IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 November 2023, accepted 8 December 2023, date of publication 12 December 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3341585

RESEARCH ARTICLE

DRL-Based Resource Allocation for NOMA-Enabled D2D Communications Underlay Cellular Networks

YUN JAE JEONG, SEOYOUNG YU, AND JEONG WOO LEE, (Member, IEEE)
School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Jeong Woo Lee (jwlee2@cau.ac.kr)

This work was supported in part by the Chung-Ang University Graduate Research Scholarship, in 2021; and in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2023-RS-2022-00156353.

ABSTRACT Since the emergence of device-to-device (D2D) communications, an efficient resource allocation (RA) scheme with low-complexity suited for high variability of network environments has been continuously demanded. As a solution, we propose a RA scheme based on deep reinforcement learning (DRL) for D2D communications exploiting cluster-wise non-orthogonal multiple access (NOMA) protocol underlay cellular networks. The goal of RA is allocating transmit power and channel spectrum to D2D links to maximize a benefit. We analyze and formulate the outage of NOMA-enabled D2D links and investigate performance measures. To alleviate system overhead and computational complexity with maintaining high benefit, we propose a sub-optimal RA scheme under a centralized multi-agent DRL framework. Each agent corresponding to each D2D cluster trains its own artificial neural networks in a cyclic manner with a timing-offset. The proposed DRL-based RA scheme enables prompt allocation of resources to D2D links based on the observation of time-varying environments. The proposed RA scheme outperforms other schemes in terms of benefit, energy efficiency, fairness and coordination of D2D users, where the performance gain becomes significant when the mutual interference among user equipments is severe. In a cell of radius 100-meter with target rates for D2D and cellular links of 2 and 8 bits/s/Hz, respectively, the proposed RA scheme improves normalized benefit, energy efficiency, fairness and coordination of D2D users by 18%, 23%, 75% and 80%, respectively, over a greedy scheme. The improvements in these performance measures over a random RA scheme are 152%, 164%, 87% and 77%, respectively.

INDEX TERMS Device-to-device communications, cellular network, deep reinforcement learning, resource allocation, non-orthogonal multiple access.

I. INTRODUCTION

As the demand for mobile and wireless communication services grows at a rapid pace, the overload on network and the shortage of communication resources become inevitable and they are regarded as critical issues for designing and operating wireless communication systems [1]. As a result, there exist increasing needs for developing communication technologies providing a large number of users with high quality of services (QoS) without imposing heavy load on networks and requiring extra spectrum resources. Device-to-device (D2D) communication was proposed as one of promising solutions to prevent network overload problem and to alleviate spectrum shortage phenomenon when serving an increasing number of mobile users in cellular networks [2], [3]. In a D2D communication protocol, mobile users in proximity can communicate with each other directly over the channels which are assigned to cellular users in the

The associate editor coordinating the review of this manuscript and approving it for publication was Hang Shen.

network [4]. D2D communication, especially, underlay cellular networks, is considered one of key technologies in the fifth-generation (5G) wireless communications [5], and its scope has been extended to vehicle-to-vehicle (V2V) and vehicle-to-everything (V2X) systems [6].

Since D2D links share spectrum resources with cellular links, and sometimes with other D2D links as well, there exist inherent mutual interferences among user equipments (UEs) in the cell, which degrades the overall performance of network. Thus, communication resources for D2D communications underlay cellular networks need to be allocated such that the performance of D2D links improves with the QoS of cellular links maintained at a desired level. It follows that resource allocation (RA) is a crucial issue for D2D communications in both theoretical and practical aspects. A huge number of research activities have been conducted to design and analyze RA schemes for D2D communications underlay cellular networks [7], [8], [9], [10], [11], [12]. Joint RA framework using convex optimization [7] and iterative power control algorithm [8] were introduced for D2D links. RA strategy in cooperative D2D communications was investigated [9], and interference management through D2D power allocation and shared channel assignment was proposed [10]. In [11], optimal power control and resource sharing mode selection algorithm depending on the application of D2D link were proposed. When different channels are allocated to distinct D2D links, efficient one-to-one mapping strategies can be applied to RA [12]. However, in case that multiple D2D links are allowed to utilize an identical cellular channel, RA becomes NP-hard and optimal RA solution cannot be found in an analytical manner. Accordingly, reduced-complexity sub-optimal RA mechanisms have been investigated to be applied to practical D2D communication systems [13], [14], [15], [16]. Two-phase channel assignment scheme with low complexity [13] and alternating optimization of subchannel and power allocation [14] were proposed for D2D communications underlay cellular network. Efficient RA scheme using convex optimization with reduced complexity was proposed for unmanned aerial vehicle (UAV)-assisted cellular network [15], and a joint RA scheme with simple lower-bound-based power control was applied to D2D-based V2X communications underlay cellular networks [16].

When multiple D2D links utilize a common cellular channel, the multiple access of D2D users becomes an important issue for a spectrum efficient network operation. Recently, non-orthogonal multiple access (NOMA) techonology has been proposed to improve the performance of multi-user communication systems in terms of spectral efficiency, transmission latency and user fairness at the cost of increasing complexity [17], [18], [19]. Multiple user signals are multiplexed into a single signal and transmitted over a single channel, then, each user detects its own signal with the aid of successive interference cancellation (SIC) at the receiver. In power-domain NOMA, which is a widely used mechanism, different power levels are assigned to distinct user signals according to which the order of detection and SIC is determined [20], [21], [22]. The performance of NOMA networks has been analyzed in various aspects. Outage probability of downlink NOMA network was analyzed [23], [24], [25], where ergodic sum rate was also analyzed in [23] and cooperative NOMA system with imperfect SIC was considered in [25]. In [26], the outage probability of uplink and downlink NOMA system over different fading channels was analyzed. RA mechanisms for NOMA networks have also been investigated [23], [27], [28], [29]. User pairing and its impact on the performance of NOMA networks were analyzed [27], [28], and iterative algorithm for channel and power allocation in NOMA system was proposed [29].

To enhance the massive connectivity of D2D UEs with maintaining a high level of spectral efficiency, it is desirable to employ a NOMA protocol for D2D communications underlay cellular networks. As a result, a NOMA-enabled D2D communication scheme was introduced and many research works have been conducted to show the potential benefits of using this technology [30], [31], [32], [33], [34]. RA algorithm based on matching theory was proposed for NOMA-enabled D2D communication systems [30], [31]. A resource management scheme using licensed and unlicensed spectrum was studied [32], and Hungarian algorithm was applied to RA for NOMA-based D2D communications [33]. Optimal channel and power allocation for D2D cluster with NOMA principle was also studied [34].

In case of considering UEs with mobility, we need a dynamic RA scheme suitable for networks whose set-up varies continuously. Dynamic RA requires a much higher network overhead and computational complexity than a static RA because resources need to be adaptively allocated according to varying network set-up. Data transmissions may be conducted over multiple time steps due to segmentation of long data frame into multiple short ones. The number of time steps may be determined by various factors, e.g., data rate, channel bandwidth and battery life of UEs. The RA becomes more complex if a sequence of resources needs to be allocated to users over multiple time steps. Thus, a sub-optimal RA scheme with low complexity is more demanding in mobile communication networks.

Recently, deep learning (DL) and reinforcement learning (RL) have attracted many researchers in a wide range of engineering fields. Deep RL (DRL) incorporates DL into RL, in which suitable decisions are made from unstructured input data, where deep Q-network (DQN) is a well-known example [35]. Various forms of learning mechanisms have been actively applied to optimization problems including RA for D2D communications [36], [37], [38], [39], [40], [41], [42], [43]. DL-based power allocation was applied to MIMO-NOMA system [36], and deep neural network (DNN) was used for sub-channel and power allocation of network with low complexity [37]. DRL-based RA schemes were proposed for D2D communications in various scenarios [38], [39], [40], [41], [42], [43], [44], where distributed DRL was employed in [38], D2D pairs were focused in [39], unicast and broadcast in V2V communications were considered in [40], and double deep-Q-network (DDQN)-based dynamic spectrum access algorithm was proposed in [41]. Joint sub-channel and power allocation scheme based on DRL was proposed for NOMA cellular network [42], and deep deterministic policy gradient (DDPG) scheme was utilized in RA for NOMA-based V2X communications [43].

Depending on who performs RA, we categorize RA mechanism into a centralized one and a decentralized one. Centralized RA achieves high QoS of communication links at the cost of high amount of system overhead and high computational complexity. On the other hand, decentralized RA requires low amount of system overhead and complexity resulting in a degraded QoS. In general, a single-agent framework is used for centralized RA [41], [42] while a multi-agent framework is used for decentralized RA [43], [45]. To obtain a high QoS with a low computational complexity, we adopt a multi-agent DRL framework in the centralized RA scheme. Agents reside in a central coordinator of the cell, where each agent corresponds to each D2D link. The ANN in a single-agent DRL is segmented into multiple smaller ones, where each agent has its own ANN. Then, each agent in multi-agent DRL requires lower computational complexities in both the training phase and testing phase than the single agent.

In this paper, we consider D2D cluster communications underlay cellular networks to serve much more D2D users than the number of available channels, where D2D links in each cluster operate with NOMA protocol. Considering randomly varying small-scale fading gain of channels, we investigate the outage to evaluate the quality of communication links. We also consider random variation of UEs' positions to reflect the high mobility of UEs in the cell. We formulate the operation of NOMA-enabled D2D communications and derive outage probabilities of D2D links as well as cellular links. To obtain D2D outage probability, we analyze the influence of success and failure of SIC for preceding D2D user signals under the NOMA framework. Then, we define an effective throughput and provide a benefit as the performance measure of the RA scheme for D2D communications underlay cellular networks, where the benefit is defined as the sum of average effective throughput of all cellular and D2D links in a cell accumulated over multiple time steps. Transmit power and channel spectrum are considered communication resources to be allocated. The objective of RA is to determine resources of each D2D link to maximize the benefit of the cell. To obtain high QoS of overall network with a reduced complexity, we construct a multi-agent DRL framework for RA operating in a centralized manner. Multi agents conduct constituent learning processes in a cyclic manner with a timing-offset in a training phase. In a testing phase, the proposed RA scheme promptly allocates resources to D2D links depending on the observation for environment including positions of UEs in the cell, which vary dynamically. This work is an extension of a previous work of the authors [39], where the scenario of D2D pairs is extended to D2D clusters employing NOMA protocol for downlink D2D communications underlay cellular networks.

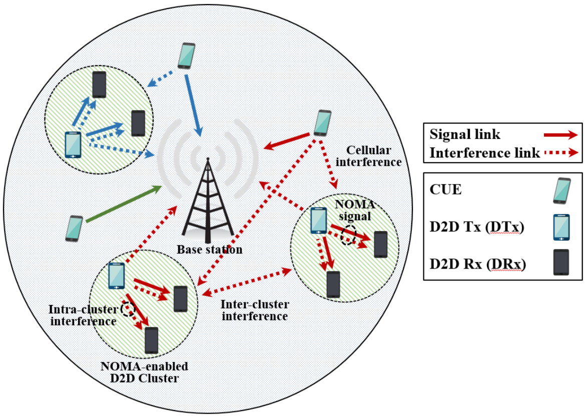


FIGURE 1. Graphical description of system model.

It is observed from simulations that the proposed DRL-based RA scheme performs well in various aspects. The usefulness of the proposed RA scheme is clear in case that UEs are distributed densely in a cell resulting in a high level of mutual interferences among UEs. The performance gain of the proposed RA scheme over other schemes is significant when cellular links have a higher priority over D2D links. The proposed RA scheme enables an energy efficient operation of whole network including D2D as well as cellular links. We can also obtain high levels of fairness and coordination of D2D users by using the proposed RA scheme. Consequently, the proposed RA scheme is considered practically efficient in the next-generation wireless communications in which a high number of UEs with high mobility exist in cellular networks.

This paper is organized as follows. In Sec. II, we present the system model of NOMA-enabled D2D cluster communications underlay cellular networks and formulate its operation. In Sec. III, we provide the performance measure of RA and define the optimal RA problem for NOMA-enabled D2D communications underlay cellular networks. In Section IV, we propose a multi-agent DRL-based RA scheme, in which multiple agents conduct constituent learning in a cyclic manner with a timing offset in a training phase. In Section V, we analyze the performance of the proposed scheme in various aspects and compare it with other RA schemes. Finally, we conclude this paper in Section VI.

