



Deep Reinforcement Learning Based Throughput Maximization Scheme for D2D Users Underlying NOMA-Enabled Cellular Network

Vineet Vishnoi¹(✉), Praveen Kumar Malik², Ishan Budhiraja³,
and Ashima Yadav³

¹ School of Electronics and Electrical Engineering, Lovely Professional University,
Phagwara, Punjab, India

vishnoivineet@gmail.com

² Lovely Professional University, Phagwara, Punjab, India

praveen.23314@lpu.co.in

³ Bennett University, Greater Noida, Uttar Pradesh, India

{ishan.budhiraja,ashima.yadav}@bennett.edu.in

Abstract. Device-to-Device (D2D) communication is a potential technology that efficiently reuses spectrum resources with CMUs in a fifth-generation (5G) underlay and even beyond the network. It improves network capacity and spectral efficiency at the cost of co-channel interference. Moreover, massive connectivity has not been fully exploited for efficient spectral efficiency usage in the existing solutions. To resolve the aforementioned issues, we combine non-orthogonal multiple access (NOMA) approaches with cellular mobile users (CMUs) in order to improve their throughput while preserving the signal-to-interference noise ratio (SINR) offered by CMUs and D2D mobile pairs (DMPs). The problem of power allocation is formulated as mixed-integer non-linear programming, which is then transformed to machine learning using the markov decision process (MDP). Then, a deep reinforcement learning (DRL) approach is proposed for solving the continuous optimisation problem in a centralised fashion. Furthermore, to achieve better performance and a faster convergence rate, the higher proximal policy optimization (PPO) scheme is employed. Numerical results reveal that the proposed algorithm outperformed state-of-the-art schemes in terms of throughput.

Keywords: D2D · DRL · MDP · NOMA · PPO · Throughput

1 Introduction

The explosive growth in the use of smart phones, smart devices, and internet-based services is causing massive data traffic on wireless networks. D2D communication is a key candidate that increases network spectrum efficiency by

sharing radio resources between DMPs and CMUs in the underlay scenario [1]. D2D communication uses the concept of low power transfer, which enhances energy efficiency. Also, in D2D enabled networks, network throughput increases due to sharing of frequency spectrum. Despite the many advantages of D2D communication, its performance has deteriorated in ultra-dense networks due to in-cooperation between nearby BSs and intra-user interference.

To improve the network performance, the authors proposed a promising technique called NOMA for 5G and beyond. The NOMA scheme serves more than one user at the same time and same frequency via power domain multiplexing [2]. Different CMUs get different power from BS based on their channel gain conditions. CMUs having high channel gain get low power signals, and users having low channel gain get high power signals. However, NOMA causes intra-user interference among CMUs. Successive interference cancellation (SIC) is used to decode and compensate for intra-user interference at the receiver end [3]. Therefore, integrating the NOMA scheme with the D2D network improves SE, throughput, sum rate, and capacity of the entire network at the cost of additional interference. Some challenges in implementing the NOMA scheme are degradation in bit error rate, complexity at the receiver side, CMUs location, and physical security.

Reinforcement Learning (RL) is a subfield of machine learning in which an agent can make decisions on a regular basis, track the outcomes, and then change its strategy automatically to obtain the best policy [4]. It has been proven that the learning process in RL converges, but it takes a very long time to arrive at the best policy. The reason for this is that RL must explore and learn about the whole system. So RL becomes unsuitable and inappropriate for large-scale networks. As a result, RL implementations in practise are extremely limited in ultra-dense networks.

DRL schemes are used to overcome RL's limitations by combining RL and deep learning. DRL schemes take advantage of deep neural networks to improve the learning process. The RL algorithm learns faster and performs better using DRL schemes. Therefore, DRL schemes are used in a variety of RL applications, including speech recognition, natural language processing, computer vision, and robotics [5]. DRL schemes are widely used in the field of wireless communication and networking to handle different challenges and issues.

