# Reconfigurable Intelligent Surface Aided Cellular Networks with Device-to-Device Users

Zelin Ji, Zhijin Qin, *Senior Member, IEEE*, and Clive G. Parini, *Member, IEEE*

***Abstract*—Reconfigurable intelligent surface (RIS) technology is promising to enhance wireless communications services by providing smart radio environment. In this paper, we investigate the RIS aided cellular networks with device-to-device (D2D) users, and maximize the sum of the transmission rate of the D2D communications and the cellular networks from a new perspective. In addition to solving the typical resource allocation problems for D2D communications, this paper further optimize the wireless environment by adjusting the position and phase shift of the RIS. To solve this non-convex problem, we propose a novel decentralized double deep Q-network ($D^3QN$) framework for the resource allocation at users and a centralized DDQN for RIS optimization at the base station (BS), which are verified to achieve the near-optimal performance with lower complexity and enhanced robustness. Simulation results illustrate that the proposed framework can achieve higher transmission rates compared to benchmarks, meanwhile meeting the quality of service (QoS) requirements at the BS and D2D users.**

***Index Terms*—Deep reinforcement learning, device-to-device communication, reconfigurable intelligent surface.**

## I. Introduction

As one of the key technologies of the fifth-generation (5G) and the beyond communication systems, device-to-device (D2D) communications permit devices to communicate with proximity devices over the licensed spectrum allocated for cellular networks. By doing so, it will enhance the communication system performance by reducing the latency as well as improve spectrum efficiency (SE), and energy efficiency (EE) [2]. D2D communications have been adopted in various applications, standards, and regulations, including the 3rd Generation Partnership Project (3GPP) proximity services [3], Internet of Things (IoT), vehicle-to-everything (V2X) communications, and wearable communications [4]. According to Cisco, the share of D2D links will increase 20 percent from 2018 to 2023 [5].

There is a rich body of literature on resource allocation for D2D communications [6]–[8]. Recently, a novel approach referred to as the smart radio environment presents a new perspective to enhance the D2D communications. Particularly, the wireless environment can be controllable and programmable. In this way, we can optimize the communication environment and resource allocation for D2D devices simultaneously, thus permitting us to control or even eliminate the interference.

One key technique to realize the smart radio environment is reconfigurable intelligent surface (RIS) [9], [10], which has attracted extensive attention in wireless communications. Equipped with an array of low-cost passive reflecting elements, the phase shift and reflection amplitude of each RIS element can be adjusted by a controller, enabling it to modify wireless communication environment proactively. Compared with the conventional relays, the advantages of RIS include lowered energy consumption, real-time phase shift adjustment and enhanced system capacity [11]. Although the control signal can be analog using varactors to achieve continuous phase shift [12], the long response time and low phase accuracy of varactors make it impractical for wireless communications. Theoretical analyses for multi-bit controlled elements have been provided to strike a tradeoff between the system performance and the complexity [13], [14]. The performance improvement has been further verified by an RIS-based wireless communication prototype [15], which shows the great potential of RIS aided wireless networks.

On the other hand, the large number of RIS elements requires optimization methods with lower complexity. Although typical optimization tools may obtain the optimal solution, their high computational cost makes them unrealistic for the real-time optimization. Fortunately, machine learning (ML) methods, especially deep learning (DL) approaches, have become promising tools to address nonlinear non-convex problems and high-computation issues, which are mathematically intractable. Particularly, deep Q-network (DQN) leverages neural network (NN) to deal with complex input state, and has shown its power in solving sophisticated decision-making problems under uncertain and dynamic environments, e.g., human-level game playing [16], [17] and AlphaGo [18]. Inspired by the remarkable performance of DQN in various areas, there have been some works exploring its application in wireless communications [19]–[23]. DQN provides a principled and robust method to tackle the dynamic environment by making decisions for discrete optimization problems, which bring it the ability to optimize the resource allocation for D2D communications in varying channel state environment (CSI) [21]. Moreover, as a new technique of DQN, the double DQN (DDQN) provides a more reasonable way to evaluate and execute the action, which avoids the overestimation challenge of legacy DQN algorithms and is more robust to time-varying environment.

### A. Related work

*1) Resource allocation in D2D:* The existing works on D2D communications mainly focus on transmit power and channel assignment optimization [6]–[8]. Mili *et. al* [7] maximize EE by optimizing the transmit power while satisfying the

Part of this work was presented at the IEEE Global Communications Conference 2020 [1].

Zelin Ji, Zhijin Qin, and Clive G. Parini are with Queen Mary University of London, London, UK, E1 4NS, email: {z.ji, z.qin, c.g.parini}@qmul.ac.uk. (Corresponding author: Zhijin Qin.)

QoS requirement for D2D and cellular users. To overcome the challenges brought by dynamic D2D channels, Liang *et. al* [8] have proposed an algorithm for robust spectrum allocation and power optimization. While there exists plenty of literature applying optimization tools to solve resource allocation problems for D2D communications, most of them are centralized and requires intensive computation at the BS to execute the optimization process [24].

One solution to realize the distributed resource allocation is game theory. To overcome the convergence challenge caused by fast varying channels, Dominic *et al.* [25] adopted the stochastic learning algorithm among users. However, the game theory based algorithms neglect the collaboration behaviour among users, i.e., each user only focus on their own benefits, which may be adverse to the overall system performance.

As discussed earlier, DL based approaches enable wireless communication users treat the dynamic environment and make their robust decisions with lower computational complexity. DL has been applied to resource allocation [24] and physical layer processing [26]. Moreover, relying on the local users information and observations, multi-agent reinforcement learning (MARL) based decentralized optimization approaches have been widely applied in wireless communications [20]–[23]. Leveraging MARL, D2D pairs can make their own decisions on transmit power and spectrum sharing policy, which also offloads the computational complexity from the BS to users. Liu *et. al* [20] deploy D2D users as MARL agents, which learn to access the channel of cellular users by collectively interacting with the communication environment and receiving the rewards. Each D2D pair chooses its transmit power level and sub-channel to minimizing long-term system cost. However, the unknown policy and information of other users cause a non-stationary environment. To overcome it, MARL algorithms with improved state observation have been proposed in [21], [22], [27]. Such decentralized optimization approach has been verified to achieve the near-optimal performance [23]. Although the above valuable works improve the performance of D2D communications significantly, they mainly focus on the transmit power allocation and channel assignment allocation under static communication environments. With the aid of RIS, we are able to actively control the communication environment and optimize the resource allocation from a brand new perspective.

*2) RIS enhanced wireless communications:* Recently, RIS has been explored in a wide range of scenarios, e.g., RIS-enhanced cellular networks beyond 5G, RIS-assisted indoor communications, and IoT applications [28]. Most work only consider the phase shift design by assuming that the RIS is deployed at a fixed location. However, the location of the RIS will affect the performance significantly [29], [30]. By considering the costs and available space to install RIS, its deployment location should be optimized. Particularly, RIS has been successfully applied in D2D networks in [31]–[33]. Many approaches have been developed to optimize RIS for achieving higher throughput or EE. To solve the non-convex maximizing problems, Cao *et. al* [31] and Fu *et. al* [32] tend to find sub-optimal solutions by using the block coordinate descent and Riemannian pursuit method, respectively. To achieve a performance-complexity tradeoff, Pradhan *et. al* [33] adopt the projected sub-gradient method for the phase shift. However, to enhance the overall system performance, the optimization of RIS becomes a critical challenge due to the huge number of reflecting elements to optimize [34]. The time-varying D2D channel also brings high transmission overhead to optimization algorithms.

A well-trained ML model is an effective approach to lower computational cost. Although the ML model requires more computations at the training stage, it could be trained offline and is robust to fast channel variations in dynamic environment. Particularly, Gao *et. al* [29] investigate the application of reinforcement learning (RL) for aerial RIS trajectory optimization. Moreover, as a novel branch of ML, DL enable the users cope the complicated environment and has been applied for channel estimation and phase shift optimization in RIS-aided communications [35]. Motivated by the applications of DL in solving sophisticated optimization problems, Taha *et. al* [36] have applied the DL method for estimating the channels and configuring of RIS. DQN has shown its potential for optimization the phase shift and location of RIS. Liu *et. al* [37] apply a DQN based algorithm to optimize the phase shift of RIS for RIS-aided unmanned aerial vehicle communications.

