An Autonomous Learning-Based Algorithm for Joint Channel and Power Level Selection by D2D Pairs in Heterogeneous Cellular Networks

Alia Asheralieva and Yoshikazu Miyanaga

Abstract – We study the problem of autonomous operation of the device-to-device (D2D) pairs in a heterogeneous cellular network with multiple base stations (BSs). The spectrum bands of the BSs (that may overlap with each other) comprise the sets of orthogonal wireless channels. We consider the following spectrum usage scenarios: i) the D2D pairs transmit over the dedicated frequency bands, ii) the D2D pairs operate on the shared cellular/D2D channels. The goal of each device pair is 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 its minimum tolerable signal-to-interference-plus-noise ratio (SINR) requirements. We formulate this problem as a stochastic non-cooperative game with multiple players (D2D pairs) where each player becomes a learning agent whose task is to "learn" its best strategy (based on the locally observed information) and develop a fully autonomous multi-agent Q-learning algorithm converging to a mixed-strategy Nash equilibrium (NE). The proposed learning method is implemented in a Long Term Evolution-Advanced (LTE-A) network and evaluated via the OPNET-based simulations. The algorithm shows relatively fast convergence and near-optimal performance after a small number of iterations.

Index Terms – Device-to-Device Communication, Heterogeneous Networks, Interference Management, Reinforcement Learning, Resource Allocation.

I. INTRODUCTION

D2D communication is a direct communication between two users without traversing the BS or core network. The main advantage of D2D access is the possibility of quality of service (QoS) provisioning for mobile users. The two-fold challenge of D2D communication is interference mitigation (between the users operating on the same frequency bands) and efficient spectrum management (so that the allocated spectrum resources be not wasted) [1]. Adequate interference management and resource allocation schemes can boost the performance of D2D-enabled cellular networks (in terms of spectrum efficiency, cellular coverage, network throughput, user experience, etc.). The majority of existing approaches in this area (e.g., [2] – [8]) consider the network as a single entity, where the resources are allocated by some BS that has a global knowledge of the precise or statistical channel state information (CSI). However, since conventional pilot signals (deployed for cellular communication) cannot be used for estimation of the D2D channels [9], the assumption of the precise CSI in a D2D-enabled cellular network is somewhat unrealistic. As a result, many researchers have argued in favor of an alternative resource allocation strategy, where the D2D users establish a secondary network that is allowed to occupy the vacant cellular bands, thereby causing no interference to primary (cellular) users. In this context, many algorithms (including [10] – [13]) have been proposed on a game-theoretical basis. Unfortunately, most of the deployed game models (e.g., [10] – [12]) need many rounds of negotiation between the nodes resulting in the large control signaling overhead [14]. Auction-theoretic algorithms (such as [13]) require at least two bidders to compete for an auction target (leading to heavy information exchange among the cellular/D2D users).

An elegant way for distributed resource allocation for D2D communication has been presented in [15] and [16]. In [15], the problem of resource sharing between the cellular and D2D users in large-scale cellular networks is modelled using a coalition formation game where the total transmission rate in the system is represented by a utility function (that all the D2D pairs and cellular user aim to maximize). Based on the formulated utility maximization problem, the authors derive the algorithm for dynamic coalition formation, prove its convergence to a Nash-stable equilibrium (which approaches to an optimal solution with geometric rate), and show that the proposed method enhances the system sum rate when compared to other resource sharing strategies without loss of fairness. A new game model, called Bayesian non-transferable utility overlapping coalition formation, for spectrum sharing between the D2D and cellular communications in the network with multiple operators (where the D2D users associated with the same operator are regarded as a coalition) is proposed in [16]. It is assumed (due to complexity of the considered interference management) that each sub-band can be occupied by at most two users (either two D2D pairs or one D2D pair and one cellular subscriber). To find a stable coalition structure in a proposed game, the authors introduce a hierarchical matching method that consists of three separate algorithms, each of which is used to reach a stable matching state of a specific matching market. In passing, we note that the techniques developed in [15] and [16] require a certain amount of cooperation (in the form of information exchange among the D2D links, cellular users, and BSs) for coalition formation, although their design presumes distributed sharing of the spectrum resources.

Another approach that has recently been studied is a learning theory. In wireless communications, reinforcement learning (RL) [17] has been studied in the context of the various spectrum access problems. In [18] and [19], 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 the pure-strategy Nash equilibrium and, therefore, the proposed approach is not realizable in a broader range of problems. A cognitive network with multiple players has been analyzed in [20]. In this work, the learning and channel selection have been separated into two different procedures, which increased the complexity of the proposed resource allocation approach. Besides, the stability of a final solution was not verified. In [9], the opportunistic D2D access has been formulated as a multi-agent learning game, where the players (D2D users) learn the optimal action from successive interactions with a dynamic environment (orthogonal cellular channel). Each D2D user is selfish and aims at optimizing its throughput performance while being allowed to transmit on the vacant channels using a carrier sense multiple access with collision avoidance (CSMA/CA). The authors study a basic network model comprising only one BS and, therefore, the possibility of the inter-cell interference between the users operating in different cells over the same frequency bands is not taken into account. Note that a CSMA/CA-based medium access presumes that there are no channels with interfering users (i.e., each orthogonal channel can be occupied by at most one cellular/D2D user). Clearly, this is rather an impractical scenario to consider since many potential benefits of D2D communication (such as spectral efficiency and network utilization improvements) are dismissed. Besides, a CSMA/CA requires some channel access delay (which is independent of service rate) before the actual data transmission, which contributes to the growth in total medium access control (MAC) delay and, consequently, decreases the overall network throughput.

The main contributions of this work are as follows. We consider a fully autonomous operation of the D2D users in a heterogeneous cellular network comprising a number of BSs that have no information exchange with the device pairs and, hence, do not possess any knowledge about the quality of their channels. It is assumed that the spectrum bands of the BSs (comprising a number of orthogonal wireless channels) may overlap with each other, creating the possibility of the inter-cell interference (ICI) between the users transmitting over the same channels. We consider two possible scenarios of D2D operation. In the first scenario, the D2D users communicate over the dedicated spectrum resources creating no interference to cellular users operating on different frequencies. In the second scenario, the D2D users transmit over the shared cellular/D2D channels. The goal of each device pair is 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 level defined for each network user. We formulate this problem as a stochastic non-cooperative game with multiple selfish and rational players (D2D pairs) that have no information about the operating conditions. In a stochastic environment, each player becomes a learning agent whose task is to learn the best (leading to NE) strategy based on its observed reward and channel state (defined by the quality of the selected channel). Consequently, we model the game dynamics using a finite-state Markov Decision Process (MDP) and derive a multi-agent Q-learning method [17] for channel selection. Note that the conventional multi-agent Q-learning techniques (such as [21], [22]) presume the global knowledge of the players' strategies in all possible states (because in a multi-agent environment, the reward of each player depends not only on the action and state of this player but also on the other players' strategies). This requires a certain amount of information exchange among the players and is not realizable in our network model because of the absence of the established control signaling links between the D2D pairs. For this reason, we develop a fully autonomous Q-learning algorithm where, instead of the actual knowledge, we estimate the players' beliefs about the strategies of all the other learning agents and prove the convergence of this method to a mixed-strategy NE.

It should be mentioned that the Q-learning based techniques have been considered in the past in the context of the various wireless networking problems (different from the one considered in this paper). For instance, in [23], the authors propose a Q-learning based dynamic channel assignment technique for mobile networks. In [24], a multi-agent Q-learning algorithm is proposed for interference avoidance in self-organized femtocell networks. In [25], the problem of channel selection in cognitive radio systems is solved using the multi-agent Q-learning. In all proposed techniques, each learning agent estimates the corresponding action-values based only on its local information (state, action, and reward) without estimating the strategies of all the other agents. Consequently, the dependency of the player's reward on the actions of other players is neglected in these algorithms (and the product of the players' strategies is omitted in algorithms' derivations). As a result of such simplistic formulation, the proposed algorithms can only guarantee the convergence to some near-optimal point [26] (unlike our method which converges to a NE state), which can be proved using the results of the stochastic approximation theory and, more precisely, the development of the ordinary differential equation (ODE) approach to stochastic approximation provided in [25], [27].

The rest of the paper is organized as follows. The network model is described in Section II. A stochastic game for joint channel and power level selection by D2D pairs and the multi-agent Q-learning method are formulated in Section III. The autonomous channel and power level selection algorithm based on player's beliefs is derived in Section IV. The implementation and performance evaluation of the proposed algorithm in LTE-A network is presented in Section V. Results of this work are summarized in Conclusion.

II. Problem Statement

A. Network Model

In this paper, we focus on the downlink (DL) operation of a D2D-enabled heterogeneous network. However, the analytical framework proposed in the paper can also be applied (with some certain modifications) to the uplink (UL) cellular channels.[1] We consider a D2D-enabled heterogeneous cellular network (illustrated in Figure 1) that consists of N BSs, numbered BS_1, …, BS_N. We denote by $\mathbf{N} = \{1, \ldots, N\}$ the set of the BSs' indices. Each BS in the network can be either a high-power BS serving the macrocell or low-power BS serving the small (micro, pico or femto) cell. It is assumed that the BSs operate on their own licensed spectrum bands that may overlap with each other. The BSs serve M D2D pairs and L cellular users. For notation consistency, the D2D pairs are numbered, interchangeably, as PU_{M+1}, …, PU_{N+M} or (U_{N+1}, $U_{N+1'}$), …, (U_{N+M}, $U_{N+M'}$), with $\mathbf{M} = \{N + 1, \ldots, N + M\}$ being the set of the D2D pairs' indices. The cellular users are denoted as U_{N+M+1}, …, U_{N+M+L}, with $\mathbf{L} = \{N + M + 1, \ldots, N + M + L\}$ being the set of the cellular users' indices.

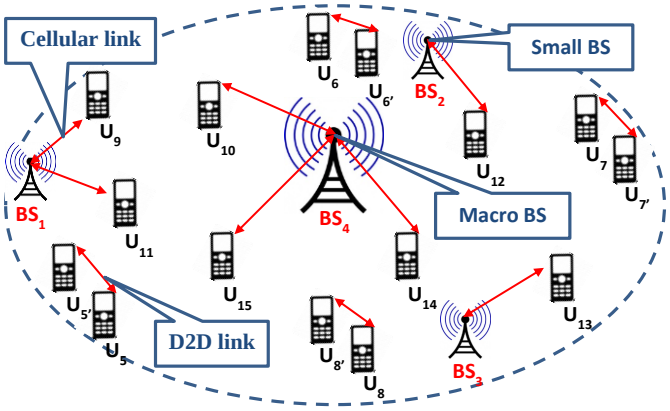

Figure 1. A D2D-enabled heterogeneous network with N = 4 BSs (one macro-BS and three small BSs), M = 4 D2D pairs and L = 7 cellular users.

The described system runs on a slotted-time basis with the time axis partitioned into the equal non-overlapping time intervals (slots) of the length T_s, with t denoting an integer-valued slot index. We follow the same lines, as [16] and assume that the communication of cellular and D2D users is synchronized by the timing signals sent by a cellular network or the global positioning system (GPS). The wireless channels are allocated to cellular users by the associated BSs according to some predetermined scheduling procedure (based on the global CSI that can be obtained through the use of conventional pilot signals [28]). Unlike cellular users, the D2D users operate fully autonomously (without any cooperation or information exchange among each other and with the BSs). Hence, they have neither precise nor statistical knowledge about the operating environment (such as channel quality or network traffic) and have to select the wireless channels and adjust their transmission power independently based only on their own local observations. It is assumed that each D2D pair decides its channel and power level at the beginning of each time slot t. We also assume that the each cellular/D2D user stays in the system for the indefinitely long time. That is, although the users eventually leave the network, they are uncertain about the exact duration of their stay.

