Dynamic Spectrum Access for D2D-Enabled Internet of Things: A Deep Reinforcement Learning Approach

Jingfei Huang, Yang Yang, *Member, IEEE*, Zhen Gao, *Member, IEEE*, Dazhong He, *Member, IEEE*, and Derrick Wing Kwan Ng, *Fellow, IEEE*

***Abstract*—Device-to-device (D2D) communication is regarded as a promising technology to support spectral-efficient Internet of Things (IoT) in beyond fifth-generation (5G) and sixth-generation (6G) networks. This article investigates the spectrum access problem for D2D-assisted cellular networks based on deep reinforcement learning (DRL), which can be applied to both the uplink and downlink scenarios. Specifically, we consider a time-slotted cellular network, where D2D nodes share the cellular spectrum resources (CUEs) with cellular users in a time-splitting manner. Besides, D2D nodes could reuse time slots preoccupied by CUEs according to a location-based spectrum access (LSA) strategy on the premise of cellular communication quality. The key challenge lies in that D2D nodes have no information on the LSA strategy and the access principle of CUEs. Thus, we design a DRL-based spectrum access scheme such that the D2D nodes can autonomously acquire an optimal strategy for efficient spectrum access without any prior knowledge to achieve a specific objective such as maximizing the normalized sum throughput. Moreover, we adopt a generalized double deep *Q*-network (DDQN) algorithm and extend the objective function to explore the resource allocation fairness for D2D nodes. The proposed scheme is evaluated under various conditions and our simulation results show that it can achieve the near-optimal throughput performance with different objectives compared to the benchmark, which is the theoretical throughput upper bound derived from a genius-aided scheme with complete system knowledge available.**

***Index Terms*—Device-to-device (D2D) communication, deep reinforcement learning (DRL), dynamic spectrum access, Internet of Things (IoT).**

Manuscript received 24 January 2022; revised 1 March 2022; accepted 14 March 2022. Date of publication 17 March 2022; date of current version 7 September 2022. This work was supported in part by the Beijing Natural Science Foundation under Grant L182024; in part by the National Natural Science Foundation of China under Grant 61801035; and in part by the Guangxi Science and Technology Base and Talent Special Project under Grant AD19110042; and in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2022D09. *(Corresponding author: Yang Yang.)*

Jingfei Huang, Yang Yang, and Dazhong He are with the School of Artificial Intelligence and the Center for Data Science, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: buptyy1015@gmail.com).

Zhen Gao is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China, and also with the Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China.

Derrick Wing Kwan Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2025, Australia.

Digital Object Identifier 10.1109/JIOT.2022.3160197

I. Introduction

THE FIFTH-GENERATION (5G) technology has been deployed commercially, which provides massive machine-type communication (mMTC) [1] and ultrareliable and low-latency communication (URLLC) services, thus greatly promoting the development of Internet of Things (IoT) applications [2]. In future sixth-generation (6G) networks, combined with some advanced technologies, such as massive access [3], mobile edge computing [4], [5], and intelligent remote management [6], the IoT is regarded as a fundamental building block for supporting reliable massive access and seamless communication among various devices, paving the way to revolutionize the industries.

As one of the typical IoT applications, device-to-device (D2D) communication has attracted extensive attention for enabling high transmission data rate and low latency. Specifically, in D2D communications, two adjacent mobile devices compose a pair of D2D nodes and communicate with each other directly without involving the base station (BS) [7]. For D2D-enabled IoT applications, D2D communication allows cooperation and information exchange among IoT devices in the proximity of each other [8], [9]. Also, it supports relaying for data gathered by IoT sensors to the BS, thus improving the communication quality and decreasing latency for IoT applications. In addition, as D2D communication can reuse the time-frequency resources of cellular communication, it is beneficial to improve the overall spectral efficiency and energy efficiency for the 6G-enabled IoT network [10]–[12]. It is anticipated that the D2D-enabled IoT communications will play a vital role in the development of future intelligent wireless communication systems. However, D2D communications among IoT devices would lead to severe interference to other cellular users (CUEs) when they transmit concurrently via the same frequency spectrum. As a result, it is important to employ proper spectrum access schemes to mitigate co-channel interference for D2D-assisted cellular communications.

A. Related Work

To address the coexistence issue of D2D nodes and CUEs under limited radio resources, numerous existing works have emerged to investigate the spectrum sharing problem for D2D communications. For instance, graph theory has been applied to address the resource allocation problem in D2D networks [13],

[14] for many years. Also, a bipartite graph was constructed in [15], where the weight of the bipartite graph was characterized as the maximization of sum rate of the associated D2D and cellular links under outage constraints. On the other hand, game theory is also a powerful handy method to deal with spectrum sharing problems. For example, based on the auction game theory, Deng *et al.* [16] investigated the dynamic spectrum sharing for hybrid access in cognitive macrofemtocell networks and Zaki *et al.* [17] proposed a resource allocation algorithm to efficiently share resources of CUEs through D2D communication. Besides, Nguyen *et al.* [18] modeled the competition among D2D nodes as a noncooperative power control game and aimed to maximize the average rate of D2D while ensuring a minimum rate for CUEs. However, such methods mainly rely on model-dependent problem settings and require more sophisticated computation, which cannot be effectively utilized to deal with real-world problems. Fortunately, reinforcement learning (RL) [19], as a branch of machine learning, has been regarded as a powerful tool for making intelligent decisions, which enables the agent to find an optimal policy to maximize the long-term reward without knowing the environment model. It is important to note that deep RL (DRL) has been applied to address dynamic spectrum access problems. For example, in [20], DRL enables secondary users (SUs) to make spectrum access decisions relying on their own spectrum sensing outcomes, which can help the SU to significantly reduce the chances of collision with primary users and other SUs. Also, there is an upsurge of research interests in exploiting RL algorithms to design resource allocation schemes for D2D-assisted cellular networks. The majority of these works pay more attention to the scenario where the common cellular resources are shared by both D2D and cellular communications simultaneously (i.e., in a nonorthogonal manner) [21]. These works usually strive to mitigate the interference between D2D and cellular communications by proposing efficient spectrum allocation or power control methods [22]–[26]. Specifically, Zia *et al.* [22] proposed a spectrum allocation scheme based on Q-learning, which aims to maximize the system sum-rate by mitigating the co-channel interference between D2D and cellular communications. Furthermore, in [23], a distributed framework for spectrum allocation based on the actor–critic (AC) algorithm was designed, which collects all users' historical information or neighbor users' historical information to train the DRL model via a centralized training process. In addition, some works jointly consider the power control and spectrum allocation for D2D-assisted cellular networks [24], [25]. Specifically, Wang *et al.* [24] aimed to maximize the system capacity and spectral efficiency while minimizing the interference to CUEs, and Guo *et al.* [25] aimed to maximize the sum throughput under the constraints of latency and reliability. Also, Ye *et al.* [26] proposed a resource allocation scheme based on DRL for vehicle-to-vehicle (V2V) communication, where V2V links aim to learn an optimal resource allocation strategy to satisfy the latency constraints of V2V links while minimizing the interference to vehicle-to-infrastructure (V2I) links.

In contrast, some works assume that the spectrum resources are allocated to D2D and cellular communications orthogonally (i.e., in an orthogonal manner) [21]. As such, they are committed to mitigating the co-channel interference among D2D nodes [27]–[29]. Particularly, Zhang *et al.* [27] investigated a joint resource allocation and communication mode selection optimization problem, aiming to maximize the system energy efficiency. Besides, Moussaid *et al.* [28] proposed a D2D transmission scheme based on the deep Q-network (DQN) algorithm to maximize the sum rate of the D2D network. In addition, Tan *et al.* [29] jointly considered the channel selection and power control optimization problem and proposed a DRL-based scheme to maximize the weighted-sum rate of the D2D network. The comparison of the related works and our work is further summarized in Table I. Although all these works have made a great contribution to the resource allocation for D2D-assisted networks, they mainly consider only a particular spectrum access manner and rarely explore the fairness while allocating resources among D2D nodes.

B. Contribution

Inspired by the aforementioned literature, this article investigates a hybrid spectrum access scheme for D2D-assisted cellular networks with the consideration of both nonorthogonal and orthogonal access manners, which can be applied to both the uplink and the downlink scenarios. More specifically, we investigate a dynamic time-slotted network where CUEs and D2D nodes are allowed to access a shared cellular spectrum channel in different time slots. Certain time slots in each frame are orthogonally preassigned among CUEs. Besides, D2D nodes are allowed to transmit in the same time slot with CUEs according to a location-based spectrum access (LSA) strategy to improve the spectral efficiency. Taking the uplink scenario as an example, when the CUE is physically close enough to the BS, the channel quality of the cellular link is sufficiently good and is tolerable to the interference caused by D2D communication. As such, D2D nodes can utilize both the resources allocated to the CUE and the remaining unused resources. As when the CUE is far away from the BS, the signal received by the BS is weak and thus, D2D nodes can only exploit the unused resources to strike a balance between the performance of cellular communication and D2D transmission. On the basis of this fact, D2D nodes can choose between two spectrum access manners (nonorthogonal access and orthogonal access) based on the position of the CUE to reduce the impacts caused by severe interference.

