Energy-Efficient Power Control and Resource Allocation Based on Deep Reinforcement Learning for D2D Communications in Cellular Networks

Sami Alenezi, Chunbo Luo, Geyong Min
Department of Computer Science, University of Exeter, Exeter, EX4 4QF, United Kingdom
E-mail: {sa596, c.luo, g.min}@exeter.ac.uk

***Abstract*—Device-to-Device (D2D) communication has become a promising and new paradigm for enhancing network performance in cellular networks. D2D communication enables users to communicate directly without passing through the base station, thereby improving spectral efficiency and reducing communication delay. However, due to the intertwined interference environment, the shared spectrum and reused frequency may limit the network performance. In this paper, We propose a Proximal Policy Optimisation (PPO) algorithm based on Markov Decision Process (MDP) to optimise resource allocation and improve energy efficiency. Resource allocation and power control are jointly considered with the aim of maximising the overall throughput of the network while guaranteeing the minimum requirement of Quality of Service (QoS). Extensive simulation experiments are conducted to validate the efficacy of our proposed scheme. The results demonstrate that our method outperforms the traditional method in terms of energy efficiency and training time.**

Reinforcement learning, D2D communications, Resource allocation, Power control, Energy efficiency.

I. Introduction

With the continuous evolution of wireless communication systems, the explosive growth of data traffic carried by mobile communications and the shortage of wireless spectrum resources have forced cellular networks to face huge challenges [1]. Device-to-Device (D2D) communication enables direct communications between two peer-to-peer user nodes to improve resource utilisation and network capacity, becoming one of the key technologies in the fifth generation (5G) cellular networks [2]. In a distributed network composed of D2D communication users, each user node can send and receive signals, and is capable of automatic routing. The user nodes of the network share part of the hardware resources bypassing through the base station (BS) to improve resource utilisation and Quality of Service (QoS) [3].

However, the sharing of spectrum resources may cause interference to user communications, making the resource allocation more complicated, thereby affecting the overall communication experience. Therefore, it is essential to allocate spectrum resources appropriately to mitigate the mutual interference and make up for the shortage of available spectrum [4]. Moreover, proper resource management should ensure effective improvement of the total throughput of D2D communication without harmful interference to the cellular network [5].

The issue of resource allocation has been treated in different ways in the literature. Some existing methods optimise resource allocation through the game theory model [6]–[10]. Rathi *et al.* [6] developed a game theory-based model to offer optimal solutions to the resource allocation problems and maximise the whole system throughput for device-to-device communications. In [7], a game-theoretic resource allocation scheme, termed GALLERY and a resource allocation protocol based on the equilibrium derivation are proposed to improve the system performance. The authors in [8] decompose the resource allocation problem into independent sub-problems. A sub-channel allocation algorithm is proposed, and unpopular D2D users are assigned with higher priority to guarantee fairness. In [9], D2D users act as relays for cellular users to set up a relationship between the total average achievable rate and resource allocation. Maximum resource allocation is achieved by maximising the total average achievable rate under the constraint of outage probability. In [10], the research considers the dynamic network condition where direct D2D communication is not possible and will require a relay. Aiming at minimising the interference, a resource allocation method for two-hop D2D communications has been proposed, and the base station will give a higher priority to the resource block that creates less interference. The result demonstrates the scheme perform better compared to the random allocation scheme.

Since the problem formulation of resource allocation involves binary channel assignment parameters, it leads to a non-convex, mixed-integer-non-linear program (MINLP), so the solution obtained by using traditional technology is not globally optimal and may not be available in real-time performance. Therefore, deep learning approaches are developed to address the resource allocation problem. Deep learning, as a subset of Machine Learning (ML), uses the hierarchical structure of artificial neural networks (ANN) to perform the learning process [11]. The hierarchical function adopts a non-linear method [12], which can effectively solve the problem of resource allocation and management in communication networks.

Extensive deep learning-based research has been conducted on resource allocation in D2D communications [13]–[16]. Feki *et al.* [13] proposed a dynamic neural Q-learning-based resource allocation algorithm for D2D-based communication in the Long Term Evolution Advanced (LTE-A) cellular networks under the constraint of the minimum QoS requirement. The work in [14] utilises Deep Q-network (DQN) to maximise the overall effective throughput by allocating channel resources to each D2D pair. In [15], the authors propose a novel QoS-based resource allocation method and adopt OFDMA to reuse the

978-1-6654-6667-7/21/$31.00 ©2021 IEEE
DOI 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00026

uplink resource. In [16], a decentralised resource allocation scheme based on deep reinforcement learning is proposed to allocate appropriate channel resources to each D2D pair iteratively for maximising the overall effective throughput.

Besides the quality of service, energy efficiency is also an important evaluation criterion of power control for D2D communications underlaying cellular networks. Many approaches of power control have been proposed based on deep learning in D2D Communications [17]–[20]. Gengtian *et al.* [17] propose a reinforcement learning algorithm for adaptive power control that helps reduce interference to increase system throughput. The simulation results are compared with the traditional algorithm in Long Term Evolution (LTE). In [18], the authors introduce an online distributed reinforcement learning algorithm to maximise network throughput while guaranteeing QoS of all D2D users' and cellular users' (CUs) considering the dynamic wireless channel environment. Ji *et al.* [19] design a novel algorithm with two parallel deep Q-networks (DQNs) to maximise the energy efficiency (EE) of the considered network. The proposed deep RL based power optimisation method with dynamic rewards achieves higher EE while satisfying the system throughput requirements. In [20], the power control problem is modelled as a Stackelberg game. Simulation results show that the scheme is efficient with low overhead.

