EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems

Research Article EAI.EU

Deep Reinforcement Learning for Intelligent Reflecting Surface-assisted D2D Communications

Khoi Khac Nguyen[1], Antonino Masaracchia[1], Cheng Yin[2]

[1]Queen's University Belfast, UK
[2]University of Surrey, UK

Abstract

In this paper, we propose a deep reinforcement learning (DRL) approach for solving the optimisation problem of the network's sum-rate in device-to-device (D2D) communications supported by an intelligent reflecting surface (IRS). The IRS is deployed to mitigate the interference and enhance the signal between the D2D transmitter and the associated D2D receiver. Our objective is to jointly optimise the transmit power at the D2D transmitter and the phase shift matrix at the IRS to maximise the network sum-rate. We formulate a Markov decision process and then propose the proximal policy optimisation for solving the maximisation game. Simulation results show impressive performance in terms of the achievable rate and processing time.

Received on 17 November 2022; accepted on 25 December 2022; published on 03 January 2023

Keywords: Intelligent reflecting surface (IRS), D2D communications, deep reinforcement learning.

doi:10.4108/eetinis.v10i1.2864

1. Introduction

Device-to-device (D2D) communications play a critical role in 5G networks by allowing users to communicate directly without the involvement of base stations. It helps reduce the latency and improve the information transmission efficiency [1, 2]. In [1], the D2D transmitters harvest energy through the simultaneous wireless information and power transfer protocol (SWIPT). Then, a game theory approach was proposed to solve the power allocation and power splitting at SWIPT with pricing strategies for maximising the network performance. In [2], the optimised power allocation was proposed to maximise the energy efficiency (EE) performance at the D2D-based vehicle-to-vehicle communications, by following a machine learning-based approach. Authors in [3] proposed a three stage wireless energy harvesting protocol for a relay-assisted network in a cognitive spectrum sharing paradigm. For the considered network scenario and algorithm they provided a closed form expression for the outage probability. Subsequently, through computer simulations they showed how the most relevant parameters like the energy harvesting constraint, the interference power constraints on the primary user network, and an interference imposed by primary user network on the secondary user cognitive network, impact on the outage probability.

Intelligent reflecting surface (IRS), referring to the technology of massive elements of flexible reflection capability controlled by an intelligent unit, has recently attracted great attention from the research community as an efficient means to expand wireless coverage. The IRS can manage the incoming signal by a controller, which allows to efficiently adapt the angle of passive reflection from the transmitters toward the receivers [4–7]. In [5], the IRS harvests energy from the access point (AP) and uses it for reflecting the signal in two phases. The AP beamforming vector, the IRS's phase scheduling, and the passive beamforming were optimised to maximise the information rate. In [6], a channel estimation scheme for a multi-user multiple-input multiple-output (MIMO) system has been designed with the support of double IRS panels.

Some research works have investigated the efficiency of the IRS in assisting the D2D communications [8–12]. In [8] and [9], two sub-problems with fixed passive beamforming vector and fixed phase shift matrix were considered. To solve the power allocation optimisation with the fixed phase shift matrix, the authors in [8] used the gradient descent method while the authors in [9] employed the Dinkelbach method. For the phase shift optimisation, a local search algorithm was proposed in [8] while fractional programming was utilised in [9]. However, these approaches assume a discrete phase shift and only reach a sub-optimal solution. Moreover, these works only consider perfect conditions, e.g., channel state information (CSI). In addition, these algorithms cause large delays due to high computational complexity.

Very recently, deep reinforcement learning (DRL) has been applied as an effective solution for solving complicated problems in wireless networks [13–17]. In [2], the DRL algorithm was used to choose the continuous transmit power level at the D2D transmitters for maximising the EE performance. In [14], discrete and continuous action spaces were considered for the beamforming vector and the IRS phase shift in multiple-input single-output (MISO) communications. Then, two DRL algorithms were used to maximise the total throughput. In [15], a method based on the DRL was used for optimising the unmanned aerial vehicle (UAV)'s altitude and the IRS diagonal matrix to minimise the sum age-of-information. In [16], the authors used the DRL technique to maximise the signal-to-noise ratio.

