Q-Learning based Joint Energy-Spectral Efficiency Optimization in Multi-Hop Device-to-Device Communication

Muhidul Islam Khan, Luca Reggiani, Muhammad Mahtab Alam, Yannick Le Moullec, Navuday Sharma, Elias Yaacoub, and Maurizio Magarini

***Abstract*—In scenarios like critical public safety communication networks, on-scene available (OSA) user equipment (UE) may be only partially connected with the network infrastructure, *e.g.*, due to physical damages or on-purpose deactivation by the authorities. In this work, we consider multi-hop Device-to-Device (D2D) communication in a hybrid infrastructure where OSA UEs connect to each other in a seamless manner in order to disseminate critical information to a deployed command center. The challenge that we address is to simultaneously keep the OSA UEs alive as long as possible and send the critical information to a final destination (e.g. a command center) as rapidly as possible, while considering the heterogeneous characteristics of the OSA UEs. We propose a dynamic adaptation approach based on machine learning to improve a joint energy-spectral efficiency (ESE). We apply a Q-learning scheme in a hybrid fashion (partially distributed and centralized) in learner agents (distributed OSA UEs) and scheduler agents (remote radio heads or RRHs), for which the next hop selection and RRH selection algorithms are proposed. Our simulation results show that the proposed dynamic adaptation approach outperforms the baseline system by approximately 67% in terms of joint energy-spectral efficiency, wherein the energy efficiency of the OSA UEs benefit from a gain of approximately 30%. Finally, the results show also that our proposed framework with C-RAN reduces latency by approximately 50% w.r.t. the baseline.**

***Index Terms*—Joint Energy-Spectral Efficiency (ESE), Device-to-Device (D2D), Public Safety Networks, Pervasive Public Safety Communication, Internet of Things (IoT).**

I. Introduction

In the context of the cellular networks, from Advanced Long Term Evolution (LTE-A) to fifth-generation of mobile communication (5G), the increasing number of devices keeps pushing the demand for higher spectral and energy efficiencies. In this direction, 5G provides a roadmap for increased resource efficiency and energy efficiency, greater reliability, and low latency solutions [1]. In particular, Device-to-Device (D2D) communication is regarded as a key technology in 5G wireless systems for providing services that include live data and video sharing [2]. D2D communication allows user devices (UEs) that are in close proximity to exchange information over a direct link, which can be operated as an underlay to LTE-A networks by reusing the spectrum resources. D2D communication is not only useful for local communication but also for extending the range of a base-station (BS) to out-of-coverage UEs. This opportunity can be provided by D2D-based relays. Relay UEs help to communicate with the BS and other out-of-coverage UEs, as standardized in 3GPP Release 13 [3]. Two key features of D2D proximity service (ProSe) are [4]:

- UEs in close physical proximity are able to discover the existence of each other through network assisted discovery;
- direct communications between two UEs, with or without inclusion of the control from the network, can be enabled using a direct interface called Sidelink (SL).

In this work, we exploit the above concepts in a scenario where UEs are partially connected with the network infrastructure, damaged or deactivated, and there is the necessity of conveying the connectivity to a given destination. This is a crucial case in public safety networks (PSN), which are mission critical wireless networks for emergency scenarios [5] that create a link between the persons in the area, including the rescue teams, and a national command center for sharing mission-critical information. However, traditional PSNs are not designed to cope with cases during which the UEs are partially connected with the network infrastructure, *e.g.*, due to on-purpose deactivation by authorities (e.g. in case of terrorist activity) or physical damage. Thus, D2D communication is seen as a solution for extending the coverage of the sites that remain active in such partial coverage scenarios.

Therefore, our proposed setup considers multi-hop D2D communication in the reference context of PSNs, *e.g.*, for disasters or terrorist attacks where the cellular BS (i.e. eNodeB or gNodeB) becomes non-functional or fully destroyed. In such situation, also unmanned aerial vehicles (UAVs) could be deployed in order to assist OSA UEs in disseminating the information such as user ID, location and images to the command center without going through the BSs.

The problem of connectivity is manifold as it includes ensuring the end-to-end network connectivity with the external command center, routing critical information to this center. In a traditional scenario, the BS would allocate the resources to the cellular and D2D users; dedicated resource allocation and proper power allocation would be applied at the D2D devices and interference and reliability would be controlled.

Muhidul Islam Khan, Muhammad Mahtab Alam, Yannick Le Moullec and Navuday Sharma are with Thomas Johann Seebeck Department of Electronics, School of Information Technologies, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia. Elias Yaacoub is with the Computer Science and Engineering Department, Qatar University, Doha, Qatar. Luca Reggiani and Maurizio Magarini are with Dipartimento di Electtronica e Informazione, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy.

However, in the underlay mode (as in our scenario), resource allocation and interference management pose a challenge due to the absence of BS control, meaning that these tasks have to be executed on the OSA UEs and UAVs (the latter act as RRHs). Moreover, multi-hop communication is essential for such a scenario to disseminate the critical information. Finding the best routing path is challenging due to harsh propagation conditions and severe links obstructions, which makes it difficult to rely only on the status of the up and down links; network parameters such as the RRH load, congestion level, link quality metric, average number of hops, and throughput should be considered as well. Taking into account all these aspects, there is a clear need for a novel approach that would allow controlling the resources, minimizing the interference, and optimizing the multi-hop routing path, while being light-weight enough to be implemented on computationally-limited equipment. Our approach is based on a joint optimization of energy and spectral efficiencies (ESE) under constraints imposed in a scenario in which multi-hop D2D connections can provide diffused connectivity without a centralized support.

Existing solutions for ESE, reliable connectivity and routing are ill suited due to severe link conditions and unique mobility patterns in the disaster scenario, i.e., harsh radio propagation conditions, variable transmission range and heterogeneous resources, e.g. [6]–[11] lack exploiting light-weight, time critical on-line mechanisms to learn the adaptive resource allocation and ESE; furthermore, such works use centralized approaches that do not exploit the OSA resources available in our scenario; [12] partly uses distributed Q-learning but focuses on system throughput.

Here we propose a joint ESE approach based on a light-weight machine learning approach as a novel and promising way to solve the connectivity problem. The originality of our work is to propose a hybrid (partially distributed and centralized) approach; such an approach exploits distributed agents i.e., OSA UEs at a local level complements the basic foundation of RL, while the deployed command center acts as a centralized radio access network (C-RAN) for coordination and optimization of RRH performance and exploit machine learning (ML) techniques. In our solution, reinforcement learning, specifically Q-Learning, is implemented as a lightweight and model-free algorithm and we remark that this is highly desirable for a disaster scenario where the UEs need to run algorithms with low complexities and where there is no specific model of the environment. The main contributions of this work are summarized as follows:

- Joint ESE optimization for D2D communication, including the impact of multiple hops: existing works such as [13], [14] and [15] have investigated the trade-off between energy and spectral efficiency in the context of PSNs, but without exploiting a cost function for the joint ESE with multiple hops. On the other hand, our work optimizes this ESE measure by applying Q-learning [16], whose reward function reflects the link quality, i.e. interference level, power consumption, and congestion. Moreover, we apply Q-learning in a hybrid fashion (partially distributed and centralized) where the learning algorithm is applied at two types of agents: learner agents, which are the distributed devices, i.e. OSA UEs, and scheduler agents, which are the RRHs.
- Most of the existing routing protocols in ad hoc networks rely just on the status of the links (up or down); on the other hand, our proposed ML-based approach finds out the best path by considering the network parameters, i.e. RRH load, congestion level, link quality metric, average number of hops and throughput. Optimizing the routing path contributes to make the agents behave so that both energy and spectral efficiency are enhanced simultaneously and dynamically. This aspect is crucial for the optimization process and it is reflected in the definition of the joint energy and spectral efficiency to be maximized, equation (14), which incorporates the number of hops in the formulation.
- As a result, our approach outperforms the baseline (non-optimized) algorithm by 67% in terms of joint ESE. This translates into more UEs being alive over the number of iterations (i.e. time) as compared to the baseline algorithm. Also, our adaptive algorithm helps to improve the energy efficiency of connected UEs by 30%.
- Context aware and reliable D2D multi-hop routing and network connections to ensure high end-to-end throughput and low end-to-end energy consumption and delay. Most of the existing routing protocols in ad hoc networks only rely on the status of the links (up or down); on the other hand, our proposed ML-based approach finds out the best path by considering the network parameters, *i.e.*, RRH load, congestion level, link quality metric, end-to-end delay, execution time, and throughput. Furthermore, in our work, the optimum path is achieved by ensuring that the quality of the links within the network is detected on a continuous basis instead of discrete times like in existing protocols. As a result, our proposed framework for a disaster scenario reduces latency by almost 50%.