The key challenge lies in that D2D nodes have no information on the system model, including the access principle of CUEs and the LSA strategy. Thus, based on the DRL technology, we design a spectrum access scheme by exploiting the double DQN (DDQN) algorithm, an extension of conventional DQN algorithm to avoid the overestimation of Q-values. The goal of the proposed scheme is to enable D2D nodes to learn an optimal strategy for spectrum access autonomously such that a specific objective such as maximizing the normalized sum throughput or maximizing a fairness objective function can be achieved without knowing any prior knowledge of the system. Specifically, by properly defining the action space, the state space, and the reward function, the

TABLE I
COMPARISON OF RELATED WORKS WITH OUR WORK

Related works	Optimization method	Optimization objective	Optimization actions	Access manner
[15]	Graph theory (bipartite graph) and Hungarian algorithm	Maximize ergodic sum rates under transmit power and outage constraints	Find the optimal pairing between D2D links and CUEs	Non-orthogonal access
[17]	Game theory (auction-based algorithm)	Maximize the system data rate	Resource blocks (RBs) selection	Non-orthogonal access
[18]	Game theory (non-cooperative power control game)	Guarantee the QoS of CUEs and maximize the data rate	Power control and RBs selection	Non-orthogonal access
[22]	Q-learning algorithm	Maximize the D2D users throughput with minimal interference to CUEs	RBs selection	Non-orthogonal access
[23]	Actor-critic (AC) algorithm	Maximize the sum rate of D2D links while guaranteeing the transmission quality of CUEs	RBs selection	Non-orthogonal access
[24]	Deep Q-network (DQN) algorithm	Maximize the system capacity and spectral efficiency while minimizing the interference to CUEs	Power control and channel selection	Non-orthogonal access
[25]	Queueing analysis and Hungarian algorithm	Maximize the sum throughput of CUEs with guarantee on DUEs' reliability and latency	Power and spectrum allocation	Non-orthogonal access
[26]	DQN algorithm	Satisfy the latency constraints on V2V links while minimizing the interference to V2I links	Power and spectrum allocation	Non-orthogonal access
[27]	Deep deterministic policy gradient (DDPG) algorithm	Maximize the energy efficiency	Communication mode selection and resource allocation	Orthogonal access
[28]	DQN algorithm	Maximize the sum-rate of D2D network	Select proper D2D links to activate	Orthogonal access
[29]	DQN algorithm	Maximize the weighted-sum-rate of D2D network	Channel selection and power control	Orthogonal access
Ours	Generalized Double DQN (DDQN) algorithm	Maximize the system sum throughput while guaranteeing the cellular communication quality and achieve the fairness while allocating resources among D2D nodes	Select proper time slots for D2D communication	Non-orthogonal and orthogonal access

proposed scheme based on the conventional DDQN algorithm can easily achieve the maximum sum throughput objective. Furthermore, the consideration of resource allocation fairness among D2D nodes is incorporated into the design of the objective function. Particularly, the majority of the existing works mainly rely on the conventional DQN algorithm, which treats the Q-function as the objective function [19], [22], [24]. In contrast, this article separates the Q-function and the objective function by reformulating the objective function with a nonlinear combination of the Q-functions, which makes a contribution to generalizing the DQN-based framework such that different objectives can be incorporated and achieved.

A part of this article for the uplink scenario has been presented in our earlier work [30], but the fairness objective in [30] is different from that in this work and this article extends the proposed scheme to the downlink scenario. In addition, extensive simulation results and discussions are provided to better interpret the DRL-based spectrum access scheme for both the uplink and the downlink scenarios, respectively. The main contributions of this article are summarized as follows.

1) We investigate a hybrid spectrum access scheme for D2D communication, which takes into account both the nonorthogonal and orthogonal access manners and exploits the DRL technique in the framework design. In particular, there is a virtual controller acting as a centralized agent, which autonomously learns to select the proper time slots for D2D communication to achieve a predefined objective by trial-and-error interactions with the environment without any prior knowledge.
2) In this article, D2D nodes not only try to avoid causing interference to CUEs but also strive to achieve a specific objective such as maximizing the sum throughput of all users or maximizing a fairness objective function. Specifically, to investigate the optimization problem about resource allocation among D2D nodes, this article adopts a more general framework of the DDQN algorithm, which defines the Q-function for each user and reformulates the objective function with a nonlinear combination of Q-functions to achieve more general objectives.
3) This article analyzes the DRL-based spectrum access scheme for both the uplink and the downlink scenarios. The performance of the proposed scheme is evaluated under different scenarios, and simulation results show that the proposed scheme can achieve near-optimal performance with great convergence.

C. Paper Organization

The remainder of the article is organized as follows. We first introduce the system model for both the uplink and

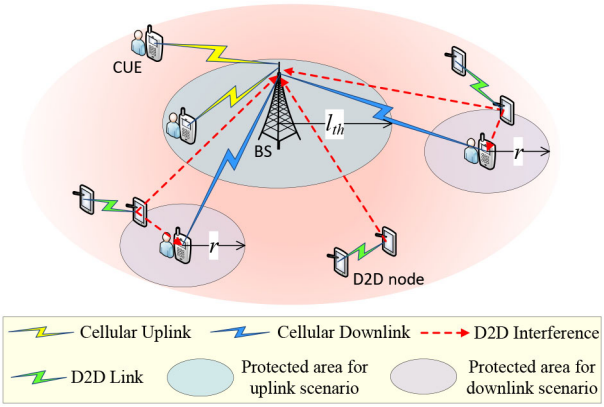


Fig. 1. System model for D2D-assisted cellular networks.

the downlink scenarios in Section II. Then, the framework design of DRL-based spectrum access scheme is presented in Section III, where the conventional DQN-based framework is extended to consider resource allocation fairness. After that, we evaluate the performance of the proposed scheme in both the uplink and the downlink scenarios, and simulation results are shown in Sections IV and V, respectively, followed by the conclusion in Section VI.

II. System Model

This section introduces the system model and spectrum access problem for both the uplink and the downlink scenarios. As shown in Fig. 1, we consider a D2D-assisted cellular network where multiple D2D nodes could share the cellular spectrum resources (CUEs) with CUEs in a time-splitting manner and the BS is located at the center of the cell. Specifically, we consider a time-slotted network where a period of time T is divided into Y time slots. X ($X \leq Y$) specific time slots within each period are assigned to CUEs in a time-division manner and the time slot assignment among CUEs is orthogonal. Meanwhile, D2D nodes are allowed to access the shared spectrum in the same time slot with CUEs under certain conditions, which will be discussed in detail in the following sections. Furthermore, we assume that the system is saturated, which means all the users always have data to transmit [31]. At the beginning of a time slot, a transmitter starts transmission and it must end transmission within this time slot, as commonly assumed in the literature. By observing the environment information, D2D nodes learn to select proper time slots for spectrum access to accomplish a specific objective such as the maximum sum throughput objective or the fairness objective while ensuring the quality of cellular communication. The main mathematical notations adopted in this section are listed in Table II.

TABLE II
Mathematical Notations

Notations	Physical interpretation
P_c, P_d, P_b	Transmit power of the CUE, D2D transmitter, and BS
P_n	AWGN power
G_c^B	Channel gain from the CUE to the BS in uplink scenario
G_d^B	Channel gain from the D2D transmitter to the BS
G_b^C	Channel gain from the BS to the CUE in downlink scenario
G_d^C	Channel gain from the D2D transmitter to the CUE
$\gamma^{[U]}$	SINR of cellular communication for uplink scenario
$\gamma^{[D]}$	SINR of cellular communication for downlink scenario
γ_{th}	The SINR threshold of cellular communication
l_1	The distance between the BS and the CUE
l_2	The distance between the D2D transmitter and the CUE
l_{th}	Radius of the "protected area" for uplink scenario
r	Radius of the "protected area" for downlink scenario
d	The D2D transmitter-receiver distance in meters
P_{LOS}	Probability of D2D nodes in a LOS connection
PL_{LOS}	Path loss when D2D nodes have a LOS connection

A. Uplink Spectrum Sharing

For the uplink scenario, considering the fact that D2D communication may cause severe interference to cellular links when they transmit in the same time slot, we expect that D2D nodes learn to decide whether to share a time slot with a CUE according to the distance between the CUE and the BS. We simplify the system model by allowing at most one D2D node to access the spectrum in each time slot. If a D2D node shares the same time slot with the CUE, the signal-to-interference-plus-noise ratio (SINR) of the cellular communication at the BS can be expressed as

$$\gamma^{[U]} = \frac{P_c \cdot G_c^B}{P_n + P_d \cdot G_d^B} \tag{1}$$

where P_c and P_d are the transmit powers of the CUE and the D2D transmitter (DT), respectively, P_n is the additive white Gaussian noise (AWGN) power, G_c^B is the power gain of the cellular link, and G_d^B is the power gain of the link between the DT and the BS. To guarantee the communication quality between the CUE and the BS, the SINR of the cellular communication is required to be higher than a specific threshold γ_{th}, i.e., $\gamma^{[U]} > \gamma_{\text{th}}$. As a result, the following inequality is satisfied:

$$G_c^B \geq \frac{\gamma_{\text{th}} \cdot \left(P_n + P_d \cdot G_d^B\right)}{P_c}. \tag{2}$$

Note that the above inequality does not explicitly consider the shadowing and multipath induced fading due to their stochastic nature, but the negative impact of fading on the communication quality can be mitigated by setting a higher SINR threshold. Considering the fact that $G_c^B \propto l_1^{-n}$ (n is the path-loss exponent with $n = 2$ in free space, and l_1 denotes the distance between the BS and the CUE), there exists a threshold l_{th} for the distance l_1 according to (2). In other words, for the uplink scenario, if the distance between the CUE and the BS is short enough (i.e., $l_1 \leq l_{\text{th}}$), D2D nodes can share the same time slots with the CUE via nonorthogonal access while ensuring the quality of the cellular communication. Otherwise (i.e., $l_1 > l_{\text{th}}$) D2D nodes can only exploit the time slots unoccupied by the CUE via orthogonal access to protect the communication of CUEs at the edge of the cell. For convenience, the coverage of the cell is simply regarded as a circular zone with a radius of R meters and a circular zone within l_{th} meters away from the BS is treated as a "protected area," as shown in Fig. 1. D2D nodes aim to automatically perceive the existence of the protected area through the interactions with the

environment. As will be illustrated later, this location-based access scheme makes a great difference when D2D nodes learn to select proper time slots for spectrum access according to the location information.

B. D2D Channel Model

Additionally, in the system, D2D links are modeled by a probabilistic path-loss model [32]. Specifically, the D2D link has a line-of-sight (LOS) or non-LOS (NLOS) connection with different probabilities based on the distance between the DT and the D2D receiver (DR). According to [32], the probability of having a LOS connection for a D2D node is given by

$$P_{\text{LOS}} = \begin{cases} 1, & d \leq 4 \\ \exp(-(d-4)/3), & 4 < d < 60 \\ 0, & d \geq 60 \end{cases} \tag{3}$$

where d is the DT-receiver distance in meters. Therefore, the probability of having a NLOS connection is $P_{\text{NLOS}} = 1 - P_{\text{LOS}}$. According to [32], the formulations of LOS and NLOS path loss can be described as

$$\text{PL}_{\text{LOS}} = 16.9\log_{10}(d[m]) + 32.8 + 20\log_{10}(f[\text{GHz}]) \tag{4}$$

$$\text{PL}_{\text{NLOS}} = 40\log_{10}(d[m]) + 79 + 30\log_{10}(f[\text{GHz}]) \tag{5}$$

respectively, where f represents the center frequency. Hence, the average path loss for the D2D link is given by $\text{PL}[\text{dB}] = P_{\text{LOS}}\text{PL}_{\text{LOS}} + P_{\text{NLOS}}\text{PL}_{\text{NLOS}}$ and the signal power received by the DR is written as $P_r[\text{mW}] = 10^{(P_d[\text{dBm}]-\text{PL}[\text{dB}])/10}$, where P_d is the transmit power of the DT. Then, the achievable transmission rate of the D2D link can be obtained by the Shannon equation $C = W\log_2(1 + P_r/P_n)$, where P_n is the noise power and W is the bandwidth of the channel. Note that this transmission rate will be used to design the reward function.