Solving the joint optimisation of power allocation and IRS configuration results to be a challenging problem. The traditional optimisation approaches mostly focus on solving sub-problems [8–12]or considering a discrete phase shift matrix at RIS [8, 10] to reduce the complexity. In contrast, as already mentioned before, the adoption of DRL based algorithms represents a very powerful and efficient approach for solving non-convex and complex problems. In this paper, we propose a DRL algorithm for solving the joint power allocation and phase shift matrix optimisation in IRS-assisted D2D communications. Firstly, we conceive a D2D communication system with the support of the IRS. The D2D channel is a combination of the direct link and the reflective link. In this context, the IRS is used to mitigate the channel interference, as well as to enhance the information transmission. Secondly, we formulate a Markov decision process (MDP) [18] for the network throughput maximisation in the IRS-assisted D2D communications, in which the optimisation variables are the power at the D2D users and the phase shifts at the IRS. In this paper, we characterise the continuous action space and propose an on-policy algorithm to search for an optimal policy for maximising the network sum-rate. Therefore, we reduce the human intervention for designing the discrete variables, reduce neural networks' size, and train them better in centralised learning. Finally, we compare the efficiency of our proposed methods with other schemes in terms of the achievable network sum-rate.

2. System Model and Problem Formulation

We consider an IRS-assisted wireless network with N pairs of D2D users distributed randomly and an IRS panel, as shown in Fig. 1. Each pair of D2D users comprises of a single-antenna D2D transmitter (D2D-Tx) and a single-antenna D2D receiver (D2D-Rx). An IRS panel with K reflective elements is deployed to enhance the signal from the D2D-Tx to the associated D2D-Rx and mitigate the interference from other D2D-Txs. The IRS with reflective elements maps the receiver's signal by the value of the phase shift matrix controlled by an intelligent unit. The received signal at the D2D-Rx is composed of a direct signal and a reflective one.

We denote the position of the nth D2D-Tx at time step t as $X_n^t(\mathrm{Tx}) = \left(x_n^t(\mathrm{Tx}), y_n^t(\mathrm{Tx})\right), n = 1, \ldots, N$ and that of the ℓth D2D-Rx as $X_\ell^t(\mathrm{Rx}) = \left(x_\ell^t(\mathrm{Rx})), y_\ell^t(\mathrm{Rx})\right), \ell = 1, \ldots, N$. The IRS is fixed at the position $(x_{\mathrm{IRS}}^t, y_{\mathrm{IRS}}^t, z_{\mathrm{IRS}}^t)$. The phase shift value of each element in the IRS belongs to $[0, 2\pi]$.

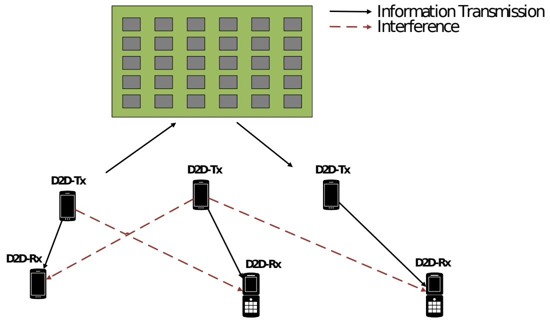

Figure 1. System model of the IRS-assisted D2D communications.

We denote the direct channel from the nth D2D-Tx to the ℓth D2D-Rx at time step t by $h_{n\ell}^t$, and the reflective channel by $H_{n\ell}^t$. The phase shift matrix at the IRS at time step t is defined by $\mathbf{\Phi}^t = \mathrm{diag}(\eta_2^t e^j\theta_1^t, \eta_2^t e^j\theta_2^t, \ldots, \eta_K^t e^j\theta_K^t)$, where $\eta_k^t \in [0, 1]$ and $\theta_k^t \in [0, 2\pi]$ represent the reflection amplitude and the phase shift value, respectively; j is the imaginary unit. In this paper, we assume that the amplitudes of all elements are set to $\eta_k^t = 1$.

The distance between the nth D2D-Tx and the ℓth D2D-Rx at time step t is defined as

$$d_{n\ell}^t = \sqrt{\left(x_n^t(\mathrm{Tx}) - x_\ell^t(\mathrm{Rx})\right)^2 + \left(y_n^t(\mathrm{Tx}) - y_\ell^t(\mathrm{Rx})\right)^2}. \quad (1)$$

Similarly, the distance between the nth D2D-Tx and the IRS is $d_{n,IRS}^t$ and the distance between the IRS and the ℓth D2D-Rx is $d_{IRS,\ell}^t$ at time step t. The direct channel is formulated as

$$h_{nm}^t = \sqrt{\beta_0 (d_{n\ell}^t)^{-\kappa_0}}, \quad (2)$$

where β_0 is the channel power gain at the reference distance $d_0 = 1$ m and κ_0 is the path-loss exponent in the line-of-sight (LoS) case. Here, we assume that the small-scale fading follows the Nakagami-m distribution with m as the fading severity parameter.