The paper is structured as follows: Sect. II discusses the state-of-the-art solutions, highlighting the differences with our approach. Sect. III presents the system model with the parameters and notations; Sect. IV develops the problem formulation, the related constraints and the final optimization function selected for our purpose. Sect. V describes in detail the approach for performing, in practice, the optimization of the cost function; Sect. VI presents the numerical results. Finally, the conclusions are presented in Sect. VII.

II. Related Work

Over the last few years, work related to disaster communication management has attracted increased attention. Several systems, *e.g.*, SafeCity and M-Urgency, are currently in use for allowing people to communicate during critical and emergency situations. SafeCity allows receiving live mobile video stream of crises and emergency situations [17]. M-Urgency allows users to use iOS or Android to stream live reports over the cellular network to a local rescue point and delivers real-time position through GPS to confirm an immediate and appropriate help for victims [18].

A stringent limitation of such systems is that they need the network infrastructure to be functional in disaster situations. However, the communication infrastructure may not be available to users in disaster areas, which makes communication difficult between victims and rescue teams.

In such a context, D2D communication can help to effectively use the radio resources for collecting useful information from different nodes in the disaster area.

In earlier works, Babun *et al.* [13], [14] developed a multi-hop based D2D communication for the public safety (PS) application. Their proposed multi-hop based communication allows for extending the coverage of the network for disaster scenarios. Simulation results show that their approach helps to increase the energy efficiency with the increasing number of hops. Moreover, they assess the trade-off between spectral and energy efficiency where with the increase of transmit power, energy efficiency decreases to achieve the same level of spectral efficiency.

However, achieving joint ESE is essential but was not considered. Ali *et al.* [15] proposed a D2D communication mechanism and the use of multi-hop communications in public safety scenarios. They modified the nature of proximity service and obtained several benefits in disaster scenarios, such as low end-to-end delays, energy savings and extension of cellular range via UE-to-UE relaying. However, their proposed method is limited only to find out the trade-off between energy efficiency and spectral efficiency, not their joint optimization.

Moreover, in this type of multi-hop scenarios, finding out the best path for routing is essential, but their proposed work does not consider any network parameters for doing so.

While joint ESE optimization has been addressed in LTE networks (*e.g.*, [19]), to the best of our knowledge, very few papers specifically address joint ESE in D2D enabled cellular networks. Recently, [20] proposed an optimal scheme for D2D-enabled mobile-traffic offloading in a D2D overlay cellular networks with the purpose of maximizing the ESE of the network considering maximal cellular user outage and D2D transmitter power constraint. While such work provides useful insight for optimizing ESE in D2D enabled cellular networks, their focus is on the overall network performance, and not on that of the D2D users and the dissemination of critical information via multi-hop routing, as in our case.

Moreover, the last few years have seen the emergence of machine learning based approaches for resource allocation and interference mitigation in D2D enabled networks (*e.g.*, [21] and the references therein). Asheralieva *et al.* [22] proposed an autonomous learning method for joint channel and power level selection by D2D pairs in heterogeneous cellular networks, where D2D pairs operate on the shared cellular/D2D channels. The goal of each device pair is to select jointly 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 requirements. In [23], the authors proposed a machine learning scheme for energy efficient resource allocation in 5G heterogeneous cloud radio access networks. Their centralized resource allocation scheme uses online learning, which guarantees interference mitigation and maximizes energy efficiency while maintaining QoS requirements for all users. Simulation results show that the proposed resource allocation solution can mitigate interference, and increase energy and spectral efficiency significantly. However, so far, such online machine learning approaches have not been applied to optimizing the joint ESE or do not consider the multi-hop routing aspects.

To summarize, the existing research works are scattered in terms of energy efficiency, spectral efficiency, and routing for multi-hop D2D communication. Contrary to the existing works, we propose a combination of light-weight reinforcement learning approach (Q-learning) and ESE optimization. The proposed approach is applied in a hybrid fashion (distributed and centralized)), which contributes to optimizing jointly the ESE in multi-hop D2D communication as well as the multi-hop routing by finding out the best path considering the network parameters.

TABLE I
NOTATIONS FOR THE SYSTEM MODEL.

K	Number of cells
C_k	Set of CU devices in cell k
D_k	Set of D2D devices in cell k
M_k	Set of available channels in cell k
Λ_k	Set of multi-hop paths for the D2D devices in cell k
$L_{\lambda,k}$	Set of D2D links in the λ-th path in Λ_k
c	Index for a generic CU belonging to C_k
d	Index for a generic D2D belonging to D_k or $L_{\lambda,k}$
λ	Index of a given multi-hop path in Λ_k
$\lvert\cdot\rvert$	Number of elements in the sets C_k, D_k, M_k, Λ_k

III. System Model

The basic architecture of the D2D communication underlaying C-RAN consists of the following elements: (a) a Base band unit (BBU) pool/proximity server, (b) some RRHs, (c) Cellular users (CUs), and (d) D2D users, similarly to the scenario in [24].

Fig. 1 shows a PS scenario where some active RRHs communicate with the C-RAN server of the deployed command center. The BBU pool/proximity server consists of BBUs that can be treated as a virtual BS (VBS) to enable the functions of the network. The fronthaul (FH) links help to connect the RRHs to the proximity server. Some UAVs, acting as RRHs, are also deployed for communication. Due to the out-of-radio coverage situation, D2D devices need to communicate with each other in a multi-hop fashion. There are several types of interference (both inter-cell and intra-cell) in this scenario, i.e., interference from D2D users to RRHs, from cellular users to D2D users and from one D2D link to another. Our goal is to maximize the overall energy and spectral efficiency of the network exploiting multi-hop D2D communication and maintaining the communication links between CUs and active RRHs. So, interference to these links should be minimized.

We assume that there are K adjacent cells, $\{k = 1, 2, 3, \ldots, K\}$. In cell k, the set of D2D and CU links are denoted as D_k and C_k, respectively. A key-point of this paper is that a link connecting two D2D users or a device to an RRH

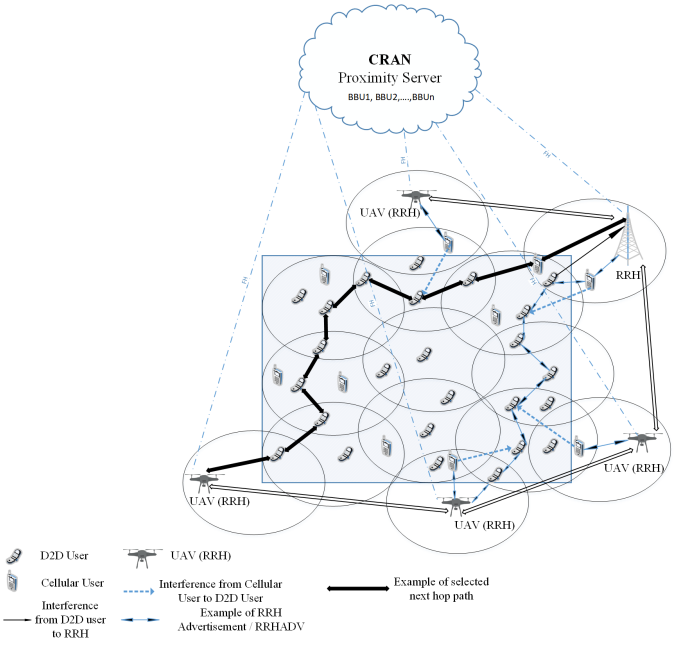


Fig. 1. The overall scenario of multi-hop D2D communications underlaid C-RAN for public safety application.

can be part of a path made of multiple links, activated by a set of D2D devices. Therefore, not all the active D2D devices are transmitting their own data but some can operate simply as relays in the multi-hop communication path. The set of multi-hop paths for the D2D devices in the k-th cell is denoted as Λ_k and, to each path $\lambda \in \Lambda_k$ it is associated the corresponding set of D2D links $L_{\lambda,k}$; since a D2D device can belong just to one multi-hop path, we have that $\bigcup_{\lambda\in\Lambda_k} L_{\lambda,k} = D_k$. In addition, D2D devices operating as relays can transmit and receive data on different channels and they are supposed to reuse the channels available for standard CUs. The set of available channels in the generic cell k is denoted as M_k, including D2D single connections, which can be carried out with LTE-Direct or WiFi-Direct technologies.

A summary of the notations is also reported in Table I.

IV. Problem Formulation

In this section, we aim to achieve the formulation of the cost function with a set of constraints in order to achieve the performance enhancement of the system energy and spectral efficiency exploiting D2D, possibly multi-hop, communication, concurrent with the standard cellular CU links. The result of the optimization, in which the limiting factor is constituted by the mutual interference, will provide:

- The assignment of the channels in M_k for each cell k; the set of channels will be common and reused by all the cells.
- The selection of the multi-hop D2D connections, taking advantage from the energy savings and from the reuse of the cellular resources.
- The assignment of the powers of CU and D2D devices, taking into account the used technology and the power/rate constraints.