C. Downlink Spectrum Sharing

In this section, we introduce the spectrum access problem in the scenario considering the downlink communication between the BS and CUEs. The possible interference for CUEs comes from nearby D2D nodes that access the shared downlink spectrum in the same time slot, and thus, the SINR of the CUE can be written as

$$\gamma^{[D]} = \frac{P_b \cdot G_b^C}{P_n + P_d \cdot G_d^C} \tag{6}$$

where P_b is the transmit power of the BS, G_b^C is the power gain of the cellular link, and G_d^C is the power gain of the interference link between the DT and the CUE. With the requirement of satisfying the minimum SINR threshold γ_{th}, we obtain the following inequality:

$$G_d^C \leq \frac{P_b \cdot G_b^C - P_n \cdot \gamma_{\text{th}}}{P_d \cdot \gamma_{\text{th}}}. \tag{7}$$

Similar to the uplink scenario, let l_2 denote the distance between the DT and the CUE and there is a threshold r for the distance l_2 for the reason of $G_d^C \propto l_2^{-n}$. Specifically speaking, for the downlink scenario, CUEs could suffer from the less interference when D2D nodes are far away from CUEs, as a result of which, D2D nodes can utilize both resources allocated to CUEs and the rest unused resources via nonorthogonal access manner. Otherwise, D2D nodes can only exploit resources unoccupied by the CUE in an orthogonal access manner to guarantee the quality of cellular communication. On the basis of this fact, CUEs have a protected area and for convenience, we assume that a circular zone within r meters away from the CUE is regarded as the protected area, as shown in Fig. 1. Only if the D2D node is outside the protected area can it have an opportunity to share the same time slot with the CUE, which should be acquired autonomously by D2D nodes through interactions with the environment.

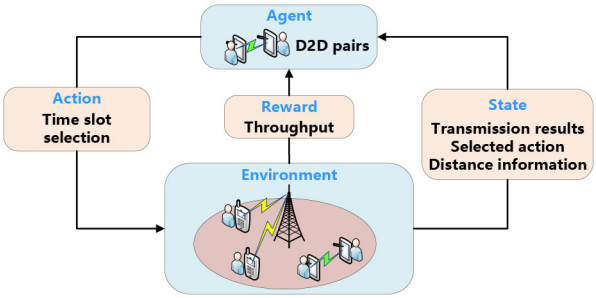


Fig. 2. Agent-environment interaction process in RL.

III. Deep Reinforcement Learning for Dynamic Spectrum Access

DRL is the underpinning technique in our proposed scheme, especially the DDQN algorithm. This section first presents an overview of RL and DDQN algorithm. Then, we introduce the framework design of the proposed scheme exploiting DDQN algorithm and further reformulate the proposed scheme by extending the objective function to achieve the resource allocation fairness among D2D nodes.

A. Underpinning Technique for DRL

The framework of RL includes the agent and the environment interacting with each other in discrete time steps, as illustrated in Fig. 2. At time t, the agent observes the state of the environment s_t from state space S and takes an action a_t from action space A according to a policy π. Then, the environment will feed back to the agent a reward r_{t+1} to evaluate the agent's performance and transits to the next state s_{t+1}. By iterative interactions, the agent will receive a series of rewards, which can be used to compute the cumulative discounted reward defined as

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \tag{8}$$

where $\gamma \in [0, 1)$ is a discount factor to indicate the importance of future rewards. The purpose of the RL is to find an optimal policy π^* to maximize some objective functions such as R_t. In Q-learning, a traditional RL algorithm, the objective function is an action-state value function named Q-function to indicate the long-term value of the agent after taking action a in the current state s. The Q-function for a state–action pair (s, a) under a certain policy π is described

as $Q^{\pi}(s,a) = \mathbb{E}_{\pi}[R_t \mid s_t = s, a_t = a]$. Additionally, there is a Q-value table in Q-learning, which stores Q-values for different state–action pairs for the purpose of finding an optimal policy π^*, i.e., $Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$. This can be described as follows according to the Bellman optimality equation:

$$Q^*(s,a) = \mathbb{E}\Big[r_{t+1} + \gamma \max_{a'} Q^*\big(s_{t+1}, a'\big) | s_t = s, a_t = a\Big]. \quad (9)$$

The agent collects the experience data in the form of $(s_t, a_t, r_{t+1}, s_{t+1})$ by interacting with the environment in discrete time steps, which will be used to update Q-value as follows:

$$Q(s_t,a_t) = Q(s_t,a_t) + \eta\Big(r_{t+1} + \gamma \max_{a} Q(s_{t+1},a) - Q(s_t,a_t)\Big) \quad (10)$$

where $\eta \in [0,1)$ is the learning rate. Furthermore, the agent adopts the ε-greedy policy while selecting an action according to Q-values. In particular, at time t, the agent selects action a_t among all allowed actions randomly with a probability ε and it selects action a_t, which maximizes $Q(s_t,a_t)$ with a probability $1-\varepsilon$.

In general, Q-learning can perform well when the state space and the action space are small scale. However, as the action space and/or the state space increases, the Q-value table maintained by Q-learning becomes large, thus causing difficulty for the algorithm convergence. As one of the most popular DRL algorithms, DQN combines Q-learning with deep learning which adopts a deep neural network (DNN) to fit Q-function rather than maintaining a Q-value table. Concretely speaking, the DNN obtains the environment state s_t as an input and outputs the predicted Q-values for all allowed actions denoted by $q(s_t, a_t; \boldsymbol{\theta}), a_t \in A$, where $\boldsymbol{\theta}$ represents the neural network parameters. The target of the DNN is to minimize the loss function defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\Big[\big(y_t - q(s_t,a_t;\boldsymbol{\theta})\big)^2\Big] \quad (11)$$

where y_t denotes the target output of the DNN given by

$$y_t = r_{t+1} + \gamma \max_{a} q(s_{t+1}, a, \boldsymbol{\theta}). \quad (12)$$

For training, we can train the DNN by iteratively updating its parameters to optimize the fitting of Q-function. The process for $\boldsymbol{\theta}$ iteration is given by the following:

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \eta \mathbb{E}\big[(y_t - q(s_t,a_t;\boldsymbol{\theta}))\nabla q(s_t,a_t;\boldsymbol{\theta})\big]. \quad (13)$$

Moreover, instead of training the DNN with a single experience data at each training step, the agent here collects experience data in the form of $(s_t, a_t, r_{t+1}, s_{t+1})$ and stores them into an experience replay pool. For training, a minibatch of experiences will be taken from the replay pool randomly to train the DNN. Also, the experience replay pool obeys a first-in–first-out (FIFO) manner. The DQN algorithm also adopts a "double neural networks" method to enhance its stability, where another target network with parameters $\boldsymbol{\theta'}$ is introduced in the training process. Thus, we replace the target output of the DNN y_t defined in (12) with

$$y_t = r_{t+1} + \gamma \max_{a} q\big(s_{t+1}, a; \boldsymbol{\theta'}\big). \quad (14)$$

Update $\boldsymbol{\theta'}$ with $\boldsymbol{\theta}$ every C training steps.

DQN is one of the most popular DRL algorithms, but there still exist some drawbacks such as overestimation. This will produce a higher Q-value estimation compared to the true value. In this article, we exploit the DDQN algorithm to formulate the proposed scheme, which separates the action selection from the estimation of the target output to alleviate overestimation. As a result, the target output of the DNN y_t given in (14) is reformulated as follows:

$$y_t = r_{t+1} + \gamma q\big(s_{t+1}, a; \boldsymbol{\theta'}\big) \quad (15)$$

where $a = \arg\max_a q(s_{t+1}, a; \boldsymbol{\theta})$.

B. Framework Design for Uplink Scenario

In the proposed scheme, we should treat each D2D node as an agent and thus, a multiagent optimization problem is presented. However, as agents select their own actions independently and simultaneously in the multiagent DRL framework, the actions of D2D nodes will conflict with each other when they make decisions independently. Hence, we set a central controller as a centralized agent for transmission coordination among all D2D nodes, which learns to select proper time slots for D2D communication to maximize the normalized sum throughput by interacting with the environment. As for the transmission of the coordination information, one of the possible solutions is to convey through a control channel, which can be implemented in a short time slot after each time slot of data transmission. Also, there are other possible implementations and we omit the discussion here since the implementation details are beyond the scope of this article. Below are listed the definitions of "action set," "state set," and "reward function" in detail.

Action Set: The agent takes an action a_t to decide whether to transmit and which one to transmit at time t. Here, we define the action space A as $\{0, 1, \ldots, K\}$, where K is the number of D2D nodes, $a_t = 0$ means that the agent selects no D2D node to transmit signals, and $a_t = i$ $(i = 1, 2, \ldots, K)$ means that the agent selects ith D2D node to transmit.

State Set: The agent expects to monitor the channel and obtain a channel observation to indicate transmission results or idleness of the channel, so we first define the channel state observed by the agent after taking action a_t as c_t. As discussed in Section II, the transmission of D2D nodes in the time slot occupied by CUE will result in severe interference if current time slot is not available for D2D nodes. So we define $c_t \in \{\mathbb{S}, \mathbb{I}, \mathbb{F}, \mathbb{R}\}$, where $\mathbb{S}$ means only one user occupies current time slot; $\mathbb{I}$ means the channel is idle; $\mathbb{F}$ means that the current time slot is reused by D2D nodes, which interfere the cellular communication severely; $\mathbb{R}$ means D2D nodes successfully share current time slot with a CUE. Besides, the agent observes the environment and maintains a distance vector $\boldsymbol{d} = [d_1, d_2, \ldots, d_L]$ to record the distance information between CUEs and the BS, in which L is the number of CUEs. Then, the environment observation at time $t+1$ is expressed as $\boldsymbol{o}_{t+1} = [a_t, c_t, \boldsymbol{d}_{t+1}]$. The past M observations of the environment constitute the environment state at time $t+1$, which is written as $s_{t+1} = [\boldsymbol{o}_{t-M+2}, \ldots, \boldsymbol{o}_t, \boldsymbol{o}_{t+1}]$, where M

denotes the length of historical observations. Intuitively, the agent will achieve better performance with a longer length of historical observations M. However, the longer the length of historical observations, the larger the state space, which may affect the convergence of the algorithm. As such, we have to set a proper M to strike a balance between the performance and the algorithm convergence.