Recently, resource management mainly considers both resource allocation and power control to minimise the interference caused by D2D communications [21]–[29]. The authors in [21] propose a D2D resource allocation and power control (DRAPC) framework. The optimisation problem is defined to maximise links supported in the network. In [22], the author provides a low complexity algorithm for spectrum reuse and power assignment. Through the demonstration of simulation results, the overall system throughput has increased, and the interference has been mitigated. However, the solutions are not intelligent enough, address the importance of deep learning approaches. In [23], the resource allocation problem is formulated as a Mixed-Integer Non-Linear Programming (MINLP) problem and transformed into a resource group (RG) allocation problem, which can be solved optimally by the Hungarian method. Luo *et al.* [24] use Q-learning, the basic RL algorithm, to finish the channel assignment and the power allocation at the same time. The system capacity has been improved. In [25], a hybrid intelligent clustering strategy (HICS) based on unsupervised learning is proposed. By maximising the total energy efficiency of the D2D multicast cluster, a joint resource allocation scheme is proposed. The simulation result demonstrates the proposed algorithm has decreased the computation complexity. Saied *et al.* [26] use an Actor-Critic Reinforcement Learning (AC-RL) approach to solve the resource management problem. A distributed multi-agent reinforcement learning (MARL) based joint SA-PC algorithm is proposed in [27] for performing spectrum allocation and power control to each D2D user in the network. In [28], the authors propose a Stackelberg game (SG) guided multi-agent deep reinforcement learning (MADRL) approach to make smart power control and channel allocation decisions in a distributed manner. In [29], the author considers the instability of the D2D communication and proposes a mobility-aware joint resource allocation and power allocation algorithm (MARP) to optimise the channel resources allocation and power allocation.

In this paper, we investigate the joint mode selection, resource allocation and power control in a D2D-enabled heterogeneous network. Specifically, Deep Q-Network (DQN)-based deep learning method is proposed for optimal resource allocation, while proximal policy optimisation (PPO) is employed for power control to improve the overall system throughput with the constraint of QoS. The main contributions of this work are as follows:

- We model the resource allocation and power control of D2D communication as a joint optimisation problem, which is formulated to maximise the overall system throughput under the constraint of minimum QoS requirements.
- Considering the complexity and non-convexity of the joint optimisation problem, a DRL-based method is proposed to optimise the resource allocation and power control policy intelligently. The joint optimisation problem is modelled as a Markov Decision Process.
- The simulation results demonstrate the proposed algorithm effectively improves the overall energy efficiency compared with the other two algorithms without affecting the overall system throughput.

The rest of this paper can be organised as follows. Section II presents the system model and optimisation problem. In Section III, we model the resource management and power control problem as MDP and formulate its basic elements, including action, agent, state and environment. Section IV shows the simulation results. Finally, section V concludes the paper.

II. System Model And Problem Formulation

In this section, the system model of the network is established and introduced as shown in Fig. 1. We consider an uplink communication single-cell scenario consisting of one eNB in D2D based cellular user network. There is a set of $M = \{1, 2, ..., m\}$ cellular user equipments (CUEs) located in the coverage area of an eNB and a set of $N = \{1, 2, ..., n\}$ D2D pairs within the cell. However, the randomly generated distance between a D2D pair is restricted to approximate the real case scenario, where the D2D communication method is chosen based upon the spatial proximity of two cellular users.

In this paper, we only focus on the uplink transmission of the cellular users. Considering each CUE occupies one resource block (RB), which can be shared by multiple D2D pairs and the D2D pairs could reuse the uplink resources of the CUE, from which the RB set is represented as $K = \{1, 2, ..., k\}$. Since D2D pairs multiplex the uplink transmission resource with cellular users, causing mutual interference that is experienced between CUEs and D2D pairs. The key parameter signal to interference plus noise ratio (SINR) should be analysed respectively.

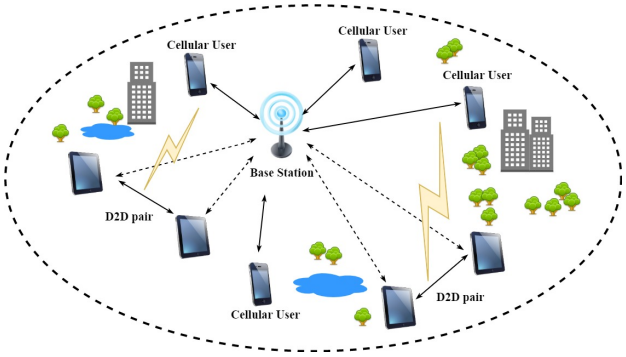

Fig. 1. The system framework of D2D communications.