### *B. Motivation and Contribution*

We consider an RIS enhanced D2D communications system to actively optimize the performance of communication. The challenges occur in several aspects. The fast channel variations of D2D communications make the conventional resource allocation approaches based on perfect CSI not applicable anymore [8]. Most of the current works for the RIS enhanced D2D communication separate the RIS optimization and resource allocation into sub-problems, then leveraging alternating optimization to solve the problem [31]–[33], [38]. However, the fast channel variations may affect the performance of the alternating optimization based approaches since the environment is varying and the algorithms are hard to converge. Additionally, the centralized alternating optimization algorithms pose high computational pressure on the center, which is adverse to the development of the massive capacity and connectivity trend for 5G and the beyond. The same situation applies to the centralized machine learning algorithms, where users need to upload all of the local CSI and other related information to the centralized computational center in real-time, bringing huge transmission overhead and computational pressure to the center.

To overcome these challenges, we jointly optimize the transmit power and the channel assignment for D2D pairs in a distributedly way. Additionally, a centralized DDQN model is adopted to optimized the RIS position, and the phase shift of RIS at the BS. The major contributions of this paper are summarized as follows:

1) To improve the sum rate of the cellular networks with D2D communications, we formulate the problem as joint optimization for the resource allocation of D2D pairs, the RIS position, and phase shift of RIS.
2) To enhance the robustness and effectiveness of the proposed algorithm, a novel DDQN algorithm is applied

where action choosing and target Q-value generation are decoupled, thus overcoming the overestimation problem in the conventional DQN.

3) To separate the RIS optimization task and the resource allocation task, a decentralized framework is designed. The RIS is optimized by a centralized DDQN at the BS. Meanwhile, D2D pairs are allowed to make their own policies locally based on a decentralized DDQN (D$^3$QN) approach, thereby offloading the computational cost at the BS and lowering the uplink transmission overhead.

The rest of this paper is organized as follows. The system model of RIS enhanced D2D communication system is presented in Section II. The proposed decentralized resource allocation is introduced in Section III. The centralized RIS optimization is presented in Section IV. Simulation results are presented in Section V. Finally, conclusions are drawn in Section VI.

## II. System Model

In this section, the system model for RIS enhanced cellular network with underlay D2D communications is described and an uplink rate maximization problem is formulated.

### A. System settings

As shown in Fig. 1, we consider the uplink transmission in a cellular network, which includes $K$ cellular users communicate with the BS in the cellular mode, and $I$ D2D pairs communicating with each other by reusing the resource blocks (RB) with cellular users. Assuming that the $i$-th D2D transmitter, $D_i^t$, communicating with the corresponding receiver, $D_i^r$, by reusing the RB assigned to the $k$-th cellular user, $U_k$, then $D_i^t$ becomes the source of interference for $U_k$. To enhance the transmission performance of the network, an RIS equipped with a uniform linear array (ULA) composed of $N$ passive elements is deployed.

Assuming that the horizontal coordinates of $D_i^t$, $D_i^r$, $U_k$, RIS and the BS are denoted as $\boldsymbol{P}_i^{D_t} = (X_i^{D_t}, Y_i^{D_t}) \in \mathbb{R}^2$, $\boldsymbol{P}_i^{D_r} = (X_i^{D_r}, Y_i^{D_r}) \in \mathbb{R}^2$, $\boldsymbol{P}_k^{U} = (X_k^{U}, Y_k^{U}) \in \mathbb{R}^2$, $\boldsymbol{P}^{RIS} = (X^{RIS}, Y^{RIS}) \in \mathbb{R}^2$ and $\boldsymbol{P}^{BS} = (X^{BS}, Y^{BS}) \in \mathbb{R}^2$, respectively. The 3D distance between $D_i^t$ and $D_i^r$, can be calculated by

$$d_i^D = \sqrt{(Z_i^{D_t} - Z_i^{D_r})^2 + ||\boldsymbol{P}_i^{D_t} - \boldsymbol{P}_i^{D_r}||^2} \tag{1}$$

where $Z_i^{D_t}$ and $Z_i^{D_r}$ represent the antenna height of the $i$-th D2D transmitter and the receiver. The distance between cellular users, D2D users, the RIS and the BS can be denoted in the similar way. The small-scale fading for the direct links, i.e., the links without the aid of the RIS, are modeled by the Rayleigh fading. The channel coefficient $g_i^D[k]$ between the D2D pair ($D_i^t$ and $D_i^r$) over the $k$ sub-channel, which is preoccupied by the $k$-th cellular user $U_k$ can be denoted as

$$g_i^D[k] = L(d_i^D) m_i^D[k], \tag{2}$$

where $m_i^D[k]$ represents Rayleigh fading between $i$-th D2D pair on $k$-th sub-channel, which is assumed to be complex Gaussian distributed with zero mean and unit variance, i.e., $m_i^D[k] \sim \mathcal{CN}(0,1), \forall i \in I, \forall k \in K$. Meanwhile, the large-scale pathloss coefficient is modeled according to the pathloss model in [39]. Similarly as (2), the channel coefficients for the direct link between $U_k$ and $D_i^r$, the link between $U_k$ and the BS, the link between $D_i^t$ and the BS over $k$-th sub-channel are denoted by $g_{k,i}[k]$, $g_k^U[k]$, and $g_i^{BS}[k]$.

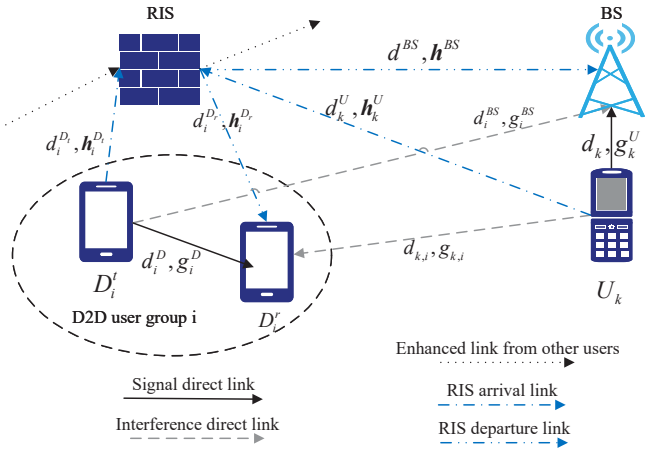



Fig. 1. System model for RIS enhanced cellular network with underlay D2D communications. Note that the RIS influences the signal link and the interference link at the same time.

The channel coefficient over the $k$-th sub-channel for the overall $i$-th D2D link, $h_i^D[k]$, is given by

$$h_i^D[k] = \underbrace{(\boldsymbol{h}_i^{D_r}[k])^{\mathrm{H}} \boldsymbol{\Theta} \boldsymbol{h}_i^{D_t}[k]}_{\text{Reflection link}} + g_i^D[k]. \tag{3}$$

Therein, the channel for RIS departure link and arrival link can be denoted by

$$(\boldsymbol{h}_i^{D_r}[k])^{\mathrm{H}} = L(d_i^{D_r}) e^{-j2\pi \frac{d_i^{D_r}}{\lambda[k]}} \boldsymbol{a}_{AoD}^{\mathrm{H}}[k],$$

and

$$\boldsymbol{h}_i^{D_t}[k] = L(d_i^{D_t}) e^{-j2\pi \frac{d_i^{D_t}}{\lambda[k]}} \boldsymbol{a}_{AoA}[k],$$

respectively, where $d_i^{D_t}$ and $d_i^{D_r}$ denote the distance between RIS and $D_i^t$, between RIS and $D_i^r$, respectively. The arrival and departure array responses of the RIS are denoted as

$$\boldsymbol{a}_{AoD}[k] = \left[1, \ldots, e^{-j2\pi \frac{d_s}{\lambda[k]}(N-1)\sin(\theta_{AoD})}\right]^{\mathrm{T}},$$

and

$$\boldsymbol{a}_{AoA}[k] = \left[1, \ldots, e^{-j2\pi \frac{d_s}{\lambda[k]}(N-1)\sin(\theta_{AoA})}\right]^{\mathrm{T}},$$

respectively. The departure angel and the arrival angle of the RIS is denoted by $\theta_{AoD}$ and $\theta_{AoA}$, respectively. The symbol $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ represent transpose and conjugate transpose operation, $\mathrm{diag}[\cdot]$ represents the diagonal matrix.

