I. J. Computer Network and Information Security, **2023, 4, 13-24**
Published Online on August 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2023.04.02

Reinforcement Learning Based Efficient Power Control and Spectrum Utilization for D2D Communication in 5G Network

Chellarao Chowdary Mallipudi
Wireless Sensor Networks Lab, Department of Electronics and Communication Engineering, National Institute of Technology Patna, Patna, Bihar, 800005, India
E-mail: chellaraom.ug18.ec@nitp.ac.in
ORCID iD: https://orcid.org/0000-0002-5630-9969

Saurabh Chandra*
Wireless Sensor Networks Lab, Department of Electronics and Communication Engineering, National Institute of Technology Patna, Patna, Bihar, 800005, India
E-mail: saurabh.ec.jrf21@nitp.ac.in
ORCID iD: https://orcid.org/0000-0003-3029-3687
*Corresponding author

Prateek Prakash
Wireless Sensor Networks Lab, Department of Electronics and Communication Engineering, National Institute of Technology Patna, Patna, Bihar, 800005, India
E-mail: prateek.ec18@nitp.ac.in
ORCID iD: https://orcid.org/0000-0002-5171-8173

Rajeev Arya
Wireless Sensor Networks Lab, Department of Electronics and Communication Engineering, National Institute of Technology Patna, Patna, Bihar, 800005, India
E-mail: rajeev.arya@nitp.ac.in
ORCID iD: https://orcid.org/0000-0002-0346-2150

Akhtar Husain
Department of Computer Science & Information Technology, Mahatma Jyotiba Phule Rohilkhand University, Bareilly, Uttar Pradesh, 243006, India
E-mail: akhtarhusain@mjpru.ac.in
ORCID iD: https://orcid.org/0000-0002-0282-8608

Shamimul Qamar
Computer Science & Engineering Department, College of Sciences & Arts, Dhahran Al Janoub Campus, King Khalid University, Abha 64261, Kingdom of Saudi Arabia, KSA
E-mail: sqamar@kku.edu.sa

Received: 31 January 2022; Revised: 27 August 2022; Accepted: 23 March 2023; Published: 08 August 2023

Abstract: There are billions of inter-connected devices by the help of Internet-of-Things (IoT) that have been used in a number of applications such as for wearable devices, e-healthcare, agriculture, transportation, etc. Interconnection of devices establishes a direct link and easily shares the information by utilizing the spectrum of cellular users to enhance the spectral efficiency with low power consumption in an underlaid Device-to-Device (D2D) communication. Due to reuse of the spectrum of cellular devices by D2D users causes severe interference between them which may impact on the network performance. Therefore, we proposed a Q-Learning based low power selection scheme with the help of multi-agent reinforcement learning to detract the interference that helps to increase the capacity of the D2D network. For the maximization of capacity, the updated reward function has been reformulated with the help of a stochastic policy environment. With the help of a stochastic approach, we figure out the proposed optimal low power consumption techniques which ensures the quality of service (QoS) standards of the cellular devices and D2D users for

D2D communication in 5G Networks and increase the utilization of resources. Numerical results confirm that the proposed scheme improves the spectral efficiency and sum rate as compared to Q-Learning approach by 14% and 12.65%.

Index Terms: Reinforcement Learning, Fifth Generation (5G), Internet of Things (IoT), Power Control, Device-to-device (D2D).

1. Introduction

With the rapid advancement of intelligent devices and utilizations of smart phones with the internet brings great enhancement in the usage of devices and convenience to society in the next generation networks. Performance of these devices needs improvement in data rate, spectral efficiency with low power consumptions [1]. Demands of users increases which creates immense traffic to the users and causes major issues to the design of next generation mobile communication system [2]. To fulfil the demands of users, a wide variety of applications need to be provided to enhance performances of the networks. In various applications such as video calling, live conferences, 4k video streaming online, multimedia content sharing, live concerts etc. [3]. To meet users' expectation, Device-to-Device (D2D) communication comes into picture for providing better data rate, more energy efficient with low power consumption and low latency. In D2D communication, direct communication between users will happen when two users are in close proximity and establish a direct D2D link between them. Features of D2D communication include higher capacity, lower delay, high spectral utilization efficiency and low power consumption with high energy efficiency. Due to reuse of spectrum from cellular users by the D2D users causes severe interference between them needs to be addressed.

In the next 10 years, when we look at the 2030's, our expectation towards the 5G Network may be successfully transformed socially through several effective applications and services in the era of enhanced mobile broadband (eMBB), connectivity of massive machines etc. In the evolution of 5G Networks, there is a small need to go beyond 5G which can offer and provide new capabilities to the users. The time has arrived to start exploring the 5G element and formulate the important issues. Machine Learning (ML) concept is fully utilized to optimize the networks, which make possible many new services to make our lives in a better way. 5G facilitates major improvements in terms of performance parameters but also puts difficult energy requirements and low power devices. Decrement in the level of energy consumption and low power associated with the cellular network operations is relatively a major focus of area and shown interest by the researchers. Energy consumption defines the main features of D2D in terms of network energy efficiency but power consumed by individual mobile devices also needs to be considered. Application such as public safety communication is required in the case of infrastructure damage (Example: Kedarnath Disaster) as well as consciousness towards the proximity users. In future, D2D may allow users to take benefits in terms of shortest communication, spectral efficient, lower latency, enhanced data rate and less energy consumption. Enhance the performances in 5G new radio (NR), significantly mitigating the power and energy consumption of the network. Focuses on several new technologies and use cases that support D2D communication. Also highlights the issues on interference management [4], power control and resource optimization in D2D communication.

