Power Control Based on Safe Q Learning for D2D Communication

Adil BOUMAALIF and Ouadoudi ZYTOUNE

***Abstract*—Nowadays, Device to Device (D2D) communication becomes a crucial technology in 5G wireless systems. It is intended to improve the system performance, enhance the user experience and offer a large variety of applications, that is why it attracts more attention. Motivated by the machine learning successful applications to many practical domains, researchers have proposed its application in wireless communication topics, especially in D2D communications. One of the scenarios of D2D communications is underlaid in-band, when each resource block is shared between the D2D and cellular users, thus co-channel interference is a challenging problem. To manage interference mitigation with power allocation technique, we propose, in this paper, a new power control algorithm based on Safe Q-learning algorithm. Our goal is to maximize, in the same time, the throughput of D2D users and the device lifetime metrics, while guaranteeing the required SINR for the cellular communications. It has been demonstrated that through our algorithm, D2D users equipments are able to learn their power in a self-organized manner, in addition to achieving better device lifetime metrics than that based on an enhanced Q-learning algorithm.**

***Index Terms*—D2D communications, Reinforcement learning, Safe Q-learning, Device lifetime optimization.**

I. Introduction

The rise of the number of mobile users has given the impulse to the demand for proximity services with high data rates. The 5th-generation (5G) networks pledge to introduce new technologies according to the future expected demands in order to provide effective and resource-efficient solutions. Device to device (D2D) communication has been proposed the first time in release 12 of 4G mobile networks, and it will play a significant role in 5G wireless systems, as it offers a variety of services with high data rate and low latency [1]. In cellular networks case, D2D communication is seen as direct communication between two proximate mobile equipments without the need of involving the Base Station (BS). Generally, D2D communications are non-transparent to the cellular infrastructure and it can take place on the cellular spectrum (that is, in-band) or unlicensed spectrum (that is, out-band). Hence, D2D communications can likely improve spectral efficiency, throughput, energy consumption efficiency, transmission delay, and fairness [2]. In an underlaid in-band setting, when the D2D users compete with the cellular users (CUs) to use the same resources, the major challenge is the existing of interference due to the aggressive frequency reuse. It is crucial to design an powerful interference management plan to mitigate the interference caused by the D2D links to the cellular links, and reciprocally [3]. The paper [4] presented achievable data rate utilities of UEs to make decision on receiving content via relay with or without reward given by the BS.

One of the solutions to mitigate the interference in the above model, is the power control, it is broadly used in current wireless systems [5]. For instance, researchers in [6] has elaborated a Green Hose-Rectangle Model to optimize power efficiency in communications networks for green computing. Authors in [7] has proposed an iterative distributed power control algorithm with the objective to allocates transmit powers which lead to minimized power consumption while meeting a sum-rate constraint. In [8], To enhance the energy efficiency (EE) of D2D communications, an effective iterative resource allocation and power control strategy is suggested. Also in [9], the authors of the paper use mathematical tool of stochastic geometry and give a channel allocation (CA) scheme jointly with a group of three power control (PC) schemes to reduce interference in a D2D underlaid cellular networks, in order to enhance D2D and cellular coverage chances, and improve spectral and power capability. In [10] considered the distribution of the D2D transmit power.

Another crucial factor to be considered is the battery lifetime of the mobile users. This subject was insensitively considered(reader can see for example references as [11], [12], [13]). As known the battery lifetime is not unlimited, as they experience degradation influenced by multiple things, including both manufacturing aspects and operating circumstances. For operating conditions, we can cite High discharging currents, impulse discharging currents, low or high operating temperature that short the battery lifetime [14]. For example, To prolong the network lifetime, authors in [15] proposed a scale-free topology based on the Node Lifetime in a Wireless Sensor Networks (WSN) environment. In our case, the decision of the power allocation mechanism should be made based on the battery residual energy information of each D2D device. For example, a node with low residual battery energy shouldn't transmit with higher transmit power even this latter can guarantee better performance in terms of throughput [16].

Improvements in artificial intelligence (AI) and machine learning (ML) give endless potentials in various science and engineering fields including computer communication networks. It is defined as the system's capability to acquire and integrate knowledge based on exhaustive observations, and to ameliorate itself by learning new knowledge instead of being programmed with that knowledge [17]. We have four main categories of machine learning methods: Supervised technique, Unsupervised technique, Semi-supervised

Manuscript received January 15, 2022; revised December 30, 2022.

