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

Initialise the policy π with the parameter θ_π
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

$$\theta^{i+1} = \arg\max \frac{1}{D} \sum^{D} \mathcal{L}^{\text{clip}}(s, a; \theta^t) \tag{19}$$

  Update value network parameters ϕ_θ using the SGD

$$\phi_\theta^{i+1} = \arg\min \frac{1}{D} \sum^{D} \left(V^\pi(s) - r \right)^2 \tag{20}$$

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

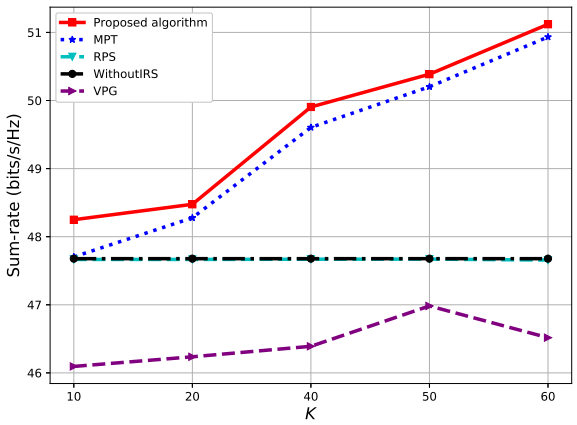

Figure 2. The network sum-rate versus the number of IRS elements, K.

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

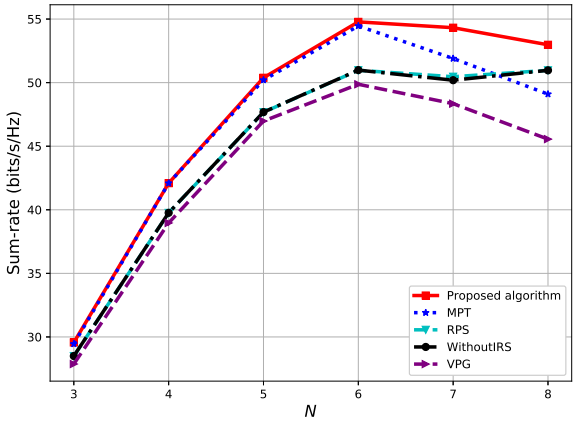

Figure 3. The network sum-rate versus the number of D2D pairs, N.

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.

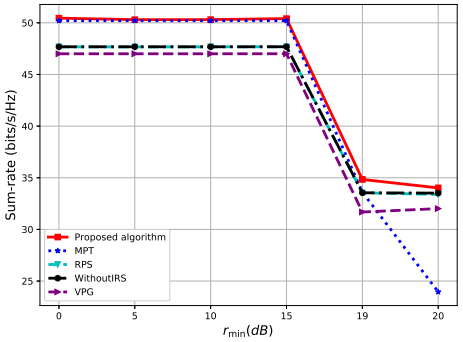

Figure 4. The network sum-rate versus the QoS threshold, $r_{\min}$.

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.

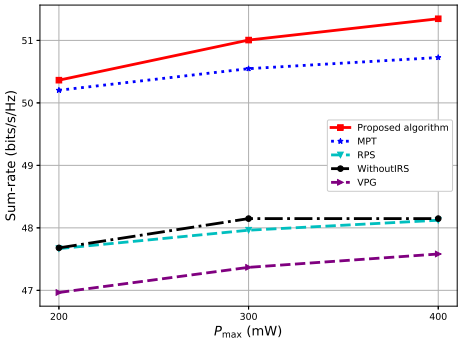

Figure 5. The network sum-rate versus the maximum transmit power, $P_{\max}$.

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.

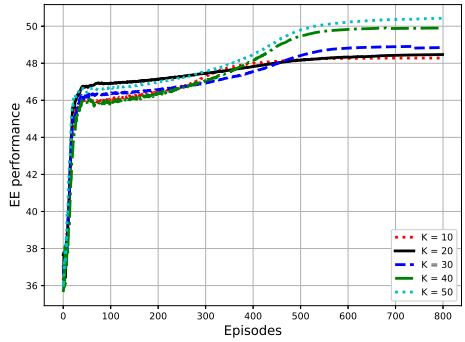

Figure 6. The network sum-rate while using the PPO algorithm.

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.