1.1 Related Work

The authors proposed incorporating the DRL scheme DQN into an overlay D2D communication network in order to reduce mutual interference among DMPs and improve the SE [6]. In [7], the authors solved the energy-efficiency resource allocation issue for underlay D2D networks to maximise the user experience by using deep queue learning (DQL), Double DQL and Duelling DQL DRL techniques. The authors solved the problem of power allocation for underlay D2D communication networks in a changing environment and proposed a DRL-based approach, DQN, to improve the capacity of the system and user experience quality for the entire network [8]. Ji *et al.* [9] studied the issue of resource allocation

in underlay D2D networks, for which the DRL based scheme DQN was used to improve energy efficiency with respect to network throughput. The authors used a DRL-based technique named deep deterministic policy gradient (DDPG) to study an energy efficiency (EE) maximisation problem in terms of mode selection and resource allocation for an uplink D2D scenario and achieved greater EE and a faster convergence rate than state-of-the-art schemes [10]. Chen *et al.* [11] addressed the issue of channel allocation in the overlay D2D communication network and then developed the DRL scheme, DQN, to maximise the sum rate. The authors mitigate the co-channel interference in overlay D2D communication networks via a distributed DRL-based algorithm [12]. To optimise the aggregate of the fairness utility function with respect to scheduling of RB and power control, the authors combined the traditional computational scheme with the DDPG [13]. Tang *et al.* [14] suggested DQL and DQN algorithms for D2D-assisted cache-enabled IoT to reduce the energy cost of efficient traffic.

1.2 Motivation and Contributions

In contrast to the previous work, we only employ the NOMA technique in this work to schedule a set of CMUs on the RBs, while DMUs reuse these resource blocks (RBs) in an orthogonal manner for the cellular tier, subject to interference protection. The motivation for using traditional DMPs rather than NOMA-based DMPs is to minimise the computation on the resource-constrained D2D devices, hence making it more feasible. Furthermore, if interference is adequately handled, then more DMPs can join the network, because in the NOMA scheme, two or more DMPs can be scheduled to sustain minimal SIC receiver complexity. This inspires us to work on a solution for underlay DMPs that coordinates with NOMA-based CMUs. Furthermore, the delay in the centralised learning and the processing time in their optimisation algorithms is huge for real-time use cases. To get the better of these aforementioned shortcomings, in this paper, we propose efficient DRL algorithms by optimising the power allocation of the BS and the DTs for maximising the network throughput.

The main contributions of this paper are as follows:

- The throughput problem is formulated for the downlink scenario with the power restrictions. To optimise the throughput network performance, we propose a centralised DRL technique for solving the power allocation at the BS and DTs.
- To improve the network performance, we introduced the PPO algorithm with a new better sampling technique.
- The numerical results demonstrated that the proposed methods efficiently solve the optimisation problem with the dynamic environmental setting and outperform the other benchmarks.

1.3 Organization

The remainder of the paper is arranged in the following manner. The system model and problem formulation are described in Sect. 2. Section 3

demonstrates the suggested scheme. The suggested scheme's performance was evaluated in Sect. 4. Section 5 contains the conclusion.

2 System Model and Problem Formulation

Consider a model comprising a BS, a set of \mathcal{I} CMUs as $\mathcal{I} = \{1, 2, \dots, i \dots I\}$, a set of \mathcal{J} D2D mobile pairs (DMPs) as $\mathcal{J} = \{1, 2, \dots, j \dots J\}$, and a set of \mathcal{N} RB as $\mathcal{N} = \{1, 2, \dots, n \dots N\}$ shown in Fig. 1. The BS provides service to a group of CMUs through the NOMA scheme in a downlink scenario. On the other hand, D2D transmitters and receivers communicate with each other through the orthogonal multiple access (OMA) scheme. In this model, CMUs and DMPs share the same RB.

In this model, the BS used NOMA to schedule with the CMUs through a RB, and the D2D transmitter communicates with the D2D receiver using the OMA scheme in each DMP. The CMUs and DMPs form a cluster. Furthermore, the total number of users can vary from 2 to $|I| + |J|$ in each cluster. In NOMA based systems, if the number of users increases in the same RB, then SIC implementation complexity at the receiver increases. So, to keep the receiver complexity to a minimum, this model considers only two CMUs in each cluster. There is no limit on DMPs on a RB. Let \mathcal{N} represent the clusters set with each RB assigned to one of them. Let \mathcal{N} is the set of clusters, i.e., each RB is allocated to each cluster. Also, assume that the BS is aware of the channel state information (CSI) of all CMUs and DMPs. Furthermore, this model considers quasi-static Rayleigh fading, in which each channel's gain is constant and follows a Gaussian complex distribution.