Let us start expressing the achievable spectral efficiency (SE) of the D2D link d in the k-th cell ($d \in D_k, k = \{1, 2, \cdots, K\}$) on the set of available channels M_k, given by

$$SE_{D2D,d,k} = \frac{1}{|M_k|} \sum_{m\in M_k} \log_2\left(1 + \frac{\gamma^m_{D2D,d,k} \cdot p^m_{D2D,d,k} \cdot g^m_{d,k}}{I^m_{ID,d,k} + I^m_{ED,d,k} + I^m_{IC,d,k} + I^m_{EC,d,k} + P_N}\right), \tag{1}$$

where $|M_k|$ is the number of channels, $p^m_{D2D,d,k}$ is the transmission power of the D2D transmitter in the d-th link of the k-th cell, $\gamma^m_{D2D,d,k}$ is the D2D binary channel selection indicator, P_N is the noise power and $g^m_{d,k}$ is the D2D signal channel gain; the channel selection indicator $\gamma^m_{D2D,d,k}$ is set to 1 when the m-th channel is reused as the transmit channel in this link, otherwise $\gamma^m_{D2D,d,k} = 0$. The terms that compose the total interference are described in the following list:

- $I^m_{ID,d,k}$, the intra-cell interference on the d-th D2D link from the other D2D links in the m-th channel since D2D devices can reuse the same channel multiple times in the same cell:

$$I^m_{ID,d,k} = \sum_{\hat{d}\in D_k\setminus\{d\}} \gamma^m_{D2D,\hat{d},k} \cdot p^m_{D2D,\hat{d},k} \cdot g^m_{\hat{d},d,k}, \tag{2}$$

  where each term in the sum is the intra-cell interference from the $\hat{d}$-th D2D interferer ($\hat{d} \neq d$).
- $I^m_{ED,d,k}$, the inter-cell interference caused by external D2D links in the m-th channel:

$$I^m_{ED,d,k} = \sum_{k'\in K\setminus\{k\}} \sum_{\hat{d}\in D_{k'}} \gamma^m_{D2D,\hat{d},k'} \cdot p^m_{D2D,\hat{d},k'} \cdot g^m_{\hat{d},d,k',k}, \tag{3}$$

  where each term in the double sum is the interference on the m-th channel from the D2D interferer in the k'-th cell to the receiver of the d-th D2D link in the k-th cell ($k' \neq k$).
- $I_{IC,d,k}$, the intra-cell interference caused by CUs in the m-th channel:

$$I_{IC,d,k} = \sum_{c\in C_k} \gamma^m_{CU,c,k} p^m_{CU,c,k} \cdot h^m_{c,d,k}, \tag{4}$$

  where $\gamma^m_{CU,c,k}$ is the CU binary channel selection indicator, equal to 1 if the c-th device in the k-th cell occupies channel m and each term in the sum is the intra-cell interference from the c-th CU interferer to the D2D receiver of the d-th link (a different symbol h is used for the channel gains between a CU transmitter and a D2D device). We remark that only one term in the sum will be different from 0 (so $\gamma^m_{CU,c,k} = 1$ for one c and 0 for the others) since the links cannot be reused by more CU terminals in the same cell.
- $I_{EC,d,k}$, the inter-cell interference caused by CUs in the m-th channel:

$$I_{EC,d,k} = \sum_{k'\in K\setminus\{k\}} \sum_{c\in C_k} \gamma^m_{CU,c,k'} p^m_{CU,c,k'} \cdot h^m_{c,d,k',k}, \tag{5}$$

where each term in the sum denotes the inter-cell interference from the c-th CU interferer in the k'-th cell to the d-th D2D link in the k-th cell.

Similarly, the SE of the c-th CU in the k-th cell can be expressed as

$$SE_{CU,c,k} = \frac{1}{|M_k|} \sum_{m \in M_k} \log_2 \left(1 + \frac{\gamma^m_{CU,c,k} \cdot p^m_{CU,c,k} \cdot h^m_{c,k}}{I^m_{IDC,c,k} + I^m_{EDC,c,k} + I^m_{EC,c,k} + P_N} \right), \quad (6)$$

where $p^m_{CU,c,k} \cdot h^m_{c,k}$ is the signal power received by the associated RRH in the m-th channel. Again, the terms that compose the total interference are classified as follows:

- $I^m_{IDC,c,k}$, the intra-cell interference caused by D2D links in the same k-th cell, namely
$$I^m_{IDC,c,k} = \sum_{d \in D_k} \gamma^m_{D2D,d,k} \cdot p^m_{D2D,d,k} \cdot \bar{h}^m_{c,d,k}, \quad (7)$$
where each term in the sum is the intra-cell interference from the d-th D2D interferer on the m-th channel. The symbol $\bar{h}^m_{c,d,k}$ denotes the channel gain between D2D transmitter and CU link (the RRH in this case), which is different from the previous $h^m_{c,d,k}$, defined between the CU transmitter and the D2D receiver.
- $I^m_{EDC,c,k}$, the inter-cell interference caused by D2D pairs in the other cells,
$$I^m_{EDC,c,k} = \sum_{k' \in K \setminus \{k\}} \sum_{d \in D_{k'}} \gamma^m_{D2D,d,k'} \cdot p^m_{D2D,d,k'} \cdot \bar{h}^m_{c,d,k,k'}, \quad (8)$$
where each term is the inter-cell interference on the m-th channel from the d-th D2D interferer in the k'-th cell.
- $I^m_{EC,c,k}$, the inter-cell interference caused by CUs in adjacent cells, or
$$I^m_{EC,c,k} = \sum_{k' \in K \setminus \{k\}} \sum_{c' \in C_{k'}} \gamma^m_{CU,c',k'} \cdot p^m_{CU,c',k'} \cdot y^m_{c,c',k,k'}, \quad (9)$$
where each term is the inter-cell interference from the c'-th cellular interferer in the k'-th cell (a different symbol y is used for the channel gains between a CU device and the RRH associated with another CU).

Now the overall SE ([bit/s/Hz]) in the generic k-th cell can be expressed as

$$SE_k = \sum_{c \in C_k} (SE_{CU,c,k}) + \sum_{\lambda \in \Lambda_k} \left(\min_{d \in L_{\lambda,k}} \{SE_{D2D,d,k}\} \right), \quad (10)$$

where the minimum of the spectral efficiencies among those associated with the links in a generic multi-hop D2D path represents the maximum achievable data stream in the path, as the final rate has to be clearly adapted to the minimum among the theoretical capacities of the consecutive links.

On the other hand, the energy efficiency (EE, [bit/J]) can be denoted as

$$EE_k = \frac{\sum_{c \in C_k} B \cdot SE_{CU,c,k} + \sum_{\lambda \in \Lambda_k} B \min_{d \in L_{\lambda,k}} \{SE_{D2D,d,k}\}}{\sum_{c \in C_k} p_{CU,c,k} + \sum_{d \in D_k} p_{D2D,d,k}}, \quad (11)$$

where $p_{CU,c,k} = \sum_{m \in M_k} p^m_{CU,c,k}$ is the total power spent by each CU, $p_{D2D,d,k} = \sum_{m \in M_k} p^m_{D2D,d,k}$ is the total transmission power spent by each D2D terminal and B is the channel bandwidth. Of course, at the second term of the denominator, related to the D2D devices with multi-hop paths, the sum takes into account all the powers spent by the D2D devices involved in a single path, including the relays.

Now, we are ready to propose a formulation for an effective joint optimization of the spectral and energy efficiency:

1) The powers $p_{D2D,d,k}$ and $p_{CU,c,k}$ are replaced by the total powers, including the circuit consumption, given by
$$p_{D2Dtot,d,k} = \frac{1}{\eta} p_{D2D,d,k} + 2 \cdot p_{cir}, \quad (12)$$
and
$$p_{CUtot,c,k} = \frac{1}{\eta} p_{CU,c,k} + p_{cir}, \quad (13)$$
respectively. The circuit power of both the D2D transmitter and receiver is denoted by $2p_{cir}$, η is the power amplifier (PA) efficiency ($0 < \eta < 1$) and the circuit power of the transmitter UE is just a single term p_{cir} [25].
2) The term $\min_{d \in L_{\lambda,k}} \{SE_{D2D,d,k}\}$, which is challenging to manage in the optimization process, is approximated by $(\sum_{d \in L_{\lambda,k}} SE_{D2D,d,k})/|L_{\lambda,k}|$, where $|L_{\lambda,k}|$ is the number of hops in the corresponding path. This approximation is equivalent to considering all the rates in each link of a multi-hop path approximately equal, at least for their impact on the optimization function (in any case the real final rate assignment will respect the limitation given by the link with the minimum rate). This assumption has an additional impact, i.e. it mitigates the adoption of a large number of hops in the system, for their impact on the interference with other devices and on the SE when we use many relays.