If CUE share its resources with D2D pairs, it will suffer interference from the D2D pairs. we define the SINR of the mth CUE at the kth RB as:

$$\gamma_m^C = \frac{p_m^C \cdot g_{m,bs}^C}{\sigma^2 + \sum_{i=1}^{N} p_i^D \cdot g_{i,bs}^C} \tag{1}$$

where p_m^C is the mth CUE uplink transmission power on the kth RB while p_i^D specifies the transmission power of ith D2D pairs on the kth RB. $g_{m,bs}^C$ and $g_{i,bs}^C$ indicate the channel gain on the kth RB from BS to mth CUE and ith D2D transmitter respectively. σ^2 is the zero-mean additive white Gaussian noise (AWGN) power variance.

Similarly, when the D2D pairs share the kth RB, the interference is caused by reusing the RB from the co-channel mth CUE. The SINR of the ith D2D pair at the kth RB is denoted as:

$$\gamma_i^D = \frac{p_i^D \cdot g_{i,i}^D}{\sigma^2 + p_m^C g_{m,i}^C + \sum_{j=1, j\neq i}^{N} p_j^D \cdot g_{j,i}^D} \tag{2}$$

where p_i^D specifies the transmission power of ith D2D user on the kth RB. $g_{i,i}^D$, $g_{m,i}^C$ and $g_{j,i}^D$, respectively, represent the link gain of the channel over the kth D2D link, from cellular transmitter m to D2D receiver i, and from D2D transmitter j to D2D receiver i, communicating over the kth resource block.

g is the gain from the transmitter to the receiver, which can be expressed as follows:

$$g = 10^{(-PL - shadowing)/10} \tag{3}$$

PL is the path loss between the transmitter and receiver.

In this paper, we consider the system performance, including system data rate and system energy efficiency (EE), where D2D users and CUEs coexist. Since the premise of the power control scheme is to guarantee the QoS of cellular users, the resource allocation and transmission power of cellular users are given and fixed. Therefore, based on the above description, we aim to maximise the system energy efficiency while guaranteeing the minimum QoS requirements of all users.

The joint optimisation resource allocation and power control problem can be formulated as the maximisation of the system energy efficiency under the QoS constraints of all CUEs and D2D users, which can be realised by solving the problem 4:

$$\begin{aligned} \arg\max_{P,k} & \sum_{k=1}^{K} \left\{ \log_2\left(1+\gamma_m^C\right) + \sum_{i\in\Re_k} \log_2\left(1+\gamma_i^D\right) \right\} \\ \text{S.t. } & \gamma_m^C \geq \tau_0 \\ & 0 \leq p_i^D \leq p_{\max}, \forall i, k \end{aligned} \tag{4}$$

where $p_{\max}$ is the maximum transmission power for each D2D pair and τ_0 is the minimum SINR of CUEs. It is obvious that p_i^D is the objective function as increasing the transmit power of D2D will cause more interference to other D2D user pairs and CUEs. If decrease the transmit power, it will decrease the overall throughput of the system.

The maximum data transmission rate in bits per second for each D2D pair is:

$$r = \frac{B}{2} \log\left(1 + \frac{P_C |h|^2}{B N_0}\right) \tag{5}$$

where B is the channel bandwidth and the one-half factor is the natural result of consuming two slots for transmission. P_C is the transmission power for each cellular user. N_0 stands for the power spectral density of the AWGN channel.

h is the channel coefficient which is expressed as:

$$|h|^2 = \frac{|h_0|^2}{\mathcal{PL}_C \cdot d^{\varphi}} \tag{6}$$

where h_0 follows a complex normal distribution $\mathcal{CN}(0,1)$ and $\mathcal{PL}_C$ is the path-loss constant. φ is denoted as the path-loss exponent.

The total power consumption during the uplink transmission is:

$$P = P_{CC} + P_C \tag{7}$$

where P_{CC} is the circuit power of each cellular user, and the value of P_C is the same for all the cellular users.

Therefore, the energy efficiency formula is denoted as:

$$EE = \frac{r}{P} = \frac{\frac{B}{2}\log\left(1+\frac{P_C|h|^2}{BN_0}\right)}{P_{CC}+P_C} \tag{8}$$

III. Deep Reinforcement Learning Algorithm for Resource Allocation and Power Selection

In our settings, a 500-meter radius single cell with the eNB locates in the centre of the cell while CUEs and D2D users are distributed randomly in the cell. The default number of CUEs and D2D pairs are 20 and 10, respectively. Snapshot for the distribution of CUEs and D2D Users in a single cell is illustrated in Fig. 2. We abstract available orthogonal communication bands as the communication resources, each resource is considered as an ideal non-interfering block. There are 25 resource blocks in default, which is larger than the number of CUEs while is less than the total number of users in the cell. Under such context, there are at least 5 resource blocks that are needed to be shared with other users, which will cause interference. Each CUE will be assigned with one