II. SYSTEM MODEL

We consider a single-cell cellular network shown in Fig. 1, in which a base station (BS) is located at the center and M cellular user equipments (CUE) and K NOMA-enabled D2D clusters are distributed over the cell. We consider a cellular uplink period, in which D2D clusters operate in a downlink mode with a NOMA strategy underlay cellular network. Each

D2D cluster is composed of multiple D2D user equipments (DUEs) in proximity, where one of DUEs operates as a transmitter (DTx), or a cluster head, and other N DUEs operate as receivers (DRx).[1] We use an index k, $1 \leq k \leq K$, to specify a D2D cluster or DTx in that cluster. We also let indices m and $n(k)$ represent a cellular link, or CUE, and DRx in the k-th D2D cluster, respectively, where $1 \leq m \leq M$ and $1 \leq n \leq N$. We let $h_{x,y}$ denote a small-scale fading gain of the channel between two nodes x and y, where nodes include CUE, DTx, DRx and BS denoted by m, k, $n(k)$ and B, respectively. We let $d_{x,y}$ denote the distance between two nodes x and y, and we use a log-distance model for large-scale fading as $d_{x,y}^{-\alpha}$ with a path loss exponent α. We model the location of UEs in the cell and the location of DRxs in each D2D cluster as a uniform binomial point process (BPP) [12]. We suppose $h_{x,y}$ is independent and identically distributed (i.i.d.) zero-mean circularly symmetric complex Gaussian with a unit variance. We define a channel access indicator $\delta_{m,k}$ as $\delta_{m,k} = 1$ if a D2D cluster k and CUE m share a channel, and $\delta_{m,k} = 0$ otherwise. In the same manner, $\delta_{j,k} = 1$ if D2D clusters j and k share a channel, and $\delta_{j,k} = 0$ otherwise.

Each cellular link is allocated a dedicated channel so that a cellular channel is not occupied by multiple cellular links. D2D clusters are allowed to use channels already occupied by cellular links, where multiple D2D clusters can share a cellular channel. Within a D2D cluster, N distinct D2D user signals are allocated different levels of transmit power and superimposed into a single downlink NOMA signal at DTx. Then, the NOMA signal is sent to all DRxs over the channel assigned to the cluster. Let $s^D_{n(k)}$ denote a D2D user $n(k)$ signal with a unit power and $P^D_{n(k)}$ be the transmit power allocated to the corresponding user signal. Then, the NOMA signal, x^D_k, transmitted from the DTx of a cluster k is expressed as

$$x^D_k = \sum_{n=1}^{N} \sqrt{P^D_{n(k)}} s^D_{n(k)}, \quad 1 \leq k \leq K. \tag{1}$$

The total transmit power of x^D_k is determined as $P^D_k = \sum_{n=1}^{N} P^D_{n(k)}$. According to NOMA protocol, transmit power of user signals is determined in the reverse order of the corresponding channel strength. Considering small-scale fading and large-scale fading of the channel, the strength of channel between two nodes x and y is determined by $|h_{x,y}|^2 d^{-\alpha}_{x,y}$. Then, the transmit power allocation for D2D user signals in the cluster by NOMA principle is summarized as the following: if $|h_{k,1(k)}|^2 d^{-\alpha}_{k,1(k)} < |h_{k,2(k)}|^2 d^{-\alpha}_{k,2(k)} < \cdots < |h_{k,n(k)}|^2 d^{-\alpha}_{k,n(k)}$, then $P^D_{1(k)} > P^D_{2(k)} > \ldots > P^D_{n(k)}$, where the user index n in the cluster k is labelled in an ascending order of channel strength or in a descending order of transmit power.

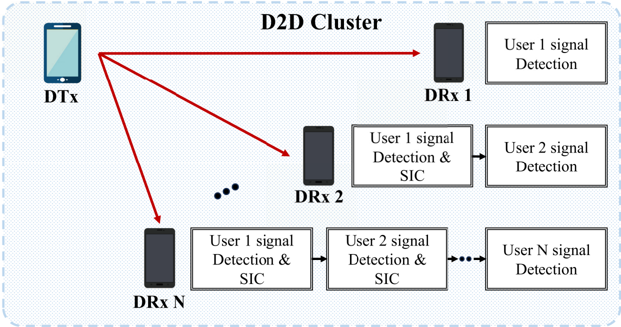


FIGURE 2. Detection and SIC for user signals in NOMA-enabled D2D downlink communications.

Each DRx detects its own signal from the received NOMA signal through SIC as depicted in Fig. 2. Before detecting its own signal, each DRx performs detection and SIC for user signals with higher transmit power than its own signal in a descending order of transmit power. Thus, each DRx detects its own signal in the presence of other user signals with lower transmit power as interferences under the assumption that all user signals with higher transmit power are completely cancelled by SIC. If SIC for preceding user signals is not perfectly done, residual error signals exist when detecting its own signal. In case that channel gains vary within a cellular uplink period or D2D downlink period, the transmit power allocation needs to be based on the mean strength of channel, i.e., $E\left\{|h_{k,n(k)}|^2 d^{-\alpha}_{k,n(k)}\right\}$. Since small-scale fading gains are considered identical for all channels, the transmit power is allocated to user signals in the order of distance from the corresponding DRx to DTx. It follows that the order of SIC at each DRx is also determined as the order of distance between DRx and DTx.

The received signal used for detecting the user $n'(k)$ signal at DRx $n(k)$ after SIC for preceding user signals is written by

$$\begin{aligned} y_{n',n(k)} = & \sum_{m=1}^{M} \delta_{m,k} \sqrt{P^C_m} x^C_m h_{m,n(k)} \sqrt{d^{-\alpha}_{m,n(k)}} \\ & + \sqrt{P^D_{n'(k)}} s^D_{n'(k)} h_{k,n(k)} \sqrt{d^{-\alpha}_{k,n(k)}} \\ & + \sum_{i=1}^{n'-1} \sqrt{P^D_{i(k)}} e^D_{i(k)} h_{k,n(k)} \sqrt{d^{-\alpha}_{k,n(k)}} \\ & + \sum_{i=n'+1}^{N} \sqrt{P^D_{i(k)}} s^D_{i(k)} h_{k,n(k)} \sqrt{d^{-\alpha}_{k,n(k)}} \\ & + \sum_{j=1:j\neq k}^{K} \delta_{j,k} x^D_j h_{j,n(k)} \sqrt{d^{-\alpha}_{j,n(k)}} + w_{n(k)}, \end{aligned} \tag{2}$$

where $n' \leq n$, $e^D_{i(k)}$ is a residual SIC error defined by the discrepancy between the symbol $s^D_{i(k)}$ and its corresponding detection, x^C_m is the m-th CUE's signal with a unit power,

[1] We may allow a DRX to download data from multiple DTXs, which is considered the scenario of overlapping D2D clusters. In this scenario, each D2D cluster has one DTx and operates with NOMA protocol while DRx may belong to multiple D2D clusters. The signal detection at DRx needs to employ multi-user detection techniques, which will result in much more complex signal detection processes. This scenario is out of scope of this paper and the RA for overlapping D2D clusters needs to be studied further. Note that there is no limit on the number of DRxs in clusters.

TABLE 1. List of notations and symbols.

Symbol	Description
P_m^C	Transmit power of the m-th CUE
$P_{n(k)}^D$	Transmit power of the n-th user signal in the k-th D2D cluster
P_k^D	Total transmit power of the k-th D2D cluster ($P_k^D = \sum_{n=1}^N P_{n(k)}^D$)
$h_{m,B}$	Small-scale fading gain between the m-th CUE and BS
$h_{k,B}$	Small-scale fading gain between DTx of the k-th cluster and BS
$h_{m,n(k)}$	Small-scale fading gain between the m-th CUE and the n-th DRx in the k-th cluster
$h_{k,n(k)}$	Small-scale fading gain between DTx and the n-th DRx in the k-th cluster
$h_{j,n(k)}$	Small-scale fading gain between DTx of the j-th cluster and the n-th DRx in the k-th cluster
$\delta_{m,k}$	Binary(1/0) indicator showing if the m-th CUE and the k-th D2D cluster share a channel
$\delta_{j,k}$	Binary(1/0) indicator showing if the j-th D2D cluster and the k-th D2D cluster share a channel
$d_{m,B}$	Distance between the m-th CUE and BS
$d_{k,B}$	Distance between DTx of the k-th cluster and BS
$d_{m,n(k)}$	Distance between the m-th CUE and the n-th DRx of the k-th cluster
$d_{k,n(k)}$	Distance between DTx and the n-th DRx in the k-th cluster
$d_{j,n(k)}$	Distance between DTx of the j-th cluster and the n-th DRx of the k-th cluster
R^C	Target rate of cellular link
R^D	Target rate of D2D link
γ_{th}^D	Threshold SINR below which an outage of D2D link is declared ($\gamma_{th}^D = 2^{R^D} - 1$)
γ_{th}^C	Threshold SINR below which an outage of cellular link is declared ($\gamma_{th}^C = 2^{R^C} - 1$)
$y_{n',n(k)}$	Received signal used for detecting the n'-th user signal at the n-th DRx in the k-th cluster after SIC for preceding signals
$\gamma_{n',n(k)}^D$	SINR measured at the n-th DRx in the k-th D2D cluster when detecting the n'-th user signal
γ_m^C	SINR of the m-th cellular link measured at BS
$p_{n',n(k)}^{out}$	Outage probability of the n'-th user signal at the n-th DRx in the k-th cluster
p_m^{out}	Outage probability of the m-th cellular link

and $w_{n(k)}$ is a zero-mean Gaussian noise with a variance of σ^2. Note that the first term in the right hand side corresponds to the interference from cellular link. The third term represents the residual interference existing due to an unsuccessful SIC for preceding signals. The fourth and fifth terms represent the intra-cluster interference and inter-cluster interference, respectively. The second term is related with the user $n'(k)$ signal. The signal-to-interference-plus-noise-ratio (SINR) measured at the n-th DRx in the k-th D2D cluster when detecting user $n'(k)$ signal is obtained as

$$\gamma_{n',n(k)}^D = \frac{P_{n'(k)}^D |h_{k,n(k)}|^2 d_{k,n(k)}^{-\alpha}}{I_{n(k)}^{CU} + I_{n',n(k)}^{RES} + I_{n',n(k)}^{NOMA} + I_{n(k)}^{D2D} + \sigma^2}, \quad (3)$$

where

$$I_{n(k)}^{CU} = \sum_{m=1}^{M} \delta_{m,k} P_m^C |h_{m,n(k)}|^2 d_{m,n(k)}^{-\alpha}$$

$$I_{n',n(k)}^{RES} = \sum_{i=1}^{n'-1} P_{i(k)}^D P_{i(k)}^e |h_{k,n(k)}|^2 d_{k,n(k)}^{-\alpha}$$

$$I_{n',n(k)}^{NOMA} = \sum_{i=n'+1}^{N} P_{i(k)}^D |h_{k,n(k)}|^2 d_{k,n(k)}^{-\alpha}$$

$$I_{n(k)}^{D2D} = \sum_{j=1:j\neq k}^{K} \delta_{j,k} P_j^D |h_{j,n(k)}|^2 d_{j,n(k)}^{-\alpha}$$

and $P_{i(k)}^e$ is the power of $e_{i(k)}^D$. Note that $I_{n(k)}^{CU}$, $I_{n',n(k)}^{RES}$, $I_{n',n(k)}^{NOMA}$ and $I_{n(k)}^{D2D}$ represent interference power caused by cellular link signal, SIC error, D2D user signals with lower transmit power in the same cluster and D2D signal of other clusters, respectively.