A. BOUMAALIF is a PhD graduate in computer sciences from ASElab, ENSA, Ibn Tofail University, Kenitra, Morocco (email: adil.boumaalif@gmail.com)

O. ZYTOUNE is a Professor of Electrical Engineering and member of ASElab, ENSA, Ibn Tofail University, Kenitra, Morocco (email: zytoune.ouadoudi@uit.ac.ma).

and Reinforcement learning ones [18].

- Supervised learning uses tagged training data and a set of training cases to conclude a pattern that maps an input data to an output. Classification that separates the data, and regression that permits to fit the data are the most common supervised tasks.
- Unsupervised learning explores unlabeled data sets. Clustering, association rules, density estimation, dimensionality reduction, anomaly detection are some applications of unsupervised learning tasks.
- Semi-supervised learning mixes supervised and unsupervised methods. It operates on labeled as well as unlabeled data.
- Reinforcement learning enables agents to automatically learn and evaluate the optimal behavior in a specific environment to enhance its efficiency, it is considered as an environment-driven approach.

Reinforcement Learning (RL), has been considered as a powerful tool in solving resource allocation problems in 5G. Reader can refer to recent works as [19] and [20] to explore some utilization of RL in optimisation problems. It is also used as a power allocation technique to minimize the overall network interference [21]. Q-learning is a basic RL algorithm, and it has many variants which they are used as an effective way for D2D power control. In [22], authors propose two RL algorithms, team-Q learning and distributed-Q learning, as a power control techniques, in order to maximize the overall system capacity while guaranteeing the requirement of quality of service(QoS) from CUs. Xu, in [23], has proposed a Hierarchical Extreme Learning Machine (H-ELM) algorithm for the D2D power allocation, and he has proved with simulation, and in comparison with other RL algorithms, i.e. distributed Q-learning and CART Decision Tree, that the proposed algorithm provides better performance in communication throughput as well as in energy efficiency with limited time consumption. In [24], researchers has combined neural networks with Q-learning, and proposed a new algorithm, Multi-Agent Deep Q algorithm, which showed higher performance in comparison with other traditional power control algorithms. In [25], a battery lifetime aware resource allocation framework in cellular-based M2M networks was proposed. This framework provides substantial network lifetime enhancement and network maintenance cost reduction in comparison with literature solutions.

In this paper, a D2D power control algorithm based on Safe Q-learning is proposed to achieve better performance in term of both throughput and device life-time, while maintaining lower interference applied to CUs. The rest of this paper is organized as follows. Section II describes the system model and the formulation of the problem. Section III introduces the Safe Q-learning algorithm. Section IV describes Safe Q-learning algorithm for D2D power control. Section V presents numerical and simulation results. Section VI concludes this paper.

II. System Model and Problem formulation

We consider M cellular users and N D2D pairs distributed uniformly at random within the coverage area of BS in a single cell. Cellular and D2D users share the same number of available resource blocks (RBs) for their uplink (UL) transmission, which is denoted by K. We assume that each RB is taken by one cellular user and can be shared with N_k D2D pairs, where $N_k \leq N$. In this scenario, we have two kinds of interference, the first one is applied to the BS from the D2D transmitters, the second is applied to the D2D receiver from the cellular user and other D2D transmitters who share the same RB. So the signal to interference plus noise ratio (SINR) of the ith D2D user on the kth RB is expressed as follows:

$$\gamma_{i,k}^{d} = \frac{p_{i,k}^{d}.G_{i,k}^{d}}{\sigma^2 + p_k^c.G_{i,k}^c + \sum\limits_{j=1,j\neq i}^{N_k} p_{j,k}^{d}.G_{j,k}^{d}} \tag{1}$$

Where $p_{i,k}^{d}$, p_k^c and $p_{j,k}^{d}$ denote the transmit power of the ith D2D transmitter, the cellular user and the other D2D transmitters sharing the same kth RB, respectively. $G_{i,k}^{d}$, $G_{i,k}^{c}$ and $G_{j,k}^{d}$ represent, respectively, the channel gain in the ith D2D link, the channel gain between cellular transmitter and the ith D2D receiver and the channel gain between one of the other D2D transmitters (the jth one), sharing the same kth RB, and the ith D2D receiver. σ^2 is the noise power. For G, it can generally expressed as follows [24]:

$$G = 10^{(-PathLoss-Shadowing)/10}$$

Likewise, the SINR of the cellular user in the kth RB, is given by:

$$\gamma_k^c = \frac{p_k^c.G_{0,k}^c}{\sigma^2 + \sum\limits_{j=1}^{N_k} p_{j,k}^{d}.G_{j,k}^{d}} \tag{2}$$

With $G_{0,k}^{c}$ is the channel gain between the BS and the cellular user.