Therefore, we express our ESE function for each cell k as

$$ESE_k = \frac{1}{\sum_{c \in C_k} p_{CU,c,k} + \sum_{d \in D_k} p_{D2Dtot,d,k}} \times \times B \left(\sum_{c \in C_k} SE_{CUtot,c,k} + \sum_{\lambda \in \Lambda_k} \sum_{d \in L_{\lambda,k}} \frac{SE_{D2D,d,k}}{|L_{\lambda,k}|} \right), \quad (14)$$

where the peculiarity is the role of the factor $|L_{\lambda,k}|$, i.e. the number of hops in each path involving D2D devices: this factor limits the adoption of the multi-hop feature to the cases in which it is really advantageous for the overall system, avoiding excessive proliferation of mutual interference phenomena and spectral efficiency losses. At the same time, we remark that D2D links (the second term in (14)) are reusing channels already occupied by CU terminals (contributing to ESE by means of the first term in (14)) and this constitutes a potential gain on the overall spectral efficiency; at the same time, the activation of links in the second term of (14) increases the interference, causing a decrease of the first term. In this trade-off, the optimization process is supposed to find the correct optimal working point.

Now, the problem formulation for each cell k is given by the maximization of (14) with the set of constraints $\{C1, C2, C3, C4, C5\}$ regarding the maximum power, the minimum service rate and the binary channel allocation indicators:

$$\begin{aligned}
\max \quad & ESE_k = \\
& \frac{\sum_{c\in C_k} SE_{CU,c,k} + \sum_{\lambda\in\Lambda_k}\sum_{d\in L_{\lambda,k}} SE_{D2D,d,k}/|L_{\lambda,k}|}{1/B(\sum_{c\in C_k} p_{CU,c,k} + \sum_{d\in D_k} p_{D2Dtot,d,k})} \\
\text{s.t.} \quad & C1 : p_{D2Dtot,d,k} \leq p_{max} \\
& C2 : p_{CUtot,c,k} \leq p_{max} \\
& C3 : SE_{D2D,d,k} \geq SE_{D2D,min} \\
& C4 : SE_{CU,c,k} \geq SE_{CU,min} \\
& C5 : \{\gamma^m_{CU,c,k}, \gamma^m_{D2D,d,k}\} \in \{0,1\}
\end{aligned} \tag{15}$$

where C1 is the maximum transmission power constraint, i.e., the transmission power should not be greater than P_{max}; C2 is the constraint for binary channel selection indicators; C3 and C4 are the constraints for the minimum level of SE for the cellular users and D2D users, respectively; and C5 and C6 are the constraints for the minimum level of power consumption for the D2D and cellular users, respectively.

V. The Proposed Method for Dynamic Adaptation of Joint Energy-Spectral Efficiency

Reinforcement learning (RL) helps to learn the optimal action in a dynamic environment. One of the challenges in RL is the trade-off between exploitation and exploration. The actions are performed by agents on a trial-and-error basis during the interaction with the environment; the agents need to exploit the information collected by the learning algorithm and they also need to explore new actions and states for finding better policies and reach learning convergence, i.e., in our case to achieve the optimal level of ESE, which maximizes the reward in the long run. Moreover, our proposed RL approach selects the best path for multi-hop D2D routing considering important network parameters, i.e., RRH load, congestion level, link quality metric, average number of hops, and throughput. By optimizing the routing path, the proposed algorithm helps the agents to behave so that both energy and spectral efficiency are optimized.

In this work, we use Q-learning, one type of RL, as it is straightforward and has lower computational complexity and execution time w.r.t. other variants [26].

Fig. 2 shows the idea behind the Q-learning process designed for maximizing the energy spectral efficiency. On the right side of Fig. 2, the Q-learning process is composed of a reward function (Rf) (defined in (22)), which is a function of link quality metrics (LQMs) (defined in (19)); LQMs are determined by interference levels, power and congestion levels (CL) (defined in (18)). The rationale behind this design is that Q-learning helps to optimize Rf, who is coherent with a maximization of the spectral and energy efficiency of each link according to the ESE formulation in (14). For increasing Rf, the LQMs of the links have to be reduced,

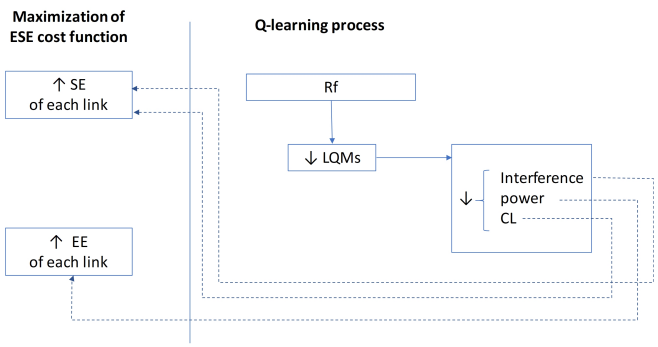


Fig. 2. Q-learning process for the maximization of energy spectral efficiency.

through the reduction of interference, powers and CL; in particular, CL is a function of the link utilization U, which is inversely proportional to the link throughput. Therefore, minimizing CL is consistent with the SE increase of link and the same can be observed for the interference levels. At the same time, the inclusion of power reductions as part of the LQMs minimization, provides the leverage for the energy efficiency enhancement. Throughput in (21) should approach the maximum spectral efficiencies defined in (1) and (6) multiplied by the bandwidth B. In the learning process, the throughput is generated considering a full queue of packets from each node in order to look at the network potential and it is maximized through the minimization of LQM. The reuse of a channel for a D2D link is considered in the process but it is limited again by the role of interference in LQM as it should be in the allocation process. Finally, it is crucial to observe, for the relation with the formulation of the final ESE cost function (14), the role of the factor $|L_{\lambda,k}|$ in the learning process: using more paths for the same link is not directly limited in the reward function Rf but it causes indirectly more interference and more congestion in the interested nodes without increasing the rate. Therefore, it is expected that the learning process will select this option only when the impact on LQM is really limited, generating an enhancement of the spectral efficiency.

A. Network Parameters and Components in the Proposed Method

1) RRH Load: Determining the RRH with the least amount of load is useful in a disaster scenario since the least loaded RRH can send feedback rapidly and can be freed for load advertising again [27]. The RRH load depends on the volume of traffic that the RRH has just processed, its current load and its previous estimated load; this is calculated, as a function of a time step or iteration t, as

$$L_c[t] = \Xi\frac{v[t]}{QL} + (1-\Xi)L_c[t-1], \tag{16}$$

where $\Xi \in [0,1]$ is a weighting coefficient for choosing the impact of the new load measurement. The term $v[t]$ is the volume of traffic that the RRH has just processed and QL is the maximum queue length of the RRH.

2) Congestion Level: The utilization U of generic link l is defined as follows:

$$U(l,t) = \frac{\sum_{i \in succ(l,t)} Size_i}{BW_l}, \quad (17)$$

where $succ(l,t)$ denotes the number of packets traversed link l successfully during time t and $Size_i$ denotes the size of packet i. BW_l is the capacity of link l, obtained by the spectral efficiencies derived in Sect. IV and multiplied by the bandwidth B. Thus, we define the congestion level (CL) as

$$CL(l,t) = \beta U(l,t) + (1-\beta) CL(l,t-1), \quad (18)$$

where the parameter $\beta \in [0,1]$.

3) Link Quality Metric: We define the link quality metric (LQM) for link l operating on a generic channel as

$$LQM_l = \left(1 - \left(\frac{1}{2}\right)^n\right)(I_{r_l} + p_l) + \left(\frac{1}{2}\right)^n CL(l,t), \quad (19)$$

where I_{r_l} is the interference level and p_l is the total power consumption of a generic D2D or CU device, as defined in (12), (13) and used in (14). The reason behind the definition of LQM_l is our intention to capture interferences, power consumption, and congestion in computing the link quality. A larger value of n gives more importance to interference; on the other hand, for $n = 1$, interference and congestion have the same weight (0.5 each). Smaller LQM_l values for a given link reflect better quality.

Then we define the path quality (PQ) of a path having L hops from a UE (D2D or CU) to an RRH as

$$PQ = \sum_{l=1}^{L} \frac{(LQM_l)^{n_H}}{1 - LQM_l}, \quad (20)$$

where n_H is the hop distance of the link starting from the RRH [27].

A low value of PQ reflects a good quality path and viceversa. The values of LQM_l vary between 0 and 1. The ratio $\frac{(LQM_l)^{n_H}}{1-LQM_l}$ increases significantly for high values of LQM_l. Thus, PQ gives more importance/weight to links in the neighborhood of RRHs. This means that paths with high quality links in the neighborhood of RRHs will be preferred over other paths.