The phase shift and amplitude attenuation $A$ for all the RIS elements can be expressed as $\boldsymbol{\Theta} \triangleq \mathrm{diag}\left[Ae^{j\theta_1}, Ae^{j\theta_2}, \ldots, Ae^{j\theta_N}\right]$, where $A \in [0,1]$ and $\theta \in [0, 2\pi)$ Similarly, the overall channel coefficient between $U_k$ and $D_i^r$, the channel coefficient between $l$-th D2D transmitter $D_l^t$ and $D_i^r$, the coefficient between $U_k$ and the BS, the coefficient between $D_i^t$ and the BS can be represented as

$h_{k,i}[k]$, $h_{l,i}^{D}[k]$, $h_{k}^{U}[k]$ and $h_{i}^{BS}[k]$, respectively. The signal $y_i[k]$ received by $D_i^r$ over the $k$-th sub-channel is denoted as

$$y_i[k] = \underbrace{h_i^D[k] \cdot x_i^D}_{\text{Desired signal}} + \underbrace{h_{k,i}[k] \cdot x_k^U}_{\text{Interference signal}} + \underbrace{z}_{\text{Noise}}, \tag{4}$$

where $x_i^D \triangleq \sqrt{p_i^D} u_i^D$ and $x_k^U \triangleq \sqrt{p_k^U} u_k^U$ denote the signal from $D_i^t$ and $U_k$, $p_i^D$ and $p_k^U$ denote the transmit power of the $D_i^t$ and $U_k$, and $u_i^D$ and $u_k^U$ represent the unit variance entries with zero mean, and $z \sim \mathcal{N}(0, \sigma^2)$ denotes the AWGN noise signal with mean 0 variance $\sigma^2$. Then, the signal-to-interference-plus-noise ratio (SINR) at $D_i^r$ and the BS for $U_k$ over the $k$-th sub-channel can be denoted as

$$\gamma_i^D[k] = \frac{p_i^D \left| h_i^D[k] \right|^2}{I_i[k] + \sigma^2}, \tag{5}$$

and

$$\gamma_k^U[k] = \frac{p_k^U \left| h_k^U[k] \right|^2}{\sum_{i=1}^{I} \rho_{k,i} p_i^D \left| h_i^{BS}[k] \right|^2 + \sigma^2}, \tag{6}$$

respectively, where $\rho_{k,i}$ is the resource reuse coefficient of $U_k$ and $i$-th D2D pair, and $\rho_{k,i} = 1$ when $i$-th D2D pair reuses the channel assigned to $U_k$. Otherwise, $\rho_{k,i} = 0$. Moreover, the interference to $D_i^r$ is given by

$$I_i[k] = \rho_{k,i} p_k^U \left| h_{k,i}[k] \right|^2 + \sum_{l=1, l \neq i}^{I} \rho_{k,l} p_l^D \left| h_{l,i}^D[k] \right|^2, \tag{7}$$

Then, the ergodic capacity for $i$-th D2D pair and for the $k$-th cellular user $U_k$ can be denoted by

$$C_i^D[k] = \mathbb{E}\left[ B[k] \log_2(1 + \gamma_i^D[k]) \right], \tag{8}$$

and

$$C_k^U[k] = \mathbb{E}\left[ B[k] \log_2(1 + \gamma_k^U[k]) \right], \tag{9}$$

respectively, where $\mathbb{E}[\cdot]$ represents the statistical expectation of $[\cdot]$, representing the expectation of the rate over the small scale fading distribution, $B_k$ is the bandwidth of $k$-th sub-channel. The channel capacity of underlay D2D networks could be expressed by

$$C_{\text{sum}} = \sum_{k=1}^{K} \left( \sum_{i=1}^{I} \rho_{k,i} C_i^D[k] + C_k^U[k] \right). \tag{10}$$

### *B. Problem formulation*

We aims to maximize the sum rate in (9) by jointly optimize the phase shift, the position of RIS, the resource reuse coefficient $\boldsymbol{\rho} = [\rho_{1,1}, \ldots, \rho_{1,I}, \ldots, \rho_{K,1}, \ldots, \rho_{K,I}]$, and the transmit power $\boldsymbol{p}^D = [p_1^D, \ldots, p_I^D]$ of D2D transmitters. The joint data rate maximization problem can be formulated as

$$\begin{aligned}
\text{P1:} \quad & \underset{\{\boldsymbol{P}^{RIS}, \boldsymbol{\Theta}, \boldsymbol{\rho}, \boldsymbol{p}^D\}}{\text{maximize}} && C_{\text{sum}} && \text{(11a)} \\
& \text{subject to} && p_i^D \leq p_{max}^D, \forall i \in I, && \text{(11b)} \\
& && \gamma_i^D \geq \gamma_{min}^D, \forall i \in I, && \text{(11c)} \\
& && \gamma_k^U \geq \gamma_{min}^U, \forall k \in K, && \text{(11d)} \\
& && \rho_{k,i} \in \{0,1\}, \forall i \in I, \forall k \in K, && \text{(11e)} \\
& && \sum_{k=1}^{K} \rho_{k,i} \leq 1, \forall i \in I, && \text{(11f)} \\
& && 0 \leq \theta_n < 2\pi, \forall n \in N, && \text{(11g)} \\
& && \boldsymbol{P}^{RIS} \in \boldsymbol{P}, && \text{(11h)}
\end{aligned}$$

where $\gamma_{min}^D$ and $\gamma_{min}^U$ are the SINR thresholds at the D2D receiver and the BS, respectively. Meanwhile, we restrict the location of the RIS in some discrete grids $\boldsymbol{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_o, \ldots, \boldsymbol{P}_O\}$, where $O$ represents the number of grids that RIS can be installed on. This is because that RIS cannot be installed everywhere so that the 2D continuous variables for the location of RIS is impractical. The grids $\boldsymbol{P}$ are distributed uniformly to simulate the distribution of RIS in reality. Constraint (11e) and (11f) assumes that each D2D pair only occupies one RB. Due to hardware limitations, RIS elements can only provide discrete phase shifts. This constraint (11e) and (11f) make (P1) non-convex. To solve the non-convex problem, we have to utilize exhaustive search, which is impractical when the number of D2D pairs, cellular users, and the number of RIS elements become large. Generally, classical optimization tools can be leveraged to acquire suboptimal solutions [31]–[33]. Alternatively, we leverage the D$^3$QN framework for resource allocation and the centralized DDQN framework for RIS optimization, which are more applicable to solve problems with high dimension inputs as well as large state and action space.

## III. Resource allocation optimization by proposed D$^3$QN framework

The optimization objective of (P1) is to jointly optimize the resource allocation for D2D pairs, plus the location and phase shift for the RIS. Rather than optimizing the configuration of RIS and the resource sharing centrally at the BS, we propose D$^3$QN framework to executed the resource allocation decentrally at each D2D pair. Meanwhile, the RIS is optimized by the centralized DDQN at the BS. By decoupling the joint optimization into sub-problems, we not only lower the computational pressure on the BS significantly, but also enable D2D pairs determine the resource sharing policy by their local information to reduce the transmission overhead.

In this section, we introduce the basic concept of RL and the proposed D$^3$QN framework for resource allocation of D2D pairs. The joint optimization for the location and phase shift of RIS is subsequently presented in Section IV.

### A. System description

Generally, the resource allocation optimization problem can be modeled as a linear sum assignment programming (LSAP) problem and can be solved by Hungarian algorithm [40] with computational complexity $O(K^3)$. The complexity is much higher if we take the transmit power of D2D transmitters into account. The high complexity of the Hungarian algorithm makes real-time optimization is impractical in the proposed D2D communications scenario. Additionally, the algorithm is required to be robust for fast channel variations and unstable CSI for different RIS implementations. Leveraging D$^3$QN, we can model the channel assignment and transmit power as a MARL problem and train the agents under different CSI conditions so that it can be adaptive to the various communications system. Noted that the D$^3$QN models at the D2D pairs should be trained offline. Actually, updating the resource allocation policy too quickly can cause challenges on convergence performance when we train the centralized DDQN for RIS optimization. This is because even if the RIS controller takes the exactly same action, the rewards would be various for different resource allocation policies, making the algorithm hard to converge. The unstable reward requires a robust resource allocation algorithm so that it can work under different RIS implementations.