D2D communication is used to achieve the increased capacity with highly spectral and energy efficient devices and optimal performance of the network. To minimize the interference is one of the major issues in underlaid D2D communication. To tackle such issues, devices' power controlling method is one of the most crucial methods for interference mitigation between the devices while keeping a signal-to-noise-plus-interference ratio (SNIR) value above threshold value. Researchers have proposed several power consumptions schemes which minimize the interference between users during direct communication. Centralized and distributed are the two different ways to control the transmission power of the devices. Distributed power control [5] is most commonly used approach because requirement of local information by the base station is very less. Stochastic modelling, traditional game theory [6] and graph theory [7-9] are the techniques used for solving the issues in D2D communication. Apart from these techniques, Machine Learning (ML) [10] is one of the most efficient techniques which provides more accuracy on the outcomes after simulations. Many researchers have shown their interests for managing the resources using Reinforcement Learning (RL) and supervised learning in wireless communication. In 5G cellular networks, Q-Learning approach is utilized to resolve the issues of resource sharing to cellular users. Objective is to identify the problems occurred during reuse of the same resources and sharing the information directly from one device to another which may cause high power consumption with less data rate and less spectral efficient devices. It still remained to be addressed. Our aim is to provide the solution to the identified issues by using the ML concept.

In this work, transmission power scheme is proposed by utilizing the reinforcement learning techniques to achieve improved capacity and optimal behaviour of the network while ensuring the quality of service (QoS) of both D2D and cellular device users. The main contribution to this work is summarized below:

- In D2D underlaying cellular network model, focus is to improve the capacity of D2D users and utilization efficiency of spectrum. The throughput maximization problem is formulated under QoS constraints and transmission power.
- We assume D2D users as an agent and propose a Q-Learning based low power selection scheme and limit the power levels to reduce the interference. Approximation of Q-value has been considered for the maintenance of Q-table of the network and by limiting the sets of power level with change in state for achieving rewards.
- We utilize the concept of stochastic approach to achieve enhanced throughput with better resource utilization efficiency of the network for the users which ensures the quality of service (QoS) standard of the communication of cellular user and D2D user.
- We evaluate the proposed power control approach using Q-Learning techniques through comprehensive simulations. Proposed approach achieves low power consumption with higher throughput and efficient spectrum utilization of all the D2D users as compared to other scheme shown in the simulation results.

The outline of the work in this paper is categorized as follows: Section 2 describes the existing literature on D2D communication to solve several issues. After identifying the Section 3 describes the scenario for D2D communication and the objective function is defined and formulated. The proposed scheme used to solve the identified problem is elaborated in Section 4. The analysis and impact of different simulation parameters have been discussed in Section 5. Finally, the limitation of the proposed work and future scope of the work is concluded in Section 6.

2. Related Works

In this section, the existing method for minimization of power consumption related to the D2D communication using reinforcement learning are discussed. Basically, reinforcement learning is a machine learning algorithm used for controlling the policy and support for making intelligent decisions. Agents learn to map environmental states to actions to maximize their long-term rewards on a trial-and-error basis. Q-Learning is a model free reinforcement learning algorithm where the agent only knows the set of possible states and which action needs to perform. Also knows the current state of the environment. By maintaining the Q-functions in an updated Q-table. From Q-Learning, agents learn the policy where we can achieve maximum rewards. Agents select the actions on the basis of maximum value of actions. The whole process of learning and updating Q-table is called Q-Learning. Authors of [11], integrated reinforcement learning in D2D overlay and underlay communication with the challenges such as power control, resource sharing and also decision making has been sorted out. Number of research work have been done to manage and diminish the interference [4] for the enhancement of spectrum efficiency in D2D communication.

Authors in [12] have investigated the capability of usage of deep reinforcement learning (DRL) in D2D communication to provide a solution by solving the subcarrier power allocation problem by reducing the delay and achieving more reliability of the network. In the dearth of full instantaneous CSI with the help of DRL based resource allocation algorithm, we can try to learn the optimal policy. Authors in [13] describe the problem of gain of proximity users, interference of mutually connected devices in IoT. To solve these issues, authors utilize the concept of DRL and proposed DRL based control schemes to improve spectral efficiency. Overall rate of the network also gets improved by jointly considering the proposed scheduling approach for the allocation of resource block and transmission power control schemes to improve. To make devices highly energy efficient and maintain the minimum requirement of QoS is not considered by the authors. In [14], authors proposed a RL based scheme to distribute the power resources and selection of relays together for making devices highly energy efficient. By utilizing the concept of Q-Learning, issues on selection of relay can be solved. Authors have focused more on solving the problem of power control to maximize the energy efficiency of the devices. Demands of users for the service enhancement of rich contents over device networks have been increased which leads to serious traffic congestion to the networks. Q-Learning is used to enhance the quality of experience (QoE), maximize the spectrum utilization efficiency, and transmission rate. To make the delivery of a content more efficient, a content delivery model is proposed by the authors in [2] for the edge users using the Stackelberg approach. Authors in [15], considered the issues of spectrum allocation to the cellular user for intercell D2D communication. To address such issues, machine learning based Nash equilibrium techniques have been utilized. But they ignored it to maximize the data rate with a low power consumption device for the better performance of the system. In [16], the authors proposed an algorithm for the allocation of those channels which is free from collision to perform concurrent transmission without interfering with multiple channels in D2D communication using channel hopping techniques. Authors in [16] mainly considered channel allocation problems but issues of low power consumption have been missed. To support and access for both LTE and D2D users, authors in [17] designed spectrum sensing based protocol used in unlicensed spectrum for maximization of sum-rate with the help of greedy and GSO algorithm. To enhance the overall transmission rate, interference between users has been managed by the authors in [18] proposed max-sum message passing approach and branch-n-bound algorithm by exploiting channel diversity technique. The problem of low power D2D communication can be resolved by using deep reinforcement learning techniques in [19], where power distribution techniques to control power have been adopted to reduce the interference by consideration of outdated channel state information.