2.1 Channel Model

Assume that CMU and DMP on the r^{th} cluster are represented by I_n and J_n , respectively. Let P_T represent the power transmitted by the BS and P_i represent the power assigned to CMU. The received message at CMUs i from BS in the n^{th} cluster is given as:

$$y_{b-i}^n = \sqrt{P_i^n} g_i^n x_i^n + \sum_{i' \neq i, i' \in I_n} \sqrt{P_{i'}^n} g_{i'}^n x_{i'}^n + \sum_{j \in J_n} \sqrt{P_j^n} g_{j-i}^n x_j^n + \zeta_i^n, \quad (1)$$

where x_i represent the transmitted symbol for CMUs, g_{j-i}^n represent the channel gain between CMUs i and DP j . P_j^n is the power of DP j and ζ_i^n is the additive white noise.

Let a_i^n is the channel coefficient for CMU i and b_i^n is the channel coefficient for D2D user j and is defined as:

$$a_i^n = \begin{cases} 1 & \text{if CMU } i \text{ is scheduled on the } n^{th} \text{ RB,} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

$$b_j^n = \begin{cases} 1 & \text{if DMP } j \text{ is scheduled on the } n^{\text{th}} \text{ RB,} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

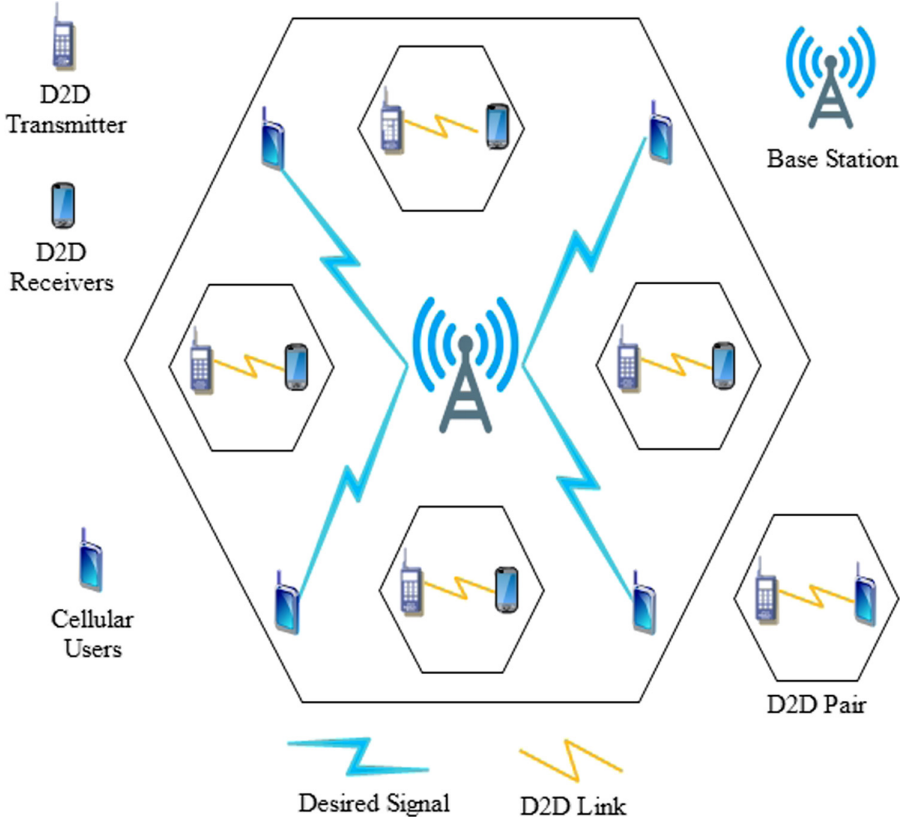


Fig. 1. System architecture.

2.2 Throughput Calculation

The desirable throughput for i^{th} CMU in the n^{th} cluster using (1) is given as:

$$D_i^n = B \log_2 \left[1 + \frac{P_i^n |g_i^n|^2}{\mathcal{I}\mathcal{F}_{i'-i}^n + \mathcal{I}\mathcal{F}_{j-i}^n + \xi_j^n} \right], \quad (4)$$

where B represents the amount of bandwidth allotted to every one RB and $\mathcal{I}\mathcal{F}_{i'-i}^n$ is the intra-user interference produced by other CMUs on CMU i and

can be given as:

$$\mathcal{IF}_{i'-i}^n = \sum_{i' \neq i, i' \in I_n} \alpha_i^n P_i^n |g_i^n|^2. \quad (5)$$

and \mathcal{IF}_{j-i}^n is the interference produced by DMP j on CMU i is specified as follows:

$$\mathcal{IF}_{j-i}^n = \sum_{j \in J} \beta_j^n P_j^n |g_{j-i}^n|^2. \quad (6)$$