The channel gain between the nth D2D-Tx and the IRS can be written as

$$h^t_{n,IRS} = \sqrt{\beta_0 (d^t_{n,IRS})^{-\kappa_1}} \left(\sqrt{\frac{\vartheta}{1+\vartheta}} \tilde{h}^{LoS}_{n,IRS} + \sqrt{\frac{1}{\vartheta+1}} \tilde{h}^{NLoS}_{n,IRS} \right) \tag{3}$$

where κ_1 is the path loss exponent, ϑ is the Rician factor; $\tilde{h}^{LoS}_{n,IRS}$ and $\tilde{h}^{NLoS}_{n,IRS}$ are the LoS and the non-line-of-sight (NLoS) components for the D2D-Tx and the IRS link, respectively. Specifically, the deterministic LoS component is defined as $\tilde{h}^{LoS}_{n,IRS} = [1, e^{-j\frac{2\pi}{\lambda}d\cos(\phi^t_{AoA})}, \ldots, e^{-j\frac{2\pi}{\lambda}d(K-1)\cos(\phi^t_{AoA})}]$, where d and λ are the IRS's element spacing and the carrier wavelength, respectively; $\cos(\phi^t_{AoA})$ is the cosine of the angle of arrival(AoA). The NLoS component $\tilde{h}^{NLoS}_{n,IRS} \sim \mathcal{CN}(0,1)$ follows i.i.d. complex Gaussian distribution with zero mean and unit variance. Similarly, the channel gain between the IRS and the ℓth D2D-Rx is $h_{IRS,\ell}$. The reflective channel via the IRS from the nth D2D-Tx toward the ℓth D2D-Rx at time step t is described by $H^t_{n\ell} = h^t_{n,IRS}\mathbf{\Phi}h^t_{IRS,\ell}$.

The received signal at the nth D2D-Rx at time step t can be written as

$$\begin{aligned} s^t_n = & \left(h^t_{nn} + h^t_{n,IRS}\mathbf{\Phi}h^t_{IRS,n}\right)\sqrt{p^t_n}u^t_n \\ & + \sum_{\ell \neq n}^{N} \left(h^t_{\ell n} + h^t_{\ell,IRS}\mathbf{\Phi}h^t_{IRS,n}\right)\sqrt{p^t_\ell}u^t_\ell + \varpi, \end{aligned} \tag{4}$$

where p^t_n is the transmit power at the nth D2D-Tx at time step t, u^t_n is the transmitted symbol from the nth D2D-Tx, and $\varpi \sim \mathcal{N}(0, \alpha^2)$ is the complex additive white Gaussian noise.

Accordingly, the received signal-to-interference-plus-noise ratio (SINR) at the nth D2D-Rx can be represented as

$$\gamma^t_n = \frac{|h^t_{nn} + h^t_{n,IRS}\mathbf{\Phi}h^t_{IRS,n}|^2 p^t_n}{\sum_{\ell \neq n, \ell \in N} |h^t_{\ell n} + h^t_{\ell,IRS}\mathbf{\Phi}h^t_{IRS,n}|^2 p^t_\ell + \alpha^2}. \tag{5}$$

The achievable sum-rate at the nth D2D pair during time step t is defined as

$$R^t_n = B\log_2(1 + \gamma^t_n), \tag{6}$$

where B is the bandwidth.

In this paper, we aim at optimising the power allocation of all N pairs of D2D users $P = \{p_1, p_2, \ldots, p_N\}$ and the phase shift matrix $\mathbf{\Phi}$ of the IRS to maximise the network sum-rate while satisfying all the constraints. The considered network optimisation can be formulated as follows:

$$\begin{aligned} \max_{P,\mathbf{\Phi}} \quad & R^t_{total} = \sum_{n=1}^{N} R^t_n \\ s.t. \quad & 0 < p_n < P_{\max}, \forall n \in N \\ & R^t_n \geq r_{\min}, \forall n \in N \\ & \theta_k \in [0, 2\pi], \forall k \in K, \end{aligned} \tag{7}$$

where $P_{\max}$ is the maximum transmit power at the D2D-Tx and the constraint $R^t_n \geq r_{\min}, \forall n \in N$ indicates the quality-of-service (QoS) of the D2D communications.

