future internet

Article

Intelligent Reflecting Surface-Aided Device-to-Device Communication: A Deep Reinforcement Learning Approach

Ajmery Sultana † and **Xavier Fernando ***

Department of Electrical, Computer and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada
* Correspondence: fernando@ryerson.ca
† Current address: Department of Computer Science, Algoma University, Brampton, ON L6V 1A3, Canada.

Abstract: Recently, the growing demand of various emerging applications in the realms of sixth-generation (6G) wireless networks has made the term internet of Things (IoT) very popular. Device-to-device (D2D) communication has emerged as one of the significant enablers for the 6G-based IoT network. Recently, the intelligent reflecting surface (IRS) has been considered as a hardware-efficient innovative scheme for future wireless networks due to its ability to mitigate propagation-induced impairments and to realize a smart radio environment. Such an IRS-assisted D2D underlay cellular network is investigated in this paper. Our aim is to maximize the network's spectrum efficiency (SE) by jointly optimizing the transmit power of both the cellular users (CUs) and the D2D pairs, the resource reuse indicators, and the IRS reflection coefficients. Instead of using traditional optimization solution schemes to solve this mixed integer nonlinear optimization problem, a reinforcement learning (RL) approach is used in this paper. The IRS-assisted D2D communication network is structured by the Markov Decision Process (MDP) in the RL framework. First, a Q-learning-based solution is studied. Then, to make a scalable solution with large dimension state and action spaces, a deep Q-learning-based solution scheme using experience replay is proposed. Lastly, an actor-critic framework based on the deep deterministic policy gradient (DDPG) scheme is proposed to learn the optimal policy of the constructed optimization problem considering continuous-valued state and action spaces. Simulation outcomes reveal that the proposed RL-based solution schemes can provide significant SE enhancements compared to the existing optimization schemes.

Keywords: device-to-device communication; overlay communication; intelligent reflecting surface (IRS); reinforcement learning (RL); spectrum efficiency (SE)

Citation: Sultana, A.; Fernando, X. Intelligent Reflecting Surface-Aided Device-to-Device Communication: A Deep Reinforcement Learning Approach. *Future Internet* **2022**, *14*, 256. https://doi.org/10.3390/fi14090256

Academic Editor: Paolo Bellavista

Received: 29 June 2022
Accepted: 25 August 2022
Published: 29 August 2022

1. Introduction

Recently, Internet of Things (IoT) systems have been rapidly deployed at an unprecedented pace with the growing demand in various emerging applications in the realms of sixth-generation (6G) networks [1]. The International Data Corporation (IDC) reveals that 41.6 billion connected IoT devices will generate 79.4 zettabytes of data by 2025 [2]. Hence, the design of 6G cellular IoT networks needs to satisfy the growing demands and expectations, e.g., huge capacity, low latency, seamless coverage, and global connectivity [3].

Direct communication between devices in close proximity (D2D communication) has emerged as a significant enabler of the 6G-based IoT networks [4]. In D2D communications, devices can communicate directly to exchange information, or they can relay data over the licensed cellular spectrum without forwarding it to a base station (BS) and without harmfully interfering with the licensed cellular users (this is called underlay communication). Thus, D2D communication is considered a promising offloading solution that enhances the overall network's performance by improving the spectral and energy efficiencies while reducing latency [5].

Due to the spectrum scarcity, there is a constant push for D2D users to share the underutilized radio spectrum with cellular users (CUs). However, this might cause degradation

in the quality-of-service (QoS) for the CUs due to unacceptable co-channel interference. Hence, interference management in underlay D2D communication has become a crucial challenge [5]. This can, however, be carefully tackled using efficient resource allocation [6], beamforming, and interference cancellation schemes [7]. On the other hand, due to transmit power limitations, the success rate of offloading is poor for those D2D users who are far away or hidden from each other. Moreover, the offloading communication links are prone to deep fade, blockages by obstacles, shadowing, etc. Therefore, an advanced framework for efficient D2D off-loading is necessary for satisfactory performance.

Recently, the usage of intelligent reflecting surfaces (IRS) has emerged as a hardware solution to improve radio frequency (RF) signal propagation issues and, in addition, to realize a smart radio environment [8]. Specifically, the IRS is a two-dimensional man-made surface of electromagnetic (EM) material, namely a metasurface, that is composed of a large array of specially designed passive scattering elements. Using software, each scattering element can be programmed to maneuver the properties of the reflected RF signal, namely the phase and spacial angles compared to the incident signal. This can create a desirable multi-path effect. For instance, the reflected RF signals can be added coherently to improve the received signal power or be combined destructively to mitigate interference. The IRS can be applied as a coating on a building wall or even be carried by aerial platforms, paving the way to several smart functionalities. Hence, the IRS can turn a radio environment into a smart space with enhanced data rate, extended coverage, low power consumption, and more secure transmissions [8]. In contrast to the traditional relay communications, IRS can characterize a fully controllable signal propagation scenario without a constant power supply and active circuitry requirements [9]. This enhances the system capacity, spectral efficiency, and energy efficiency substantially.

Therefore, an IRS-assisted D2D framework underlaying cellular system is very promising in eliminating the interference between the cellular and D2D users. Thus, this framework has the ability to boost the signal strengths of the cellular radio links, and it can help to meet the required data rate [10]. The research in this area has not yet been fully explored, which is the motivation behind this study on the IRS-assisted D2D framework.

Most of the resource allocation problems in the literature are constructed as nonlinear non-convex optimization problems that are mathematically intractable [10]. Therefore, traditional optimization approaches (e.g., exhaustive search) are infeasible for real-time optimization. The large number of IRS elements in an IRS-aided D2D framework require real-time optimization schemes with low complexity. Fortunately, reinforcement learning (RL), a subset of machine learning (ML), is very effective for controlling relevant policies and supporting intelligent decisions under uncertain and dynamic environments [11]. Q-learning, deep Q-learning, actor-critic framework, and other RL approaches have been enormously adopted for resource management in wireless networks [12–15]. Therefore, we propose an RL-based scheme to solve the optimization problem of the IRS-assisted D2D underlay cellular network.

1.1. Related Works

There have been several related works in this area with different parameters. The paper [16] investigates the task offloading and resource management for an IRS-assisted D2D cooperative computing system. Here, the convex optimization theory and semi-definite relaxation method are utilized to provide the optimal solution. On the other hand, an IRS-aided D2D offloading system is studied in [17], where an IRS is employed to assist in the offloading of computations from a group of intensive users to a group of idle users. A mixed-integer stochastic successive convex approximation scheme is proposed to tackle this problem. The physical layer security and data transmission for underlay D2D networks have been investigated in [18], where a combination of IRS and a full-duplex jamming receiver is considered for robustness and security enhancements. Another application of IRS in D2D communications is presented in [19], where a double deep Q-network (DDQN) structure is applied to optimize the transmit power and channel

assignment for the D2D pairs, the RIS position, and the phase shift to maximize the sum rate of the D2D network. An IRS-assisted single cell uplink D2D network has been studied in [20], where the problem of maximizing the total system rate is formulated by jointly optimizing the transmission power of all links and the discrete phase shifts of the surface; then, the optimization problem is decomposed into two sub-problems and solved iteratively. Another IRS-empowered underlay D2D network is presented in [10], where the spectrum efficiency (SE) and the energy efficiency of the network are maximized by jointly optimizing the resource reuse indicators, the transmit power, the IRS's passive beamforming, and the BS's receive beamforming. An efficient relative-channel-strength-based user-pairing scheme is proposed to determine the resource reuse indicators, and then other variables are jointly optimized by iterative algorithms. In [21], the IRS is introduced into a D2D communication system to improve the throughput of the D2D network, where the block coordinate descent algorithm and a semidefinite relaxation technique were utilized to optimize the beamforming vector, power allocation, and phase shift matrix. A two-timescale IRS-aided D2D ergodic rate maximization problem is studied in [22] subject to a given outage probability constrained by a cellular link QoS requirement. Here, the optimization problem is decoupled into two sub-problems and then solved iteratively with closed-form expressions.