Similarly, the desired throughput for DMP j on the n^{th} cluster is specified as:

$$D_j^n = B \log_2 \left[1 + \frac{P_j^n |g_j^n|^2}{\mathcal{IF}_{j'-j}^n + \mathcal{IF}_{BS-j}^n + \xi_j^n} \right], \quad (7)$$

where $\mathcal{IF}_{j'-i}^n$ is the co-tier interference caused by other DMPs on DMP j which is given as:

$$\mathcal{IF}_{j'-i}^n = \sum_{j' \neq j, j' \in J_n} \beta_{j'}^n P_{j'}^n |g_{j'}^n|^2. \quad (8)$$

and \mathcal{IF}_{BS-j}^n represents the cross-tier interference produced by BS on all DMPs and is defined as:

$$I_{BS-j}^n = \sum_{j \in J} \beta_j^n P_j^n |g_{BS-j}^n|^2, \quad (9)$$

where g_{BS-j}^n represent the channel gain between BS and j^{th} DMP.

Now, the total throughput of the overall network obtained from (4) and (7) is given as:

$$D_T^n = \sum_{n=1}^N \left[\sum_{i=1}^I D_i^n + \sum_{j=1}^J D_j^n \right]. \quad (10)$$

2.3 Problem Formulation

The aim of this paper is to increase the total network's throughput by reducing interference. The following is the problem's mathematical formulation:

$$\begin{aligned} & \max_{P_i^n, P_j^n} D_T^n, & (11) \\ \text{s.t. } & V_1 : P_i^n \leq P_i^{n, \max}, \\ & V_2 : P_j^n \leq P_j^{n, \max}, \\ & V_3 : P_i^n, P_j^n \geq 0, \\ & V_4 : 2 \leq \sum_{j=1}^J b_j^n \leq |I| + |J|, \\ & V_5 : \sum_{j=1}^J b_{j,i}^n P_j^n |g_{j,i}|^2 \leq I_i^{\text{threshold}}, \end{aligned}$$

where the constraints V_1 and V_2 ensure that the power transmitted by the BS and the DT must be less than the maximum transmission power. Constraint V_3 shows that the transmitted power must be a non-negative integer. The constraints V_4 shows that the total number of users can vary from 2 to $|I| + |J|$ in each cluster. The highest interference threshold assigned by CMUs to a resource block is represented by Constraints V_5 .

3 Proposed Solution

3.1 Centralised Optimisation

In this model, consider that information is processed at a centralised location in a centralised manner (e.g., at the base station). In each sharing resource block, the next action for each system element will be transferred. As a result, we consider the central processing point as an agent (BS) for optimising throughput at CMUEs and DMPs. The optimisation problem can be defined by MDP as:

$$MDP = (\mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{R}, \Gamma). \quad (12)$$

With respect to the above model, the game with a centralised optimisation approach is described as:

State Space: In order to achieve maximum throughput, the agent interacts with the environment. Therefore, the agent is solely aware of local information such as different channel gains and interferences. The state space is defined as:

$$\mathbb{S} = [g_i^n, g_{BS-j}^n, g_{j-i}^n, \mathcal{I}\mathcal{F}_{j-i}^n, \mathcal{I}\mathcal{F}_{BS-j}^n, \mathcal{I}\mathcal{F}_{i'-i}^n]. \quad (13)$$

Action Space: In NOMA based systems, our aim is to optimise the throughput at BS. So, action space is represented as:

$$\mathbb{A} = \left[(P_1^n, P_2^n, \dots, P_i^n); (P_1^n, P_2^n, \dots, P_j^n) \right]. \quad (14)$$

At the state s^t , agent perform the action a^t . After performing action a^t , agent moves at the next state s^{t+1} .

Reward Function: To maximise the throughput, the reward function is expressed as:

$$\mathbb{R} = \sum_{i=1}^I \sum_{j=1}^J D_T^n. \quad (15)$$

After defining the throughput model, a DRL approach is proposed to identify the optimal policy. The DDPG is a hybrid model with a value function-based actor component and a policy search-based critic component. To enhance the convergence speed and reduce unnecessary calculations, we apply experienced replay buffer and target network approaches to the DDPG algorithm. A finite memory of capacity C is utilised to store the executed transition (s^t, a^t, r^t, s^{t+1})

in the experience replay buffer. We select a small batch E at random from the finite size memory C after collecting enough samples. This small batch trains the neural network. For updating the new sample and deleting the old ones, memory C is assigned to a finite size. The target value is estimated by using target networks for both the critic and the actor network.