We define the network lifetime, or its service duration, as the time duration from the starting reference time till the moment when the network is considered to be non-functional. However, the network is considered to be non-operational is application dependent. In safety-critical applications, for example, where the death even one node worsens the performance or coverage. Also, in sensor deployments with low densities, where correlation between the reading of different nodes is weak, the shortest individual lifetime (SIL) may designate the network lifetime. However in situations, e.g. where high correlation exists between data collected by different nodes, the largest individual lifetime (LIL) or the average of individual nodes lifetime (AIL) may be used as the network lifetime. In this work, we consider the first context, i.e., the smallest individual lifetime is used as the network lifetime [25], that is: $DL_{net} = min_i DL_i$

$$DL_i = \frac{E_{i0}}{\mathbb{E}[p_i^d]} \tag{3}$$

Where E_{i0} is the initial energy of i^{th} D2D transmitter, and $\mathbb{E}[p_i^d]$ is its transmit power expectation. Assuming that, for simplicity, each message transmission occurs in unity of time.

In this study, we assume that the objective of our D2D power control algorithm is to maximize, in the same time: (i)

the overall throughput of the D2D communications and (ii) the device lifetime metrics, while guaranteeing the minimum SINR for the cellular network. To achieve this objective, we have to solve the following optimization problem:

$$\begin{aligned} \max_{p_K} \quad & \sum_{k=1}^{K}(\sum_{i=1}^{N_k} log_2(1+\gamma_{i,k}^d) + C.DL_{net}) \\ s.t. \quad & \gamma_k^c \geq \tau \\ & 0 \leq p_{i,k}^d \leq p_{max}, \forall i,k \end{aligned} \tag{4}$$

Where $p_K = (p_{1,k}^d, ..., p_{i,k}^d, ..., p_{N,k}^d)$ is the vector of D2D transmit powers. C is an arbitrary positive constant which maximize our multi-objective optimization problem. The value of C is determined based on a specific choice of the relative weight of the objectives, i.e., in our model, is it more important for the throughput to be close to the maximal value than for network life time, or the inverse?

We can see obviously that when the transmit power of the D2D users increases, the D2D throughout will increase automatically, while the D2D lifetime will decrease and the cellular communications will experience more interference. On the other hand, to ensure a minimum QoS of cellular users, D2D transmit powers should be limited. To find the optimal D2D transmit power, a Safe Q-learning based power control algorithm will be introduced.

III. Safe Q-Learning

The idea of reinforcement learning is simply an agent in an environment, performing actions based on its observations of this environment. Generally, the agent chooses actions depending on a policy. The agent receives immediate rewards from the environment, which signalizes how well the agent behave. The goal of the agent is to maximize its cumulative reward by observing its environment and the reward information received, and then executing the optimal actions. The Q-Learning algorithm is an off-policy control algorithm, meaning that it does not depend on the policy the agent uses to explore the environment. It is defined by the following update equation:

$$Q_{t+1}(S_i, a_j) \leftarrow Q_t(S_i, a_j) + \alpha[R_t + \gamma \max_{a_j'} Q_t(S', a_j') - Q_t(S_i, a_j)] \tag{5}$$

Where α is the learning rate and γ is the discount factor of the Q-Learning algorithm.

Without the need of any policy being followed and using such update scheme for action-value pairs, Q-learning can reach the best approximation of the optimal action-value function[26]. However, as stated in [22], the distributed Q-learning gives benefits than classic one. It enables agents to learn independently and then reduces the complexity of Q-value table. Its principle is to split the large Q-value table to multiple small ones. Thus, multiple Q-value tables are maintained. The Q-values in each Q-table will be updated only when the next Q-value is greater than current Q-value. In distributed Q-learning, the missions of learning optimal action policy are decentralized to each agent in team, that is, there is no central control mechanism. So, the state-action space becomes smaller even the number of agents increases. Consequently, the convergence time of the algorithm becomes more speed [31].

The Q-values are updated as follows (6):

One of the major drawbacks of Q-learning (either in the classic or in its deep version) is regarding to the method of selecting actions from quite uncertain parts of state space, which happens often after occurrence of an severe event. Let consider, For example, a system that has been in a stable state for a large time, that part of the Q-table is well-known, however, the other parts can be highly uncertain. If an extraordinary event occurs, it forces the system to those areas, so in this case, how can transform the random exploration of the learning into a safe one [27].