1.2. Contributions

The resource allocation problem for an IRS-aided underlay D2D cellular network is studied in this paper. The main contributions are as follows:

- The objective of this paper is to maximize the network's SE, i.e., sum rate of both the CUs and the D2D pairs, by jointly optimizing the transmit power for both the CUs and the D2D pairs, the resource reuse indicators, and the IRS reflection coefficients. The optimization problem is subjected to the signal-to-interference-plus-noise ratio (SINR) constraints to meet the minimum data rate requirements to ensure the QoS for both the CUs and the D2D radio links. Since the constructed problem in this paper is a mixed integer non-linear optimization problem and poses challenges to being solved optimally, RL-based approaches are utilized.
- The IRS-assisted D2D underlay cellular network is structured by a Markov Decision Process (MDP) in the RL framework. At first, a Q-learning-based solution scheme is utilized. Then, to make a scalable solution in the large dimensional state and action spaces, a deep Q-learning-based solution scheme using experience replay is adopted. Lastly, an actor-critic framework based on the deep deterministic policy gradient (DDPG) scheme is proposed to learn the optimal policy of the constructed optimization problem with the consideration of the continuous-valued state and action spaces.
- Simulation outcomes under various representative parameters are provided to prove the effectiveness of the proposed RL-based solution schemes, and an analysis of the impact of different parameters on the system performance is also provided. It is certainly observed that the proposed RL-based solution schemes can provide significantly higher SE compared to the traditional underlay D2D network without IRS and without RL under different network parameters.

1.3. Organization

The paper is organized as follows: the system model for the IRS-assisted D2D underlay cellular network is described and the optimization problem is constructed in Section 2. The proposed RL-based solution schemes are presented in Section 3. The performance analysis and simulation outcomes of the proposed RL-based solution schemes are provided in Section 4. Finally, Section 5 concludes the paper.

2. System Model and Problem Formulation

2.1. System Model

In this paper, a single-cell cellular network-assisted D2D communication is considered, where an IRS is employed to elevate the link quality when a D2D user uses the cellular spectrum in an underlay manner. Please note that the IRS is not needed for direct D2D communication, which is of short range and the IRS is also not needed for communication between cellular users, as it happens through the BS. As shown in Figure 1, the basic system consists of one base station (BS), one IRS (more IRS units will be added as needed at locations that will provide best coverage to every user), a number of CUs, and a number of D2D pairs. The BS, CUs, and D2D pairs are equipped with a single antenna, while the IRS has N ($N > 1$) reflecting elements. A controller is attached to the IRS to control the reflection coefficients of each element. Note that we assume each reflecting element acts as an independent unit; therefore, one IRS can handle N user pairs.

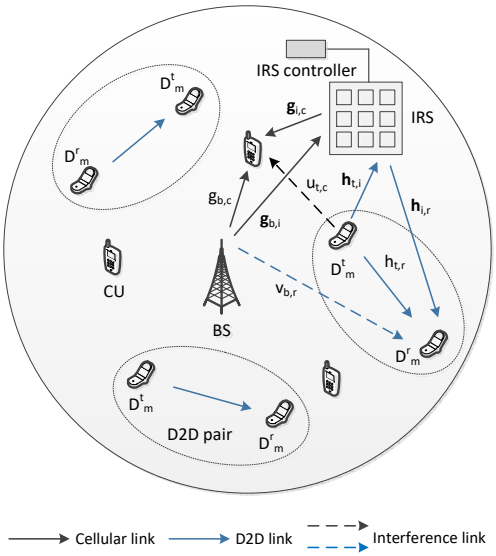

Figure 1. An IRS-assisted D2D underlay communication network.

In this system, K CUs coexist with M D2D pairs, and the D2D pairs share the downlink stream of the cellular network. Orthogonal frequency division multiple access (OFDMA) is used for the multiple access technique for both the CUs and D2D pairs, where resource blocks are employed for the spectrum allocation. In order to tackle the interference, it is assumed that the spectrum of a CU can be shared by at most one D2D pair, and a D2D pair can share at most one CU's resource.

Let $\mathcal{K}$ be the set for CUs, where $\mathcal{K} = \{1, 2, \ldots, K\}$, $\mathcal{M}$ is the set for D2D pairs, and where $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{N}$ are the set for IRS reflecting elements, where $\mathcal{N} = \{1, 2, \ldots, N\}$. We denote the k^{th} CU in the system by C_k, where $k \in \mathcal{K}$, and the m^{th} D2D pair by D_m, where $m \in \mathcal{M}$. The transmitter and the receiver of the D2D pair D_m are represented by D^t_m and D^r_m, respectively.

The channel model for all links in this system is considered quasi-static flat-fading. Let $\boldsymbol{g}_{b,i} \in \mathbb{C}^{N\times 1}$ be the channel gain of the cellular link from the BS to the IRS, $\boldsymbol{g}_{i,c} \in \mathbb{C}^{1\times N}$ be the channel gain from the IRS to CU C_k, and $g_{b,c} \in \mathbb{C}$ be the channel gain from the BS to CU C_k. On the other hand, the channel gain of the D2D link from D2D transmitter D^t_m to the IRS is denoted as $\boldsymbol{h}_{t,i} \in \mathbb{C}^{N\times 1}$, the channel gain from the IRS to D2D receiver D^r_m is denoted as $\boldsymbol{h}_{r,i} \in \mathbb{C}^{1\times N}$, and the channel gain from D2D transmitter D^t_m to D2D receiver D^r_m is denoted as $h_{t,r} \in \mathbb{C}$. Let $u_{t,c} \in \mathbb{C}$ be the interference link from D2D transmitter D^t_m to CU C_k and $v_{b,r} \in \mathbb{C}$ be the interference link from the BS to D2D receiver D^r_m.

Let $\boldsymbol{\Phi} \triangleq \text{diag}\,[\Phi_1, \Phi_2, \ldots, \Phi_N] \in \mathbb{C}^{N\times N}$ be the reflection coefficient matrix at the IRS, where $\Phi_n = \alpha_n e^{j\theta_n}$ represents an amplitude coefficient $\alpha_n \in (0, 1]$ and a phase shift

coefficient $\theta_n \in (0, 2\pi]$. Here, $\alpha_n = 1, \forall n \in \mathcal{N}$ is considered to achieve a maximum signal reflection.

The instantaneous SINR of the received signal at CU C_k from the BS can be written as

$$\gamma_k^c = \frac{P_k^c \, |\boldsymbol{g}_{i,c}^H \boldsymbol{\Phi} \boldsymbol{g}_{b,i} + g_{b,c}^H|^2}{\sum_{m=1}^{M} x_{k,m} P_m^d \, |\boldsymbol{g}_{i,c}^H \boldsymbol{\Phi} \boldsymbol{h}_{t,i} + u_{t,c}^H|^2 + \sigma^2}, \tag{1}$$

where P_k^c is the transmit power of CU k, and P_m^d is the transmit power of D2D pair m; $x_{k,m}$ is a binary decision variable of resource reuse indicator, where k represents cellular communication link and m D2D communication link, $x_{k,m} = 1$ when m D2D link reuses the spectrum of CU k, $x_{k,m} = 0$ otherwise, and σ^2 is the additive white Gaussian noise (AWGN) power.

The SINR of the received signal at the D2D receiver D_m^r from the D2D transmitter D_m^t can be written as

$$\gamma_m^d = \frac{P_m^d \, |\boldsymbol{h}_{i,r}^H \boldsymbol{\Phi} \boldsymbol{h}_{t,i} + h_{t,r}^H|^2}{\sum_{k=1}^{K} x_{k,m} P_k^c \, |\boldsymbol{h}_{i,r}^H \boldsymbol{\Phi} \boldsymbol{g}_{b,i} + v_{b,r}^H|^2 + \sigma^2}. \tag{2}$$

Therefore, the network's SE (i.e., sum rate of both CU and D2D pairs) in bps/Hz is given as

$$R(\boldsymbol{x}, \boldsymbol{P}, \boldsymbol{\Phi}) = \sum_{k=1}^{K} \log_2(1 + \gamma_k^c) + \sum_{m=1}^{M} \log_2(1 + \gamma_m^d), \tag{3}$$

where $\boldsymbol{x} = [x_{1,1}, \ldots, x_{1,M}, x_{2,1}, \ldots, x_{K,M}]^T$ is the resource reuse indicator vector, and $\boldsymbol{P} = [P_1^c, \ldots, P_K^c, P_1^d, \ldots, P_M^d]^T$ is the power allocation vector.