Let $Q(s, a; \phi_x)$ represents the critic network along with variable ϕ_x and $Q'(s, a; \phi_{x'})$ represents the target critic network along with variable $\phi_{x'}$. Similarly, $\nu(s, a; \phi_\nu)$ represents the actor network along with the variable ϕ_ν and $\nu'(s, a; \phi_{\nu'})$ represents the target actor network along with variable $\phi_{\nu'}$. Stochastic gradient descent (SGD) is used to train the actor and critic network over a small batch of E samples. Now the critic is updated by minimising

$$M = \frac{1}{E} \sum_i^E (y^i - Q(s^i, a^i, \phi_x))^2. \quad (16)$$

with the target

$$y^i = r^i(s^i, a^i) + \Gamma Q(s^{i+1}, a^{i+1}; \phi_{x'}|_a^{i+1} = \nu'(s^{i+1}; \phi_{\nu'}). \quad (17)$$

The actor network parameter is updated as follows:

$$\vec{\nabla}_{\phi_\nu} K \approx \frac{1}{E} \sum_i^E \vec{\nabla}_{a^i} Q(s^i, a^i; \phi_x)|_{a^i=\nu(s^i)} \vec{\nabla}_{\phi_\nu} \nu(s^i; \phi_\nu). \quad (18)$$

Soft target updates are used to update the target actor network parameters ϕ_x and the target critic network parameters $\phi_{\nu'}$ as follows:

$$\phi_{x'} \leftarrow \chi \phi_x + (1 - \chi) \phi_{x'} \quad (19)$$

$$\phi_{\nu'} \leftarrow \chi \phi_\nu + (1 - \chi) \phi_{\nu'}, \quad (20)$$

where χ is defined as a hyperparameter and has a range between 0 and 1.

The deterministic policy is trained in an off-line manner in the DDPG approach. So, a noise process is added and defined as $\mathcal{Z}[0, 1]$. Therefore, the target actor network is defined as follows:

$$\nu'(s^t; \phi_{\nu'}^t) = \nu(s^t; \phi_\nu^t) + \mathcal{Z}\chi(0, 1). \quad (21)$$

In the suggested algorithm, we give the of the DDPG algorithm-based method for power allocation and the NOMA-based BS in the downlink scenario. In the proposed algorithm, Θ represent the number of maximum episodes and \mathcal{T} denotes time step.

3.2 Proximal Policy Optimization

In order to achieve better performance, we consider a policy approach denoted as proximal policy optimization (PPO) in this model. Current and obtained

Algorithm 1. Throughput Maximization for D2D Users Using Centralised Optimization Technique.

Input

- Environment: (a) DMPs (b) CUEs (c) NOMA-integrated BS.
- $D_i^n \geq D_i^{n,\min}$: Minimum requirement of CUE
- $D_j^n \geq D_j^{n,\min}$: Minimum Requirements of DPs

Initialization:

- $\mathbb{Q}(s, a, \phi_x)$ = Critic network along with variable ϕ_x .
- $\nu(s; \phi_\nu)$ = Actor network along with variable ϕ_ν .
- $\mathbb{Q}'(s, a, \phi'_x)$ = Target critic network along with variable $\phi_{x'}$
- $\nu'(s; \phi_{\nu'})$ = Target actor network along with variable $\phi_{\nu'}$.
- C = Experience Replay

Output: α, β

- 1: **for** episode = 1, ..., Θ **do**
 - 2: Begin a process of action exploration
 - 3: Obtain the starting state of observation s^0
 - 4: **for** iteration = 1, ..., \mathcal{T} **do**
 - 5: Execute the action a^t achieved at state s^t
 - 6: Modify the reward r^t in accordance with (27)
 - 7: Notice the next state s^{t+1}
 - 8: save the transition (s^t, a^t, r^t, s^{t+1}) in the replay buffer
 - 9: replay buffer
 - 10: Sample randomly a mini-batch of E transitions
 - 11: (s^i, a^i, r^i, s^{i+1}) from C
 - 12: Update critic parameter by stochastic gradient descent using loss function in (27)
 - 13: Update the actor policy parameter in (30)
 - 14: Update the target critic network parameters
 - 15: $(\phi_{x'}, \phi_{\nu'})$ according to (31) and (32)
 - 16: Update the state $s^t = s^{t+1}$
 - 17: Update the state $s^t = s^{t+1}$
 - 18: **end for**
 - 19: **end for**
-