Reward Function: For the case of the maximum sum throughput objective, the reward function is designed to be related to the sum throughput and it is determined by the channel state. Specifically, the reward of CUEs is simply set to 1 as long as CUEs communicate with the BS successfully. As for D2D nodes, taking the path loss of D2D links into account, we utilize a normalized transmission rate $\hat{C}^{(i)} = C^{(i)}/C_{\max}$ $(i = 1, 2, \ldots, K)$ to denote the reward of the ith D2D node, where $C^{(i)}$ is the transmission rate of the ith D2D node and $C_{\max}$ is the largest transmission rate when the D2D link is always in LOS connection. Then, the reward r_{t+1} from environment feedback after action a_t is written as

$$r_{t+1} = \begin{cases} 0, & c_t = \mathbb{I} \quad \text{or} \quad \mathbb{F} \\ 1, & c_t = \mathbb{S} \quad \text{and} \quad a_t = 0 \\ \hat{C}^{(i)}, & c_t = \mathbb{S} \quad \text{and} \quad a_t = i \\ 1 + \hat{C}^{(i)}, & c_t = \mathbb{R}. \end{cases} (i = 1, 2, \ldots, K) \tag{16}$$

C. Fairness Objective

For the original proposed scheme discussed above, the objective of the agent is only to maximize the sum throughput of all CUE and D2D users. Considering the fact that the Q-function is calculated according to a series of rewards received by the agent, we design the reward function as defined in (16) to denote the total transmission results of all users in each time slot, which ensures that the Q-function indicates the objective of maximum sum throughput. However, this framework with the maximum sum throughput objective does not facilitate the agent to properly assign time slots among D2D nodes, which will be discussed in Section IV-C in detail. For instance, if the current time slot is available for multiple D2D nodes, the agent will always choose the D2D node with the largest transmission rate to maximize the system sum throughput, thus leading to resource allocation unfairness among D2D nodes. As a remedy, we separate the Q-function and the objective function to generalize the algorithm and raise a new fairness objective function as

$$\lambda v^{(0)} + \sum_{i=1}^{K} \log\left(v^{(i)}\right) \quad (\lambda > 1) \tag{17}$$

where $v^{(0)}$ denotes the total throughput of all CUEs, $v^{(i)}$ $(i = 1, 2, \ldots, K)$ is the individual throughput of the ith D2D node, and λ indicates the importance of cellular communication, which is set to 5 in the following simulation experiments. Since the new objective function requires the individual throughput of each user while the Q-function in the original algorithm is related to the overall throughput, the original framework design cannot be directly applied to the new cases with fairness objective. Thus, we modify the original DDQN-based algorithm as follows.

In the original algorithm, the agent receives an overall reward from the environment to indicate the total transmission result of all users. As for the Q-function, regarded as the objective function, is computed based on the overall reward. In contrast, in the new algorithm, the reward from the environment feedback is a $K + 1$ dimension vector denoted by $[r^{(i)}]_{i=0}^{K}$, where $r^{(0)}$ denotes the total transmission results of all CUEs and $r^{(i)}$ $(i = 1, 2, \ldots, K)$ is the transmission result of the ith D2D node. Only if the ith $(i = 0, 1, \ldots, K)$ user successfully transmits data at time t can $r^{(i)}$ be set to 1 or $\hat{C}^{(i)}$, which is also expressed as follows in detail:

$$r_{t+1}^{(i)} = \begin{cases} 1, & (a_t = 0 \quad \text{and} \quad c_t = \mathbb{S}) \quad \text{or} \quad c_t = \mathbb{R} \\ & (i = 0) \\ \hat{C}^{(i)}, & a_t = i \quad \text{and} \quad (c_t = \mathbb{S} \quad \text{or} \quad c_t = \mathbb{R}) \\ & (i = 1, \ldots, K) \\ 0, & \text{others} \\ & (i = 0, \ldots, K). \end{cases} \tag{18}$$

Correspondingly, the agent maintains a $K + 1$ dimension vector defined as $[Q^{(i)}(s, a)]_{i=0}^{K}$ to denote the Q-function, where $Q^{(i)}(s, a)$ is the expected cumulative discounted reward of the ith user, and thus, it is related to the individual throughput of the ith user. Since it is hard to get the real throughput of each user, we exploit the element Q-value $Q^{(i)}(s, a)$ to substitute for $v^{(i)}$ in the new objective function. Therefore, the agent selects action a_t, which can maximize $\lambda Q^{(0)}(s_t, a_t) + \sum_{i=1}^{K} \log(Q^{(i)}(s_t, a_t))$ with a probability $1 - \varepsilon$ in the ε-greedy algorithm. Since the agent maintains $K + 1$ element Q values $Q^{(i)}(s, a)$ $(i = 0, 1, \ldots, K)$, the agent has to parallelly update them as follows:

$$\begin{aligned} Q^{(i)}(s_t, a_t) = {} & Q^{(i)}(s_t, a_t) \\ & + \eta\left(r_{t+1}^{(i)} + \gamma Q^{(i)}\left(s_{t+1}, a'\right) - Q^{(i)}(s_t, a_t)\right) \end{aligned} \tag{19}$$

where

$$a' = \arg\max_{a}\left(\lambda Q^{(0)}(s_{t+1}, a) + \sum_{i=1}^{K} \log\left(Q^{(i)}(s_{t+1}, a)\right)\right).$$

More importantly, compared with the traditional DQN-based algorithms mainly regarding the Q-function as the objective function, the algorithm here produces multiple Q functions to separate the Q-function and objective function, each of which is part of the objective function. In other words, the objective function is a nonlinear combination of multiple Q functions [33], which provides an approach to achieve more general objectives for DQN-based algorithms.

Considering the DNN, let $q^{(i)}(s, a; \boldsymbol{\theta})$ approximated by DNN denote the estimation of $Q^{(i)}(s, a)$, and thus, DNN has to output $K + 1$ dimension Q-values for all allowed actions with the environment state as the input. Besides, the agent collects and stores the experience data in the form of $(s_t, a_t, [r_{t+1}^{(i)}]_{i=0}^{K}, s_{t+1})$ rather than $(s_t, a_t, r_{t+1}, s_{t+1})$. Accordingly, we should replace the loss function and target

Algorithm 1 Proposed Scheme With Fairness Objective

Initialize the experience replay pool $\mathcal{D}$ with capacity D
Randomly initialize the parameters of current DNN $\boldsymbol{\theta}$
Randomly initialize the parameters of target DNN $\boldsymbol{\theta'}$
Initialize the state
for $t = 0, 1, ...$ **do**
Input s_t to DNN and output $Q = \{q^{(i)}(s_t, a; \boldsymbol{\theta}) | a \in A, i = 0, 1, ..., K\}$
Select action a_t based on the ϵ-greedy algorithm
Execute a_t and observe reward $\left[r_{t+1}^{(i)}\right]_{i=0}^{K}$ and next state s_{t+1}
Collect experience data $\left(s_t, a_t, \left[r_{t+1}^{(i)}\right]_{i=0}^{K}, s_{t+1}\right)$ to $\mathcal{D}$
Update $\boldsymbol{\theta'}$ with $\boldsymbol{\theta}$ in every C training steps
TRAIN DNN
end for
procedure TRAIN DNN
Sample N_E experience data randomly from $\mathcal{D}$ as $\mathcal{E}$
for each tuple $e = \left(s, a, \left[r^{(i)}\right]_{i=0}^{K}, s'\right)$ in $\mathcal{E}$ **do**

$$a' = \arg\max_a \lambda q^{(0)}(s', a; \boldsymbol{\theta}) + \sum_{i=1}^{K} \log(q^{(i)}(s', a; \boldsymbol{\theta}))$$

Compute $y_e^{(i)} = r + \gamma q^{(i)}\left(s', a'; \boldsymbol{\theta'}\right)$
end for
Update $\boldsymbol{\theta}$ through Gradient Descent:

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \frac{\eta}{N_E} \sum_{e \in E} \frac{1}{K+1} \sum_{i=0}^{K} (y_e^{(i)} - q^{(i)}(s, a; \boldsymbol{\theta})) \nabla q^{(i)}(s, a; \boldsymbol{\theta})$$

end procedure

output of the DNN defined in (11), (15) with the following:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{1}{K+1} \sum_{i=0}^{K} \left(y_t^{(i)} - q^{(i)}(s_t, a_t; \boldsymbol{\theta})\right)^2\right] \tag{20}$$

$$y_t^{(i)} = r_{t+1} + \gamma q^{(i)}\left(s_{t+1}, a; \boldsymbol{\theta'}\right) \tag{21}$$

where

$$a = \arg\max_a \left(\lambda q^{(0)}(s_{t+1}, a; \boldsymbol{\theta}) + \sum_{i=1}^{K} \log\left(q^{(i)}(s_{t+1}, a; \boldsymbol{\theta})\right)\right).$$

The update of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \eta \mathbb{E}\left[\frac{1}{K+1} \sum_{i=0}^{K} \left(y_t^{(i)} - q^{(i)}(s_t, a_t; \boldsymbol{\theta})\right) \nabla q^{(i)}(s_t, a_t; \boldsymbol{\theta})\right]. \tag{22}$$

The training process pseudocode for the reformulated scheme with fairness objective is shown in Algorithm 1.

