Sum Throughput Maximization Scheme for NOMA-Enabled D2D Groups Using Deep Reinforcement Learning in 5G and Beyond Networks

Mohammad Aftab Alam Khan, Hazilah Mad Kaidi, *Senior Member, IEEE*, Norulhusna Ahmad, and Masood Ur Rehman *Senior Member, IEEE*

***Abstract*— Device-to-Device (D2D) communication underlaying cellular network is a capable system for advancing the spectrum's efficiency. However, in this condition, D2D generates cross-channel and co-channel interference for cellular and other D2D users, which creates an excessive technical challenge for allocating the spectrum. Despite this, massive connectivity is another issue in the 5G and beyond networks that need to be addressed. To overcome this problem, non-orthogonal multiple access (NOMA) is integrated with the D2D groups (DGs). In this paper, our target is to maximize the sum throughput of the overall network while maintaining the signal-to-interference noise ratio (SINR) of the cellular and D2D users. To achieve the target, a discriminated spectrum distribution framework dependent on multi-agent deep reinforcement learning (MADRL), termed a deep deterministic policy gradient (DDPG) is proposed. Here, it shares the global historical states, actions, and policies using the duration of central training. Furthermore, the proximal online policy scheme (POPS) is used to decrease the computation complexity of training. It utilized the clipping substitute technique for the modification and reduction of complexity at the training stage. The simulation results demonstrated that the proposed scheme POPS attains 16.67%, 24.98%, and 59.09% higher performance than the DDPG, Deep Dueling and deep Q-network (DQN).**

***Index Terms*— D2D, NOMA, DGs, SINR, MADRL, DDPG, POPS, and DQN**

I. Introduction

DEVICE-to-device communication (DDC) is an optimistic technologies for 5G and beyond networks. It benefits the system by lowering network latency and enhancing throughput. DDC can increase energy efficiency (EE) of wireless networks by off-loading data traffic in cellular networks to prevent congestion [1]. Despite these potential advantages co-channel and cross-channel interference poses a significant challenge that must be addressed to increase spectrum efficiency and meet the end-user quality of service (QoS) requirements [2], [3]. Regardless of these interferences, massive connectivity is another issue in the 5G and beyond networks that need to be addressed. To solve this problem, researchers from both academia and industry proposed NOMA technique.

The NOMA is an emerging technique for 5G and beyond networks, which have the ability to tackle the pressure of massive data traffic [4]. The NOMA technique allows several users to share the same kind of frequency resources at the same span of time while using different power levels. Therefore, it can address massive connectivity opportunities as well as improve overall network throughput compared to traditional multiplexing techniques [5]. Also, the presence of successive interference cancellation (SIC) at the receiver's side mitigates the effect of intra-user interference. Additionally, it has been shown that SIC can improve broadcast quality at the expense of a more complex receiver design [6]. However, the performance of NOMA is degraded in the condition of a dynamic environment. To tackle this challenge, the model-free reinforcement learning (RL) approach is used [7].

The RL is an effective method in which an autonomous agent makes successive judgments utilising a variety of mathematical operations. In RL, an agent learns via trial and error how to interact with the changing environment. Also, an agent improves the efficiency of its prior activity using previous learning environment. The agent then performs new action, evaluates the results of the encounter, and makes a choice using interactive learning techniques [7]. The RL method, however, performs sluggishly in large, stochastic, and unpredictable networks. It is therefore inappropriate for massive networks [8]. The deep learning (DL) methods have recently been employed to address this issue. These methods are capable of providing versatile and efficient solutions for reducing computational complexity even with big datasets. To overcome RL's limitations, DL is combined with RL and

This work is supported by the Ministry of Higher Education MoHE under Fundamental Research Grant Scheme R.K130000.7856.5F506 and supported by the Universiti Teknologi Malaysia under UTM Transdisciplinary Research Grant Q. K130000.3556.06G45.

Mohammad Aftab Alam Khan is with the Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur, 54100 Malaysia. email: (aftab1982@graduate.utm.my)

Hazilah Mad Kaidi is with the Department of Engineering and Technology, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur, 54100, Malaysia. (e-mail: hazilah.kl@utm.my).

Norulhusna Ahmad is with the Department of Engineering and Technology, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur, 54100, Malaysia. (e-mail: norulhusna.kl@utm.my).

Masood Ur Rehman is with Electronics and Telecommunication Engineering, James Watt School of Engineering, University of Glasgow, G128QQ. email:(masood.UrRehman@glasgow.ac.uk).

referred to as deep reinforcement learning (DRL) [8].

The DRL is a powerful technique for embedded optimization that can operate quickly in wireless communication networks [9]. DRL approaches train neural networks offline before deploying them on terminal devices or controllers. It uses the learned model to forecast the optimal transmission power scheduling with minimum processing complexity for resource management. Also, the DRL technique learn in an online fashion find the best relationship between each state-action pair and its aggregate reward [9], [10].

A. Related work

In this section, the existing techniques to manage cellular users (CUs) and D2D transmitters (DTs) transmit power in order to mitigate interference between D2D users (DUs) and CUs to enhance overall sum-rate have been investigated and studied. In a centralised architecture, the agent is each D2D link. The training centre (usually the base state or local AP) receives channel state information (CSI) from the agent and conducts training and testing in a centralised fashion. In [8], the authors designed a DRL-based joint resource block (RB) scheduling and power control scheme to improve the sum-rate of the network while considering the user's fairness among all the links. In [11], the authors developed statistical-feature-based power regulation to increase overall sum throughput by minimizing co-channel interference induced by DTs and CUs. The authors of [12] investigated the resource allocation problem using interference control and proposed a heuristic technique to optimize the system's sum throughput while meeting the interference limitations. In [13], [14], the authors examined the resource allocation problem of DDCs in diverse environments by merging millimeter wave (mmWave) and cellular bands, proposing a coalition building method and a heuristic approach to increase system's sum throughput. In [15], authors used the static and repeated game model to examined the resource allocation problem of multi-cell D2D communications. Here, each player's transmission information is kept secret from other players and the D2D links use shared resources of multiple cells. The power allocation issue for DDCs embedded in cellular networks in the context of SWIPT was examined in [16]. In this, a game model has proposed in which each D2D link selects the transmit power and power splitting ratio that will maximise its utility. In [17], authors maximized the energy efficiency for content sharing with Collaborative Mobile Clouds.

In [18], authors suggested a lightweight blockchain to help swarms of heterogeneous unmanned aircraft systems (UASs) enhance routing security while working with limited computing resources. The swarm UASs can reduce assaults from malicious UASs and restrict their connections to the swarm UAS networking by using lightweight blockchain. Similar to the authors of [18], authors in [19] conducted a thorough review of the literature that has already been published in the field of UAS detection and mitigation, identified the difficulties in preventing the use of unauthorised or unsafe UAS, and assessed the trends in detection and mitigation for defending against UAS-based threats. The authors investigated a bio-inspired routing for UAS swarm networking in [20]. Each UAS swarm exhibits the key traits of cell wall construction, which was modelled after the biological cell paradigm. To increase the viability and throughput of heterogeneous UAS swarm networking, the authors in [21] developed a cell wall structure for intercommunication between the networks. In [22], authors proposed an ideal cell wall model to increase the throughput in heterogeneous UAS swarm networking. In [23], authors developed the model of biological cell wall communication for heterogeneous swarm UAS networking. Also, the authors addressed the edge-coloring problem of cell wall communication scheduling with the use of reinforcement learning in order to obtain the highest throughput possible between the heterogeneous swarm UAS networking on a global scale. The methodologies proposed in the aforementioned literature required almost accurate network knowledge to adequately solve the optimization challenge. Also, traditional optimization methodologies are greatly hampered by the increasing complexity and variety of wireless networks. In particular, resource allocation problems in complicated wireless networks are frequently described as non-convex, combinatorial, or mixed integer non-linear programming problems. Furthermore, in a dynamic wireless communication environment, the unpredictability of channel status information is readily detrimental to the performance of traditional systems [24].

In order to solve the resource allocation problem in D2D communication, a more adaptable architecture is required. RL has been a potent strategy for resource management concerns, particularly in wireless communication networks [25]. Additionally, the agent may interact directly with the environment, manage resources, and interfere to maximise its own strategy [26]. Using a distributed power allocation system based on Q-learning, the transmission rate of DTs may be enhanced while keeping the QoS of CUs was investigated in [27]. However, it was observed that the increases in state and action spaces result in the curse of dimensionality, making it more difficult for Q-learning to store all state-activity values in a table form. Therefore, DRL has been found to be suitable to solve the challenge of intelligent resource management because DRL yields more robust learning in high dimensions state and action spaces.

In [28], the authors studied a DQN approach to improve the total data rate by investigating user association and power allocation, which investigated the challenge of increasing non-LoS transmission performance in 5G wireless communication. In [29], the authors developed a DQN-based system for energy-efficient resource allocation in ultradense networks. The DRL-based resource management systems proposed in [28], [29] were implemented in a centralised architecture, with the central controller controlling the optimization problem. Large transmission overheads caused by rising network sizes place an additional computing burden on the central controller. To address this issue, distributed resource allocation systems based on DRL was developed in order to decrease computing complexity. In [30], the authors addressed the problem of throughput optimization. Here, the proximal online policy technique is used for quick sampling. Also, to the enhance capacity and quality of service (QoS) of the network, a DRL technique based on DQN is considered. In [31], the authors describe a multiagent DQL strategy to handle the problem of throughput maximization. The proposed technique delivers better performance that is comparable to the fractional programming technique, although it needs discrimination of the power management parameters. The authors in [32] investigated the application of DQN, RL, and DDPG approaches to solve the sum throughput optimization problem. The DQL and RL algorithms need discretization of the power allocation parameters, which creates uncertainty because no known process for selecting the best discretization parameter exists. In [33], authors proposed a deep reinforcement learning approach to tackle the resource allocation problem for one-to-many DDCs underlaying cellular networks.