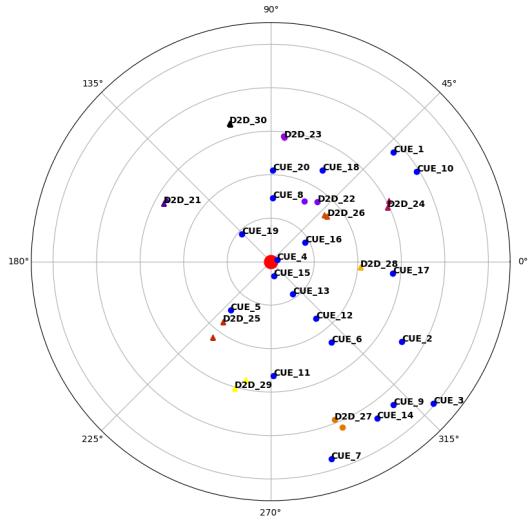

Fig. 2. Snapshot for CUEs and D2D users distribution in the cell with radius 500m where M = 20 and N = 10.

static resource block, while D2D users can be assigned with a new resource block based on the policy.

Similar to [30], the minimum QoS requirements for cellular users and D2D users are set to be 0 and 5 dB, respectively. In the proposed algorithm, the actor-network of PPO is a neural network with three hidden layers. The number of neurons in the PPO hidden layers are 64, 128, and 64, respectively, and the neurons are activated by the Rectifier Linear Unit (ReLU) function. The hyperparameters of our algorithm are listed in Table. I.

A. *Reinforcement Learning*

As the resource allocation is non-convex and non-linear, the reinforcement learning implements agents to interact with the environment constantly to find the optimal policy, which could be a feasible solution for the optimisation problem. Reinforcement learning is based on the Markov Decision Process (MDP). The standard MDP can be represented by a five-tuple $< S, A, P, R, \gamma >$ where S, A and R denote the sets of states, actions and rewards. P is the transition probability, describing the probability when the agent takes the action from the state s to the new state s'. γ is the discount factor which $\gamma \in (0, 1)$.

Each agent learns and makes the decision by interacting with the environment, and a strategy π, which is defined as the process of choosing actions from the initial state, is updated to obtain the optimal policy.

The policy is a function that decides the action selection with the given state. $V^{\pi}(s)$ denotes the state-value function, which denotes a cumulative discounted reward.

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\left[\sum_{t=0}\gamma^{t} r_{t}\left(s_{t}, a_{t}\right) \mid s_{0}=s, \pi\right] \\ &= E_{\pi}\left[r\left(s, a_{t}\right)+\gamma \sum_{s' \in S} P\left(s' \mid s, a_{t}\right) V^{\pi}\left(s'\right)\right] \end{aligned} \tag{9}$$

The optimal policy $V^{*}(s)$ satisfies the Bellman equation and is the maximum value of cumulative discounted reward:

$$V^{*}(s)=\max_{a \in A}\left\{E_{\pi^{x}}\left[r(s, a)+\gamma \sum_{S' \in S} P\left(s' \mid s, a\right) V^{\pi^{x}}\left(s'\right)\right]\right\} \tag{10}$$

We use the MDP to find the optimal policy which contributes to the resource management scheme in D2D communication cellular networks.

The basic parameters of our learning system are as follows:

- Agent

In this system, eNB is able to observe the communication related information as a whole and is responsible for allocating communication resources. Therefore, eNB is trained to manage the cell communication information as an agent. In this mode, the observations and the actions of the agent are the sets of all D2D pairs. The eNB can observe the global information and give a global decision, which is in the form of an array of observations or actions for or from each D2D transmitter. For the independent mode algorithm that is used to test and validate, each D2D user is the agent. Different from the eNB agent, global information can be observed by each D2D user, and individual actions will be made by each agent. After all the decisions are made, the actions will be made to the environment to obtain a new state.

- State

At time t, the mth CUE state $s_t^m \in S_t$ contains the information of assigned resource block assigned to each D2D agent in $t-1$, the user's QoS satisfaction degree. The state can be defined as:

$$s^{m}(t)=\{\gamma_{i}^{D}, P_{inter}^{m,k}, \xi^{m}(t)\} \tag{11}$$

which is a set of D2D link SINR on kth resource block. P_{inter} is defined as:

$$P_{inter}^{m,k}=p_{n}^{C} g_{n,i}^{C}+\sum_{j=1, j \neq i}^{N} p_{j}^{D} \cdot g_{j,i}^{D} \tag{12}$$

A continuous variable $\xi^{m}(t)$ representing the user's QoS satisfaction degree is computed by:

$$\xi^{m}(t)=\begin{cases} \frac{R_{m}(t)}{R_{\min}} & R_{m}(t)<R_{\min} \\ 1.0 & R_{m}(t) \geq R_{\min} \end{cases} \tag{13}$$

where $R_m(t)$ represents the system throughput of the mth CUE at time t. R_{min} is the minimum throughput requirements of the mth CUE. $R_m(t)/R_{min}$ denotes the QoS satisfaction

degree of the mth CUE, and $R_m(t)/R_{min} = 1.0$ represents that the system throughput of the mth CUE achieves basic QoS satisfaction degree.

- Action

In our algorithm, the agent needs to determine the power transmission and communication resource allocation strategy for optimising the system energy efficiency while ensure the QoS requirement of CUEs. The action is a set of transmission power decisions for each D2D transmitter, which is defined as:

$$a^m(t) = \{a^p(t), a^k(t)\}, \forall p \in P, \forall k \in K, \forall n \in N \tag{14}$$

Where P represents the power level set from which the D2D transmitter can choose, K stands for the available resource block, and N stands for the total number of D2D pairs.

