IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 December 2022, accepted 27 December 2022, date of publication 29 December 2022,
date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3233028

RESEARCH ARTICLE

Optimal Power Allocation With Multiple Joint Associations in Multi-User MIMO Full-Duplex Systems

KUNBEI PAN[1,2,3], BIN ZHOU[1,3], AND ZHIYONG BU[1,3]

[1]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Wireless Sensor Network and Communications, Chinese Academy of Sciences, Shanghai 200050, China

Corresponding author: Bin Zhou (bin.zhou@mail.sim.ac.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1803101.

ABSTRACT Optimum power allocation is an effective way to mitigate residual self-interference and inter-user interference in multiple input multiple output full-duplex (FD) systems. However, current research mainly considers parts of influencing factors and sets service models fixed. Given this, we comprehensively focus on three perspectives in a novel power allocation method, which involve the muting management (MM) and the assignment of both base station antennas and subcarriers in the FD system. Then, we formulate an optimization problem to maximize the total spectrum efficiency. According to the categories of variables in the nonconvex objective function, we first propose a hierarchical algorithm, which is decomposed into the first-order Taylor approximation (FOTA) method and the greedy algorithm. The continuous and discrete variables related subproblems are solved through FOTA and greedy algorithm, respectively. Among them, the greedy algorithm is an alternative to a traditional method of exhaustive search. Considering the high complexity of the greedy algorithm, we further introduce deep reinforcement learning (DRL) instead to solve the corresponding subproblem. Thus, two Double Deep Q-learning Networks are constructed to train the samples in each sub-slot. Simulation results validate that the hybrid DRL-convex method outperforms the hybrid greedy-convex method. Meanwhile, the MM introduced scheme's performance gain is more evident than that of the method without MM in many scenarios.

INDEX TERMS Full-duplex (FD), power allocation, spectrum efficiency (SE), smart antennas, muting management (MM), subcarrier assignment, deep reinforcement learning (DRL), jamming user (JU).

I. INTRODUCTION

With the development of big data, blockchain, stream media, and others, related businesses have been incorporated into the enriched public life. According to the International Telecommunication Union (ITU) forecast, global mobile data will increase by about 55% annually from 2020 to 2030 and reach 5,000 Ebits in 2030 [1]. Although spectrum efficiency (SE) for conventional communication technologies has approximated the Shannon capacity bottleneck, the higher SE requirement for future wireless communications has been put forward. Compared with traditional half-duplex (HD) technology, full-duplex (FD) (i.e., co-time co-frequency full-duplex in-band) technology can significantly improve SE [2]. Because of this, FD technology is regarded as one of the candidates for enhanced wireless air interface technologies of 6G [3], which has a broad prospect. Nevertheless, it is beset with severe self-interference (SI) (i.e., interference from its emitting to receiving antennas). If the SI is not addressed appropriately, the expected receiving signal will result in decoded failure due to it being submerged in the transmitting signal [4].

Currently, there are active and passive self-interference cancellation (SIC) technologies [2], but partial SI remains

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

because of inaccurate SI channel estimation and hardware impairment [5]. Moreover, residual SI is further aggravated caused by the extensively used multiple input multiple output (MIMO) technology, which would make residual SI compound and challenging to eliminate [6]. When multiple users are located in the cellular network, interference from user to base station (BS) and that from user to user also exist. All the aforementioned interference together deteriorates the FD performance.

As is known, lifting power could increase capacity but decrease it in turn due to intensifying interference. The revenue depends on the weights of each desired signal and interference, which is a typical allocation problem. Therefore, appropriate power allocation can effectively address the tradeoff between performance gain and loss, and is mainly fulfilled through an objective function (e.g., maximized SE, energy efficiency (EE)). In view of this, many scholars focus on power allocation methods to alleviate the FD performance reduction caused by multiform interference [7]. More importantly, the diversity gain can be improved with the application of smart antennas rather than deploying the traditional fixed antennas [8]. With the aid of smart antennas, the FD technique is further enhanced. Considering that spectrum resource scarcity is an international problem [1], for this reason, implementing power allocation along with FD technique in a limited spectrum resource is very meaningful in the current situation.

A. MOTIVATION

Despite of aforementioned benefits provided by the power allocation method in the FD systems, current research seldomly considers both smart antennas and spectrum resource scarcity in a power allocation method [9], [10], [11], [12], [13], [14], [15]. Meanwhile, other scholars set the precondition of users' service models fixed to optimize the objective function, which may not be an overall optimum [16]. Inspired by this, we propose a power allocation method with multiple joint associations (i.e., smart antennas, scheduled users, and subcarriers) to improve FD performance from each aspect. To be specific, the layer of smart antennas works on lifting diversity gain. The muting management (MM) for scheduled users is to restrain interference caused by parts of users, which generate more disturbances than others. The rational subcarrier assignment aims to reduce competition in spectrum resources. All three elements are bound up with SE performance. In our work, the MM is realized through a newly designed frame structure, which considers both muting and compensation for a small group of users. This field is different from the previous work. Additionally, we integrate the assignment of antennas and subcarriers in power allocation and work the joint optimization problem out through a hierarchical algorithm, which is another new scheme to the existing works.

B. MAIN CONTRIBUTIONS

In this paper, we review the current investigations about power allocation in FD systems. Accordingly, we propose a novel power allocation method. The main contributions of our work are summarized as follows:

1. Considering the different types of interference, we divide the scheduled users by service type. To realize service enabled from the user level, we define the user identifier and devise MM by adding a trigger region and a muting indication in sub-slot 1.
2. To integrate the affecting factors that involve the MM and assignment of antennas and subcarriers, we formulate an objective function of the power allocation method, which considers the above three elements to optimize the overall SE fully.
3. The proposed optimization problem is decomposed into two subproblems in terms of continuous and discrete variables. With the first-order Taylor approximation (FOTA) method, the continuous part is converted into convexity. Then we employ the greedy algorithm based on the traditional method (exhaustive search) to tackle the discrete part as a benchmark.
4. Because deep reinforcement learning (DRL) is more appropriate for solving nonconvex problems of discrete variables, we design another hybrid method based on two Double Deep Q-learning Networks (DDQNs) instead of the greedy algorithm. Simulations demonstrate that the hybrid DRL-convex method outperforms the hybrid greedy-convex method. Also, our proposal with MM achieves performance enhancement in comparison to that without MM.