4) Throughput: The average throughput is defined as the sum of the total amount of bits successfully received by all active users in the system divided by the product of the number of cells in the system and the transmission time interval (TTI) (which for LTE is 1 ms),

$$Throughput = \frac{\sum_{c=1}^{C} \sum_{u=1}^{U} \beta_u^c}{K \times T_{sim}}, \quad (21)$$

where K is the total number of cells, T_{sim} is the simulation time per run, β_u^c is the number of bits received with success by user u in cell c.

5) State: The set of states is represented by $S = \{S_n, S_p\}$ where S_n is the set of an UE's neighboring nodes, and S_p is the set of packets to be forwarded.

Algorithm 1 RRH selection algorithm

Input: RRHADV
Output: Least loaded RRH
while B **do**attery lifetime is not equal to zero
  Receive RRHADV from a RRH
  Update corresponding entry in RRHTable
  Calculate the RRH load:

$$L_c[t] = \Xi \frac{v[t]}{QL} + (1-\Xi) L_c[t-1]$$

  BestRRH $\leftarrow$ RRH with the minimum load
end while

6) Actions:

- **Forward:** Forwards a packet with selected transmit power level, $A_f = a(s_j|s_i), j \in I$; execution of $a(s_j|s_i)$ means that UE i forwards a packet to UE $j \in I$ with a given transmit power level, and sends feedback to the predecessor. I denotes the set of the i's neighboring UEs. The transmit power p_{dc} is selected in the range of $[0, p_{max}]$.
- **Drop:** Drop, A_d, drops the data packet.

7) Reward Function: We define the reward function (Rf) as follows:

$$Rf = \begin{cases} \frac{LQM_l}{1-LQM_l} + LQM_l \times A & \text{for } n \leq N, \\ \frac{LQM_l}{1-LQM_l} & n > N, \end{cases} \quad (22)$$

where N is the maximum number of hops in the link and n is the current hop starting from the RRH. A assumes a predefined constant value; a large enough value of A (*e.g.*, 100) allows to differentiate good paths from poor paths and eliminate the poor ones from the routing tables.

B. Proposed reinforcement learning for dynamic adaptation

One of the RL strategies for multi-agent based scheduling is based on multiple independent learners. Each agent learns independently based on local states and local rewards. This strategy may lead to an anomalous situation for action selection because there is no communication among the agents and they do not have any real view of the entire system. In this case, an agent learns according to its local information and a coordination mechanism is required. We consider two types of agents in our environment, *e.g.*, scheduler agent (the distributed UEs) and learner agent (the RRHs). The scheduler agents submit their local rewards to a learner agent. The learner agent collects the rewards and updates an utility table that holds the corresponding efficiency of executing action. Then, the learner agent sends the updated utility table to the scheduler agents which can then make their decisions. In each case, the data transmission energy has been considered. In the sequel, each step of the learning method is explained.

1) RRH Selection Algorithm: RRHs advertise the load to the network. The first step is to find out the RRH with the minimum load. Algorithm 1 shows the steps of the RRH selection algorithm where the input is RRH advertisements and the output is the least loaded RRH.

Algorithm 2 Next hop selection algorithm at D2D device i

Input: Packet with RRH destination RRH_d
Output: Best next hop to RRH_d
Variables: RoutingTable, j
while Battery lifetime is not equal to zero **do**
Receive a packet with destination RRH
Determine the next-hop corresponding to the path with the smallest path quality (PQ):

$$PQ = \sum_{l=1}^{L} \frac{(LQM_l)^{n_H}}{1 - LQM_l}$$

Send packet to j with selected level of p_{dc}
Receive feedback/reward, Rf from j

$$Rf = \begin{cases} \frac{LQM_l}{1-LQM_l} + LQM_l \times A & \text{for } n \leq N. \\ \frac{LQM_l}{1-LQM_l} & n > N. \end{cases}$$

Update the Q value for Q-learning
Update the corresponding entry in the table, RRHTable
end while

2) Next hop Selection Algorithm: D2D devices have the packet with RRH destination RRH_d. UEs need to find out the next hop to RRH_d. Now, until the battery lifetime is not equal to zero, D2D devices receive a packet with the destination RRH. Then, the agents/UEs determines the next-hop corresponding to the path with the smallest path quality and sends the packet to the neighbor, j. After that, the sending agents receive feedback/reward from the neighbor j. Finally, the Q-value for Q-learning (see sec. V-B3) is updated. Algorithm 2 shows the steps for selecting the next hop for a given D2D device i.

Figure 3 depicts the multi-hop routing scenario considering Algorithms 1 and 2. The figure illustrates the RRHADV (RRH advertisements) in the network where UEs receive RRHADV and select the least loaded RRH, as well as the hop selection based on the path quality (PQ), where UEs select the next hop corresponding to the path with the smallest PQ. For example, in the figure there are two paths with $PQ = 0.25$ and $PQ = 0.50$; UEs select the hop with the smallest one ($PQ = 0.25$) according to the proposed algorithm.

3) Q-learning for dynamic adaptation: We apply Q-learning for optimizing the routing path and ESE. In our case, the components are:

- *Agent:* Each UE denotes an agent responsible for executing the online learning algorithm.
- *Environment:* The application represents the environment in our approach. Interaction between the agents and the environment is achieved by executing actions and receiving a reward function.
- *Action:* Agent action is the currently executed application task on the UEs.
- *State:* A state describes a particular scenario of the environment based on some application oriented variables.
- *Policy:* Agent policy determines what action will be selected in a particular state. This policy determines which action to execute at the perceived state and focuses more on exploration or exploitation, depending on the selected setting of the learning algorithm.
- *Value function:* This function defines what is good for an agent over the long run. It is built upon the reward function values over time and its quality exclusively depends on the reward function.
- *Reward function:* This function provides a mapping of the agent's state and the corresponding action with a reward value that contributes to the performance.

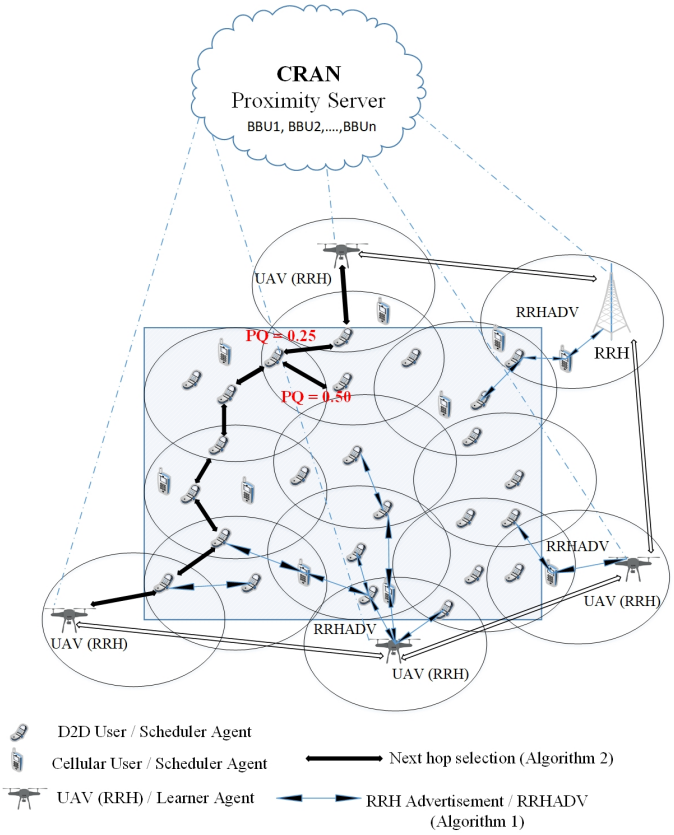


Fig. 3. Illustration for **Algorithm 1** and **Algorithm 2**. **Algorithm 1:** RRH advertisement/RRHADV from RRHs. Scheduler agents/UEs receive RRHADV from RRHs to select the least loaded RRH/BestRRH. **Algorithm 2:** Next hop selection for UEs/Scheduler agents and routing toward the destination RRH. UEs select the next hop to the path with the smallest value of PQ.

In the Q learning framework, the agents learn the utility of performing various actions over time steps using the local information [28]. As shown in Algorithm 3, here the UEs try to maximize the utility and maintain a Q matrix for the value functions. Initially, all the Q values of the matrix are zeros. At each state the UEs perform an action; depending on the reward function and the Q value, they move to the next state which maximizes the system performance.