- Reward Function

In the considered optimisation problem, the goal is to achieve the maximum system energy efficiency while reducing interference on cellular users sharing the same resource block. Therefore, the violation of the last rule should be negatively correlated with the value of the reward function while positively correlated with the system throughput.

The reward function is defined as:

$$r(t) = \alpha_1 \sum_{n\in N} \left(I_n^{\text{QoS}}(t)\right) + \alpha_2 \sum_{n\in N} \left(I_n^{\text{self}}(t)\right) + \alpha_3 \sum_{n\in N} \left(I_n^{\text{switch}}(t)\right) + EE \tag{15}$$

where part 1 is the reward of the overall system throughput, part 2 indicates the penalty term of the unsatisfied latency and unsatisfied reliability of D2D link, part3 and part 4 are the cost functions in terms of the unsatisfied minimum sum data rate requirements of cellular link and D2D link, respectively. Where E_{QoS} contains information about whether the QoS of the cellular user device in the same resource block satisfies the minimal requirement γ_{min}. It can be defined as:

$$I_n(t) = \begin{cases} 0 & \gamma_m^C \geq \gamma_{min} \\ 1 & \gamma_m^C < \gamma_{min} \end{cases} \tag{16}$$

If the SINR of the cellular user is less than the minimum requirement, a punishment factor would be added for the learning process of the reinforcement learning algorithm.

Normally, for the reinforcement learning algorithm, it is not wise to subtract a constant value in the reward function, which may influence the agent to find a strategy to end the game as fast as possible. In our case, however, the reward is not sparse, and the agent can get a reward on each step. In addition, we fix the number of steps for each episode, and therefore, it is plausible to subtract a constant value from the reward function, which may lead to a faster rate of convergence.

B. *Deep Q-Network*

Deep Q-Network integrates neural networks on the basis of Q-learning. The neural network is used as an approximation to find the optimal behaviour value function, which is expressed as:

$$Q(s_t, a_t; \Theta) \approx Q(s_t, a_t) \tag{17}$$

Θ is the neural network weight and s_t and a_t specify the states and actions respectively.

In the iterative process, the neural network minimises the loss function by updating the neural network weights:

$$\text{Loss}(\theta) = \frac{1}{n} \sum_{a_t=1}^{n} \left(y - Q(s_t, a_t)\right)^2 \tag{18}$$

where

$$y = r_t + \max Q(s_t, a_t) \tag{19}$$

r_t is the corresponding reward.

Deep Q-Network algorithm could store the previous experiences (state, action and reward) into memory. In each learning process, the random extraction of previous experiences could disrupt the correlation relationship between each experience.

C. *Proximal Policy Optimisation Method*

Proximal Policy Optimisation (PPO) is an on-policy and model-free algorithm which belongs to the policy gradient methods. PPO proposes a new objective function that can be updated in multiple training steps in small batches, which solves the problem that the step size is difficult to determine in the Policy Gradient algorithm. Moreover, PPO can support continuous input and output at a fast convergence speed. The detailed PPO algorithm is shown in Algorithm 1.

TABLE I
SIMULATION PARAMETERS

Parameter	Definition	Value
M	Number of CUEs	20
N	Number of D2D pairs	10
K	Number of Resource blocks	25
P	Maximum Transmission Power	250 mW
p_c	Circuit Power	150 mW
d	Distance	50 meters
B	Bandwidth	10 MHz
N_0	Noise	-174 dBm/Hz
$\rho\mathcal{L}_{\text{BS}}$	Base Station Path-loss Constant	$15.3 + 37.6 \log_{10}^{d}$
$\rho\mathcal{L}_{\text{User}}$	User Path-loss Constant	$28 + 40 \log_{10}^{d}$
φ	Path-loss Exponent	4
$\lvert h_0 \rvert$	The Channel Coefficients	$\mathcal{CN}(0, 1)$
γ_{PPO}	The Discount Factor	0.9
γ_{DQN}	The Reward Discount	0.99
α_{Actor}	Learning Rate for Actor	0.0001
α_{Critic}	Learning Rate for Critic	0.0002
ϵ	Update Batch Size	32

IV. SIMULATION RESULTS

In this section, the simulation parameters are illustrated in Table I. The parameters are mostly based on [31], [32] and [33]. These three articles have achieved great optimisation results in power control and share relatively similar simulation parameters. In this paper, we assume there is only one eNB within the single-cell surrounding by 20 CUEs and 10 D2D