In this paper, we study two different scenarios of D2D communication. In the first scenario, the D2D pairs and cellular users operate on different frequency bands (hence, creating no interference to each other). In the second scenario, the D2D pairs and cellular users transmit over the same channels. Consequently, we consider the network where K orthogonal channels, numbered C_1, ..., C_K, are available for D2D communication, with $\mathbf{K} = \{1, \ldots, K\}$ being the set of the corresponding channel indices. Note that in the first scenario, only the D2D users are allowed to transmit over the channels from the set $\mathbf{K}$. The cellular users operate on their dedicated channels belonging to some set $\mathbf{K^C}$, such that $\mathbf{K^C} \cap \mathbf{K} = \varnothing$, and the total available network bandwidth consists of the channels from the union set $\mathbf{K^C} \cup \mathbf{K}$.[2] In the second scenario, both the cellular and D2D users can transmit over the channels in the set $\mathbf{K}$ (that is, the total spectrum band of the network comprises the channels in the set $\mathbf{K}$).[3]

In the following, for each D2D pair PU_m and cellular user U_m, we define the binary channel allocation vector $\mathbf{c}^m_t = (c^m{}_1(t), \ldots, c^m{}_K(t))$, $m \in \mathbf{M} \cup \mathbf{L}$, with the elements $c^m{}_k(t)$, $k \in \mathbf{K}$, equalling 1 if PU_m/U_m transmits over the channel C_k at slot t, and 0 otherwise. For each cellular user U_l, we also define the binary BS association vector $\mathbf{b}^l_t = (b^l{}_1(t), \ldots, b^l{}_N(t))$, $l \in \mathbf{L}$, with the elements $b^l_n(t)$, $n \in \mathbf{N}$, equalling 1 if U_l is associated with the BS_n at slot t, and 0 otherwise. In our network model, the number of D2D pairs operating (simultaneously) on the same channels is unlimited. However, at any slot t, each D2D user can select at most one channel within the available bandwidth. That is,

$$\sum_{k \in \mathbf{K}} c^m_k(t) \le 1, \forall m \in \mathbf{M}. \tag{1a}$$

In the first scenario, all the channels in the set $\mathbf{K}$ are reserved for D2D communication only and we have

$$c^l_k(t) = 0, \forall l \in \mathbf{L}, \forall k \in \mathbf{K}, \tag{1b}$$

at any slot t. We also introduce the finite sets of possible channel selection decisions made by each PU_m, defined as

$$\mathbf{C}^m = \left\{ \mathbf{c}^m_t \;\middle|\; \sum_{k \in \mathbf{K}} c^m_k(t) \le 1 \right\}, \forall m \in \mathbf{M}. \tag{2}$$

B. Channel Model

Let $G^k{}_{m,j}(t)$, $j \in \mathbf{M} \cup \mathbf{L}$, $m \in \mathbf{M}$, $k \in \mathbf{K}$ denote the link gain of the channel between U_m and U_j operating on C_k at slot t and let $G^k{}_{n,j}(t)$, $n \in \mathbf{N}$ be the link gain of the channel between the BS_n and U_j operating on C_k at slot t. Note that the instantaneous values of $G^k{}_{n,l}(t)$ can be measured by all cellular users and the BSs for any $n \in \mathbf{N}$, $l \in \mathbf{L}$, $k \in \mathbf{K}$, through the use of the pilot

[1] The implementation of a proposed autonomous scheme for joint channel and power level selection by D2D users in the UL channels of a cellular network is discussed in Section IV.

[2] Since the focus of this paper is on D2D communication, we don't provide any details about the channels of cellular users and their operation.

[3] Note that the total bandwidth of the network is composed of the spectrum bands of different BSs (which may overlap with each other).

signals [28]. The D2D users, however, do not possess any prior information on the wireless channel quality. Therefore, the exact values of $G^k_{m,j}(t)$ and $G^k_{n,j}(t)$ are unknown to all D2D pairs (and the BSs).

In the first scenario, the interference to each D2D pair is created by the other D2D users operating on the same channel. Consequently, the SINR for a D2D pair PU_m transmitting over the wireless channel C_k equals

$$SINR^m_k(t) = \frac{S^m_k(t)}{I^{\mathbf{D}}_{m,k}(t) + \sigma^2}, \forall m \in \mathbf{M}, \forall k \in \mathbf{K} \quad (3a)$$

at any slot t. In (3a), σ^2 is the variance of zero-mean additive white Gaussian noise (AWGN) power; $S^m{}_k(t)$ is the useful signal power of PU_m over the channel C_k at slot t, given by

$$S^m_k(t) = G^k_{m,m'}(t) c^m_k(t) P^m(t), \quad (3b)$$

where $P^m(t) \le P_{\max}$ is the transmission power (that should not exceed some maximal allowed power level $P_{\max}$[4]) selected by PU_m at slot t; $I^{\mathbf{D}}{}_{m,k}(t)$ is the interference to PU_m from the D2D users operating on the channel C_k, defined, at any slot t, as

$$I^{\mathbf{D}}_{m,k}(t) = \sum_{j \in \mathbf{M} \setminus \{m\}} G^k_{j,m'}(t) c^j_k(t) P^j(t). \quad (3c)$$

In the second scenario, the interference to the D2D pairs is caused not only by other D2D users but also by cellular users operating on the same channels and associated with different BSs. The SINR for a D2D pair PU_m transmitting over the wireless channel C_k is equal

$$SINR^m_k(t) = \frac{S^m_k(t)}{I^{\mathbf{D}}_{m,k}(t) + I^{\mathbf{C}}_{m,k}(t) + \sigma^2}, \forall m \in \mathbf{M}, \forall k \in \mathbf{K} \quad (4a)$$

at any slot t, with $I^{\mathbf{C}}{}_{m,k}(t)$ being the interference to PU_m from the cellular users operating on the channel C_k, given by

$$I^{\mathbf{C}}_{m,k}(t) = \sum_{i \in \mathbf{N}} \sum_{j \in \mathbf{L}} G^k_{i,m'}(t) b^j_i(t) c^j_k(t) P^j(t), \quad (4b)$$

where $P^j(t)$ is the instantaneous transmission power in the DL channel between the cellular user U_j and its associated BS.

To be compatible with the proposed multi-agent Q-learning framework, we discretize the transmission power of the D2D pairs, as follows. For each PU_m, we consider J power levels, numbered P_1, …, P_J, and define the binary power level selection vector $\mathbf{p}^m{}_t = (p^m{}_1(t), \ldots, p^m{}_J(t))$, with the elements $p^m{}_j(t)$, $j \in \mathbf{J} = \{1, \ldots, J\}$, equalling 1 if PU_m selects to transmit at a power level P_j at slot t, and 0 otherwise. Since only one power level can be selected at any slot t, we have

$$\sum_{j \in \mathbf{J}} p^m_j(t) \le 1, \forall m \in \mathbf{M}. \quad (5a)$$

Given (5a), the instantaneous transmission power of each PU_m can be calculated according to

$$P^m(t) = \frac{P_{\max}}{J} \sum_{j \in \mathbf{J}} j \cdot p^m_j(t), \forall m \in \mathbf{M}. \quad (5b)$$

We also define a finite set of possible power level selection decisions made by PU_m, as

$$\mathbf{P}^m = \left\{ \mathbf{p}^m_t \;\middle|\; \sum_{j \in \mathbf{J}} p^m_j(t) \le 1 \right\}, \forall m \in \mathbf{M}. \quad (6)$$

[4] The settings of a maximal allowed power level $P_{\max}$ for D2D users will be discussed in Section IV.

C. Problem Statement

The main goal of the proposed autonomous scheme for joint channel and power level selection is to ensure that there are no D2D pairs with the SINRs falling below some predefined thresholds. That is,

$$SINR^m(t) = \sum_{k \in \mathbf{K}} SINR^m_k(t) \ge SINR^m_{\min}, \forall m \in \mathbf{M} \quad (7)$$

where $SINR^m{}_{\min}$ is the minimal satisfactory SINR level of PU_m. If, at the current slot t, the constraint in (7) is satisfied for some PU_m then this D2D pair obtains a reward $u^m{}_t$, defined as the difference between the throughput and the cost of power consumption achieved by selecting the specific channel and power level.[5] Otherwise (if the constraint in (7) is not satisfied for PU_m), it receives a zero reward. Consequently, we can express the reward $u^m{}_t$ obtained by PU_m at slot t, as

$$u^m_t = \begin{cases} R^m(t) - \upsilon^m P^m(t), & \text{if } SINR^m(t) \ge SINR^m_{\min}, \\ 0, & \text{otherwise} \end{cases} \quad (8a)$$

for all $m \in \mathbf{M}$. In (8a), $\upsilon_m \ge 0$ is the cost per unit (Watt) level of power, the throughput $R^m(t)$ is given by

$$R^m(t) = \omega \log(1 + SINR^m(t)) = \omega \log\left(1 + \sum_{k \in \mathbf{K}} SINR^m_k(t)\right), \quad (8b)$$

where ω is the channel bandwidth, and the transmission power $P^m(t)$ is calculated using (5b). Note that at any slot t, the instantaneous reward of PU_m in (8a) depends on

- the individual channel and power level decisions of PU_m, $\mathbf{c}^m{}_t$ and $\mathbf{p}^m{}_t$, and the current link gain in its channel $G^k{}_{m,m}(t)$ (which are known or can be measured by PU_m);
- the channels and power levels selected by the other users and the link gains in their channels (which cannot be observed by PU_m).

In our network, at any slot t, each PU_m selects the transmission channel and power level to maximize its long-term reward $U^m{}_t$, defined as a sum of the instantaneous rewards that this D2D pair receives over the infinite (indefinitely-long) period in the future.[6] To ensure the finiteness of this sum, we introduce the discounting [29] of the future rewards relative to the earlier rewards, which is used to model the situations where the users are uncertain about the duration of their stay in the system [29]. A common assumption is that a user wants to maximize a weighted sum of its future instantaneous rewards where the weights of the later periods are less than the weights of the earlier periods. For simplicity, this assumption often takes a particular form that the sequence of weights forms a geometric progression: for some fixed $\gamma \in [0, 1)$, each weighting factor is γ times the previous weight (γ is called the discount rate) [29]. In this

[5] Note that achieving the high SINR regime (and, consequently, throughput) requires each D2D user to transmit at a maximal power level which, in turn, results in increased power consumption and magnifies the interference to the other network users. Therefore, when formulating the reward of the D2D pairs, we should also take into account the cost of power consumption to quantify the tradeoff (as in [25]) between the achieved throughput and power level.

[6] Note that in our model, the users are uncertain about the time when they will leave the system. Hence, we are not able to predict the duration of their stay in the system.

case, the long-term reward of the D2D pair PU_m can be expressed as

$$U_t^m = \sum_{\tau=t}^{+\infty} \gamma^{t-\tau} u_\tau^m. \quad (9)$$

We now introduce the set of possible channel and power level decisions made by each D2D pair PU_m, defined as $\mathbf{A}^m = \mathbf{C}^m \times \mathbf{P}^m$ (where × denotes the Cartesian product). Then, the objective of each PU_m is to select, at any slot t, the pair $(\bar{\mathbf{c}}_t^m, \bar{\mathbf{p}}_t^m) \in \mathbf{A}^m$ which maximizes its long-term reward in (9). That is,

$$(\bar{\mathbf{c}}_t^m, \bar{\mathbf{p}}_t^m) = \arg\max_{(\mathbf{c}_t^m, \mathbf{p}_t^m) \in \mathbf{A}^m} U_t^m. \quad (10)$$

In the following, we will use the subscript t to denote the values of the parameters and functions at slot t. We will ignore a subscript t when focusing on one slot of the decision process to simplify the notation.

III. Stochastic Non-Cooperative Game for Joint Channel and Power Level Selection by D2D Pairs

A. Game Formulation

Let us formulate the considered joint channel and power level selection problem as a dynamic non-cooperative game with M players (D2D pairs) having no information about the operating environment (channel quality and network traffic). It is assumed that all players are selfish and rational and, at any slot t, execute their actions (determined by the players' decisions on their transmission channels and power levels) with the objective to maximize their long-term rewards in (9). Consequently, the action space of the player PU_m is equivalent to the finite set of its possible channel and power level decisions, $\mathbf{A}^m = \mathbf{C}^m \times \mathbf{P}^m$. The action executed by PU_m at slot t, $a^m{}_t = (\mathbf{c}^m{}_t, \mathbf{p}^m{}_t) \in \mathbf{A}^m$, consists of two parts representing, respectively, the channel and power level selected by this player at a given time slot. For each PU_m, we also introduce the vector of actions taken by the other $M-1$ players at slot t, defined as $\mathbf{a}^{-m}{}_t = (a^1{}_t, \ldots, a^{m-1}{}_t, a^{m+1}{}_t, \ldots, a^M{}_t) \in \mathbf{A}^{-m}$, where $\mathbf{A}^{-m} = \times_{i \in \mathbf{M} \setminus \{m\}} \mathbf{A}^i$.

In our network, at any slot t, the throughput $R^m(t)$ of the player PU_m depends on the current SINR in its channel (see equation (8b)) which is determined by the players' actions $(a^m{}_t, \mathbf{a}^{-m}{}_t)$ and the instantaneous values of the link gain coefficient matrix $\mathbf{G}^m{}_t$, given by

$$\mathbf{G}_t^m = \begin{bmatrix} G^1_{1,m'}(t) & \ldots & G^K_{1,m'}(t) \\ \vdots & \vdots & \vdots \\ G^1_{N+M,m'}(t) & \ldots & G^K_{N+M,m'}(t) \end{bmatrix}, \forall m \in \mathbf{M}. \quad (11)$$

Indeed, the instantaneous SINR for each D2D pair (i.e., $SINR^m(t)$) in the first and second scenarios, can be expressed, for all $m \in \mathbf{M}$, as

$$SINR^m(a_t^m, \mathbf{a}_t^{-m}, \mathbf{G}_t^m) = \sum_{k \in \mathbf{K}} \frac{S_k^m(a_t^m, \mathbf{G}_t^m)}{I^{\mathbf{D}}_{m,k}(\mathbf{a}_t^{-m}, \mathbf{G}_t^m) + \sigma^2} \quad (12a)$$

and

$$SINR^m(a_t^m, \mathbf{a}_t^{-m}, \mathbf{G}_t^m) = \sum_{k \in \mathbf{K}} \frac{S_k^m(a_t^m, \mathbf{G}_t^m)}{I^{\mathbf{D}}_{m,k}(\mathbf{a}_t^{-m}, \mathbf{G}_t^m) + I^{\mathbf{C}}_{m,k} + \sigma^2}, \quad (12b)$$

respectively, where $S^m{}_k(t)$ and $I^{\mathbf{D}}{}_{m,k}(t)$ are presented explicitly as the functions of $a^m{}_t$, $\mathbf{a}^{-m}{}_t$, and $G^m{}_t$. Note that at any slot t, each player PU_m can measure its current SINR level $SINR^m(t)$.