The Q value is updated as follows for the (state, action) pair

$$Q_{t+1}(s_t, a_t) = (1-\varphi)Q_t(s_t, a_t) + \varphi(r_{t+1}(s_{t+1}) + \Gamma V_t(s_{t+1})) \quad (23)$$

$$V_{t+1}(s_t) = \max_{a \in A} Q_{t+1}(s_t, a) \quad (24)$$

Algorithm 3 Proposed reinforcement learning algorithm

Initialize $Q(s,a)=0$ where s is the set of states and a is the set of actions

while Battery lifetime is not equal to zero **do**

Determine current state

Select action a based on policy

$$\frac{e^{Q(s,a)}/\omega}{\sum_a (e^{Q(s,a)/\omega})}$$

Execute the selected action

Calculate the reward

$$Rf = \begin{cases} \frac{LQM_l}{1-LQM_l} + LQM_l \times A & \text{for } n \leq N. \\ \frac{LQM_l}{1-LQM_l} & n > N. \end{cases}$$

Calculate the learning rate

$$\varphi = \frac{Z}{visited(s,a)}$$

Calculate Q value for the executed action

$$Q_{t+1}(s_t,a_t) = (1-\varphi)Q_t(s_t,a_t)+\varphi(Rf(s_{t+1})+\Gamma V_t(s_{t+1}))$$

Calculate the value function for the executed action

$$V_{t+1}(s_t) = \max_{a \in A} Q_{t+1}(s_t,a)$$

Update the utility table of the scheduler agent

$$U(q) = (1-\Upsilon)U(q) + \Upsilon \sum_i Rf_i$$

Move to the next state based on the executed action

end while

TABLE II
REINFORCEMENT LEARNING PARAMETERS

Parameter	*Value*
$Q_{t+1}(s_t,a_t)$	update of the Q value at time $t+1$
$\max_{a \in A} Q_{t+1}(s_t,a)$	Maximum Q value
r_{t+1}	Immediate Reward
v_t	Value function
Γ	Discount factor
φ	Learning rate
ω	Temperature constant

where $Q_{t+1}(s_t,a_t)$ is the update of the Q value at time $t+1$ after executing action a at time step t. r_{t+1} is the immediate reward after executing the action a at time t, V_t is the value function for node at time t and V_{t+1} is the value function at time $t+1$. The term $\max_{a \in A} Q_{t+1}(s_t,a)$ is the maximum Q value after performing an action from the action set A for the agent i. The parameter Γ is the discount-factor which can be set to a value in $[0,1]$; for higher Γ values, the agent relies more on the future than the immediate reward. Finally φ is the learning rate parameter which can be set to a value in $[0,1]$ [29]; it controls the rate at which an agent tries to learn by giving more or less weight to the previously learned utility value. When φ is close to 1, the agent gives more priority to the previously learned utility value.

Here we use soft-max strategy and Boltzmann distributions for the exploration and exploitation [30]. The probability of selecting an action a in state s is proportional to $e^{Q(s,a)}\omega$. That is, at state s, agents select an action based on the probability

$$\frac{e^{Q(s,a)}/\omega}{\sum_a (e^{Q(s,a)/\omega})}, \tag{25}$$

where ω is the temperature constant. If $\omega > 0$, the agents will focus on choosing the actions randomly, i.e., exploration. On the other hand, if $\omega \to 0$, the best action based on Q-values is chosen, i.e., exploitation. The learning rate, φ is slowly decreased in order to take into account the impact of the visited state-action pair, i.e.

$$\varphi = \frac{Z}{visited(s,a)}, \tag{26}$$

where Z is the positive constant and $visited(s,a)$ is the number of visited state-action pairs so far.

4) Updating the utility table: In each time step, the learner agent receives the reward from all scheduler agents and updates the utility table, $U(q)$:

$$U(q) = (1-\Upsilon)U(q) + \Upsilon \sum_i Rf_i \tag{27}$$

where Υ is learning factor and Rf_i is the reward vector generated by the i-th agent [31].

After updating the utility table, the learner agent sends it back to the scheduler agents; they will generate the rewards and submit the reward vector to the learner agent. Table II shows the main parameters of the reinforcement learning.

VI. PERFORMANCE EVALUATION

In this section, we present the simulation results in order to discuss the performance of our proposed approach in terms of joint ESE and remaining energy of the network.

We evaluate our approach against existing variants of RL, i.e., state-action-reward-state-action (SARSA), Q(λ), and SARSA(λ) [32] and observe that we outperform them all. SARSA shows an increase of EE until a similar number of stages as with our approach, but then EE decreases very rapidly due to the number of depleted UEs. Indeed, in Q-learning, the Q-value is updated with the maximum valued action at the next state; on the other hand, in SARSA, the update is dependent upon the action that is actually taken at the next state. While SARSA is suitable when the agents focus on exploration [33], it is risky for dynamic scenarios like ours, where there exist a number of depleted UEs over time as well as interferences. The Q(λ) algorithm is similar to Q-learning except for the eligibility traces. Q(λ) stops learning at the iteration where the agent selects the exploratory action and eligibility traces for all state-action pairs are set to zero, which is also not suitable for dynamic scenarios like ours [34].

A. Simulation setup

Rouil *et al.* [35] provide an implementation of LTE D2D functionalities (direct communication, direct discovery, and out of coverage D2D synchronization) for NS3. We extend their implemented model by adding C-RAN functionalities

TABLE III
SIMULATION PARAMETERS

Parameter	*Value*
Area	1000×1000 m
Total Number of UEs	60
Cell radius	300 m
Intercell distance	500 m
Number of RRHs	5
Maximum Tx power p_{max}	23 dBm
Bandwidth B	180 KHz
Constant circuit power p_{cir}	10 dBm
Thermal noise power P_N	10^{-7} W
Battery capacity	800 mAh
p	0.5
β	0.8
φ	0.5
Γ	0.5
e_{tx}	0.3
τ_{tx}	0.3
e_{rx}	0.4
τ_{rx}	0.4
k_{tx}	0.2
λ_{tx}	0.2
k_{rx}	0.2
λ_{rx}	0.2
λ	0.5
$SE_{D2D,min}$	1.90
$SE_{CU,min}$	0.30
Z	1
Υ	0.5
ξ	-6 dB
ϱ	4
n	1

(BBUs and RRH/UAVs), for communication purposes and an ad-hoc routing protocol for multi-hop D2D communication.

The channel gain between the transmitter i and the receiver j is proportional to $d_{i,j}^{-2}|h_{i,j}|^2$, where $d_{i,j}$ is the distance between the transmitter i and the receiver j. $h_{i,j}$ is the complex Gaussian channel coefficient that satisfies $h_{i,j} \sim CN(0,1)$ [36]. Each simulation starts with the UEs of random initial amounts of battery charge between 1 and 800 [mAh]. The simulation results are averaged over 10 simulation runs. The location of the UEs are generated randomly in each simulation run.

Table III shows the simulation parameters used in the simulation.

B. Results for Stand-alone EE and SE

In this subsection, stand-alone denotes the optimization of either EE or SE.

1) Stand-alone EE performance evaluation: Fig. 4 shows the stand-alone EE optimization (calculated by Equation (11)) and non-optimized SE (calculated by Equation (10)). We can observe that our proposed method helps to increase EE until 160 iterations. Then, with the increasing number of depleted UEs, EE progressively decreases.

The Q(λ) algorithm is similar to Q-learning except for the eligibility traces. Q(λ) stops learning at the iteration where the agent selects the exploratory action and eligibility traces for all state-action pairs are set to zero, which is also not suitable for dynamic scenarios like ours [34].

The algorithm Q(λ) shows lower EE as compared to both our approach and SARSA up to 210 iterations, after which it is a bit better than SARSA (Q(λ) takes a bit of time to take over SARSA due to the accumulating traces of non-greedy actions).

SARSA(λ) is similar to SARSA with the eligibility traces as well. That is why using SARSA(λ), EE is lower as compared to the other methods, although a similar trend can be observed for both SARSA(λ) and Q(λ). SARSA(λ) takes over SARSA after 205 iterations due to the accumulating eligibility traces.

Furthermore, we can observe that, from the $50-th$ iteration till to the $250-th$, all learning algorithms behave similarly, until reaching the convergence level. After reaching the convergence learning level, our proposed reinforcement learning outperforms the other methods drastically.

Finally, since SE is not optimized, it rapidly decreases after 50 iterations due to the depleted number of UEs and interferences. This motivates the need for joint ESE optimization, as shown later in Section VI-C.

2) Stand-alone SE performance evaluation: Fig. 5 shows the stand-alone SE optimization and unoptimized EE. We can observe that the optimized SE remains maximized for a longer time as compared to the baseline technique. However, with the increasing number of depleted UEs, the SE decreases (after approx. 170 iterations). But EE without optimization goes down even more rapidly. Our proposed learning method outperforms the three other learning variants. SARSA performs better than Q(λ) and SARSA(λ) until 270 iterations; then SE decreases rapidly due to SARSA's focus on exploration. Q(λ) performs better than SARSA(λ) at every iteration. Our approach benefits from the exploration-exploitation strategy and heuristic learning rate update mechanism, which helps the agents to behave in an adaptive manner in the environment for achieving higher SE. On the other hand, the baseline (unoptimized) EE decreases rapidly after only 75 iterations as it does not have any dynamic adaptation to face the changes in the environment.