Algorithm 1 Proposed Algorithm for Resource Block Selection and Power Control

Initialise the global parameters θ_p, θ_v and thread parameters ${\theta_p}'$, ${\theta_v}'$.
Initialise the game episode $k = 0$, the step in an episode $t = 0$, the update steps T_{update} and maximum steps $steps$.
Initialise maximum global shared resource allocation episode counter $K_{allocate}$ and power selection episode counter K_{power}..
Initialise the action-value function network $Q(s_t, a_t)$ with random weights ω.
Initialise the target network $Q(s_t, a_t)$ with weights $\omega' = \omega$.
$k \geq k_{allocate}$ k in $K_{allocate}$
t in steps
Get a(t) by policy $\pi\left(a(t) \mid s(t); \theta_p'\right)$.
Perform the a(t) and Update the environment
Obtain the immediate reward r(t) and the next state s(t+1).
t = T_{update} Get a(t) by policy $\pi\left(a(t) \mid s(t); \theta_p'\right)$.
Get advantage estimates $A(s(t), a(t); \theta_p, \theta_v)$.
Update thread parameters ${\theta_p}'$ and ${\theta_v}'$ by maximising the PPO objectives:

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{|\mathcal{D}_k| T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min$$

$$\operatorname{clip}\left(r_k\left({\theta_p}'\right), 1-\varepsilon, 1+\varepsilon\right) A(s(k), a(k); \theta_p, \theta_v)$$

via stochastic gradient ascent with Adam Update global parameters θ_p and θ_v.
$k \geq K_{power}$ k in K_{power}
t in steps
Get a(t) using ϵ-greedy policy from $\max Q(s_t, a_t)$.
Obtain the immediate reward r(t) and the next state s(t+1).
Store transition $(s(t), a(t), r(t), s(t+1))$ to form a experience replay buffer
Sample random minibatch of transitions $(s(t), a(t), r(t), s(t+1))$ from the buffer
Update DQN policy with Equation(19)

pairs sets. The number of available resource blocks is 25, and all CUEs and D2D pairs move dynamically and randomly in the area of the cell. It is easy to find out the total number of CUEs is between the number of D2D pairs and resource blocks which results in the sharing process between D2D pairs and CUEs.

Assuming the available resource blocks will be assigned to the CUEs first, and the remaining resources will leave to D2D pairs. As for the situation of resource reusing, it only happens when there is no free resource block. Based on the calculation formulas given in section II, we have modelled the D2D pairs communication links and CUEs communication links with path loss and shadows to simulate the real situation.

As shown in Fig. 3, the learning process of the three approaches in terms of the reward functions when the number of D2D users is 10. It can be observed that our proposed

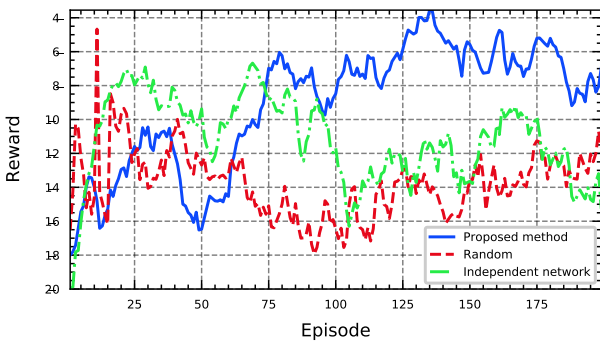

Fig. 3. Learning process comparisons of algorithms.

approaches greatly outperform the independent mode and the random one. The independent network is able to find a larger reward function in the early stage of the training process. However, it would diverge with more training episodes.

By observing actions of the independent network in each episode, we found that the independent network is able to find sub-optimal resource block choices. As this scenario is generally better than the original resource block assignment, the reward increases. However, with more training done to the algorithm, a better choice, an unused resource block or resource block occupied by a cellular user that is far away from all the other D2D devices, will come up. However, this is catastrophic when some D2D devices would choose the same resource block at the same time, and as they reuse the same resource block at the same time, the opposite reward would be delivered to each independent network. Two plausible solutions are proposed and tested to improve the independent network:

- By adding negative element when reusing the same resource block with other D2D devices and positive element when the agent is the only D2D device use the resource block with the reward function.
- By sharing the network parameters for a faster convergence rate.

The first method has similar results as in Fig. 3. This is due to two main reasons: the more carefully-designed reward function has a similar element. The reward on whether reuse the same resource block with other devices can directly influence the throughput of all devices that use this resource block, and all the agents choose the same action at the same time without communication. This means that the element is redundant and cannot help to solve the convergence problem. The second method, similarly, cannot convergent and will lead to a stable, sub-optimal value. As the definition of the Markov process, the action is abstracted and regarded as an integrated input to the environment at the same time. Therefore, even with the same parameters, agents have great chances when finding an optimal resource block and choose the resource block as its action at the same time, which would lead to an opposite reward.

Fig. 4 demonstrates the reward of D2D users with different algorithms versus the number of D2D users. In our proposed algorithm, it can promote D2D users to utilise the remaining available resources block to D2D users that are in high interference areas, such as a D2D pair that is close to other CUEs. However, it is worth noting that D2D users have limited resource blocks and when the number of D2D users increases, the higher chances the resource blocks will need to be reused by more than one user causing greater interference. For independent assigning mode, such an increase in the number of D2D users will cause greater chaos in the information delivered by the reward function, which will lead to a further drop in the performance.