B. Motivation

In the above-mentioned works, it has been observed that the author's aim is to optimize the resources for D2D users underlaying cellular networks. Firstly, the authors from [11]- [24] used conventional algorithms to maximize the sum throughput of the overall network. Also, the authors want to mitigate cross and co-channel interference by optimizing the resources and power of the DUs and CUs. The solutions provided by these authors are not scalable due to the time-varying environment. Secondly, the authors from [26]- [32] aimed to maximize the sum throughput of the D2D network while maintaining the SINR of the CUs. To achieve the target, these authors applied DRL models such as DQN, deep dueling, and DDPG with the D2D network. Here, the continuous characteristics in space and action were discretized. The discretized in continuous space generates the quantization noise into state and action space, making it challenging for these models to attain the best policy. Moreover, in all these papers, the authors did not focus on the massive connectivity. To overcome all these issues, we integrated NOMA with DGs underlaying cellular networks in this paper. Additionally, we gather historical data to train the model rather than using it to make decisions. Here, each agent develops a generalizable model that allows it to make decisions based on actual state observations in real time. The states addressed the minimum SINR requirements, co-channel interference, and channel gain. To reduce the co-channel interference, we also optimize the power of the DTs. Furthermore, we applied the POPS to reduce the computation complexity of training by utilising the clipping substitute technique.

C. Contributions

To handle the challenges mentioned above, we examined the sum throughput maximization for NOMA-enabled DGs underlaying cellular networks. Firstly, the DDPG is engaged in learning the optimum policy from continuous and action space. Then, the POPS is used to decrease the computation complexity of training by utilizing the clipping substitute technique. The key contributions of the paper are described below:

- First of all, the joint channel scheduling and power control problem in NOMA-based DGs underlay cellular networks is formulated. The aim of the formulated problem is to maximize the sum throughput of the overall network while minimizing cross and co-channel interference.
- The formulated optimization problem is transformed into a machine learning (ML) form using the Markov decision process (MDP) model. Here, we consider that the DT in DGs acts as an agent that learns experiences in a trial-and-error way to acquire the best optimal policy, without having the complete information from a time-varying wireless environment.
- Now, to characterize the agent's behavior with respect to the environment we used continuous states and actions. Also, to achieve the best optimal policy, the DDPG technique is applied. The suggested approach blends value-based and determination-policy-based approaches to architecture while taking into account both actor and critic networks. Actor networks are responsible for creating deterministic actions, and critic networks are responsible for assessing actor networks.
- Lastly, POPS is applied that used the clipping substitute technique to reduce the complexity and difficulty that arise at the training stages.

D. Organization

The remainder of the paper is arranged as described. Section II elaborates on the system model and problem formulation. The proposed approach is discussed in Section III. The performance of the suggested system is evaluated, and its results are contrasted with those of the most recent state-of-the-art schemes, in Section IV. Finally, the conclusion of the paper is stated in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

A unicellular underlay cellular network with the base station (BS) denoted as b, $\mathcal{M}$ CUs, and $\mathcal{G}$ DGs in an uplink transmission scenario as shown in Fig. 1. The BS is in the center, while the $\mathcal{M} \in \{1, 2, \dots, m, \dots, M\}$ CUs and $\mathcal{G} \in \{1, 2, \dots, g, \dots, G\}$ DGs are evenly and randomly distributed around the cell. There is one DT and $\mathcal{D} \in \{1, 2, \dots d, \dots, D\}$ DUs in each DG. Since the velocity of DTs and DUs is very low, their positions are constantly changing. Let DTs be denoted as $\mathcal{G} \in \{1, 2, \dots, g, \dots, G\}$, implying that the number of DGs equals the number of DTs. To interact with the BS, $\mathcal{M}$ CUs and $\mathcal{G}$ DGs used OMA technique, and DT in each DG used the NOMA technique to deliver services to DUs. Consider that $\mathcal{K} \in \{1, 2, \dots, k, \dots, K\}$ RBs are allocated to CUs and DTs for communicating with the BS and corresponding DUs. Let B be the network's total bandwidth, which is divided equally into K RBs. Also, the quasi-static Rayleigh fading channel model is utilized where the constant coefficients of channels pursue a Gaussian complex distribution.

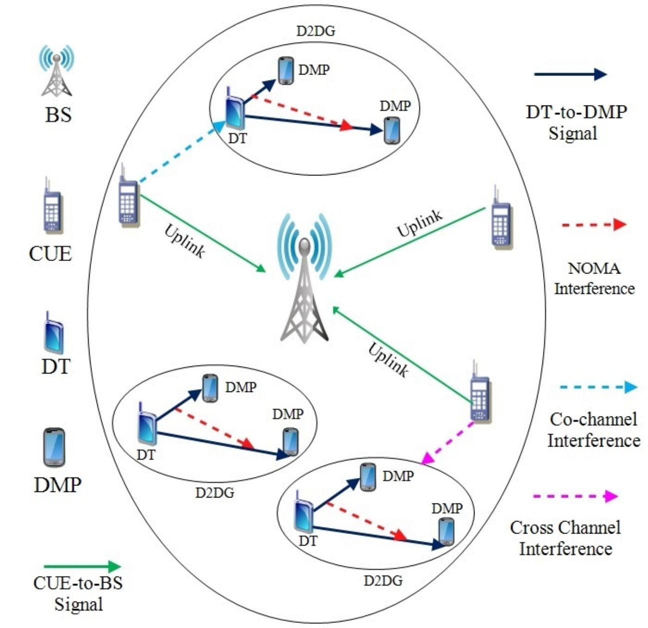

Fig. 1. System Model.

B. Channel Model

To develop a NOMA-based DG, each group must have at least two DUs. This condition can be shown as follows:

$$2 < \sum_{d=1}^{D} \lambda_{g,d}^{k} < D, \tag{1}$$

where D represents the total number of DUs in each DG, and $\lambda_{g,d}^{k}$ represents the DT-DUs group index, respectively. $\lambda_{g,d}^{k}$ can be expressed as follows:

$$\lambda_{g,d}^{k} = \begin{cases} 1, & \text{if the } d^{th} \text{ DU is associated with the } g^{th} \text{ DT,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

1) CU Channel Model: The m^{th} CU send a signal to the BS, which could be shown as below:

$$\mathbb{Y}_{m,b}^{k} = |h_{m,b}^{k}|\sqrt{P_m^k}x_{m,b}^{k} + \sum_{g=1}^{G}\lambda_{g,d}^{k}\theta_{g,m}^{k}|h_{g,b}^{k}|\sqrt{P_g^k}x_{g,b}^{k} + \zeta_{m,b}^{k}, \tag{3}$$

where

$$\theta_{g,m}^{k} = \begin{cases} 1, & \text{if the } m^{th} \text{ CU and the } g^{th} \text{ DT are scheduled} \\ & \text{across the } k^{th} \text{ RB,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In Eq.(3) P_m^k and P_g^k represents transmitted power of m^{th} CU, and g^{th} DT, $h_{m,b}^{k}$ represents channel gain between m^{th} CU and BS, $h_{g,b}^{k}$ represents channel gain between g^{th} DT and BS, $x_{m,b}^{k}$ and $x_{g,b}^{k}$ represents the transmitted message for m^{th} CU and g^{th} DT, respectively. $\zeta_{m,b}^{k}$ is additive white Gaussian noise (AWGN) with mean = 0 and variance $\sigma = 1$.

Now, using Eq. (3) , the SINR is determined as follows:

$$\gamma_{m,b}^{k} = \frac{P_m^k|h_{m,b}^{k}|^2}{IF_{DG} + \sigma_{m,b}^2}, \tag{5}$$

where IF_{DG} represents the interference caused by DGs and defined as follows:

$$IF_{DG} = \sum_{g=1}^{G}\lambda_{g,d}^{k}\theta_{g,m}^{k}P_g^k|h_{g,b}^{k}|^2. \tag{6}$$

Defining $|h_{m,b}^{k}|^2 = |\widehat{h}_{m,b}^{k}|^2 z_{m,b}^{-\beta}$, and $|h_{g,b}^{k}|^2 = |\widehat{h}_{g,b}^{k}|^2 z_{g,b}^{-\beta}$. $\widehat{h}_{m,b}^{k}$ and $\widehat{h}_{g,b}^{k}$ represent small scale fading along with $\hat{h}_{g,b}^{k} \sim \mathcal{CN}(0,1)$ and $\hat{h}_{g,b}^{k} \sim \mathcal{CN}(0,1)$, respectively. The distance from the m^{th} CU to b and from the g^{th} DT to b is denoted by $z_{m,b}$ and $z_{g,b}$, respectively. The path loss exponent is β.