The instantaneous reward $u^m{}_t$ obtained by PU_m after taking some action $a^m{}_t$ equals zero if the current SINR in its channel is less than the predefined threshold $SINR^m{}_{\min}$ (see equation (8a)). Accordingly, at any slot t, for each PU_m we can define the state of the game $s^m{}_t$, as

$$s_t^m = \begin{cases} 1, & \text{if } SINR^m(a_t^m, \mathbf{a}_t^{-m}, \mathbf{G}_t^m) \geq SINR^m_{\min}, \forall m \in \mathbf{M}. \\ 0, & \text{othewise} \end{cases} \quad (13)$$

The state $s^m{}_t$ is fully observable (since it depends on the values of $SINR^m(t)$ that can be measured at any slot t). Using (13), the instantaneous reward $u^m{}_t$ received by PU_m at slot t, can be expressed as

$$u_t^m = u^m(a_t^m, \mathbf{a}_t^{-m}, s_t^m) = s_t^m [R^m(a_t^m, \mathbf{a}_t^{-m}, \mathbf{G}_t^m) - \upsilon^m P^m(a_t^m)], \quad (14a)$$

for all $m \in \mathbf{M}$, where

$$R^m(a_t^m, \mathbf{a}_t^{-m}, \mathbf{G}_t^m) = \omega \log(1 + SINR^m(a_t^m, \mathbf{a}_t^{-m}, \mathbf{G}_t^m)), \quad (14b)$$

$SINR^m(a^m{}_t, \mathbf{a}^{-m}{}_t, \mathbf{G}^m{}_t)$ is determined from (12a) or (12b) (depending on the considered scenario), and the instantaneous transmission power $P^m(t)$ is expressed explicitly as a function of $a^m{}_t$.

It follows from (14a) that, at any slot t, the reward $u^m{}_t$ obtained by PU_m depends on the current (fully-observable) state $s^m{}_t$ and partially-observable actions $(a^m{}_t, \mathbf{a}^{-m}{}_t)$. At the next slot t, the game moves to a new random state $s^m{}_{t+1}$ whose distribution depends on the previous state $s^m{}_t$ and the selected actions $(a^m{}_t, \mathbf{a}^{-m}{}_t)$. This procedure repeats for the indefinite number of slots. At any t, the player PU_m can observe its state $s^m{}_t$ and action $a^m{}_t$ but does not know the actions of other players, $\mathbf{a}^{-m}{}_t$, and the precise values of $\mathbf{G}^m{}_t$. The state transition probabilities (from the current state $s^m{}_t$ to next state $s^m{}_{t+1}$ given the executed actions $(a^m{}_t, \mathbf{a}^{-m}{}_t)$) are also unknown to each player PU_m. Consequently, the considered game is stochastic, as it follows from its definition in [30].[7] Hence, we can characterize this game by the tuple $\mathbf{\Gamma} = (\mathbf{S}, \mathbf{M}, \mathbf{A}, T, \mathbf{U})$ comprising the following respective components:

- the finite set of possible states, $\mathbf{S} = \times_{m \in \mathbf{M}} \mathbf{S}^m$, with $\mathbf{S}^m = \{0, 1\}$, for all $m \in \mathbf{M}$;
- the finite set of players $\mathbf{M}$;
- the finite set of action profiles, $\mathbf{A} = \times_{m \in \mathbf{M}} \mathbf{A}^m$;
- the state transition probability function $T(s^m, \mathbf{a}, s'^m) = \Pr\{s'^m \mid s^m, \mathbf{a}\}$, defined as the probability of transitioning to

[7] In game theory, a stochastic game is a dynamic game with the probabilistic transitions. It generalizes both MDPs and repeated games and is played in a sequence of stages. At the beginning of each stage, the game is in some state. The players select actions and each player receives a reward that depends on the current observable state and the chosen actions. The game then moves to a new random state whose distribution depends on the previous state and actions. The procedure is repeated at the new state and play continues for a finite or the infinite number of stages. The total reward received by the player is often taken to be the discounted sum of the stage rewards or the limit inferior of the averages of the stage rewards. Note that, unlike Bayesian games where some states are partially-observable (and to deal with such incomplete information about the states, the players form their beliefs about these states) [31], in a stochastic game the states are fully-observable (although the probabilities of transitioning from one state to another in a single stage are known) [30].

the next state $s'^m = s^m{}_{t+1} \in \mathbf{S}^m$ given the joint action $\mathbf{a} = (a^1{}_t, \ldots, a^M{}_t) \in \mathbf{A}$, executed in state $s^m = s^m{}_t \in \mathbf{S}^m$;
- the vector of player rewards, $\mathbf{U} = (u^1, \ldots, u^M)$.

A solution to this game is represented by some action vector $(\bar{a}^m, \bar{\mathbf{a}}^{-m}) \in \mathbf{A}$, that corresponds to the NE where the following inequalities hold for each PU_m in any $s^m \in \mathbf{S}^m$ [29]:

$$u^m(\bar{a}^m, \bar{\mathbf{a}}^{-m}, s^m) \geq u^m(a^m, \bar{\mathbf{a}}^{-m}, s^m), \forall a^m \in \mathbf{A}^m. \quad (15)$$

That is, in the NE state, the action of each player is the best response to the actions of the other players. Hence, no players can gain by unilateral deviation [29].

We also introduce a mixed strategy profile $\pi^m : \mathbf{S}^m \rightarrow \mathbf{A}^m$ (mapping from states to actions) of the player PU_m in state s^m, denoted as $\pi^m(s^m) = \{\pi^m(s^m, a^m)\}_{a^m \in \mathbf{A}^m}$. Each element $\pi^m(s^m, a^m)$ of $\pi^m(s^m)$ is the probability with which a D2D pair PU_m selects an action $a^m \in \mathbf{A}^m$ in state $s^m \in \mathbf{S}^m$. For each PU_m, we also define the vector of strategies $\boldsymbol{\pi}^{-m}(s) = (\pi^1(s^1), \ldots, \pi^{m-1}(s^{m-1}), \pi^{m+1}(s^{m+1}), \ldots, \pi^M(s^M))$ of the other $M - 1$ players. The presence of incomplete information (about the reward structure and state transition probabilities) in a formulated non-cooperative game provides opportunities for the players to learn their optimal strategies through repeated interactions with the stochastic environment and each player PU_m becomes a learning agent whose task is to find a NE strategy $\bar{\pi}^m$ for any state s^m.[8] In the next subsection, we model the dynamic of our game using a finite-state MDP and derive the multi-agent Q-learning process [17] for joint channel and power level selection by D2D pairs.[9]

B. Multi-Agent Q-Learning

Q-learning is a RL method for solving the problems modeled after MDPs, where a learning agent operates in an unknown stochastic environment [17]. In multi-agent Q-learning, each m^{th} agent i) observes its environment, using some representation of the environment state s^m, ii) selects one of the available actions a^m, and iii) receives, as a consequence of its action a^m, the immediate response from the environment in the form of reward u^m. When the rewards are based only on the current state and action (without any influence of previous states and actions), the task of the m^{th} agent is said to satisfy the Markov property and can be formally defined as the MDP consisting of i) a discrete set of environment states $\mathbf{S}^m$, ii) a discrete set of possible actions $\mathbf{A}^m$, iii) a one-slot dynamics of the environment given by the state transition probabilities $T_{s^m s'^m} = T(s^m, \mathbf{a}, s'^m)$, for all $a^m \in \mathbf{A}^m$ and $s^m, s'^m \in \mathbf{S}^m$.

The solution to the above MDP is then to select, for each state $s^m \in \mathbf{S}^m$, an optimal strategy $\bar{\pi}^m(s^m)$ that maximizes its value-state function V, formally defined as [33]

$$V(s^m, \pi^m, \boldsymbol{\pi}^{-m}) = \mathrm{E}\left\{ \sum_{\tau=t}^{+\infty} \gamma^{t-\tau} u^m_\tau \;\middle|\; s^m_t = s^m \right\}$$
$$= \mathrm{E}\left\{ \sum_{\tau=t}^{+\infty} \gamma^{t-\tau} u^m[\pi^m(s^m_\tau), \boldsymbol{\pi}^{-m}(s^m_\tau), s^m_\tau] \;\middle|\; s^m_t = s^m \right\} \quad (16a)$$
$$= \mathrm{E}\left\{ u^m[\pi^m(s^m), \boldsymbol{\pi}^{-m}(s^m), s^m] \right\}$$
$$+ \gamma \sum_{s'^m \in \mathbf{S}^m} T_{s^m s'^m}[\pi^m(s^m), \boldsymbol{\pi}^{-m}(s^m)] V(s'^m, \pi^m, \boldsymbol{\pi}^{-m}),$$

for all $m \in \mathbf{M}$, where $\mathrm{E}\{x\}$ is the expectation of x, and

$$\mathrm{E}\left\{ u^m(\pi^m(s^m), \boldsymbol{\pi}^{-m}(s^m), s^m) \right\} = \sum_{s'^m \in \mathbf{S}^m} T_{s^m s'^m}[\pi^m(s^m), \boldsymbol{\pi}^{-m}(s^m)]$$
$$\times \sum_{(a^m, \mathbf{a}^{-m}) \in \mathbf{A}} \left\{ u^m(a^m, \mathbf{a}^{-m}, s^m) \prod_{i \in \mathbf{M}} \pi^i(s^i, a^i) \right\}. \quad (16b)$$

Note that the value-state function determines the expected return (a sum of the expected discounted rewards that an agent receives over the infinite time in the future) that can be obtained from each state on every possible policy.

Consequently, a strategy tuple $(\bar{\pi}^m, \bar{\boldsymbol{\pi}}^{-m})$, with $\bar{\boldsymbol{\pi}}^{-m} = (\bar{\pi}^1, \ldots, \bar{\pi}^{m-1}, \bar{\pi}^{m+1}, \ldots, \bar{\pi}^M)$, is a NE if the following inequalities hold for any π^m:

$$V(s^m, \bar{\pi}^m, \bar{\boldsymbol{\pi}}^{-m}) \geq V(s^m, \pi^m, \bar{\boldsymbol{\pi}}^{-m}), \forall s^m \in \mathbf{S}^m. \quad (17)$$

Since any finite game in a strategic form has the mixed-strategy equilibrium [29], there always exists a NE in our game which satisfies the following Bellman's optimality equation [33]:

$$\bar{V}(s^m, \pi^m, \boldsymbol{\pi}^{-m}) = V(s^m, \bar{\pi}^m, \bar{\boldsymbol{\pi}}^{-m})$$
$$= \max_{a^m \in \mathbf{A}^m} \Big[\mathrm{E}\left\{ u^m[a^m, \bar{\boldsymbol{\pi}}^{-m}(s^m), s^m] \right\} \quad (18a)$$
$$+ \gamma \sum_{s'^m \in \mathbf{S}^m} T_{s^m s'^m}[a^m, \bar{\boldsymbol{\pi}}^{-m}(s^m)] V(s'^m, \bar{\pi}^m, \bar{\boldsymbol{\pi}}^{-m}) \Big]$$

for all $m \in \mathbf{M}$, where

$$\mathrm{E}\left\{ u^m(a^m, \bar{\boldsymbol{\pi}}^{-m}(s^m), s^m) \right\} = \sum_{s'^m \in \mathbf{S}^m} T_{s^m s'^m}[a^m, \bar{\boldsymbol{\pi}}^{-m}(s^m)]$$
$$\times \sum_{\mathbf{a}^{-m} \in \mathbf{A}^{-m}} \left\{ u^m(a^m, \mathbf{a}^{-m}, s^m) \prod_{i \in \mathbf{M} \setminus \{m\}} \bar{\pi}^i(s^i, a^i) \right\}. \quad (18b)$$

Let us further define the function

$$\bar{Q}(s^m, a^m) = \mathrm{E}\left\{ \sum_{\tau=t}^{+\infty} \gamma^{t-\tau} u^m_\tau \;\middle|\; s^m_t = s^m, a^m_t = a^m \right\}$$
$$= \mathrm{E}\left\{ u^m[a^m, \bar{\boldsymbol{\pi}}^{-m}(s^m), s^m] \right\} \quad (19)$$
$$+ \gamma \sum_{s'^m \in \mathbf{S}^m} T_{s^m s'^m}[a^m, \bar{\boldsymbol{\pi}}^{-m}(s^m)] V(s'^m, \bar{\pi}^m, \bar{\boldsymbol{\pi}}^{-m})$$

which is called an optimal action-value function (or simply, an action-value) and measures the maximum expected return for taking an action a^m in state s^m, and thereafter following an optimal strategy [17], [33]. Although more than one NE strategy $\bar{\pi}^m(s^m)$ can exist for each learning agent, all of them will have the same action-values $\bar{Q}(s^m, a^m)$ and

$$V(s^m, \bar{\pi}^m, \bar{\boldsymbol{\pi}}^{-m}) = \max_{a^m \in \mathbf{A}^m} \bar{Q}(s^m, a^m). \quad (20)$$

[8] Note that if the structure of reward functions and state transitions were known to all the players, we could find the NE strategy using any integer programming (IP) method. However, since each player has no information about the rewards of other players and probabilities of state transitions, it has to learn its environment to find the NE strategy for each state of the game.