The remainder of this work is organized as follows. The related works are presented in Section II. Our system model, followed by MM and uplink (UL)/downlink (DL) interference model in each sub-slot, is described in Section III. According to the system model, we formulate the optimization problem for maximizing SE in Section IV. Section V presents two proposed hierarchical algorithms to tackle the nonconvex problem, and detailed complexity analysis is presented. In Section VI, numerical results demonstrate our proposal. Section VII remarks on the conclusion of our work.

II. RELATED WORKS

Based on existing SIC technology, optimizing the formulated objective function in a power allocation method is the mainstream for improving the FD system's performance recently.

Some scholars design optimized power allocation methods in MIMO FD relaying. In [17], to satisfy the requirements of each user's signal to interference plus noise ratio (SINR), along with saving the allocated power at the FD relay, the authors formulate an EE-optimization problem in a block fading channel and work the issue out through the geometric programming method. Based on [17], the authors in [18] mainly focus on the relationship between antenna number and SE/EE. The optimal number of antennas to maximize the objective function is derived. They demonstrate that a performance bottleneck confines the FD antenna scale due to the distortion noise.

Moreover, other scholars apply the power control method in FD cognitive radio networks to further promote the degrees of freedom in the MIMO system. The authors in [19] put forward four cognitive radio modes, where the secondary users adopt different power strategies. Additionally, the authors make a performance contrast between FD and HD in four schemes, respectively.

The above authors in [17], [18], and [19] set the FD antennas' operational modes fixed, and other scholars consequently research adaptive FD antennas in power allocation. Unlike [18], the authors in [9] give an optimum ratio between emitting and receiving antennas instead of equal numbers to maximize the sum rate in the power allocation method. Some scholars study power allocation with flexible antennas under secure transmission in FD systems. In [10], the antenna selection coefficient has been put forward to regulate the number of emitting/receiving antennas in the FD system. Then a power allocation method based on quantum calculation is applied to maximize the security capability and EE.

The authors in [9] and [10] regard the emitting/receiving antennas as a group, while others in [11] and [12] treat each FD antenna as an individual. In [11], the authors propose a scheme that can dynamically select each emitting/receiving antenna according to various channel conditions, thus raising the FD diversity gain for SE enhancement in power allocation. To further research the diversity gain of FD, the authors in [12] introduce a binary matrix to define the operating modes of each antenna. Using the assignment matrix, they construct a two-stage SE objective function, which is solved through successive convex approximation.

All the above studies have not considered the FD network with spectrum resource intensive. Some scholars take bandwidth or subcarrier as a power allocation factor. The authors in [13] adopt a three-stage Stackelberg game in power allocation, which takes bandwidth and EE as pricing and utility, respectively. They attempt to acquire the optimum utility value through the game. In [14], the authors present auxiliary variables and penalty factors to handle the discrete subcarrier assignment variables. With the problem reconstruction, the optimization solution of EE in power allocation has been acquired through the Lagrange method. In [15], the authors propose a power allocation method based on successive convex approximation in the FD distributed antenna system. The system includes several user-centric virtual cells that share limited subcarriers.

The sequential decision problem is known to be solved by reinforcement learning (RL) [20]. Since RL can find an appropriate compromise between performance and complexity in the case of massive samples [21], it has attracted tremendous attention from academia. Therefore, many scholars have attempted to solve the power allocation problem with RL [22], [23], [24], which has also been used in FD systems recently [25], [26], [27], [28], [29], [30]. To name a few, in [25], based on the underlay mode referring to [19], the authors employ DRL for power control, which increases the secondary user's SINR by improving its perception accuracy.

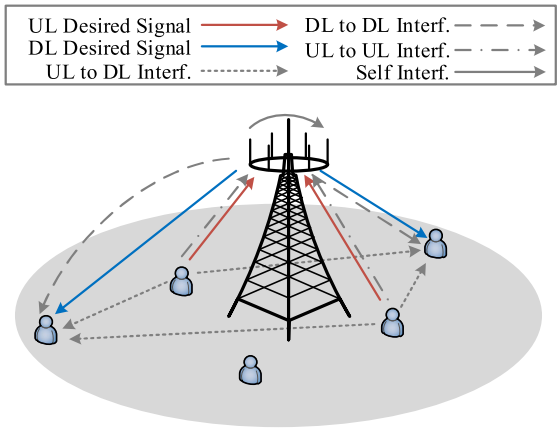

FIGURE 1. Example of a multi-user cellular network with four scheduled users (i.e., two UL and two DL users) and one unscheduled user (i.e., non-service user) at the moment.

The times of satisfaction for the capacity requirement at both primary and secondary users are defined as rewards, which can be maximized through a training process. The authors in [26] focus on a pair of terminals with FD capability. By setting applicable states, actions, and rewards, they propose the hybrid RL scheme to maximize the sum of SE and energy transmission efficiency. Meanwhile, the influence of different antenna numbers and power budgets on performance is also studied. In [27], the authors adopt RL in an unmanned aerial vehicle FD relay scenario to maximize secrecy capacity. At the same time, different RL techniques are compared in terms of secrecy rate and convergence.

In view of the above research, the researchers in [9], [10], [11], and [12] aim to improve the spatial diversity gain in power allocation but do not consider the case that both emitting and receiving antennas are shared. Also, subcarrier assignment in the objective function is not involved concurrently. Although authors in [13], [14], and [15] consider subcarrier assignment, the FD antennas are invariable. To the best of our knowledge, the joint optimization problem of FD antenna and subcarrier assignment in power allocation has not been investigated integrally. Also, investigations in [9], [10], [11], [14], and [15] set users' service models fixed, ignoring the service of muting. In response to this, we put forward our proposal. Meanwhile, considering the advantage of RL, we adopt DRL to enhance the algorithm in the FD system.

III. SYSTEM MODEL

Fig. 1 depicts a BS working in FD mode, equipped with N smart antennas in the cell. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the set of BS antennas. Each antenna is connected with an analog circulator device to isolate radio UL and DL. As a result, the operational mode of single transmitting, single receiving, or co-transmitting co-receiving in the same band can be selected [31], [32].

We suppose that Z users are uniformly distributed in the cellular network. The set of users represented by

$\mathcal{Z} = \{1, 2, \ldots, Z\}$ is classified into the subsets of UL users, DL users, and non-service users at the moment. Each user is equipped with one antenna and can transmit or receive data at a different time due to working in HD mode. The network spectrum resources are divided into M mutually orthogonal subcarriers, the set of which is denoted as $\mathcal{M} = \{1, 2, \ldots, M\}$. We assume that scheduled users in different subcarriers do not interfere with each other. As the number of scheduled users (i.e., service users) is larger than the number of subcarriers, scheduled users reuse part of the subcarriers, which incurs interference. In view of this, we will describe MM in the following subsection.