2.2. Problem Formulation

Our objective is to maximize the SE shown in (3) by jointly optimizing the transmit power, the resource reuse indicator, and the IRS reflection coefficients. Therefore, the optimization problem (OP) can be formulated as follows:

$$\begin{aligned}
&(\text{OP}) : \max_{\boldsymbol{x}, \boldsymbol{P}, \boldsymbol{\Phi}} R(\boldsymbol{x}, \boldsymbol{P}, \boldsymbol{\Phi}) \qquad (4)\\
&\text{subject to:}\\
&C1 : \ \gamma_k^c \geq \gamma_{\min}^c; \ \ \forall k \in \mathcal{K},\\
&C2 : \ \gamma_m^d \geq \gamma_{\min}^d; \ \ \forall m \in \mathcal{M},\\
&C3 : \ \sum_{k=1}^{K} x_{k,m} \leq 1,\\
&C4 : \ \sum_{m=1}^{M} x_{k,m} \leq 1,\\
&C5 : \ 0 \leq P_k^c \leq P_{\max}^c; \ \ \forall k \in \mathcal{K},\\
&C6 : \ 0 \leq P_m^d \leq P_{\max}^d; \ \ \forall m \in \mathcal{M},\\
&C7 : \ |\Phi_n| = 1, 0 < \theta_n \leq 2\pi; \ \ \forall n \in \mathcal{N}.
\end{aligned}$$

Here, constraints $C1$ and $C2$ denote the required minimum SINRs to ensure the QoS for the cellular communication links and the D2D communication links, respectively. Constraints $C3$ and $C4$ indicate that a D2D pair shares at most one CU's spectrum and the spectrum of a CU can be shared by at most one D2D pair, respectively. Constraints $C5$ and $C6$ represent the maximum transmission power constraints for the CUs and the D2D pairs, respectively, and constraint $C7$ denotes the practical reflecting coefficients of the IRS.

Note that the formulated OP in (4) is a mixed integer non-linear optimization problem and thus is NP-hard [23]. The variables $\boldsymbol{x}, \boldsymbol{P}, \boldsymbol{\Phi}$ in the objective function are all coupled,

which makes their joint optimization computationally intractable and thus difficult to solve. In addition, in practical network scenarios, all the parameters associated with the network will change dynamically. Therefore, utilization of traditional optimization solutions to solve the formulated OP generally leads to an infeasible solution. Hence, an RL-based model-free architecture is used for learning the optimization policy to obtain the feasible $\boldsymbol{x}, \boldsymbol{P}, \boldsymbol{\Phi}$.

3. Reinforcement Learning-Based Solution

Firstly, the formulated OP in (4) is reformulated using the MDP model to be solved by RL. Secondly, a Q-learning-based solution scheme is proposed. Then, to reach a scalable solution appropriate for the large dimension state and action spaces, a deep Q-learning-based scheme using experience replay is adopted. Lastly, an actor-critic framework based on the deep deterministic policy gradient (DDPG) scheme is proposed to learn the optimal policy of the constructed optimization problem with the consideration of the continuous-valued state and action spaces.

3.1. Reinforcement Learning (RL)

The interactions between the agent and the environment in RL can be well described by the MDP model. The MDP model is structured as a tuple $(\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \delta)$, where $\mathcal{S}$ be the state space set, $\mathcal{A}$ the action space set, r the immediate reward function, $\mathcal{P}$ the state transition probability, and δ the discount factor. As shown in Figure 2, in this model, each D2D pair is regarded as a learning agent in the IRS-aided D2D communication system environment, and there is no need for prior knowledge about the environment. The learning agent (i.e., the controller) continuously learns through interactions with the environment E in discrete timesteps by taking an action and then observing instant rewards and the state transitions in the environment. Thereby, it gradually derives its best action. The structure of the MDP model is given in detail as follows:

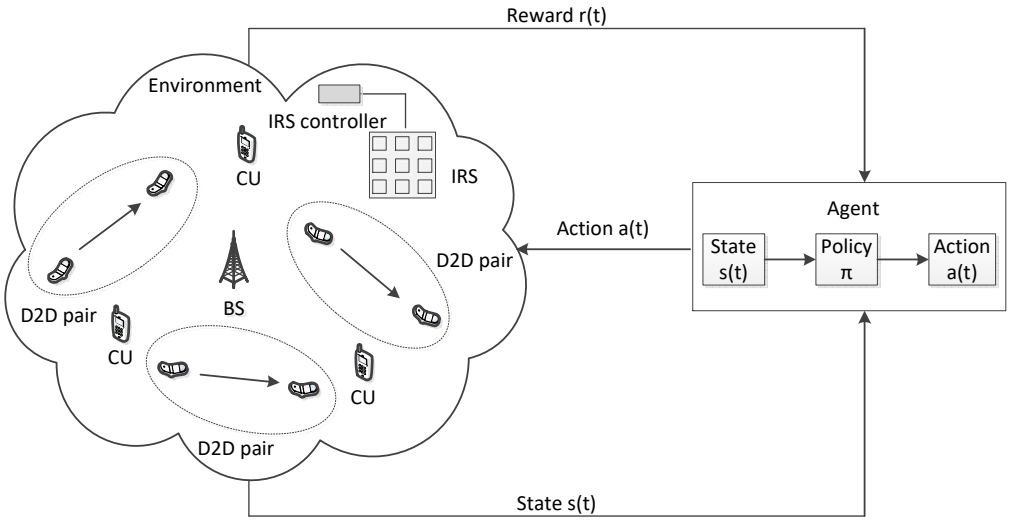

Figure 2. The RL framework for the IRS-assisted D2D communication network.

- *Agent*: A pair of every D2D transmitter and receiver.
- *State*: The observed information in the environment constitutes the system state $s(t) \in \mathcal{S}$, which characterizes the environment at current time t. It includes the channel information and all D2D agents' behaviors (e.g., SINR requirement). Therefore, the system state $s(t)$ can be defined by the following expression:

$$s(t) = \left\{ \left\{ G_k^{c,t} \right\}_{k \in \mathcal{K}}, \left\{ H_m^{d,t} \right\}_{m \in \mathcal{M}}, \left\{ \gamma_k^{c,t} \right\}_{k \in \mathcal{K}}, \left\{ \gamma_m^{d,t} \right\}_{m \in \mathcal{M}} \right\}, \tag{5}$$

where $G_k^{c,t}$ denotes the instantaneous channel information of the k^{th} cellular communication link that includes $\boldsymbol{g}_{b,i}$, $\boldsymbol{g}_{i,c}$, and $g_{b,c}$. $H_m^{d,t}$ represents the instant channel information of the m^{th} D2D communication link that includes $\boldsymbol{h}_{t,i}$, $\boldsymbol{h}_{r,i}$, and $h_{t,r}$.

- *Action*: The learning agent (D2D pair) does the appropriate action $a(t) \in \mathcal{A}$ at time t during the learning process following the current state, $s(t) \in \mathcal{S}$, based on the policy π. Hence, the action is defined by constraints C3 and C4 (a D2D pair shares at most one CU's spectrum and the spectrum of a CU can be shared by at most one D2D pair, respectively), constraints C5 and C6 (maximum transmission power constraints for the CUs and the D2D pairs, respectively), and constraint C7 (practical reflecting coefficients of the IRS). Based on the constraints in the formulated optimization problem, the action in this reinforcement learning framework depends on the transmit power, the resource reuse indicator, and the IRS reflection coefficients. Thus, the action $a(t) \in \mathcal{A}$ can be defined as follows:

$$a(t) = \left\{ \left\{ P_k^{c,t} \right\}_{k \in \mathcal{K}}, \left\{ P_m^{d,t} \right\}_{m \in \mathcal{M}}, \left\{ x_{k,m}^t \right\}_{k \in \mathcal{K}, m \in \mathcal{M}}, \left\{ \theta_n^t \right\}_{n \in \mathcal{N}} \right\}. \tag{6}$$