Given the arbitrary location and phase shift of the RIS, the resource allocation optimization problem can be simplified into

$$\begin{aligned}
\text{P2: } \underset{\{\boldsymbol{\rho}, \boldsymbol{p}^D\}}{\text{maximize}} \quad & C_{\text{sum}} && \text{(12a)}\\
\text{subject to} \quad & p_i^D \le p_{max}^D, \forall i \in I, && \text{(12b)}\\
& \gamma_i^D \ge \gamma_{min}^D, \forall i \in I, && \text{(12c)}\\
& \gamma_k^U \ge \gamma_{min}^U, \forall k \in K, && \text{(12d)}\\
& \rho_{k,i} \in \{0,1\}, \forall i \in I, \forall k \in K, && \text{(12e)}\\
& \sum_{k=1}^{K} \rho_{k,i} \le 1, \forall i \in I. && \text{(12f)}
\end{aligned}$$

### B. Concept of reinforcement learning and DQN

RL is a branch of ML paradigm that allows agents to learn the optimal policy by the trial-and-error interaction with the environment to maximize the desired reward. Mathematically, the RL can be modeled as an markov decision process (MDP), including environment state $\mathcal{S}$, actions $\mathcal{A}$, and the reward $\mathcal{R}$ which can be determined for each state-action pair. During each training step $t$, each agent observes the state $s_t \in \mathcal{S}$ and then take an action $a_t \in \mathcal{A}$ according to a certain policy $\pi$. Then the agent receives the corresponding reward $r_t$ and turn to the next state $s_{t+1}$, which is determined by current state $s_t$ and action $a_t$ but independent of the past states. Formally, this process can be denoted by a transition tuple $e_t = (s_t, a_t, r_t, s_{t+1})$. The interaction process is shown in Fig. 2.

During each training step $t$, the objective of RL is to maximize the cumulative desired return from time $t$ to the future, which can be expressed by

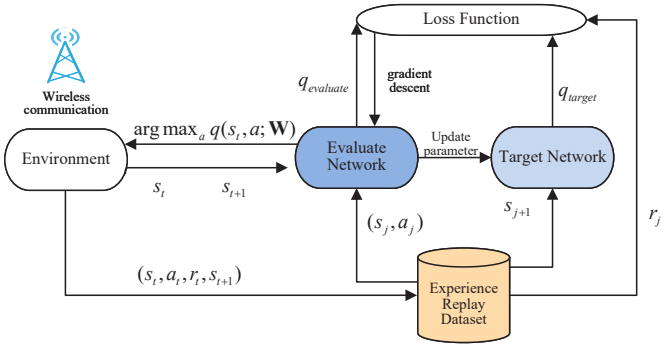



Fig. 2. The interaction of the DDQN at each agent with the environment. The parameters of the evaluate network will be updated to the target network periodically.

$$R_t = \sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau}, \tag{13}$$

where $\gamma \in (0,1)$ represents the discount factor which represents the impact of the future reward. The expectation reward for a state-action pair $(s,a)$, the action-value function, is defined as

$$q^\pi(s,a) = \mathbb{E}_\pi[R_t | s_t = s, a_t = a], \tag{14}$$

where policy $\pi$ is defined as a mapping from state $\mathcal{S}$ to the probability of choosing each action in $\mathcal{A}$.

The objective of RL is to find a optimal policy $\pi^* = \arg\max_\pi q^\pi(s,a)$. The optimal action-value function obeys an important identity known as the Bellman equation. The optimal policy is to select the action that maximizes the expected Q-value [16]:

$$q^*(s_t, a_t) = \mathbb{E}[r_t + \gamma \max_{a' \in \mathcal{A}} q^*(s_{t+1}, a') | s_t, a_t]. \tag{15}$$

Authors in [41] have shown that, $q(s_t, a_t) \to q^*(s_t, a_t)$ as $t \to \infty$. However, it is impractical since the training step is discrete. Instead, the NNs $\boldsymbol{W}$ are applied to be function approximator to estimate the action-value function, i.e., $q(s_t, a_t; \boldsymbol{W}) \approx q^*(s_t, a_t)$, which is the basic idea of the DQN. When the state and action space become large, this method does not need to maintain the large Q-table as conventional RL approaches do, thereby expanding the applications of RL in wireless communications greatly.

The training data set, also named replay memory $\mathcal{D} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_t, \ldots]$ for NN is stored according to agent's experience at each training step $t$, where the experience $\boldsymbol{e}_t = (s_t, a_t, r_t, s_{t+1})$ is called transition, including the current state, action, reward and next state information. The training minibatch $(s_j, a_j, r_j, s_{j+1})$ is sampled from the training data set. During the training process, parameters are updated to the Q estimation network at each step to generate the estimated Q-value. Q target network is updated after every $\upsilon$ steps according to the parameters in the Q estimation network. The training process is to minimize the error function which

represents the estimated Q-value and the realistic Q-value, which can be given by:

$$Loss(\boldsymbol{W}) = \mathbb{E}[(q_{target} - q(s_j, a_j; \boldsymbol{W}))^2], \quad (16)$$

where $q_{target} = r_j + \gamma \max_{a'} q(s_{j+1}, a'; \boldsymbol{W}^-)$ is the target Q-value for minibatch $j$, which is the output of Q target network in conventional DQN algorithm, $\boldsymbol{W}$ and $\boldsymbol{W}^-$ denotes the weights of the evaluation network and the target network, respectively. The weights are optimized by the gradient descent method [16].

### C. Double deep Q-network algorithm

The DQN algorithm can achieve a near-optimal performance in some scenarios, while sometimes it causes the overestimate problem. The target Q-value is approximately generated by the target network by maximizing the action-value function, while this target value is even higher than the true optimal action-value. The overestimate problem is severest when the number of actions becomes large, affecting the convergence and performance of the learned strategies. The idea of DDQN is decomposing action selection and action evaluation [19] to reduce overestimations. Unlike DQN that uses the evaluate network to estimate the action-value function and select action at the same time, DDQN uses the target network when evaluating the action-value function. In other words, the DDQN uses the evaluate network to select the action, while using the target network to fairly evaluate this action. The updated target Q-value function in DDQN is defined as

$$q_{target} = r_j + \gamma q(s_{j+1}, \arg\max_{a' \in \mathcal{A}} q(s_{j+1}, a'; \boldsymbol{W}); \boldsymbol{W}^-). \quad (17)$$

Note that DDQN is a model-free algorithm, which guarantees its robustness for different scenarios. Meanwhile, it is an off-policy algorithm which learns from the greedy policy and choose the action according to $\epsilon$-greedy algorithm to make a tradeoff between exploitation and exploration. The agent will choose actions uniformly from $\mathcal{A}$ with a probability of $\epsilon$, while choosing the action $a = \max_a q(s, a; W)$ which maximizes the Q-value with a probability of (1-$\epsilon$). In this paper, we leverage an improved algorithm called decaying $\epsilon$-greedy algorithm as shown in [42], so that we can achieve a better explored performance at the beginning and converged performance in the end.

### D. Observation state for D$^3$QN

In this subsection, the proposed D$^3$QN framework is introduced in details. As a decentralized framework, each D2D pair is modeled as an agent, concurrently making there own decision based on the local observation. An agent cannot directly acquire the global environment state $s_t$ which contains the global channel information and agents behaviour, thus the state design in [43] based on global SINR information is not applicable. Given the current environment state $s_t^{RA}$, the $i$-th D2D pair $D_i$ generates the unique observation $o(s_t^{RA}, i)$ from $s_t^{RA}$ at each training step $t$. Then it takes an action $a_t^{(i)}$, forming the joint action $\boldsymbol{a}_t$ with all the other agents. Then $D_i$ will receive an reward $r_t^{RA}$ and the environment turn to the next state $s_{t+1}^{RA}$. Observations $o(s_{t+1}^{RA}, i)$ in the next training step will then be generated by $D_i$.

Rather than the location information based state definition, the CSI based state definition enhances the robustness of the model. In other words, for $D_i$, the observation space includes:

1) Local channel information $h_i^D[k]$;
2) The interference channel from other D2D transmitters $h_{l,i}^D[k], \forall l \neq i, l \in I$;
3) The interference channel to the BS $h_i^{BS}[k]$;
4) The interference from cellular users $h_{k,i}[k], \forall k \in K$;
5) The interference power $I_i[k]$.

The information of channel $h_i^D[k]$, $h_{l,i}^D[k]$ and $h_{k,i}[k]$ can be estimated by D2D receiver accurately, while $h_i^{BS}[k]$ can be estimated and broadcast by the BS. Additionally, interference power $I_i[k]$ can be measured by D2D receiver. Hence, the observation space of $D_i$ at time $t$ can be denoted by $o(s_t^{RA}, i) = \{\{H_i[k]\}_{\forall k \in K}, \{I_i[k]\}_{\forall k \in K}\}$, where $H_i[k] = \{h_i^D[k], \{h_{l,i}^D[k]\}_{\forall l \in I, l \neq i}, h_i^{BS}[k], h_{k,i}[k]\}$.