A. MUTING MANAGEMENT

The authors in [16] propose a concept of Interference Aware Muting that forces the mobile terminal to turn off due to causing severe interference to BS. We call such users the jamming users (JUs). As shown in Fig. 1, if one UL and one DL user are deemed as JUs, the UL JU would aggravate UL to DL and UL to UL interference, while DL JU exacerbates SI and DL to DL interference. We can observe that muting the JUs is a tractable and explicit strategy. However, the muting process causes a service interruption to the JU that suffers a performance loss. In order to minimize the side effect of muting, the muted JUs will return to regular service at the next time slot. That is, the muting orders will be invalid at the subsequent slot until the new arrival of orders. Under the above operations, albeit with performance partially reduced from outages, we still attempt to reach a state where the advantage outweighs its drawback compared with unmuted before. Motivated by this, we introduce MM into the system model.

Because the service type of users depends on the service scheduled from BS, we assume that each time slot contains a control region and a data region for simplicity [33]. The schedule information, which determines the service type, is monitored in the control region by a user. When a user has detected a downlink control information that is relevant to UL or DL schedule information, the data will be transferred during the related data region. The affiliated data region is subject to the specific frame pattern that BS has configured [34]. To better evaluate the proposed MM, we combine two consecutive time slots (called sub-slots 1 and 2) into one schedule unit, in which the schedule information for the two time slots keeps the same to ensure users' service continuity for a while. Sub-slot 1 is added with a trigger region and a muting indication based on the primitive frame structure, as shown in Fig. 2, where sub-slot 2 is the default. The intention of this configuration is that we expect to keep the minimal possible change in order to ensure compatibility. BS judges the service muting decisions through power policy adjustment in the trigger region, and the muting indication bears the muting order that delivers to related scheduled users. During sub-slot 2, the muting order will not work so that the silenced users can restore the service. Meanwhile, the appropriate compensation should be considered at sub-slot 2 in terms of fairness. The offset process is unrelated to muting order as long as sub-slot 2 has acquired the muted users information. It is evident that the operation in sub-slot 2 is aligned with the design framework.

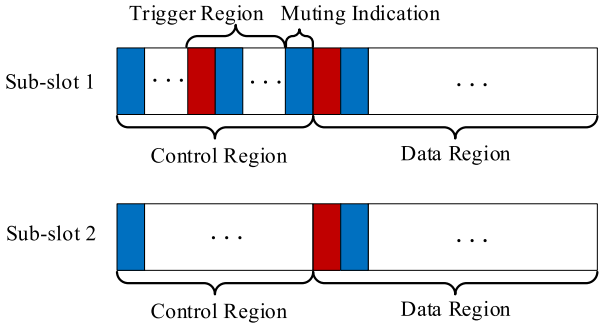

FIGURE 2. Frame structure of sub-slots 1 and 2 for each user. The dark-blue and red-brown frames stand for DL and UL subframes, respectively. The ellipsis indicates the specific frame pattern, which is not our focal point in this study.

To manage muting from the user level, we define service identifiers for each user in the cell.

For instance, $\mathcal{I}(\tau_\kappa) = \{\alpha_1^{\mathrm{u}}(\tau_\kappa), \alpha_1^{\mathrm{d}}(\tau_\kappa), \alpha_2^{\mathrm{u}}(\tau_\kappa), \alpha_2^{\mathrm{d}}(\tau_\kappa), \ldots, \alpha_Z^{\mathrm{u}}(\tau_\kappa), \alpha_Z^{\mathrm{d}}(\tau_\kappa)\}$ is denoted as service identifiers for Z users at κ^{th} schedule unit, where $\tau_\kappa \in \{t_\kappa, t_\kappa + \Delta t\}$. Δt is the length of a sub-slot, and $\tau_\kappa = t_\kappa$ or $t_\kappa + \Delta t$ means the sub-slot 1 or 2 in the κ^{th} unit. $\alpha_z^{\mathrm{u}}(\tau_\kappa)$ and $\alpha_z^{\mathrm{d}}(\tau_\kappa)$ signifies the UL and DL service identifier for user z at the related sub-slot, respectively. For ease of writing, sub-slot 1 or 2 at the κ^{th} unit is recorded as $t_{\kappa,1}$ or $t_{\kappa,2}$. In this paper, we mainly analyze one unit, so we abbreviate $t_{\kappa,1}$ and $t_{\kappa,2}$ to t_1 and t_2, respectively.

In conclusion, identifiers for user z at sub-slots 1 and 2 can be expressed as

$$\alpha_z^{\chi}(t_1) = \begin{cases} 1, & \mathrm{Sch}_z^{\chi}(t_1) = \mathrm{Mu}_z(t_1) = 1, \\ 2, & \mathrm{Sch}_z^{\chi}(t_1) = 1,\ \mathrm{Mu}_z(t_1) = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{1a}$$

$$\alpha_z^{\chi}(t_2) = \begin{cases} 2, & \mathrm{Sch}_z^{\chi}(t_2) = \mathrm{Sch}_z^{\chi}(t_1) = 1, \\ 0, & \mathrm{Sch}_z^{\chi}(t_2) = \mathrm{Sch}_z^{\chi}(t_1) = 0, \end{cases} \tag{1b}$$

in which $\chi \in \{\mathrm{u}, \mathrm{d}\}$ expresses the service type of UL or DL. Mu and Sch are short for Muting and Schedule, respectively. Since users work in HD mode, the muting indicator is not attentive to the specific service type. $\mathrm{Mu}_z(t_1) = 1$ or 0 denotes that the muting order has delivered to user z or not at sub-slot 1. Similarly, $\mathrm{Sch}_z^{\chi}(t_{1\mathrm{or}2}) = 1$ or 0 indicates that user z concerning service type χ is scheduled or not at the whole unit. Notably, users can not be scheduled for two types of service simultaneously due to operation under HD mode.

To sum up, $\alpha_z^{\chi}(t_1) = 1$ or 2 means that the scheduled users have been silenced or not at sub-slot 1. $\alpha_z^{\chi}(t_2) = 2$ guarantees the continuity of the same schedule information in a unit. In addition, $\alpha_z^{\chi}(t_{1\mathrm{or}\,2}) = 0$ indicates that the user is not scheduled at the unit.