- *Transition probability:* $\mathcal{P}(s'|s(t), a(t))$ is the probability of transitioning from a current state $s(t) \in \mathcal{S}$ to a new state $s' \in \mathcal{S}$ after executing an action $a(t) \in \mathcal{A}$.
- *Policy:* The D2D agent's behavior is defined by a policy π, which maps states $s(t)$ to a probability distribution over the actions $a(t)$, e.g., $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$. Note that the policy function needs to satisfy $\sum_{(a(t) \in \mathcal{A})} \pi(s(t), a(t)) = 1$.
- *Reward:* The reward r is the immediate return to the D2D agent after taking the action $a(t) \in \mathcal{A}$ given the state $s(t) \in \mathcal{S}$. It is also a performance indicator that indicates how good the action $a(t)$ is in a given state s_m^t at time instant t. Hence, considering the interactions with the environment, each D2D agent takes its decision to maximize its reward. Now, the reward function for the D2D pair can be written as

$$r = R - \Upsilon_1 \sum_{k \in \mathcal{K}} (\gamma_k^c - \gamma_{\min}^c) - \Upsilon_2 \sum_{m \in \mathcal{M}} \left(\gamma_m^d - \gamma_{\min}^d \right), \tag{7}$$

where the first part is the network's SE at time t and the second and third parts denote the minimum data rate requirement that is not satisfied for the CUs and the D2D pairs, respectively. The parameters Υ_1 and Υ_2 are the weights of the second and third parts, and

$$r = \begin{cases} R, & \gamma_k^c \geq \gamma_{\min}^c \ \& \ \gamma_m^d \geq \gamma_{\min}^d, \\ \text{Negative}, & \text{otherwise.} \end{cases} \tag{8}$$

The reward function remains positive for the D2D agent if it satisfies both the conditions $\gamma_k^c \geq \gamma_{\min}^c$ and $\gamma_m^d \geq \gamma_{\min}^d$; otherwise, it will be negative.

The framework for the RL-based MDP model works as follows: the D2D agent observes a state $s(t) \in \mathcal{S}$, and considering the interaction with the environment, it takes an action $a(t) \in \mathcal{A}$, selecting the transmit power, the resource reuse indicators, and the passive beamforming based on the policy π. When an action $a(t) \in \mathcal{A}$ is executed, the D2D agent makes a transition from the current state $s(t)$ of the environment to a new state s', and the D2D agent receives an immediate reward r.

3.2. Q-Learning-Based Solution Scheme

The objective of the D2D agent is to find a policy π^* which maximizes the *cumulative* discounted reward denoted as

$$\sum_{\tau \geq 0} \delta^\tau r(t + \tau), \tag{9}$$

where $r(t)$ is the immediate reward for the D2D agent at time t, $\delta \in (0, 1]$ represents the discount factor, and t denotes the time slot.

The optimal policy π^*, which maximizes the expected sum of rewards, can be written as

$$\pi^* = \arg \max_{\pi} \mathbb{E}\left\{ \sum_{\tau \geq 0} \delta^{\tau} r(t+\tau) | \pi \right\}. \tag{10}$$

Now, the value function at state s following the policy can be expressed as

$$V^{\pi}(s) = \mathbb{E}\left\{ \sum_{\tau \geq 0} \delta^{\tau} r(t+\tau) | \pi, s = s(t) \right\}. \tag{11}$$

The Q-value function of considering action a in state s following the policy can be written as

$$Q^{\pi}(s,a) = \mathbb{E}\left\{ \sum_{\tau \geq 0} \delta^{\tau} r(t+\tau) | \pi, s = s(t), a = a(t) \right\}. \tag{12}$$

Now, the optimal Q-value function Q^* from a given (state, action) pair can be represented as

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left\{ \sum_{\tau \geq 0} \delta^{\tau} r(t+\tau) | \pi, s = s(t), a = a(t) \right\}. \tag{13}$$

The Q^* that satisfies the following Bellman equation [24] is utilized to find the optimal policy π^* by considering the best action in any state:

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}_{s' \sim E}\left[r + \delta \max_{a'} Q^*(s',a') | s, a \right] \\ &= \sum_{s',r} \mathcal{P}(s',r|s,a)\left[r + \delta \max_{a'} Q^*(s',a') \right]. \end{aligned} \tag{14}$$

Now, the Bellman equation (14) is updated with an iterative approach as follows:

$$Q_{i+1}(s,a) = \mathbb{E}\left[r + \delta \max_{a'} Q_i(s',a') | s, a \right], \tag{15}$$

where Q_i converges to Q^* as iteration $i \to \infty$.

The iterative update on the Bellman equation is then utilized to find the optimal $Q^*(s,a)$, which can be written as

$$Q^*(s,a) \leftarrow (1-\alpha) Q^*(s,a) + \alpha \left[r + \delta \max_{a'} Q^{\pi}(s',a') \right], \tag{16}$$

where $\alpha \in (0,1]$ denotes the learning rate.

In a Q-learning-based solution scheme, the ε-greedy scheme [24] is used to choose the action based on the current action-value estimation, which can be written as follows:

$$\pi = \begin{cases} \arg \max_{a(t) \in \mathcal{A}} Q(s(t), a(t)), & \text{with probability } 1-\varepsilon \\ \text{random}\{a_i\}_{a_i \in \mathcal{A}}, & \text{with probability } \varepsilon. \end{cases} \tag{17}$$

The Q-learning-based solution scheme is provided in Algorithm 1. At each iteration step of this algorithm, each D2D agent selects an action and updates its own policy independently.

Algorithm 1: Q-learning Based Solution Scheme

Initialization:
Initialize the Q-value function $Q(s,a)$ with random weights
Parameter Setting and Updating:
for *each episode* **do**
  Initial states $s^0 = \{s_1^0, \ldots, s_M^0\}$ observed by all D2D agents
  for *each time slot* **do**
    Actions $a(t)$ are selected using the ε-greedy scheme (17) and then executed by all D2D agents ;
    Perform observation of the rewards $r(t)$ and the new state s';
    Update $Q(s(t), a(t))$ using (16);
    Update $\pi(s(t), a(t))$;

3.3. Deep Q-Learning-Based Solution Scheme

In this solution scheme, a deep neural network is used as a function approximator to estimate the action-value function, where the Q-value is determined as follows:

$$Q(s,a;\omega) \approx Q^*(s,a), \tag{18}$$

where ω is the weighting (bias) parameter in the neural network. The optimal $Q^*(s,a)$ is obtained by updating ω using stochastic optimization methods, i.e.,

$$\omega(t+1) = \omega(t) - \lambda \Delta_\omega L(\omega), \tag{19}$$

where λ denotes the learning rate to update ω, and Δ_ω represents the gradient of the loss function $L(\omega)$ with respect to ω. Here, $L(\omega)$ is the difference between the predicted value and the actual target value of the neural network. The action-value function $Q(s,a)$ with a parameter vector ω is then optimized by minimizing the loss function for iteration i, and this can be expressed as

$$L_i(\omega_i) = \mathbb{E}_{s,a}[(y_i - Q(s,a);\omega_i))^2], \tag{20}$$

where y_i denotes the target value, which can be estimated as

$$y_i = \mathbb{E}_{s' \sim E}\left[r + \delta \max_{a'} Q(s', a'; \omega_{i-1}) | s, a \right]. \tag{21}$$

In particular, the gradient update with respect to the Q-function parameters ω is computed as

$$\begin{aligned} \Delta_{\omega_i} L_i(\omega_i) = \mathbb{E}_{s,a;s' \sim E}\Big[& r + \delta \max_{a'} Q(s', a'; \omega_{i-1}) \\ & - Q(s,a;\omega_i) \Delta_{\omega_i} Q(s,a;\omega_i) \Big]. \end{aligned} \tag{22}$$

During the training phase of the Q-network, the learning from different batches of consecutive samples is challenging due to the correlation among these samples. This leads to inefficient learning. On the other hand, since the current Q-network parameters define the upcoming training samples, this can introduce bad feedback loops. To tackle these issues, the experience buffer method is utilized where the Q-network is trained on random mini batches of transitions $(s(t), a(t), r(t), s')$ as experienced from the replay buffer, instead of consecutive samples. Let B_T be the total experiences that can be saved in the replay buffer. At each learning step, a random mini batch of B experiences is sampled uniformly from the replay buffer pool, and then the Q-network's weights are updated by the stochastic

gradient descent scheme. The deep Q-learning-based solution scheme using experience replay is summarized in Algorithm 2.