2) DUs Channel Model: For multiplexing of power signals, DT uses superposition coding in the PD-NOMA approach. On the other hand, DUs use SIC, resulting in a reduction of interference (intra-user). The signal sent by DT via superposition coding to the $\mathcal{D}$ DUs having various power allocation factors is now as follows:

$$\eta_{g,1}x_{g,1}^{k} + \eta_{g,2}x_{g,2}^{k} + \cdots + \eta_{g,d_s}x_{g,d_s}^{k} + \eta_{g,d_w}x_{g,d_w}^{k} \cdots + \eta_{g,D}x_{g,D}^{k}, \tag{7}$$

where $\{\eta_1, \eta_2, \ldots, \eta_D\}$ and $\{x_1, x_2, \ldots, x_r\}$ is the power allocation factor and messages for DUs. The strongest and weakest DU related to the DT are denoted by d_s and d_w, respectively. According to [4], [9], the signal received at d_w DUs in the g^{th} DG across the k^{th} RB is given as follows:

$$\begin{aligned} \mathbb{Y}_{g,d_w}^{k} &= \sqrt{P_g\eta_{g,d_w}}|h_{g,d_w}^{k}|x_{g,d_w}^{k} + \sum_{d=2}^{D}\sqrt{P_g\eta_{g,d_s}}|h_{g,d_s}^{k}|x_{g,d_s}^{k} \\ &+ \sum_{g' \neq g}^{G}\theta_{g',g}^{k}\sqrt{P_{g'}}|h_{g',g,d_w}^{k}|x_{g',g,d_w}^{k} \\ &+ \lambda_{g,d}^{k}\sqrt{P_m}|h_{m,g,d_w}^{k}|x_{m,g,d_w}^{k} + \zeta_{g,d_w}^{k}, \end{aligned} \tag{8}$$

where ζ_{g,d_w}^{k} is the AWGN, and $\theta_{g',g}^{k}$ denotes interference (co-channel), that is expressed as:

$$\theta_{g',g}^{k} = \begin{cases} 1, & \text{if the } (g')^{th} \text{ and the } g^{th} \text{ DT are scheduled} \\ & \text{across the } k^{th} \text{ RB,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Defining $|h_{g,d_w}^{k}| = |\widehat{h}_{g,d_w}^{k}|z_{g,d_w}^{-\beta}$, $|h_{g,d_s}^{k}| = |\widehat{h}_{g,d_s}^{k}|z_{g,d_s}^{-\beta}$, $|h_{g',g,d_w}^{k}| = |\widehat{h}_{g',g,d_w}^{k}|z_{g',g,d_w}^{-\beta}$,and $|h_{m,g,d_w}^{k}|=|\widehat{h}_{m,g,d_w}^{k}|z_{m,g,d_w}^{-\beta}$. Here, $\widehat{h}_{g,d_w}^{k} \sim \mathcal{CN}(0,1)$, $\widehat{h}_{g,d_s}^{k} \sim \mathcal{CN}(0,1)$, $\widehat{h}_{g',g,d_w}^{k} \sim \mathcal{CN}(0,1)$ and $\widehat{h}_{m,g,d_w}^{k} \sim \mathcal{CN}(0,1)$. The distance from the g^{th} DG to the d_w DU and from the $(g')^{th}$ DG to d_w DU in the g^{th} DG are denoted by $z_{g,d_w}^{-\beta}$ and $z_{g',g,d_w}^{-\beta}$, respectively. Similarly, the distance from the g^{th} DG to the d_s DU and from the m^{th} CU to d_w DU in the g^{th} DG are denoted by g, d_s and $z_{m,g,d_w}^{-\beta}$, respectively.

Now, using Eq. (8) , the SINR for d_s DU is determined as follows:

$$\gamma_{g,d_s}^{k} = \frac{|h_{g,d_s}^{k}|^2 P_g\eta_{g,d_s}}{IF_{g,d_s}^{NO} + IF_{g,d_s}^{CO} + IF_{g,d_s}^{CR} + \sigma^2}, \tag{10}$$

where IF^{NO}_{g,d_s} is the NOMA interference and expressed as follows:

$$IF^{NO}_{g,d_s} = \sum_{d=2}^{D} \sqrt{P_g \eta_{g,d_s}} |h^k_{g,d_s}| x^k_{g,d_s} \tag{11}$$

IF^{CO}_{g,d_s} is the co-channel interference and expressed as follows:

$$IF^{CO}_{g,d_s} = \sum_{g' \neq g}^{G} \theta^k_{g',g} \sqrt{P_{g'}} |h^k_{g',g,d_w}| x^k_{g',g,d_w}. \tag{12}$$

IF^{CR}_{g,d_s} is the cross-channel interference and expressed as follows:

$$IF^{CR}_{g,d_s} = \lambda^k_{g,d} \theta^k_{g,m} \sqrt{P_m} |h^k_{m,g,d_w}| x^k_{m,g,d_w}. \tag{13}$$

Now, $\gamma_{(d^{th} DU)} \geq \gamma_{(d' \neq d DUs)}$ for efficient SIC, i.e.

$$\frac{|h^k_{g,d_s}|^2 P_g \eta_{g,d_s}}{IF^{NO}_{g,d_s} + IF^{CO}_{g,d_s} + IF^{CR}_{g,d_s} + \sigma^2} \geq \sum_{d=2}^{D} \frac{|h^k_{g,d_w}|^2 P_d \eta_{g,d_w}}{IF^{NO}_{g,d_w} + IF^{CO}_{g,d_w} + IF^{CR}_{g,d_w} + \sigma^2}. \tag{14}$$

So, Eq. (14) can be reformulated as follows:

$$\Delta(\theta) = |h^k_{g,d_s}|^2 (IF^{NO}_{g,d_w} + IF^{CO}_{g,d_w} + IF^{CR}_{g,d_w} + \sigma^2). - \sum_{d_w=2}^{D} |h^k_{g,d_w}|^2 (IF^{d,NO}_{g,d_s} + IF^{CO}_{g,d_s} + IF^{CR}_{g,d_s} + \sigma^2) \geq 0 \tag{15}$$

To decrypt its own data, d_s DU first performs a successful SIC, and then it eliminates interference from d_w DUs. As a result, using Eq. (8), the received SINR at the d_s DU is represented as follows:

$$\gamma^k_{g,d_s} = \left(\frac{|h^k_{g,d_s}|^2 P_g \eta_{g,d_s}}{IF^{CO}_{g,d_s} + IF^{CR}_{g,d_s} + \sigma^2} \right). \tag{16}$$

Accordingly, for the rest DUs, the SINR is computed as follows:

$$\gamma^k_{g,d_w} = \sum_{d=2}^{D} \left(\frac{|h^k_{g,d_w}|^2 P_g \eta_{g,d_w}}{IF^{d,NO}_{g,d_w} + IF^{CO}_{g,d_w} + IF^{CR}_{g,d_w} + \sigma^2} \right). \tag{17}$$

C. Data Rate and Overall Sum Throughput Estimation

As per the Shannon capacity theorem, m^{th} CU's throughput considering Eq. (5) is defined as follows:

$$\mathbb{T}^k_{m,b} = \log_2(1 + \gamma^k_{m,b}). \tag{18}$$

Accordingly, the throughputs of DUs using Eq. (16) and Eq. (17) are represented as:

$$\mathbb{T}^k_{g,d_s} = (1 + \gamma^k_{g,d_s}). \tag{19}$$

$$\mathbb{T}^k_{g,d_w} = \sum_{d_w=2}^{D} (1 + \gamma^k_{g,d_w}). \tag{20}$$

The sum throughput of $\mathcal{M}$ CUs and $\mathcal{G}$ DGs with $\mathcal{D}$ DUs is expressed as:

$$\mathbb{ST}^k_m = \sum_{m=1}^{M} \log_2(1 + \gamma^k_{m,b}). \tag{21}$$

$$\mathbb{ST}^k_g = \sum_{g=1}^{G} \left(\lambda^k_{g,d} \theta_{g,m} (R^k_{g,d_s} + \sum_{d_w=2}^{D} R^k_{g,d_w}) \right). \tag{22}$$

Now, the total sum-rate is computed as follows:

$$\mathbb{ST}^k_{m+g} = \sum_{m=1}^{M} \left(R^k_{m,b} + \sum_{g=1}^{G} \lambda^k_{g,d} \theta_{g,m} (R^k_{g,d_s} + \sum_{d_w=2}^{D} R^k_{g,d_w}) \right). \tag{23}$$

D. Problem Formulation

The main aim of this paper is to achieve maximum overall network's sum throughput while retaining the SINR of CUs and DGs. Therefore, the sum-rate optimisation problem can be described as follows:

$$\begin{array}{rcll}
P.F. & : & \max\limits_{\theta,P} \mathbb{ST}^k_{m+g}, & \quad (24) \\
s.t.\ \mathbb{V}_1 & : & 2 < \sum\limits_{d=1}^{D} \lambda^k_{g,d} < D, & \forall \mathcal{G}, \\
\mathbb{V}_2 & : & \sum\limits_{g=1}^{G} \lambda^k_{g,d} \leq 1, & \forall \mathcal{G}, \\
\mathbb{V}_3 & : & \mathbb{R}^k_{m,b} \geq \mathbb{R}^{k,\min}_{m,b}, & \forall \mathcal{M}, \mathcal{K}, \\
\mathbb{V}_4 & : & \sum\limits_{g=1}^{G} \theta_{g,m} p_g |h^k_{g,m}|^2 \leq IF^{th}_m, & \forall \mathcal{M}, \\
\mathbb{V}_5 & : & \Delta(\theta) \geq 0, & \forall \mathcal{G}, \\
\mathbb{V}_6 & : & \mathbb{R}^k_{g,d_s} \geq \mathbb{R}^{\min}_{g,d_s}, \mathbb{R}^k_{g,d_w} \geq \mathbb{R}^{\min}_{g,d_w}, & \forall \mathcal{K}, \\
\mathbb{V}_7 & : & \theta^k_{g,m}, \theta^k_{g',g} \in \{0,1\}, & \forall \mathcal{M}, \mathcal{G}, \mathcal{K}, \\
\mathbb{V}_8 & : & P^k_m \leq P^{k,\max}_m, & \forall \mathcal{M}, \\
\mathbb{V}_9 & : & \eta_{g,d_s} + \sum\limits_{d_w=2}^{D} \eta_{g,d_w} \leq 1, & \forall \mathcal{G}, \mathcal{D}, \\
\mathbb{V}_{10} & : & \eta_{g,d_s} \geq 0, \sum\limits_{d_w=2}^{D} \eta_{g,d_w} \geq 0, & \forall \mathcal{G}, \mathcal{D}. \\
\mathbb{V}_{11} & : & P^k_m \geq 0, & \forall m \in \mathcal{M} \\
\mathbb{V}_{12} & : & P^k_g \geq 0, & \forall g \in \mathcal{G}
\end{array}$$