As the above discussions, the scheduled users can be mathematically categorized into two types. One is a collective of UL users, and the other is a set of DL users, denoted by

$$\mathcal{J}=\{\alpha_z^{\mathrm{u}}(t_1)=1,2,\alpha_z^{\mathrm{u}}(t_2)=2;z\in\mathcal{Z}\}=\{1,2,\ldots,J\} \quad (2)$$

and

$$\mathcal{K}=\{\alpha_z^{\mathrm{d}}(t_1)=1,2,\alpha_z^{\mathrm{d}}(t_2)=2;z\in\mathcal{Z}\}=\{1,2,\ldots,K\}, \quad (3)$$

respectively.

To be specific, we use $\mathcal{U}$ and $\mathcal{D}$ (set of muted UL and DL users at sub-slot 1) to represent $\{\alpha_z^{\mathrm{u}}(t_1)=1,\alpha_z^{\mathrm{u}}(t_2)=2;z\in\mathcal{Z}\}$ and $\{\alpha_z^{\mathrm{d}}(t_1)=1,\alpha_z^{\mathrm{d}}(t_2)=2;z\in\mathcal{Z}\}$ in $\mathcal{J}$ and $\mathcal{K}$, respectively.

Let $\mathcal{G}$ denote the set of scheduled users, which satisfies $\mathcal{G}=\mathcal{J}\cup\mathcal{K}=\{1,2,\ldots,G\}$. Note that we only consider users in $\mathcal{G}$ of the interference model below.

B. INTERFERENCE MODEL

In this paper, we apply a composite fading channel and can acquire complete channel state information (CSI) [35], [36]. Considering the channel's frequency characteristic, we suppose that the CSI between two nodes in one schedule unit will remain unchanged [10]. Accordingly, the difference in transmission between two sub-slots pivots on MM.

1) TRANSMISSION AT SUB-SLOT 2

We first construct mathematical modeling at sub-slot 2 for ease of analysis because no MM is applied.

For the modeling's sake, we initially assume that all users in $\mathcal{G}$ share the same subcarrier, and BS fixes N_t emitting and N_r receiving antennas, satisfying $N_t=N_r=N$.

The signal received at BS from user j at sub-slot 2 can be written as

$$\begin{aligned}\boldsymbol{y}_j^{\mathrm{u}}(t_2)=&\boldsymbol{h}_j^{\mathrm{u}}(t_2)\sqrt{p_j(t_2)}x_j^{\mathrm{u}}(t_2)\\&+\sum_{j'\in\mathcal{J},j'\neq j}\boldsymbol{h}_{j'}^{\mathrm{u}}(t_2)\sqrt{p_{j'}(t_2)}x_{j'}^{\mathrm{u}}(t_2)\\&+\sum_{k\in\mathcal{K}}(\boldsymbol{w}_k(t_2))^H\mathbf{H}_{\mathrm{SI}}(t_2)x_k^{\mathrm{d}}(t_2)+\boldsymbol{n}_j^{\mathrm{u}}(t_2),\end{aligned} \quad (4)$$

where

$$\boldsymbol{h}_j^{\mathrm{u}}(t_2)=\sqrt{\boldsymbol{d}_j^{-\gamma}\circ\boldsymbol{\mathfrak{a}}_j(t_2)}\circ\boldsymbol{\mathfrak{w}}_j(t_2)\in\mathbb{C}^{1\times N} \quad (5)$$

represents the channel vector from user j to BS. $\boldsymbol{d}_j\in\mathbb{R}^{1\times N}$ denotes the distance vector from user j to each BS antenna. $\boldsymbol{\mathfrak{a}}_j(t_2)\in\mathbb{R}^{1\times N}$ and $\boldsymbol{\mathfrak{w}}_j(t_2)\in\mathbb{C}^{1\times N}$ indicate the lognormal shadow fading and small-scale fading vector, respectively. Both $\boldsymbol{\mathfrak{a}}_j$ and $\boldsymbol{\mathfrak{w}}_j$ obey independently identical distribution as $\boldsymbol{\mathfrak{a}}_j,\boldsymbol{\mathfrak{w}}_j\sim\mathcal{CN}(0,\mathbf{1}_{1\times N})$ [15], where $\mathbf{1}_{1\times N}$ stands for $1\times N$ dimensional vector with elements all 1.

The first term of (4) implies the desired signal. The second term signifies the interference caused by other UL users except user j (namely, UL to UL interference), and the third term indicates the residual SI, which has been mitigated by DL precoding (i.e., DL power allocation).

$p_j(t_2)$ in the first term represents the transmitted power of user j, which satisfies

$$p_j(t_2)\in\mathcal{P}(t_2)=\{p_1(t_2),p_2(t_2),\ldots,p_J(t_2)\}, \quad (6)$$

where $\mathcal{P}(t_2)$ is a set of transmitted powers for all UL users. Meanwhile, $x_j^{\mathrm{u}}(t_2)$ stands for the transmission symbol from user j, which follows $\mathbb{E}[x_j^{\mathrm{u}}(t_2)(x_j^{\mathrm{u}}(t_2))^*]=1$.

$\mathbf{H}_{\mathrm{SI}}(t_2)\in\mathbb{C}^{N\times N}$ in the third term is the residual SI matrix and follows

$$\mathbf{H}_{\mathrm{SI}}\sim\mathcal{CN}\left(\sqrt{a\sigma_{\mathrm{SI}}^2/(a+1)}\mathbf{1}_{N\times N},\mathbf{I}_N\sigma_{\mathrm{SI}}^2/(a+1)\right), \quad (7)$$

where a is the Rician factor and σ_{SI}^2 is the SI power ratio of pre-SIC to post-SIC [35]. Additionally, $x_k^{\mathrm{d}}(t_2)$ denotes the received symbol of user k from BS, which also satisfies $\mathbb{E}[x_k^{\mathrm{d}}(t_2)(x_k^{\mathrm{d}}(t_2))^*]=1$.

DL precoding vector for user k is expressed as $\boldsymbol{w}_k(t_2)\in\mathbb{C}^{N\times 1}$, which satisfies

$$\boldsymbol{w}_k(t_2)\in\mathcal{W}(t_2)=\{\boldsymbol{w}_1(t_2),\boldsymbol{w}_2(t_2),\ldots,\boldsymbol{w}_K(t_2)\}. \quad (8)$$

$\mathcal{W}(t_2)$ means the set of DL precoding vectors of all DL users.