On the basis of the existing literatures, to solve different issues using RL based algorithms were the main objectives of the researchers. Such as sharing of resources, low power consumption, throughput enhancement with lower delay of D2D link etc. Also, keeping the QoS requirement for the D2D and cellular user to maximize capacity as per the demand by the users are also one of the main objectives. To achieve maximal throughput, the authors in [20] worked on distributed multi-agent learning techniques for the better distribution of resources. Using RL, the resource allocation problem being solved was the objective by the authors. However, the concept of RL in D2D is mainly utilized for power control, and sharing of resources from cellular users to device-to-device users to achieve maximum throughput and make devices with high spectral utilization and energy efficient. Even so, there are still many issues in D2D communication yet to be solved. There is a need of study to minimize power consumption from devices by D2D user and cellular user while maintaining the threshold SNIR to improve the capacity of D2D links. Meanwhile, managing the QoS, throughput maximization and achieving highly spectral efficiency of D2D and cellular users via Q-learning based power level selection approach motivates us to present this work.

3. Network Model and Problem Formulation

In this section, first we describe the scenario in the network model sub-section, followed by defining the problem formulation for the identified problem for achieving the higher capacity by consuming lesser power during transmission of signal from the devices in the next sub-section.

3.1. Network Model

In this subsection, scenario illustration of the D2D model is shown in Fig. 1. From Fig. 1, considered D2D Model consists with a cell for uplink D2D underlaid communication, where, we take an assumption that the users associate with D2D link and users associates with channel link of cellular users are in random position inside cell with single base station which is centrally located in a cell. We also consider that the number of D2D users are positioned inside the boundary of the cell. When users with closest proximity and satisfies the criteria of threshold distance between D2D links. Then, establishing the direct link between them for communication. There are index sets of M number of users indicating users of cellular network and index sets of N number of D2D user pairs representing the sets of cellular devices and D2D users such as $C_P = \{1, 2, \ldots\ldots, \mathrm{M}\}$, and $S_P = \{1, 2, \ldots\ldots, \mathrm{N}\}$. Base station allocates the spectrum to cellular devices which requires better information of the state of the channel. Sharing of resources by the base station to the cellular link will be done orthogonally. The bandwidth provided to the resource block can be denoted as α. Base stations that share the spectrum to the cellular user will be denoted as $\Re = \{1, 2, \ldots.., k,\ k+1, \ldots., K\}$. The channel gets affected by the fading due to shadowing, pathloss and also from propagation due to multiple paths.

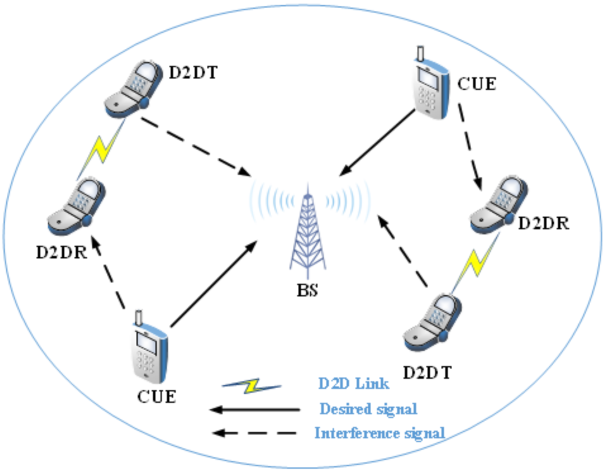

Fig.1. Scenario illustration of D2D users and cellular users

Here, we considered the impact of the multipath fading and shadowing during transmission of signal. In this work, the log-distance pathloss model and the gain and interference of the channel link of cellular device user and the D2D user at the receiver side can be expressed as in (1):

$$\Omega_i = \ell \lambda_i \varphi_i \partial_i^{-\omega} \tag{1}$$

where, ℓ represents the pathloss constant, channel gain with an exponential distribution with mean equals to 1 due to fast fading is represented as λ_i, φ_i represents the shadowing, ω is the factor due to pathloss, and ∂ is the cellular-D2D user distance. The distribution of spectrum in the form of multiple resource blocks to cellular users for communication with Base Station and share the spectrum with D2D users. The reception of signal at the receiver side of D2D users and at the base station with the help of equations expressed in (2):

$$y_m^{k,c} = \sqrt{\rho_m^k} h_{m,b}^k \phi_{m,b}^k + \sum_{m\in M, n\in N}^{k\in K} \chi_{m,n}^k \sqrt{\rho_n^k} h_{n,b}^k \phi_{n,b}^k + \eta \tag{2}$$

where, ρ_m^k is the power transmitted by the cellular devices, ρ_n^k is the power transmitted by the D2D users, $\phi_{m,b}^k$ and $\phi_{n,b}^k$ are the information transmitted by the cellular device and D2D users to the Base Station, η is the power with noise generated by the device, and $h_{m,b}^k$, $h_{n,b}^k$ are the gain of channel link from the m-number of users belong to cellular device to the BS and number of D2D user pair at transmitter to the BS. $\chi_{m,n}^k$ indicating the resource sharing block which is being shared to the cellular device to make communication possible using spectrum, and it is given by (3):

$$\chi_{m,n}^k = \begin{cases} 1 & \text{if D2D user gets resources} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Channels of D2D link and cellular devices suffers with large and small-scale fading. Based on (1), the expression for SNIR of the cellular device is given in (4):

$$\zeta_m^k = \frac{\rho_k^m * \Omega g_{m,k}^0}{\eta + \sum_{k=1,\, j\in\mathbb{R}_s}^{K, n\in N} \rho_k^{nj} * \Omega g_{j,k}^{j0}}, \tag{4}$$

where ρ_k^{nj} and ρ_k^m denotes the power emitted from the D2D pairs at the transmitter side and cellular devices, respectively, and η is the power with noise generated by the device.

Similarly, the expression for the SNIR of the D2D pairs will be given in (5)

$$\zeta_n^k = \frac{\rho_k^{ni} * \Omega g_{ni,t}^{ii}}{\eta + \rho_k^m * \Omega g_k^m + \sum_{k=1,\, j\in\mathbb{R}_s}^{K, n\in N} \chi_{n.m}^k \rho_k^{nj} * \Omega g_{nj,k}^{ji}}, \tag{5}$$

where $\Omega g_{ni,k}^{ii}, \Omega g_{nj,k}^{ji}$, and Ωg_k^{mi} represent the gain of the i^{th} D2D link at transmitter, gain between i^{th} D2D pair at transmitter and j^{th} D2D pairs at receiver of the different pair, and gain of the channel link of number of users of cellular device user to the i^{th} number of of D2D user pair at receiver, respectively. χ is indicating the resource sharing block which is being shared to the cellular device to make communication possible using spectrum.