Algorithm 2: Deep Q-learning Based Solution Scheme using Experience Replay

Initialization:
Initialize the Q-network and the target Q-network with the parameterized function ω and the experience replay buffer $\mathcal{B}$.
Parameter Setting and Updating:
for *each episode* **do**
  Initial states $s^0 = \{s_1^0, \ldots, s_M^0\}$ observed by all D2D agents
  for *each time slot* **do**
    Actions $a(t) = \eta(s(t))$ with the parameterized function ω^η is selected and executed by all D2D agents ;
    Perform observation of the rewards $r(t)$ and the new state s';
    if *the transition number* $< B_T$ **then**
      Save the transition $\{s(t), a(t), r(t), s'\}$ in the experience replay buffer;
    else
      Restore the first stored transition with $\{s(t), a(t), r(t), s'\}$ in the experience replay buffer;
      A mini-batch of B transitions is selected from the experience replay buffer;
      Update the gradient of the loss function $L(\omega)$ with respect to the parameterized function ω using (22)

3.4. Actor-Critic-Based Solution Scheme

To learn the optimal policy for the OP formulated in (4), an actor-critic framework based on the DDPG scheme [25] is adopted, as shown in Figure 3, with consideration of the continuous-valued state and action spaces. In the DDPG scheme, the D2D agent passes through the learning process and targets finding the optimal policy π by achieving the maximum long-term reward $\mathbb{E}\{Q^\pi(s,a)\}$. The DDPG scheme combines deterministic policy gradients and deep Q-learning approaches to train the actor and the critic network. Similar to DQN, a replay memory and a separate target network are also utilized here. In order to form the actor-critic framework, two similarly structured deep neural networks (DNNs) with different parameter settings are considered, where an evaluation network is used for real-time parameter updating, and a target network is included for soft parameter updating. The utilization of experience replay buffer and batch normalization procedures not only helps to improve the performances of both the evaluation and the target network but also provides stable and faster convergence. Each transition $(s(t), a(t), r(t), s')$ of the D2D agent is kept in the experience replay buffer. When the experience replay buffer becomes full, the first recorded transitions are overwritten by the new-coming transitions that initiate the learning process. A random mini batch of B transitions, denoted as (s_i, a_i, r_i, s_i') $\forall\ i = 1, \ldots, B$, is chosen during the learning process by sampling uniformly from the experience replay buffer to update the actor and the critic networks.

In the critic structure, the evaluation and target networks are used to estimate the Q-value functions as follows:

$$Q^\pi(s,a) = \mathbb{E}_{r,s' \sim E}\left[r + \delta \mathbb{E}_{a' \sim \pi} Q^\pi(s', a')\right]. \tag{23}$$

The critic structure parameter is updated as $Q(a) \approx Q^\pi(s,a)$ during the learning period since the achievable Q-value function under the deterministic policy π ($\pi : \mathcal{S} \rightarrow \mathcal{A}$) is $Q^\pi(s,a)$, where $\pi(s) = \arg\max_a Q(s,a)$ [26]. Let ω^Q and $\omega^{Q'}$ be the parameterized functions; $Q(\cdot)$ denotes the network function of the evaluation network, and $Q'(\cdot)$ represents the network functions of the target network for the critic structure. Then, the mini batch of B transitions is chosen during the learning process, and the weights (ω^Q) are updated by minimizing the loss function stated as

$$L(\omega^Q) = \mathbb{E}\{(\xi_i)^2\}, \tag{24}$$

where ξ_i is the temporal-difference (TD) error, which denotes the deference between the estimated Q-value and the target Q-value, which can be written as follows:

$$\xi_i = Q(a_i) - (r_i + \delta Q'(a')). \tag{25}$$

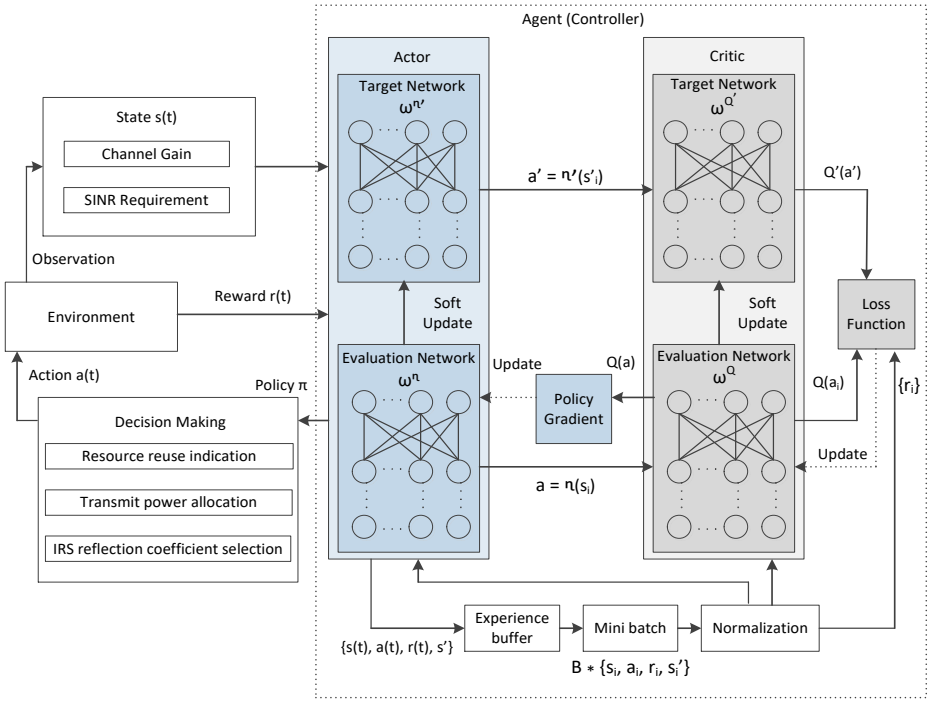

Figure 3. The actor-critic framework based on the DDPG scheme.

Let the loss function $L(\omega^Q)$ be continuously differentiable with respect to ω^Q. Then, the critic structure can update ω^Q with the gradient of $L(\omega^Q)$ as follows:

$$\Delta\omega^Q = \lambda^c \frac{1}{B} \sum_{\mathcal{B}} \xi_i \Delta_{\omega^Q} Q(a_i), \tag{26}$$

where λ^c denotes the learning rate of the critic structure, and $\Delta_{\omega^Q} Q(a_i)$ is the derivative of $L(\omega^Q)$ with respect to ω^Q.

Let ω^η and $\omega^{\eta'}$ be the parameterized functions; $\eta(\cdot)$ denotes the network function of the evaluation network, and $\eta'(\cdot)$ represents the network function of the target network in the actor structure. The actor structure utilizes the policy gradient scheme to update the parameterized function ω^η during the learning process to make $\eta(s_i)$ approximate the optimal policy π so that the maximum $\mathbb{E}\{Q^\pi(s,a)\}$ can be achieved. Thus, the policy under the parameterized actor function ω^η can be represented as $J(\omega^\eta) = \mathbb{E}\{Q(a)\}$, as in [27], where, $a = \eta(s_i)$. Since $J(\omega^\eta)$ is continuously differentiable with respect to ω^η, ω^η can be updated by the chain rule to the expected return from the start distribution J with the gradient of $J(\omega^\eta)$ as follows:

$$\Delta\omega^\eta = \lambda^a \frac{1}{B} \sum_{\mathcal{B}} \Delta_a Q(a)|_{a=\eta(s_i)} \Delta_{\omega^\eta} \eta(s_i), \tag{27}$$

where λ^a denotes the learning rate of the actor structure; $\Delta_a Q(a)|_{a=\eta(s_i)}$ is the derivative of $Q(a)$ with respect to a, where $a = \eta(s_i)$; and $\Delta_{\omega^\eta} \eta(s_i|\omega^\eta)$ is the derivative of $\eta(s_i)$

with respect to ω^{η}. Then, the following equations (28) and (29) are constructed by the derivative scheme:

$$\omega^{\eta'} = \varsigma^{a}\omega^{\eta} + (1-\varsigma^{a})\omega^{\eta'}, \tag{28}$$

$$\omega^{Q'} = \varsigma^{c}\omega^{Q} + (1-\varsigma^{c})\omega^{Q'}, \tag{29}$$

where $\varsigma^{a} \ll 1$ and $\varsigma^{c} \ll 1$. This allows the values of the target networks to be constrained to change slowly, which greatly improves the stability of the learning [25].