C. Performance Evaluation of joint ESE

Fig. 6 (a) shows the joint ESE using our proposed approach over the number of iterations. It yields the best results among all evaluated approaches, achieving convergence at iteration 138 and remaining at the same level until iteration 165. Then joint ESE decreases because of the energy depletion in the UEs. At this point, we can sacrifice delay and focus on joint ESE to keep the network alive as long as possible. Here, we observe the same trend as in all methods for stand-alone EE and stand-alone SE. Again, our approach performs better than the others thanks to the algorithm's exploration-exploitation strategy and heuristic learning rate. Initially, SARSA performs better than Q(λ) and SARSA(λ), but is overtaken by Q(λ) at approx. 260 iterations and by SARSA(λ) at approx. 300 iterations thanks to their accumulating traces.

On the other hand, for the baseline (unoptimized) method, the joint ESE decreases very sharply after only 50 iterations. Our proposed method outperforms the joint ESE by 67% as compared to the baseline algorithm.

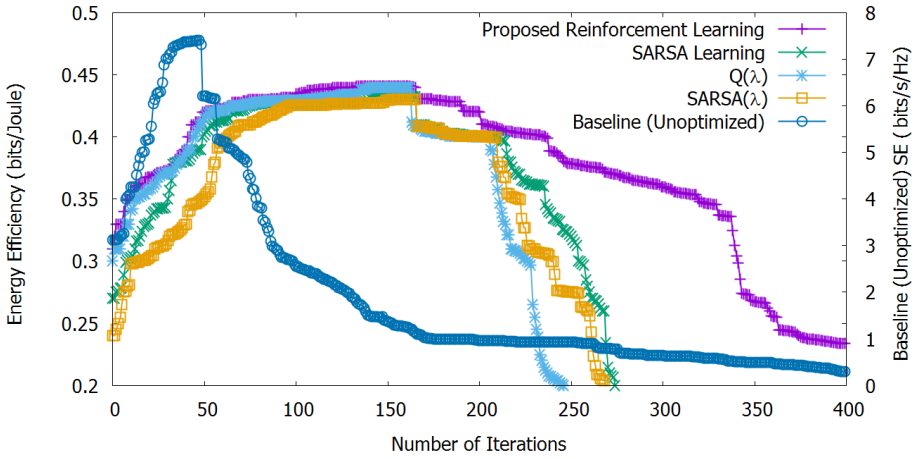


Fig. 4. Stand-alone energy efficiency performance evaluation and non-optimized spectral efficiency. The proposed method helps to increase EE until 160 iterations; afterwards, energy efficiency progressively decreases due to the increasing number of depleted UEs.

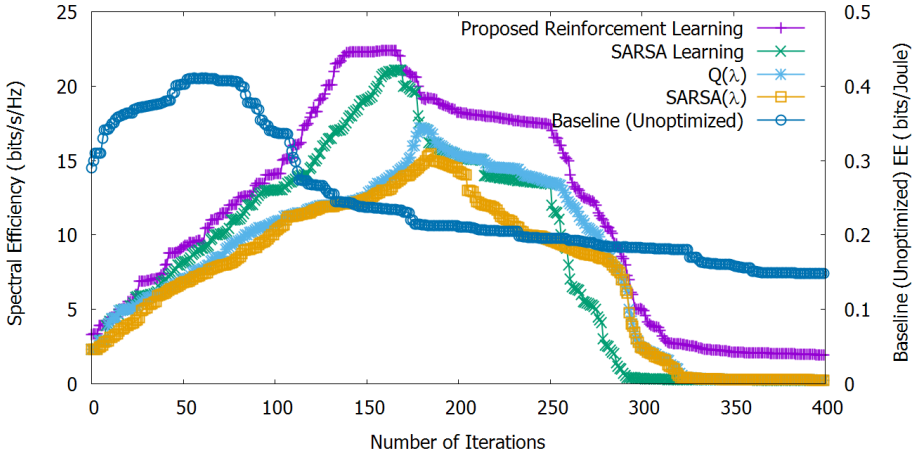
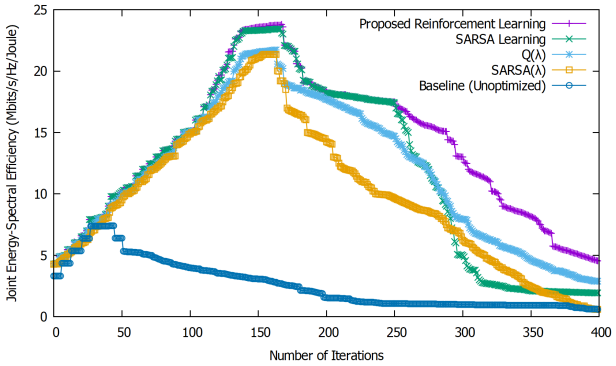
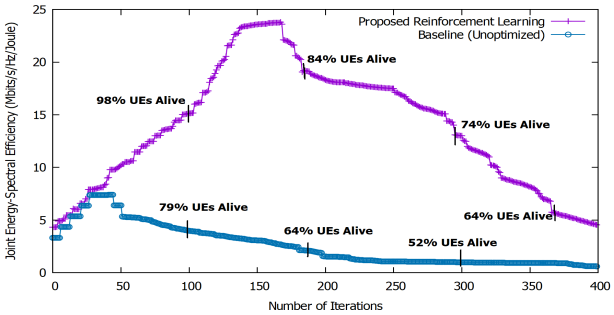


Fig. 5. Stand-alone spectral efficiency performance evaluation and unoptimized energy efficiency. Spectral efficiency optimized by means of the proposed method remains maximized for a longer time as compared to the baseline technique; it decreases after approximately 170 iterations due to the increasing number of depleted UEs.



(a) Performance evaluation of joint ESE.



(b) Percentage of alive UEs using the proposed method and baseline (unoptimized).

Fig. 6. Performance evaluation of joint ESE and percentage of alive UEs. (a) Performance evaluation of joint ESE. (b) Percentage of alive UEs using the proposed method and baseline (unoptimized).

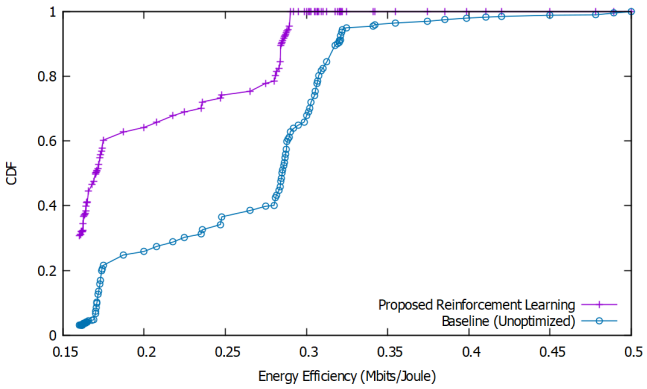


Fig. 7. Cumulative distribution function (CDF) of energy efficiency of the connected UEs.

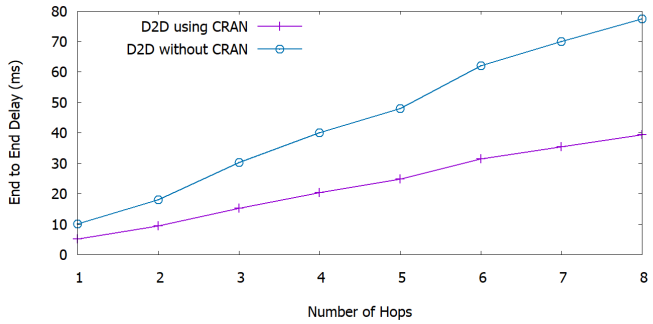


Fig. 8. End-to-end delay of the network over number of hops.

Moreover, our proposed approach outperforms all others in terms of the percentage of alive UEs. Fig. 6 (b) shows that for the joint ESE, optimized by means of our proposed approach, 98% of UEs are alive after 125 iterations whereas only 79% are alive at the same number of iterations without optimization.

Fig. 7 shows the cumulative distribution function (CDF) for the energy efficiency of the connected UEs. We adopt the Monte-Carlo method to calculate the CDF to obtain the numerical results. Here, we observe that our proposed method for energy efficiency outperforms the baseline protocol/unoptimized one by almost up to 30%.

D. Delay and Execution Time

Fig. 8 shows the end-to-end delay of the network as a function of the number of hops. Here, end-to-end delay is estimated as the time between the availability of a packet at the transmitter and at the receiver in ms. We observe that, when increasing the number of hops, the end-to-end delay increases linearly in both frameworks, *i.e.* D2D using C-RAN and D2D without C-RAN. We compare our framework with the D2D without C-RAN framework and we observe that the end-to-end delay decreases in our proposed framework. The reason for this improvement is that the data traffic passes through BBUs that have higher processing power and can deliver the data to the UEs more rapidly, while D2D communication without C-RAN functionality is characterized by higher latency. Our proposed framework yields 50% reduced latency compared to the traditional infrastructure. This result highlights the contribution of our proposed framework in delay-sensitive disaster scenario, *e.g.* a zone subject to a terrorist attack.