Particularly, the multi-agent learning process can be described as Markov game. The state transition depends on actions taken by all of the agents, i.e., the joint action contributes to the state shift. Apart from the action taken by an agent itself, the actions of other agents can impact the reward of the agent, forming an unstable environment. The nonstationary environment from the view of each agent leads to nonstationary Q-function, making RL hard to converge. The nonstationarity challenge is tackled in [44] with a unique state, which includes view-based positional distribution and shared position information by each vehicle. However, the nonstationarity challenge is severer when combining with deep learning. The proposed D$^3$QN models use experience replay to feed the NN, while the environment that generated the data in the agent's replay memory is different from the current environment, and the convergence performance of the learning process is affected. To enable replay memory in MARL, authors in [27] designed a low dimensional *fingerprint* which includes the information of policy changes of other agents. The policy change is highly correlated with epoch $e$ and the exploration rate $\epsilon$. In other words, the observation space for $i$-th agent can be expressed as

$$z_t^{(i)} = \{o(s_t^{RA}, i), e, \epsilon\}. \quad (18)$$

Such *fingerprint* allows an agent to expect the policy change of other agents, thus improving the stationarity of the environment.

### E. Actions and rewards definition for D$^3$QN

As aforementioned, cellular users communicate with the BS on disjoint channels. Each D2D pair can choose one of $K$ subchannel which is preoccupied with a cellular user. The range of D2D transmit power including $A_p$ multiple discrete levels is $[0, p_{max}^D]$. As the result, the dimension of the action space is equal to $A_p \times K$. The actions of all agents form a joint action $\boldsymbol{a}_t$ which represents the resource reuse scheme.

Reward represents the objective of the optimization. All agents receive the same reward $r_t^{RA}$ according to the joint action $\boldsymbol{a}_t$ such that encouraging cooperative behaviors. To guarantee the QoS and the SINR requirement of D2D communications and cellular networks, the reward $r_t^{RA}$ is directly proportional to the sum rate for the successful transmission, i.e., (12c), (12d) and (12f) are satisfied. Note that a constant of proportionality is set as $\xi^{RA}$ to guarantee the training performance.

$$r^{RA} = \begin{cases} \xi^{RA} C_{\text{sum}}, & \text{if (12c), (12d) and (12f) are satisfied;} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

### F. Training algorithm for D$^3$QN

As introduced above, each D$^3$QN model at the D2D pair takes its observation state as the input. Several fully connected layers are leveraged as the hidden layer. During the training and testing phases, the RIS is randomly implemented and updated at the beginning of each epoch. One training epoch contains several training steps during which the D2D pairs interact with the wireless communication environment and store the experience in the training data sets, i.e., replay memories. The details of the training algorithm for the D$^3$QN models at the D2D pairs is shown in **Algorithm 1**.

**Algorithm 1** D$^3$QN training algorithm for the resource allocation.

**Input**: The observation space $z^{(i)}, \forall i \in I$;
Initialize the D$^3$QN models $\boldsymbol{W}^{(i)}, \forall i \in I$, for each D2D pair;
**for** each epoch **do**
  Initialize the implementation of RIS randomly;
  Update the large-scale fading channel;
  **for** each training step $t$ **do**
    **for** each D2D pair $i$ **do**
      Observe $z_t^{(i)}$;
      Choose action $a_t^{(i)}$ according to the observation $z_t^{(i)}$ and $\epsilon$-greedy algorithm;
    **end for**
    Form the joint action $\boldsymbol{a}_t$ and receive reward $r_t^{RA}$;
    Update the small-scale fading channel;
    **for** each D2D pairs $i$ **do**
      Observe $z_{t+1}^{(i)}$;
      Store transition $e_t^{RA} = (z_t^{(i)}, a_t^{(i)}, r_t^{RA}, z_{t+1}^{(i)})$ in $\mathcal{D}_i^{RA}$;
    **end for**
  **end for**
  **for** each D2D pair $i$ **do**
    Replay memory:
    Sample random minibatch of transitions $e_j^{RA} = (z_j^{(i)}, a_t^{(i)}, r_j^{RA}, z_{j+1}^{(i)})$ in $\mathcal{D}_i^{RA}$;
    Calculate $q_{target}^{RA}$ by (17)
    Perform a gradient descent step on $(q_{target}^{RA} - q^{RA}(z_j^{(i)}, a_t^{(i)}; \boldsymbol{W}^{(i)}))^2$;
  **end for**
**end for**

## IV. Joint Optimization for the Location and Phase Shift of RIS

As shown in Fig. 3, after resource allocation decisions are made by D2D pairs, the resource sharing policy will be sent to the BS as a part of the input information of the centralized DDQN to optimize the RIS. Based on the resource sharing information, the capacity optimization problem at the BS can be formulated as

$$\begin{aligned} \text{P3:} \quad & \underset{\{\boldsymbol{\Theta}, \boldsymbol{P}^{RIS}\}}{\text{maximize}} && C_{\text{sum}} && \text{(20a)} \\ & \text{subject to} && \gamma_i^D \geq \gamma_{min}^D, \forall i \in I, && \text{(20b)} \\ & && \gamma_k^U \geq \gamma_{min}^U, \forall k \in K, && \text{(20c)} \\ & && 0 \leq \theta_n < 2\pi, \forall n \in N, && \text{(20d)} \\ & && \boldsymbol{P}^{RIS} \in \boldsymbol{P}, && \text{(20e)} \end{aligned}$$

Based on the resource allocation information, a centralized DDQN model is proposed at the BS to solve the joint RIS location and phase shift optimization problem. Particularly, the DDQN components are introduced first, then the training algorithm and performance analysis are presented in the following.

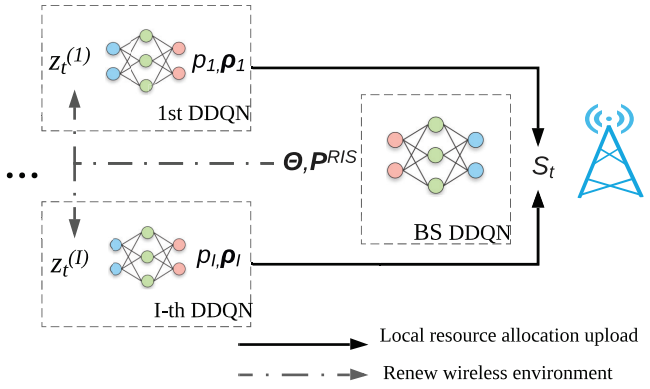



Fig. 3. Architecture of the proposed D$^3$QN framework and centralized DDQN. The framework contains $I$ decentralized DDQN models for the resource allocation optimization and one centralized DDQN models for RIS optimization.

### A. Reinforcement learning components definition for centralized DDQN

For the centralized DDQN, the input state contains the channel assignment coefficient $\boldsymbol{\rho}$, the transmit power of each D2D transmitters, the RIS phase shift $\boldsymbol{\Theta}$, as well as the location information $\boldsymbol{P}^{RIS}$ of RIS, which is denoted as $\mathcal{S} = [\boldsymbol{\rho}, \boldsymbol{p}^D, \boldsymbol{P}^{RIS}, \boldsymbol{\Theta}]$.

Action set $\mathcal{A}$ represents the possible action choice for the RIS controller. Generally, the location of RIS need to be optimized before installation, while the phase shift can be adjusted, so the action space contains the phase shift and location of RIS. At training step $t$, action $a_t$ consists of two parts: i) the variable quantity of phase shift matrix, $\Delta\boldsymbol{\Theta} =$

$\{\Delta\theta_1, \ldots, \Delta\theta_N\}$, where $\Delta\theta_n \in \{-\delta, 0, +\delta\}, \forall n \in N$; ii) the location choice of RIS, $\boldsymbol{P}^{RIS} \in \boldsymbol{P}$. Formally, the action $a_t = [\Delta\boldsymbol{\Theta}; \boldsymbol{P}^{RIS}]$, which has a cardinality $|a|$ of $(N+2)$. Action set $\mathcal{A}$ includes all possible actions with the cardinality $|\mathcal{A}| = 3^N \times O$.