The DDPG-based actor-critic algorithm with the parameter updating processes is summarized in Algorithm 3. In this scheme, the evaluation network of the actor structure performs the action for each state in the given environment. A random mini batch of transitions are chosen in each learning process by sampling uniformly from the experience replay buffer, and they are input one by one to the D2D agent. With each input experience, the actor updates its parameterized function ω^{η} of the evaluation network based on the policy gradient to maximize $\mathbb{E}\{Q(a)\}$, and the critic updates its parameterized function ω^{Q} of the evaluation network based on the loss function to minimize the TD error. For both the actor and critic, the parameterized functions of the target networks are soft updated instead of directly copying the weights with that of their evaluation networks.

Algorithm 3: DDPG-based Actor-Critic Solution Scheme

Initialization:
Initialize the parameterized functions ω^{η}, $\omega^{\eta\prime}$, ω^{Q}, and $\omega^{Q\prime}$ of the evaluation and target networks of the actor-critic structure and the experience replay buffer $\mathcal{B}$.
Parameter Setting and Updating:
for *each episode* **do**
  Initial states $s^{0} = \{s_1^0, \ldots, s_M^0\}$ observed by all D2D agents
  for *each time slot* **do**
    Actions $a(t) = \eta(s(t))$ with the parameterized function ω^{η} is selected and executed by all D2D agents ;
    Perform observation of the rewards $r(t)$ and the new state s';
    if *the transition number* $< B_T$ **then**
      Save the transition $\{s(t), a(t), r(t), s'\}$ in the experience replay buffer;
    else
      Restore the first stored transition with $\{s(t), a(t), r(t), s'\}$ in the experience replay buffer;
      A mini-batch of B transitions is selected from the experience replay buffer;
      Update the parameterized functions of the critic structure:
      $\omega^{Q} \leftarrow \omega^{Q} + \Delta\omega^{Q}$;
      $\omega^{Q'} = \varsigma^{c}\omega^{Q} + (1-\varsigma^{c})\omega^{Q'}$;
      Update the parameterized functions of the actor structure:
      $\omega^{\eta} \leftarrow \omega^{\eta} + \Delta\omega^{\eta}$;
      $\omega^{\eta'} = \varsigma^{a}\omega^{\eta} + (1-\varsigma^{a})\omega^{\eta'}$;

4. Performance Analysis and Simulation Outcomes

In this section, the simulation outcomes under various scenarios are provided for the IRS-aided D2D cellular network to prove the effectiveness of the RL-based proposed solution schemes and to analyze the impact of different parameters on the system performance.

4.1. Simulation Setup

A single cell of a cellular wireless network is considered, where the BS is placed at the center, and both the CUs and D2D pairs are uniformly distributed within the cell. The IRS is located between the D2D pairs. The locations of the CUs, D2D pairs, and IRS are generated initially (in one realization) and then kept fixed throughout the simulation. Furthermore, each antenna of both the CUs and the D2D pairs is assumed to have an isotropic radiation pattern with 0 dB antenna gain, where each reflecting element of the IRS is assumed to

have a 3 dB gain for a fair comparison since each IRS reflects signals only in its front half-space [28].

Two categories of fading, i.e., large-scale and small-scale fading, are considered for the channel model. The large-scale fading based on the distance-dependent path loss model is considered. The small-scale fading components of $h_{t,r}$, $v_{b,r}$, $g_{b,c}$, and $u_{t,c}$ follow the Rayleigh fading model, while the small-scale fading components of $g_{b,i}$, $h_{t,i}$, $g_{i,c}$, and $h_{i,r}$ follow the Rician fading model [28], i.e.,

$$g_{b,i} = \sqrt{\frac{\mu_{b,i}}{\mu_{b,i}+1}}\bar{g}^{b,i} + \sqrt{\frac{1}{\mu_{b,i}+1}}\hat{g}_{b,i}, \tag{30}$$

where $\bar{g}_{b,i}$ denotes the line-of-site (LOS) component, $\hat{g}_{b,i}$ denotes the non-LOS (NLOS) component, and μ represents the Rician factor. The simulation parameters used for the IRS-aided D2D underlay cellular network are provided in Table 1.

Table 1. Simulation Parameters for IRS-aided D2D system.

Parameters	Values
Cellular cell radius	500 m
D2D link distance	10–50 m
Number of CUs	5–25
Number of D2D users	5–35
Number of reflecting elements	50
CUs' maximum transmit power	24 dBm
D2D transmitters' maximum transmit power	24 dBm
CUs' minimum SINR requirement	0.3 bps/Hz
D2D receivers' minimum SINR requirement	0.3 bps/Hz
Resource block bandwidth	180 kHz
Pathloss exponent	4
Pathloss constant	10^{-2}
Shadowing	8 dB
Multi-path fading	1
Noise spectral density	−144 dBm/Hz

The simulation procedures of the RL-based proposed solution schemes consist of two segments, i.e., learning and testing. First, the RL-based framework learns from the constructed models, and then, the learned framework is tested under different parameter settings to evaluate the system performance. Unless specified otherwise, the parameters for the learning phase and the testing phase remain the same. In the proposed actor-critic solution scheme, a three-layer fully connected neural network with two hidden layers is considered as the actor. The actor network uses a rectified linear unit (ReLU) function for the first and second layers and uses *tanh* for the final layer. For the critic, a four-layer fully connected neural network with three hidden layers is considered. The critic network uses the ReLU function for the first, second, and third layers and leaner activation for the final layer. To simulate the model, the Python 4.5 platform is utilized. All the network parameters are trained using the Adam optimizer [29]. The simulation parameters related to the RL model are provided in Table 2.

Table 2. Simulation Parameters for RL model.

Parameters	Values
Learning rates of the actor/critic	0.0001/0.001
Discount factor	0.99
Greedy rate	0.1
Soft target update parameter	0.001
Replay buffer size	10,000
Mini-batch size	62

To analyze the system performance, the RL-based proposed solution schemes are compared with the following two benchmark schemes:

- Underlaying D2D without RL (scheme-1): An IRS-empowered underlay D2D communication network is considered scheme-1 [28]. Here, a user-pairing scheme determines the resource reuse indicator, and then, the transmit power and the passive beamforming are jointly optimized by iterative algorithms.
- Underlaying D2D without IRS without RL (scheme-2): A traditional cellular system underlaying the D2D network without IRS is considered scheme-2, where a two-stage approach is proposed to solve the optimization problem instead of RL [6].

4.2. Evaluation of the Proposed Solution Schemes

Figure 4 compares the convergence performance of the RL-based proposed schemes in terms of the total reward achieved per episode during the learning stage. It is observed from Figure 4 that these rewards fluctuate rapidly and are apparently small in the first part of the episodes during the learning period and then become relatively stable and large. The total reward performance of the proposed Q-learning-based scheme (Scheme-Q) is the worst due to the large state-action space. The proposed *deep* Q-learning-based solution scheme (Scheme-DQ) performs better than scheme-Q. This is because it approximates the complex mapping between the state-action spaces using a deep neural network with experience replay. The proposed actor-critic scheme (Scheme-AC) performs the best. This is because all parameters of this scheme are initialized, and the experiences are stored in the replay buffer. These parameters are then updated with the joint process of policy value learning during the learning stage. Here, the proposed Scheme-AC optimizes the policy with good convergence properties. Therefore, the performance of the proposed Scheme-AC is better than that of Scheme-Q and Scheme-DQ.