$\mathbb{V}_1$ assures that each DT provides service to at least two DUs to accomplish downlink NOMA. $\mathbb{V}_2$ assures that each DU must be linked to one DT. CU's minimal data rate demand is denoted by $\mathbb{V}_3$. The $\mathbb{V}_4$ defines the combined interference threshold for CUs. $\mathbb{V}_5$ specifies the effective SIC, whereas $\mathbb{V}_6$ specifies the minimal SINR need of d_s and d_w DUs for each DG. $\mathbb{V}_7$ guarantees that each RB can only be used by one CU and one DG at a time. The $\mathbb{C}_8$ and $\mathbb{C}_9$ imply that $S_{\max}$ shares the same RB and $S_{\max}$ is the maximal number of DGs and $T_{\max}$ is the maximal number of RBs that could be utilized again by the constraints of DGs. The major limitation of the transmitted power of CUs and DTs is ensured by $\mathbb{V}_8$ and $\mathbb{V}_9$. $\mathbb{V}_{10}$ assures that transmission power must be positive. $\mathbb{V}_{11}$ and $\mathbb{V}_{12}$ specify that the power of CUs and DTs must be a positive integers.

III. Proposed Scheme

In this section, first of all, the optimization problem formulated in (24) is transformed into RL form using the MDP concept. After that, the DDPG is used to lower the computing burden in order to find the best policy. Also, it improves the training process stability. Finally, the POPS is applied to reduce the computation complexity of the training process by using the clipping substitute technique.

A. Model for Markov Decisions Process

Let the MDP consists of five tuples as $(\mathbb{S}, \mathbb{A}, \mathbb{P}, \digamma, \Gamma)$. In MDP, tuple $\mathbb{S}$ stands for a set of states, tuple $\mathbb{A}$ stands for actions, tuple $\mathbb{P}$ stands for mapping association between states s and s', tuple $\digamma$ stands for rewards, and tuple $\Gamma \in [0, 1)$ stands for the discount factor. The MDP model's detailed description is as follows:

1) Agent: In the proposed system model each DT is an agent and there are G such agents.

2) State Space: The agent (DT) examines the state in order to characterise the environment, which is comprised of various parts: channel gain of different links, link's previous interference, and D2D link's RB. Also, DT optimize resource allocation by exploiting local information and previous non-local information. The state space with respect to environmental parameters is defined as follows:

$$s^k_{g,t} =$$

$$\left[h^k_{m,b}, h^k_{g,b}, h^k_{g,d_w}, h^k_{g,d_s}, IF^{NO}_{g,d_s}, IF^{CO}_{g,d_s}, IF^{CR}_{g,d_s}, R_g\right]. \quad (25)$$

The state space can be given as $S = \{s^k_{g,t} | g = 1, \ldots 2, \ldots G\}$.

3) Action Space: Depending on the current state and decision policy, the g^{th} agent takes an action $a^k_g \in \mathbb{A}$. The agent's overall action can be defined as follows:

$$a^k_{g,t} =$$

$$\left[(\theta^k_{1,m}, \theta^k_{g,m}, \theta^k_{G,m}); (P^k_1, P^k_m, \ldots, P^k_M); (P^k_1, P^k_g, \ldots, P^k_G)\right]. \quad (26)$$

Agent executes the action $a^k_{g,t}$ in the state $s^k_{g,t}$. Agent goes to the next state $s^k_{g,t+1}$ after executing the action $a^k_{g,t}$. $A = \{a^k_{g,t} | g = 1, \ldots 2, \ldots G\}$ is used for representing the action space.

4) Reward Function: The reward function in RL techniques control the training process. With the help of the environment's interactions, each agent decides how to maximize its reward. The reward function is defined as:

$$\digamma = \begin{cases} \mathbb{ST}^k_{m+g}, & \text{if constraints satisfied,} \\ \digamma_{neg}, & \text{otherwise.} \end{cases} \quad (27)$$

The Eqn.(27) shows that the reward is positive, i.e., $\mathbb{ST}^k_{m+g}$ for an agent when it satisfies all the constraints, otherwise it is negative, which is a penalty for the DGs.

B. Multi-Agent Deep Deterministic Policy Gradient (MAD2PG) Scheme

We used the DDPG [31] for obtaining the best policy for the reward described in (27). The benefits of actor-critic (AC), deterministic policy gradient (DPG), and DQN are integrated into the DDPG. The actor-critic concept is used in DDPG and is integrated with the DPG to lower the computing burden on the agent in order to find the best policy. The configuration in multi-agent DDPG (MAD2PG) is done by an AC network, and each of them consists of two DNNs. Moreover, we settle the experience replay buffer for gathering experiences, so as the reduction the relevance of sample data. Moreover, the MAD2PG algorithm improves the training process stability by including DQN's network with the AC model.

The DDPG algorithm improves the training process stability by including DQN's network model for the actor and critic models as well as the updating of policy and Q networks, which are defined below:

1) Value Function: When an agent moves from a random state s to a state $S(T)$, the predicted return following the policy π is known as the state-value function and is defined as:

$$\mathbb{V}_\pi(s) = E_\pi\left[\sum_{t=1}^{T} \Gamma^t \digamma(t) | s_0 = 0\right], \quad (28)$$

where s_0 represents the initial state and Γ is used to mitigate the effect of future awards on the present one. Now, for the previous state s and action a, the state value function formulates the predicted return for a policy and is defined as:

$$\mathbb{Q}_\pi(s, a) = \mathbb{E}_\pi\big[\digamma(t) + \Gamma \mathbb{V}_\pi\big(s(t+1)\big)\big]. \quad (29)$$

The $\mathbb{Q}_\pi(s, a)$ can be transformed to the Bellman equation as per MDP [34].

$$\mathbb{Q}_\pi(s, a) = \mathbb{E}\big[\digamma(t) + \Gamma \mathbb{E}\big[\mathbb{Q}_\pi\big(s(t+1), a(t+1)\big)\big]\big]. \quad (30)$$

Consider a Performance Objective Function (POF), which is described on the basis of performance under the policy π and given as:

$$\begin{aligned} \mathbb{J}(\pi) &= \mathbb{E}_\pi\big(\mathbb{Q}_\pi(s, a)\big) \\ &= \int_{\mathbb{S}} D_\pi(s) \int_{\mathbb{A}} \pi(a|s) \mathbb{Q}_\pi(s, a) d_a d_s. \end{aligned} \quad (31)$$

2) Actor Method: The policy gradient approach is used in the actor method to estimate and enhance parametric policies, with the gradient ascent approach supports the improvement of policies and updating the specification of the policy network in an iterative manner. The actor framework consists of two DNNs: an Online-Policy (OP) network and a Target Policy (TP) network. Let $\phi^{\varpi} = \big[\phi_1, \phi_2, \ldots, \phi_n\big]$ denotes OP network's parameters, $\phi^{\varpi'} = \big[\phi_1^{'}, \phi_2^{'}, \ldots, \phi_n^{'}\big]$ denotes TP network's parameters and $\varpi_\phi(s, a)$ denotes policy. At each time slot t, an action is generated by the OP network, which depends on DPG. The agent then collects state vectors from the environment and is represented as:

$$a(t) = \varpi\big(s(t) | \phi^{\varpi'}\big) + \mathcal{W}_N, \quad (32)$$

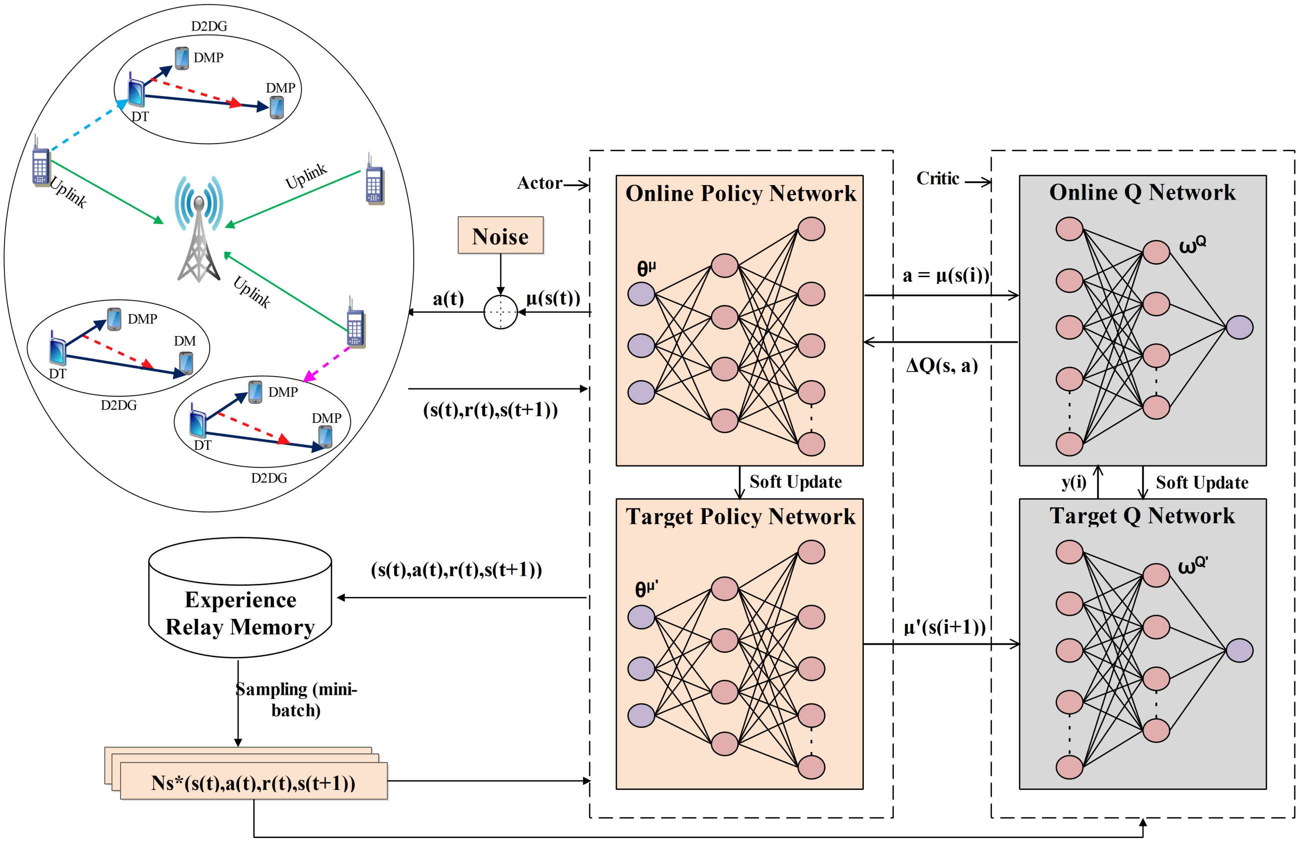

Fig. 2. DQN Model.

where $\mathcal{W}_N \in (0,1)$ is white noise, which is added to make exploration process easier in continuous action space. Using Eq.(32) POF is reformulated as follows [35]:

$$\begin{aligned}\mathbb{J}(\varpi_\phi) &= \int_{\mathbb{S}} D_{\varpi_\phi}(s)\mathbb{Q}\big((s,\varpi_\phi(s))\big)ds\\ &= \mathbb{E}_{\varpi_\phi}\big[\mathbb{Q}\big(s,\varpi_\phi(s)\big)\big].\end{aligned} \tag{33}$$

The actor calculates the gradient of the POF in order to improve policy. The DPG of Eq. (33) according to [17] is calculated as follows:

$$\begin{aligned}\nabla_\phi\mathbb{J}(\varpi_\phi) &= \int_{\mathbb{S}} D_{\varpi_\phi}(s)\nabla_\phi\varpi_\phi(s)\nabla_\phi\mathbb{Q}\big((s,a)\big|_{a=\varpi_\phi}ds\\ &= \mathbb{E}_{\varpi_\phi}\big[\nabla_\phi\varpi_\phi(s)\nabla_\phi\mathbb{Q}\big(s,a\big)\big|_{a=\varpi_\phi}ds\big].\end{aligned} \tag{34}$$

The training set is considered to be independent and uniformly distributed when employing neural networks to estimate the state value function. But, the data gathered by the agent frequently contains a high degree of correlation, making the RL model is unstable if such data is considered for training. Experience replay technique has the ability to break the correlation between the collected data.

Let experience at time slot t is represented by $\big(s^k_{g,t}, a^k_{g,t}, F^k_{g,t}, s^k_{g,t+1}\big)$ and it is recorded in the experience replay buffer using a bounded storage capacity of U. The buffer updates the experience on a regular basis by gathering new samples and removing old ones. The Monte-carlo approach [32] is used to compute the prediction in the replay buffer by randomly sampling mini-batch of capacity V. Therefore, Eq. (43) is reformulated as:

$$\begin{aligned}&\nabla_\phi\mathbb{J}(\varpi_\phi)\\ &\approx \frac{1}{V}\sum_{v=1}^{V}\big(\nabla_\phi\mathbb{Q}(s,a)\big)\big|_{s=s^k_{e,t},a=\varpi(s^k_{e,t})}\nabla_\phi\varpi_\phi(s)\big|_{s=s^k_{e,t}},\end{aligned} \tag{35}$$

where $\mathbb{Q}(s,a)$ represents the state value function created by the critic network.

C. Critic Method

The actor network's performance is evaluated by the critic network. It consists of two DNN networks similar to the actor network: the critic Q network and the target Q' network. Conventional-table based RL techniques perform well when state space is small and action space is discrete. However, the value function approximation process estimates the value function using a set of parameters. It reduces computation complexity, dimension of input samples, enhances generality and eliminates over-fitting. DNN is applied in the critic network for estimation of the value function and is defined as $\mathbb{Q}_\psi(s,a) \approx \mathbb{Q}(s,a)$.Let $\psi^Q = \big[\psi_1, \psi_2, \ldots, \psi_n\big]$ denotes critic

Q network's parameters, $\psi^{Q'} = [\psi'_1, \psi'_2, \ldots, \psi'_n]$ denotes target Q' network's parameters. Now mini-batch of e^{th} transition samples, i.e., $s(e)$ and $a(i)$, is extracted from the replay buffer in critic Q network. Then these samples are given to the DNN to evaluate the Q value $Q\big(s(e), a(e)|_{\psi Q}\big)$. Concurrently in the target Q' network, the mini-batch sample $\digamma(e)$ and $s(e+1)$ are given to DNN to create the target Q' value $y(e)$, which is determined as follows:

$$y(e) = \digamma(e)\Gamma Q'\big(s(e+1), \varpi'\big(s(e+1)|\psi^{Q'}\big)\big), \tag{36}$$

where $\varpi'\big(s(e+1)|\psi^{Q'}$ indicates the estimation of $a(t+1)$ and is generated by the target actor network ϖ'.

The loss function must be minimized for each learning step in order to modify the critic network and defined as follows:

$$\mathbb{L}_f \cong \frac{1}{V} \sum_{v=1}^{V} \big(y(e) - \mathbb{Q}\big(s(e), a(e)\big)|_{\phi Q}\big)^2 \tag{37}$$

D. The Network Update Procedure

The learning method of model-free RL approaches is relies on policy iterative process. policy iterative process is categorized into two stages, known as policy assessment and policy improvement. In policy assessment, Monte-Carlo evaluation and temporal difference training methods are used to estimate action-value functions.

On the other hand greedy policy is used in policy improvement to optimize the value of the action function. But optimization of the action value function is not possible in policy evaluation because of infinite states and action values in continuous space. Therefore, the value of the action function is estimated with the use of DNNs, and policy is improved by modifying the DNN policy functions' parameters. The OP network's DNN parameters are modified as follows:

$$\phi(t+1) = \phi(t) + \chi_\phi \nabla_\phi \varpi_\phi\big(s(t), a(t)\big)\big|_{a=\varpi_\phi(s(t))}, \tag{38}$$

where OP network's learning rate is denoted by χ.

In contrast to OP networks, online Q networks update their parameters using a gradient-based approach and single-step temporal difference (TD) error. In comparison to the Monte-Carlo approach, the TD method is more efficient. The Monte-Carlo only modifies the value function once in each episode. However, the TD method integrates Monte-Carlo using dynamic programming to improve the performance. The TD error is now calculated by taking the deviation from the target Q value and is defined as follows:

$$\delta(t) = \digamma(t) + \Gamma\mathbb{Q}^{\psi'}\big(s(t+1), \varpi_{\phi'}\big(s(t+1)\big)\big) - \mathbb{Q}^{\psi}\big(s(t) - a(t)\big). \tag{39}$$

In Eq. (48) $\mathbb{Q}^{\psi'}\big(s(t+1), \varpi_{\phi'}\big(s(t+1)\big)\big)$ is the target $\mathbb{Q}$ value and $\mathbb{Q}^{\psi}\big(s(t) - a(t)\big)$ is the estimated online $\mathbb{Q}$ value. The updated parameters are define as follows:

$$\psi(t+1) = \psi(t) + \chi_\psi \delta(t) \nabla_\phi \mathbb{Q}^{\psi}\big(s(t), a(t)\big). \tag{40}$$

Now, Soft-update technique is used to modify the parameter ϕ^{ϖ} and ψ^{Q}. Therefore updated parameters are defined as follows:

$$\phi^{\varpi'} \leftarrow \alpha\phi^{\varpi} + (1-\alpha)\phi^{\varpi} \tag{41}$$

$$\psi^{Q'} \leftarrow \alpha\psi^{Q} + (1-\alpha)\psi^{Q}, \tag{42}$$

where α is the soft updated step size.

The description of the MAD2PG scheme is shown in the Algorithm 1. As the BS has a larger processing capacity than the CUs and DUs, the training part of the algorithm is finished there, and users need to download the weights of the trained target actor network $\phi^{\varpi'}$ from the BS in order to utilize the actor component of the algorithm to disperse its execution. In order to train the DNNs for the actor part and critic portion, the MA-DDPG technique leverages historical data, and it then returns the target actor network's weights. When the algorithm is run, fresh information is produced, which may be added to the experience replay buffer V to further adjust weights.

Algorithm 1 DDPG Algorithm for Sum Throughput Network.