[9] Note that in case if some states in the game were partially-observable (as in Bayesian games), the game dynamics should be modeled using partially-observable MDPs (POMDPs).

$$Q_{t+1}(s^m,a^m)\leftarrow Q_t(s^m,a^m)+\begin{cases}\alpha_t\left[\sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}}\left\{u^m[a^m,\mathbf{a}^{-m},s^m]\prod_{i\in\mathbf{M}\setminus\{m\}}\pi^i(s^i,a^i)\right\}+\gamma\max_{a'^m\in\mathbf{A}^m}Q_t(s'^m,a'^m)-Q_t(s^m,a^m)\right], & \text{if } s_t^m=s^m,a_t^m=a^m;\\ 0, & \text{otherwise}\end{cases} \quad (22)$$

$$\hat{u}^m_{t+1}(s^m,a^m)\leftarrow\hat{u}^m_t(s^m,a^m)+I_{s_t^m=s^m,a_t^m=a^m}\,\alpha_t\left[\sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}}\left\{u^m[a^m,\mathbf{a}^{-m},s^m]Y(s^m,\mathbf{a}^{-m})\right\}+\gamma\max_{a'^m\in\mathbf{A}^m}\hat{u}^m_t(s'^m,a'^m)-\hat{u}^m_t(s^m,a^m)\right],\ \forall m\in\mathbf{M}. \quad (23)$$

Combining (18a) and (20), we get

$$\overline{Q}(s^m,a^m)=\mathrm{E}\left\{u^m[a^m,\overline{\boldsymbol{\pi}}^{-m}(s^m),s^m]\right\}+\gamma\sum_{s'^m\in\mathbf{S}^m}T_{s^m s'^m}[a^m,\overline{\boldsymbol{\pi}}^{-m}(s^m)]\max_{a'^m\in\mathbf{A}^m}\overline{Q}(s'^m,a'^m). \quad (21)$$

That is, an optimal action-value function can be obtained recursively from the corresponding action-values. In Q-learning [33], each agent learns the optimal action-values using the updating rule, described by (22) for all $s^m\in\mathbf{S}^m$, $a^m\in\mathbf{A}^m$, $m\in\mathbf{M}$, with $\alpha_t\in[0, 1)$ and Q_t being, respectively, the learning rate and action-value at slot t.

IV. Learning Based Algorithm for Autonomous Channel and Power Level Selection by D2D Pairs

Note that the multi-agent Q-learning algorithm given by (22) implies that each D2D pair has to possess not only the local information about its own strategy and reward but also the information regarding the strategies of the other players. In the following, we develop a fully autonomous channel selection algorithm (which will not rely on the availability of the global information about the strategies of the players) based on the multi-agent Q-learning process in (22).

A. Autonomous Q-Learning Based on Players' Beliefs

The Q-learning method formulated in (22) can be applied to our channel selection game to learn the players' rewards. The corresponding recursive learning process for each player PU_m is given by (23), for all $s^m\in\mathbf{S}^m$, $a^m\in\mathbf{A}^m$. In (23), $\hat{u}_t^m(s^m,a^m)\leftarrow Q_t(s^m,a^m)$ is the estimated reward of a player PU_m obtained by playing the action $a^m{}_t$ in state $s^m{}_t$, I is the indicator function, and the function $Y(s^m,\mathbf{a}^{-m})$ is given by

$$Y(s^m,\mathbf{a}^{-m})=\prod_{i\in\mathbf{M}\setminus\{m\}}\pi^i(s^i,a^i),\ \forall m\in\mathbf{M}. \quad (24)$$

In a NE state, the strategies of the players remain unaltered. Each player in this state can be viewed as a learning agent behaving optimally with respect to his beliefs about the strategies of all the other players (that is, with respect to $Y(s^m,\mathbf{a}^{-m})$). The probability that PU_m achieves a reward $u^m(a^m,\mathbf{a}^{-m},s^m)$ at slot t is equal to the product $\eta^m=\pi^m(s^m,a^m)Y(s^m,\mathbf{a}^{-m})$. In the following, we use notation t_c to denote the number of slots between any two consecutive events of PU_m achieving the same reward value $u^m(a^m,\mathbf{a}^{-m},s^m)$. Note that t_c has an independent and identical distribution with η^m and, hence,

$$E\{\eta^m\}=\frac{1}{E\{t_c\}}. \quad (25a)$$

Based on (25a), each PU_m can estimate $Y(s^m,\mathbf{a}^{-m})$ according to

$$\hat{Y}(s^m,\mathbf{a}^{-m})\approx\frac{1}{E\{t_c\}\,\pi^m(s^m,a^m)},\ \forall m\in\mathbf{M}. \quad (25b)$$

Let us further introduce the reference points $\dot{Y}(s^m,\mathbf{a}^{-m})$ and $\dot{\pi}^m(s^m,a^m)$ which represent some specific values of the belief $Y(s^m,\mathbf{a}^{-m})$ and strategy $\pi^m(s^m,a^m)$, respectively. Then,

$$\begin{aligned}\hat{Y}(s^m,\mathbf{a}^{-m})-\dot{Y}(s^m,\mathbf{a}^{-m})&\approx\frac{1}{\mathrm{E}\{t_c\}}\left\{\frac{1}{\pi^m(s^m,a^m)}-\frac{1}{\dot{\pi}^m(s^m,a^m)}\right\} \quad (26a)\\ &=\beta_t^m[\dot{\pi}^m(s^m,a^m)-\pi^m(s^m,a^m)],\ \forall m\in\mathbf{M}\end{aligned}$$

where

$$\beta_t^m=\frac{1}{\mathrm{E}\{t_c\}\,\pi^m(s^m,a^m)\dot{\pi}^m(s^m,a^m)},\ \forall m\in\mathbf{M} \quad (26b)$$

is some positive scalar. When operating fully autonomously, the players can modify the reference points based on their past local observations. In this paper, we propose to update the reference points as

$$\dot{Y}(s^m,\mathbf{a}^{-m})=Y_{t-1}(s^m,\mathbf{a}^{-m}),\ \dot{\pi}^m(s^m,a^m)=\pi^m_{t-1}(s^m,a^m) \quad (27)$$

where $Y_{t-1}(s^m,\mathbf{a}^{-m})$ and $\pi^m{}_{t-1}(s^m,a^m)$ are, respectively, the belief and strategy reference points at past slot $t-1$. Combining (26a) and (27), we get the following updating rule for belief estimation:

$$\hat{Y}(s^m,\mathbf{a}^{-m})=Y_{t-1}(s^m,\mathbf{a}^{-m})-\beta_t^m[\pi^m(s^m,a^m)-\pi^m_{t-1}(s^m,a^m)] \quad (28)$$

for all $m\in\mathbf{M}$. Then, the autonomous multi-agent Q-learning method given by (23) transforms to (29) for all $s^m\in\mathbf{S}^m$, $a^m\in\mathbf{A}^m$, where the belief estimates are obtained from (28).

Another challenging issue of Q-learning is the trade-off between exploration and exploitation. To obtain the higher reward, a learning agent must exploit the actions that had been effective in the past. But to discover such actions, it has to explore the actions which have not been selected previously. Hence, both the exploration and exploitation should be used conjointly to complete the learning task [17]. A rather typical approach to deal with the problem of exploration/exploitation is ε-greedy selection [34]. With this approach, the best action is selected for a proportion $1-\varepsilon$ of the trials; any other action is chosen uniformly at random for a proportion ε.

$$\hat{u}^m_{t+1}(s^m,a^m)\leftarrow\hat{u}^m_t(s^m,a^m)+I_{s_t^m=s^m,a_t^m=a^m}\,\alpha_t\left[\sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}}\left\{u^m[a^m,\mathbf{a}^{-m},s^m]\hat{Y}(s^m,\mathbf{a}^{-m})\right\}+\gamma\max_{a'^m\in\mathbf{A}^m}\hat{u}^m_t(s'^m,a'^m)-\hat{u}^m_t(s^m,a^m)\right],\ \forall m\in\mathbf{M}. \quad (29)$$

Algorithm: Autonomous multi-agent Q-learning Process Based on Players' Beliefs

Initialization:

Set time $t \leftarrow 0$ and the parameters γ, c_α, c_β, ζ_α;

For all $m \in \mathbf{M}$ **do**

{

Set the parameter $\zeta^m{}_\beta$;

For all $s^m \in \mathbf{S}^m$, $a^m \in \mathbf{A}^m$ **do**

{

Initialize the action-value $\hat{u}_t^m(s^m, a^m) \leftarrow 0$, strategy $\pi^m(s^m, a^m) \leftarrow \frac{1}{|\mathbf{A}^m|} = \frac{1}{KJ}$, and belief $\hat{Y}(s^m, \mathbf{a}^{-m}) \leftarrow 0$;

Initialize the state $s^m \leftarrow s^m{}_t = 0$;

}

Main Loop:

While ($t < T$) **do**

{

For all PU_m, $m \in \mathbf{M}$ **do**

{

Update the learning rate α_t and the parameter $\beta^m{}_t$ according to **(31)**;

Select an action a^m according to the strategy $\pi^m(s^m)$;

Measure the achieved SINR at the receiver;

If ($SINR^m(t) \geq SINR^m{}_{\min}$) **then set** $s^m{}_t \leftarrow 1$ **else set** $s^m{}_t \leftarrow 0$;

Update the instantaneous reward $u^m{}_t$ according to **(14a)**;

Update the action-value $\hat{u}^m_{t+1}(s^m, a^m)$ according to **(29)**;

Update the strategy $\pi^m(s^m, a^m)$ according to **(30)**;

Update the belief $\hat{Y}(s^m, \mathbf{a}^{-m})$ according to **(28)**;

Update time $t \leftarrow t + 1$ and the state $s^m \leftarrow s^m{}_t$;

}

}

Figure 2. Autonomous multi-agent Q-learning Process Based on Players' Beliefs.

On the other hand, the probability of selecting an action can be expressed as a function of the respective action-value (in our case, estimated reward $\hat{u}_t^m(s^m, a^m)$).Then, all actions will be weighted according to their action-value, so that the best action will be selected with higher probability. The most common way for such action selection is through the use of the Boltzmann-Gibbs (BG) distribution, given by [17]

$$\pi^m(s^m, a^m) = \frac{e^{\hat{u}_t^m(s^m, a^m)/T_B}}{\sum_{a'^m \in \mathbf{A}^m} e^{\hat{u}_t^m(s^m, a'^m)/T_B}}, \forall m \in \mathbf{M} \quad (30)$$

where T_B is the Boltzmann temperature. High temperatures make all actions almost equiprobable, whereas low temperatures result in the large difference of selection probabilities [17].

The proof of the convergence of the proposed autonomous multi-agent Q-learning method given by (29) is provided in Appendix B. The learning rate α_t and the parameters $\beta^m{}_t$ are set in accordance with the conditions specified in [35], as

$$\alpha_t = \frac{1}{(t + c_\alpha)^{\zeta_\alpha}}, \beta_t^m = \frac{1}{(t + c_\beta)^{\zeta^m_\beta}}, \forall m \in \mathbf{M} \quad (31)$$

where $c_\alpha > 0$, $c_\beta > 0$, $\zeta_\alpha \in (1/2, 1]$, $\zeta^m{}_\beta \in (1/2, 1]$. The algorithm for autonomous channel and power level selection by D2D pairs is outlined in Figure 2.

B. Implementation Issues

In this subsection, we consider the settings of a maximal allowed power level $P_{\max}$ for D2D users and discuss the implementation of the proposed Q-learning based channel and power level selection method in the UL direction of a D2D-enabled cellular network. Note that in the first scenario, the D2D pairs operate on the dedicated spectrum creating no interference to cellular users. Consequently, the maximal allowed power level of each D2D user is limited only by the physical characteristics of its mobile terminal (described by some maximal possible power level $\mathbf{P}_{\max}$). Hence, we can set $P_{\max} = \mathbf{P}_{\max}$, for all D2D users. On the other hand, to prevent high interference caused by D2D communication to cellular users, a great care should be taken when setting the maximal allowed power level in the second scenario.