The last term $\boldsymbol{n}_j^{\mathrm{u}}(t_2)\in\mathbb{C}^{N\times 1}\sim\mathcal{CN}(0,\sigma_{\mathrm{u},j}^2\mathbf{I}_N)$ indicates additive white gaussian noise (AWGN) vector related to user j at BS.

The signal received at user k from BS at sub-slot 2 is similarly expressed as

$$\begin{aligned}y_k^{\mathrm{d}}(t_2)=&\left(\boldsymbol{h}_k^{\mathrm{d}}(t_2)\right)^H\boldsymbol{w}_k(t_2)x_k^{\mathrm{d}}(t_2)\\&+\sum_{k'\in\mathcal{K},k'\neq k}\left(\boldsymbol{h}_k^{\mathrm{d}}(t_2)\right)^H\boldsymbol{w}_{k'}(t_2)x_{k'}^{\mathrm{d}}(t_2)\\&+\sum_{j\in\mathcal{J}}g_{k,j}(t_2)\sqrt{p_j(t_2)}x_j^{\mathrm{u}}(t_2)+n_k^{\mathrm{d}}(t_2),\end{aligned} \quad (9)$$

where $\boldsymbol{h}_k^{\mathrm{d}}(t_2)\in\mathbb{C}^{N\times 1}$ denotes the channel vector from BS to user k.

Similar to (4), the first term in (9) represents the expected signal. The second term indicates the interference caused by receiving other DL users' signals, which is the aforementioned DL to DL interference. The third term means user k is interfered with by UL users, namely, the UL to DL interference. In the third term, $g_{k,j}(t_2)$ represents channel gain from UL user j to DL user k. The final term $n_k^{\mathrm{d}}(t_2)\sim\mathcal{CN}(0,\sigma_{\mathrm{d},k}^2)$ stands for AWGN at user k.

According to (4) and (9), the target UL or DL user signal mingles with different categories of interference, as shown in Fig. 1, thus inducing undesirable channel conditions. It results in an optimization bottleneck of FD performance for the case of all BS radiating/receiving antennas inflexible [32]. Hence, we refine each BS antenna's working mode, which covers reception/transmission independence mode and coexistence mode.

The smart antennas are modeled with an assignment vector $\mathbf{Q} = [\mathbf{q}^{\mathrm{u}}, \mathbf{q}^{\mathrm{d}}]$, in which $\mathbf{q}^{\mathrm{u}}$ and $\mathbf{q}^{\mathrm{d}}$ are subvectors of receiving and emitting antennas, respectively. The subvectors at sub-slot 2 is written as

$$\mathbf{q}^{\mathrm{u}}(t_2) = \left[q_1^{\mathrm{u}}(t_2), q_2^{\mathrm{u}}(t_2), \ldots, q_N^{\mathrm{u}}(t_2)\right], \tag{10a}$$

$$\mathbf{q}^{\mathrm{d}}(t_2) = \left[q_1^{\mathrm{d}}(t_2), q_2^{\mathrm{d}}(t_2), \ldots, q_N^{\mathrm{d}}(t_2)\right], \tag{10b}$$

where $q_l^{\chi}(t_2)$ is the state of antenna l ($\forall l \in \mathcal{N}$), defined as

$$q_l^{\chi}(t_2) = \begin{cases} 1, & \text{antenna } l \text{ is used for service } \chi, \\ 0, & \text{antenna } l \text{ is not used for service } \chi. \end{cases} \tag{11}$$

Accordingly, the vector $\mathbf{Q}$ acts on the channel vector as the following (12) and (13) to realize the adaptive antennas.

Moreover, the interference model only applies to the users that share the same subcarrier. We reconstruct the UL/DL interference model by assigning BS antennas and subcarriers. So (4) and (9) can be rewritten as

$$\begin{aligned} &\boldsymbol{y}_{j,m}^{\mathrm{u}}(t_2) \\ &= \mathbf{q}^{\mathrm{u}}(t_2) \circ \boldsymbol{h}_j^{\mathrm{u}}(t_2)\sqrt{p_j(t_2)}x_j^{\mathrm{u}}(t_2)b_{j,m}^{\mathrm{u}}(t_2) \\ &\quad + \sum_{j' \in \mathcal{J}, j' \neq j} \mathbf{q}^{\mathrm{u}}(t_2) \circ \boldsymbol{h}_{j'}^{\mathrm{u}}(t_2)\sqrt{p_{j'}(t_2)}x_{j'}^{\mathrm{u}}(t_2)b_{j',m}^{\mathrm{u}}(t_2)b_{j,m}^{\mathrm{u}}(t_2) \\ &\quad + \sum_{k \in \mathcal{K}} (\boldsymbol{w}_k(t_2))^H \left(\left(\mathbf{q}^{\mathrm{u}}(t_2)\right)^T \mathbf{q}^{\mathrm{d}}(t_2)\right) \\ &\quad \circ \mathbf{H}_{\mathrm{SI}}(t_2)x_k^{\mathrm{d}}(t_2)b_{k,m}^{\mathrm{d}}(t_2)b_{j,m}^{\mathrm{u}}(t_2) \\ &\quad + \mathbf{q}^{\mathrm{u}}(t_2) \circ \mathbf{n}_j^{\mathrm{u}}(t_2)b_{j,m}^{\mathrm{u}}(t_2) \end{aligned} \tag{12}$$

and

$$\begin{aligned} y_{k,m}^{\mathrm{d}}(t_2) &= \left(\boldsymbol{h}_k^{\mathrm{d}}(t_2)\right)^H \circ \mathbf{q}^{\mathrm{d}}(t_2)\boldsymbol{w}_k(t_2)x_k^{\mathrm{d}}(t_2)b_{k,m}^{\mathrm{d}}(t_2) \\ &\quad + \sum_{k' \in \mathcal{K}, k' \neq k} \left(\boldsymbol{h}_k^{\mathrm{d}}(t_2)\right)^H \\ &\quad \circ \mathbf{q}^{\mathrm{d}}(t_2)\boldsymbol{w}_{k'}(t_2)x_{k'}^{\mathrm{d}}(t_2)b_{k',m}^{\mathrm{d}}(t_2)b_{k,m}^{\mathrm{d}}(t_2) \\ &\quad + \sum_{j \in \mathcal{J}} g_{k,j}(t_2)\sqrt{p_j(t_2)}x_j^{\mathrm{u}}(t_2)b_{j,m}^{\mathrm{u}}(t_2)b_{k,m}^{\mathrm{d}}(t_2) \\ &\quad + n_k^{\mathrm{d}}(t_2)b_{k,m}^{\mathrm{d}}(t_2), \end{aligned} \tag{13}$$

respectively, where operator "$\circ$" represents Hadamard product. $b_{z,m}^{\chi}(t_2)$ is the assignment state of subcarrier m to user z for service χ, represented as

$$b_{z,m}^{\chi}(t_2) = \begin{cases} 0, & \mathrm{Sch}_z^{\chi}(t_1) = \mathrm{Sch}_z^{\chi}(t_2) = 0, \\ 1, & \mathrm{Sch}_z^{\chi}(t_1) = \mathrm{Sch}_z^{\chi}(t_2) = 1. \end{cases} \tag{14}$$