The received signal at BS for the m-th CUE's uplink signal is written by

$$y_m = \sqrt{P_m^C} x_m^C h_{m,B} \sqrt{d_{m,B}^{-\alpha}} + \sum_{k=1}^{K} \delta_{m,k} x_k^D h_{k,B} \sqrt{d_{k,B}^{-\alpha}} + w_B, \quad (4)$$

where P_m^C is the transmit power of x_m^C and w_B is a noise at the receiver in BS which is zero-mean Gaussian with a variance of σ^2. The SINR of the m-th cellular link, measured at BS, is defined by

$$\gamma_m^C = \frac{P_m^C |h_{m,B}|^2 d_{m,B}^{-\alpha}}{\sum_{k=1}^K \delta_{m,k} P_k^D |h_{k,B}|^2 d_{k,B}^{-\alpha} + \sigma^2}. \quad (5)$$

We analyze the performance of RA scheme by using SINRs of D2D links and cellular links obtained above in the following section. Note that notations and symbols used in this paper are listed in Table 1.

III. SYSTEM PERFORMANCE ANALYSIS

Since $h_{k,n(k)}$ are i.i.d. for all k and n, the average strength of D2D downlink channel is determined by $d_{k,n(k)}^{-\alpha}$. Suppose $d_{k,1(k)} > d_{k,2(k)} > \cdots > d_{k,N(k)}$, then the downlink channel for the user $l'(k)$ is weaker than that for the user $l(k)$ on the average if $l' < l$. Thus, the transmit power is allocated as $P_{1(k)}^D > P_{2(k)}^D > \cdots > P_{N(k)}^D$ when forming a NOMA signal. At DRx $1(k)$, user $1(k)$ signal is detected in the presence of all other user signals as an interference and SIC is not needed.

At DRx $n(k)$, $n > 1$, user $n'(k)$ signals, $n' < n$, are detected and cancelled earlier in a successive manner so that user $n(k)$ signal is detected in the presence of user $l(k)$ signals, $l > n$, as interferences.

We define an outage of a user link as an event that the achievable data rate of the corresponding user signal does not exceed a target rate [46]. Since the achievable rate is defined by $\log_2(1 + \text{SINR})$ for a given SINR, the outage probability of user n' signal at DRx n in the cluster k is written by

$$\begin{aligned} p^{out}_{n',n(k)} &= \Pr\left\{\log_2(1+\gamma^D_{n',n(k)}) < R^D\right\} \\ &= \Pr\{\gamma^D_{n',n(k)} < \gamma^D_{th}\}, \end{aligned} \tag{6}$$

where R^D is a target rate of D2D link and γ^D_{th} is a threshold SINR satisfying $R^D = \log_2(1+\gamma^D_{th})$. We declare a successful detection of a signal if a target rate is achieved [23]. Thus, the detection probability of a user signal is equal to the probability that the user signal is not in outage. It follows that the probability that user $n'(k)$ signal is correctly detected at DRx $n(k)$ is written as $1 - p^{out}_{n',n(k)}$ and it is expanded by the law of total probability as

$$\begin{aligned} 1 - p^{out}_{n',n(k)} &= \Pr\{\gamma^D_{n',n(k)} \geq \gamma^D_{th}\} \\ &= \Pr\{\gamma^D_{n',n(k)} \geq \gamma^D_{th} \mid \gamma^D_{n'-1,n(k)} \geq \gamma^D_{th}\} \cdot \Pr\{\gamma^D_{n'-1,n(k)} \geq \gamma^D_{th}\} \\ &\quad + \Pr\{\gamma^D_{n',n(k)} \geq \gamma^D_{th} \mid \gamma^D_{n'-1,n(k)} < \gamma^D_{th}\} \cdot \Pr\{\gamma^D_{n'-1,n(k)} < \gamma^D_{th}\}. \end{aligned} \tag{7}$$

According to the power allocation policy in NOMA, the transmit power in the D2D cluster k is assigned to two successive user signals $n'-1$ and n' as $P^D_{n'-1(k)} > P^D_{n'(k)}$. If $\gamma^D_{n'-1,n(k)} < \gamma^D_{th}$, the user $n'-1$ signal is not correctly detected and thus SIC for this signal is not successful. This causes the existence of residual SIC error interference for user $n'-1$ signal when detecting user n' signal. Considering this fact and the power allocation $P^D_{n'-1(k)} > P^D_{n'(k)}$ introduced above, it is inferred from the definition of SINR given in (3) that we have $\gamma^D_{n'-1,n(k)} > \gamma^D_{n',n(k)}$ with high probability. Then, we may approximate $\Pr\{\gamma^D_{n',n(k)} \geq \gamma^D_{th} \mid \gamma^D_{n'-1,n(k)} < \gamma^D_{th}\} \approx 0$ and rewrite (7) as

$$\begin{aligned} &\Pr\{\gamma^D_{n',n(k)} \geq \gamma^D_{th}\} \\ &\approx \Pr\{\gamma^D_{n',n(k)} \geq \gamma^D_{th} \mid \gamma^D_{n'-1,n(k)} \geq \gamma^D_{th}\} \cdot \Pr\{\gamma^D_{n'-1,n(k)} \geq \gamma^D_{th}\}. \end{aligned} \tag{8}$$

Then, we have the following sequence of approximations:

$$\begin{aligned} \Pr\{\gamma^D_{n'-1,n(k)} \geq \gamma^D_{th}\} &\approx \Pr\{\gamma^D_{n'-1,n(k)} \geq \gamma^D_{th} \mid \gamma^D_{n'-2,n(k)} \geq \gamma^D_{th}\} \\ &\quad \cdot \Pr\{\gamma^D_{n'-2,n(k)} \geq \gamma^D_{th}\}, \\ \Pr\{\gamma^D_{n'-2,n(k)} \geq \gamma^D_{th}\} &\approx \Pr\{\gamma^D_{n'-2,n(k)} \geq \gamma^D_{th} \mid \gamma^D_{n'-3,n(k)} \geq \gamma^D_{th}\} \\ &\quad \cdot \Pr\{\gamma^D_{n'-3,n(k)} \geq \gamma^D_{th}\}, \\ &\vdots \\ \Pr\{\gamma^D_{2,n(k)} \geq \gamma^D_{th}\} &\approx \Pr\{\gamma^D_{2,n(k)} \geq \gamma^D_{th} \mid \gamma^D_{1,n(k)} \geq \gamma^D_{th}\} \\ &\quad \cdot \Pr\{\gamma^D_{1,n(k)} \geq \gamma^D_{th}\}. \end{aligned}$$

By plugging this sequence of approximations into (8) and recalling $1 - p^{out}_{n',n(k)} = \Pr\{\gamma^D_{n',n(k)} \geq \gamma^D_{th}\}$, we obtain

$$\begin{aligned} &1 - p^{out}_{n',n(k)} \\ &\approx \Pr\{\gamma^D_{1,n(k)} \geq \gamma^D_{th}\} \cdot \prod_{l=2}^{n'} \Pr\{\gamma^D_{l,n(k)} \geq \gamma^D_{th} \mid \gamma^D_{l-1,n(k)} \geq \gamma^D_{th}\}. \end{aligned} \tag{9}$$

By using $1 - p^{out}_{l,n(k)} = \Pr\{\gamma^D_{l,n(k)} \geq \gamma^D_{th}\}$, we claim that $\gamma^D_{l,n(k)} \geq \gamma^D_{th}$ implies the successful detection of user l signal. It is inferred from $\Pr\{\gamma^D_{l,n(k)} \geq \gamma^D_{th} \mid \gamma^D_{l-1,n(k)} < \gamma^D_{th}\} \approx 0$ that the successful detection of user $l-1$ signal and the perfect cancellation of corresponding signal are mandatory condition for successful detection of user l signal. By considering sequential relationship, we can claim that when detecting user l signal, meeting the condition $\gamma^D_{l-1,n(k)} \geq \gamma^D_{th}$ can be interpreted as perfect SIC for all preceding user signals, i.e., users $l-1, l-2, \cdots, 1$. Thus, we express (9) in a more intuitive form as

$$\begin{aligned} &1 - p^{out}_{n',n(k)} \\ &\approx \prod_{l=1}^{n'} \Pr\{\gamma^D_{l,n(k)} \geq \gamma^D_{th} \mid \text{perfect SIC for preceding user signals}\}. \end{aligned} \tag{10}$$

Note that user $1(k)$ signal is detected first. So, without loss of generality, we can include the detection of user $1(k)$ signal in a generalized form given in (10). Since $\Pr\{\gamma^D_{l,n(k)} \geq \gamma^D_{th}\} = 1 - p^{out}_{l,n(k)}$ as introduced earlier, we express (10) as

$$\begin{aligned} &1 - p^{out}_{n',n(k)} \\ &\approx \prod_{l=1}^{n'} \left(1 - p^{out}_{l,n(k)} \mid \text{perfect SIC for preceding user signals}\right). \end{aligned} \tag{11}$$

Under the condition of perfect SIC for preceding user signals, we obtain $1 - p^{out}_{l,n(k)}$ as (12), shown at the bottom of the next page, where the detailed derivation is provided in Appendix A.

Then, by plugging the result of (12) obtained for all $l \leq n'$ into (11), we obtain a closed form expression for the probability that user $n'(k)$ signal is correctly detected at DRx $n(k)$, i.e., $1 - p^{out}_{n',n(k)}$, where $n' \leq n$.

The outage probability of a cellular link m is obtained by using (5) as

$$\begin{aligned} p^{out}_m &= \Pr\left\{\log_2\left(1+\gamma^C_m\right) < R^C\right\} \\ &= \Pr\left\{\frac{P^C_m |h_{m,B}|^2 d^{-\alpha}_{m,B}}{\sum_{k=1}^{K} \delta_{m,k} P^D_k |h_{k,B}|^2 d^{-\alpha}_{k,B} + \sigma^2} < \gamma^C_{th}\right\}, \end{aligned} \tag{13}$$

where $\gamma^C_{th} = 2^{R^C} - 1$. It is known from [39] that for exponential random variables $a_0, \cdots, a_N$, we have

$\Pr\left\{\frac{a_0}{\sum_{i=1}^{N} a_i+b} \leq \gamma\right\} = 1-e^{-\frac{b}{\mu_0}\gamma} \prod_{i=1}^{N}\left(1+\frac{\mu_i}{\mu_0}\gamma\right)^{-1}$, where $\mu_i = E\{a_i\}$. Thus, (13) can be rewritten as [39]

$$\begin{aligned} p_m^{out} &= 1-\exp\left(-\frac{\sigma^2\gamma_{th}^C}{P_m^C d_{m,B}^{-\alpha}}\right)\cdot\prod_{k=1}^{K}\left(1+\gamma_{th}^C\frac{\delta_{m,k}P_k^D d_{k,B}^{-\alpha}}{P_m^C d_{m,B}^{-\alpha}}\right)^{-1} \\ &= 1-\exp\left(-\frac{\sigma^2\gamma_{th}^C}{P_m^C d_{m,B}^{-\alpha}}\right)\cdot\prod_{k=1}^{K}\left(1+\gamma_{th}^C\frac{P_k^D}{P_m^C}\left(\frac{d_{k,B}}{d_{m,B}}\right)^{-\alpha}\right)^{-\delta_{m,k}}. \end{aligned} \tag{14}$$

We define an effective throughput as a target rate multiplied by the probability that the user is not in outage. Thus, the effective throughput of cellular link m and D2D link $n(k)$ is obtained by $R^C\left(1-p_m^{out}\right)$ and $R^D\left(1-p_{n,n(k)}^{out}\right)$, respectively, where $p_{n,n(k)}^{out}$ is the outage probability of the downlink communication for user n in the k-th cluster. The goal of RA is determining the channel spectrum and transmit power of all D2D links at each time step to maximize the cumulative sum of average effective throughputs of cellular link and D2D link over multiple time steps T. This can be mathematically expressed by

$$\begin{aligned} \max_{\substack{P_{n(k)}^D[t],\delta_{m,k}[t] \\ \forall m,k,n,t}} &\sum_{t=1}^{T}\left\{\frac{1}{M}\sum_{m=1}^{M} R^C\left(1-p_m^{out}[t]\right)\right. \\ &\left.+\frac{1}{KN}\sum_{k=1}^{K}\sum_{n=1}^{N} R^D\left(1-p_{n,n(k)}^{out}[t]\right)\right\} \\ \text{subject to } &\sum_{m=1}^{M}\delta_{m,k}[t] \leq 1 \quad \text{for each } k,t \\ &P_{\min}^D \leq P_{n(k)}^D[t] \leq P_{\max}^D \quad \text{for each } n,k,t \\ &P_{i(k)}^D[t] \leq P_{j(k)}^D[t] \text{ if } i>j \quad \text{for each } k,t, \end{aligned} \tag{15}$$

where the time step is specified in variables as $[t]$ with a slight abuse of notation. The first constraint enforces each D2D cluster to reuse one cellular channel at most. The last constraint guides power allocation for downlink NOMA signals depending on the order of channel strength as introduced above.