Consider the interference to cellular users in the case when all D2D users transmit with the power level $P_{\max}$ and all cellular users transmitting at the maximal possible power $\mathbf{P}_{\max}$. For cellular user U_l associated with some BS_n and operating on a shared cellular/D2D channel C_k, the SINR equals

$$SINR^l_{n,k}(t) = \frac{S^l_{n,k}(t)}{I^{\mathbf{D}}_{l,k}(t) + I^{\mathbf{C}}_{n,l,k}(t) + \sigma^2}, \forall n \in \mathbf{N}, \forall k \in \mathbf{K}, \forall l \in \mathbf{L} \quad (32a)$$

where $S^l_{n,k}(t)$ is the current useful signal power of U_l over the channel C_k associated with the BS_n, given by

$$S^l_{n,k}(t) = \mathbf{P}_{\max} G^k_{n,l}(t) b^l_n(t) c^l_k(t), \quad (32b)$$

$I^{\mathbf{D}}_{l,k}(t)$ is the interference to U_l from the D2D users operating on the channel C_k at slot t, defined as

$$I^{\mathbf{D}}_{l,k}(t) = P_{\max} \sum_{m \in \mathbf{M}} G^k_{m,l}(t) c^m_k(t), \quad (32c)$$

and $I^{\mathbf{C}}_{n,l,k}(t)$ is the instantaneous interference to U_l associated with the BS_n from other cellular users operating on the channel C_k and associated with different BSs, described by

$$I^{\mathbf{C}}_{n,l,k}(t) = \mathbf{P}_{\max} \sum_{i \in \mathbf{N} \setminus \{n\}} \sum_{m \in \mathbf{L} \setminus \{l\}} G^k_{i,m}(t) b^m_i(t) c^m_k(t). \quad (32d)$$

Let $SINR^l_{\min}$ be some minimum tolerable SINR level below which the transmission of cellular user U_l is considered unsatisfactory. Then, $P_{\max}$ should be kept at the level

$$P_{\max} \le \frac{S^l_{n,k}(t) - SINR^l_{\min}(I^{\mathbf{C}}_{n,l,k}(t) + \sigma^2)}{SINR^l_{\min} \sum_{m \in \mathbf{M}} G^k_{m,l}(t) c^m_k(t)}. \quad (33)$$

The right-hand side of inequality in (33) depends on i) the BS and channel allocations to all cellular users, $\mathbf{b}^l_t$ and $\mathbf{c}^l_t$, $l \in \mathbf{L}$, (which are determined by the BSs), ii) the autonomous channel selection decisions of the D2D pairs, $\mathbf{c}^m_t$, $m \in \mathbf{M}$, iii) the values of $G^k_{n,l}(t)$ and the link gain coefficient matrix $\mathbf{G}^l_t$, given by

$$\mathbf{G}^l_t = \begin{bmatrix} G^1_{1,l}(t) & \dots & G^K_{1,l}(t) \\ \vdots & \vdots & \vdots \\ G^1_{N+M,l}(t) & \dots & G^K_{N+M,l}(t) \end{bmatrix}, \forall l \in \mathbf{L}, \quad (34)$$

and iv) the AWGN power variance σ^2.

Since we do not know the BS and channel allocations to cellular users, the channel selection decisions made by the D2D pairs, and the instantaneous values of the link gains, we can only make some simplistic assumptions about these parameters. In particular, we assume that the probability that a cellular user U_l is associated with a particular BS and wireless channel and the probability that the D2D pair PU_m selects a specific wireless channel are distributed uniformly at random, so that

$$\Pr\left\{b^l_n = 1, c^l_k = 1\right\} \approx \frac{1}{N \mid \mathbf{K}_n \mid}, \forall n \in \mathbf{N}, \forall k \in \mathbf{K}_n, \forall l \in \mathbf{L}; \quad (35a)$$

$$\Pr\left\{c^m_k = 1\right\} \approx \frac{1}{K}, \forall n \in \mathbf{N}, \forall k \in \mathbf{K}, \forall m \in \mathbf{M}. \quad (35b)$$

Next, we assume that the link gains in all channels are determined from the statistical information (about their past values). That is,

$$G^k_{i,l}(t) \approx \frac{\sum_{\tau=1}^{t-1} G^k_{i,l}(\tau)}{t-1} = G^k_{i,l}, \forall i \in \mathbf{N} \cup \mathbf{M} \cup \{l\}, \forall l \in \mathbf{L}, \forall k \in \mathbf{K}. \quad (35c)$$

Based on the assumptions specified in (35a) – (35c), the maximal allowed power level in the second scenario can be set according to

$$P_{\max} \le \min_{n \in \mathbf{N},\, k \in \mathbf{K}_n,\, l \in \mathbf{L}} \frac{\frac{G^k_{n,l}}{SINR^l_{\min} \mid \mathbf{K}_n \mid} - \sum_{i \in \mathbf{N} \setminus \{n\}} \frac{(L-1) G^k_{i,l}}{\mid \mathbf{K}_i \mid} - \frac{N\sigma^2}{\mathbf{P}_{\max}}}{\frac{1}{K\, \mathbf{P}_{\max}} \sum_{m \in \mathbf{M}} G^k_{m,l}}. \quad (36)$$

We now discuss the implementation of the proposed RL-based algorithm for D2D users in the UL direction of a considered heterogeneous cellular system. First, note that the analytical framework developed in the first scenario is applicable to both the UL and DL directions of cellular network (since there is no interference from cellular to D2D communications and vice versa). However, to apply the results of this work to the UL cellular channels in the second scenario, the following modifications should be made to the formulation of the considered resource allocation problem. Particularly, the interference from cellular communications to each D2D pair PU_m transmitting over the channel C_k (denoted $I^{\mathbf{C}}_{m,k}(t)$) in the SINR expression for D2D users given by (4a) will take the form

$$I^{\mathbf{C}}_{m,k}(t) = \sum_{i \in \mathbf{N}} \sum_{j \in \mathbf{L}} G^k_{j,m'}(t) b^j_i(t) c^j_k(t) P^j(t), \quad (37)$$

at any slot t, where $P^j(t)$ is the instantaneous transmission power of a cellular user U_j. The link gain coefficient matrix $\mathbf{G}^m_t$ in (11) should be redefined as

$$\mathbf{G}^m_t = \begin{bmatrix} G^1_{N+1,m'}(t) & \dots & G^K_{N+1,m'}(t) \\ \vdots & \vdots & \vdots \\ G^1_{N+M+L,m'}(t) & \dots & G^K_{N+M+L,m'}(t) \end{bmatrix}, \forall m \in \mathbf{M}. \quad (38)$$

Next, the following changes have to be made to the SINR formulation for cellular users, given by (32a). The instantaneous interference from the D2D pairs to each cellular user U_l transmitting over the channel C_k in the UL directions, will take the form

$$I^{\mathbf{D}}_{l,k}(t) = P_{\max} \sum_{m \in \mathbf{M}} G^k_{n,m}(t) c^m_k(t), \quad (39a)$$

and the interference to U_l associated with the BS_n from other cellular users transmitting in the UL direction over the channel C_k (denoted $I^{\mathbf{C}}_{n,l,k}(t)$) will be described by

$$I^{\mathbf{C}}_{n,l,k}(t) = \mathbf{P}_{\max} \sum_{i \in \mathbf{N} \setminus \{n\}} \sum_{m \in \mathbf{L} \setminus \{l\}} G^k_{n,m}(t) b^m_i(t) c^m_k(t). \quad (39b)$$

at any slot t. Consequently, the maximal allowed power level of the D2D users in the second scenario will be given by

$$P_{\max} \le \min_{n \in \mathbf{N},\, k \in \mathbf{K}_n,\, l \in \mathbf{L}} \frac{\frac{G^k_{n,l}}{SINR^l_{\min} \mid \mathbf{K}_n \mid} - \sum_{i \in \mathbf{N} \setminus \{n\}} \sum_{m \in \mathbf{L} \setminus \{l\}} \frac{G^k_{n,l}}{\mid \mathbf{K}_i \mid} - \frac{N\sigma^2}{\mathbf{P}_{\max}}}{\frac{1}{K\, \mathbf{P}_{\max}} \sum_{m \in \mathbf{M}} G^k_{n,m}}. \quad (40)$$

V. Performance Evaluation

In this paper, we show the deployment of the proposed autonomous RL-based algorithm for joint channel and power level selection by D2D users in a heterogeneous 3rd Generation Partnership Project (3GPP) LTE-A network. The simulation model of a network has been implemented using the DL spectrum resources of a standard Time Division Duplex (TDD) LTE-A system using the OPNET development package [36]. The model consists of N = 3 BSs (LTE evolved NodeBs or eNBs): BS_1 (pico-eNB), BS_2 (macro-eNB) and BS_3 (micro-eNB), with the antenna pattern specified in [37], placed as shown in Figure 3. In the first scenario, the total network bandwidth is composed of the channels in the union set $\mathbf{K}^{\mathbf{C}} \cup \mathbf{K}$ = {1, …, 50} comprising 50 LTE resource blocks (RBs). Among these RBs, K = 25 RBs from the set $\mathbf{K}$ = {1, …, 10, 26, …, 40} are reserved for D2D communication and the other 25 RBs in the set $\mathbf{K}^{\mathbf{C}}$ = {11, …, 25, 41, …, 50} are available for cellular users. In the second scenario, the total spectrum band of the network consists of 50 RBs belonging to the set $\mathbf{K}$ = {1, …, 50}, all of which can be used by both the cellular and D2D communications. The spectrum bands of the

BSs in the first and second scenarios are specified in Tables I and II, respectively. Here each k^{th} LTE resource block (RB_k) corresponds to a cellular channel C_k, $\mathbf{K}_n$ denotes the set of the channels within the licensed band of BS_n available for D2D communications, $\mathbf{K}^{\mathbf{C}}{}_n$ denotes the set of the channels belonging to BS_n reserved for cellular users in the first scenario, the channel bandwidth is ω = 180 kHz.

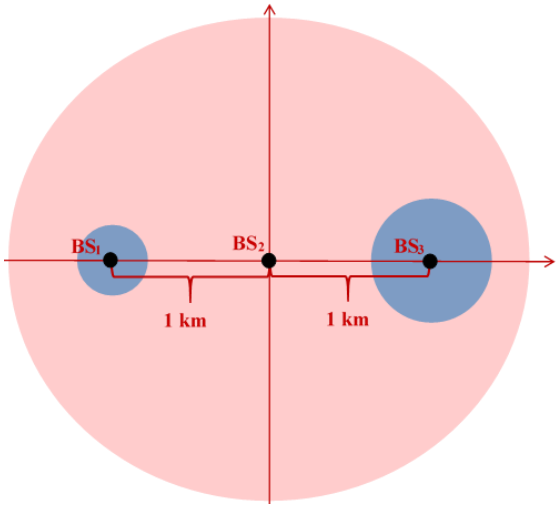

Figure 3. The placement of the BSs in simulation model. Pink circles show the service areas of BS_2, blue circles show the service areas of BS_1 and BS_3.

TABLE I. THE SPECTRUM BANDS OF THE BSS IN THE FIRST SCENARIO

BS_n	Bandwidth, $\lvert\mathbf{K}^{\mathbf{C}}{}_n \cup \mathbf{K}_n\rvert$	D2D channels C_k, $k \in \mathbf{K}_n$	Cellular channels C_k, $k \in \mathbf{K}^{\mathbf{C}}{}_n$
BS_1	25 RBs (5 MHz)	$\mathbf{K}_1 = \{1, \ldots, 10\}$	$\mathbf{K}^{\mathbf{C}}{}_1 = \{11, \ldots, 25\}$
BS_2	50 RBs (10 MHz)	$\mathbf{K}_2 = \{1, \ldots, 10, 26, \ldots, 40\}$	$\mathbf{K}^{\mathbf{C}}{}_2 = \{11, \ldots, 25, 41, \ldots, 50\}$
BS_3	25 RBs (5 MHz)	$\mathbf{K}_3 = \{26, \ldots, 40\}$	$\mathbf{K}^{\mathbf{C}}{}_3 = \{41, \ldots, 50\}$

TABLE II. THE SPECTRUM BANDS OF THE BSS IN THE SECOND SCENARIO

BS_n	Bandwidth, $\lvert\mathbf{K}_n\rvert$	Shared cellular/D2D channels C_k, $k \in \mathbf{K}_n$
BS_1	25 RBs (5 MHz)	$\mathbf{K}_1 = \{1, \ldots, 25\}$
BS_2	50 RBs (10 MHz)	$\mathbf{K}_2 = \{1, \ldots, 50\}$
BS_3	25 RBs (5 MHz)	$\mathbf{K}_3 = \{26, \ldots, 50\}$

The BSs serve the sets of the D2D pairs and cellular users randomly positioned inside the network service area. It is assumed that the cellular users are allocated the RBs by the BSs according to a standard LTE packet scheduling algorithm [38] based on the precise CSI. In all simulation experiments, the number of cellular users is fixed and equal L = 100. All user devices 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, the user traffic is modeled as a full buffer with load 10 packets per second and packet size 1500 bytes. The minimum SINR levels for the cellular/D2D users are set as $SINR^m{}_{\min} = SINR_{\min}$ = 0 dB, for all $m \in \mathbf{M} \cup \mathbf{L}$. In all the experiments, the number of power levels is equal J = 10, whereas the maximal possible power level (and, consequently, the maximal allowed power level P_{max} in the first scenario) is equal $\mathbf{P}_{\max}$ = 23 dBm. In the second scenario, the maximal allowed power level is calculated according to (36) (when applied to our settings, (36) leads to expression $P_{\max} = 20 \text{ dBm} - 10\log_{10}(M)$). The main simulation parameters are listed in Table III; other parameters are set in accordance with 3GPP specifications [39].