D. DRL for Downlink Spectrum Access

As discussed in Section III-B, to avoid potential conflicts when D2D nodes make decisions independently, we set a central controller as a centralized agent for transmission coordination among all D2D nodes. For the downlink scenario, at time t, the agent observes the environment state s_t and takes an action $a_t \in \{0, 1, \ldots, K\}$, where K is the number of D2D nodes, $a_t = 0$ means no D2D node transmits signals, and $a_t = i$ $(i = 1, 2, \ldots, K)$ means the agent selects the ith D2D node to transmit. Similar to the uplink condition, the state of the environment s_t consists of the past M observations of the environment, which can be defined as $s_t = [\boldsymbol{o}_{t-M+1}, \ldots, \boldsymbol{o}_{t-1}, \boldsymbol{o}_t]$. The environment observation at time $t+1$ is written as $\boldsymbol{o}_{t+1} = [a_t, c_t, \boldsymbol{d}_{t+1}]$, where $c_t \in \{\mathbb{S}, \mathbb{I}, \mathbb{F}, \mathbb{R}\}$ is the channel state observed by the agent after taking action a_t. Also, $\mathbb{S}$ means only one user occupies current time slot; $\mathbb{I}$ means the channel is idle; $\mathbb{F}$ means that the current time slot is reused by D2D nodes, which interfere the cellular communication severely; $\mathbb{R}$ means that D2D nodes successfully share the current time slot with a CUE. $\boldsymbol{d}_{t+1} = [d_{11}, \ldots, d_{ij}, \ldots, d_{KL}]_{t+1}$ is a distance vector, in which d_{ij} indicates the distance between the ith D2D node and the jth CUE. As for the objective, we still consider two objectives: 1) the maximum sum throughput objective and 2) the fairness objective. For the former, the agent expects to acquire an optimal strategy for spectrum access to maximize the normalized sum throughput of all users, and the reward r_{t+1} from environment feedback after action a_t is the same as defined in (16). For the latter, the agent is expected to maximize the new objective function $\lambda v^{(0)} + \sum_{i=1}^{K} \log(v^{(i)})$ $(\lambda > 1)$ to achieve the fairness while allocating resources among D2D nodes, where $v^{(i)}$ $(i = 1, 2, \ldots, K)$ is the individual throughput of the ith D2D node, $v^{(0)}$ is the total throughput of all CUEs, and λ indicates the importance of cellular communication. Correspondingly, the reward function and Q-function have to be reformulated as discussed in Section III-C, which generalizes the framework of DQN-based algorithms so that a more general class of objectives can be achieved than can conventional algorithms.

IV. Performance Evaluation for Uplink Scenario

This section evaluates the performance of the proposed scheme for the uplink scenario with the maximum sum throughput objective and fairness objective. We first introduce the setup for simulation experiments and then analyze the simulation results in different cases.

A. Simulation Setup

We established the DNN for the agent based on Google TensorFlow and utilized CPU Intel i7-6700 to conduct the interactions between the agent and the environment. In the system, the radius R of the cell is set to 500 m. The transmit power P_d is set to 23 dBm and the noise power P_n is -114 dBm. The centre frequency f is 2 GHz and the bandwidth W is set to 20 MHz [32]. In the following, we introduce the hyper-parameters for simulations, the performance metric, and the benchmark used in this article.

Hyperparameters: As shown in Fig. 3, we construct the DNN with 6 convolutional neural networks as hidden layers and the output layer is a fully connected neural network. Each

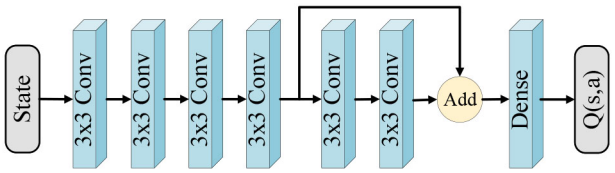


Fig. 3. Architecture of DNN with a shortcut.

TABLE III
PARAMETERS FOR SIMULATION

Parameters	Values
Radius of the cell R	500 m
Centre frequency f	2 GHz
Bandwidth W	20 MHz
D2D transmit power P_d	23 dBm
Noise power P_n	-114 dBm
Length of historical state M	10
Learning rate η	0.01
Discount factor γ	0.3
Initial exploration ϵ	1
Final exploration ϵ	0.005
Decay rate for ϵ	0.995
Mini-batch size	64
Update frequency for target network C	20
Adopted λ in the fairness objective	5

hidden layer adopts 3×3 convolution kernel and uses ReLU function as the activation function. Besides, referring to the residual units introduced in [34], we add one shortcut from the input of the fifth hidden layer to the output layer, which aims to reduce the deterioration of the algorithm performance as the depth of the neural network increases. Moreover, the Adam optimizer [35] is adopted to update the parameter $\boldsymbol{\theta}$ with the minibatch size set to 64. Other hyperparameters for simulation are summarized in Table III. Note that the discount factor γ is set to 0.3, which will be discussed in details in the following section. To select a proper value for the adopted λ in the fairness objective function, we compare the throughput results for different λ values when there are six CUEs coexisting with two D2D nodes. Simulation results are shown in Fig. 4 where the transmitter–receiver distance of two D2D nodes are different in cases 1, 2, and 3, from which we can see that the algorithm with $\lambda = 5$ can achieve the highest sum throughput.

Performance Metric: In this article, we assume that only one packet is transmitted in each time slot for cellular communication and the throughput for cellular communication is defined as the number of packets successfully transmitted per time slot, while the throughput for D2D communication is defined as the normalized transmission rate. Here, we adopt the normalized throughput as the performance metric and the average of sum throughput at time t is derived from the reward values of the past N time slots, which is written as $\sum_{\tau=t-N+1}^{t} r_\tau/N$ or $\sum_{\tau=t-N+1}^{t}\sum_{i=0}^{K} r_\tau^{(i)}/N$ in the case of sum throughput objective or fairness objective, respectively. In the following simulations, N is set to 1000. If one time slot is 1 ms in duration, then the throughput is the total number of packets over the past second.

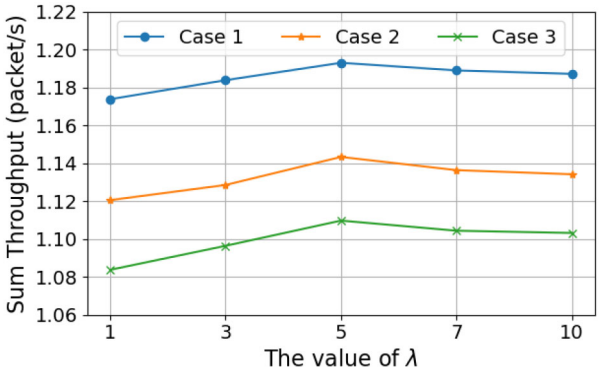


Fig. 4. Sum throughput results with different λ values. The transmitter-receiver distance of two D2D nodes are (4 m, 15 m), (4 m, 30 m), and (4 m, 45 m) in cases 1, 2, and 3, respectively.

Benchmark: In the proposed scheme, the D2D nodes have no idea of any prior knowledge, such as access principles of the CUEs and the location related information. In contrast, for the benchmark, we consider a genius-aided scheme where the D2D nodes are considered to be model-aware nodes and perform an optimal spectrum access policy derived from a prior knowledge. Specifically, model-aware D2D nodes are aware of how to select proper time slots for spectrum access according to the access principle of CUEs and the location information. As a result, the genius-aided scheme can reach the theoretical upper bound of the throughput.

B. Discount Factor

This section explores the influence of discount factor γ on the performance of the proposed scheme. We set γ to 0.1, 0.3, 0.7, and 0.9, and the sum throughput over the training period is shown in Fig. 5(a), from which we can see that the algorithms with lower $\gamma \in \{0.1, 0.3\}$ enjoy a better convergence. According to the Bellman equation given in (9), as the discount factor γ increases, the variance of the Q-value will become larger. In RL, variance usually refers to a noisy but on average accurate value estimate, while the additional noise on data can slow down the convergence speed, and can deteriorate the performance of trained DNN [36]. Then, we evaluate the proposed scheme with different γ values in the cases where multiple D2D nodes have different transmitter-receiver distance and one of them changes its transmitter–receiver distance. The results are presented in Fig. 5(b), from which we can see that the algorithm with $\gamma = 0.3$ can achieve the highest sum throughput while the lowest value is obtained by the one with the highest γ value. All the simulation results show that a high γ value has a negative influence on the performance of the proposed scheme, which illustrates that the future rewards make little contributions to the action selection for the agent at current time. As a result, we recommend a low discount factor here and set γ to 0.3 in the sequel simulations.

C. Throughput Analysis

As for the uplink scenario, the optimal policy for D2D nodes is to exploit all the idle time slots within a period and share the time slot with CUEs when the CUE is physically

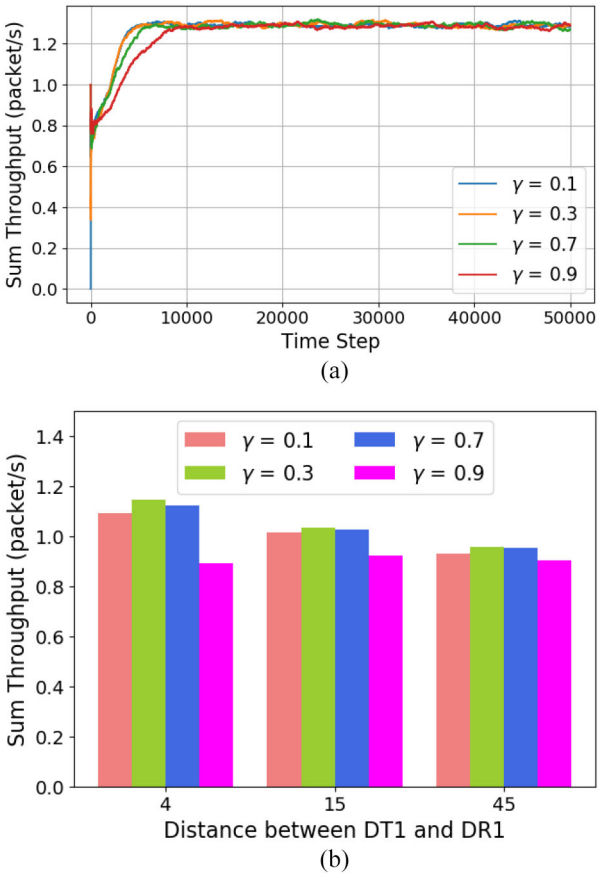


Fig. 5. Algorithm performance with different discount factor γ. (a) Recorded throughput during the training period with different γ values. (b) Throughput results versus DT-receiver distance with different γ values.