In order to solve this problematic, Safe Q-learning is driven from the constrained Markov decision processes (CMDP). This variant of Q-Learning algorithms adds a restriction function to the classic objective function so that it splits the action space into feasible action space that fulfills the constraint condition and infeasible action space that does not fit the constraint term. Under these conditions, the agent only executes feasible actions and goes into safe states so as to abstain from unnecessary damage caused by incorrectly executing the infeasible action [28]. The Fig.1 gives a simple illustration of Safe Q-learning model.

As a result, Safe-RL can maximize the expectation value of reward as well as guaranteeing sufficient performance and dealing with safe constraints. we can distinguish two kins of Safe-RL, the first one is based on the amendment of the optimality criterion, the standard discounted finite/infinite horizon, with a safety aspect. The principle of the second one is the modification of the exploration process via the integration of external knowledge or the guidance of a risk metric [29].

We can describe the algorithm as a 5-tuple $< S, A, T, R, C >$ where S represents the set of states, A represents the set of actions, $T(s, a, s') = Pr(s'|s, a)$ is a transition model that captures the probability of passing to state s' by executing action a at state s, $R(s, a, s')$ is an immediate reward got when executing action a at state s. C is the set of constraint functions which represents that the action space is constrained. A policy $\pi : S \rightarrow A$ is the mapping function from states to actions [30].

IV. Safe Q-learning Power control algorithm

Consider learning on the kth RB, we focus on N_k D2D transmitters who share this RB as the **Agents**. Our objective is that the agent learn its transmit power from the feasible actions taking into account the predefined constraint. In the above scenario, states, actions, reward, constraint functions are defined as follows:

State: We specify the state as:

$$S_t^{i,k} = I_t^k$$

Where i indicates the D2D user, k is the considered RB, at time t, and I_t^k represents whether the level of interference applied the cellular communications is acceptable or not, i.e. The cellular SINR is above the minimum threshold τ:

$$I_t^k = \begin{cases} 1 & if \ \gamma_k^c \geq \tau \\ 0 & otherwise \end{cases}$$

We suppose that the D2D user gets the actual value of SINR from the BS.

$$Q_{t+1}(S_t, a_t) = \begin{cases} max\{Q_t(S_t, a_t); R_{t+1} + \gamma \max_{a'} Q_t(S_{t+1}, a'_{t+1})\} & \text{if } S = S_t \text{ and } a = a_t \\ Q_t(S_t, a_t) & \text{otherwise,} \end{cases} \quad (6)$$

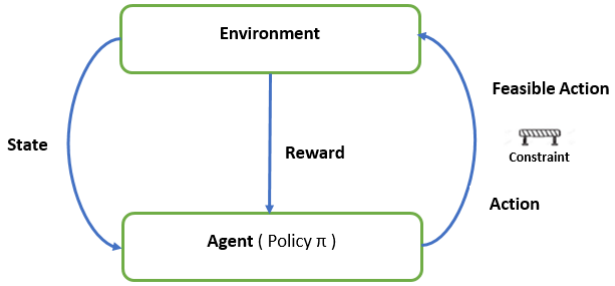

Fig. 1. The illustration of Safe Q-learning

Action: The action of each agent is composed of a set of transmitting power levels. We denote actions by the set:

$$A = (a_1^k, a_2^k, ..., a_l^k)$$

where l means that every agent has l power levels.

In this article we utilize the $\epsilon - greedy$ strategy to select actions based on the actual $Q-$value estimation, which is described as the following:

• choose action randomly with probability ϵ from the feasible action space,

• choose action according to $a = arg\ max_{a \in A} Q(s, a)$ with probability $1 - \epsilon$

Reward: The reward function reflects the learning objectives of RL.so we define the reward as

$$R = \begin{cases} log_2(1 + \gamma_{i,k}^d) & if\ \gamma_k^c \geq \tau \\ -1 & otherwise \end{cases}$$

Constraint: The constraint function represents that the action space (power levels) is constrained to a reduced set of actions considering the device battery level. i.e $A_c = (a_1^k, ..., a_m^k)$ where $m \leq l$

The power control strategy is illustrated in the following algorithm.1. in the beginning, we associate for each D2D transmitter a set of feasible transmit power levels based on its initial energy reserve. Then, in the learning process, we use $\epsilon - greedy$ algorithm to select the next action to perform. This selection is done in the set of feasible actions related the considered D2D transmitter. The selected action is then executed, and the related reward is earned. Thus, the Q-table is updated. This process is iterated till the maximal number of iteration is achieved.