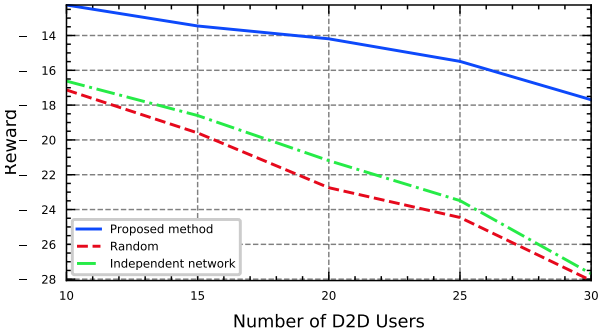

Fig. 4. Reward of D2D users versus the numbers of D2D users.

Fig. 5 shows the influence of the distance between the D2D pairs on the system throughput of different algorithms. It can be observed from the figure that the system throughput of all algorithms decreases with larger maximum D2D pairs distance. This means that a higher D2D pairs distance will directly affect the energy efficiency by affecting the system throughput, which is the denominator of equation (8). In addition, the transmission power of the D2D transmitter has to be increased to guarantee the quality of service, which will also affect the energy efficiency by increasing the numerator of the equation. Therefore, the selection of the maximum distance of D2D communication should not be too large, which may lead to low energy efficiency, nor being too small which the serviceability of D2D communication will be impaired.

V. Conclusion

To maximise the overall system throughput and energy efficiency in D2D communications, we have formulated the optimisation problem into a joint mode selection, considering both resource assignment and power control with the constraints of QoS requirements. Deep Q-Network is proposed for resource block allocation, and a proximal policy optimisation (PPO) algorithm is proposed based on MDP for energy control. With the guidance of the deep learning algorithm, D2D could make intelligent selections and thereby improve the performance of the cellular network under the premise of ensuring the communication quality of cellular users, reducing interference and improving system throughput. Experimental

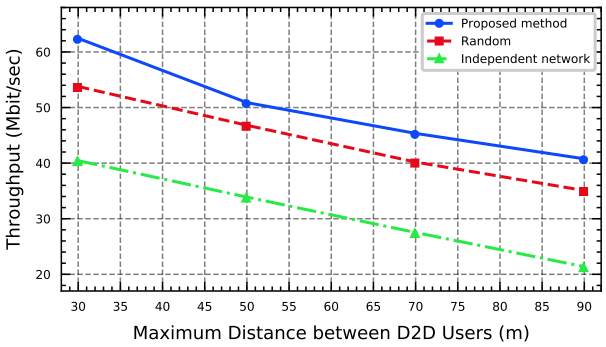

Fig. 5. Overall system throughput for different D2D users distances

results show that the algorithm has better performance than the traditional algorithm. The simulation results show that the proposed solution can efficaciously guarantee the quality of service and improve the overall throughput, which outperforms other existing algorithms by having better convergence and less executing time.

Acknowledgment

Sami Alenezi thanks Northern Border University (NBU), Saudi Arabia and the Saudi Arabia Cultural Bureau in the UK for sponsoring this study via the PhD Scholarship.

References

[1] Wang, C. X., Haider, F., Gao, X., You, X. H., Yang, Y., Yuan, D., Aggoune, H. M., Haas, H., Fletcher, S., and Hepsaydir, E., "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, 2014.

[2] Doppler, K., Rinne, M., Wijting, C., Ribeiro, C. B., and Hug, K., "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, 2009.

[3] Feng, D., Lu, L., Yuan-wu, Y., Li, G. Y., Feng, G., and Li, S., "Device-to-Device Communications Underlaying Cellular Networks," vol. 61, no. 8, pp. 3541–3551, 2013.

[4] Gandotra, P. and Jha, R. K., "Device-to-Device Communication in Cellular Networks: A Survey," *Journal of Network and Computer Applications*, vol. 71, no. 4, pp. 99–117, 2016.

[5] Zhao, W. and Wang, S., "Resource Sharing Scheme for Device-to-Device Communication Underlaying Cellular Networks," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4838–4848, 2015.

[6] Rathi, R. and Gupta, N., "A review of D2D communication with game-theoretic resource allocation models," *Proceedings - 2017 International Conference on Next Generation Computing and Information Systems, ICNGCIS 2017*, pp. 153–158, 2018.

[7] Huang, J., Sun, Y., and Chen, Q., "GALLERY: A Game-Theoretic Resource Allocation Scheme for Multicell Device-to-Device Communications Underlaying Cellular Networks," *IEEE Internet of Things Journal*, vol. 2, no. 6, pp. 504–514, 2015.

[8] Dun, H., Ye, F., and Jiao, S., "A Noval Fast Resource Allocation Scheme for D2D- enabled Cellular Networks," pp. 2020–2021, 2020.

[9] Lee, J., Member, S., and Lee, J. H., "Performance Analysis and Resource Allocation for Cooperative D2D Communication in Cellular Networks With Multiple D2D Pairs," vol. 23, no. 5, pp. 2019–2022, 2019.

[10] Mishra, P. K., Kumar, A., and Pandey, S., "Minimum Interference Based Resource Allocation Method in Two-Hop D2D Communication for 5G Cellular Networks," no. Iciss, pp. 1191–1196, 2017.

[11] Mao, Q., Hu, F., and Hao, Q., "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 4, pp. 2595–2621, 2018.