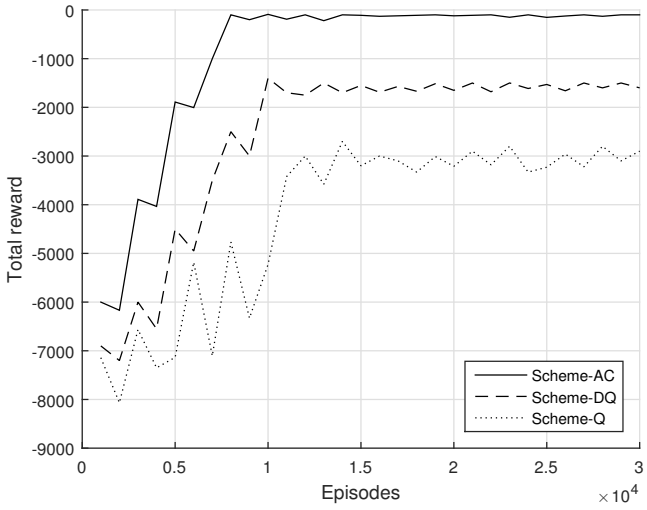

Figure 4. Convergence of the RL-based proposed solution schemes.

Figure 5 shows the SE performance with respect to the number of CUs for different solution schemes. The result implies that the SE performance of the network increases with the number of CUs with a fixed number of D2D pairs. This is because when the number of CUs increases, the D2D pairs have a better opportunity to access the best spectrum resources with less interference. It is also observed that the SE performance of the RL-based proposed schemes outperforms that of the other schemes. The SE performance of Scheme-Q and Scheme-AC remains at the top because they use the experience replay buffers to learn faster and benefit from the potential deployment of the IRS elements.

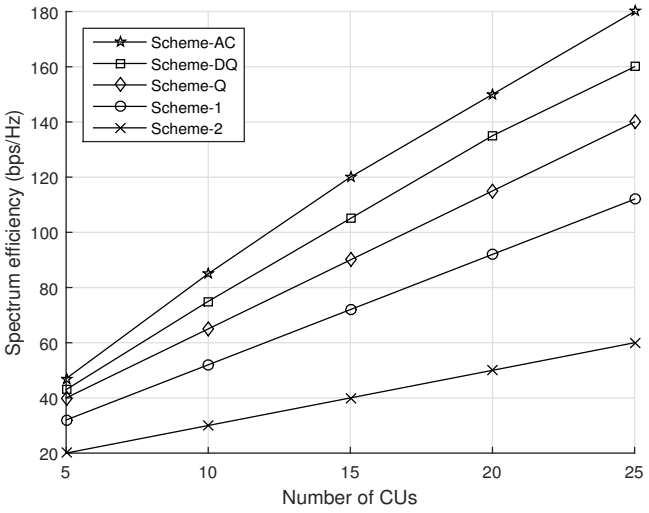

Figure 5. SE versus the number of CUs.

Figure 6 presents the SE performance with respect to the number of D2D pairs for different solution schemes. It is observed that as the number of D2D pairs increases, the SE performance for all the schemes monotonically increases but slowly saturates. This is because of higher interference with more D2D pairs. However, the SE performance of the RL-based proposed solution schemes outperforms that of the other two schemes.

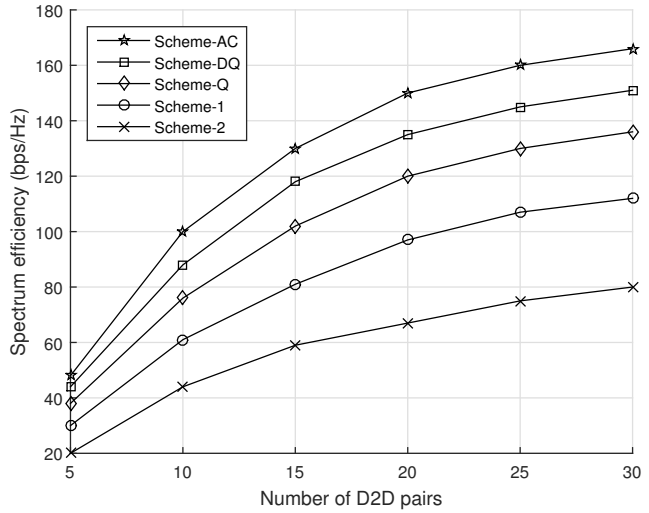

Figure 6. SE versus number of D2D pairs.

Figure 7 shows the SE with the minimum SINR requirement of the D2D pairs under different solution schemes. It is observed that when the SINR requirement of the D2D pairs is less, the SE performance of all the solution schemes is high. However, at a higher SINR requirement for D2D pairs, the SE performance starts decreasing for all the schemes. This is because a higher SINR requires lower interference, which naturally decreases the

SE. It can also be seen that the RL-based schemes yield better SE performance than the other schemes.

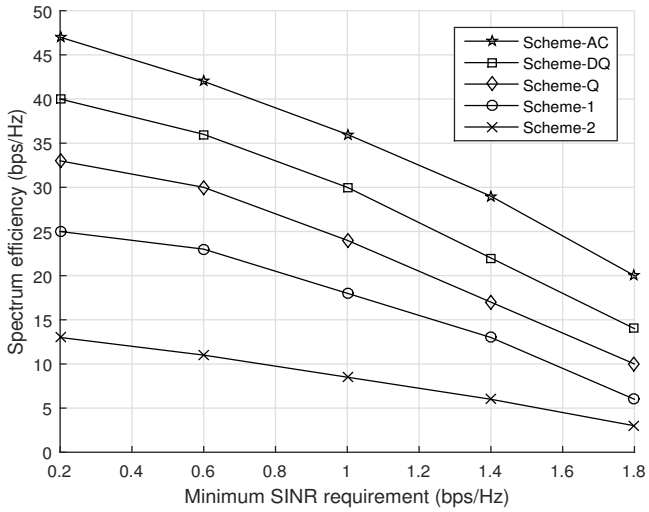

Figure 7. SE versus minimum SINR requirement for D2D pairs.

Figure 8 shows that the SE performance increases for all four schemes with the number of IRS elements. The SE performance of scheme-2 (without IRS) remains unchanged for obvious reasons. Again, the RL-based proposed schemes outperform the other schemes. The reason behind this increasing trend is that more reflecting elements can further enhance the channel strength by passive beamforming and suppress the undesired channel interference. However, the increase in the SE performance slows down as the number of IRS elements becomes too large, which is caused by the increase in interference signal paths between the CUs and D2D pairs.

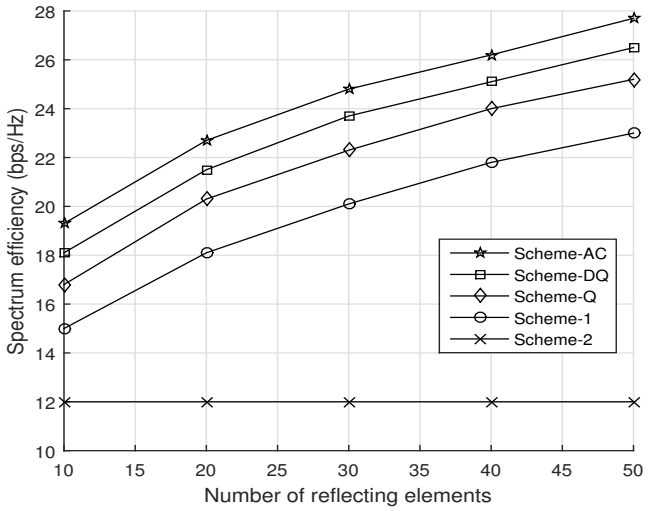

Figure 8. SE versus number of reflecting elements of IRS.

In Figure 9, it can be observed that the SE performances of the four schemes first increase with the increase of the Rician factor and then remain almost constant. However, the SE performance of scheme-2 (without IRS) remains unchanged. Again, the proposed RL-based schemes outperform the other two schemes. This is because as the Rician factor increases at first, the slowly varying LoS components between the users and the IRS are enhanced to boost the SE performance of the proposed RL-based solution schemes. However, when the Rician factor is relatively large, the LoS components are dominant and therefore have less effect on the channel strength.