TABLE III. SIMULATION PARAMETERS OF LTE-A MODEL

Parameter	Value
Frame Structure	Type 2 (TDD)
Slot duration, T_s	1 ms
TDD configuration	0
eNodeB Tx power	46 dBm
Max Tx power of the users	23 dBm
Noise power	-174 dBm/Hz
Path loss, cellular link	128.1 + 37.6 log(d), d[km]
NLOS path loss, D2D link	40 log(d) + 30 log(f) + 49, d[km], f[Hz]
LOS path loss, 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 proposed autonomous channel and power level selection method for D2D communication is abbreviated as ACS. In the following, we evaluate the effectiveness of ACS by comparing its performance with the performance of the following learning algorithms: i) Multi-agent autonomous ε-greedy Q-learning (ε-GQL) [34] where the action-values are updated according to (29) and the strategies are selected using the ε-greedy method (in the examples below, we set the value of ε in accordance with most common suggestions (see [34]) as ε = 0.1); ii) Uniform random selection (URS) where the actions are selected uniformly at random at any slot t; iii) Parallel fictitious play (FP) [40], [41] where, at any t, each player PU_m updates his belief about the other players' strategies based on the observed actions $\mathbf{a}^{-m}{}_{t-1}$ at a previous slot $t - 1$, using such beliefs, each player selects an action that maximizes its expected long-term reward in (9); iv) Parallel best response dynamics (BRD) [40], [41] where, at any t, each PU_m selects the optimal strategy to maximize its expected long-term reward based on the precise knowledge of the channel state $\mathbf{G}^m{}_t$ and past actions $\mathbf{a}^{-m}{}_{t-1}$; v) Optimal centralized strategy (OCS) where the channels are allocated by the BSs to both the cellular and D2D users based on the global CSI knowledge to maximize their total long-term rewards. Although the scenarios considered in FP, BRD, and OCS are not realizable in practical wireless networks (since they require a massive amount of control signaling in the network), we use them only to benchmark the performance of the proposed ACS algorithm.

Figures 4 – 6 demonstrate simulation results in the first scenario with $c_\alpha = c_\beta = 0.5$, $\zeta_\alpha = 0.7$, $\zeta^m{}_\beta = 0.8$, T_B = 10, 100, 1000 (in ACS), γ = 0.5 (in ACS and ε-GQL), and $\upsilon^m = \upsilon$ = 10 (in all simulated schemes), for all $m \in \mathbf{M}$. Figure 4 shows the average instantaneous reward of the D2D pairs, calculated as

$$u_t = \frac{1}{M} \sum_{m \in \mathbf{M}} u_t^m,$$

depending on the number of iterations with fixed M = 100 D2D pairs. Figure 5 illustrates the difference between the maximal possible instantaneous reward $u_t^{\max}$ (obtained in OCS) and the average instantaneous reward u_t received using the other methods as a function of the number of D2D pairs M after t = 50 algorithm iterations. Figure 6 shows the average convergence time in different algorithms (URS is not shown in

this figure since it does not converge to a NE state) depending on the number of D2D pairs M.

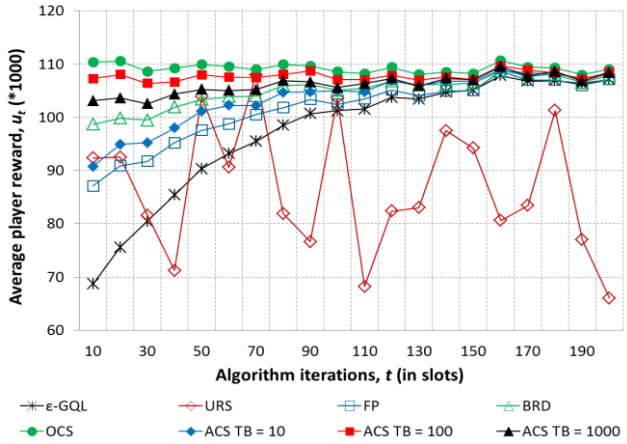

Figure 4. The average instantaneous reward u_t depending on the number of algorithm iterations t at a fixed M = 100.

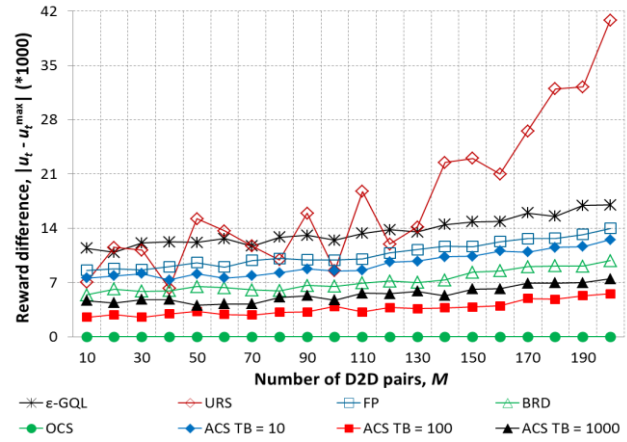

Figure 5. Reward difference $|u_t - u_t^{\max}|$ as a function of the number of D2D pairs M after t = 50 algorithm iterations.

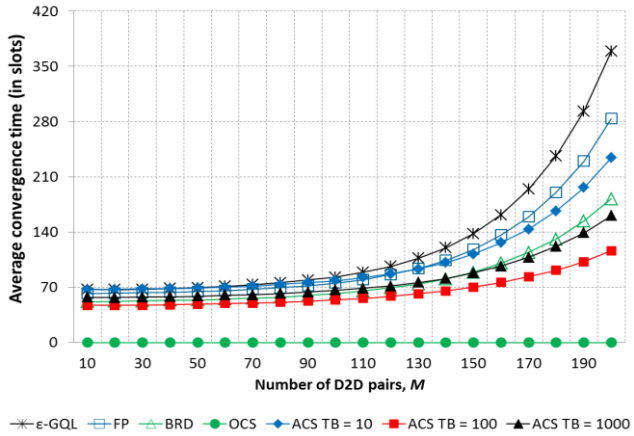

Figure 6. The average number of iterations necessary for the algorithm convergence (to a NE state) as a function of the number of D2D pairs M.

It follows from these figures that the performance of the proposed channel selection method is heavily influenced by the settings of T_B:

- For low temperatures (T_B = 10), the performance of ACS is getting closer to the performance of FP and ε-GQL. Note that the average instantaneous reward in ACS is upper bounded by the maximal reward $u_t^{\max}$ corresponding to the difference between the maximal possible throughput and the minimal transmission cost, i.e. $u_t^{\max} = R_t^{\max} - \upsilon P_{\max}/J \approx 120 - 2 = 118$ and, hence, the values of $e^{\hat{u}_t^m(s^m,a^m)/T_B}$ (in equation (30) vary between 133252.35 and 1 (for $u_t = 0$), for all $m \in \mathbf{M}$. This means that there is a very large difference in strategy selection probabilities leading to the predominance of the exploitation over exploration and resulting (as it follows from the graphs in Figures 4 – 6) in a relatively slow convergence speed of a corresponding algorithm.
- As temperature increases (T_B = 100), the performance of ACS improves, showing the near-optimal results after a relatively small number of iterations. Note that the values of $e^{\hat{u}_t^m(s^m,a^m)/T_B}$ fluctuate between 3.25 and 1 which indicates that there is a moderate difference in strategy selection probabilities resulting in the good balance between exploitation and exploration.
- At higher temperatures (T_B = 1000), the values of $e^{\hat{u}_t^m(s^m,a^m)/T_B}$ fluctuate between 1.125 and 1 which implies that there is almost no difference in strategy selection probabilities leading to the predominance of exploration over exploitation.

Based on the above, we come up with the following recommendations for setting the temperature in BG distribution given by (30): the values of T_B should be comparable to the maximal possible instantaneous reward $u_t^{\max}$ (that is, $1 \le u_t^{\max}/T_B \le 10$). The values of $u_t^{\max}$ can be (roughly) estimated is we know the average load Ω generated by each user and the average cost of power consumption υ per user, as $u_t^{\max} = \Omega - \upsilon P_{\max}/J$, Ω[bps], $P_{\max}$[W]. In our case, the best performance was achieved with T_B = 100, $u_t^{\max}$ = 118 and $u_t^{\max}/T_B$ = 1.18. Note that according to the Theorem (stated in Appendix) the Q-learning process in (29) converges given that the two conditions, (42b) and (42c), are satisfied. From Proposition 2, the first condition, given by (42b), is satisfied only if T_B is sufficiently large (greater than the difference between the achieved action values). On the other hand, the difference $|Q_t(s^m, a^m) - Q'_t(s^m, a^m)|$ is decreasing over time, for all $m \in \mathbf{M}$. Consequently,

$$\lim_{t\to\infty} \varphi\left(\left[\frac{Q_t(s^m,\mathbf{a}^{-m}) - Q'_t(s^m,\mathbf{a}^{-m})}{T_B} \right]^d \right) = 0, \forall m \in \mathbf{M}$$

which implies that the inequality in (53b) may yet hold for small values of T_B. Hence, even if the values of T_B are smaller than the difference between the corresponding action values at the beginning of simulations (e.g., with T_B = 10), the algorithm in (29) still converges to a NE state (since the values of $|Q_t(s^m, a^m) - Q'_t(s^m, a^m)|$ are becoming much smaller than the values of T_B after a certain time period).

Figures 7 and 8 show the average instantaneous throughput R_t and the average instantaneous transmit power P_t of the D2D pairs, calculated according to

$$R_t = \frac{1}{M}\sum_{m\in\mathbf{M}} R_t^m,\ P_t = \frac{1}{M}\sum_{m\in\mathbf{M}} P_t^m,$$

collected in the first and second scenarios after t = 50 algorithm iterations with M = 100, 200, 300, 400, $c_\alpha = c_\beta = 0.5$, $\zeta_\alpha = 0.7$, $\zeta^m{}_\beta = 0.8$, $T_B = 100$, $\gamma = 0.5$, and $\upsilon^m = \upsilon$ varying from 10 to 200, for all $m \in \mathbf{M}$. It follows from these figures that both R_t and P_t decreases with υ, which is rather predictable because if the cost of power consumption for D2D users is high then they have to reduce their power levels (and, consequently, the achieved throughput). We also note that in the first scenario, R_t and P_t are higher than that in the second scenario, for υ < 160 and lower, for υ > 160. This can be explained by the fact that in this scenario, the D2D users are allowed to transmit at high power levels while in the second scenario their transmission power is limited (to protect the cellular users from heavy interference). This results in increased R_t and P_t for lower values of υ. However, for υ > 160, the high cost per unit power does not allow the D2D users to transmit at high power levels causing degraded levels of R_t and P_t (since the amount of spectrum resources accessible for the D2D pairs in the first scenario is less than that in the second scenario).

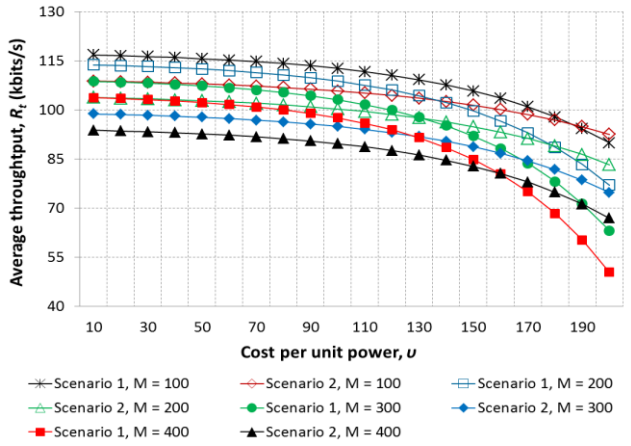

Figure 7. The average instantaneous throughput R_t in different scenarios as a function of the cost per unit power $\upsilon^m = \upsilon$ with fixed values of M after t = 50 algorithm iterations.