policies are compared in the PPO algorithm and then the objective function is maximised as:

$$\begin{aligned} \mathcal{F}(s, a; \phi) &= \mathcal{E} \left[\frac{\pi(s, a; \phi)}{\pi(s, a; \phi_{old}} W^\pi(s, a) \right] \\ &= \mathcal{E} P_\phi^t W^\pi(s, a), \end{aligned} \quad (22)$$

where P_ϕ^t represents the probability ratio and $W^\pi(s, a) = \mathbb{Q}^\phi(s, a) - \mathcal{V}^\pi s$ is a function that approximates the advance function in. To maximise the goal, SGD is applied for training networks with a mini-batch E . As a result, the policy is updated via

$$\phi^{t+1} = \arg \max \mathcal{E} [\mathcal{F}(s, a; \phi^t)] \quad (23)$$

To avoid excessive changes, we employ the clipping method function $\text{clipp}(p_\pi^t, 1 - \lambda, 1 + \lambda)$ in this work to limit the objective value as follows:

$$\begin{aligned} & \mathcal{F}^{\text{clipp}}(s, a; \phi) \\ &= \mathcal{E} [\min(p_\phi^t, W^\pi(s, a), \text{clipp}(p_\phi^t, 1 - \lambda, 1 + \lambda)W^\pi(s, a))], \end{aligned} \quad (24)$$

where λ is a constant of low value. When the advantage $W^\pi(s, a)$ is greater than zero then upper bound is defined as $1 + \lambda$. In this condition, the objective is defined as:

$$\mathcal{F}^{\text{clipp}}(s, a; \phi) = \min \left[\frac{\pi(s, a; \phi)}{\pi(s, a; \phi_{old})}, (1 + \lambda) \right] W^\pi(s, a). \quad (25)$$

When the advantage $W^\pi(s, a)$ is less than zero then lower bound is defined as $1 - \lambda$. In this condition, the objective is defined as:

$$\mathcal{F}^{\text{clipp}}(s, a; \phi) = \min \left[\frac{\pi(s, a; \phi)}{\pi(s, a; \phi_{old})}, (1 - \lambda) \right] W^\pi(s, a). \quad (26)$$

In (25), if advantage $W^\pi(s, a)$ is greater than zero, then the value of the objective increases. But the minimum term puts a limit on the increased value. When $\pi(s, a; \phi) > (1 + \lambda)\pi(s, a; \phi_{old})$, then factor $(1 + \lambda)W^\pi(s, a)$ limits the objective value within the range.

Similarly, in (26), if advantage $W^\pi(s, a)$ is less than zero, then the value of the objective decreases. But the maximum term puts a limit on the decreased value. When $\pi(s, a; \phi) < (1 - \lambda)\pi(s, a; \phi_{old})$, then factor $(1 - \lambda)W^\pi(s, a)$ limits the objective value. Thus, the minimum and maximum terms put conditions on the objective in such a way that the new policy does not deviate from the old policy. An advantage function is denoted as [15]:

$$W^\pi(s, a) = r^t + \mathbb{I}\mathcal{V}^\pi(s^{t+1}) - \mathcal{V}^\pi(s^t). \quad (27)$$

4 Performance Evaluation

The performance of the suggested strategy is examined and described in this section. It is divided into two sections: (i) Numerical Settings (ii) Results and Discussion.

4.1 Numerical Settings

Simulation Parameters for DMPs and CMUs. The BS is considered to be deployed at a fixed location in the simulation, and I CMUs and J DMPs are deployed according to a homogeneous Poisson point process (PPP). The main parameters used in simulations are taken from [3, 13] and are presented in Table 1.

Table 1. Underlay D2D network simulation parameters.