It is evident that a scheduled user corresponds to the user assigned a subcarrier and vice versa. In conclusion, the assignment states for each user constitute the subcarrier allocation matrix, which is defined as

$$\begin{aligned} \mathbf{B}(t_2) = [&\boldsymbol{b}_1^{\mathrm{u}}(t_2), \boldsymbol{b}_1^{\mathrm{d}}(t_2), \boldsymbol{b}_2^{\mathrm{u}}(t_2), \boldsymbol{b}_2^{\mathrm{d}}(t_2), \\ &\ldots, \boldsymbol{b}_Z^{\mathrm{u}}(t_2), \boldsymbol{b}_Z^{\mathrm{d}}(t_2)]^T \in \mathbb{R}^{2Z \times M}, \end{aligned} \tag{15}$$

where $\boldsymbol{b}_z^{\chi}(t_2) = [b_{z,1}^{\chi}(t_2), b_{z,2}^{\chi}(t_2), \ldots, b_{z,M}^{\chi}(t_2)]$ is a submatrix for user z.

Note that for an arbitrary subcarrier m, one and only one $\boldsymbol{y}_{j,m}^{\mathrm{u}}(t_2)$ has practical significance. It results from the fact that each scheduled user is only assigned one subcarrier. Hence, we could substitute the expression $\boldsymbol{y}_j^{\mathrm{u}}(t_2)$ for $\boldsymbol{y}_{j,m}^{\mathrm{u}}(t_2)$ in the paper below for simplification. Similarly, $y_{k,m}^{\mathrm{d}}(t_2)$ is simplified to $y_k^{\mathrm{d}}(t_2)$.

2) TRANSMISSION AT SUB-SLOT 1

Since the transmission at sub-slot 1 involves an additional factor related to MM, we reformulate the interference UL/DL models at sub-slot 1 according to (12) and (13) as

$$\begin{aligned} \boldsymbol{y}_j^{\mathrm{u}}(t_1) &= \mathbf{q}^{\mathrm{u}}(t_1) \circ \boldsymbol{h}_j^{\mathrm{u}}(t_1)\sqrt{p_j(t_1)}x_j^{\mathrm{u}}(t_1)b_{j,m}^{\mathrm{u}}(t_1)e_j(t_1) \\ &\quad + \sum_{j' \in \mathcal{J}, j' \neq j} \Big(\mathbf{q}^{\mathrm{u}}(t_1) \circ \boldsymbol{h}_{j'}^{\mathrm{u}}(t_1)\sqrt{p_{j'}(t_1)}x_{j'}^{\mathrm{u}}(t_1) \\ &\quad \cdot b_{j',m}^{\mathrm{u}}(t_1)e_{j'}(t_1)b_{j,m}^{\mathrm{u}}(t_1)e_j(t_1)\Big) \\ &\quad + \sum_{k \in \mathcal{K}} \Big((\boldsymbol{w}_k(t_1))^H \left(\left(\mathbf{q}^{\mathrm{u}}(t_1)\right)^T \mathbf{q}^{\mathrm{d}}(t_1)\right) \\ &\quad \circ \mathbf{H}_{\mathrm{SI}}(t_1)x_k^{\mathrm{d}}(t_1)b_{k,m}^{\mathrm{d}}(t_1)e_k(t_1)b_{j,m}^{\mathrm{u}}(t_1)e_j(t_1)\Big) \\ &\quad + \mathbf{q}^{\mathrm{u}}(t_1) \circ \boldsymbol{n}_j^{\mathrm{u}}(t_1)b_{j,m}^{\mathrm{u}}(t_1)e_j(t_1) \end{aligned} \tag{16}$$

and

$$\begin{aligned} &y_{k,m}^{\mathrm{d}}(t_1) \\ &= \left(\boldsymbol{h}_k^{\mathrm{d}}(t_1)\right)^H \circ \mathbf{q}^{\mathrm{d}}(t_1)\boldsymbol{w}_k(t_1)x_k^{\mathrm{d}}(t_1)b_{k,m}^{\mathrm{d}}(t_1)e_k(t_1) \\ &\quad + \sum_{k' \in \mathcal{K}, k' \neq k} \Big(\left(\boldsymbol{h}_k^{\mathrm{d}}(t_1)\right)^H \circ \mathbf{q}^{\mathrm{d}}(t_1)\boldsymbol{w}_{k'}(t_1)x_{k'}^{\mathrm{d}}(t_1)b_{k',m}^{\mathrm{d}}(t_1) \\ &\quad \cdot e_{k'}(t_1)b_{k,m}^{\mathrm{d}}(t_1)e_k(t_1)\Big) \\ &\quad + \sum_{j \in \mathcal{J}} g_{k,j}(t_1)\sqrt{p_j(t_1)}x_j^{\mathrm{u}}(t_1)b_{j,m}^{\mathrm{u}}(t_1)e_j(t_1)b_{k,m}^{\mathrm{d}}(t_1)e_k(t_1) \\ &\quad + n_k^{\mathrm{d}}(t_1)b_{k,m}^{\mathrm{d}}(t_1)e_k(t_1), \end{aligned} \tag{17}$$

where $e_z(t_1)$ is the service muting state written as

$$e_z(t_1) = \begin{cases} 0, & \alpha_z^{\chi}(t_1) = 1,\ \alpha_z^{\chi}(t_2) = 2, \\ 1, & \alpha_z^{\chi}(t_1) = 2,\ \alpha_z^{\chi}(t_2) = 2. \end{cases} \tag{18}$$

It is obvious that value 0 means muting. Similar to the subcarrier assignment, $e_z(t_1)$ is a part of the vector $\mathbf{e}(t_1) = [e_1(t_1), e_2(t_1), \ldots, e_Z(t_1)]$.