The reward represents whether we encourage or punish an action, so it is defined based on the objective function given in (5). The reward for RIS optimization $r^{RIS}$ can be defined as a similar way as $r^{RA}$ in (19). For a successful transmission at training step $t$, i.e., the constraints (20b) and (20c) are satisfied, the reward can be defined as directly proportional to the sum rate at this training step, otherwise a punishment reward 0 will be received by the BS.

### *B. Proposed centralized DDQN algorithm for the control of RIS*

Leveraging the NN, the centralized DDQN model can find the relationship between the input state and the corresponding deployment of RIS. The components in centralized DDQN is defined as

- **Agent**: The agent for the centralized DDQN model is the BS. The BS processes the inputs and executes the outputs of DDQN to control RIS.
- **Input**: At each training step $t$, the centralized DDQN model takes the states $\mathcal{S}_t$ as the input, which includes the resource allocation, current RIS position, and current phase shift information.
- **Output**: The output of the centralized DDQN model is the evaluated Q-value for state-action pairs. The output layer contains $|\mathcal{A}|$ units, which represents the number of possible actions. As shown in Fig. 2, two identical networks are set: evaluation network and target network. In the evaluation network, the current state $\mathcal{S}_j$ is the input information, and the output is the evaluate Q-value for each action. In the target network, the next expected state $\mathcal{S}_{j+1}$ is the input, while the output is the Q-value for each action in the next state.

By receiving the input information, the RIS controller can train the weights and update NNs to estimate the action-value function. The proposed centralized DDQN algorithm for RIS optimization is shown in **Algorithm 2**.

**Algorithm 2** Centralized DDQN algorithm for the RIS optimization at the BS.

**Input**: Current RIS position and phase, resource allocation policy;
Initialize: action-value function Q with random weights $\boldsymbol{W}$, replay memory $\mathcal{D}^{\mathcal{RIS}}$, Trained D$^3$QN framework for resource allocation;
**for** each epoch **do**
  Update the large-scale fading channel;
  **for** each training step $t$ **do**
    Observe state $\mathcal{S}_t$;
    Choose $a_t \in \mathcal{A}$ according to $\epsilon$-greedy algorithm;
    Execute $a_t$ and update the small-scale fading channel;
    Execute D$^3$QN and perform resource allocation;
    Calculate reward $r_t^{RIS}$ by (19);
    Observe $s_{t+1}$;
    Store transition $e_t^{RIS} = (\mathcal{S}_t, a_t, r_t^{RIS}, \mathcal{S}_{t+1})$ in $\mathcal{D}$;
  **end for**
  **if** learning begins **then**
    Replay memory:
    Sample random minibatch of transitions $(\mathcal{S}_j, a_j, r_j^{RIS}, \mathcal{S}_{j+1})$ in $\mathcal{D}^{\mathcal{RIS}}$;
    Calculate $q_{target}^{RIS}$ according to (17);
    Perform a gradient descent step on $(q_{target}^{RIS} - q(\mathcal{S}_j, a_j; \boldsymbol{W}))^2$;
  **end if**
**end for**
**Return**: action-value function and optimized action $a$.

### *C. Performance analysis for D$^3$QN and centralized DDQN*

*1) Computational complexity:* Generally, the concept of floating point operations (FLOPs) is used to measure the computational complexity for NNs. For each fully connected layer, the number of FLOPs is $[N_{in} + (N_{in} - 1) + 1] \times N_{out}$, where $N_{in}$ and $N_{out}$ represents the number of neurons. For the centralized DDQN at the BS, the number of FLOPs is $\text{FLOPs}(BS) = 2[|\mathcal{S}| \times N_1^{BS} + N_1^{BS} \times N_2^{BS} + N_2^{BS} \times N_3^{BS} + N_3^{BS} \times (3^N \times O)]$, where $|\mathcal{S}|$ denotes the dimension of the input state, $N_\mu^{BS}$ represents the number of neurons in $\mu$-th layer of the DDQN at the BS. For the D$^3$QN model at each D2D pair, the number of FLOPs is $\text{FLOPs}(i) = 2[|z^{(i)}| \times N_1(i) + N_1(i) \times N_2(i) + N_2(i) \times (A_p \times K)]$, where $|z^{(i)}|$ the dimension of the observation space of $D_i$. Therefore, the overall computational complexity can be expressed by

$$\text{FLOPs} = \underbrace{\text{FLOPs}(BS)}_{\text{At the BS}} + \underbrace{\sum_{i=1}^{I} \text{FLOPs}(i)}_{\text{At D2D pairs}}. \tag{21}$$

Note that the proposed D$^3$QN models for resource allocation could be trained offline because it is robust to the dynamic environment. Compared to the alternative maximization (AM) approaches [31]–[33] that optimize the resource allocation and RIS configuration by iterations, the proposed trained model only requires a little computational complexity to generate solutions.

*2) Communication cost:* Compared with the centralized algorithms that the users need to upload the local information to the BS and receive the optimized control signals from the BS, the proposed decentralized resource allocation algorithm enables users complete the resource allocation process locally, thus reducing the communication cost significantly. Due to the fast channel variation, the AM approaches via several iterations are non-applicable, while the centralized DDQN algorithm for resource allocation requires the global real-time CSI. Local users will upload the local observation information to the BS in each training step, which causes heavy transmission overhead. Assuming that the fast fading is updated every 1ms, while the pathloss and shadowing are updated every 100ms, which means we also set $T_s = 100$ training steps in each training epoch. The fast fading channel and

large scale fading channel are renewed in each training step and epoch, respectively. Therefore, for the resource allocation task, the uplink communication cost for the centralized DDQN algorithm is denoted as

$$C_{\text{CRL}} = T_e \times T_s \times \sum_{i=0}^{I} |z^{(i)}|, \tag{22}$$

where $T_e$ and $T_s$ represents the number of maximum training epochs and the number of training steps in each training epoch. On the other hand, for proposed D$^3$QN framework, D2D pairs choose their own access channel and transmit power under the guidance of the local NN, and only upload the optimized transmit power and channel assignment results to the BS in each training epoch. Thereby, the corresponding uplink communication cost can be given by

$$C_{\text{D}^3\text{QN}} = T_e \times I \times (K+1). \tag{23}$$

In a nutshell, the proposed D$^3$QN framework outperforms the traditional mathematical solutions and centralized algorithms in terms of communication cost and computational complexity.

## V. Numerical Results

In this section, the performance of the proposed D$^3$QN framework for resource allocation and centralized DDQN for RIS optimization are evaluated by comparing it with the benchmark algorithms. Assuming that the cellular users and D2D pairs are distributed in a $100\text{m}\times100\text{m}$ square. The whole area is divided into $O = 16$ identical grids, where RIS can be installed in any of them. We apply the simulation settings in [39] to model the channel, and the simulation parameters are listed in TABLE I.

In the proposed D$^3$QN framework, each DDQN consists of a 5-layer fully connected (FC) NN with 3 hidden layers. The number of neurons in the three hidden layers are set to 500, 250, and 120, respectively. We apply the rectified linear unit (ReLU) function as the activate function, which is defined as $f(x) = \max(0, x)$, while the RMSProp optimizer [45] is applied to train the NNs. Note that the trained resource allocation model only needs to be updated when the wireless communication system experience significant changes, thus the resource allocation model is first trained and remains unchanged during the optimization of the RIS. We train the centralized DDQN for 3000 epochs and the exploration rate $\epsilon$ decreases from 1 to 0.02 over 2700 epochs linearly. The discount factor $\gamma$ is set to 0.95.

We compare the proposed D$^3$QN framework with the following benchmarks:

1) **Exhaustive search**: Exhaustive search is adopted to acquire the optimal resource allocation and RIS deployment.
2) **No RIS and random RA**: This scheme does not deploy RIS for enhancement, and the resource allocation is execute randomly.
3) **Fully random baseline**: The random RIS deployment with random resource allocation scheme is adopted in this scheme.