Considering mobilities of CUEs and DUEs, distances between pairs of UEs may vary for every transmission period. CUEs may change their dedicated channels depending on network conditions. Then, transmit power and channel allocation for all D2D user signals need to be determined optimally every transmission period. This may incur tremendously heavy network load and require extremely high amount of computational complexity, which makes optimal RA infeasible in practical network environment. Thus, there exists a strong need for designing practically feasible RA scheme for D2D communications underlay cellular networks showing reasonably high effective throughput with low computational complexity.

As a solution to resolve this problem, we propose a multi-agent DRL based RA scheme which is introduced in detail in the following section.

IV. DRL-BASED RESOURCE ALLOCATION

RL is a mechanism of learning how to make suitable decision for a given situation in order to maximize a return through a trial-and-error search. Markov decision process (MDP) is adopted to formalize sequential decision making in RL, in which agents interact with environment, observe states and take actions. MDP is represented by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions taken for a given state, $\mathcal{P} : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \rightarrow [0,1]$ is a probability that a pair of state and action is mapped to a next state, and $\mathcal{R}$ is a set of rewards. In RL, at a certain time step t, an agent observes a state $s[t] \in \mathcal{S}$ from an environment and takes an action $a[t] \in \mathcal{A}$ based on a policy π, by which the state $s[t]$ transits to a new state $s'[t]$, where $s'[t]$ is used as a state $s[t+1]$ at the next time step. The agent obtains a reward $r[t]$ and evaluates an expected return obtained by starting from a state s, taking an action a, and following a policy π thereafter in the form of an action-value funciton $q_\pi(s,a) = \mathbb{E}\{G[t] \mid s[t]=s, a[t]=a, \pi\}$, where a return is defined by $G[t] = \sum_{k=0}^{\infty}\beta^k r[t+k+1]$ with a discount factor $0 \leq \beta \leq 1$.

Q-learning is an off-policy RL algorithm handling stochastic transitions without environment model. At each time step, the agent at a certain state selects an action based on an action selection rule, e.g., greedy, ϵ-greedy, and soft-max method. The quality of the state and action pair is evaluated by a state-action value $Q(s,a)$, which is recursively updated by

$$\begin{aligned} Q(s[t],a[t]) \leftarrow &(1-\mu)Q(s[t],a[t]) \\ &+\mu\left(r[t]+\beta\max_{a} Q(s'[t],a)\right), \end{aligned}$$

$$1-p_{l,n(k)}^{out} = \begin{cases} \exp\left(-\frac{\sigma^2}{P_{l(k)}^D d_{k,n(k)}^{-\alpha}}D_{l(k)}^{-1}\right)\cdot\prod\limits_{m=1}^{M}\left(1+\frac{P_m^C d_{m,n(k)}^{-\alpha}}{P_{l(k)}^D d_{k,n(k)}^{-\alpha}}D_{l(k)}^{-1}\right)^{-\delta_{m,k}}\cdot\prod\limits_{j=1:j\neq k}^{K}\left(1+\frac{P_j^D d_{j,n(k)}^{-\alpha}}{P_{l(k)}^D d_{k,n(k)}^{-\alpha}}D_{l(k)}^{-1}\right)^{-\delta_{k,j}}, & \text{if } D_{l(k)} > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where $D_{l(k)} = \frac{1}{\gamma_{th}^D} - \sum_{i=l+1}^{N}\frac{P_{i(k)}^D}{P_{l(k)}^D}$.

where μ is the learning rate and $s'[t] = s[t+1]$. The optimal action-value function, $q^*(s, a) = \max_\pi q_\pi(s, a)$, is approximated by the state-action value $Q(s, a)$, independent of the policy being followed. In MDP, the state-action value converges with probability 1 to the optimal action-value function if each action is executed at each state during the infinite run times and the learning rate μ decays appropriately.

With a large state space $\mathcal{S}$, evaluating $Q(s, a)$ for all $s \in \mathcal{S}$ requires high computational complexity. As a solution, DQN, or DRL, adopting an artificial neural network (ANN) as a function approximator for state-action values has been introduced [35]. In the training phase of DQN, two ANNs, called a prediction network and a target network, are used. Observations are defined as states and fed to input nodes of ANNs. For a given state s, the prediction network computes $Q(s, a)$ for each realization of action a at each output node. The agent takes an action based on an action selection rule, by which a reward r and a new state s' are obtained. Then, the transition vector $\{s, a, r, s'\}$ is stored in the experience replay memory. Random samples of prior transition vectors are picked from an experience replay memory and used to evaluate a loss function in the prediction network and target network. The prediction network is updated at every time step while the target network is updated periodically or updated softly at every time step. After a training phase is completed, the testing phase begins, in which the agent takes an action a resulting in the greatest $Q(s, a)$ for a given state s.

We propose a DRL-based centralized RA scheme for NOMA-enabled D2D cluster communications underlay cellular networks. To reduce the high computational complexity inherent to centralized RA schemes, we adopt a multi-agent structure [39] in DRL, where each agent corresponds to each D2D cluster and operates its own DQN while agents exist physically in a central coordinator at BS. The state defined by observations of environment is shared by all agents while the action is defined distinctly for each agent. Since each agent has its own action, the size of action space is exponentially reduced compared with that in a single-agent DRL framework. Consequently, the ANN defined in a single-agent DRL can be segmented into multiple smaller ones, and the multi-agent DRL requires tremendously lower amount of computational complexity in both training and testing phases.

Let $z[t]$, $\mathbf{c}^C[t]$ and $\mathbf{P}^C[t]$ denote a vector of locations of all UEs in the cell, a vector of indices for channel spectrum occupied by CUEs and a vector of transmit powers of CUEs, respectively, at time step t. We also let $c_k^D[t] \in \{1, \cdots, M\}$ and $P_k[t] = \{P^D_{1(k)}[t], \cdots, P^D_{N(k)}[t]\}$ denote an index of channel spectrum assigned to the k-th D2D cluster and a vector of transmit power assigned to user signals in the k-th D2D cluster, respectively, at time step t. Then, we define a state $s[t]$ at time step t as

$$s[t] = \{\mathbf{z}[t], \mathbf{c}^C[t], \mathbf{c}^D[t], \mathbf{P}^C[t], \mathbf{P}_1[t], \cdots, \mathbf{P}_K[t]\}, \quad (16)$$

where $\mathbf{c}^D[t] = \{c_1^D[t], \cdots, c_K^D[t]\}$, and we define an action of the k-th agent at time step t as

$$a_k[t] = \left\{c_k^D[t], \mathbf{P}_k[t]\right\}. \quad (17)$$

Note that $\delta_{m,k}[t]$, $\forall m, k$, can be obtained from a given $\{\mathbf{c}^C[t], \mathbf{c}^D[t]\}$. For a practical implementation, we choose the value of $P^D_{n(k)}[t]$ out of pre-defined L discrete values, i.e., $P^D_{n(k)}[t] \in \{p_1, \cdots, p_L\}$. The instantaneous reward at time step t, denoted by $r[t]$, is defined as the sum of average effective throughputs of D2D and cellular links in the cell, i.e.,

$$r[t] = \frac{1}{M}\sum_{m=1}^{M} R^C\left(1 - p_m^{out}[t]\right) + \frac{1}{KN}\sum_{k=1}^{K}\sum_{n=1}^{N} R^D\left(1 - p^{out}_{n,n(k)}[t]\right). \quad (18)$$

We define an episode as a time duration T for which a sequence of data transmissions from DTx to DRxs is complete. In practice, the length of T may be determined according to data frame length, battery life, channel estimation cycle, etc. The reward accumulated during an episode will be called a benefit and expressed as

$$\text{Benefit} = \sum_{t=1}^{T} r[t]. \quad (19)$$

The goal of RA is to maximize the benefit of communications over cellular and D2D links under existing constraints introduced in (15).

Let us define constituent learning in RL as a sequence of operations by a single agent, i.e., observing a state s, taking an action a, observing a reward r and a new state s', and updating weights of prediction and target networks. To evaluate explicitly the influence of individual agent's action on the environment, we conduct constituent learning of multiple agents in a cyclic manner with a timing-offset [39] as described in Fig. 3. Without loss of generality, labeling agents is based on the order of performing constituent learning. After conducting constituent learning, each agent keeps idling until its turn comes around again. We define a time step as a period of learning by the agent 1 as depicted in Fig. 3. All UEs may change their locations at the beginning of every time step. Due to the existence of timing-offset, different agents may have distinct state values even within the identical time step. Thus, we specify the agent index in defining a state as $s_k[t]$. After the agent $k-1$ takes an action, a new state is observed as a result of environmental change. This new state is used as a state initiating constituent learning by the next agent k, i.e., $s_k[t] = s'_{k-1}[t]$, if $k \neq 1$. Note that agent 1 does not use $s'_K[t-1]$ observed by agent K at time step $t-1$ as $s_1[t]$ because UEs may change locations at the beginning of a new time step and thus the state is reset partly. The collection of constituent learning of all agents composes one learning iteration.

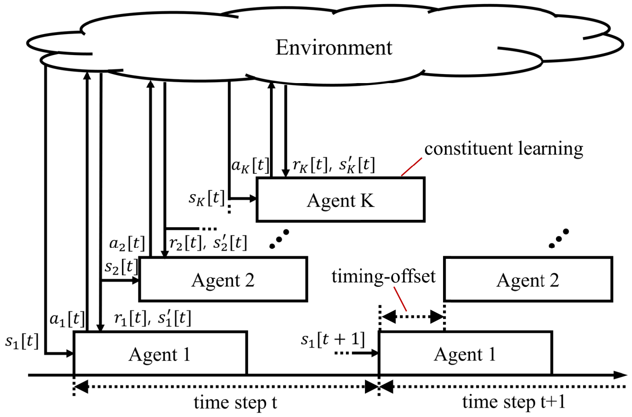


FIGURE 3. Learning of multiple agents in a cyclic manner.