The achievable capacity of the cellular device with the help of Shannon capacity formula applied using (4) is given in (6)

$$C_c = \alpha \log_2(1+\zeta_m^k) \tag{6}$$

Similarly, the achievable throughput for the D2D user using (5) is given in (7)

$$C_d = \alpha \log_2(1+\zeta_n^k), \tag{7}$$

where α is the bandwidth.

3.2. Problem Formulation

Due to reuse of spectrum by D2D users which is allocated to the cellular users causes interference between them, which may have a huge impact in their quality of communication. Therefore, spectrum sharing and low power devices for ensuring the performance of the network is more important. In this paper, performance parameters such as quality of service (QoS) with maximum power constraints and maximization of throughput of both cellular as well as D2D users is considered as the main objective. The problem is modeled as given in (8):

$$C_{sum}^{k} = \max(\sum_{k=1}^{\mathrm{K}} \sum_{i \in \mathbb{R}_t} \chi_{n,m}^{k} * (\alpha \log_2(1+\zeta_m^k) + \alpha \log_2(1+\zeta_{n,i}^k)) , \tag{8}$$

$$\text{s.t., } \sum_{k=1}^{K} \chi_{n,m}^{k} \le 1 \text{ and } \chi \in \{0,1\} ,$$

$$\zeta_n^k \ge \zeta_n^{k\min} \text{ and } \zeta_m^k \ge \zeta_m^{k\min} ,$$

$$\rho_c \le \rho_{c\max} \text{ and } 0 \le \rho_d \le \rho_{d\max}$$

where $\zeta_n^{k\min}$ and $\zeta_m^{k\min}$ are the target threshold of SNIR of D2D pairs and cellular devices. $\rho_{c\max}$ and $\rho_{d\max}$ are the limitation of power which can be maximally transmitted by the cellular devices and D2D pair at the transmitter side. The main target is to formulate the objective function to solve the identified issues of low power consumption, less spectral efficient devices, which maximizes the overall capacity for the users which is highly dependable on the requirement of QoS constraints and maximum power constraints. To address these issues, we will discuss proposed approach for low power consumption by consideration of multiple users. With the help of stochastic approach, we figure out the proposed optimal low power consumption techniques for D2D users which will be given in the upcoming section.

4. Q-Learning Based Low Power Selection Schemes and Throughput Maximization

In this section, we are going to discuss our proposed work based on reinforcement learning (RL). To solve the problem of low power consumption to minimize the interference and maximize the data rate by the help of proposed approach. It is a two-stage process, where the Q-Learning algorithm is applied first. The selection of power to perform an action by the agents to bound the power levels to a certain limit and diminish the interference by the selection of the best action (power) for getting the positive reward. If we know the user's location by calculating the distance from the reference point, we can calculate the impact of performance parameters such as channel gain, interference, SNIR, data rate etc. on the performance of the network using (3,4, and 5). For the maximization of throughput, the updated reward function has been reformulated with the help of a stochastic policy environment in the next stage discussed in the subsections.

4.1. Q-Learning Based Low Power Consumption Scheme for Multiple User

We know that Q-Learning is a model free RL algorithm. Each D2D transmitter tried their level best to get the maximum rewards by adjusting the level of actions within a certain given limit in terms of the sum data rate depending on the environment. To optimize the throughput by maintaining QoS and reduce the power consumption for efficient utilization of spectrum from cellular users shared across each resource block, we opt the multi-agent reinforcement learning technique due to consideration of number of D2D pairs reuse the spectrum from the resource block of cellular user using Markov decision process model. The Markov decision process is defined with the standard elements, which are the important components of the Q-Learning approach. It is described as follows:

Agent: Agent understands the environment and learns from trial-and-error basis. It is nothing but a user which describes the D2D users at each transmitter of the D2D user.

Action Space: Consider that all the users need to choose resources from the resource block of cellular users without any involvement of the base station. The action space of each D2D user can be defined as $\psi_{t,n}^{k} = \{\Re, \rho_n\}$ at the given time slot t . D2D user must satisfy the constraints $\zeta_m^{c,k} \ge \zeta_m^{c,k\min}$ when it selects the actions from the given action levels, given by (9)

$$\psi_{t,n}^{k} = (\rho_1, \rho_2, \ldots.., \rho_k) \tag{9}$$

State Space: We define the state space $\Delta_{t,n}^{k}(t)$ in (10) to mentioned whether each user (agent) meets its QoS demand at time t.

$$\Delta_{t,n}^{k} = \left\{ \Omega_{t,n_i}^{k}, \Omega_{t-1,m,n_j}^{c,k}, \Omega_{t-1,n_i,n_j}^{d,k}, C_{t,n}^{d,k}, K_{t-1,n} \right\} \tag{10}$$

Where $\Delta_{t,n}^{k}$ belongs to {0,1}, $\Delta_{t,n}^{k}(t) = 0$ means when users don't meet the minimum QoS requirement, we will get penalized by -1 and $\Delta_{t,n}^{k}(t) = 1$ means min requirement is maintained and we will get positive rewards. While doing this, we noticed that the number of possible states are 2^n and this is huge for users.