Parameters	Values
Cellular radius of a cell	300 m
The distance between D2D links	20 30 m
Count of CMUs	20
Count of RBs, (N)	20
Count of DMPs	10, 20, 30, \dots , 180
Each RB's bandwidth	180 kHz
Frequency of carrier	2 GHz
Density of noise power spectrum	-174 dBm/Hz
D2D pathloss exponent	4
Shadowing standard deviation	8 dB
CMU highest power, (P_i^n)	25 dBm
DMP highest power, (P_j^n)	10-25 dBm
CMU links pathloss model	$128.1 + 37.6 \log d$
DMP links pathloss model	$148 + 40 \log d$

DRL Simulation Parameters. A totally connected neural system is used in the DQN learning model. There are three layers in this network system: an input layer, a hidden layer, and an output layer. There are 250 neurons in the input layer, 250 neurons in the hidden layer, and 150 neurons in the output layer, respectively. The ReLu is employed as an activation function in the suggested model, while the adaptive moment is used as an optimizer. In the PPO algorithm, we use the learning rate = 0.00001. Table 2 contains the other parameters associated with the DQN model. Tensorflow 2.23 on the Python 5 platform is used to simulate the model.

4.2 Results and Discussion

The performance of the entire network's throughput is evaluated in relation to several factors, such as the number of DMPs, the number of CMUs, and the interference threshold.

The convergence behaviour of the proposed algorithm in relation to the number of iterations is shown in Fig. 2(a). Our suggested algorithm obtains maximum throughput in fewer than 30 iterations. The cause for this behaviour is that the proposed algorithm maximises the power of both the CMU and the DT, minimising co-channel interference. As a result, each agent educated themselves multiple times and utilised previously trained networks to acquire the best policy in less time.

The change in the throughput of the entire network in relation to the number of DMPs is shown in Fig. 2(b). The statistics suggest that as the number of DMPs increases, the network's throughput decreases due to increased co-channel

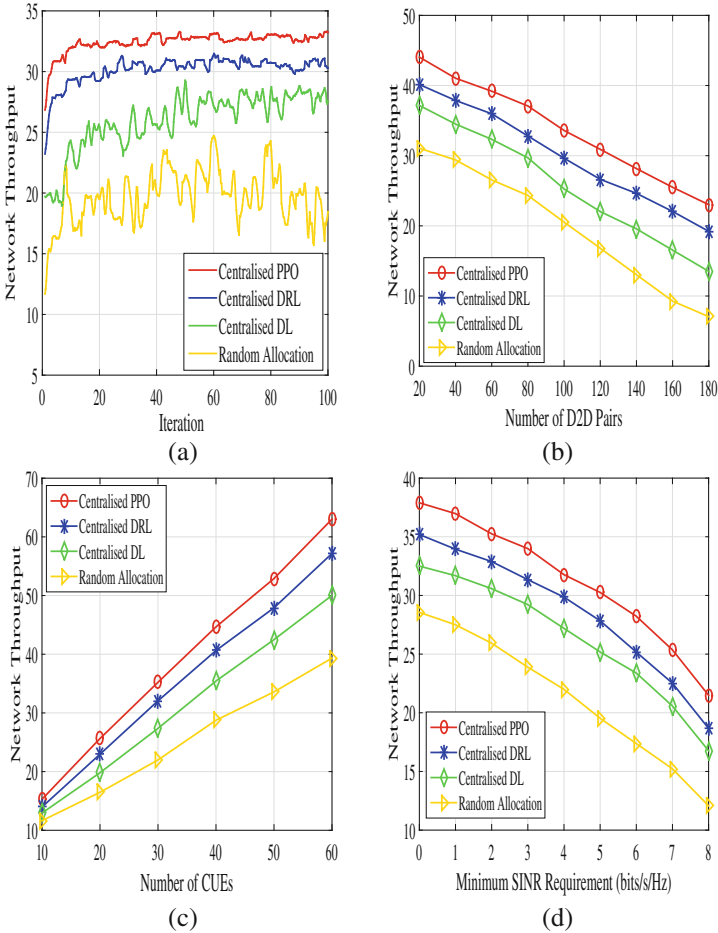


Fig. 2. Comparative Analysis (a) Network Throughput v/s number of Iterations (b) Network Throughput v/s number of D2D pairs (c) Network Throughput v/s number of CUEs (d) Network Throughput v/s Minimum SINR requirement

interference among DMPs. Figure 2(b) also shows the comparison of the PPO algorithm with other algorithms. The results suggest that the existence of multi-DQN and prioritised experience replay in PPO reduces the size of the action spaces and eliminates redundant samples, resulting in increased throughput. Also, when the number of DMPs in a cell reaches 100, the suggested scheme achieves 10.25%, 18.97%, and 32.34% higher throughput than the compared algorithms.