Algorithm 1 Training Phase of Multi-Agent DRL

```
Initialization:
for $k = 1, \ldots, K$ do
    Randomly initialize weights of prediction network $\theta_k$
    Initialize weights of target network by $\theta'_k \leftarrow \theta_k$
    Initialize experience replay memory $\Omega_k$
end for
for $e = 1, \ldots, E$ do
    Randomly initialize $\mathbf{c}^D[1], \mathbf{P}_1[1], \cdots \mathbf{P}_K[1]$
    for $t = 1, \ldots, T$ do
        Randomly set $\mathbf{z}[t], \mathbf{c}^C[t], \mathbf{P}^C[t]$
        Define $s'_0[t] = \{\mathbf{z}[t], \mathbf{c}^C[t], \mathbf{c}^D[t], \mathbf{P}^C[t], \mathbf{P}_1[t], \cdots, \mathbf{P}_K[t]\}$
        for $k = 1, \ldots, K$ do
            Set state as $s_k[t] \leftarrow s'_{k-1}[t]$
            Determine action as $a_k[t] = \{c_k^D[t], \mathbf{P}_k[t]\}$ based on the chosen action selection rule
            D2D cluster $k$ takes an action accordingly
            Observe reward $r_k[t]$ and next state $s'_k[t]$
            Store transition vector $\{s_k[t], a_k[t], r_k[t], s'_k[t]\}$ in $\Omega_k$
            Randomly sample transition vectors from $\Omega_k$
            Obtain $Q(s_k[j], a_k[j]; \theta_k)$ from the prediction network
            Obtain $y_k[j]$ by (20) in the target network
            Compute the loss function $\mathcal{L}_k$ by (21)
            Update $\theta_k$ by a chosen optimization rule and $\theta'_k$ by (22)
        end for
    end for
end for
```

The learning procedure of agent k over multiple time steps is described as follows, which is also summarized in Algorithm 1. First, we randomly initialize weights of prediction network θ_k, and set weights of target network as $\theta'_k = \theta_k$. We also initialize experience replay memory Ω_k by running a random policy. The agent k observes a state $s_k[t]$ and takes an action $a_k[t]$, by which the environment changes and the agent k obtains a reward $r_k[t]$ by (18) and observes a new state $s'_k[t]$. The transition vector $T_k[t] = \left\{s_k[t], a_k[t], r_k[t], s'_k[t]\right\}$ is stored in Ω_k. A batch of transition vectors, which have been stored in Ω_k, are sampled randomly and used to evaluate a loss function. Suppose $T_k[j]$ is one sample included in a batch $\mathcal{B}$ picked up from Ω_k, where we use j to represent an index at which the transition vector is stored in Ω_k with a slight abuse of notation. The predicted state-action value $Q(s_k[j], a_k[j]; \theta_k)$ is obtained at the output node corresponding to the action $a_k[j]$ in the prediction network. The target state-action value $y_k[j]$ is obtained by a target network as

$$y_k[j] = r_k[j] + \beta \max_a Q(s'_k[j], a; \theta'_k). \tag{20}$$

A loss function $\mathcal{L}_k$ is computed as the mean squared error (MSE) between target and predicted state-action values by

$$\mathcal{L}_k = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \Big(y_k[j] - Q(s_k[j], a_k[j]; \theta_k) \Big)^2, \tag{21}$$

where $|\mathcal{B}|$ represents the batch size. We update weights of the prediction network θ_k by an appropriately chosen optimization algorithm and update weights of target network θ'_k softly as

$$\theta'_k \leftarrow (1-\tau)\theta'_k + \tau\theta_k, \tag{22}$$

where $\tau \ll 1$. We repeat the above process over time steps in each episode and repeat the whole process over E episodes.

After a training phase is completed, the RA system enters a testing phase, which corresponds to the actual operation of D2D communication system in the cellular network. At every time step, observations of BS, which are defined as a state, are input to the trained prediction network of each agent. Then, each agent chooses the action resulting in the maximum state-action value at the output of the prediction network. The chosen actions are reported to D2D clusters and used as resources for the downlink D2D communications. In the testing phase, RA for all agents may be executed simultaneously without timing-offset at each time step.

V. NUMERICAL RESULT

We consider a single circular-shaped cellular network with a radius r_c, in which BS is located at the center and M CUEs as well as K DTxs are distributed randomly following the uniform BPP model [12]. Around each DTx, a circular-shaped D2D cluster with a radius r_d is formed, in which DTx is positioned at the center and N DRxs are also distributed randomly following the uniform BPP model. We consider an uplink period of cellular links during which D2D links operate in a downlink mode with a NOMA protocol underlay cellular network. In actual operations of D2D communications underlay cellular networks, all UEs are allowed to change their positions at the beginning of every time step and thus D2D clusters need to determine the channel spectrum and transmit power for downlink D2D user signals at each time step by using the RA scheme.

We compare performances of RA schemes in terms of a benefit, where the proposed DRL-based RA scheme,

TABLE 2. System parameters used in simulations.

Symbol	Parameter	Value
M	The number of CUEs	4
K	The number of D2D clusters	4, 6, 8, 10, 12, 14
N	The number of DRxs in a cluster	2, 3
r_c	Radius of cell [m]	50, 100, 200
r_d	Radius of D2D cluster [m]	10
α	Pathloss exponent	3.5
$P^D_{n(k)}$	Transmit power for D2D user signals [dBm]	off, -70, -60, -50, -40, -30, -20, -10, 0
R^C	Target rate of cellular link [bits/s/Hz]	4, 6, 8, 10
R^D	Target rate of D2D link [bits/s/Hz]	2
σ^2	Noise power [dBm]	-104

TABLE 3. Hyperparameters used for DRL of the proposed RA scheme.

Symbol	Parameter	Value
μ	Learning rate	0.001
β	Discount factor	0.1
$\lvert\mathcal{B}\rvert$	Batch size	64
ϵ	Exploration rate	1 to 0.1
τ	Weight for soft update of target network	0.01
$\lvert\Omega_k\rvert$	Experience replay memory size	200000

a random RA scheme and a greedy RA scheme are compared. A random RA allocates the channel spectrum and transmit power of D2D links randomly at every time step. In a greedy RA scheme, the channel spectrum and transmit power for downlink D2D user signals are determined through a greedy search to maximize the instantaneous reward defined in (18) for each time step. We evaluate performances of RA schemes by Monte Carlo simulations, where Python 3.9.16 and PyTorch 1.13.0 are used as simulation softwares. We test various values for system parameters to analyze the behavior of RA schemes in various aspects. We list parameters characterizing the environment used for simulations in Table 2. Values of hyper-parameters used for DRL of the proposed RA scheme are also listed in Table 3. In order for D2D links to be adapted to varying environment in a real time manner, we let RA for D2D clusters be performed simultaneously without timing-offset in actual operations or in a testing phase. Note that the training of DRL for the proposed RA scheme is conducted in a cyclic manner with a timing-offset as explained in Sec. IV before the actual operation starts.

We let the transmit power of each CUE be controlled adaptively such that the corresponding SNR measured at BS without considering interference results in the outage probability of 0.001. Since transmit power of CUE is determined by the distance from BS, the state defined in (16) is reduced to $s[t] = \{\mathbf{z}[t], \mathbf{c}^C[t], \mathbf{c}^D[t], \mathbf{P}_1[t], \cdots, \mathbf{P}_K[t]\}$ and consequently, the size of ANN in prediction and target networks can be reduced thanks to the use of smaller number of input nodes. Each ANN in the prediction and target networks has five fully connected layers, which are an input layer, three hidden layers and an output layer. Each hidden layer has 512 neurons equipped with ReLU activation function, and an Adam optimizer is used for updating the weight of ANNs. Experience replay memories are initially filled with transition vectors obtained by running random policies. The training phase of DRL is completed by 100000 iterations, and the ϵ-greedy policy with linear annealing is applied as an action selection rule, where ϵ decays linearly during half of total learning iterations. It is observed that the performance in testing phase improves as the number of learning iterations in training phase grows. However, the performance improvement starts to be saturated from 100000 learning iterations. Considering the balance between the computational complexity in trainng phase and the performance in testing phase, we choose the number of learning iterations as 100000. It is observed from Fig. 4 that DQN is updated well during the training phase and it converges. In simulation results, we plot normalized benefits which are defined by the benefit divided by an episode time-duration T.

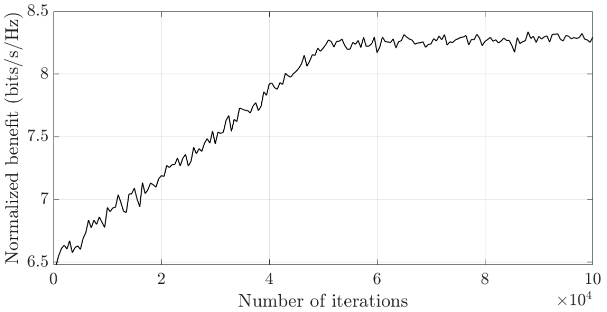
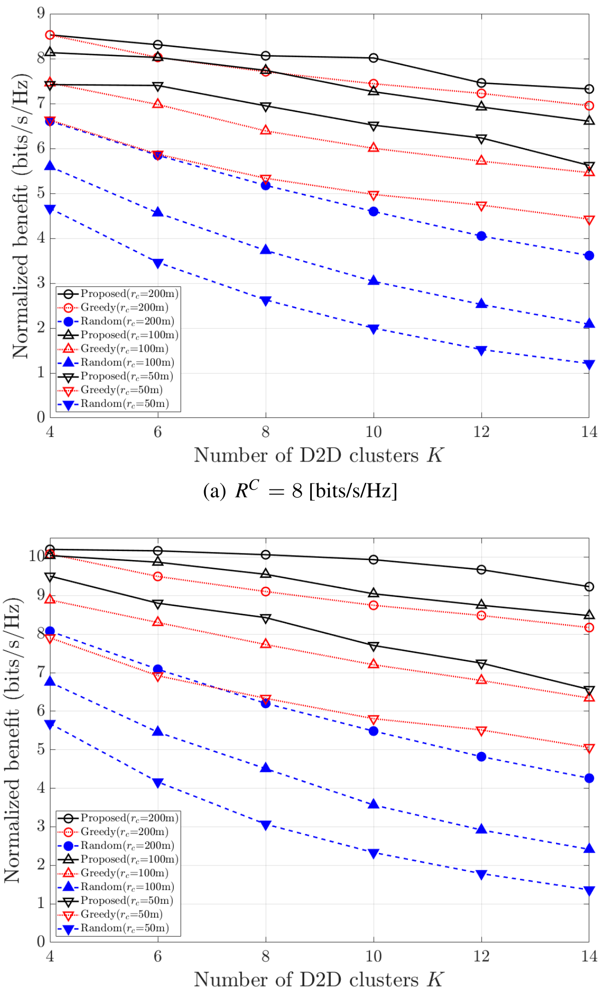


FIGURE 4. Evolution of normalized benefit in a training phase with $R^C = 8$ [bit/s/Hz], $r_c = 200$ [m] and $K = 4$, where normalized benefits averaged over every 50 episodes are plotted.

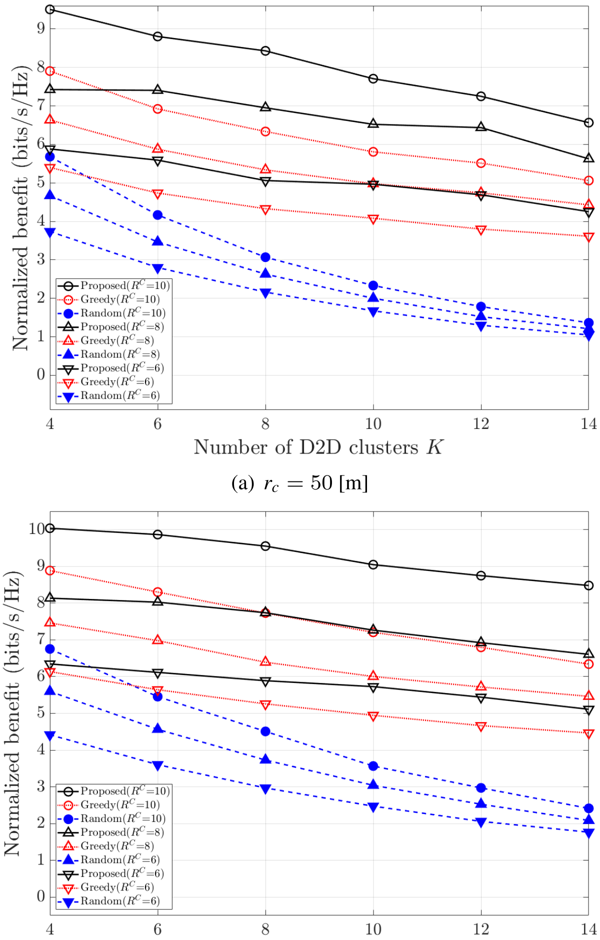
In Fig. 5 and Fig. 6, we plot the normalized benefits obtained by RA schemes under comparison with respect to the number of D2D clusters (K) existing in the cell, where various radii of cell (r_c) and target rates of cellular link (R^C) are considered. It is observed that the proposed DRL-based RA scheme results in higher benefit than other schemes in all situations. As the number of D2D clusters in the cell grows, all RA schemes show lower benefits due to a resulting severer mutual interference among UEs. However, the performance degradation of the proposed DRL-based RA scheme is less sensitive to the growth of K than other RA schemes. Thus, the performance gain of the proposed RA scheme over others becomes significant as the number of D2D clusters, or equivalently the number of DUEs, in the cell increases. It is also observed that the performance gain of the proposed DRL-based RA scheme over others increases as the radius of cell decreases. From these observations, it is obviously inferred that the proposed RA scheme is quite useful especially when UEs are distributed densely in a cell and suffer from a high level of mutual interference from other UEs. It is clear that the proposed RA scheme would be a good solution to resolve the spectrum shortage problem in the next-generation communication systems in which a high number of UEs are deployed densely over the cell.