Strategy: Policy is basically the strategy that the agent uses to find out his next action based on his current state. Policy is just the strategy with which we can approach the tasks. ε-greedy strategy is applied to this work. Let P_r be the action which performs in a random manner. Exploration is about exploring and capturing more information about the environment. And exploitation is about using the already known exploited information to heighten the rewards. Basically, policy is the path taken to achieve the target. It can be represented as in (11)

$$\pi_{\psi}^{\Delta} = \begin{cases} select\, P_r : E\left(P_r\right) = \varepsilon & \forall \varepsilon \in [0,1] \\ select\ \arg\max_{\Delta_k \in P_k} Q(\Delta_t^k, \psi_t^k) : E\left(\psi\right) = 1 - \varepsilon & \forall \varepsilon \in [0,1] \end{cases} \tag{11}$$

Reward: RL agent that is a D2D user must be trained in such a way that he takes the best action so that the reward is maximized. Reward is in terms of capacity, which is given in (12):

$$C_k = \begin{cases} \alpha \log_2(1 + \zeta_n^d), & \zeta_m^{c,k} \geq \zeta_m^{c,k\min} \\ -1, & otherwise \end{cases}. \tag{12}$$

Agent understands the environment. Based on understanding, the agent starts taking action to reach the rewards. Each time an agent is performing an action, the environment gives the reward what he got from the agent. Again, the agent starts understanding the changed environment and this process will continue until the rewards get maximized. The action level is selected where we get maximum reward. Then, update the Q-value which is given in (13):

$$Q_{t+1}(\Delta_{t+1,}^k, \psi_{t+1}^k, t) = \begin{cases} \max(Q_t(\Delta_{t,}^k, \psi_t^k, t), \\ (\mathrm{R}_{t+1} + \gamma * \max(Q_t(\Delta_{t+1,}^k, \psi_{t+1}^k, t)) & if\ s = \Delta_t^k\ and\ \rho = \psi_t^k \\ Q_t(\Delta_t^k, \psi_t^k, t) & otherwise \end{cases} \tag{13}$$

where $Q_{t+1}(\Delta_{t+1,}^k, \psi_{t+1}^k)$ is denotes the Q-value, state represents the current condition denoted by $\Delta_{t+1,}^k$, ψ_{t+1}^k denotes all the possible steps that agents can take, and discount factor is used to make balance to immediate and future rewards denoted by γ.

4.2. Throughput Maximization for the Enhancement of the Network Performance Using Stochastic Approach

In this subsection, the concept of stochastic approach has been utilized. D2D users make their own decision on the basis of their interaction with the environment to maximize their rewards which is dependent upon immediate reward function. This is nothing but the learning process of reinforcement learning. The immediate reward function at time t is defined as in (14):

$$\vartheta_t^k = C_{sum}^k - W_1 \sum_{m=1}^{M} \left(\zeta_m^{c,k} - \zeta_m^{c,k\min} \right) - W_2 \sum_{m=1}^{M} \left(\zeta_n^{d,k} - \zeta_n^{d,k\min} \right) \tag{14}$$

In (14) shows that the first term is the objective function describing the problem formulation for maximizing the capacity (throughput) of all users, and remaining terms show the required minimal information of the users corresponds to cellular device and D2D. W_1 and W_2 are the weights respectively.

$$\vartheta_t^k = \begin{cases} C_{sum}^k & \zeta_m^{c,k} \geq \zeta_m^{c,k\min}\ and\ \zeta_n^{d,k} \geq \zeta_n^{d,k\min} \\ -1 & otherwise \end{cases} \tag{15}$$

Above equation (15) represents the positive reward, only when we satisfy the condition of minimum data rate requirement otherwise, we may get penalized for the D2D users. According to the Markov decision process model, D2D users observe a state from the state space and select power from sets of action space accordingly. After that, selecting resources from the allocated resource block of cellular devices and transmitted power is controlled on the basis of policy environment $\pi\left(\Delta_{t,n}^k\right)$. The selected action level, and the condition of the environment can be perceived by the D2D users when its transition on to a next state $\Delta_{t+1,n}^k$ and the D2D users receives a reward.

At each step, the actions and states get updated after each iteration. Agents can’t learn much after a single iteration. But after enough exploring, it may converge and learn optimal Q-value. Agents select an action by referencing Q-table with maximal Q-value or by random value. The updated Q-value is expressed with the help of (13) given in (16):

$$Q_{t+1}(\Delta^k_{t+1,}, \psi^k_{t+1}, t) = \begin{cases} \max(Q_t(\Delta^k_{t,}, \psi^k_t, t), & \\ \vartheta^k_{t+1} + \gamma * \max(Q_t(\Delta^k_{t+1,}, \psi^k_{t+1}, t)) & \text{if } \Delta = \Delta^k_t \text{ and } \psi = \psi^k_t \\ Q_t(\Delta^k_t, \psi^k_t, t) & \text{otherwise} \end{cases} \tag{16}$$

Stochastic policy environment can be determined for selecting actions from the sets of power levels. Gaussian probability distribution function (pdf) is used [21], which is expressed as in (17)

$$\pi\left(\Delta^k_{t,n}\right) = \frac{1}{\sqrt{2\pi}\sigma^k_{t,n}} \exp\left(-\frac{\left(\psi^k_{t,n} - r\left(\Delta^k_{t,n}, \sigma^2\right)\right)^2}{2\sigma^2} \right) \tag{17}$$

Where $r\left(\Delta, \sigma\right)$ is the mean of the normal distribution and standard deviation of the pdf is denoted by σ .

5. Results and Discussion

In this work, we proposed a Q-Learning based low power consumption scheme with the help of reinforcement learning techniques to solve the power control problem with better utilization of spectrum using the stochastic approach. The simulated results are illustrated and the analysis of the result is to demonstrate the dominance of the proposed method and also to seek-for the efficacy on the throughput and spectral efficiency of D2D communication with respect to minimum requirement of QoS. Here, we are assuming that the N number of users corresponds to D2D and M number of users represents users of cellular devices distributed randomly inside a single cell at a radius of 500m, and the maximum D2D distance is $D_{max} = 50m$ that is D2D link distance. For the channel gain calculation, we assume a log-distance model to simulate pathloss model, and for multipath propagation, exponential distribution model with unit means and 8dB standard deviation due to shadowing. The noise power spectral density at any location is considered as -174dBm/Hz. Bandwidth available for the network is set as 10MHz. The threshold SNIR of cellular users is set to 6dB. In this work, the proposed power control approach is compared with a multi-agent Q-Learning algorithm. Parameters used for determining the channel gain, SNIR, throughput, spectral efficiency etc. System simulation parameter is shown in Table 1.