Algorithm 1 Q learning algorithm

1. Initialization:
for each D2D transmitter i **do**
  Define the feasible action set $\mathcal{A}_c^i$ based on the constraint
end for
for each state $S_i \in S$ and each action $a_j \in \mathcal{A}_c^i$ **do**
  Initialize $Q(S_i, a_j)$ arbitrarily
end for
evaluate the starting state $S_i \in S$
2. Learning:
while MaxIteration not reached **do**
  choose $a_j \in \mathcal{A}_c$ using the ϵ-greedy policy based on Q
  Take action a_j and observe the immediate reward R_t and next state S'
  $Q_{t+1}(S_i, a_j) \leftarrow max\{Q_t(S_i, a_j); R_{t+1} + \gamma \max_{a'_j} Q_t(S', a'_j)\}$
  $S_i \leftarrow S'$
end while

V. Numerical and Simulation Results

A. Simulation Parameters

In the simulation, we compare our proposed algorithm with the distributed Q-learning algorithm for a multi-agent scenario. We consider a single macrocell with a covering radius of $500m$. The spectrum is shared to 20 RBs, where multiple cellular and D2D users coexist. Cellular and D2D users are uniformly distributed on the cell that a Base station is located in the center. The distance between a D2D pair (Transmitter and receiver) have a random value in the range $[5, 50]m$. Data packets are assumed to have 180000 bits length. The Q-learning parameters are as follows: the learning rate is $\alpha = 0.5$, the discount factor is $\gamma = 0.7$, the ϵ greedy parameter $\epsilon = 0.2$. The rest of simulation parameters used in this work are summarized in Table I. We define the action space as a vector of discrete values of transmit power levels, i.e. $A = 2 : 25$ with step of $1dBm$. We distinguish six battery level ranges. Thus, when the battery level is under 17% the action space is minimized to $A_1 = 2 : 5$, and when the battery level is between 17% and 33% the action space becomes $A_2 = 2 : 9$, and so on. The D2D battery level is uniformly distributed. Thus, the $i-th$ D2D transmitter initial energy is: $E_i \in \{5, 9, 13, 17, 21, 25\} \times 100j$.

TABLE I
Simulation parameters

Parameter	Value
p_{max}, maximal transmission power	$23dBm$
Noise Power per RB	$-116dBm$
D2D pair distance	$50m$
Pathloss model between BS and users	$15.3 + 37.6log_{10}(d(km))(dB)$
Pathloss model between users	$28 + 40log_{10}(d(km))(dB)$
Macro BS antenna gain	$17dBi$
User antenna gain	$4dBi$
resource block bandwidth	$180kHz$

B. Results and Discussion

In Fig.2, we plot the D2D transmitter lifetime as a function of D2D pairs number. The minimum lifetime is defined as the number of transmissions before the first D2D transmitter runs out its total energy. As we depicted in this Fig. we can observe that the proposed algorithm shows an enhancement of the performance compared to the distributed Q-learning in terms of minimum lifetime when the number of D2D pairs is in $\{5, 10, 15\}$.

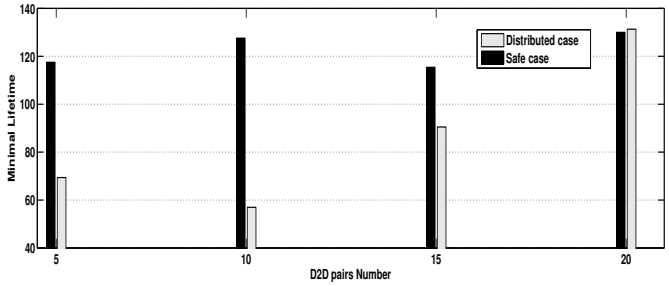

Fig. 2. Devices Lifetime vs D2D pairs, with $\tau = 6dB$

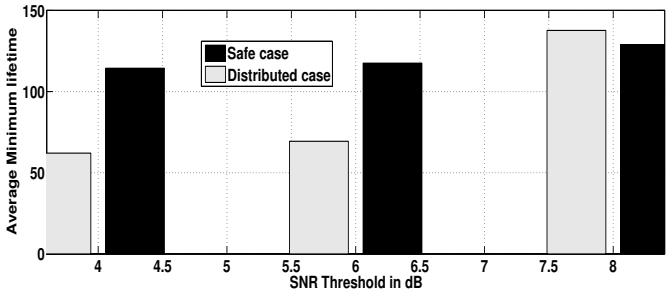

Fig. 3. Device Lifetime vs SNR threshold for 5 D2D pairs number

In Fig.3, the D2D transmitter lifetime is given with regards to the minimum cellular $SINR$ threshold. It is shown that average minimum device lifetime presents remarkable amelioration in our algorithm in comparison with distributed Q-learning where the $SINR$ is less than $8dB$.