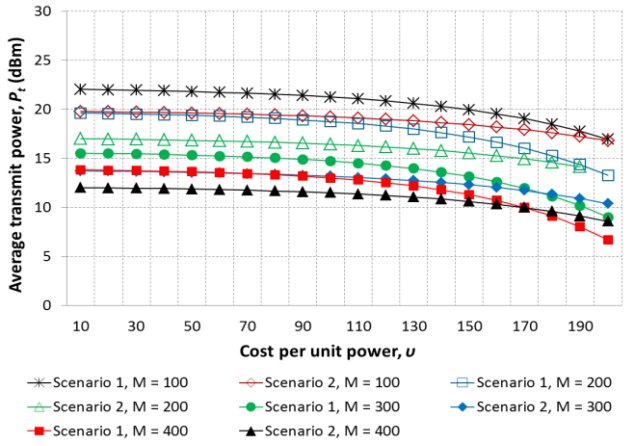

Figure 8. The average instantaneous transmit power P_t in different scenarios as a function of the cost per unit power $\upsilon^m = \upsilon$ with fixed values of M after t = 50 algorithm iterations.

Figures 9 and 10 demonstrate, respectively, the total average instantaneous throughput of cellular users $R^C{}_t$ and the total average instantaneous throughput of D2D users $R^D{}_t$ in the first and second scenarios depending on the number of D2D pairs M plotted after t = 50 algorithm iterations. Note that the average instantaneous throughput of cellular users in the second scenario decreases with the number of D2D users. This is because with the growing number of D2D pairs, the average number of D2D users occupying the shared cellular/D2D channels increases, resulting in higher interference to cellular communication and the decreasing average cellular throughput $R^C{}_t$.

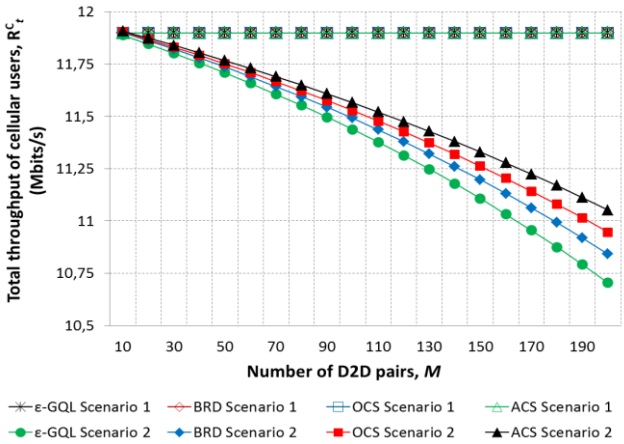

Figure 9. The total average throughput of cellular users $R^C{}_t$ in different scenarios as a function of the number of D2D users M after t = 50 algorithm iterations.

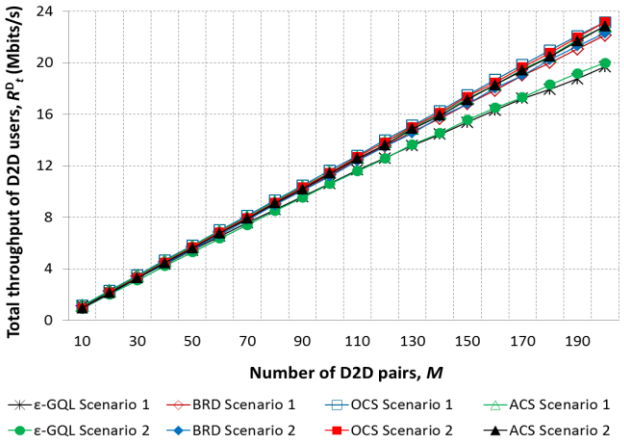

Figure 10. The total average throughput of D2D users $R^D{}_t$ in different scenarios as a function of the number of D2D users M after t = 50 algorithm iterations.

Figures 11 and 12 illustrate the network performance in the second scenario with varying minimum SINR levels. Here $R^C{}_t$ and $R^D{}_t$ are presented as the functions of $SINR_{\min}$ plotted after t = 50 algorithm iterations for a fixed number of D2D pairs, M = 100. It follows from these figures that the total throughputs of cellular and D2D users are described by some concave functions of $SINR_{\min}$, with the maximal level depending on the applied channel selection method (the highest throughput level is achieved with OCS and the lowest – with URS). Such results are rather predictable – with very low settings of

$SINR_{\min}$ ($SINR_{\min}$ < 0 dB), the throughput of users decreases because of the bad channel conditions. When $SINR_{\min}$ is too high ($SINR_{\min}$ > 8 dB), the throughput degrades due to the shortage of available bandwidth - the number of channels with suitable channel conditions decreases since not all of them satisfy the requirements of the users.

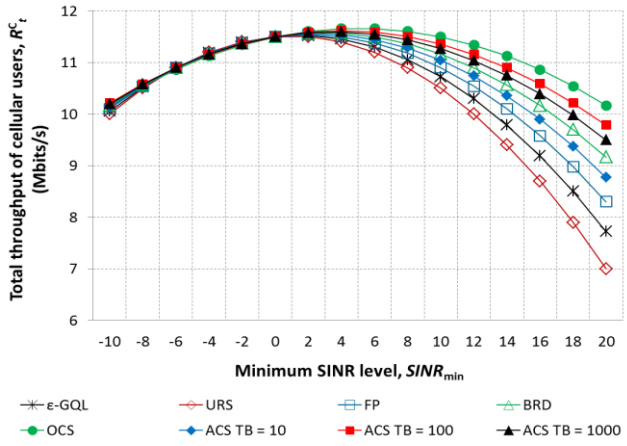

Figure 11. The total average throughput of cellular users R^C_t in the second scenario as a function of the minimum tolerable SINR level $SINR_{\min}$ after t = 50 algorithm iterations at a fixed M = 100.

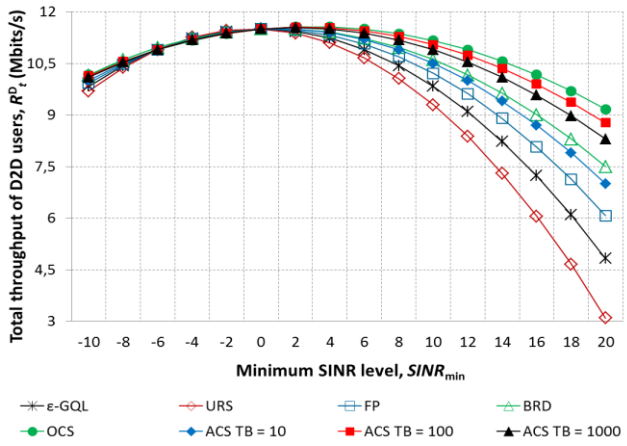

Figure 12. The total average throughput of D2D users R^D_t in the second scenario as a function of the minimum tolerable SINR level $SINR_{\min}$ after t = 50 algorithm iterations at a fixed M = 100.

VI. Conclusion

In this paper, we propose a Q-learning based method for autonomous channel and power level selection by D2D users in a multi-cell network (with the possibility of ICI). In the presented framework, the problem of distributed resource allocation for D2D pairs is formulated as a stochastic non-cooperative game with multiple selfish players whose task is to learn their best strategies based on the locally observed actions, states, and rewards. Based on this framework, we develop a fully autonomous multi-agent Q-learning algorithm (that does not require any coordination or information exchange among the users and between the D2D pairs and the BSs) and implement it in a heterogeneous LTE-A network. It is shown (via simulations) that the proposed learning method shows relatively fast convergence to a NE state and the near-optimal results given the appropriate settings of the temperature parameter T_B.

Appendix

We now prove the convergence of the proposed autonomous multi-agent Q-learning method given by (29). Let $\mathbf{Q}$ be the space of all Q-values. The Lemma below uses the sufficient conditions for convergence of a general (single-agent) Q-learning method given by

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha_t\,[u^m(a^m,\mathbf{a}^{-m},s^m) + \gamma \max_{a'\in\mathbf{A}} Q_t(s',a') - Q_t(s,a)], \tag{41}$$

which have been established in [42].

Lemma: Assume that the learning rate α_t in (41) satisfies the sufficient conditions of the convergence theorem in [42]. That is,

$$\sum_{t\ge 0}\alpha_t = +\infty, \sum_{t\ge 0}\alpha_t^2 < +\infty \tag{42a}$$

and the mapping H_t: $\mathbf{Q} \to \mathbf{Q}$ meets the following condition: there exists a number $\delta \in (0, 1)$ and a sequence $\xi_t \ge 0$ converging to zero with probability 1, such that

$$\left\|H_tQ_t - H_t\overline{Q}\right\| \le \delta\left\|Q_t - \overline{Q}\right\| + \xi_t, \forall Q_t \in \mathbf{Q} \tag{42b}$$

where $\overline{Q}$ is the optimal action-value and

$$\overline{Q} = \mathrm{E}\left\{H_t\overline{Q}\right\}. \tag{42c}$$

Then the iteration defined by

$$Q_{t+1} \leftarrow Q_t + \alpha_t(H_tQ_t - Q_t) \tag{43}$$

converges to the optimal value $\overline{Q}$ with probability 1.

The proof of the above Lemma (signifying the convergence of a general Q-learning process) has been presented in [43]. In the following, we use the Lemma to establish the convergence of the proposed autonomous Q-learning method in (29).

Proposition 1: Let us define

$$H_tQ_t(s^m,a^m) = \sum_{a^{-m}\in\mathbf{A}^{-m}}\left\{u^m[a^m,\mathbf{a}^{-m},s^m]\hat{Y}(s^m,\mathbf{a}^{-m})\right\} + \gamma \max_{a'^m\in\mathbf{A}^m} Q_t(s'^m,a'^m). \tag{44a}$$

Then, for the M-player stochastic game,

$$\overline{Q}(s^m,a^m) = \mathrm{E}\left\{H_t\overline{Q}(s^m,a^m)\right\}, \forall m\in\mathbf{M}, \forall a^m\in\mathbf{A}^m. \tag{44b}$$

Proof: Note that from (19) we have

$$\overline{Q}(s^m,a^m) = \mathrm{E}\left\{u^m[a^m,\overline{\boldsymbol{\pi}}^{-m}(s^m),s^m]\right\} + \gamma\sum_{s'^m\in\mathbf{S}^m}T_{s^ms'^m}[a^m,\overline{\boldsymbol{\pi}}^{-m}(s^m)]\max_{a'^m\in\mathbf{A}^m}Q(s'^m,a'^m) = \sum_{s'^m\in\mathbf{S}^m}T_{s^ms'^m}[a^m,\overline{\boldsymbol{\pi}}^{-m}(s^m)]\left[\sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}}\left\{u^m[a^m,\mathbf{a}^{-m},s^m]\hat{Y}(s^m,\mathbf{a}^{-m})\right\} + \gamma\max_{a'^m\in\mathbf{A}^m}Q(s'^m,a'^m)\right], \forall m\in\mathbf{M}. \tag{45a}$$

Then, using definition (40a), we arrive at

$$\overline{Q}(s^m,a^m) = \sum_{s'^m\in\mathbf{S}^m}T_{s^ms'^m}[a^m,\overline{\boldsymbol{\pi}}^{-m}(s^m)]\,H_tQ_t(s^m,a^m) = \mathrm{E}\left\{H_t\overline{Q}(s^m,a^m)\right\}, \tag{45b}$$

for all $s^m\in S^m$, $a^m\in\mathbf{A}^m$, $m\in\mathbf{M}$, which concludes the proof. □

Proposition 2: There is a number $\delta\in(0, 1)$, such that

$$\left\|H_tQ - H_tQ'\right\| \le \delta\left\|Q - Q'\right\|. \tag{46}$$

Proof: Let us define the distance between any two Q-values Q and Q' belonging to space $\mathbf{Q}$, as

$$\| Q - Q' \| = \max_{m\in\mathbf{M}} \max_{s^m\in\mathbf{S}^m} \max_{a^m\in\mathbf{A}^m} \left| Q(s^m,a^m) - Q'(s^m,a^m) \right|. \tag{47}$$

According to the above definition, we have

$$\| H_t \cdot Q - H_t \cdot Q' \| =$$

$$\max_{m\in\mathbf{M}} \max_{s^m\in\mathbf{S}^m} \max_{a^m\in\mathbf{A}^m} \left| H_t Q(s^m,a^m) - H_t Q'(s^m,a^m) \right| =$$

$$\max_{m\in\mathbf{M}} \max_{s^m\in\mathbf{S}^m} \max_{a^m\in\mathbf{A}^m} \left| \sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}} u^m[a^m,\mathbf{a}^{-m},s^m]\{ \hat{Y}(s^m,\mathbf{a}^{-m}) - \hat{Y}'(s^m,\mathbf{a}^{-m}) \} \right|$$
$$+ \gamma \max_{m\in\mathbf{M}} \max_{s^m\in\mathbf{S}^m} \max_{a'^m\in\mathbf{A}^m} \left| Q(s'^m,a'^m) - Q'(s'^m,a'^m) \right| \tag{48a}$$

and, therefore,

$$\| H_t Q - H_t Q' \| \le$$

$$\max_{m\in\mathbf{M}} \max_{s^m\in\mathbf{S}^m} \max_{a^m\in\mathbf{A}^m} \left| \sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}} u^m[a^m,\mathbf{a}^{-m},s^m]\{ \hat{Y}(s^m,\mathbf{a}^{-m}) - \hat{Y}'(s^m,\mathbf{a}^{-m}) \} \right|$$
$$+ \gamma \| Q - Q' \|. \tag{48b}$$