Table 1. System Simulation Parameters

Simulation Parameter	Value
Cell Radius	500m
Maximum D2D link distance, D_{max}	10-50m
Number of D2D users, N	10-100
Number of Cellular users, M	5-50
Number of Resource Block	50
Bandwidth, α	10MHz
Pathloss constant, κ	10^{-2}
Pathloss for cellular device link	15.3 + 37.6log10(d/km) dB
Pathloss for D2D links	22 + 40log10(d/km) dB
Shadowing (Standard deviation), δ	8dB log-normal
Multipath fading (exponential distribution), η	Unit mean
Threshold SNIR of D2D user	6dB
Threshold SNIR of Cellular user	6dB
Max power transmitted by cellular device and D2D user, $\rho^c_{\max}$ and $\rho^d_{\max}$	23dBm and 10-23dBm
Noise power,	-174dBm/Hz
Pathloss Exponent α	4

In Fig. 2, we analyze the graph of spectral efficiency of D2D users with respect to the number of D2D users by keeping varying the maximum D2D link distance. Spectral Efficiency of D2D users slightly decreases when the D2D link distance increases by 10, 20, 25 and 30m while $P_{max} = 23\text{dBm}$. Because longer D2D link distance, causes high power consumption for making D2D communication possible. Hence, it lowers the spectral efficiency with increase in D2D link distance. Fig. 3 illustrates some fluctuation of throughput of D2D users for $P_{max} = 10, 15, 20$ and 23dBm while the D2D link distance varies from 10 to 100m. From Fig. 3, we found that throughput decreases with increase of

D_{max}. Since a higher P_{max} will results in degradation of throughput as D_{max} increases. Longer the distance between D2D users causes high power consumption for making D2D communication possible. Hence, it lowers the throughput with increase in D2D link distance.

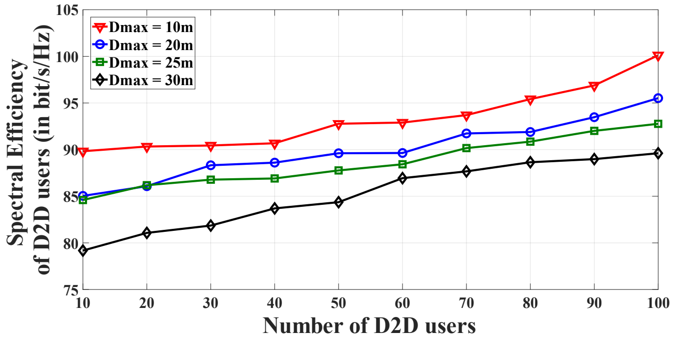

Fig.2. Spectral Efficiency versus Number of D2D users

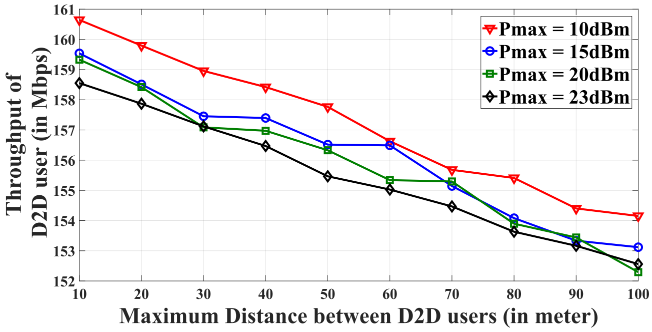

Fig.3. Throughput of D2D users versus Maximum distance between D2D users

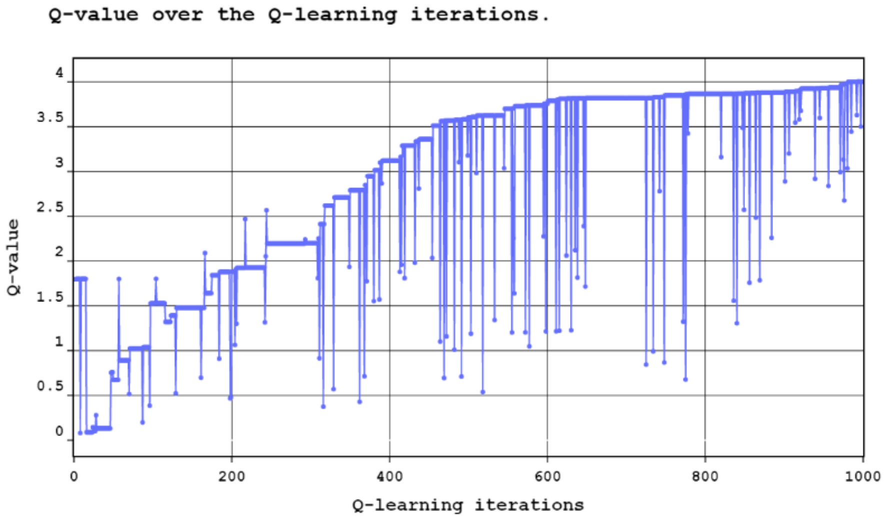

Fig.4. Q-value over the Q-Learning iterations

In Fig. 4, the graph has been plotted between Q-value v/s Q-Learning iterations. We observed from the Fig. 4, simulation results shows that the Q-value has been improved when learning rate, which makes an impact on the convergence speed of Q-value with respect to increase in number of iterations.

Fig. 5 illustrates the graph plotted for spectral efficiency and number of D2D users for the comparison of proposed power control approach with standard Q-Learning approach. The number of D2D users varies from 10 to 100 and $P_{max} = 23\text{dBm}$ with $D_{max} = 50\,\text{m}$. As we can be observed from Fig. 5, the proposed power control approach provides achievable spectral efficiency as compared to the Q-Learning approach by 14% when $P_{max} = 23\text{dBm}$ and $D_{max} = 50\,\text{m}$. Fig. 6, graph has been plotted for the comparison of throughput of D2D users between proposed power control approach and Q-Learning approach. From Fig. 6, we found that throughput decreases with increase of D_{max}. Throughput is slightly decreased when the distance between D2D transmitter user and D2D receiver user of one pair varies from 10 to 30m while $P_{max} = 23\text{dBm}$. Because longer the distance between D2D users causes high power consumption making D2D communication possible. Hence, it lowers the throughput with increase in D2D link distance. And the proposed power control approach performs well compared to Q-Learning power control approach by 12.65% when $P_{max} = 23\text{dBm}$.