We plot, in Fig. 4, the network average throughput for the D2D transmitters. As we can see, Our algorithm gives better D2D transmitters throughput than distributed Q-learning one for D2D pairs numbers between 5 and 20. These performances are important when the number of D2D transmitters-receivers is relatively low.

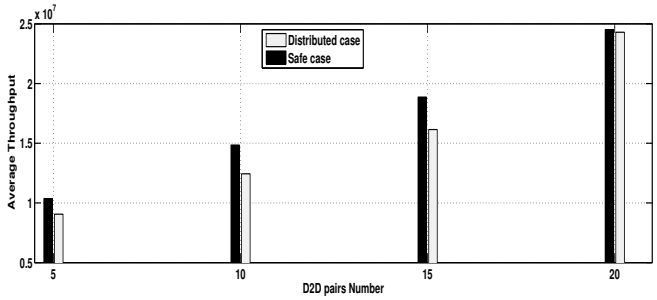

Fig. 4. Network average throughput for $\tau = 6dB$

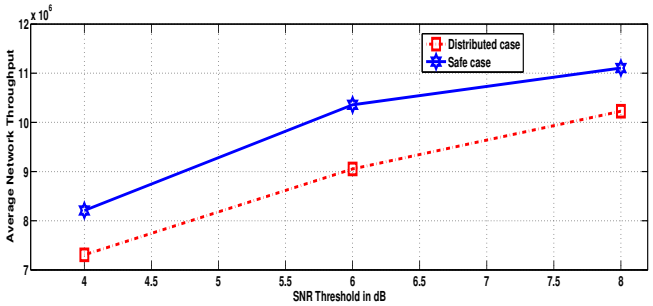

Fig. 5. Network average throughput for 5 D2D pairs number

TABLE II
SIMULATION RESULTS FOR $\epsilon = 0.1$ FOR MULTIPLE VALUES OF D2D PAIRS NUMBER

D2D pairs number	Safe Average Minimum lifetime	Distributed Average Minimum lifetime	Safe Throughput (packets)	Distributed Throughput (packets)
5	105	36	63	63
10	100	90	97	90
15	100	55	115	113
20	100	91	120	117

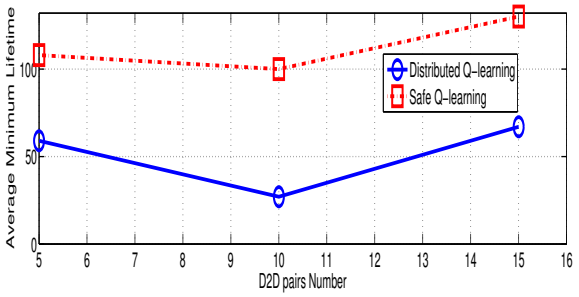

Fig. 6. Devices Lifetime vs D2D pairs

Fig. 5 gives the network throughput for 5 D2D pairs when the SNR threshold varies from $4dB$ to $8dB$. As depicted, we can observe that the proposed algorithm out performs the distributed q-learning for different $SINR$ values. As depicted in this figure, we can remark that our proposition gives an extension of the throughput by up to 14% as compared to the distributed Q-learning.

Table II gives the obtained results for learning with ϵ-greedy parameter equals to 0.1. Our algorithm always gives best performance in terms of lifetime and Throughput with regards to Distributed Q-learning.

In the last assessment, we consider a sever path loss channel model. In this situation, the channel path loss used is as follows: $28 + 50log_{10}(d(km))(dB)$. The parameter ϵ is kept as 0.1. The obtained results are given in Fig. 6 and Fig. 7. As we can observe, the network lifetime is well performed in the Safe q-learning than Distributed q-learning. We recall that with sever path loss channel, more energy is required to achieve transmission between device pairs.