TABLE IV
EXECUTION TIMES FOR THE METHODS.

Method	*Execution Time*
Proposed method	2.13 s
Q(λ)	2.50 s
SARSA	2.85 s
SARSA(λ)	2.90 s

Finally Table IV shows the execution times for the different methods: we can observe that our proposed method outperforms the others in terms of execution time. We use the hardware configuration of Intel(R) Core (TM) i7-4790 3.60 GHz CPU, 8 GB RAM and in a 64-bit Windows 10 operating system. In addition, Simulator::Schedule is used in NS-3 for calculating the execution time. The execution time is the average of five complete simulation runs consists of 400 iterations each.

VII. CONCLUSIONS

In a disaster scenario where the network coverage is not fully ensured, our proposed method for D2D communication underlaying C-RAN helps to provide improved joint ESE as compared to the baseline approach. Simulation results show that our optimized network outperforms the baseline (unoptimized) one by almost 67% in terms of joint energy-spectral efficiency and our algorithm helps to achieve approx. 30% energy efficiency for connected UEs compared to the baseline technique. Moreover, using C-RAN infrastructure, our adaptive algorithm helps to reduce the latency by 50%. We also evaluated our proposed methods against other variants of reinforcement learning and showed that our proposed method outperforms them in terms of both stand-alone and joint ESE.

In the future, we will explore other adaptive optimization methods, i.e., mixed integer optimization with adaptive partition and multi-directional search in the network. Our intention with such methods is to optimize the parameters, i.e., energy efficiency, spectral efficiency, latency, outage probability etc. in a way that some parameters can be relaxed and we can focus on the most critical parameters depending on the requirements of the environment. For example, when the network is almost going down then we need to focus on energy efficiency and we can relax delay to keep the network alive as long as possible. Moreover, considering temporal death [37] [38] of the UEs, varying the number of UAVs and perform some experiments will add further realism to our scenario.

ACKNOWLEDGMENT

This research was supported by the Estonian Research Council through the Institutional Research Project IUT19-11, and by the Horizon 2020 ERA-chair Grant "Cognitive Electronics COEL" H2020-WIDESPREAD-2014-2 (Agreement number: 668995; project TTU code VFP15051).

REFERENCES

[1] A. Gohil, H. Modi, and S. K. Patel, "5g technology of mobile communication: A survey," in *Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on*. IEEE, 2013, pp. 288–292.

[2] R. I. Ansari, C. Chrysostomou, S. A. Hassan, M. Guizani, S. Mumtaz, J. Rodriguez, and J. J. Rodrigues, "5g D2D networks: Techniques, challenges, and future prospects," *IEEE Systems Journal*, 2017.

[3] H. Holma, A. Toskala, and J. Reunanen, *LTE Small Cell Optimization: 3GPP Evolution to Release 13*. John Wiley & Sons, 2016.

[4] X. Lin, J. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3gpp device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, 2014.

[5] A. Alnoman and A. Anpalagan, "On D2D communications for public safety applications," in *Humanitarian Technology Conference (IHTC), 2017 IEEE Canada International*. IEEE, 2017, pp. 124–127.

[6] M. I. Khan, M. M. Alam, Y. Le Moullec, and E. Yaacoub, "Throughput-aware cooperative reinforcement learning for adaptive resource allocation in device-to-device communication," *Future Internet*, vol. 9, no. 4, p. 72, 2017.

[7] ——, "Cooperative reinforcement learning for adaptive power allocation in device-to-device communication," in *Proceedings of the IoT World Forum*. IEEE, 2018, pp. 476–481.

[8] S. Nie, Z. Fan, M. Zhao, X. Gu, and L. Zhang, "Q-learning based power control algorithm for D2D communication," in *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2016 IEEE 27th Annual International Symposium on*. IEEE, 2016, pp. 1–6.

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

[10] S. Lhazmir, A. Kobbane, and J. Ben-Othman, "Channel assignment for D2D communication: a regret matching based approach," in *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE, 2018, pp. 322–327.

[11] M. Bennis and D. Niyato, "A q-learning based approach to interference avoidance in self-organized femtocell networks," in *2010 IEEE Globecom Workshops*. IEEE, 2010, pp. 706–710.

[12] W. Chen and J. Zheng, "A multi-agent reinforcement learning based power control algorithm for d2d communication underlaying cellular networks," in *International Conference on Artificial Intelligence for Communications and Networks*. Springer, 2019, pp. 77–90.

[13] L. Babun, A. I. Yürekli, and I. Güvenç, "Multi-hop and D2D communications for extending coverage in public safety scenarios," in *Local Computer Networks Conference Workshops (LCN Workshops), 2015 IEEE 40th*. IEEE, 2015, pp. 912–919.

[14] L. Babun, "Extended coverage for public safety and critical communications using multi-hop and D2D communications," 2015.

[15] K. Ali, H. Nguyen, Q. Vien, P. Shah, and Z. Chu, "Disaster management using D2D communication with power transfer and clustering techniques," *IEEE Access*, 2018.

[16] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both?" *arXiv preprint arXiv:1902.02647*, 2019.

[17] P. Rawat, M. Haddad, and E. Altman, "Towards efficient disaster management: 5g and device to device communication," in *Information and Communication Technologies for Disaster Management (ICT-DM), 2015 2nd International Conference on*. IEEE, 2015, pp. 79–87.

[18] S. Krishnamoorthy and A. Agrawala, "M-urgency: a next generation, context-aware public safety application." in *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 2011, pp. 647–652.

[19] S. Boumard, I. Harjula, T. Kanstren, and S. J. Rantala, "Comparison of spectral and energy efficiency metrics using measurements in a LTE-A network," in *2018 Network Traffic Measurement and Analysis Conference (TMA)*, 06 2018, pp. 1–8.

[20] G. Zhao, S. Chen, L. Qi, L. Zhao, and L. Hanzo, "Mobile-traffic-aware offloading for energy- and spectral-efficient large-scale D2D-enabled cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3251–3264, June 2019.

[21] K. Zia, N. Javed, M. N. Sial, S. Ahmed, H. Iram, and A. A. Pirzada, "A survey of conventional and artificial intelligence / learning based resource allocation and interference mitigation schemes in D2D enabled networks," *CoRR*, vol. abs/1809.08748, 2018. [Online]. Available: http://arxiv.org/abs/1809.08748

[22] 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.

[23] I. AlQerm and B. Shihada, "Enhanced machine learning scheme for energy efficient resource allocation in 5g heterogeneous cloud radio access networks," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2017, pp. 1–7.

[24] Z. Zhou, M. Dong, K. Ota, G. Wang, and L. T. Yang, "Energy-efficient resource allocation for D2D communications underlaying cloud-ran-based LTE-A networks," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 428–438, 2016.

[25] Y. Zhang, J. Zhang, Y. Sun, and D. W. K. Ng, "Energy-efficient transmission for wireless powered D2D communication networks," in *Communications (ICC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–7.

[26] S. Koenig and R. G. Simmons, "Complexity analysis of real-time reinforcement learning applied to finding shortest paths in deterministic domains," CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, Tech. Rep., 1992.

[27] M. Boushaba, A. Hafid, A. Belbekkouche, and M. Gendreau, "Reinforcement learning based routing in wireless mesh networks," *Wireless networks*, vol. 19, no. 8, pp. 2079–2091, 2013.

[28] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[29] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[30] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 2, pp. 201–212, 2012.

[31] M. Moradi, "A centralized reinforcement learning method for multi-agent job scheduling in grid," in *Computer and Knowledge Engineering (ICCKE), 2016 6th International Conference on*. IEEE, 2016, pp. 171–176.

[32] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[33] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[34] J. Leng, C. Fyfe, and L. C. Jain, "Experimental analysis on sarsa (λ) and q (λ) with different eligibility traces strategies," *Journal of Intelligent & Fuzzy Systems*, vol. 20, no. 1, 2, pp. 73–82, 2009.

[35] R. Rouil, F. J. Cintrón, A. Ben Mosbah, and S. Gamboa, "Implementation and validation of an LTE D2D model for ns-3," in *Proceedings of the Workshop on ns-3*. ACM, 2017, pp. 55–62.

[36] Y. He, X. Luan, J. Wang, M. Feng, and J. Wu, "Power allocation for D2D communications in heterogeneous networks," in *Advanced Communication Technology (ICACT), 2014 16th International Conference on*. IEEE, 2014, pp. 1041–1044.

[37] L. Tan and S. Tang, "Energy harvesting wireless sensor node with temporal death: Novel models and analyses," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 896–909, 2017.

[38] S. Tang and L. Tan, "Reward rate maximization and optimal transmission policy of eh device with temporal death in eh-wsns." *IEEE Trans. Wireless Communications*, vol. 16, no. 2, pp. 1157–1167, 2017.