3. Joint Optimisation of Power Allocation and Phase Shift Matrix

Given the optimisation problem (7), we formulate the MDP with the agent, the state space $\mathcal{S}$, the action space $\mathcal{A}$, the transition probability $\mathcal{P}$, the reward function $\mathcal{R}$ and the discount factor ζ. Let us denote $\mathcal{P}_{ss'}(a)$ as the probability when the agent takes action $a^t \in \mathcal{A}$ at the state $s = s^t \in \mathcal{S}$ and transfers to the next state $s' = s^{t+1} \in \mathcal{S}$. In particular, we formulate the MDP game as follows:

- *State space*: The channel gain of the D2D users forms the state space as

$$\begin{aligned} \mathcal{S} = \Big\{ & |h_{11} + h^t_{1,IRS}\mathbf{\Phi}h^t_{IRS,1}|^2, \ldots, |h_{1N} + \\ & h^t_{1,IRS}\mathbf{\Phi}h^t_{IRS,N}|^2, \ldots, |h_{n\ell} + h^t_{n,IRS}\mathbf{\Phi}h^t_{IRS,\ell}|^2, \\ & \ldots, |h_{nN} + h^t_{n,IRS}\mathbf{\Phi}h^t_{IRS,N}|^2, \ldots, |h_{N1} + \\ & h^t_{n,IRS}\mathbf{\Phi}h^t_{IRS,1}|^2, \ldots, |h_{NN} + h^t_{n,IRS}\mathbf{\Phi}h^t_{IRS,N}|^2 \Big\}. \end{aligned} \tag{8}$$

- *Action space*: The D2D-Txs adjust the transmit power and the IRS changes the phase shift for maximising the expected reward. Thus, the action space for the D2D users and the IRS is considered as follows:

$$\mathcal{A} = \{p_1, p_2, \ldots, p_N, \theta_1, \theta_2, \ldots, \theta_K\}. \tag{9}$$

- *Reward function*: The agent needs to find an optimal policy for maximising the reward. In our problem, our objective is to maximise the network sum-rate; thus, the reward function is defined as

$$\mathcal{R} = \sum_{n=1}^{N} B\log_2\left(1 + \gamma^t_n\right) \tag{10}$$

In this paper, we consider a centralised optimisation where the agent is considered as a central processor, for example, at a base station, on a powered D2D user or on the cloud. At the beginning of each time step, the agent transfers the action toward the D2D pairs and the IRS.

By following the MDP, the agent interacts with the environment and receives the response to achieve the best expected reward. Particularly, the state of the agent at time step t is $s^t \in \mathcal{S}$. The agent chooses and executes the action $a^t \in \mathcal{A}$ under the policy π. The environment responds with the reward $r^t \in \mathcal{R}$. After taking the action a^t, the agent moves to the new state s' with probability $P_{ss'}(a)$. The interactions are iteratively executed and the policy is updated for the optimal reward.

Next, we propose a DRL approach to search for an optimal policy for maximising the reward value in (10). The optimal policy can be obtained by modifying the estimation of the value function or directly by the objective. We use an on-policy algorithm for our work, namely proximal policy optimisation (PPO) with the clipping surrogate technique [19]. There are several advantages when designing the state space and action space in a continuous form. Firstly, we can reduce the human intervention while we do not need to decide the number of discrete variables. Secondly, we can reduce the size of neural networks and train them better. For example, if we have N D2D pairs with the power of each D2D pair being discretised into J level and K IRS elements with the phase shift of each element being divided into K values, we need to define the output of the action-chosen neural network by $N \times J + K \times L$ in the centralised optimisation. In the meantime, we need only $N + K$ units for the output layer in the network when we use the continuous action space. Consider the probability ratio of the current policy and obtained policy $p^t_\theta = \frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}$, we need to find the optimal policy to maximise the total expected reward as follows:

$$\mathcal{L}(s,a;\theta)=\mathbb{E}\left[\frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}A^\pi(s,a)\right]=\mathbb{E}\left[p^t_\theta A^\pi(s,a)\right], \quad (11)$$

where $\mathbb{E}[\cdot]$ is the expectation operation and $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ denotes the advantage function [20]; $V^\pi(s)$ denotes the state-value function while $Q^\pi(s,a)$ is the action-value function.

In the PPO method, we limit the current policy such that it does not go far from the obtained policy by using different techniques, e.g., the clipping technique and Kullback-Leiber [20]. In this work, we use the clipping surrogate method to prevent the excessive modification of the objective value, as follows:

$$\mathcal{L}^{\text{clip}}(s,a;\theta) = \mathbb{E}\Bigg[\min\Big(p^t_\theta A^\pi(s,a), \\ \text{clip}(p^t_\theta, 1-\epsilon, 1+\epsilon)A^\pi(s,a)\Big)\Bigg], \quad (12)$$

where ϵ is a hyperparameter.