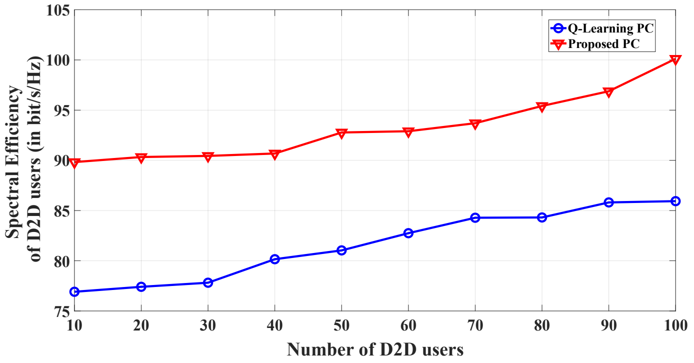

Fig.5. Spectral Efficiency of D2D users' compare with Proposed PC algorithm and Q-Learning PC algorithm for different number of D2D users for P_{max}=23dBm and D_{max}=50m

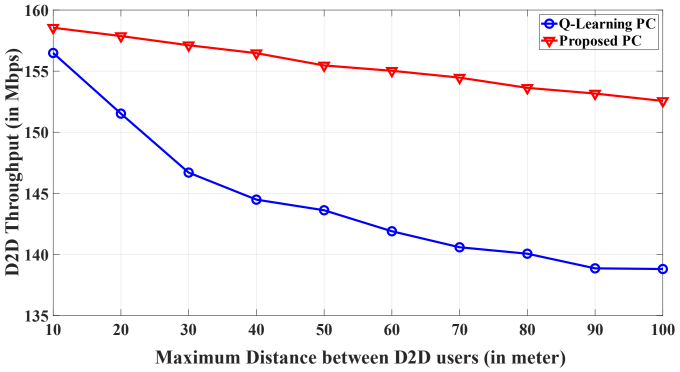

Fig.6. Throughput of D2D users' compare with Proposed PC algorithm and Q-Learning PC algorithm for different number of D2D link distance users for P_{max}=23dBm

6. Conclusions

The scope of this work is limited to overloading the states which may huge impact on the results and takes lot of computation time due to slow convergence. Complexity was not evaluated which is outside the reach of present discussion because it involves the topological parameter analysis.

This article shows a view of reinforcement learning based efficient power control and spectrum utilization approach is proposed for D2D communication in two stages. To maximize the capacity as per the demand of user may rises by reducing the interference. First to control the transmission power, which is an important mechanism for interference mitigation and with the help of stochastic policy environment in the second phase, throughput of the network gets incremented. When D2D users achieves maximum throughput, at that time D2D user aim is to transmit low power from the set of power level using Q-Learning techniques. Simulation results shows the proposed method is significantly enhance the spectral efficiency and throughput of D2D users as compared to Q-Learning approach by 14% and 12.65% while maintaining the requirement of QoS of cellular user and D2D user. Future scopes of this article need to focus on how to jointly optimize the power, spectrum reuse and channel selection by using to enhance the throughput as well as high utilization of spectrum.

References

[1] X. Shen, "Device-to-device communication in 5G cellular networks," *IEEE Network*. 2015.

[2] Z. Su, M. Dai, Q. Xu, R. Li, and S. Fu, "Q-Learning-Based Spectrum Access for Content Delivery in Mobile Networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 1, pp. 35–47, 2020.

[3] D. Wu, L. Zhou, Y. Cai, H. C. Chao, and Y. Qian, "Physical-Social-Aware D2D Content Sharing Networks: A Provider-Demander Matching Game," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7538–7549, 2018.

[4] R. S and S. G K, "Interference Mitigation and Mobility Management for D2D Communication in LTE-A Networks," *Int. J. Wirel. Microw. Technol.*, vol. 9, no. 2, pp. 20–31, 2019.

[5] M. H. Faridi, A. Jafari, and E. Dehghani, “An Efficient Distributed Power Control in Cognitive Radio Networks,” *Int. J. Inf. Technol. Comput. Sci.*, vol. 8, no. 1, pp. 48–53, 2016.

[6] E. Ogidiaka, F. N. Ogwueleka, and M. Ekata Irhebhude, “Game-Theoretic Resource Allocation Algorithms for Device-to-Device Communications in Fifth Generation Cellular Networks: A Review,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 13, no. 1, pp. 44–51, 2021.

[7] Y. Wei, Y. Qu, M. Zhao, L. Zhang, and F. Richard Yu, “Resource allocation and power control policy for device-to-device communication using multi-agent reinforcement learning,” *Comput. Mater. Contin.*, vol. 63, no. 3, pp. 1515–1532, 2020.

[8] M. Abrar, R. Masroor, I. Masroor, and A. Hussain, “IOT based efficient D2D communication,” *Moscow Work. Electron. Netw. Technol. MWENT 2018 - Proc.*, vol. 2018-March, pp. 1–7, 2018.

[9] D. Singh and S. C. Ghosh, “A distributed algorithm for D2D communication in 5G using stochastic model,” *2017 IEEE 16th Int. Symp. Netw. Comput. Appl. NCA 2017*, vol. 2017-Janua, pp. 1–8, 2017.

[10] M. Mamdouh, M. Ezzat, and H. Hefny, “Optimized Planning of Resources Demand Curve in Ground Handling based on Machine Learning Prediction,” *Int. J. Intell. Syst. Appl.*, vol. 13, no. 1, pp. 1–16, 2021.

[11] H. Yang, X. Xie, and M. Kadoch, “Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, 2019.