VI. CONCLUSION

In this work, we presented a Q-learning algorithm with safety constraint to control the transmit power of D2D

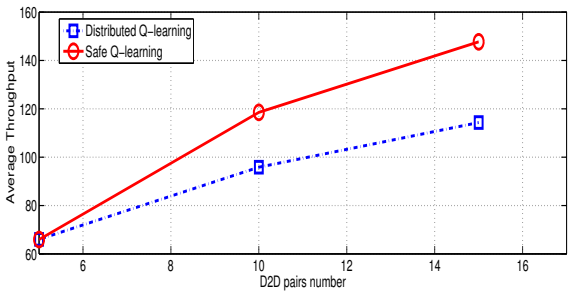

Fig. 7. Network average throughput vs D2D pairs number ×180000 bits

transmitters in order to maximize the overall throughput of D2D communications and extend the device lifetime, while ensuring the minimum SINR for the cellular communications. In our case, we defined the constraint as the battery energy level of D2D transmitter. So, devices with reduced battery energy are constrained to use limited transmit powers. Based on numerical assessment, we proved that our proposition permits a lifetime extension of D2D transmitters as compared to the distributed Q-learning algorithm. Hence, simulation shows an important improvement of the minimum device lifetime, which corresponds to the depletion of the battery of the first device.

References

[1] R. I. Ansari et al., "5G D2D Networks: Techniques, Challenges, and Future Prospects," in IEEE Systems Journal, **vol. 12**, no. 4, pp. 3970-3984, (2018).

[2] A. Asadi, Q. Wang and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks" in IEEE Communications Surveys & Tutorials, **vol. 16**, issue. 4, pp. 1801-1819, (2014).

[3] J. Yang, M. Ding, G. Mao and Z. Lin, "Interference Management in In-Band D2D Underlaid Cellular Networks," in IEEE Transactions on Cognitive Communications and Networking, **vol. 5**, no. 4, pp. 873-885, (2019).

[4] Bista, Bhed Bahadur. "Promoting Relay-Assisted Device-to-Device Communication in Cellular Networks By Reward Mechanisms." IAENG International Journal of Computer Science, **vol**. 47, no 3, pp. 431-435, (2020).

[5] N. Lee, X. Lin, J. G. Andrews and R. W. Heath, "Power Control for D2D Underlaid Cellular Networks: Modeling, Algorithms, and Analysis," in IEEE Journal on Selected Areas in Communications, **vol. 33**, pp. 1-13, (2015).

[6] Das, Bimal Chandra, et al. "Green Hose-Rectangle Model Approach for Power Efficient Communication Networks." IAENG International Journal of Computer Science, **vol**. 48, no.3, pp. 760-769, (2021).

[7] G. Fodor and N. Reider, "A Distributed Power Control Scheme for Cellular Network Assisted D2D Communications," 2011 IEEE Global Telecommunications Conference - GLOBECOM 2011, pp. 1-6, (2011).

[8] Y. Jiang, Q. Liu, F. Zheng, X. Gao and X. You, "Energy-Efficient Joint Resource Allocation and Power Control for D2D Communications," in IEEE Transactions on Vehicular Technology, vol. 65, no. 8, pp. 6119-6127, Aug. 2016.

[9] A. Abdallah, M. M. Mansour and A. Chehab, "Power Control and Channel Allocation for D2D Underlaid Cellular Networks," in IEEE Transactions on Communications, vol. 66, no. 7, pp. 3217-3234, July 2018.

[10] A. Boumaalif and O. Zytoune, "Power Distribution of Device-to-Device Communications Under Nakagami Fading Channel," in IEEE Transactions on Mobile Computing, vol. 21, no. 6, pp. 2158-2167, 1 June 2022, doi: 10.1109/TMC.2020.3035543.

[11] Ouadoudi Zytoune, Hacene Fouchal, Sherali Zeadally, A realistic relay selection scheme for cooperative MIMO networks, Ad Hoc Networks, Volume 124, 2022, 102706, ISSN 1570-8705, https://doi.org/10.1016/j.adhoc.2021.102706.

[12] Zytoune, O., Aboutajdine, D. Energy usage analysis of digital modulations in wireless sensor networks with realistic battery model. Wireless Netw 22, 2713-2725 (2016). https://doi.org/10.1007/s11276-015-1115-9.

[13] Zytoune, O., Fakhri, Y. and Aboutajdine, D. (2010), "A fairly balanced clustering algorithm for routing in wireless sensor networks", Sensor Review, Vol. 30 No. 3, pp. 242-249. https://doi.org/10.1108/02602281011051434.