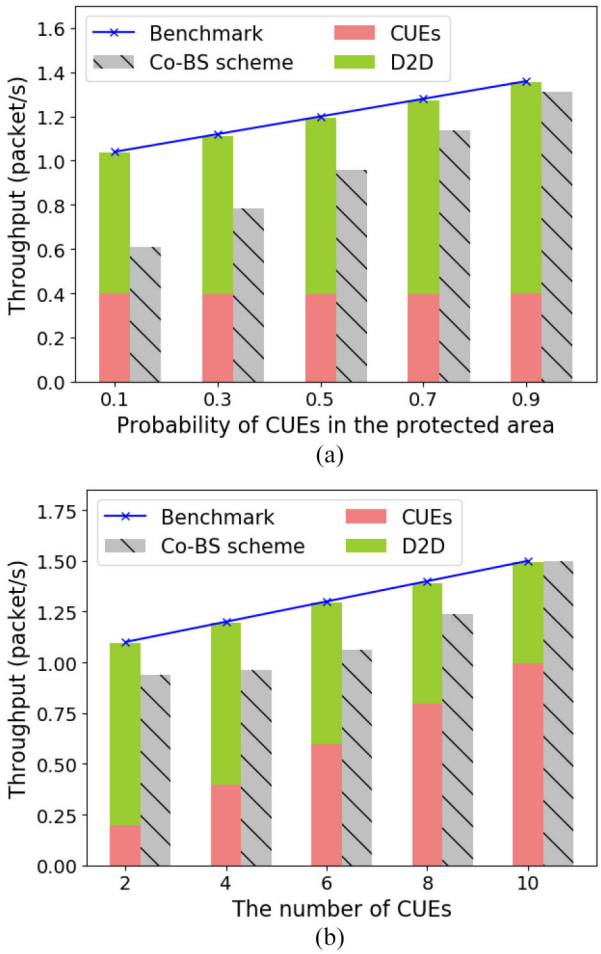


Fig. 6. Sum and individual throughput for different cases in the uplink scenario. (a) Throughput versus the probability of CUEs in the protected area. (b) Throughput versus the number of CUEs.

close enough to the BS. For convenience of analysis, let x represent the proportion of time slots preassigned to the CUEs and let p represent the probability of CUEs in the protected area. In the considered system, CUEs are randomly deployed within the coverage of the cell, so p is calculated by $([\pi * l_{\text{th}}^2]/[\pi * R^2])$. For the objective of maximizing the sum throughput, the benchmark can be calculated directly from the formula $x+(1-x+\text{xp})C_m$, where C_m denotes the throughput of D2D node with the highest normalized transmission rate. As for the fairness objective, it is hard to derive the benchmark value directly through the formula. So, we assume that D2D nodes are model-aware nodes when we perform the simulation experiments to obtain the benchmark, just as discussed in Section IV-A, which is actually impossible in reality.

In the system, each CUE is preassigned 10% of time slots within a period, and the DT-receiver distance is in the range of 4–60 m, which meets the requirements in (3). We first consider the scenario where multiple D2D nodes have the same transmitter–receiver distance. Given the objective of maximum sum throughput, we evaluate the proposed scheme in the cases of different probabilities p and different numbers of CUEs. Under this setting, we compare our scheme with a scheme named co-BS scheme where D2D nodes lack learning ability and can only get partial information of the environment relying on the cooperation of the BS. Specifically, Fig. 6(a) presents the throughput results with p varying from 0.1 to 0.9 in the case of four CUEs, and Fig. 6(b) demonstrates the throughput results versus the number of CUEs with a fixed $p = 0.5$. As we can see from Fig. 6, the throughput results achieved by the proposed scheme are very close to the benchmark, which is the theoretical upper bound of throughput. Also, the performance of the proposed scheme considerably outperforms the co-BS scheme, which proves that the agent can learn an optimal strategy for spectrum access autonomously without any prior knowledge.

Furthermore, to analyze the resource allocation among D2D nodes sharing the same transmitter–receiver distance, we plot the sum and individual throughput results over the training period with the maximum sum throughput objective in the case of the coexistence of six CUEs with two D2D nodes. The results are shown in Fig. 7(a), which prove that the system can arrive at the maximum of sum throughput and guarantee the throughput of CUEs. However, when we focus on the individual throughput of each D2D node, we find that there is a wide range of fluctuation. This means the agent has no idea of which one is the optimal choice to maximize the normalized sum throughput if all D2D nodes have the same transmission rate. This phenomenon motivates us to investigate the fairness

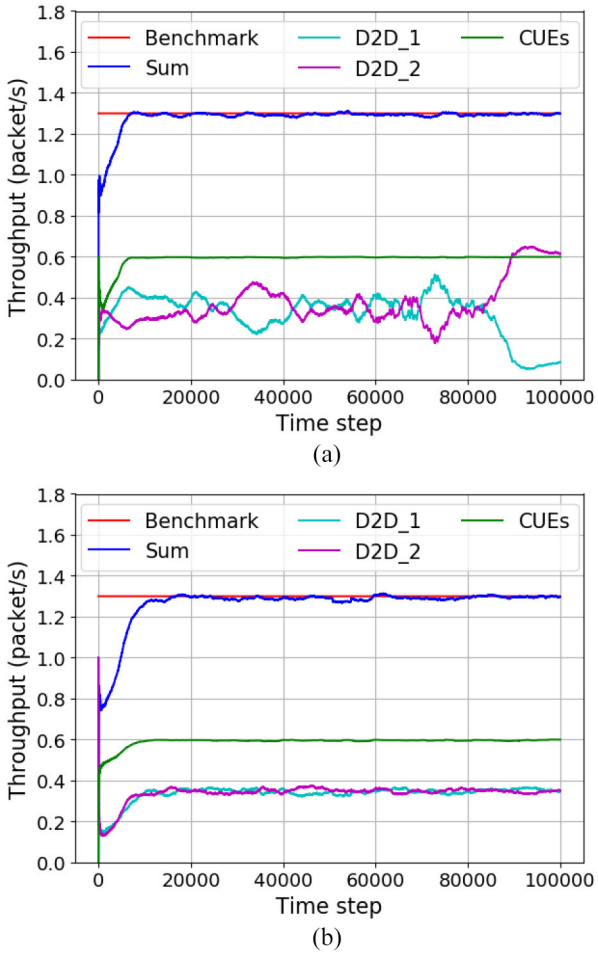


Fig. 7. Throughput results over the training period with different objectives under the uplink scenario. (a) Sum and individual throughput over the training period with the maximum sum throughput objective. (b) Sum and individual throughput over the training period with fairness objective.

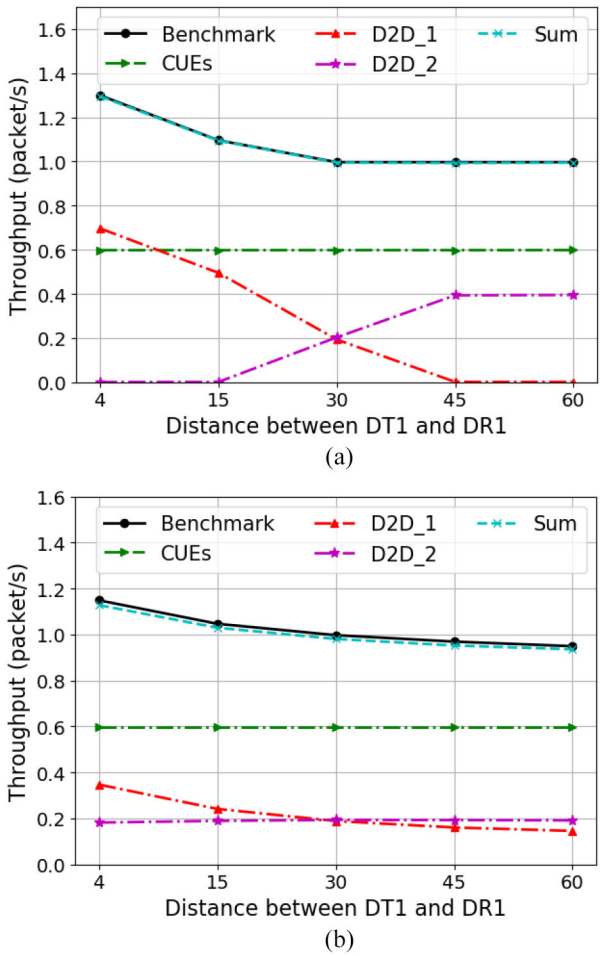


Fig. 8. Throughput results for different DT-receiver distances under the uplink scenario. (a) Throughput versus DT-receiver distance with maximum sum throughput objective. (b) Throughput versus DT-receiver distance with fairness objective.

objective as discussed in Section III-C. Specifically, under the same settings with that in Fig. 7(a), we evaluate the proposed scheme with the fairness objective. Also, Fig. 7(b) exhibits the throughput results, which proves that not only can the system sum throughput arrive at the benchmark level, but the individual throughput of each D2D node is equalized for fairness provisioning.

Then, we evaluate the proposed scheme for the cases of different DT-receiver distances. For convenience, the number of CUEs is 6 and the variable p is fixed to 0.5. In the system of including two D2D nodes (i.e., D2D_1 and D2D_2), we change the transmitter-receiver distance of D2D_1 from 4 to 60 m and the transmitter–receiver distance of D2D_2 is fixed to 30 m. The throughput results are presented in Fig. 8(a), from which we can see that the system sum throughput can still arrive at the maximum and the cellular communication throughput is guaranteed, but the resources are mostly allocated to the D2D node with the shortest transmitter–receiver distance. The reason for this phenomenon is that as the transmitter–receiver distance increases, the probability of D2D_1 in LOS connection decreases and thus, the transmission rate of D2D_1 decreases. With the maximum sum throughput objective, the agent would always select the D2D node with the largest transmission rate to transmit, thus leading to the resource allocation unfairness among D2D nodes. Then, as analyzed in Section III-C, we incorporate the consideration of fairness while designing the objective function. Under the same settings as that in Fig. 8(a), we evaluate the proposed scheme with the fairness objective and the results are exhibited in Fig. 8(b). Meanwhile, the benchmark here is derived from the genius-aided scheme with fairness objective. As we can see from Fig. 8(b), the time slots are evenly allocated to D2D nodes and the throughput of different D2D nodes is basically the same, which means that the agent can learn the optimal spectrum access scheme to guarantee the fairness while allocating resources among D2D nodes.