[12] B. Gu, X. Zhang, Z. Lin, and M. Alazab, “Deep Multiagent Reinforcement-Learning-Based Resource Allocation for Internet of Controllable Things,” *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3066–3074, 2021.

[13] I. Budhiraja, N. Kumar, and S. Tyagi, “Deep-Reinforcement-Learning-Based Proportional Fair Scheduling Control Scheme for Underlay D2D Communication,” *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3143–3156, 2021.

[14] X. Wang, T. Jin, L. Hu, and Z. Qian, “Energy-Efficient Power Allocation and Q-Learning-Based Relay Selection for Relay-Aided D2D Communication,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6452–6462, 2020.

[15] J. Huang, Y. Yin, Y. Zhao, Q. Duan, W. Wang, and S. Yu, “A Game-Theoretic Resource Allocation Approach for Intercell Device-to-Device Communications in Cellular Networks,” *IEEE Trans. Emerg. Top. Comput.*, vol. 4, no. 4, pp. 475–486, Oct. 2016.

[16] H. Zhao, K. Ding, N. I. Sarkar, J. Wei, and J. Xiong, “A Simple Distributed Channel Allocation Algorithm for D2D Communication Pairs,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10960–10969, Nov. 2018.

[17] H. Zhang, Y. Liao, and L. Song, “D2D-U: Device-to-Device Communications in Unlicensed Bands for 5G System,” *IEEE Trans. Wirel. Commun.*, vol. 16, no. 6, pp. 3507–3519, Jun. 2017.

[18] D. Della Penda, A. Abrardo, M. Moretti, and M. Johansson, “Distributed Channel Allocation for D2D-Enabled 5G Networks Using Potential Games,” *IEEE Access*, vol. 7, pp. 11195–11208, 2019.

[19] J. Shi, Q. Zhang, Y.-C. Liang, and X. Yuan, “Distributed Deep Learning for Power Control in D2D Networks With Outdated Information,” *IEEE Trans. Wirel. Commun.*, vol. 20, no. 9, pp. 5702–5713, Sep. 2021.

[20] K. Zia, N. Javed, M. N. Sial, S. Ahmed, A. A. Pirzada, and F. Pervez, “A Distributed Multi-Agent RL-Based Autonomous Spectrum Allocation Scheme in D2D Enabled Multi-Tier HetNets,” *IEEE Access*, vol. 7, no. Figure 1, pp. 6733–6745, 2019.

[21] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, “Energy Efficiency and Spectrum Efficiency of Multihop Device-to-Device Communications Underlaying Cellular Networks,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 367–380, Jan. 2016.

Authors’ Profiles

Chellarao Mallipudi is currently pursuing his Bachelor of Technology degree in Electronics and Communication Engineering from National Institute of Technology Patna, Bihar India. His research interests include Device-to-Device Communication, Internet of Things and Wireless Communication.

Saurabh Chandra received his B.E. degree in ETC from CSVTU Bhilai, India in 2014 and M.Tech in Communication System Engineering from KIIT Bhubaneswar, India in 2017. He is currently a research fellow in the Department of Electronics and Communication Engineering at National Institute of Technology Patna, Bihar India. His research interests include Wireless Communication and Device to Device Communication.

Prateek Prakash received his B.Tech degree in ECE from SPSU Udaipur India in 2014 and M.Tech in Communication System Engineering from KIIT Bhubaneswar, India in 2017. He is currently pursuing PhD in the Department of Electronics and Communication Engineering at National Institute of Technology Patna, Bihar India. His research interests include Wireless Communication and Soft Computing Techniques.

Rajeev Arya received the Engineering Degree in Electronics & Communication Engineering from Government Engineering College, Ujjain, (RGPV University, Bhopal) India in 2008, and the Master of Technology in Electronics & Communication Engineering from Indian Institute of Technology (ISM), Dhanbad, India in 2012. He received the Ph.D. in Communication Engineering from Indian Institute of Technology (IIT Roorkee), Roorkee, India in 2016. He has received Ministry of Human Resource Development Scholarship (MHRD India) during M.Tech and Ph.D. He is currently an Assistant Professor with the Department of Electronics & Communication Engineering at National Institute of Technology, Patna, India. His current research interests are Communication Systems & Wireless Communication.

Akhtar Husain received his B.E. degree in Computer Science & Engineering from Uttarakhand Technical University, Dehradun, India in 1996 and M.Tech in Computer Science & Engineering from NITTTR, Panjab University, Chandigarh, India in 2009. He received the Ph.D. from Indian Institute of Technology (IIT Roorkee), Roorkee, India in 2017. He is currently an Associate Professor with the Department of Computer Science & Engineering at MJP Rohilkhand University, Bareilly, Uttar Pradesh, India, India. His research interests include Fuzzy Logic, Cloud Computing, Wireless Communication.

Shamimul Qamar has done his B.Tech from MMMEC Gorakhpur, M.Tech from AMU, Aligarh and earned his Ph.D. degree from IIT Roorkee. Prof. Qamar has a wide teaching experience in various Engineering colleges. He has research interests in Communication & Computer network, Computer Networks, Multimedia applications, Internet applications, Satellite network, DSP and Image Processing. He has published several research papers in reputed national/international Journals and conference. He served as a Consultant in Jackson State University, the USA. He has written some text books and chapters in the field of Electronics & Computer Engineering. He is also a technical programme committee member in the International Mobility Conference, Singapore. He is a life member of International Association of Engineers and a life member of Indian Society of Technical Educational. His technical endeavours helped to set up a research lab according to latest technical innovations. He has actively participated in various technical courses workshops, seminar etc. at the IITs.

How to cite this paper: Chellarao Chowdary Mallipudi, Saurabh Chandra, Prateek Prakash, Rajeev Arya, Akhtar Husain, Shamimul Qamar, "Reinforcement Learning Based Efficient Power Control and Spectrum Utilization for D2D Communication in 5G Network", International Journal of Computer Network and Information Security(IJCNIS), Vol.15, No.4, pp.13-24, 2023. DOI:10.5815/ijcnis.2023.04.02