Input

- Environment: (a) DGs,DUs and CUs (b) BS with OFDMA scheme(uplink).
- OP network's learning rate: χ_ϕ
- Online Q network's learning rate: χ_ψ
- Soft updated step size: α

Initialization:

- OP network: ϕ^{ϖ}.
- TP network: $\phi^{\varpi'}$.
- Critic Q network: $\psi^{Q'}$.
- Target Q network: $\psi^{Q'}$.
- Replay buffer: V
- Maximum Episodes: Φ.
- Time Slot: T.

```
for episode = 1,...,Φ do
    Initialize the state $s_0$ by generating QoS demands for all users (CUs and DUs) randomly.
    for iteration = 1,..., $\mathbb{T}$ do
        Choose action a(t) in accordance with (320.
        Allocate $\lambda^k_{g,d}$ according to (2).
        Allocate $\theta^k_{g,m}$ according to (4).
        Calculate sum-rate of $\mathcal{M}$ CUs according to (21).
        Calculate sum-rate of $\mathcal{M}$ DGs according to (22).
        Formulate the total sum throughput as in (23) to generate the reward function $\digamma$.
        Collect all UEs' QoS requirements and compute the next state, $s^k_{g,t}(t+1)$.
        Store transition $\big(s^k_{g,t}, a^k_{g,t}, \digamma^k_{g,t}, s^k_{g,t+1}\big)$ in experience replay buffer of capacity $U$
        Select randomly a small batch of transitions $(s^k_e, a^k_e, \digamma^k_e, s^k_{e+1})$ from the replay buffer of capacity $V$.
        Set the value of $y(e)$ according to (36).
        Update $\mathbb{L}_f$ according to (37).
        Update POF in actor netwok according to (35).
        Renew $\phi^{\varpi'}$ according to (41)
        Renew $\psi^{Q'}$ according to (42).
        Renew the state $s^k_e = s^k_{e+1}$
        if (Present state s(t) = (1.0,1.0,...,1.0 )) then,
            break
        end if
    end for
end for
Output: $\alpha$
```

E. Proximal Online Policy Scheme

To boost performance, we used the proximal online policy scheme (POPS) in this section. Here, the present and past policies are compared in order to maximize the objective function, which is provided as:

$$\mathcal{F}(\mathtt{s},\mathtt{a},\Phi_{)} = \mathscr{E}\left[\frac{\Pi(\mathtt{s},\mathtt{a},\Phi)}{\Pi(\mathtt{s},\mathtt{a},\Phi_{old})}\right]\Upsilon^{\Pi}(\mathtt{s},\mathtt{a})$$

$$\mathcal{F}(\mathtt{s},\mathtt{a},\Phi_{)} = \mathscr{E}\mathbb{P}^{t}_{\Phi}\Upsilon^{\Pi}(\mathtt{s},\mathtt{a}), \tag{43}$$

where $\mathbb{P}^{t}_{\Phi}$ signifies probability ratio and $\Upsilon^{\Pi}(\mathtt{s},\mathtt{a}) = \mathbb{O}^{\Phi}(\mathtt{s},\mathtt{a}) - \mathcal{V}^{\Pi}(\mathtt{s})$ is the approximates advance function [36].

SGD is used to maximize the objective of training networks with a mini-batch Ω. As a result, the policy is updated with the assistance of

$$\Phi^{t+1} = \arg\max \mathscr{E}\left[\mathcal{F}(\mathtt{s},\mathtt{a};\Phi^{t})\right] \tag{44}$$

In order to limit the objective value in this method, we utilize the clip function $(f^{t}_{\Pi}, 1-\Theta, 1+\Theta)$ as follows:

$$\mathcal{F}^{clip}(\mathtt{s},\mathtt{a};\Phi) = \mathscr{E}\left[\min(f^{t}_{\Phi}, \Upsilon^{\Pi}(\mathtt{s},\mathtt{a}), clip(f^{t}_{\Phi}, 1-\Theta, 1+\Theta)\Upsilon^{\Pi}(\mathtt{s},\mathtt{a})\right], \tag{45}$$

where Θ is a low value constant.

If $\Upsilon^{\Pi}(\mathtt{s},\mathtt{a})$ becomes greater than zero with respect to the upper bound $1+\Theta$, then the objective is reformulated as follows:

$$\mathcal{F}^{clip}(\mathtt{s},\mathtt{a};\Phi) = \min\left[\frac{\Pi(\mathtt{s},\mathtt{a};\Phi)}{\Pi(\mathtt{s},\mathtt{a};\phi_{old})}, (1+\Theta)\right]\Upsilon^{\Pi}(\mathtt{s},\mathtt{a}). \tag{46}$$

In (46), the objective's value rises if the advantage $\Upsilon^{\Pi}(\mathtt{s},\mathtt{a})$ becomes greater than zero. The minimal term, however, limits the rising value. When $(1+\Theta)\Pi(\mathtt{s},\mathtt{a};\Phi_{(old)})$ $\Pi(\mathtt{s},\mathtt{a};\Phi) > (1+\Theta)\Upsilon^{\Pi}(\mathtt{s},\mathtt{a};\Phi)$, the objective value is limited by factor $(1+\Theta)\Upsilon^{\Pi}(\mathtt{s},\mathtt{a};\Phi)$.

On the other side, when $\Upsilon^{\Pi}(\mathtt{s},\mathtt{a})$ becomes less than zero at the lower limit $1-\Theta$. In this situation, the objective is redefined as follows:

$$\mathcal{F}^{clip}(\mathtt{s},\mathtt{a};\Phi) = \max\left[\frac{\Pi(\mathtt{s},\mathtt{a};\Phi)}{\Pi(\mathtt{s},\mathtt{a};\phi_{old})}, (1-\Theta)\right]\Upsilon^{\Pi}(\mathtt{s},\mathtt{a}). \tag{47}$$

Similarly, when the advantage $\Upsilon^{\Pi}(\mathtt{s},\mathtt{a})$ falls below zero in (47), the objective value decreases. However, the decreased value is constrained by the maximum term when $\Pi(\mathtt{s},\mathtt{a};\Phi)$ $< (1-\Theta)\Pi(\mathtt{s},\mathtt{a};\Phi_{(old)})$, then factor $(1-\Theta)\Upsilon^{\Pi}(\mathtt{s},\mathtt{a})$ restricts the objective value.

The objective is constrained by the minimum and maximum terms of (46) and (47), ensuring that the new tactic doesn't stray from the previous one. Consequently, the following definition of "advantage function" is defined as follows [31]:

$$\Upsilon^{\Pi}(\mathtt{s},\mathtt{a}) = \mathtt{r}^{t} + [\mathcal{V}^{\Pi}(\mathtt{s}^{t+1}) - \mathcal{V}^{\Pi}(\mathtt{s}^{t})]. \tag{48}$$

Now, the policy is trained and the parameters are updated using a mini-batch Ω as follows:

$$\Phi^{t+1}_{i} = \arg\max_{\Phi_{\Pi}} \mathscr{E}\left[\mathcal{F}^{clip}(\mathtt{s},\mathtt{a};\Phi^{t})\right] \tag{49}$$

Algorithm 2 POPS Based Throughput Enhancement Algorithm.

```
Initialization:
  • $\Pi$ = Policy with parameter $\Phi$
  • $\Theta$ = Penalty Parameter.
Output: $\theta(t), Q(t)$
for ($b = 1, b \leq \mathbb{B}, b++$) do
    Set an initial state as $s^0$
    for ($\psi = 1, \psi \leq \Psi, \psi++$) do
        Perform the action $\mathtt{a}^t$ achieved at state $\mathtt{s}^t$
        Modify the reward $\mathtt{r}^t$ in accordance with (26)
        Check the next state $s^{t+1}$
        Update the state $s^t = s^{t+1}$
        Collect a set of partial trajectories with $\mathscr{E}$ transitions
        Calculate the advantage function using (50)
    end for
    Modify policy parameters using SGD & mini-batch $\mathscr{E}$ using (51)
end for
```

IV. Performance Evaluation

The performance of the proposed approach is estimated and discussed in this section. It comprises three sections: (i) Numerical Settings (ii) Baseline Schemes (iii) Results and Discussions

A. Numerical Settings

1) Simulation Parameters of D2D Underlaying Cellular networks: The BS b is assumed to be deployed at centre of the cell, and $\mathcal{M}$ CUs and $\mathcal{G}$ DGs are evenly and randomly distributed around the BS. The radius of the cell and DGs are set to be 500 m and 50 m, respectively [37]. The DT used NOMA protocol to communicate with the DUs, and on the other hand, the CUs used OMA to transmit data to the BS. The carrier frequency and BS transmission power are set to be 5MHz and 5W, respectively. The details of the remaining parameters are shown in Table 1 [38], [39].

2) Simulation Parameters DQN Model: The DQN training model is built on a fully connected neural network. This network contains an input layer, a hidden layer, and an output layer. Each of the three connected layers uses 500, 500, and 250 neurons respectively. The proposed model uses the ReLu as an activation function and the adaptive moment as an optimizer. With the POPS approach, we apply a learning rate of lr = 0.00001. Tensorflow 2.5 on Python 5 is utilized to simulate the model [40]. The final simulation settings are listed in Table I.