[14] S. Jin, X. Huang, X. Sui, S. Wang, R. Teodorescu and D. -I. Stroe, "Overview of Methods for Battery Lifetime Extension," 2021 23rd European Conference on Power Electronics and Applications (EPE'21 ECCE Europe), pp. 1-8, (2021).

[15] Yin, Wenxiao, et al. "The Research on WSNs Scale-free Topology for Prolonging Network Lifetime." Engineering Letters, vol. 29, no.1, pp. 238-243, (2021).

[16] S. Gupta and J. Chakareski, "Lifetime Maximization in Mobile Edge Computing Networks," in IEEE Transactions on Vehicular Technology, **vol. 69**, no. 3, pp. 3310-3321, (2020).

[17] Nur Zincir-Heywood; Marco Mellia; Yixin Diao, "Overview of Artificial Intelligence and Machine Learning," in Communication Networks and Service Management in the Era of Artificial Intelligence and Machine Learning , IEEE, pp.19-32, (2021).

[18] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

[19] Hukmani, Kavish, Sucheta Kolekar, and Sreekumar Vobugari. "Solving Twisty Puzzles Using Parallel Q-learning." Engineering Letters, vol. 29, no.4, pp. 1535-1543, (2021).

[20] Xuan, Hejun, et al. "VNF Service Chain Deployment Algorithm in 5G Communication based on Reinforcement Learning." IAENG International Journal of Computer Science, vol. 48, no.1, pp. 1-7, (2021).

[21] K. Zia, N. Javed et al., "A Distributed Multi-Agent RL-Based Autonomous Spectrum Allocation Scheme in D2D Enabled Multi-Tier HetNets," in IEEE Access, **vol. 7**, pp. 6733-6745, (2019).

[22] S. Nie, Z. Fan, M. Zhao, X. Gu and L. Zhang, "Q-learning based power control algorithm for D2D communication," IEEE 27th International Symposium on PIMRC, (2016).

[23] J. Xu, X. Gu and Z. Fan, "D2D Power Control Based on Hierarchical Extreme Learning Machine," 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 1-7, (2018).

[24] S. Gengtian et al., "Power Control Based on Multi-Agent Deep Q Network for D2D Communication," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 257-261, (2020).

[25] A. Azari and G. Miao, "Network Lifetime Maximization for Cellular-Based M2M Networks," in IEEE Access, **vol. 5**, pp. 18927-18940, (2017).

[26] J. Jagannath, N. Polosky, A. Jagannath, F. Restuccia, T. Melodia, "Machine learning for wireless communications in the Internet of Things: A comprehensive survey," Ad Hoc Networks, Elsevier, **Volume 93**, (2019).

[27] M. Memarzadeh, M. Pozzi, "Model-free reinforcement learning with model-based safe exploration: Optimizing adaptive recovery process of infrastructure systems," Structural Safety, Elsevier, **Vol. 80**, pp. 46-55, (2019).

[28] Y. Ge, F. Zhu, X. Ling and Q. Liu, "Safe Q-Learning Method Based on Constrained Markov Decision Processes," in IEEE Access, **vol. 7**, pp. 165007-165017, (2019).

[29] T. Umemoto, T. Matsui, A. Mutoh, K. Moriyama and N. Inuzuka, "Safe Reinforcement Learning in Continuous State Spaces," 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), pp. 402-406, (2019).

[30] D. Chen, L. Jiang et al., "Autonomous Driving using Safe Reinforcement Learning by Incorporating a Regret-based Human Lane-Changing Decision Model," 2020 American Control Conference (ACC), pp. 4355-4361, (2020).

[31] J. Huang, B. Yang and D. Liu, "A Distributed Q-Learning Algorithm for Multi-Agent Team Coordination," 2005 International Conference on Machine Learning and Cybernetics, pp. 108-113, (2005).

Adil BOUMAALIF graduated as a Telecommunications engineer from INPT institute - Rabat, Morocco in 2010. This year, he holds a PhD degree in computer sciences from ENSA, Ibn Tofail University, Kenitra, Morocco. His research interests include Wireless Communications in next generations, Machine to Machine communications, IoT and Machine learning applications.

Ouadoudi ZYTOUNE received the Ph.D. degree in Computer Sciences and Telecommunications from the Mohammed V University in Rabat, Morocco in 2010. He is currently a Professor at the National School of Applied Sciences, Ibn Tofail University in Kenitra, Morocco. His research interests include Wireless Communications and networking, QoS in Wireless communication and Wireless Sensor Networks.