (version), 1, 2006.

[13] Bluetooth Specification, Bluetooth Specification version 1.1, 2001, http://www.bluetooth.com.

[14] A. Asadi and V. Mancuso, "Energy efficient opportunistic uplink packet forwarding in hybrid wireless networks," in *Proceedings of the 4th ACM International Conference on Future Energy Systems (e-Energy '13)*, pp. 261–262, Berkeley, Calif, USA, May 2013.

[15] A. Asadi and V. Mancuso, "On the compound impact of opportunistic scheduling and D2D communications in cellular networks," in *Proceedings of the 16th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '13)*, pp. 279–287, November 2013.

[16] A. Asadi and V. Mancuso, "WiFi Direct and LTE D2D in action," in *Proceedings of the 6th IFIP/IEEE Wireless Days Conference (WD '13)*, pp. 1–8, Valencia, Spain, November 2013.

[17] Q. Wang and B. Rengarajan, "Recouping opportunistic gain in dense base station layouts through energy-aware user cooperation," in *Proceedings of the IEEE 14th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '13)*, pp. 1–9, June 2013.

[18] B. Zhou, S. Ma, J. Xu, and Z. Li, "Group-wise channel sensing and resource pre-allocation for LTE D2D on ISM band," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '13)*, pp. 118–122, IEEE, Shanghai, China, April 2013.

[19] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: basic principles and system performance," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176–189, 2016.

[20] H. Cai, I. Koprulu, and N. B. Shroff, "Exploiting double opportunities for deadline based content propagation in wireless networks," in *Proceedings of the 32nd IEEE Conference on Computer Communications (IEEE INFOCOM '13)*, pp. 764–772, Turin, Italy, April 2013.

[21] A. Asadi, P. Jacko, and V. Mancuso, "Modeling multi-mode D2D communications in LTE," https://arxiv.org/abs/1405.6689.

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, Mass, USA, 1998.

[23] S. M. Perlaza, H. Tembine, and S. Lasaulce, "How can ignorant but patient cognitive terminals learn their strategy and utility?" in *Proceedings of the IEEE 11th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC '10)*, June 2010.

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

[25] L. Rose, S. Lasaulce, S. M. Perlaza, and M. Debbah, "Learning equilibria with partial information in decentralized wireless networks," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 136–142, 2011.

[26] Y. Xu, Q. Wu, L. Shen, J. Wang, and A. Anpalagan, "Opportunistic spectrum access with spatial reuse: graphical game and uncoupled learning solutions," *IEEE Transactions on Wireless Communications*, vol. 12, no. 10, pp. 4814–4826, 2013.

[27] Y. Xu, Q. Wu, J. Wang, L. Shen, and A. Anpalagan, "Opportunistic spectrum access using partially overlapping channels: graphical game and uncoupled learning," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3906–3918, 2013.

[28] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[29] S. Maghsudi and S. Stańczak, "Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309–1322, 2015.

[30] A. Asheralieva and Y. Miyanaga, "An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in heterogeneous cellular networks," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3996–4012, 2016.

[31] 3rd Generation Partnership Project, "Physical channels and modulation," Technical Specification 3GPP TS 36.211 V9.1.0, 2010.

[32] 3rd Generation Partnership Project; Technical Specification, "E-UTRA; MAC protocol specification," 3GPP TS 36.211 V12.5.0, 2015.

[33] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Communications*, vol. 19, no. 3, pp. 96–104, 2012.

[34] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*, John Wiley & Sons, New York, NY, USA, 2011.

[35] I. F. Akyildiz, D. M. Gutierrez-Estevez, and E. C. Reyes, "The evolution to 4G cellular systems: LTE-Advanced," *Physical Communication*, vol. 3, no. 4, pp. 217–244, 2010.

[36] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: a game-theoretic stochastic learning solution," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1380–1391, 2012.

[37] H. Li, "Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: a two by two case," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '09)*, pp. 1893–1898, October 2009.

[38] T. Berthold, *Heuristic Algorithms in Global MINLP Solvers*, Verlag Dr. Hut, 2014.

[39] S. Burer and A. N. Letchford, "Non-convex mixed-integer nonlinear programming: a survey," *Surveys in Operations Research and Management Science*, vol. 17, no. 2, pp. 97–106, 2012.

[40] M. Fischetti and A. Lodi, "Local branching," *Mathematical Programming*, vol. 98, no. 1–3, pp. 23–47, 2003.

[41] A. Lodi, "The heuristic (dark) side of MIP solvers," in *Hybrid Metaheuristics*, vol. 434 of *Studies in Computational Intelligence*, pp. 273–284, Springer, Berlin, Germany, 2013.

[42] C. D'Ambrosio, A. Frangioni, L. Liberti, and A. Lodi, "A storm of feasibility pumps for nonconvex MINLP," *Mathematical Programming B*, vol. 136, no. 2, pp. 375–402, 2012.

[43] M. Fischetti and D. Salvagnin, "Feasibility pump 2.0," *Mathematical Programming Computation*, vol. 1, no. 2-3, pp. 201–222, 2009.

[44] C. D'Ambrosio, A. Frangioni, L. Liberti, and A. Lodi, "A storm of feasibility pumps for nonconvex MINLP," *Mathematical Programming*, vol. 136, no. 2, pp. 375–402, 2012.

[45] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[46] S. Leyffer, "Integrating SQP and branch-and-bound for mixed integer nonlinear programming," *Computational Optimization and Applications*, vol. 18, no. 3, pp. 295–309, 2001.

[47] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: computation and complexity," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1040–1054, 2015.

[48] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and resource allocation in heterogeneous networks via sparsity pursuit," https://arxiv.org/abs/1408.5091.

[49] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1100–1113, 2014.

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

[51] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 248–257, 2013.

[52] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo, "Optimal joint base station assignment and beamforming for heterogeneous networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 8, pp. 1950–1961, 2014.

[53] Q. Han, B. Yang, X. Wang, K. Ma, C. Chen, and X. Guan, "Hierarchical-game-based uplink power control in femtocell networks," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 6, pp. 2819–2835, 2014.

[54] IEEE Standards Association, *IEEE 802®: Local and Metropolitan Area Network Standards: IEEE 802.11 Standard*, 2012.

[55] OPNET, http://www.opnet.com.

[56] "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)," 3GPP TS 36.300, version V9.4.0, 2010.

[57] M. Bennis, S. Guruacharya, and D. Niyato, "Distributed learning strategies for interference mitigation in femtocell networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '11)*, pp. 1–5, Houston, Tex, USA, December 2011.