B. State-of-art Schemes

For analysing the proposed scheme's performance, we contrasted our proposed scheme with three algorithms which are described as follows:

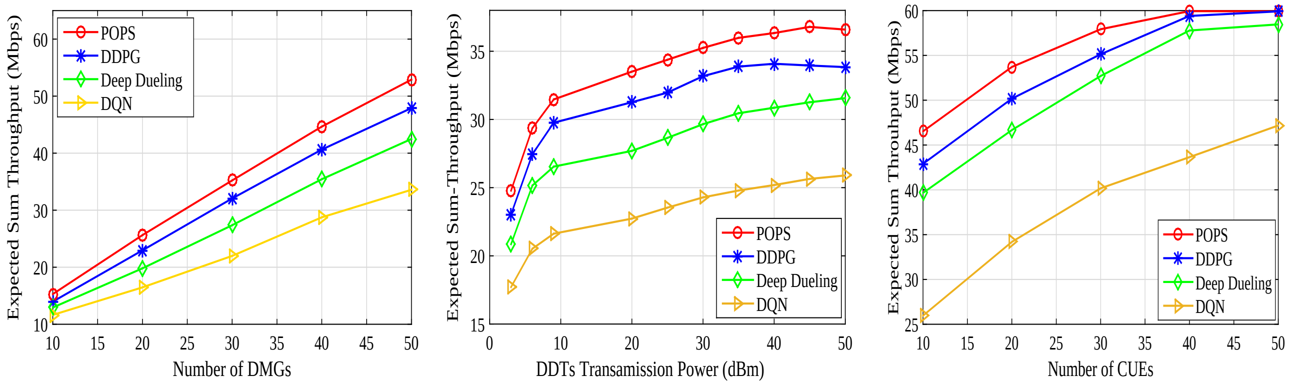

Fig. 3. Comparative Analysis (a) Expected Sum Throughput v/s Number of DGs (b) Expected Sum Throughput v/s DTs Transmission Power (c) Expected Sum Throughput v/s Number of CUs.

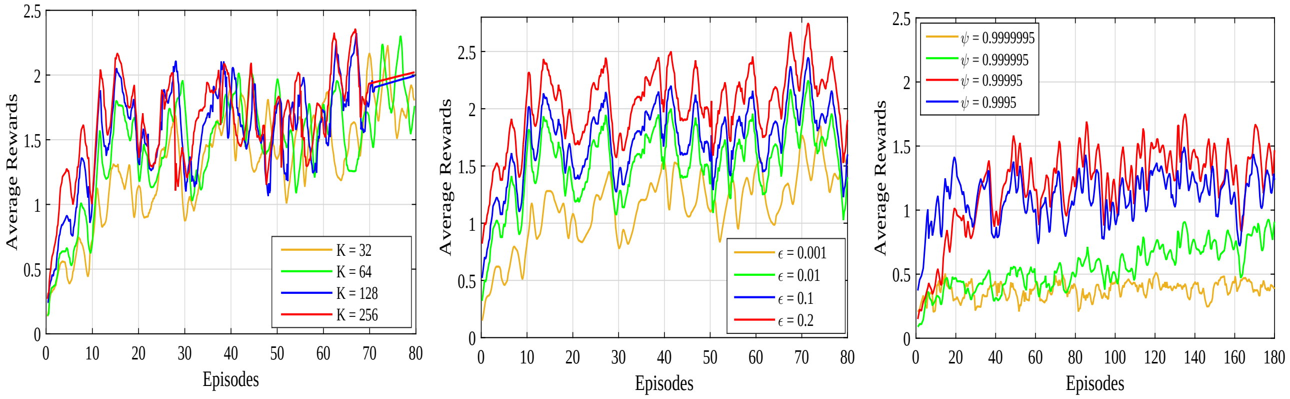

Fig. 4. Comparative Analysis (a) Average Rewards v/s Episodes (b) Average Rewards v/s Episodes (c) Average Rewards v/s Episodes.

TABLE I
SIMULATION PARAMETERS

Parameters	Values
Radius of Cellular cell	500 m
Radius of DGs	50 m
Carrier Frequency	5 MHz
Bandwidth	180KHz
BS transmission power	5 W
Number of DGs	10 ∼ 50 dBm
Number of DGs	10 ∼ 50 dBm
DTs transmission power	10 ∼ 50 dBm
Channel Gain	-30 dB
Noise spectrum density	-174 dBm/Hz
Path loss exponent	4
Actor's Learning Rate	0.01
Critic's Learning Rate	0.01
Discount Factor	0.9
Initial Exploration	1
Final Exploration	0.01
Total Exploration steps	1000
Replay Size	1000
Small-batch Size	32
Number of Steps in Each Epoch	20
Power discretization level	10
Clipping Parameter	0.2
Episodes	100

1) DDPG [41]: In this, the multi-agent actor-critic framework was suggested as a distributed spectrum allocation method based on DDPG. During centralised training, DDPG used global historical states, actions, and policies. It requires no signal involvement during the execution, and relies on user cooperation to enhance system performance.

2) Deep Dueling DQN [35]: In this, an autonomous transmission system for D2D communication networks employing dueling DQN was presented. The dueling DQN learns the value of every state except study the consequence of each action. This method was effective when actions do not have a recursive effect on the environment.

3) DQN Scheduling [42]: In this, the investigation on the selection of joint channels and problem of power control for multi-channel Device-to-Device (D2D) networks was done. Here, the algorithm based on DRL for each D2D pair for learning the diverse patterns of its radio environment from the local information only and the former estimations.

C. Results and Discussion

1) Performance Comparison: The estimated cumulative throughput of the proposed system in comparison to the baseline schemes is investigated and discussed in Fig. 3.

The variation in predicted cumulative throughput is shown in Fig. 3(a) as a function of the number of DGs in the cell. The graph shows that compared to DDPG, Deep Dueling and

DQN scheduling, the proposed technique offers a greater sum throughput. This is so because the recommended method is based on an online training policy, which improves the ability to choose how to allocate resources in real-time.

Fig. 3(b) depicts a graph of sum throughput versus DTs transmission power. According to the results, the recommended strategy outperforms DDPG, Deep Dueling, and DQN scheduling. The proposed technique optimises the power of DTs to regulate co-channel interference more than state-of-the-art solutions due to its online learning methodology. Furthermore, the proposed method reduces co-channel and cross-channel interferences received from other DTs.

Fig. 3(c) depicts the effect of the number of CUs on total throughput. The graph shows that the proposed scheme provides greater throughput than DDPG, Deep Dueling, and DQN scheduling. The key reason for this circumstance is that the proposed scheme trains the BS and CUs quickly to recognize resources owing to their online policy mechanism. The online neural network is capable to detect resources more quickly because it has an ability to train itself more faster in a dynamic environment.

2) Parameter Analysis: In Fig. 4, the proposed scheme's average rewards with respect to the variation in different parameters are examined.

The variance in average reward with regard to the episode for different batch size values is shown in Fig. 4(a). The graph shows that lower batch sizes result in more frequent changes in the proposed scheme's policy settings. As a result, the suggested system performs at its peak more rapidly.

Fig. 4(b) describes the impact of adjusting the value of ϵ on average reward with regard to episode. The graph indicates that when $\epsilon = 0.2$, the suggested POPS algorithm performs better. This happened because the POPS is based on the clipping approach, which makes it possible to get optimal performance more quickly than using state-of-the-art methods.

The average rewards v/s episode with a variable ψ is shown in Fig. 4(c). The graph illustrates that the proposed scheme at $\psi = 0.9995$ provides superior performance and achieves the best solution at a faster pace than the other ψ values. Beyond this, the proposed technique for $\psi = 0.9995$ allows the agents to learn at a faster pace, improving their optimum value and convergence speed.

V. Conclusion

This study investigates the management of resources for D2D links underlying cellular networks and formulates the problem of spectrum distribution as a decentralized multi-agent deep reinforcement learning model. The primary objective of this model is to maximize the sum throughput of D2D links, while simultaneously ensuring the quality of service (QoS) for CUs. The proposed approach towards realizing the desired outcome involves the development of a discriminated spectrum distribution framework, MAD2PG, that is based on the principles of MADRL. In this context, the technique involves sharing global historical states, actions, and policies across the duration of central training. This approach not only enhances network efficiency but also facilitates faster convergence. Moreover, to decrease the computational complexity of training, the proposed approach employs the POPS which utilizes a clipping substitute technique. The simulation results reveal that POPS outperforms DDPG, DDQN, and DQN by 16.67%, 24.98%, and 59.09%, respectively. As a future direction, we suggest the integration of the proposed approach with continuous-valued power control, to develop a combined DRL framework that repeatedly employs resource block and power transmission, to enhance the algorithm's efficiency and robustness.

Acknowledgement

This work is supported by the Ministry of Higher Education MoHE under Fundamental Research Grant Scheme R.K130000.7856.5F506 and supported by the Universiti Teknologi Malaysia under UTM Transdisciplinary Research Grant Q. K130000.3556.06G45.

References

[1] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2133–2168, April 2018.

[2] C. Kai, H. Li, L. Xu, Y. Li, and T. Jiang, "Energy-efficient device-to-device communications for green smart cities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1542–1551, April 2018.

[3] I. Budhiraja, N. Kumar, and S. Tyagi, "Cross-layer interference management scheme for d2d mobile users using noma," *IEEE Systems Journal*, vol. 15, no. 2, pp. 3109–3120, June 2021.

[4] I. Budhiraja, N. Kumar, S. Tyagi, S. Tanwar, Z. Han, M. J. Piran, and D. Y. Suh, "A systematic review on noma variants for 5g and beyond," *IEEE Access*, vol. 9, pp. 85 573–85 644, May 2021.

[5] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. Wiley- Interscience, 2006.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[8] I. Budhiraja, N. Kumar, and S. Tyagi, "Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay d2d communication," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3143–3156, Dec. 2021.

[9] Y. Wei, F. R. Yu, M. Song, and Z. Han, "Joint optimization of caching, computing, and radio resources for fog-enabled iot using natural actor–critic deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2061–2073, Oct. 2019.

[10] J. Huang, Y. Yang, Z. Gao, D. He, and D. W. K. Ng, "Dynamic spectrum access for d2d-enabled internet-of-things: A deep reinforcement learning approach," *IEEE Internet of Things Journal*, pp. 1–1, Apr. 2022.