[12] Hussain, F., Hassan, S. A., Hussain, R., and Hossain, E., "Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 2, pp. 1251–1275, 2020.

[13] Feki, S., Belghith, A., and Zarai, F., "A reinforcement learning-based radio resource management Algorithm for D2D-based V2V communication," *2019 15th International Wireless Communications and Mobile Computing Conference, IWCMC 2019*, pp. 1367–1372, 2019.

[14] Yu, S., Jeong, Y. J., and Lee, J. W., "Resource Allocation Scheme Based on Deep Reinforcement Learning for Device-to-Device Communications," *International Conference on Information Networking*, vol. 2021-Janua, pp. 712–714, 2021.

[15] Chen, Q., Zhao, S., and Shao, S., "QoS-based resource allocation scheme for Device-to-Device (D2D) communication underlaying cellular network in uplink," *2013 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2013*, pp. 13–16, 2013.

[16] Yu, S., Jeong, Y. J., and Lee, J. W., "Resource Allocation Scheme Based on Deep Reinforcement Learning for Device-to-Device Communications," pp. 712–714, 2021.

[17] Gengtian, S., Koshimizu, T., Saito, M., Zhenni, P., Jiang, L., and Shimamoto, S., "Power Control Based on Multi-Agent Deep Q Network for D2D Communication," *2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020*, pp. 257–261, 2020.

[18] Sun, Z. and Nakhai, M. R., "Channel Selection and Power Control for D2D Communication via Online Reinforcement Learning," pp. 1–6, 2021.

[19] Ji, Z., Kiani, A. K., Qin, Z., and Ahmad, R., "Power Optimization in Device-to-Device Communications : A Deep," vol. 10, no. 3, pp. 508–511, 2021.

[20] Yuan, Y., Yang, T., Feng, H., and Hu, B., "An iterative matching-stackelberg game model for channel-power allocation in D2D underlaid cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7456–7471, 2018.

[21] Lai, W. K., Wang, Y. C., Lin, H. C., and Li, J. W., "Efficient Resource Allocation and Power Control for LTE-A D2D Communication with Pure D2D Model," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 3202–3216, 2020.

[22] Guizani, Z. and Hamdi, N., "Spectrum resource block reuse and power assignment for D2D communications underlay 5G uplink network," *2016 24th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2016*, pp. 7–11, 2016.

[23] Li, Y., Gursoy, M. C., Velipasalar, S., and Tang, J., "Joint mode selection and resource allocation for D2D communications via vertex coloring," *2017 IEEE Global Communications Conference, GLOBECOM 2017 - Proceedings*, vol. 2018-Janua, no. 2, pp. 1–6, 2017.

[24] Luo, Y., Shi, Z., Zhou, X., Liu, Q., and Yi, Q., "Dynamic resource allocations based on Q-learning for D2D communication in cellular networks," *2014 11th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2014*, pp. 385–388, 2014.

[25] Jiang, F., Zhang, L., Sun, C., and Yuan, Z., "Clustering and Resource Allocation Strategy for D2D Multicast Networks with Machine Learning Approaches," no. January, 2021.

[26] Saied, A., Qiu, D., and Swessi, M., "Resource management based on reinforcement learning for D2D communication in cellular networks," *2020 International Symposium on Networks, Computers and Communications, ISNCC 2020*, 2020.

[27] Chen, W. and Zheng, J., *A Reinforcement Learning Based Joint Spectrum Allocation and Power Control Algorithm for D2D Communication Underlaying Cellular Networks*. Springer International Publishing, 2019, vol. 286, no. Ml.

[28] Shi, D., Li, L., Ohtsuki, T., Pan, M., Han, Z., and Poor, V., "Make Smart Decisions Faster: Deciding D2D Resource Allocation via Stackelberg Game Guided Multi-Agent Deep Reinforcement Learning," *IEEE Transactions on Mobile Computing*, vol. 1233, no. c, pp. 1–12, 2021.

[29] Yuan, X., Tian, H., and Fan, B., "Mobility-Aware Joint Resource Allocation and Power Allocation for D2D Communication," *IEEE Wireless Communications and Networking Conference, WCNC*, vol. 2019-April, pp. 0–5, 2019.

[30] Ding, H., Zhao, F., Tian, J., Li, D., and Zhang, H., "Ad Hoc Networks A deep reinforcement learning for user association and power control in heterogeneous networks," *Ad Hoc Networks*, vol. 102, p. 102069, 2020. [Online]. Available: https://doi.org/10.1016/j.adhoc.2019.102069

[31] Nie, S., Fan, Z., Zhao, M., Gu, X., and Zhang, L., "Q-learning based power control algorithm for D2D communication," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, 2016.

[32] Zhao, M., Wei, Y., Song, M., and Da, G., "Power Control for D2D Communication Using Multi-Agent Reinforcement Learning," *2018 IEEE/CIC International Conference on Communications in China, ICCC 2018*, no. Iccc, pp. 563–567, 2019.

[33] Sheng, M., Li, Y., Wang, X., Li, J., and Shi, Y., "Energy efficiency and delay tradeoff in device-to-device communications underlaying cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 92–106, 2016.