Consider the positive value of the advantage $A^\pi(s,a)$ function and once $\pi(s,a;\theta) > (1+\epsilon)\pi(s,a;\theta_{old})$, the term $(1+\epsilon)$ takes action and the objective is limited by $(1+\epsilon)A^\pi(s,a)$. We have

$$\mathcal{L}^{\text{clip}}(s,a;\theta) = \min\left(\frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}, (1+\epsilon)\right)A^\pi(s,a). \quad (13)$$

Meanwhile, when the advantage $A^\pi(s,a)$ is negative and $\pi(s,a;\theta) < (1-\epsilon)\pi(s,a;\theta_{old})$, the term $(1-\epsilon)$ puts a ceiling to the objective value and the objective is limited by $(1-\epsilon)A^\pi(s,a)$. We have

$$\mathcal{L}^{\text{clip}}(s,a;\theta) = \max\left(\frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}, (1-\epsilon)\right)A^\pi(s,a). \quad (14)$$

Moreover, for the advantage function $A^\pi(s,a)$, we use [21]:

$$A^\pi(s,a) = r^t + \zeta V^\pi(s^{t+1}) - V^\pi(s^t), \quad (15)$$

where the state-value function $V^\pi(s)$ is obtained at the state s under the policy π as follows:

$$V^\pi(s) = \mathbb{E}\Big[\mathcal{R}|s,\pi\Big]. \quad (16)$$

To train the policy network, we store the transition into a mini-batch memory D and then use the stochastic policy gradient (SGD) method to maximise the objective. By denoting the policy parameter by θ, it is updated as

$$\theta^{i+1} = \arg\max \mathbb{E}\Big[\mathcal{L}(s,a;\theta)\Big]. \quad (17)$$

In this work, we use a policy search algorithm to search for an optimal policy π^* with the policy parameter θ_π. The PPO algorithm is an on-policy method; thus, we initialise a network for the policy π. After each interaction with the environment, the transition (s^t, a^t, r^t, s') is stored in a buffer D. Then, the policy network is trained by the SGD with Adam optimiser [22] over D samples. The policy parameters are updated by executing (17). Moreover, we use the advantage function to define the PPO objective as in (15). Thus, we define a network with the parameter ϕ_θ to calculate the value function (16). The value network parameters ϕ_θ are updated by mean-square error using the SGD algorithm as follows:

$$\phi^{i+1}_\theta = \arg\min \frac{1}{D}\sum^{D}\Big(V^\pi(s) - r\Big)^2 \quad (18)$$

The PPO algorithm for joint optimisation of the transmit power and the phase shift matrix in the IRS-aided D2D communications is presented in Algorithm 1, where M denotes the maximum number of episodes and T is the number of iterations during a period of time.

Algorithm 1 Proposed approach based on the PPO algorithm for the IRS-assisted D2D communications.

```
Initialise the policy $\pi$ with the parameter $\theta_\pi$
Initialise other parameters
for episode = $1, \ldots, M$ do
    Receive initial observation state $s^0$
    for iteration = $1, \ldots, T$ do
        Obtain the action $a^t$ at state $s^t$ by following the current policy
        Execute the action $a^t$
        Receive the reward $r^t$ according to (10)
        Observe the new state $s^{t+1}$
        Update the state $s^t = s^{t+1}$
        Collect set of partial trajectories with $D$ transitions
        Estimate the advantage function according to (15)
    end for
    Update policy parameters using SGD with mini-batch $D$
        $$\theta^{i+1} = \arg\max \frac{1}{D} \sum^{D} \mathcal{L}^{\text{clip}}(s, a; \theta^t) \quad (19)$$
    Update value network parameters $\phi_\theta$ using the SGD
        $$\phi_\theta^{i+1} = \arg\min \frac{1}{D} \sum^{D} \left( V^\pi(s) - r \right)^2 \quad (20)$$
end for
```

4. Simulation Results

For numerical results, we use Tensorflow 1.13.1 [23]. The IRS is deployed at the center $(0, 0, 0)$, while the D2D devices are randomly distributed within a circle of 100 m from the center. The maximum distance between the D2D-Tx and the associated D2D-Rx is set to 10 m. We assume $d/\lambda = 1/2$, and set the learning rate for the PPO algorithm to 0.0001. For the neural networks, we initialise two hidden layers with 128 and 64 units, respectively. All other parameters are provided in Table 1. We consider the following algorithms in the numerical results.

- **The proposed algorithm**: We use the PPO algorithm with the clipping surrogate technique to solve the joint optimisation of the power allocation at the D2D user and the IRS's phase shift matrix.

- **Maximal power transmission (MPT)**: We apply the equal power allocation for the transmission of D2D-Tx, where each D2D-Tx transmits with maximal power $P_{\max}$. We use the PPO algorithm to optimise the IRS's phase shift matrix.