(a) $R^C = 8$ [bits/s/Hz]

(b) $R^C = 10$ [bits/s/Hz]

FIGURE 5. Normalized benefit with respect to K obtained by applying RA schemes under comparison, where $N = 2$ and $r_C = 50, 100, 200$ [m] are considered.



(a) $r_c = 50$ [m]

(b) $r_c = 100$ [m]

FIGURE 6. Normalized benefit with respect to K obtained by applying RA schemes under comparison, where $N = 2$ and $R^C = 6, 8, 10$ [bits/s/Hz] are considered.

We can obtain higher benefit when the ratio of target rates for cellular link and D2D link (R^C/R^D) gets higher. The amount of improvement in benefit resulting from increasing R^C/R^D is more noticeable with the proposed RA scheme than with others. Note that high R^C/R^D implies a high priority of cellular link over D2D link because, in case of high R^C/R^D, improving quality of cellular link can improve the overall throughput more easily. Thus, we can claim that the performance gain of the proposed RA scheme over others is significant when the priority of cellular link over D2D link is high.

In case of high R^C/R^D, an efficient RA mechanism had better keep cellular effective throughput high at the cost of low D2D effective throughput. Fig. 7 shows the fraction of sum effective throughput of all D2D links over the total effective throughput of all UEs in the cell. The lower fraction of D2D effective throughput is observed for higher R^C/R^D with all RA schemes, which implies that D2D links sacrifice themselves to achieve overall high effective throughput of the cell. With the proposed RA scheme, the fraction of D2D effective throughput does not grow much as the number of D2D clusters increases, which differs from other schemes. It is inferred from this observation that the D2D links are controlled more efficiently by the proposed RA scheme to improve the overall performance than by other schemes. Fig. 8 shows the effect of the number of users or DRxs (N) participating in a NOMA-enabled D2D downlink transmission underlay cellular networks on the benefit. The benefit decreases as N grows no matter what RA scheme is used. This is obvious because a higher number of users sharing a given communication resource experiences a higher level of mutual interference and thus a degraded QoS for each user. It is notable that the proposed RA scheme results in a higher benefit than others irrespective of the number of users.

Fig. 9 shows the ratio of outage probabilities of individual user signals over the mean outage probability of all user signals. It is observed that outage probabilities of all individual D2D user signals transmitted in each D2D cluster

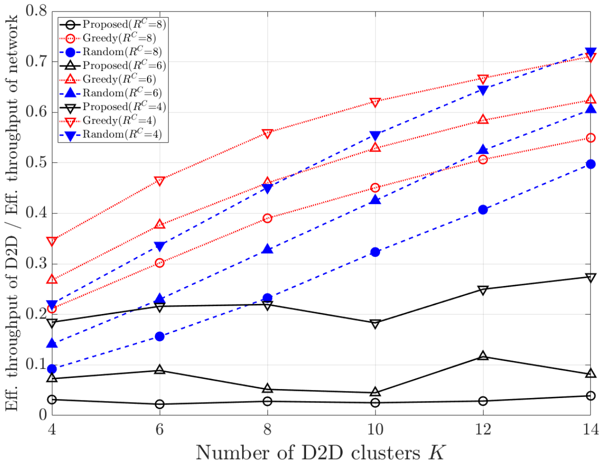


FIGURE 7. Fraction of effective throughput achieved by D2D links over the total effective throughput achieved by all cellular and D2D links, where r_c = 100 [m], N = 2 and R^C = 4, 6, 8 [bits/s/Hz] are considered.

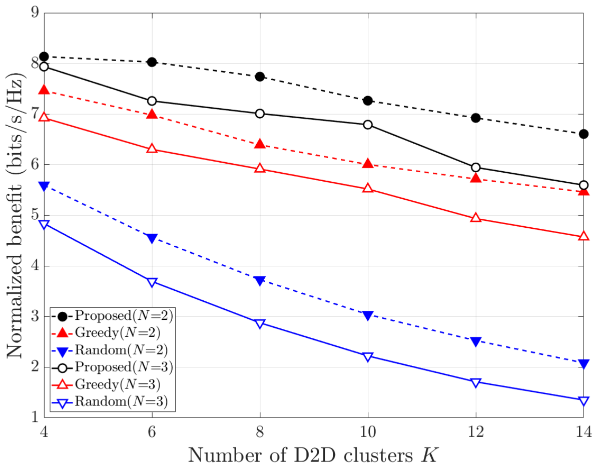


FIGURE 8. Normalized benefit with respect to K obtained by applying RA schemes under comparison, where r_c = 100 [m], R^C = 8 [bits/s/Hz] and N = 2, 3 are considered.

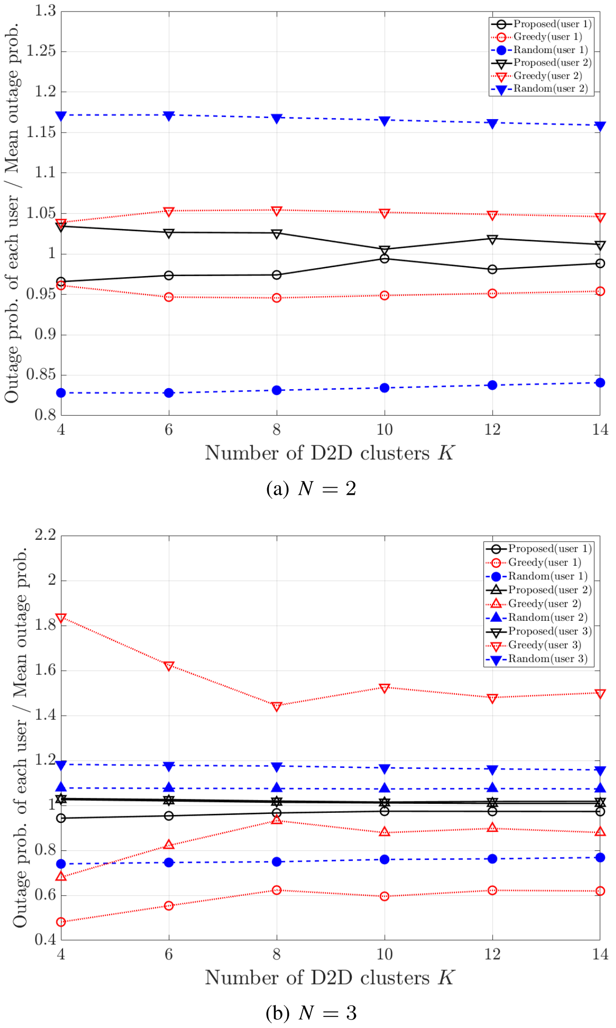


FIGURE 9. Fairness of NOMA-enabled D2D downlink communications based on RA schemes under comparison in terms of the ratio of outage probabilities of individual D2D user signals over the mean value, where r_c = 100 [m] and R^C = 8 [bits/s/Hz].

based on the proposed RA scheme do not deviate much from the mean value. Other RA schemes are observed to result in high deviations of individual outage probabilities from the mean value. This observation indicates that the proposed RA scheme achieves higher level of fairness among participating D2D users than other RA schemes. Users in each D2D cluster gain fair QoS through the proposed RA scheme and improve the overall benefit of the cell. Fairness is one of the most important criteria for designing NOMA-based communication system. Thus, the proposed DRL-based RA scheme is a suitable solution for the NOMA-enable D2D communication underlay cellular networks.

We compare the energy efficiency of RA schemes under comparison in terms of a benefit achieved by average transmit power of cellular and D2D links. It is observed from Fig. 10 that the proposed RA scheme results in the highest energy efficiency, It implies that the lowest average transmit power is required to obtain the same level of benefit. Thus, the proposed RA scheme is a reasonable solution for designing an energy efficient communication system.

We observe how evenly channels are allocated to D2D links. At every time step, we count the number of D2D links occupying each channel spectrum and sort these count numbers in a descending order. We repeat the same process multiple times to obtain a statistically meaningful result. Then, we add count numbers belonging to each order and compute the proportion of summed count numbers for each order over the total count number. We use bar graphs to show proportions of D2D clusters occupying the most crowded channel, the second most crowded channel, etc. as plotted in Fig. 11. It is observed that the proposed RA scheme and a random RA scheme result in even channel occupations by D2D links, where a random RA is inherently results in a uniform spectrum occupation. With a greedy RA scheme, on the other hand, a dominant channel occupation by D2D links is observed. It is inferred from this observation that

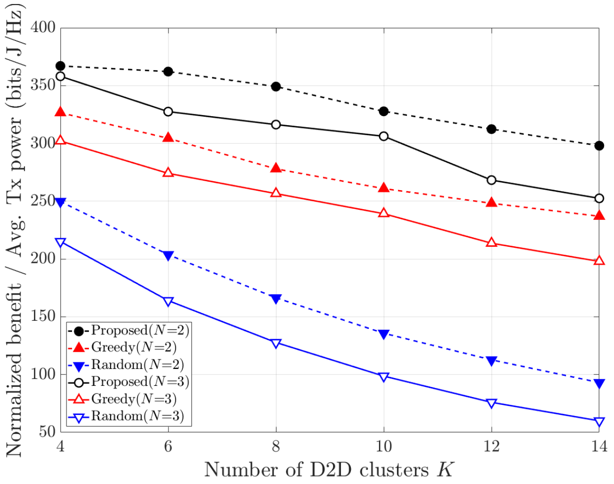


FIGURE 10. Energy efficiency obtained by RA schemes under comparison in terms of the normalized benefit per average transmit power of cellular and D2D links, where $r_c = 100$ [m], $R^C = 8$ [bits/s/Hz] and $N = 2, 3$ are considered.

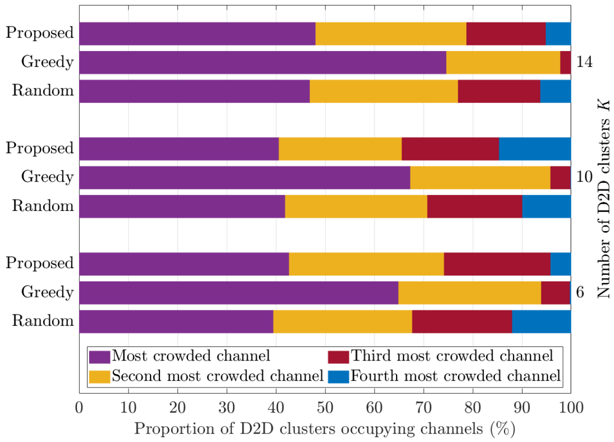
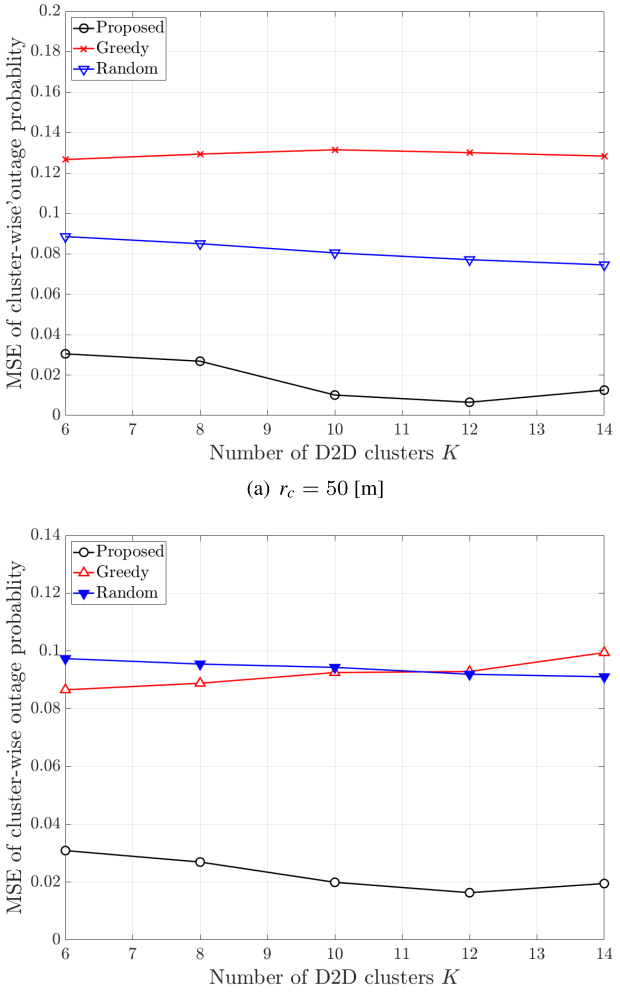


FIGURE 11. Evenness of channel occupation by D2D links resulting from applying RA schemes, where $r_c = 100$ [m], $R^C = 8$ [bits/s/Hz] and $N = 2$.

the proposed RA scheme utilizes the spectrum resources efficiently.