D. Number of States

This section further presents the results to analyze how the number of states visited by the agent during the training period vary with the radius of the protected area. There are six CUEs coexisting with two D2D nodes. As shown in Fig. 9, the vertical axis is the total number of distinct states visited by the agent after 50 000 time steps and the horizontal axis is the radius of the protected area in the uplink scenario denoted by

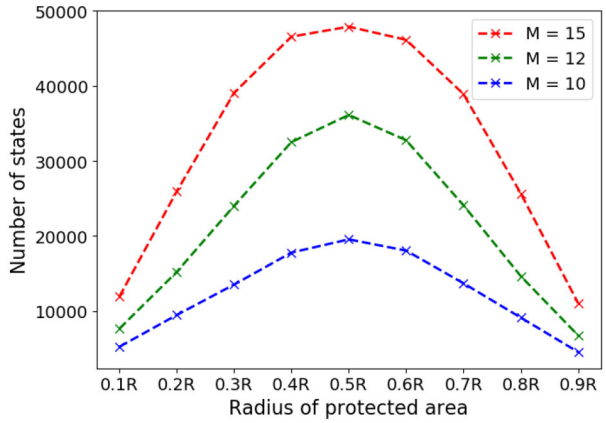


Fig. 9. Total number of distinct states visited by the agent.

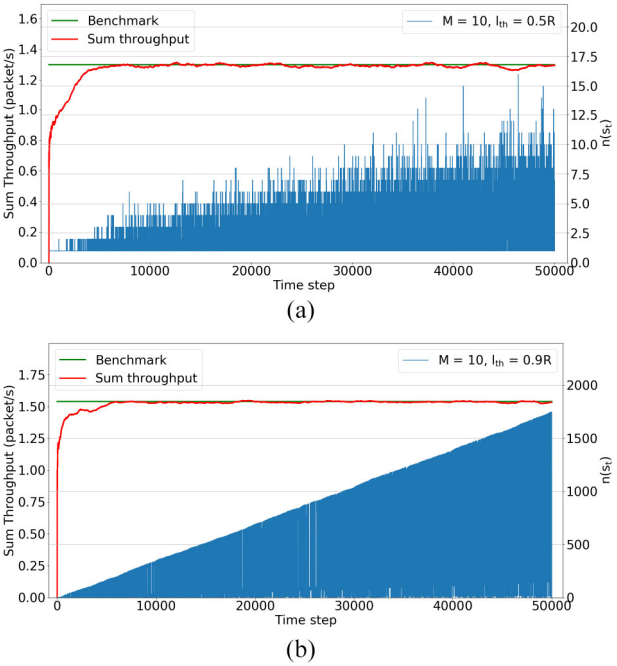


Fig. 10. Sum throughput and $n(s_t)$ during the training period. (a) $M = 10$ and $l_{\text{th}} = 0.5R$. (b) $M = 10$ and $l_{\text{th}} = 0.9R$.

l_{th}, where R denotes the radius of the cell. As the length of historical state M increases, the number of states in state space will increase and thus, the total number of distinct states visited by the agent will increase, which is also demonstrated in Fig. 9. Moreover, in Fig. 9, we find that the number of distinct states visited by the agent increases first and then decreases as the radius of the protected area increases.

To further dig into this phenomenon, we define $n(s_t)$ to denote the number of previous visits to state s_t before time step t, and draw the curve of $n(s_t)$ during the training period in Fig. 10(a) and (b) where the radius of the protected area l_{th} is $0.5R$ and $0.9R$, respectively. Comparing Fig. 10(a) with (b), we find that the agent experiences the same state with a small probability when l_{th} is $0.5R$, while the same state visited by the agent appears many times when l_{th} is $0.9R$. This phenomenon may be caused by the fact that cellular users are more likely to locate in the protected area with l_{th} approaching the radius of the cell R, and thus, the states including the

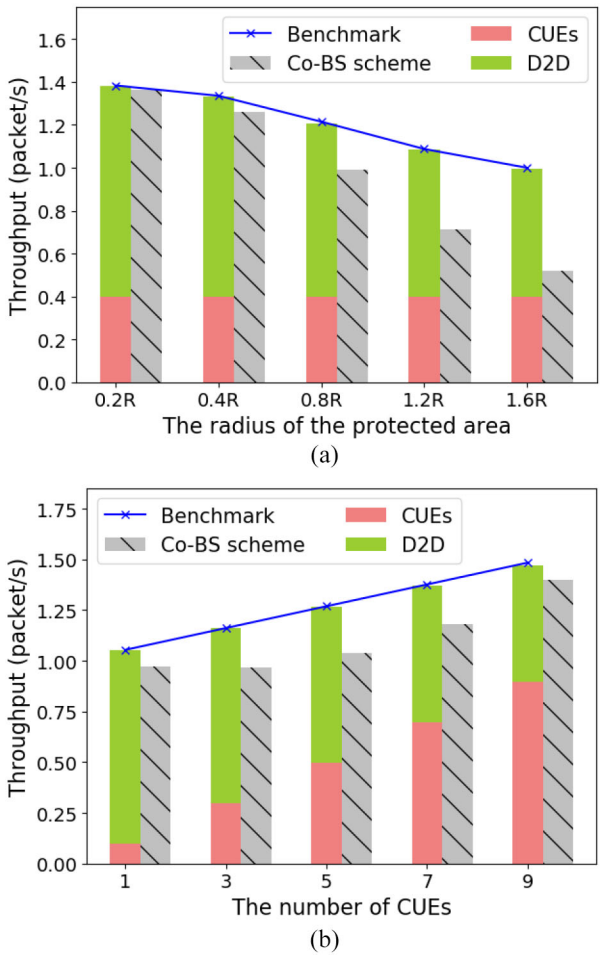


Fig. 11. Sum and individual throughput for different cases in the downlink scenario. (a) Throughput versus the radius of the protected area. (b) Throughput versus the number of CUEs.

channel state $\mathbb{R}$ will appear frequently, which is similar to the scenario when l_{th} approaches zero. Nevertheless, the results in Fig. 10 show that the algorithm still can achieve the optimal throughput results with good convergence, which means that the proposed scheme can be applied to scenarios with different probability distributions of state space and proves the robustness of the proposed scheme.

V. Performance Evaluation for Downlink Scenario

A. Throughput Analysis

The simulation setup are the same as that in Section IV-A. As for the benchmark in the downlink scenario, similar to the uplink scenario, we still consider the D2D nodes to be model-aware users when we perform simulation experiments, where the agent performs an optimal spectrum access policy derived from fully known prior knowledge. For the convenience of analysis, let r denote the radius of the protected area in the downlink scenario and each CUE is preassigned ten percent of the time slots within a period. Similar to the uplink scenario, we evaluate the proposed scheme against with different radii r and different numbers of CUEs. Specifically, Fig. 11(a) demonstrates the throughput results when r is set

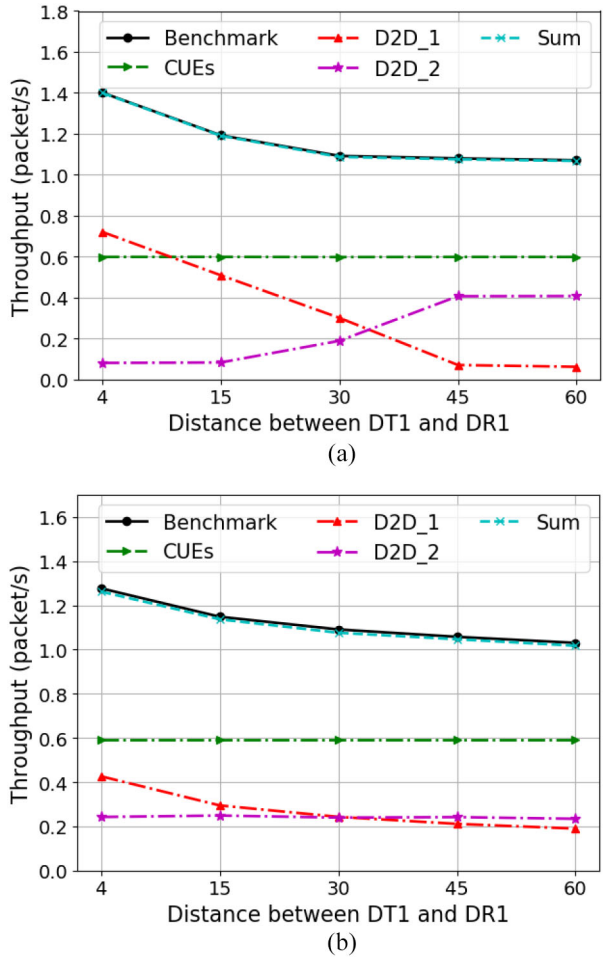


Fig. 12. Throughput results for different DT-receiver distances under the downlink scenario. (a) Throughput versus DT-receiver distance with maximum sum throughput objective. (b) Throughput versus DT-receiver distance with fairness objective.

to $0.2R$, $0.4R$, $0.8R$, $1.2R$, and $1.6R$ in the case of four CUEs, and Fig. 11(b) shows the sum and individual throughput versus the number of CUEs with a fixed $r = 0.8R$. As we can see from Fig. 11, the proposed scheme can achieve the optimal performance in terms of throughput compared to the benchmark and the performance considerably outperforms the co-BS scheme, which means that the agent can acquire an optimal strategy for downlink spectrum access autonomously without any prior knowledge.

We then evaluate the proposed scheme for the cases with different DT-receiver distances. For convenience, the number of CUEs is 6 and variable r is fixed to $0.8R$. In the system of including two D2D nodes (D2D_1 and D2D_2), the transmitter-receiver distance of D2D_1 is set to 4, 15, 30, 45, and 60 m, and the transmitter–receiver distance of D2D_2 is fixed to 30 m. Fig. 12(a) presents the throughput results with the maximum sum throughput objective, which proves that the system can still reach the maximum of sum throughput and guarantee the cellular communication throughput. However, the throughput of the D2D node with the shortest transmitter–receiver distance is much greater than that of another D2D node, which is similar to the analysis in the uplink case. As a result, as discussed in Section III-C, the fairness is incorporated into the design of the objective function. Under the same settings as that in Fig. 12(a), we evaluate the performance of the proposed scheme with fairness objective versus different DT-receiver distances, and the results are presented in Fig. 12(b). At the same time, the benchmark here is derived from the genius-aided scheme with fairness objective. As we can see from Fig. 12(b), the system throughput can arrive at the benchmark level and the individual throughput of different D2D nodes is basically the same, which means the agent can learn the optimal spectrum access scheme to guarantee the fairness while allocating resources among D2D nodes.