- **Random phase shift matrix selection (RPS)**: We optimise the power allocation at the D2D-Tx with random selection of the phase shift matrix $\mathbf{\Phi}$.

- **Without IRS**: The D2D-Tx transmits information without the support of the IRS. We optimise the power allocation by using the PPO algorithm.

- **Vanilla policy gradient method (VPG)**[24]: We use neural networks for deploying a classical policy gradient method to optimise the power allocation of the D2D-Txs and the IRS's phase shift matrix.

Table 1. SIMULATION PARAMETERS.

Parameters	Value
Bandwidth (W)	1 MHz
Path-loss parameters	$\kappa_0 = 2.5, \kappa_1 = 3.6$
Channel power gain	−30 dB
Fading parameter	$\mu = 3$
Rician factor	$\vartheta = 4$
Noise power	$\alpha^2 = -110$ dBm
Clipping parameter	$\epsilon = 0.2$
Discount factor	$\zeta = 0.9$
Max number of D2D pairs	10
Initial batch size	$K = 128$

Firstly, we compare the achievable network sum-rate provided by our proposed algorithm with that of other schemes. Fig. 2 plots the sum-rate versus different numbers of the IRS elements, K, where the number of D2D pairs is set to $N = 5$. As can be observed from this figure, the PPO algorithm-based technique outperforms other schemes and is followed by the MPT technique. The RPS, WithoutIRS and VPG schemes show poorer performance in terms of the network sum-rate. The achievable network sum-rate using our proposed algorithm and MPT improves with increasing the number of IRS elements. The results show that with the monotonic increase in the value of K, the communication quality between the D2D-Tx and associated D2D-Rx is enhanced, while the interference from other D2D-Txs is suppressed.

Next, the performance of the previously mentioned four schemes is compared while varying the number of D2D pairs, N, in Fig. 3. We set the number of IRS element to $K = 50$ and take the average over 500 episodes to obtain the results. Our proposed algorithm shows better performance, followed by MPT. With higher number of D2D users, $N \geq 6$, the performance

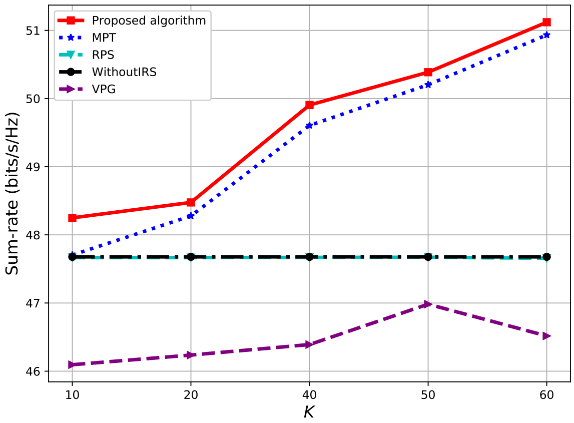

Figure 2. The network sum-rate versus the number of IRS elements, K.

attained by the proposed algorithm still stables while it decreases significantly for the other schemes. The RPS and WithoutIRS models show the worse performance.

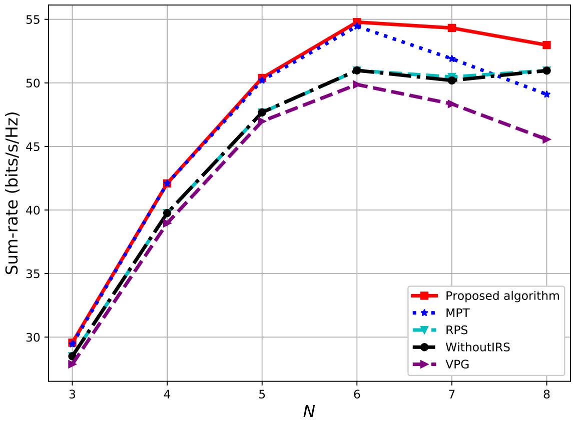

Figure 3. The network sum-rate versus the number of D2D pairs, N.

Further, we set $N = 5$, $K = 50$ and compare the performance results of the four schemes while changing the value of the threshold, $r_{\min}$, in Fig. 4. When the value of $r_{\min}$ increases towards infinity, the number of D2D pairs that satisfies the QoS constraints decreases and the sum-rate of all schemes tends to 0. The proposed algorithm outperforms the other schemes for all values of $r_{\min}$. The gap between our algorithm and others increases following the increase in $r_{\min}$ when $r_{\min} \geq 15$ dB. The MPT algorithm exhibits the worst performance when $r_{\min} = 20$ dB. This suggests that the optimisation of power allocation is important for efficient D2D communications.