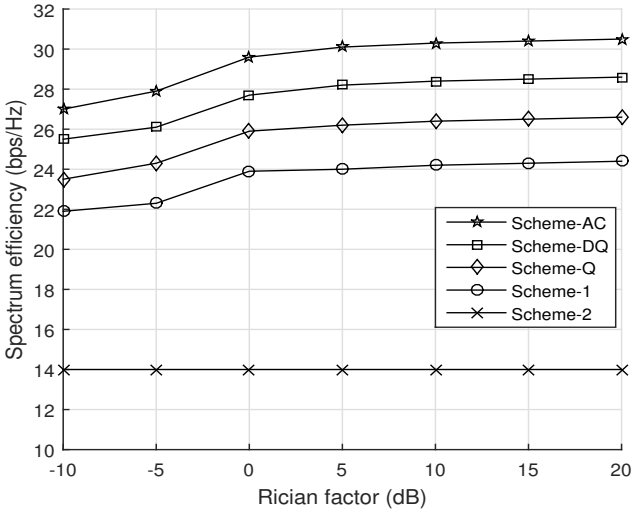

Figure 9. SE versus Rician factor (dB).

5. Conclusions

In this work, an IRS-assisted D2D underlay cellular network is investigated and structured by the MDP model in the RL framework. The network's spectral efficiency is maximized by jointly optimizing the transmit power, the resource reuse indicators, and the IRS reflection coefficients. An RL-based solution architecture is proposed to solve the constructed optimization problem. First, the Q-learning-based solution scheme is utilized. Then, to make a scalable solution, a deep Q-learning solution scheme with experience replay is proposed. Lastly, an actor-critic framework based on the deep deterministic policy gradient (DDPG) scheme is proposed to learn the optimal policy of the constructed optimization problem with the consideration of the continuous-valued state and action spaces. Simulation outcomes clearly demonstrate that the proposed RL-based solution schemes achieve a significant SE performance gain compared to the traditional underlay D2D network without RL under different parameter settings. This research can be continued to explore complex cases, such as multi-antenna base stations, multiple IRSs, and the multi-agent framework of the MDP model.

Author Contributions: Conceptualization, A.S.; methodology, A.S.; software, A.S.; validation, A.S.; formal analysis, A.S.; investigation, A.S.; writing—original draft preparation, A.S.; writing—review and editing, A.S. and X.F.; visualization, A.S.; supervision, X.F.; funding acquisition, X.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guo, F.; Yu, F.R.; Zhang, H.; Li, X.; Ji, H.; Leung, V.C.M. Enabling massive IoT Toward 6G: A comprehensive survey. *IEEE Internet Things* **2021**, *8*, 11891–11915. [CrossRef]
2. Worldwide Global Datasphere IoT Device and Data Forecast, 2021–2025. IDC: US48087621. Available online: https://www.idc.com/getdoc.jsp?containerId=US48087621 (accessed on 24 August 2022).
3. Qi, Q.; Chen, X.; Zhong, C.; Zhang, Z. Integration of energy, computation and communication in 6G cellular internet of things. *IEEE Commun. Lett.* **2020**, *24*, 1333–1337. [CrossRef]
4. Zhang, S.; Liu, J.; Guo, H.; Qi, M.; Kato, N. Envisioning Device-to-Device communications in 6G. *IEEE Netw.* **2020**, *34*, 86–98. [CrossRef]
5. Jameel, F.; Hamid, Z.; Jabeen, F.; Zeadally, S.; Javed, M.A. A survey of device-to-device communications: Research issues and challenges. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2133–2168. [CrossRef]
6. Sultana, A.; Zhao, L.; Fernando, X. Efficient resource allocation in device-to-device communication using cognitive radio technology. *IEEE Trans. Veh. Technol.* **2017**, *66*, 10024 –10034. [CrossRef]

7. Xu, W.; Liang, L.; Zhang, H.; Jin, S.; Li, J.C.; Lei, M. Performance enhanced transmission in device-to-device communications: Beamforming or interference cancellation? In Proceedings of the IEEE GLOBECOM, Anaheim, CA, USA, 3–7 December 2012; pp. 4296–4301.
8. Wu, Q.; Zhang, R. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network. *IEEE Commun. Mag.* **2020**, *58*, 106–112. [CrossRef]
9. Gong, S.; Lu, X.; Hoang, D.T.; Niyato, D.; Shu, L.; Kim, D.I.; Liang, Y.C. Toward Smart Wireless Communications via Intelligent Reflecting Surfaces: A Contemporary Survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2283–2314. [CrossRef]
10. Yang, G.; Liao, Y.; Liang, Y.C.; Tirkkonen, O.; Wang, G.; Zhu, X. Reconfigurable intelligent surface empowered Device-to-Device communication underlaying cellular networks. *IEEE Trans. Commun.* **2021**, *69*, 7790–7805. [CrossRef]
11. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
12. Luong, N.C.; Hoang, D.T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.C.; Kim, D.I. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3133–3174. [CrossRef]
13. Zappone, A.; Renzo, M.D.; Debbah, M. Wireless networks design in the era of deep learning: model-based, AI-based, or both? *IEEE Trans. Commun.* **2019**, *67*, 7331–7376. [CrossRef]
14. Kumar, A.S.; Zhao, L.; Fernando, X. Multi-Agent deep reinforcement learning-empowered channel allocation in vehicular networks. *IEEE Trans. Veh. Technol.* **2022**, *71*, 1726–1736. [CrossRef]
15. Shah, S.A.A.; Illanko, K.; Fernando, X. Deep learning based traffic flow prediction for autonomous vehicular mobile networks. In Proceedings of the IEEE VTC2021-Fall, Norman, OK, USA, 27–30 September 2021; pp. 1–5.
16. Mao, S.; Chu, X.; Wu, Q.; Liu, L.; Feng, J. Intelligent reflecting surface enhanced D2D cooperative computing. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 1419–1423. [CrossRef]
17. Liu, Y.; Hu, Q.; Cai, Y.; Juntti, M. Latency minimization in intelligent reflecting surface assisted D2D offloading systems. *IEEE Wirel. Commun. Lett.* **2021**, *9*, 3046–3050. [CrossRef]
18. Khalid, W.; Yu, H.; Do, D.T.; Kaleem, Z.; Noh, S. RIS-aided physical layer security with full-duplex jamming in underlay D2D networks. *J. Commun. Inf. Netw.* **2021**, *9*, 99667–99679. [CrossRef]
19. Ji, Z.; Qin, Z.; Parini, C.G. Reconfigurable intelligent surface assisted Device-to-Device communications. *arXiv* **2021**, arXiv:2107.02155v1.
20. Chen, Y.; Ai, B.; Zhang, H.; Niu, Y.; Song, L.; Han, Z.; Poor, H.V. Reconfigurable intelligent surface assisted Device-to-Device communications. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 2792–2804. [CrossRef]
21. Zhang, C.; Chen, W.; He, C.; Li, X. Throughput maximization for intelligent reflecting surface-aided Device-to-Device communications system. *J. Commun. Inf. Netw.* **2020**, *5*, 403–410. [CrossRef]
22. Cai, C.; Yang, H.; Yuan, X.; Liang, Y.C. Two-timescale optimization for intelligent reflecting surface aided D2D underlay communication. In Proceedings of the IEEE GLOBCOM, Taipei, Taiwan, 7–11 December 2020; pp. 1–6.
23. Plaisted, D.A. Some polynomial and integer divisibility problems are NP-HARD. In Proceedings of the Symposium on Foundations of Computer Science sfcs, Houston, TX, USA, 25–27 October 1976; pp. 264–267.
24. Wiering, M.; Otterlo, M. *Reinforcement Learning: Stateof-the-Art*; Springer Publishing Company, Incorporated: Berlin, Germany, 2014.
25. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2019**, arXiv:1509.02971v6.
26. Grondman, I.; Busoniu, L.; Lopes, G.A.; Babuska, R. A survey of actorcritic reinforcement learning: Standard and natural policy gradients. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **2012**, *42*, 1291–1307. [CrossRef]
27. Bhatnagar, S.; Sutton, R.S.; Ghavamzadeh, M.; Lee, M. Natural actorcritic algorithms. *Automatica* **1976**, *45*, 2471–2482. [CrossRef]
28. Yang, G.; Liao, Y.; Liang, Y.C.; Tirkkonen, O. Reconfigurable intelligent surface empowered underlying Device-to-Device communication. In Proceedings of the IEEE WCNC, Nanjing, China, 29 March–1 April 2021.
29. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.