B. Discussion

In conclusion, for both the uplink ad downlink scenarios, the DRL-based scheme can easily figure out the association between the environment observations and the optimal policy for maximizing the cumulative discounted reward through a trial-and-error manner, while traditional approaches require fully known prior information of the environment to characterize this association. However, the prior information is usually hard to obtain. Furthermore, the computation complexity is a major concern for the deployment of deep learning algorithms, which is mainly related to the complexity of the neural network. Considering the fact that the DNN used here is small scale and GPU cannot be fully utilized with a small-scale DNN, we utilize CPU Intel i7-6700 as the simulation platform and the average time cost for each execution is 4.12×10^{-4} s, which is acceptable for the delay requirements of D2D communication.

VI. Conclusion

Dynamic spectrum access is a crucial issue for D2D communications. This article investigated this problem based on DRL technology. Specifically, we considered a time-slotted cellular network where CUEs and D2D nodes try to access a shared cellular spectrum channel in a time-splitting manner while the access principle of CUEs is not known to D2D nodes. Besides, D2D nodes are allowed to share the same time slots with CUEs under certain circumstances according to an LSA strategy, which is also not known to D2D nodes in advance. Then, based on DRL, a hybrid spectrum access scheme was proposed, which enables D2D nodes to learn an optimal strategy for spectrum access autonomously to maximize the normalized sum throughput of all users while prior knowledge is not available.

Furthermore, we found that the resource allocation among D2D nodes is unfair and then we reformulated the proposed scheme by considering the fairness while designing the objective function. In particular, we defined the Q-function for each user and then replaced the objective function with a nonlinear combination of Q functions to achieve the resource allocation fairness for D2D nodes. In addition, we presented the framework design of the proposed scheme for both the uplink and the downlink scenarios. The performance of the proposed scheme was evaluated under several conditions, and simulation results showed that it can achieve the optimal

performance with different objectives compared to the benchmark, which is derived from a genius-aided scheme with fully known prior knowledge. For future work, we plan to investigate the spectrum access problems in the scenario where CUEs adopt diverse access principles and D2D nodes aim to learn more advanced spectrum access strategies with carrier sensing. Moreover, considering the priority of data transmission among D2D nodes, heterogeneous networks with the coexistence of saturated and unsaturated users are also important for further investigation.

REFERENCES

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[2] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[3] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 764–779, Jan. 2020.

[4] S. Yu, X. Chen, Z. Zhou, X. Gong, and D. Wu, "When deep reinforcement learning meets federated learning: Intelligent multitimescale resource management for multiaccess edge computing in 5G ultra-dense network," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2238–2251, Feb. 2021.

[5] S. Yu, X. Chen, L. Yang, D. Wu, M. Bennis, and J. Zhang, "Intelligent edge: Leveraging deep imitation learning for mobile edge computation offloading," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 92–99, Feb. 2020.

[6] D. C. Nguyen *et al.*, "6G Internet of Things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, Jan. 2022.

[7] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, 4th Quart., 2014.

[8] Y. Li, Y. Liang, Q. Liu, and H. Wang, "Resources allocation in multicell D2D communications for Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 4100–4108, Oct. 2018.

[9] W. Wang, L. Yang, A. Meng, Y. Zhan, and D. W. K. Ng, "Resource allocation for IRS-aided JP-CoMP downlink cellular networks with underlaying D2D communications," *IEEE Trans. Wireless Commun.*, early access, Dec. 1, 2021, doi: 10.1109/TWC.2021.3128711.

[10] W. Lu *et al.*, "SWIPT cooperative spectrum sharing for 6G-enabled cognitive IoT network," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15070–15080, Oct. 2021.

[11] W. Lee, M. Kim, and D.-H. Cho, "Transmit power control using deep neural network for underlay device-to-device communication," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 141–144, Feb. 2019.

[12] Y. Kai, J. Wang, H. Zhu, and J. Wang, "Resource allocation and performance analysis of cellular-assisted OFDMA device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 416–431, Jan. 2019.

[13] H. Tamura, M. Sengoku, K. Nakano, and S. Shinoda, "Graph theoretic or computational geometric research of cellular mobile communications," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 1999, pp. 153–156.

[14] A. Checco and D. J. Leith, "Learning-based constraint satisfaction with sensing restrictions," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 811–820, Oct. 2013.

[15] L. Wang, H. Tang, H. Wu, and G. L. Stüber, "Resource allocation for D2D communications underlay in Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1159–1170, Feb. 2017.

[16] Q. Deng *et al.*, "Dynamic spectrum sharing for hybrid access in OFDMA-based cognitive femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10830–10840, Nov. 2018.

[17] F. W. Zaki, S. Kishk, and N. H. Almofari, "Distributed resource allocation for D2D communication networks using auction," in *Proc. IEEE Nat. Radio Sci. Conf. (NRSC)*, Mar. 2017, pp. 284–293.

[18] H.-H. Nguyen, M. Hasegawa, and W.-J. Hwang, "Distributed resource allocation for D2D communications underlay cellular networks," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 942–945, May 2016.

[19] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[20] H.-H. Chang, H. Song, Y. Yi, J. Zhang, H. He, and L. Liu, "Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1938–1948, Apr. 2019.

[21] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[22] K. Zia, N. Javed, M. N. Sial, S. Ahmed, A. A. Pirzada, and F. Pervez, "A distributed multi-agent RL-based autonomous spectrum allocation scheme in D2D enabled multi-tier HetNets," *IEEE Access*, vol. 7, pp. 6733–6745, 2019.

[23] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1828–1840, Feb. 2020.

[24] D. Wang, H. Qin, B. Song, K. Xu, X. Du, and M. Guizani, "Joint resource allocation and power control for D2D communication with deep reinforcement learning in MCC," *Phys. Commun.*, vol. 45, pp. 1–9, Apr. 2021.

[25] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for vehicular communications with low latency and high reliability," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3887–3902, Aug. 2019.

[26] H. Ye, G. Y. Li, and B. H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[27] T. Zhang, K. Zhu, and J. Wang, "Energy-efficient mode selection and resource allocation for D2D-enabled heterogeneous networks: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1175–1187, Feb. 2021.

[28] A. Moussaid, W. Jaafar, W. Ajib, and H. Elbiaze, "Deep reinforcement learning-based data transmission for D2D communications," in *Proc. Int. Conf. Wireless Mobile Comput. Netw. Commun. (WiMob)*, Oct. 2018, pp. 1–7.

[29] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.

[30] J. Huang, Y. Yang, G. He, Y. Xiao, and J. Liu, "Deep reinforcement learning-based dynamic spectrum access for D2D communication underlay cellular networks," *IEEE Commun. Lett.*, vol. 25, no. 8, pp. 2614–2618, Aug. 2021.

[31] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1277–1290, Jun. 2019.

[32] "Study on LTE device to device proximity services; radio aspects, release 12," 3GPP, Sophia Antipolis, France, Rep. TR 36.843, Mar. 2014.

[33] C. Liu, X. Xu, and D. Hu, "Multiobjective reinforcement learning: A comprehensive overview," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 3, pp. 385–398, Mar. 2015.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.

[36] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

Jingfei Huang received the B.Eng. degree in communication engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2021, where she is currently pursuing the master's degree.

Her research interests include wireless resource allocation, machine learning, and AI-based wireless networks.

Yang Yang (Member, IEEE) received the master's and Ph.D. degrees from the School of Telecommunications Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2012 and 2015, respectively.

From 2015 to 2017, he was a Postdoctoral Researcher with Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing. He is currently an Associate Professor with the School of Artificial Intelligence, BUPT. His research interests include AI-based wireless networks, machine learning, and ultradense networks.

Dr. Yang was a recipient of the Exemplary Reviewer of IEEE TRANSACTIONS ON COMMUNICATIONS in 2018.

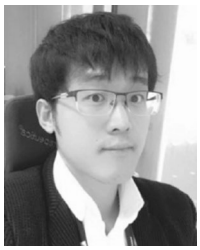
Zhen Gao (Member, IEEE) received the B.S. degree in information engineering from Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in communication and signal processing from Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, in 2016.

He is currently an Assistant Professor with Beijing Institute of Technology, Beijing, and also works as a Researcher with Southeast University, Nanjing, China. His research interests are in wireless communications, with a focus on multicarrier modulations, multiple antenna systems, and sparse signal processing.

Dr. Gao was the recipient of the IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (best paper), the Exemplary Reviewer of IEEE COMMUNICATIONS LETTERS in 2016, the *IET Electronics Letters* Premium Award (Best Paper) in 2016, and the Young Elite Scientists Sponsorship Program from 2018 to 2021 by China Association for Science and Technology.

Dazhong He (Member, IEEE) received the B.E. and M.E. degrees in information engineering and the Ph.D. degree in information and communication engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1994, 1997, and 2020, respectively.

He is currently a Senior Engineer with the School of Artificial Intelligence, BUPT. His current area of interest is in network traffic research, big data analysis, and hardware-based deep learning.



Derrick Wing Kwan Ng (Fellow, IEEE) received the bachelor's (First Class Hons.) and M.Phil. degrees in electronic engineering from Hong Kong University of Science and Technology, Hong Kong, in 2006 and 2008, respectively, and the Ph.D. degree from the University of British Columbia, Vancouver, BC, Canada, in November 2012.

He was a Senior Postdoctoral Fellow with the Institute for Digital Communications, Friedrich–Alexander-University Erlangen–Nürnberg (FAU), Erlangen, Germany. He is currently working as a Scientia Associate Professor with the University of New South Wales, Sydney, NSW, Australia. His research interests include convex and non-convex optimization, physical layer security, IRS-assisted communication, UAV-assisted communication, wireless information and power transfer, and green (energy-efficient) wireless communications.

Dr. Ng received the Australian Research Council Discovery Early Career Researcher Award 2017, the Best Paper Awards at the WCSP 2020 and 2021, the IEEE TCGCC Best Journal Paper Award 2018, INISCOM 2018, IEEE International Conference on Communications 2018 and 2021, IEEE International Conference on Computing, Networking and Communications 2016, IEEE Wireless Communications and Networking Conference 2012, the IEEE Global Telecommunication Conference 2011, and the IEEE Third International Conference on Communications and Networking in China 2008. He has been listed as a Highly Cited Researcher by Clarivate Analytics since 2018. He has been serving as an Editorial Assistant to the Editor-in-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS from January 2012 to December 2019. He is currently serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and an Area Editor for the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.