The network throughput is shown in Fig. 2(c) in relation to the number of CMUs. The results suggest that as the number of CMUs increases, the network’s throughput decreases due to an increase in the number of RBs, with constant

Table 2. DRL simulation parameters.

Parameters	Values
Learning rate of actor network	0.01
Learning rate of critic network	0.01
Discount factor	0.9
Exploration in the starting	1
Final exploration	0.01
Number of exploratory steps	1000
Capacity of the replay buffer	1000
Size of mini-batch	32
In every epoch, number of steps	20
Reward function weights	1, 1
Discretization level of power l	10
Updated interval of weights	10
Activation function	ReLu
Optimizer	Adam

DMUs. Thus, DMPs have the option of obtaining the best RBs in order to meet their data rates with less interference.

Figure 2(d) depicts the throughput variation in relation to the lowest SINR requirements for DMPs. The result indicates that while the SINR need of DMPs is low, the network's throughput is likewise high. However, when the SINR need of DMPs is high, the network's throughput begins to fall at a quicker rate. This occurred as the number of D2D transmission pairs increased in response to increased SINR requirements, resulting in increased co-channel interference between the DMPS. Furthermore, the results reveal that the suggested algorithm gives improved results than the compared existing algorithms because, in existing approaches, transmission power is solely managed by the DDPG.

5 Conclusion

In this paper, our main goal is to optimise the total network throughput while keeping the SINR of the CMUs and DMPs as high as possible. To achieve the target, a power allocation scheme is designed. First of all, the centralised optimisation scheme is applied across the NOMA-based BS and DTs to reduce cross channel and co-channel interference. Next, to achieve better performance, train the model quickly, and faster convergence rate, the PPO is used to optimise the power of the BS and DTs. The evaluated results reveal that the proposed approach accomplishes better throughput than the state-of-the-art schemes.

References

1. Shafi, M., et al.: 5G: a tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE J. Sel. Areas Commun.* **35**(6), 1201–1221 (2017)
2. Budhiraja, I., et al.: A systematic review on NOMA variants for 5G and beyond. *IEEE Access* **9**, 85573–85644 (2021)
3. Budhiraja, I., et al.: ISHU: interference reduction scheme for D2D mobile groups using uplink NOMA. *IEEE Trans. Mob. Comput.* (2021)
4. Goodfellow, I., et al.: *Deep Learning*. MIT Press, Cambridge (2016)
5. Luong, N.C., et al.: Applications of deep reinforcement learning in communications and networking: a survey. *IEEE Commun. Surv. Tutor.* **21**(4), 3133–3174 (2019)
6. Budhiraja, I., Kumar, N., Tyagi, S., Tanwar, S., Han, Z.: An energy efficient scheme for WPCN-NOMA based device-to-device communication. *IEEE Trans. Veh. Technol.* **70**(11), 11935–11948 (2021)
7. Nguyen, K.K., et al.: Non-cooperative energy efficient power allocation game in D2D communication: a multi-agent deep reinforcement learning approach. *IEEE Access* **7**, 100480–100490 (2019)
8. Bi, Z., et al.: Deep reinforcement learning based power allocation for D2D network. In: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), pp. 1–5, March 2020
9. Ji, Z., et al.: Power optimization in device-to-device communications: a deep reinforcement learning approach with dynamic reward. *IEEE Wirel. Commun. Lett.* **10**(3), 508–511 (2021)
10. Zhang, T., et al.: Energy-efficient mode selection and resource allocation for D2D-enabled heterogeneous networks: a deep reinforcement learning approach. *IEEE Trans. Wirel. Commun.* **20**(2), 1175–1187 (2021)
11. Chen, M., et al.: Continuous incentive mechanism for D2D content sharing: a deep reinforcement learning approach. In: 2020 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6, June 2020
12. Tan, J., et al.: Deep reinforcement learning for joint channel selection and power control in D2D networks. *IEEE Trans. Wirel. Commun.* **20**(2), 1363–1378 (2021)
13. Budhiraja, I., et al.: Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay D2D communication. *IEEE Internet Things J.* **8**(5), 3143–3156 (2021)
14. Tang, J., et al.: Energy minimization in D2D-assisted cache-enabled Internet of Things: a deep reinforcement learning approach. *IEEE Trans. Ind. Inf.* **16**(8), 5412–5423 (2020)
15. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937, June 2016