We observe how well coordinated participating D2D clusters are to obtain a high level of QoS for overall network. At each time step, we compute the cluster-wise outage probability of each D2D cluster, which is defined as the average outage probability of D2D receivers in each cluster. Then, we evaluate the MSE of cluster-wise outage probabilities after obtaining their mean value. We repeat the same procedure for multiple time steps and obtain the time-averaged MSE, which is plotted in Fig. 12. The lower is MSE, the better are D2D clusters are coordinated to achieve a good performance of the overall cell cooperatively. It is observed that the proposed RA scheme results in a well-coordinated operation of UEs.

We provide performance improvements in percentage (%) attained by using the proposed RA scheme over greedy and random RA schemes, whose summary is given in Table 4 and Table 5. When evaluating the percentage improvement in fairness presented in Fig. 9, we compute the gap between the highest and the lowest points for each N and K obtained by using the proposed, greedy and random RA schemes. Since a smaller gap indicates higher fairness, we evaluate the performance improvement by looking at the relative decrease of the gap. To evaluate the improvement in the evenness of channel occupation depicted in Fig. 11, we investigate the variance of the numbers of D2D clusters occupying each channel, where smaller variance indicates higher evenness. The coordination of D2D users is measured by MSE of cluster-wise outage probabilities of D2D clusters As observed from Table 4, the average amount of percentage improvement in benefit attained by the proposed RA scheme over a greedy RA scheme and a random RA scheme is 17.68% and 139.37%, respectively. The benefit improvement is significant when the cell radius is small and the number of D2D clusters in a cell is high. Thus, we can claim that the proposed RA scheme has a higher gain when the cell



(a) $r_c = 50$ [m]

(b) $r_c = 100$ [m]

FIGURE 12. MSE of outage probabilities of D2D clusters resulting from applying RA schemes, where $R^C = 8$ [bits/s/Hz] and $N = 2$.

TABLE 4. Average percentage improvement in the normalized benefit attained by the proposed RA scheme over greedy and random RA schemes with $N = 2$.

Reference	r_c [m]			R^C [bits/s/Hz]			K						Average
	50	100	200	6	8	10	4	6	8	10	12	14	
Greedy	24.63	17.96	6.92	14.80	16.79	20.94	8.56	15.46	19.04	21.27	21.59	20.48	17.68
Random	198.67	140.29	68.19	145.99	144.47	141.23	48.99	81.46	117.13	159.72	205.97	249.08	139.37

TABLE 5. Average percentage improvement in various performance measures attained by the proposed RA scheme over greedy and random RA schemes in case of $r_c = 100$ [m] and $R^C = 8$ [bits/s/Hz].

Reference	Performance Measure				
	Normalized benefit	Energy Efficiency	Fairness of D2D users	Coordination of D2D users(MSE)	Evenness of D2D channel occupation
Greedy	18.09	23.06	75.34	80.79	75.75
Random	152.45	164.22	87.07	77.58	-24.60

is more crowded with UEs. As observed in Table 5, the proposed RA scheme improves all performance measures over a greedy RA scheme. The proposed RA scheme results in a degraded evenness of D2D channel occupation over a random RA scheme. This is inevitable because a random scheme inherently aims to allocate channel spectra to D2D clusters uniformly.

In DRL-based RA scheme, ANNs are trained to find a globally optimal solution of a RA problem for an arbitrarily given input data. On the other hand, a greedy algorithm is used for each agent to find a locally optimal solution for an optimization problem with a given set of RA results of preceding agents. Thus, the proposed DRL-based RA scheme inherently outperforms a greedy RA scheme contingent on the high level of accuracy in training of ANNs. It is inferred from performance comparisons that ANNs of agents are well trained and thus the proposed RA scheme shows a better performance than greedy and random RA schemes.

We analyze computational complexities required for each agent to run the proposed RA scheme in terms of floating point operations (FLOPs), which is dominantly determined by the forward message propagation through ANN. Suppose ANN has an input layer with n_{in} nodes, N_H hidden layers each of which has n_h nodes and an output layer having n_{out} nodes. Then, a forward message propagation requires $2\left\{n_{in}n_h + (N_H - 1)n_h n_h + n_h n_{out}\right\} + N_H n_h$ FLOPs. Each input node of ANN corresponds to each entry of observation vector, i.e., locations of K DTxs, KN DRxs and M CUEs, transmit power for M CUEs and KN DRxs, and channel spectra of M CUEs and K D2D clusters. Thus, ANN has $n_{in} = 3M + 2KN + 2K$. The number of output nodes of ANN equals the number of possible sets of resource allocation vectors. For each agent, channel spectrum is chosen out of M candidates and transmit power level for each DRx is chosen out of L possible values. The number of power level selection is formulated as a combination with repetition so that $n_{out} = \frac{M(L+N-1)!}{(L-1)!N!}$. Consequently, the number of FLOPs required for each agent to run the proposed RA scheme is determined as $2\left\{n_h(3M + 2KN + 2K) + (N_H - 1)n_h^2 + n_h\frac{M(L+N-1)!}{(L-1)!N!}\right\} + N_H n_h$.

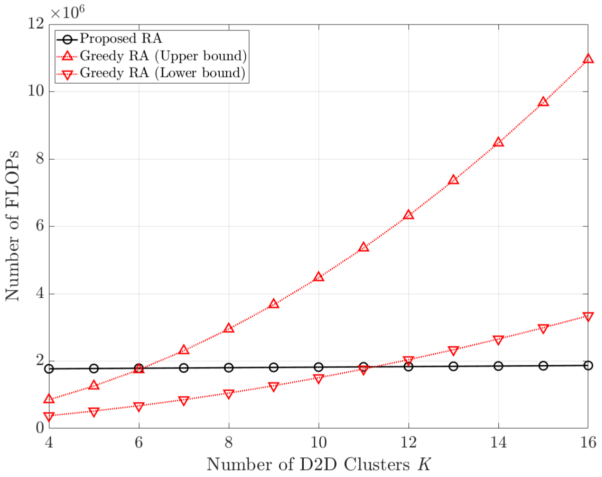


FIGURE 13. The number of FLOPs required for each D2D cluster to perform the proposed RA scheme and the greedy RA scheme, where $N = 3$, $L = 9$ and $M = 4$.

The computational complexity required for each agent to run a greedy RA scheme is dominantly determined by computation of objective function defined in (15). In this process, we compute $1 - p_{n,n(k)}^{out}$, $\forall n, k$ as well as $1 - p_m^{out}$, $\forall m$, for all resource allocation vectors. Let $\mathcal{K}_m$ denote the set of D2D clusters sharing the channel with a cellular link m, where $|\mathcal{K}_m|$ represents the cardinality of $\mathcal{K}_m$ and $\sum_{m=1}^{M} |\mathcal{K}_m| = K$. By tedious manipulations, we can obtain the total number of FLOPs required to compute $1 - p_{n,n(k)}^{out}$ for all $n \leq N$ in a cluster $k \in \mathcal{K}_m$ as $\frac{N(N+1)}{2}\left\{\frac{4(N-1)}{3} + 10|\mathcal{K}_m| + 8\right\}$. Thus, computing $\sum_{k=1}^{K}\sum_{n=1}^{N}(1 - p_{n,n(k)}^{out})$ requires $\frac{N(N+1)}{2}\left\{\frac{4K(N-1)}{3} + 8K + 10\sum_{m=1}^{M}|\mathcal{K}_m|^2\right\} + KN - 1$ FLOPs. In computing $1 - p_m^{out}$ for each m by (14), we require $8|\mathcal{K}_m| + 6$ FLOPs, and thus we require $8K + 7M - 1$ FLOPs to obtain $\sum_{m=1}^{M}(1 - p_m^{out})$. Considering all resource allocation vectors, the total number of FLOPs required for each agent to perform a greedy RA scheme by evaluating objective function is obtained as $\frac{M(L+N-1)!}{(L-1)!N!}\left\{\frac{N(N+1)}{2}\left(\frac{4K(N-1)}{3} + 8K + 10\sum_{m=1}^{M}|\mathcal{K}_m|^2\right) + KN + 8K + 7M + 4\right\}$ FLOPs. By using the property $\frac{K^2}{M} \leq \sum_{m=1}^{M}|\mathcal{K}_m|^2 \leq K^2$, we can obtain the lower and upper bounds on the number of FLOPs as well.

We plot the number of FLOPs required for the proposed RA scheme and the upper and lower bounds on the number of FLOPs required for greedy RA scheme in Fig. 13. The computational complexity of greedy RA scheme grows parabolically with respect to the number of D2D clusters. On the other hand, increasing the number of D2D clusters has a minor effect on FLOPs of the proposed RA scheme. The dominant factor determining FLOPs for the proposed scheme is the number of nodes in ANN. Consequently, the proposed RA scheme has a significant gain over the greedy scheme in terms of computational complexity as a higher number of D2D users are involved in D2D communications underlay cellular networks.

VI. CONCLUSION

We proposed a DRL-based RA scheme for D2D communications exploiting cluster-wise NOMA protocol underlay cellular networks, where transmit power and channel spectrum are considered communication resources to be allocated to D2D links. Multiple D2D links are allowed to utilize a common channel and the performance accumulated over multiple time steps are of concern, which results in a high computational complexity required for performing RA especially when a high number of D2D links are involved. In order to achieve high benefit with reduced complexity, we adopted a centralized RA scheme based on multi-agent DRL, each of which agent operates its own ANNs. Thanks to the segmented structure of ANNs, the proposed RA scheme requires reduced amount of computations compared to the RA scheme based on single-agent DRL. In actual operations, the proposed scheme allocates communication resources to D2D links adaptively in a real-time manner based on the observation for the updated network environment by using the pre-trained ANNs. We derived the outage probabilities of D2D links and cellular links analytically and provided a benefit as a performance measure of RA scheme. For this purpose, we investigated analytically the impact of successful and unsuccessful SIC for preceding D2D user signals on the outage probability of D2D links. It was observed that the proposed RA scheme outperforms others, especially when UEs are distributed densely with a high level of mutual interferences and the QoS of the cellular link has a higher priority than the D2D link. We also found that the proposed RA scheme is energy efficient and achieves a high level of fairness among D2D users in the cluster. The channel occupation of D2D users based on the proposed RA is observed to be even, which shows a well-coordinated operation of D2D users.