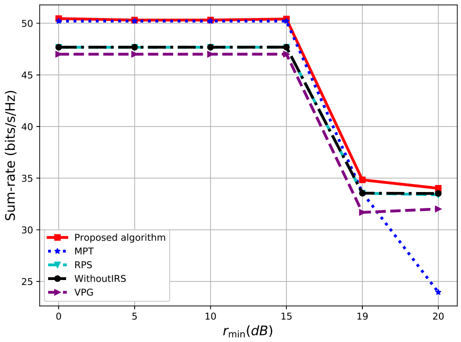

Figure 4. The network sum-rate versus the QoS threshold, $r_{\min}$.

Next, we compare the total sum-rate of the four schemes by setting different maximum transmission powers at the D2D-Tx, $P_{\max}$, in Fig. 5, with $N = 5$, $K = 50$. As $P_{\max}$ varies from 200 mW to 400 mW, the performance of the five schemes increases in the same upward trend. The gap between our proposed algorithm and the other schemes increases with the increase value of $P_{\max}$ as we jointly optimise both power allocation at the D2D-Tx and the IRS's phase shift matrix. It is clear that the proposed algorithm is more effective for mitigating interference and providing a better communication quality.

Furthermore, we use neural networks for establishing the DRL algorithm. Thus, after iterative interactions with the environment, the neural networks are trained for achieving an optimal solution. After training offline, the neural network can be deployed to the system for online execution. The online neural networks can determine the proper action for the IRS phase shift value and the D2D-Tx power allocation for maximising the network sum-rate in real-time.

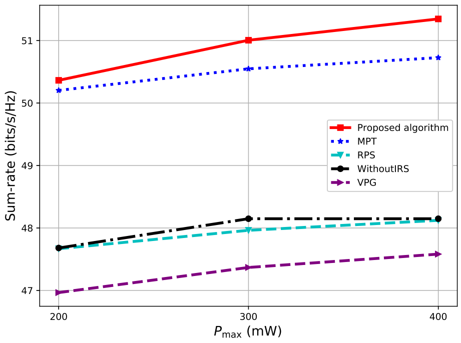

Figure 5. The network sum-rate versus the maximum transmit power, $P_{\max}$.

In Fig. 6, we compare the convergence speed of the PPO algorithm while varying the number of IRS elements, K. The PPO algorithm converges faster with the lower value of K. The slower convergence speed with the higher value of K is mainly caused due to the higher number of optimisation variables.

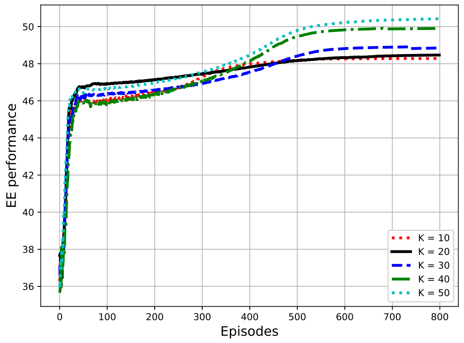

Figure 6. The network sum-rate while using the PPO algorithm.

5. Conclusion

In this paper, we have presented a DRL-based optimal resource allocation scheme for IRS-assisted D2D communications. The PPO algorithm with the clipping surrogate technique has been proposed for joint optimisation of the D2D-Tx power and the IRS's phase shift matrix. Numerical results have showed a significant improvement in the achievable network sum-rate performance compared with the benchmark schemes. Our proposed scheme demonstrates the superiority of using IRS in mitigating the interference in the D2D communications when compared with other existing schemes.

References

[1] Huang, J., Xing, C.C. and Guizani, M. (2020) Power allocation for D2D communications with SWIPT. *IEEE Trans. Wireless Commun.* **19**(4): 2308–2320.

[2] Nguyen, K.K., Duong, T.Q., Vien, N.A., Le-Khac, N.A. and Nguyen, L.D. (2019) Distributed deep deterministic policy gradient for power allocation control in D2D-based V2V communications. *IEEE Access* **7**: 164533–164543.

[3] Mousavifar, S.A., Liu, Y., Leung, C., Elkashlan, M. and Duong, T.Q. (September 2014) Wireless energy harvesting and spectrum sharing in cognitive radio. In *Proc. IEEE 80th Vehicular Technology Conference (VTC2014-Fall), Vancouver, BC, Canada*: 1–5.