TABLE I
Simulation Parameters

| Parameter | Value |
|---|---|
| Number of D2D users $2 \times I$ | 4 |
| Number of Cellular users (sub-channels) $K$ | 4 |
| Phase shift variable quantity $\delta$ | $\frac{\pi}{4}$ |
| Number of RIS elements $N$ | 8 |
| Cellular transmit power range | 23dBm |
| D2D transmit power range | $[0, 24]$dBm |
| Number of discrete levels $A_p$ | 9 |
| Minimum SINR requirements for D2D receiver $\gamma^D_{min}$ | 3dB |
| Minimum SINR requirements for the BS $\gamma^U_{min}$ | 5dB |
| Carrier frequency | 2GHz |
| Bandwidth of each sub-channel | 1MHz |
| Cellular antenna height $Z^U$ | 1.5m |
| D2D antenna height $Z^D$ | 1.5m |
| BS antenna height $Z^{BS}$ | 25m |
| RIS antenna height $Z^{RIS}$ | 10m |
| Bandwidth of each sub-channel | 1MHz |
| Noise power $\sigma^2$ | $-115$dBm |

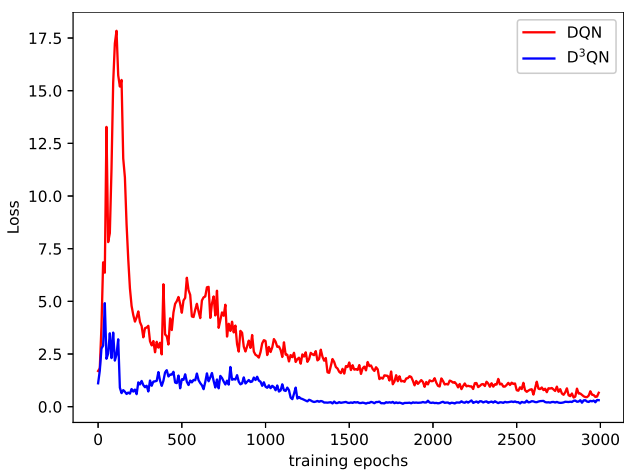
4) **RIS random baselines**: These group of baselines adopt the random deployment of RIS, where the resource allocation tasks are optimized by exhaustive search, conventional DQN algorithm, and the proposed D$^3$QN framework, respectively.
5) **DQN**: This scheme adopts the conventional DQN for RIS optimization, while the resource allocation is performed by decentralized DQN framework.

### A. Training performance of proposed algorithms

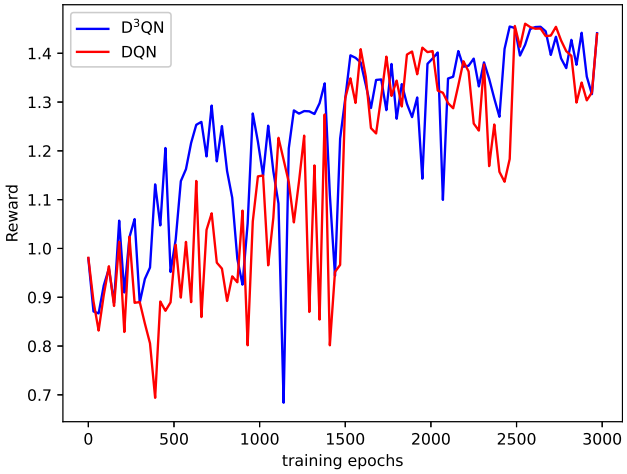
Fig. 4(a) and Fig. 4(b) demonstrate the training performance of proposed framework. It is noted that these two figures shows the loss and the reward of the NN at the BS for the optimization of RIS. Fig. 4a illustrates the loss comparison. The loss of the proposed algorithm is lower than the DQN training method during the whole training epochs and can achieve faster convergence since it avoids the overestimation problem caused by the DQN based approach. A lower loss of the proposed algorithm leads to a higher training reward since the estimated maximum value of NN is closer to the practical maximum value. As shown in Fig. 4b, both average rewards per epoch of proposed and DQN algorithm improve as training continues, while the proposed DDQN framework outperforms the DQN method. Due to the high training loss at the beginning, the DQN method achieve lower reward than the proposed framework.

### B. Effectiveness and robustness testing

In the testing phase, we verify the effectiveness of the proposed D$^3$QN framework and centralized DDQN model. The number of testing step is set as 30 epochs and the exploration rate $\epsilon$ is set as 0 which means the D2D paris always takes the action that has the highest Q-value. In each testing epoch, the phase shift and position of RIS are configured randomly to test the robustness of the proposed scheme. As illustrated in Fig. 5, the proposed algorithm can achieve near-optimal performance, reaching over 90% of the




(a) Loss comparison.




(b) Reward comparison.

Fig. 4. Training performance comparison for the RIS optimization at the BS. The trained D$^3$QN models are adopted for the resource allocation task. The reward curve shows the averaged reward per 30 epochs to enhance the readability.

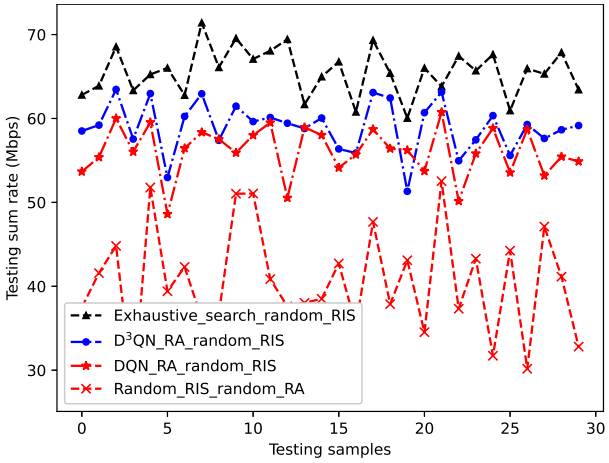



Fig. 5. Testing performance of proposed decentralized structure.

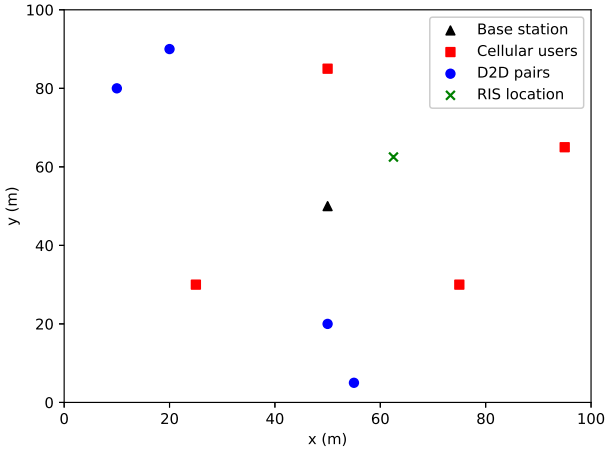



Fig. 6. The RIS location optimization result.

optimal solution. Note that the decentralized DQN framework achieves less sum rate than the proposed framework because the DQN based algorithms suffers from the overestimation problem, thus cannot take appropriate action to maximize the objective function and guarantee the SINR requirements for the BS and D2D receivers.

We also illustrate the effectiveness of the proposed algorithm. Fig. 6 illustrates a snapshot of the optimized RIS location. We can see that the RIS is located near the BS to enhance the cellular uplink rate, while stay far away from the D2D pairs to lower the interference to cellular networks caused by D2D communications.

Furthermore, we compare the proposed framework with other benchmarks in various wireless communication scenarios to verify the robustness and effectiveness of the proposed algorithm. Fig. 7 shows the sum rate performance under various SINR scenarios. It is observed that the sum rates of all schemes monotonically decrease with higher noise power. As expected, the proposed algorithm shows near-optimal performance, and outperforms the decentralized DQN and random scheme in all the cases. On the other hand, even if the resource allocation have been optimized, the random RIS scheme has significant performance loss, which verify the effectiveness of RIS phase shift and location optimization. The fully random scheme and the conventional wireless communication scheme fail to leverage the enhancement of RIS, providing less than 60% of the optimal sum rate.

Fig. 8 demonstrates the sum rate with different number of cellular users. With the increasing of accessible channels, the sum rates of all schemes increases. We also investigate the sum rate improvement over 4, 8 and 16 RIS elements. It is obvious that more RIS elements can improve the sum rate of the considered network, while the performance improvement slows down as the number of elements becomes larger. This is because the interference between D2D communications and cellular networks increases as well. Meanwhile, the computational complexity for RIS optimization at the BS also increases with more RIS elements. In practice, it is worth determining the number of RIS elements to strike a tradeoff between the sum rate performance and the training complexity.