APPENDIX A
DERIVATION OF (10)

Consider i.i.d. exponential random variables $\alpha_0, \alpha_1, \cdots, \alpha_X$, whose probability density function (PDF) is given by $f_{\alpha_i}(a_i) = \lambda_i e^{-\lambda_i a_i}$ with mean λ_i^{-1} and variance λ_i^{-2}. Let us define γ as

$$\gamma = \frac{\alpha_0}{c\alpha_0 + \sum_{i=1}^{X} \alpha_i + b}, \tag{A1}$$

where c and b are positive constants. Then,

$$\begin{aligned}\Pr\{\gamma < r\} &= \Pr\left\{\frac{\alpha_0}{c\alpha_0 + \sum_{i=1}^{X} \alpha_i + b} < r\right\} \\ &= \Pr\left\{(1 - rc)\alpha_0 < r\left(\sum_{i=1}^{X} \alpha_i + b\right)\right\} \\ &= \Pr\left\{\alpha_0' < \sum_{i=1}^{X} \alpha_i + b\right\},\end{aligned} \tag{A2}$$

where $\alpha_0' = (1/r - c)\alpha_0$. If $1/r - c \leq 0$, then $\alpha_0' < \sum_{i=1}^{X} \alpha_i + b$ holds always so $\Pr\{\gamma \leq r\} = 1$. Otherwise,

$$\begin{aligned}&\Pr\{\gamma < r\} \\ &= \int_0^\infty \cdots \int_0^\infty \Pr\left\{\alpha_0 < \left(\frac{1}{r} - c\right)^{-1}\left(\sum_{i=1}^{X} \alpha_i + b\right)\right\} \\ &\quad \cdot \prod_{i=1}^{X} f_{\alpha_i}(a_i) da_i \\ &= \int_0^\infty \cdots \int_0^\infty \left(1 - e^{-\lambda_0'(\sum_{i=1}^{X} a_i + b)}\right) \prod_{i=1}^{X} f_{\alpha_i}(a_i) da_i \\ &= 1 - \int_0^\infty \cdots \int_0^\infty e^{-\lambda_0'(\sum_{i=1}^{X} a_i + b)} \prod_{i=1}^{X} \lambda_i e^{-\lambda_i a_i} da_i \\ &= 1 - e^{-\lambda_0' b} \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{X} \lambda_i e^{-(\lambda_i + \lambda_0')a_i} da_i \\ &= 1 - e^{-\lambda_0' b} \prod_{i=1}^{X} \int_0^\infty \lambda_i e^{-(\lambda_i + \lambda_0')a_i} da_i,\end{aligned} \tag{A3}$$

where $\lambda_0' = \lambda_0(1/r - c)^{-1}$. Since $\int_0^\infty \lambda_i e^{-(\lambda_i + \lambda_0')a_i} da_i = \frac{\lambda_i}{\lambda_i + \lambda_0'}$, we rewrite (A3) as

$$\begin{aligned}\Pr\{\gamma < r\} &= 1 - e^{-\lambda_0' b} \prod_{i=1}^{X} \frac{\lambda_i}{\lambda_0' + \lambda_i} \\ &= 1 - e^{-\frac{b}{\mu_0}(1/r - c)^{-1}} \prod_{i=1}^{X} \left(1 + \frac{\mu_i}{\mu_0}(1/r - c)^{-1}\right)^{-1},\end{aligned} \tag{A4}$$

where $\mu_i = \mathbb{E}\{\alpha_i\} = \lambda_i^{-1}$. Consequently,

$$\Pr\{\gamma < r\} = \begin{cases} 1 - e^{-\frac{b}{\mu_0}(1/r - c)^{-1}} \prod_{i=1}^{X} \left(1 + \frac{\mu_i}{\mu_0}(1/r - c)^{-1}\right)^{-1}, & \text{if } 1/r - c > 0 \\ 1, & \text{otherwise.} \end{cases} \tag{A5}$$

It is clear that (3) has the same form as (A1), where $\alpha_0 = P^D_{n'(k)}|h_{k,n(k)}|^2 d^{-\alpha}_{k,n(k)}$, $c = \sum_{i=n'+1}^{N} \frac{P^D_{i(k)}}{P^D_{n'(k)}}$, $\sum_{i=1}^{X} \alpha_i = \sum_{m=1}^{M} \delta_{m,k} P^C_m |h_{m,n(k)}|^2 d^{-\alpha}_{m,n(k)} + \sum_{j=1:j\neq k}^{K} \delta_{k,j} P^D_j |h_{j,n(k)}|^2 d^{-\alpha}_{j,n(k)}$ and $b = \sigma^2$. Since $\mathbb{E}\{|h_{x,y}|^2\} = 1$ for any node x and y, we have $\mu_0 = P^D_{n'(k)} d^{-\alpha}_{k,n(k)}$ and $\mu_i = \delta_{m,k} P^C_m d^{-\alpha}_{m,n(k)}$ or $\delta_{k,j} P^D_j d^{-\alpha}_{j,n(k)}$. Then, $p^{out}_{n',n(k)} = \Pr\{\gamma^D_{n',n(k)} < \gamma^D_{th}\}$ is obtained in a form of (A5). By using the property $(1+\delta x)^{-1} = (1+x)^{-\delta}$ with $\delta = 1$ or 0, and by replacing n' by l, we obtain $1 - p^{out}_{l,n(k)} = \Pr\{\gamma^D_{l,n(k)} \geq \gamma^D_{th}\}$ as provided in (12).

REFERENCES

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*, Cisco, San Jose, CA, USA, Jun. 2017.

[2] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 42–49, Dec. 2009.

[3] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Turányi, "Design aspects of network assisted device-to-device communications," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.

[4] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, G. Feng, and S. Li, "Device-to-device communications underlaying cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3541–3551, Aug. 2013.

[5] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[6] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, "Challenges and solutions for cellular based V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 222–255, 1st Quart., 2021.

[7] R. Yin, C. Zhong, G. Yu, Z. Zhang, K. K. Wong, and X. Chen, "Joint spectrum and power allocation for D2D communications underlaying cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2182–2195, Apr. 2016.

[8] J. Hu, W. Heng, X. Li, and J. Wu, "Energy-efficient resource reuse scheme for D2D communications underlaying cellular networks," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2097–2100, Sep. 2017.

[9] J. Lee and J. H. Lee, "Performance analysis and resource allocation for cooperative D2D communication in cellular networks with multiple D2D pairs," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 909–912, May 2019.

[10] T. Huynh, T. Onuma, K. Kuroda, M. Hasegawa, and W.-J. Hwang, "Joint downlink and uplink interference management for device to device communication underlaying cellular networks," *IEEE Access*, vol. 4, pp. 4420–4430, 2016.

[11] X. Ma, J. Liu, and H. Jiang, "Resource allocation for heterogeneous applications with device-to-device communication underlaying cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 15–26, Jan. 2016.

[12] J.-H. Kim, J. Joung, and J. W. Lee, "Resource allocation for multiple device-to-device cluster multicast communications underlay cellular networks," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 412–415, Feb. 2018.

[13] P. Mach, Z. Becvar, and M. Najla, "Resource allocation for D2D communication with multiple D2D pairs reusing multiple channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1008–1011, Aug. 2019.

[14] W. Zhao and S. Wang, "Resource allocation for device-to-device communication underlaying cellular networks: An alternating optimization method," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1398–1401, Aug. 2015.

[15] H. Wang, J. Wang, G. Ding, L. Wang, T. A. Tsiftsis, and P. K. Sharma, "Resource allocation for energy harvesting-powered D2D communication underlaying UAV-assisted networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 14–24, Mar. 2018.

[16] X. Li, L. Ma, Y. Xu, and R. Shankaran, "Resource allocation for D2D-based V2X communication with imperfect CSI," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3545–3558, Apr. 2020.

[17] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf.*, Jun. 2013, pp. 1–5.

[18] Q. Wang, R. Zhang, L.-L. Yang, and L. Hanzo, "Non-orthogonal multiple access: A unified perspective," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 10–16, Apr. 2018.

[19] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[20] Y. Iraqi and A. Al-Dweik, "Power allocation for reliable SIC detection of rectangular QAM-based NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8355–8360, Aug. 2021.

[21] Z. Ding, R. Schober, and H. V. Poor, "Unveiling the importance of SIC in NOMA systems—Part 1: State of the art and recent findings," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2373–2377, Nov. 2020.

[22] A. S. de Sena, F. R. M. Lima, D. B. da Costa, Z. Ding, P. H. J. Nardelli, U. S. Dias, and C. B. Papadias, "Massive MIMO-NOMA networks with imperfect SIC: Design and fairness enhancement," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 6100–6115, Sep. 2020.

[23] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[24] X. Wang, J. Wang, L. He, and J. Song, "Outage analysis for downlink NOMA with statistical channel state information," *IEEE Wireless Commun. Lett.*, vol. 7, no. 2, pp. 142–145, Apr. 2018.

[25] G. Im and J. H. Lee, "Outage probability for cooperative NOMA systems with imperfect SIC in cognitive radio networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 692–695, Apr. 2019.

[26] A. Agarwal, R. Chaurasiya, S. Rai, and A. K. Jagannatham, "Outage probability analysis for NOMA downlink and uplink communication systems with generalized fading channels," *IEEE Access*, vol. 8, pp. 220461–220481, 2020.

[27] L. Zhu, J. Zhang, Z. Xiao, X. Cao, and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 328–331, Apr. 2019.

[28] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[29] W. Yin, L. Xu, P. Wang, Y. Wang, Y. Yang, and T. Chai, "Joint device assignment and power allocation in multihoming heterogeneous multicarrier NOMA networks," *IEEE Syst. J.*, vol. 16, no. 1, pp. 671–682, Mar. 2022.

[30] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, "Joint subchannel and power allocation for NOMA enhanced D2D communications," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 5081–5094, Nov. 2017.

[31] I. Budhiraja, N. Kumar, and S. Tyagi, "ISHU: Interference reduction scheme for D2D mobile groups using uplink NOMA," *IEEE Trans. Mobile Comput.*, vol. 21, no. 9, pp. 3208–3224, Sep. 2022.

[32] M. Sun, X. Xu, X. Tao, P. Zhang, and V. C. M. Leung, "NOMA-based D2D-enabled traffic offloading for 5G and beyond networks employing licensed and unlicensed access," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4109–4124, Jun. 2020.

[33] S. Alemaishat, O. A. Saraereh, I. Khan, and B. J. Choi, "An efficient resource allocation algorithm for D2D communications based on NOMA," *IEEE Access*, vol. 7, pp. 120238–120247, 2019.

[34] G. Wu, G. Chen, and G. Chen, "Energy-efficient utility function-based channel resource allocation and power control for D2D clusters with NOMA enablement in cellular networks," *IEEE Access*, vol. 11, pp. 45001–45010, 2023.

[35] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, and A. Graves, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[36] H. Huang, Y. Yang, Z. Ding, H. Wang, H. Sari, and F. Adachi, "Deep learning-based sum data rate and energy efficiency optimization for MIMO-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5373–5388, Aug. 2020.

[37] H. Zhang, H. Zhang, K. Long, and G. K. Karagiannidis, "Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2406–2415, Oct. 2020.

[38] H. Xiang, Y. Yang, G. He, J. Huang, and D. He, "Multi-agent deep reinforcement learning-based power control and resource allocation for D2D communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1659–1663, Aug. 2022.

[39] S. Yu and J. W. Lee, "Deep reinforcement learning based resource allocation for D2D communications underlay cellular networks," *Sensors*, vol. 22, no. 23, p. 9459, Dec. 2022.

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

[41] 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.

[42] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

[43] Y.-H. Xu, C.-C. Yang, M. Hua, and W. Zhou, "Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications," *IEEE Access*, vol. 8, pp. 18797–18807, 2020.

[44] D. Ron and J.-R. Lee, "DRL-based sum-rate maximization in D2D communication underlaid uplink cellular networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11121–11126, Oct. 2021.

[45] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[46] D. Tse and P. Viswanath, "Capacity of wireless channels," in *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge Univ. Press, 2005, p. 187.

YUN JAE JEONG received the B.S. degree in electrical engineering from Chung-Ang University, Seoul, South Korea, in 2021, where he is currently pursuing the M.S. degree. His research interests include wireless communications and AI-based communications.

SEOYOUNG YU received the B.S. and M.S. degrees in electrical engineering from Chung-Ang University, Seoul, South Korea, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree. His research interests include wireless communications and AI-aided communications.

JEONG WOO LEE (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1994 and 1996, respectively, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2003.

From 2003 to 2004, he was a Postdoctoral Research Associate with the Coordinated Science Laboratory, Urbana, IL, USA. Since 2004, he has been a Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul. From 2017 to 2018, he was a Visiting Scholar with UC San Diego, San Diego, CA, USA. His research interests include MIMO systems, wireless communications, coding and information theory, and AI-aided communications. From 2016 to 2020, he was an Associate Editor of *IET Electronics Letters*. He is currently the Guest Editor of *Sensors*. From 2019 to 2022, he was the Director of IEIE. From 2021 to 2022, he served as a Treasurer of the IEEE Seoul Section.