[4] Yu, H., Tuan, H.D., Nasir, A.A., Duong, T.Q. and Poor, H.V. (2020) Joint design of reconfigurable intelligent surfaces and transmit beamforming under proper and improper Gaussian signaling. *IEEE J. Select. Areas Commun.* **38**(11): 2589–2603.

[5] Zou, Y., Gong, S., Xu, J., Cheng, W., Hoang, D.T. and Niyato, D. (2020) Wireless powered intelligent reflecting surfaces for enhancing wireless communications. *IEEE Transactions on Vehicular Technology* **69**(10): 12369–12373.

[6] Zheng, B., You, C. and Zhang, R. (2021) Efficient channel estimation for double-IRS aided multi-user MIMO system. *IEEE Trans. Commun.* **69**(6): 3818–3832.

[7] Nguyen, K.K., Khosravirad, S., Costa, D.B.D., Nguyen, L.D. and Duong, T.Q. (2022) Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning. *IEEE J. Selected Topics in Signal Process.* **16**(3): 358–368.

[8] Chen, Y., Ai, B., Zhang, H., Niu, Y., Song, L., Han, Z. and Poor, H.V. (2021) Reconfigurable intelligent surface assisted device-to-device communications. *IEEE Trans. Wireless Commun.* **20**(5): 2792–2804.

[9] Jia, S., Yuan, X. and Liang, Y.C. (2021) Reconfigurable intelligent surfaces for energy efficiency in D2D communication network. *IEEE Wireless Commun. Lett.* **10**(3): 683–687.

[10] Pradhan, C., Li, A., Song, L., Li, J., Vucetic, B. and Li, Y. (2020) Reconfigurable intelligent surface (RIS)-enhanced two-way OFDM communications. *IEEE Transactions on Vehicular Technology* **69**(12): 16270–16275.

[11] Cao, Y., Lv, T., Ni, W. and Lin, Z. (2021) Sum-rate maximization for multi-reconfigurable intelligent surface-assisted device-to-device communications. *IEEE Trans. Commun.* **69**(11): 7283–7296.

[12] Yang, G., Liao, Y., Liang, Y.C., Tirkkonen, O., Wang, G. and Zhu, X. (2021) Reconfigurable intelligent surface empowered device-to-device communication underlaying cellular networks. *IEEE Trans. Commun.* **69**(11): 7790–7805.

[13] Nguyen, K.K., Vien, N.A., Nguyen, L.D., Le, M.T., Hanzo, L. and Duong, T.Q. (2021) Real-time energy harvesting aided scheduling in UAV-assisted D2D networks relying on deep reinforcement learning. *IEEE Access* **9**: 3638–3648.

[14] Huang, C., Mo, R. and Yuen, C. (2020) Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE J. Select. Areas Commun.* **38**(8): 1839–1850.

[15] Shokry, M., Elhattab, M., Assi, C., Sharafeddine, S. and Ghrayeb, A. (2021) Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach. *IEEE Transactions on Vehicular Technology* **70**(4): 3978–3983.

[16] Feng, K., Wang, Q., Li, X. and Wen, C.K. (2020) Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems. *IEEE Wireless Commun. Lett.* **9**(5): 745–749.

[17] Nguyen, K.K., Duong, T.Q., Do-Duy, T., Claussen, H. and Hanzo, L. (2022) 3D UAV trajectory and data collection optimisation via deep reinforcement learning. *IEEE Trans. Commun.* **70**(4): 2358–2371.

[18] Bertsekas, D.P. (1995) *Dynamic Programming and Optimal Control*, **1** (Athena Scientific Belmont, MA).

[19] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017), Proximal policy optimization algorithms. URL https://arxiv.org/abs/1707.06347.

[20] Schulman, J., Moritz, P., Levine, S., Jordan, M.I. and Abbeel, P. (2016) High-dimensional continuous control using generalized advantage estimation. In *Proc. 4th International Conf. Learning Representations (ICLR)*.

[21] Mnih, V. *et al.* (2016) Asynchronous methods for deep reinforcement learning. In *Proc. Int. Conf. Mach. Learn.* (PMLR): 1928–1937.

[22] Kingma, D.P. and Ba, J.L. (2014), Adam: A method for stochastic optimization. URL arXivpreprintarXiv:1412.6980.

[23] Abadi, M. *et al.* (2016) Tensorflow: A system for large-scale machine learning. In *Proc. 12th USENIX Sym. Opr. Syst. Design and Imp. (OSDI 16)*: 265–283.

[24] Sutton, R.S., McAllester, D., Singh, S. and Mansour, Y. (2000) Policy gradient methods for reinforcement learning with function approximation. In *Adv. Neural Inf. Process. Syst.*: 1057–1063.