Note that because of (26a),

$$\sum_{a^{-m}\in\mathbf{A}^{-m}} u^m[a^m,\mathbf{a}^{-m},s^m]\{ \hat{Y}(s^m,\mathbf{a}^{-m}) - \hat{Y}'(s^m,\mathbf{a}^{-m}) \} =$$
$$-\beta_t^m \sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}} u^m[a^m,\mathbf{a}^{-m},s^m]\{ \pi^m(s^m,\mathbf{a}^{-m}) - \pi'^m(s^m,\mathbf{a}^{-m}) \}. \tag{49}$$

Now let us consider (30) and apply the well-known Taylor series expansion of exponential function [44] to the nominator of (30). Specifically, we get

$$e^{Q(s^m,a^m)/T_B} = \sum_{i=0}^{+\infty} \frac{(Q(s^m,a^m)/T_B)^i}{i!}$$
$$= 1 + \frac{Q(s^m,a^m)}{T_B} + \Phi\left(\frac{Q(s^m,a^m)}{T_B} \right), \forall m \in \mathbf{M}. \tag{50}$$

Using (50), we arrive at

$$\sum_{a'^m\in\mathbf{A}^m} e^{Q(s^m,a'^m)/T_B} = \left| \mathbf{A}^m \right|$$
$$+ \sum_{a'^m\in\mathbf{A}^m} \left[\frac{Q(s^m,a'^m)}{T_B} + \Phi\left(\frac{Q(s^m,a'^m)}{T_B} \right) \right], \forall m \in \mathbf{M} \tag{51}$$

where $| \mathbf{A}^m | = KJ$. If T_B is sufficiently large then it is rather straightforward to verify that

$$\pi^m(s^m,a^m) = \frac{e^{Q(s^m,a^m)/T_B}}{\sum_{a'^m\in\mathbf{A}^m} e^{Q(s^m,a'^m)/T_B}}$$
$$= \frac{1}{KJ}\left[1 + \frac{Q(s^m,a^m)}{T_B} + \varphi\left(\left[\frac{Q(s^m,a^m)}{T_B} \right]^d \right) \right], \forall m \in \mathbf{M} \tag{52a}$$

$$\pi'^m(s^m,a^m) = \frac{e^{Q'(s^m,a^m)/T_B}}{\sum_{a'^m\in\mathbf{A}^m} e^{Q'(s^m,a'^m)/T_B}}$$
$$= \frac{1}{KJ}\left[1 + \frac{Q'(s^m,a^m)}{T_B} + \varphi\left(\left[\frac{Q'(s^m,a^m)}{T_B} \right]^d \right) \right], \forall m \in \mathbf{M} \tag{52b}$$

where $\varphi(x^d)$ is a polynomial of some order $O(x^d)$.

Substituting (52a) and (52b) to (49), we get

$$\sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}} u^m[a^m,\mathbf{a}^{-m},s^m]\{ \hat{Y}(s^m,\mathbf{a}^{-m}) - \hat{Y}'(s^m,\mathbf{a}^{-m}) \} =$$
$$-\beta_t^m \sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}} \frac{u^m[a^m,\mathbf{a}^{-m},s^m]}{KJ} \left\{ \frac{Q(s^m,\mathbf{a}^{-m}) - Q'(s^m,\mathbf{a}^{-m})}{T_B} \right.$$
$$\left. + \varphi\left(\left[\frac{Q(s^m,\mathbf{a}^{-m}) - Q'(s^m,\mathbf{a}^{-m})}{T_B} \right]^d \right) \right\}, \forall m \in \mathbf{M} \tag{53a}$$

which implies that we can always take sufficiently large T_B (greater than the difference $| Q(s^m, a^m) - Q'(s^m, a^m) |$ between the corresponding action-values), such that

$$\left| \sum_{\mathbf{a}^{-m}\in\mathbf{A}^{-m}} u^m[a^m,\mathbf{a}^{-m},s^m]\{ \hat{Y}(s^m,\mathbf{a}^{-m}) - \hat{Y}'(s^m,\mathbf{a}^{-m}) \} \right|$$
$$\le \frac{1-\gamma}{KJ} \left| Q(s^m,\mathbf{a}^{-m}) - Q'(s^m,\mathbf{a}^{-m}) \right|, \forall m \in \mathbf{M} \tag{53b}$$

and, therefore,

$$\| H_t Q - H_t Q' \|$$
$$\le \max_{m\in\mathbf{M}} \max_{s^m\in\mathbf{S}^m} \max_{a^m\in\mathbf{A}^m} \frac{1-\gamma}{KJ} \left| Q(s^m,\mathbf{a}^{-m}) - Q'(s^m,\mathbf{a}^{-m}) \right| + \gamma \| Q - Q' \|$$
$$\le \frac{1-\gamma}{KJ} \| Q - Q' \| + \gamma \| Q - Q' \| = \frac{1+\gamma(KJ-1)}{KJ} \| Q - Q' \|$$
$$= \delta \| Q - Q' \|. \tag{53c}$$

Now, since

$$\delta = \frac{1+\gamma(KJ-1)}{KJ} \le 1, \tag{54}$$

with the equality if and only if $KJ = 1$, the inequality (54) is satisfied, which concludes the proof. □

Theorem: Regardless the initial settings of $Q_0(s^m, a^m)$, if T_B is sufficiently large and the learning rate α_t satisfies the condition (42a) of Lemma, the autonomous Q-learning algorithm given by (29) converges to the optimal value with probability 1. That is,

$$Q_t(s^m,a^m) \to \overline{Q}(s^m,a^m), \text{as } t \to \infty, \forall m \in \mathbf{M}. \tag{55}$$

Proof: Note that according to Lemma, the Q-learning process converges given that the conditions (42b) and (42c) are satisfied. In our case, the mapping operator condition (42b) is satisfied by Proposition 2. The fixed point condition (42c) is ensured by Proposition 1. Hence, the learning algorithm in (29) converges to the optimal value with probability 1, which concludes the proof. □

REFERENCES

[1] A. Asadi et. al., "A Survey on Device-to-Device Communication in Cellular Networks," IEEE Commun. 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 Trans. Commun.*, 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," to appear in *IEEE Trans. Veh. Tech.*, 2016.

[4] D. D. Penda, L. Fu and M. Johansson, "Energy efficient D2D communications in dynamic TDD systems," arXiv preprint arXiv:1506.00412, 2015.

[5] K. Yang et. al., "Energy-Efficient Resource Allocation for Device-to-Device Communications Overlaying LTE Networks," to be published in Proc. IEEE VTC, 2015.

[6] A. Asadi, P. Jacko, and V. Mancuso, "Modeling Multi-mode D2D Communications in LTE," arXiv preprint arXiv:1405.6689, 2014.

[7] B. Guo et. al., "Graph-Based Resource Allocation for D2D Communications Underlying Cellular Networks in Multiuser Scenario," Int. Journal of Antennas and Propagation, pp. 1 – 6, 2014.

[8] F. Malandrino et. al., "Uplink and Downlink Resource Allocation in D2D-Enabled Heterogeneous Networks," in Proc. IEEE WCNC, 2014, pp. 87 - 92.

[9] S. Maghsudi and S. Stanczak, "Channel Selection for Network-Assisted D2D Communication via No-Regret Bandit Learning With Calibrated Forecasting," IEEE Trans. Wireless Commun., vol. 14, no. 3, pp. 1309 – 1322, 2015.

[10] X. Chen et. al., "Exploiting social ties for cooperative D2D communications: a mobile social networking case," IEEE/ACM Trans. Networking, vol. 23, no. 5, pp. 1471- 1484, 2015.

[11] M. Hasan and E. Hossain, "Distributed Resource Allocation for Relay-Aided Device-to-Device Communication Under Channel Uncertainties: A Stable Matching Approach," arXiv preprint arXiv:1508.01459, 2015.

[12] B.-Y. Huang et.al., "Resource Allocation in D2D Communication – A Game Theoretic Approach," In Proc. IEEE ICC, 2014, pp. 483 - 488.

[13] M. Hasan and E. Hossain, "Distributed Resource Allocation in D2D-Enabled Multi-tier Cellular Networks: An Auction Approach," arXiv preprint arXiv:1501.04199, 2015.

[14] L. Song et. al., "Game-theoretic Resource Allocation Methods for Device-to-Device Communication," IEEE Wireless Commun., vol. 21, no. 3, pp. 136 – 144, 2014.

[15] Y. Li et al., "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, 2014.

[16] Y. Xiao et. al., "A Bayesian Overlapping Coalition Formation Game for Device-to-Device Spectrum Sharing in Cellular Networks," IEEE Trans. Wireless Commun., vol. 14, no. 7, pp. 4034 – 4051, 2015.

[17] "Reinforcement Learning: An Introduction," by R. S. Sutton and A. G. Barto, MIT Press, Cambridge MA, 1998.

[18] Y. Xu et. al., "Opportunistic spectrum access with spatial reuse: Graphical game and uncoupled learning solutions," IEEE Trans. Wireless Commun., vol. 12, no. 10, pp. 4814 – 4826, 2013.

[19] Y. Xu et. al., "Opportunistic spectrum access using partially overlapping channels: Graphical game and uncoupled learning," IEEE Trans. Commun., vol. 61, no. 9, pp. 3906 – 3918, 2013.

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

[21] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," Journal of Machine Learning Research, vol. 4, pp. 1039-1069, 2003.

[22] A. Greenwald, and K. Hall, "Correlated-Q learning," in Proc. ICML, 2003.

[23] N. Junhong and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," IEEE Trans. on Veh. Tech., vol. 48, no. 5, pp. 1676 – 1687, 1999.

[24] B. Mehdi and D. Niyato, "A Q-learning based approach to interference avoidance in self-organized femtocell networks," In Proc. IEEE GLOBECOM, 2010, pp. 706 – 710.

[25] L. Husheng, "Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: A two by two case," In Proc. IEEE SMC, 2009, pp. 1893 – 1898.

[26] S. M. Perlaza, H. Tembine, and S. Lasaulce, "How can ignorant but patient cognitive terminals learn their strategy and utility," in Proc. SPAWC, 2010.

[27] M. Benaïm, "Dynamics of stochastic approximations," Lectures Notes in Mathematics, vol. 1709, 1999.

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

[29] "A Course in Game Theory," by M. J. Osborne and A. Rubinstein, MIT Press, Cambridge MA, 1994.

[30] "Stochastic Games and Applications," by A. Neyman and S. Sorin, Kluwer Academic Press, Dordrecht, 2003.

[31] J. C. Harsanyi, John C., "Games with Incomplete Information Played by Bayesian Players," Management Science, vol. 14, no. 3, pp. 159 – 183 (Part I), vol. 14, no. 5, pp. 320 – 334 (Part II), vol. 14, no. 7, pp. 486 – 502 (Part III), 1967/1968.

[32] G. Chalkiadakis and C. Boutilier, "Coordination in Multiagent Reinforcement Learning: A Bayesian Approach," in Proc. AAMAS, 2003.

[33] G. Tesauro, "Extending Q-Learning to General Adaptive Multi-Agent Systems," in Advances in Neural Information Processing Systems, no. 16, 2003.

[34] E. R. Gomes and R. Kowalczyk, "Dynamic analysis of multiagent Q-learning with ϵ-greedy exploration," in Proc. ICML, 2009.

[35] OPNET. Available online: http://www.opnet.com.

[36] Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN). 3GPP TS 36.300. (Release 8).

[37] E-UTRA; MAC protocol specification, 3GPP TS 36.321. (Release 8).

[38] A. Orsino, "Efficient Data Uploading Supported by D2D Communications in LTE-A Systems," arXiv preprint arXiv:1503.09076, 2015.

[39] Overview of 3GPP Release 12. V0.1.4, September 2014.

[40] "Mechanisms and Games for Dynamic Spectrum Allocation," by T. Alpcan et. al., MIT Press, Cambridge MA, 2014.

[41] L. Rose et. al., "Learning equilibria with partial information in decentralized wireless networks," IEEE Commun. Mag., vol. 49, no. 8, pp. 136 – 142, 2011.

[42] C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, no. 8, pp. 279-292, 1992.

[43] C. Szepesvari and M. L. Littman, "A unified analysis of value function-based reinforcement learning algorithms," Neural Computation, vol. 11, no. 8, pp. 2017-2060, 1999.

[44] "Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables," by M. Abramowitz and I. A. Stegun, New York: Dover Publications, Ninth printing, 1970.