As seen from (18), the muting indicator only acts on the scheduled user, parameters $e_z(t_1)$ and $b_{z,(\cdot)}^{\chi}(t_1)$ are highly correlated. The notation "$(\cdot)$" in $b_{z,(\cdot)}^{\chi}(t_1)$ means the allocated specific subcarrier to user z.

In view of the correlation, we introduce a new parameter $\tilde{b}_z(t_1)$ called a service-enabled state, which is defined as

$$\tilde{b}_z(t_1) = b_{z,(\cdot)}^{\chi}(t_1)e_z(t_1), \tag{19}$$

and satisfies

$$\tilde{b}_z(t_1) = \begin{cases} 1, & \text{user } z \text{ is not muted for the assigned} \\ & \text{subcarrier at sub-slot 1,} \\ 0, & \text{user } z \text{ is muted for the assigned} \\ & \text{subcarrier at sub-slot 1.} \end{cases} \tag{20}$$

The service-enabled states of each user also form the vector $\tilde{\mathbf{B}}(t_1) = [\tilde{b}_1(t_1), \tilde{b}_2(t_1), \ldots, \tilde{b}_Z(t_1)]$.

We rewrite (16) and (17) using parameter $\tilde{\mathbf{B}}(t_1)$ and can readily get the new expressions at sub-slot 1 that resemble (12) and (13) at sub-slot 2. The difference between UL/DL interference models at each sub-slot lies in the varied parameters $\mathbf{B}$ and $\tilde{\mathbf{B}}$ below

$$\mathbf{y}_j^{\mathrm{u}}(t_i) = \begin{cases} \boldsymbol{f}_j(\mathbf{Q}, \tilde{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_1), & j \in \mathcal{J}, \\ \boldsymbol{f}_j(\mathbf{Q}, \mathbf{B}, \mathcal{W}, \mathcal{P}, t_2), & j \in \mathcal{J}, \end{cases} \tag{21a}$$

$$y_k^{\mathrm{d}}(t_i) = \begin{cases} g_k(\mathbf{Q}, \tilde{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_1), & k \in \mathcal{K}, \\ g_k(\mathbf{Q}, \mathbf{B}, \mathcal{W}, \mathcal{P}, t_2), & k \in \mathcal{K}, \end{cases} \tag{21b}$$

where $\boldsymbol{f}_j$ and g_k are the functions of the received UL and DL signals at the matched sub-slot, respectively.

To strive for simplification of (21a) and (21b), we regulate new variate $\hat{\mathbf{B}}$ to substitute for $\mathbf{B}$ and $\tilde{\mathbf{B}}$ below

$$\hat{\mathbf{B}}(t_i) = \begin{cases} \tilde{\mathbf{B}}(t_i), & i = 1, \\ \mathbf{B}(t_i), & i = 2. \end{cases} \tag{22}$$

$\hat{b}_z(t_i)$ is the element of $\hat{\mathbf{B}}(t_i)$, which follows

$$\hat{b}_z(t_i) = \begin{cases} b^{\chi}_{z,(\cdot)}(t_1) e_z(t_1), & (\cdot) = \text{assigned subcarrier}, \\ b^{\chi}_{z,m}(t_2), & \forall m \in \mathcal{M}. \end{cases} \tag{23}$$

As a result, we create a single standard formula instead of the two expressions at each sub-slot for the sake of problem formulation.

IV. PROBLEM FORMULATION

The UL SINR of user j at sub-slot 1 or 2 is written as

$$\gamma_j^{\mathrm{u}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i) = \frac{\hat{b}_j(t_i) p_j(t_i) \left(\tilde{\boldsymbol{h}}_j^{\mathrm{u}}(t_i)\right)^H \tilde{\boldsymbol{h}}_j^{\mathrm{u}}(t_i)}{\Phi_{\mathrm{UU}}(t_i) + \boldsymbol{\Phi}_{\mathrm{SI}}(t_i) + \sigma^2_{\mathrm{u},j} \hat{b}_j(t_i) \left(\mathbf{q}^{\mathrm{u}}(t_i)\right)^T \mathbf{q}^{\mathrm{u}}(t_i)}, \tag{24}$$

where

$$\tilde{\boldsymbol{h}}_j^{\mathrm{u}}(t_i) = \mathbf{q}^{\mathrm{u}}(t_i) \circ \boldsymbol{h}_j^{\mathrm{u}}(t_i). \tag{25}$$

$\boldsymbol{\Phi}_{\mathrm{UU}}(t_i)$ and $\boldsymbol{\Phi}_{\mathrm{SI}}(t_i)$ are the covariances matrices of UL to UL interference and SI, respectively. They are given by

$$\boldsymbol{\Phi}_{\mathrm{UU}}(t_i) = \sum_{j' \in \mathcal{J}, j' \neq j} \hat{b}_{j'}(t_i) p_{j'}(t_i) \left(\tilde{\boldsymbol{h}}_{j'}^{\mathrm{u}}(t_i)\right)^H \tilde{\boldsymbol{h}}_{j'}^{\mathrm{u}}(t_i), \tag{26}$$

$$\boldsymbol{\Phi}_{\mathrm{SI}}(t_i) = \sum_{k \in \mathcal{K}} \hat{b}_k(t_i) \left(\boldsymbol{w}_k(t_i)\right)^H \tilde{\mathbf{H}}_{\mathrm{SI}}(t_i) \left(\tilde{\mathbf{H}}_{\mathrm{SI}}(t_i)\right)^H \boldsymbol{w}_k(t_i), \tag{27}$$

in which

$$\tilde{\mathbf{H}}_{\mathrm{SI}}(t_i) = \left(\left(\mathbf{q}^{\mathrm{u}}(t_i)\right)^T \mathbf{q}^{\mathrm{d}}(t_i)\right) \circ \mathbf{H}_{\mathrm{SI}}(t_i). \tag{28}$$

Similarly, the DL SINR of user k is expressed as

$$\gamma_k^{\mathrm{d}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i) = \frac{\hat{b}_k(t_i) \left(\tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i)\right)^H \boldsymbol{w}_k(t_i) \left(\boldsymbol{w}_k(t_i)\right)^H \tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i)}{\phi_{\mathrm{DD}}(t_i) + \phi_{\mathrm{UD}}(t_i) + \hat{b}_k(t_i) \sigma^2_{\mathrm{d},k}}, \tag{29}$$

where

$$\tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i) = \boldsymbol{h}_k^{\mathrm{d}}(t_i) \circ \mathbf{q}^{\mathrm{d}}(t_i). \tag{30}$$