[11] P. Sun, K. G. Shin, H. Zhang, and L. He, "Transmit power control for d2d-underlaid cellular networks based on statistical features," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4110–4119, Oct. 2017.

[12] K. Wen, Y. Chen, and Y. Hu, "A resource allocation method for d2d and small cellular users in hetnet," in *3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China*. IEEE, 2017, pp. 628–632.

[13] Y. Chen, B. Ai, Y. Niu, K. Guan, and Z. Han, "Resource allocation for device-to-device communications underlaying heterogeneous cellular networks using coalitional games," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 4163–4176, Aug. 2018.

[14] Y. Chen, B. Ai, Y. Niu, R. He, Z. Zhong, and Z. Han, "Resource allocation for device-to-device communications in multi-cell multi-band heterogeneous cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4760–4773, 2019.

[15] J. Huang, C.-C. Xing, Y. Qian, and Z. J. Haas, “Resource allocation for multicell device-to-device communications underlaying 5g networks: A game-theoretic mechanism with incomplete information,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2557–2570, Oct. 2017.

[16] J. Huang, C.-c. Xing, and M. Guizani, “Power allocation for d2d communications with swipt,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2308–2320, Jan. 2020.

[17] J. Huang, C. Huang, C.-C. Xing, Z. Chang, Y. Zhao, and Q. Zhao, “An energy-efficient communication scheme for collaborative mobile clouds in content sharing: Design and optimization,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 10, pp. 5700–5707, May 2019.

[18] J. Wang, Y. Liu, S. Niu, and H. Song, “Lightweight blockchain assisted secure routing of swarm uas networking,” *Computer Communications*, vol. 165, pp. 131–140, Jan. 2021.

[19] J. Wang, Y. Liu, and H. Song, “Counter-unmanned aircraft system (s)(c-uas): State of the art, challenges, and future trends,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 3, pp. 4–29, Mar. 2021.

[20] J. Wang, Y. Liu, S. Niu, and H. Song, “Bio-inspired routing for heterogeneous unmanned aircraft systems (uas) swarm networking,” *Computers and Electrical Engineering*, vol. 95, p. 107401, Oct. 2021.

[21] ——, “Extensive throughput enhancement for 5g-enabled uav swarm networking,” *IEEE Journal on Miniaturization for Air and Space Systems*, vol. 2, no. 4, pp. 199–208, Mar. 2021.

[22] J. Wang, Y. Liu, S. Niu, W. Jing, and H. Song, “Throughput optimization in heterogeneous swarms of unmanned aircraft systems for advanced aerial mobility,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2752–2761, May 2021.

[23] J. Wang, Y. Liu, S. Niu, and H. Song, “Reinforcement learning optimized throughput for 5g enhanced swarm uas networking,” in *IEEE International Conference on Communications (ICC), Montreal, QC, Canada*. IEEE, Aug. 2021, pp. 1–6.

[24] L. Liang, H. Ye, G. Yu, and G. Y. Li, “Deep-learning-based wireless resource allocation with application to vehicular networks,” *Proceedings of the IEEE*, vol. 108, no. 2, pp. 341–356, 2020.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, Nov. 2018.

[26] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, “Applications of deep reinforcement learning in communications and networking: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, May 2019.

[27] R. Amiri, M. A. Almasi, J. G. Andrews, and H. Mehrpouyan, “Reinforcement learning for self organization and power control of two-tier heterogeneous networks,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3933–3947, Aug. 2019.

[28] C. Luo, J. Ji, Q. Wang, L. Yu, and P. Li, “Online power control for 5g wireless communications: A deep q-network approach,” in *IEEE International Conference on Communications (ICC), Kansas, USA*, pp. 1-6, May 2018, pp. 1–6.

[29] H. Li, H. Gao, T. Lv, and Y. Lu, “Deep q-learning based dynamic resource allocation for self-powered ultra-dense networks,” in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2018, pp. 1–6.

[30] V. Vishnoi, P. K. Malik, I. Budhiraja, and A. Yadav, “Deep reinforcement learning based throughput maximization scheme for d2d users underlaying noma-enabled cellular network,” in *International Advanced Computing Conference*. Springer, 2021, pp. 318–331.

[31] Y. S. Nasir and D. Guo, “Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, Nov. 2019.

[32] F. Meng, P. Chen, L. Wu, and J. Cheng, “Power allocation in multi-user cellular networks: Deep reinforcement learning approaches,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6255–6267, June 2020.

[33] J. Huang, C.-C. Xing, S. Gu, and E. Baker, “Drop maslow’s hammer or not: machine learning for resource management in d2d communications,” *ACM SIGAPP Applied Computing Review*, vol. 22, no. 1, pp. 5–14, 2022.

[34] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Advances in neural information processing systems*, vol. 12, pp. 1057–1063, Dec. 2000.

[35] T.-W. Ban, “An autonomous transmission scheme using dueling dqn for d2d communication networks,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 348–16 352, Dec. 2020.

[36] Z. Ji, A. K. Kiani, Z. Qin, and R. Ahmad, “Power optimization in device-to-device communications: A deep reinforcement learning approach with dynamic reward,” *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 508–511, Aug. 2021.

[37] G. T. . V1.20.01, “Study on lte device to device proximity services; radio aspects,” Technical report, Tech. Rep.

[38] S.-A. Ciou, J.-C. Kao, C. Y. Lee, and K.-Y. Chen, “Multi-sharing resource allocation for device-to-device communication underlaying 5g mobile networks,” in *IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong*. IEEE, 2015, pp. 1509–1514.

[39] S. Abeta, “Evolved universal terrestrial radio access (eutra); further advancements for e-utra physical layer aspects,” Technical report (TR) 36.814. 3GPP, Tech. Rep.

[40] H. Zhang, S. Chong, X. Zhang, and N. Lin, “A deep reinforcement learning based d2d relay selection and power level allocation in mmwave vehicular networks,” *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 416–419, June 2020.

[41] Z. Li and C. Guo, “Multi-agent deep reinforcement learning based spectrum allocation for d2d underlay communications,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1828–1840, Dec. 2019.

[42] J. Tan, L. Zhang, and Y.-C. Liang, “Deep reinforcement learning for channel selection and power control in d2d networks,” in *IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA*, pp. 1-6, 2019.

Mohammad Aftab Alam Khan received the B.Tech and M.Tech degrees in Electronics and Communication Engineering in 2005 and 2012, respectively from India. He is currently pursuing PhD at Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia (UTM), Malaysia. His research interests include mobile and Wireless Communications, Error Control Coding, Internet of Things, D2D Communication, NOMA, and Energy Harvesting.

Hazilah Mad Kaidi is a senior lecturer in the Razak Faculty of Technology and Informatics. She received her M.Sc. degree in Telecommunication and Information Engineering at Universiti Teknologi MARA Malaysia in 2008, her B.Eng (Horns) in Electrical Engineering in Telecommunication and her PhD from Universiti Teknologi Malaysia in 2006 and 2015. She is also an associate member of the Wireless Communication Centre (WCC), one of Malaysia’s Higher Institution Centres of Excellence (HICoE). She is a committee member of the National 5G sub-working group and Crowdsource application group under the IMT and Future Networks working group in The Malaysian Technical Standards Forum Bhd (MTFSB). She is an IEEE senior member. Her research interests include mobile and Wireless Communications, Error Control Coding, Relay Networks, Cooperative Communications, Hybrid ARQ, Cross-Layer Design, Internet of Things, and Green Technology.

Norulhusna Ahmad graduated from Universiti Teknologi Malaysia (UTM) in 2001 with BSc in Electrical Engineering. She joined UTM as a staff and later pursued her study at the same university. She received her Master’s degree in Electrical Engineering (Telecommunication) and PhD in Electrical Engineering in 2003 and 2014, respectively. Currently, she is a lecturer at Razak Faculty of Technology and Informatics, UTM KL. During her PhD, she did an attachment in Japan Advanced Institute of Science and Technology (JAIST) under the supervision of Prof. Dr. Tadashi Matsumoto and Asst. Prof. Dr. Khoirul Anwar on the project in non-orthogonal frequency division multiplexing (n-OFDM) system. Her expertise is in the area of digital signal processing and wireless communication. Her research interests are in future communication, such as 5G and cognitive radio focusing on error correcting codes, turbo equalization, OFDM, resource allocation, network coding and cooperative communication.

Masood Ur Rehman (S'06, M'10, SM'16) received the B.Sc. degree in electronics and telecommunication engineering from University of Engineering and Technology, Lahore, Pakistan in 2004 and the M.Sc. and Ph.D. degrees in electronic engineering from Queen Mary University of London, London, UK, in 2006 and 2010, respectively. He worked at Queen Mary University of London as a postdoctoral research assistant till 2012 before joining the Centre for Wireless Research at University of Bedfordshire as a Lecturer. He served briefly at the University of Essex and then moved to the James Watt School of Engineering at University of Glasgow in the capacity of an Assistant Professor in 2019. He currently works as an Associate Professor at University of Glasgow. His research interests include compact antenna design, radiowave propagation, satellite navigation system antennas in cluttered environment, electromagnetic wave interaction with human body, wireless sensor networks in healthcare and environmental monitoring, mmWave and nano communications for body-centric networks and D2D/H2H communications. He is acting as an editor of PeerJ Computer Science, associate editor of IEEE Sensors Journal, IEEE Access, IET Electronics Letters and Microwave & Optical Technology Letters, topic editor for MDPI Sensors, editorial advisor to CambridgeScholars Publishing, and lead guest editor of numerous special issues of renowned journals.