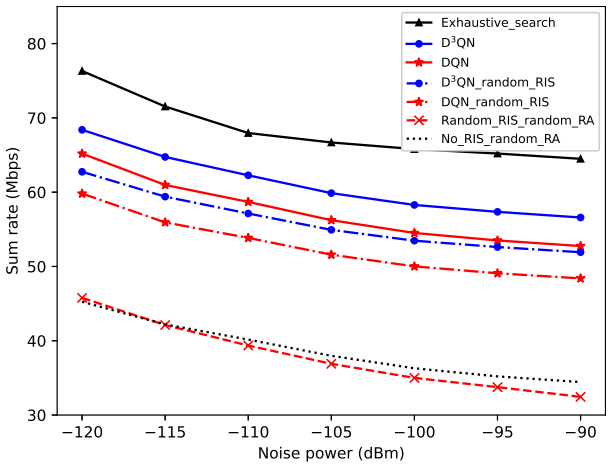



Fig. 7. Sum rate comparison over different noise power the optimized and random RIS deployment.

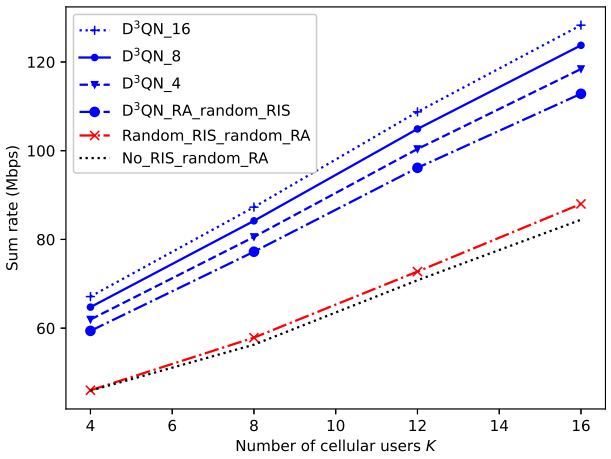



Fig. 8. Sum rate over the number of cellular users $K$ with different number of RIS elements.

## VI. Conclusion

In this paper, we aim to maximize the sum rate of the device-to-device users and the cellular users. To solve the non-convex problem with lower complexity, a novel decentralized double deep Q-network framework for resource allocation and a centralized double deep Q-network framework for RIS optimization have been proposed. We decoupled the joint optimization task into sub-problems and reduced the computational pressure at the central BS by decentralized resource allocation problem. Leveraging the double deep Q-network algorithm, the overestimation challenge has been overcome. Communication cost and computational complexity analysis and simulation results show that the proposed algorithm achieved the near-optimal performance while offloading the computational pressure from the base station significantly. The proposed framework is verified to be robust to fast channel variations and various noise scenarios. It also outperforms the other benchmarks in terms of the achieved sum rate.

## References

[1] Z. Ji and Z. Qin, "Reconfigurable intelligent surface enhanced device-to-device communications," in *Proc. IEEE Global Commun. Conf.*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

[2] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 3, pp. 2133–2168, Apr. 2018.

[3] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[4] M. Höyhtyä, O. Apilo, and M. Lasanen, "Review of latest advances in 3GPP standardization: D2D communication in 5G systems and its energy consumption models," *Future Internet*, vol. 10, no. 1, p. 3, Jan. 2018.

[5] Cisco, "Cisco annual internet report (2018–2023) white paper," 2020. [Online]. Available: https://www.cisco.com

[6] C. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[7] M. Robat Mili, P. Tehrani, and M. Bennis, "Energy-efficient power allocation in OFDMA D2D communication by multiobjective optimization," *IEEE Wireless. Commun. Lett.*, vol. 5, no. 6, pp. 668–671, Dec. 2016.

[8] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for D2D-enabled vehicular communications," *IEEE Trans. on Commun.*, vol. 65, no. 7, pp. 3186–3197, Apr. 2017.

[9] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable ai meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, May 2019.

[10] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546–1577, third quarter 2021.

[11] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[12] X. Tan, Z. Sun, J. M. Jornet, and D. Pados, "Increasing indoor spectrum sharing capacity using smart reflect-array," in *Proc. IEEE Internat. Conf. on Commun. (ICC)*, Kuala Lumpur, Malaysia, May. 2016, pp. 1–6.

[13] B. Wu, A. Sutinjo, M. E. Potter, and M. Okoniewski, "On the selection of the number of bits to control a dynamic digital MEMS reflectarray," *IEEE Antennas Wirel. Propag. Lett.*, vol. 7, pp. 183–186, Mar. 2008.

[14] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. on Commun.*, vol. 68, no. 3, pp. 1838–1851, Dec. 2020.

[15] L. Dai, B. Wang, M. Wang, X. Yang, J. Tan, S. Bi, S. Xu, F. Yang, Z. Chen, M. D. Renzo, C. Chae, and L. Hanzo, "Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45 913–45 923, Mar. 2020.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," in *Proc. NIPS Deep Learn.Workshop*, Dec. 2013.

[17] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, Nov. 2019.

[18] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[19] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. of the AAAI conference on artificial intelligence*, vol. 30, no. 1, Mar. 2016.

[20] X. Liu, J. Yu, Z. Feng, and Y. Gao, "Multi-agent reinforcement learning for resource allocation in iot networks with edge computing," *China Commun.*, vol. 17, no. 9, pp. 220–236, Sep. 2020.

[21] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Aug. 2019.

[22] H. V. Vu, Z. Liu, D. H. N. Nguyen, R. Morawski, and T. Le-Ngoc, "Multi-agent reinforcement learning for joint channel assignment and power allocation in platoon-based C-V2X systems," *arXiv preprint arXiv:2011.04555*, Nov. 2020.

[23] L. Wang, H. Ye, L. Liang, and G. Y. Li, "Learn to compress CSI and allocate resources in vehicular networks," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3640–3653, Mar. 2020.

[24] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 341–356, Feb. 2020.

[25] S. Dominic and L. Jacob, "Distributed resource allocation for D2D communications underlaying cellular networks in time-varying environment," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 388–391, Nov. 2018.

[26] Z. Qin, H. Ye, G. Y. Li, and B. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[27] J. Foerster *et al.*, "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 1146–1155.

[28] M. Di Renzo, A. Zappone, M. Debbah, M. S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[29] N. Gao, S. Jin, X. Li, and M. Matthaiou, "Aerial ris-assisted high altitude platform communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2096–2100, Jun. 2021.

[30] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, "Reconfigurable intelligent surface (RIS) assisted wireless coverage extension: RIS orientation and location optimization," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 269–273, Jan. 2021.

[31] Y. Cao and T. Lv, "Sum rate maximization for reconfigurable intelligent surface assisted device-to-device communications," *arXiv preprint arXiv:2001.03344*, Jan. 2020.

[32] M. Fu, Y. Zhou, and Y. Shi, "Reconfigurable intelligent surface for interference alignment in MIMO device-to-device networks," *arXiv preprint arXiv:2005.06766*, May 2020.

[33] C. Pradhan, A. Li, L. Song, J. Li, B. Vucetic, and Y. Li, "Reconfigurable intelligent surface RIS-enhanced two-way OFDM communications," *arXiv preprint arXiv:2005.01910*, May 2020.

[34] M. A. El Mossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Trans. on Cogn. Commun. and Netw.*, p. 1–1, May 2020.

[35] H. Gacanin and M. D. Renzo, "Wireless 2.0: Towards an intelligent radio environment empowered by reconfigurable meta-surfaces and artificial intelligence," *arXiv preprint arXiv:2002.11040*, Dec. 2020.

[36] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *arXiv preprint arXiv:1904.10136*, Apr. 2019.

[37] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Areas Commun.*, pp. 1–1, Dec. 2020.

[38] G. Yang, Y. Liao, Y.-C. Liang, and O. Tirkkonen, "Reconfigurable intelligent surface empowered device-to-device communication underlaying cellular networks," *arXiv preprint arXiv:2006.02103*, Jul. 2020.

[39] *Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz (Release 16)*, 3GPP TR 38.901 V16.1.0, 3rd Generation Partnership Project, Dec. 2019.

[40] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[41] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 2018.

[42] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *arXiv preprint arXiv:2001.10363*, Jan. 2020.

[43] Z. Ji, A. K. Kiani, Z. Qin, and R. Ahmad, "Power optimization in device-to-device communications: A deep reinforcement learning approach with dynamic reward," *IEEE Wireless. Commun. Lett.*, pp. 1–1, Nov. 2020.

[44] A. Gündogan, H. M. Gürsu, V. Pauli, and W. Kellerer, "Distributed resource allocation with multi-agent deep reinforcement learning for 5G-V2V communication," *arXiv preprint arXiv:2010.05290*, Oct. 2020.

[45] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint, arXiv:1609.04747v2*, Jun. 2016.