$\phi_{\mathrm{DD}}(t_i)$ and $\phi_{\mathrm{UD}}(t_i)$ are the variances of DL to DL and UL to DL interference, respectively. They are written as

$$\phi_{\mathrm{DD}}(t_i) = \sum_{k' \in \mathcal{K}, k' \neq k} \hat{b}_{k'}(t_i) \left(\tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i)\right)^H \boldsymbol{w}_{k'}(t_i) \left(\boldsymbol{w}_{k'}(t_i)\right)^H \tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i), \tag{31}$$

$$\phi_{\mathrm{UD}}(t_i) = \sum_{j \in \mathcal{J}} g^2_{k,j}(t_i) p_j(t_i) \hat{b}_j(t_i). \tag{32}$$

Finally, we substitute (24) and (29) into the Shannon formula to acquire the SE of UL user j and DL user k as

$$R_j^{\mathrm{u}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i) = \log \det \left(\mathbf{I}_N + \gamma_j^{\mathrm{u}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i)\right) \tag{33}$$

and

$$R_k^{\mathrm{d}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i) = \log \left(1 + \gamma_k^{\mathrm{d}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i)\right), \tag{34}$$

respectively, where $\det(\cdot)$ is the determinant operator.

Thus, the total SE of all scheduled users in the cell at the schedule unit is defined as

$$\begin{aligned} \mathrm{R}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}) &= \sum_{i=1}^{2} R(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i) \\ &= \sum_{i=1}^{2} \left(\sum_{j \in \mathcal{J}} R_j^{\mathrm{u}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i) \right. \\ &\quad \left. + \sum_{k \in \mathcal{K}} R_k^{\mathrm{d}}(\mathbf{Q}, \hat{\mathbf{B}}, \mathcal{W}, \mathcal{P}, t_i) \right) \end{aligned} \tag{35}$$

Given the (35), it is noteworthy that the total SE is related to multiple influencing factors. The parameters $\mathcal{P}$ and $\mathcal{W}$ apparently work for UL and DL power allocation to mitigate interference. In contrast, the parameters $\mathbf{Q}$ and $\hat{\mathbf{B}}$ are indirectly concerned with that, which is explained below.

In (24), the powers of a specific user j' or k can impact the $\boldsymbol{\Phi}_{\mathrm{UU}}$ or $\boldsymbol{\Phi}_{\mathrm{SI}}$ term of user j, which brings side effects to the SE of user j. Similarly, in (29), the user k' or j will decrease the SE of user k via the increase of ϕ_{DD} or ϕ_{UD}. For ease of presentation, we suppose two users as an entirety, one user at numerator referring to (24) or (29) with a higher ratio of throughput to power (also called EE) is interfered with by another. If the performance loss is higher than the SE obtained by the lower-EE user, the total performance will degrade.

Hence, it is easily acquired that the appropriate parameter $\hat{\mathbf{B}}$ to mute the lower-EE user can effectively mitigate the residual SI (equivalent to $\mathbf{\Phi}_{\mathrm{SI}}$) or multi-user interference (same as $\mathbf{\Phi}_{\mathrm{UU}}$, ϕ_{DD}, or ϕ_{UD}). Also, parameter $\mathbf{Q}$ is correlated with the composite channel gains, such as $\tilde{\boldsymbol{h}}_j^{\mathrm{u}}$, $\boldsymbol{h}_{j'}^{\mathrm{u}}$, and $\tilde{\boldsymbol{h}}_k^{\mathrm{d}}$, which directly affect the powers. It shows a better performance than fixed channel gains $\boldsymbol{h}_j^{\mathrm{u}}$, $\boldsymbol{h}_{j'}^{\mathrm{u}}$, and $\boldsymbol{h}_k^{\mathrm{d}}$ when they are in poor condition. In fact, the essence of parameters $\mathbf{Q}$ and $\hat{\mathbf{B}}$ comes down to a power allocation issue.

In order to implement a comprehensive power allocation method, we take (35) as an objective function and formulate the optimization problem subject to several constraints as

$$\max_{\{\mathbf{Q},\hat{\mathbf{B}},\mathcal{W},\mathcal{P}\}} \quad \mathrm{R}(\mathbf{Q},\hat{\mathbf{B}},\mathcal{W},\mathcal{P}) \tag{36a}$$

$$s.t.\ i \in \{1,2\}, \tag{36b}$$

$$\chi \in \{\mathrm{u},\mathrm{d}\}, \tag{36c}$$

$$\sum_{m,m\in\mathcal{M}} b_{z,m}^{\chi}(t_i) \le 1, \tag{36d}$$

$$\alpha_z^{\chi}(t_i) \in \{1,2\}, \tag{36e}$$

$$\tilde{b}_z(t_1) \in \{0,1\}, \quad \forall z \in \mathcal{Z}, \tag{36f}$$

$$q_l^{\chi}(t_i) \in \{0,1\}, \quad \forall l \in \mathcal{N}, \tag{36g}$$

$$q_l^{\mathrm{u}}(t_i) + q_l^{\mathrm{d}}(t_i) \in \{1,2\}, \tag{36h}$$

$$\sum_{k\in\mathcal{K}} \left|\mathbf{q}^{\mathrm{d}}(t_i)\boldsymbol{w}_k(t_i)\right| \le P_{\max}, \quad \forall k \in \mathcal{K},$$

$$0 \le \mathbf{q}^{\mathrm{d}}(t_i)\boldsymbol{w}_k(t_i) \le p_{\max}, \quad \boldsymbol{w}_k(t_i) \in \mathcal{W}(t_i), \tag{36i}$$

$$0 \le p_j(t_i) \le p_{\max}, p_j(t_i) \in \mathcal{P}(t_i), \quad \forall j \in \mathcal{J}, \tag{36j}$$

$$R_j^{\mathrm{u}}(\mathbf{Q},\hat{\mathbf{B}},\mathcal{W},\mathcal{P},t_i) \ge \mathrm{R}_{\mathrm{req}}^{\mathrm{u}}, \quad \forall j \in \mathcal{J}/\mathcal{U}, \tag{36k}$$

$$R_j^{\mathrm{u}}(\mathbf{Q},\hat{\mathbf{B}},\mathcal{W},\mathcal{P},t_2) \ge \beta\mathrm{R}_{\mathrm{req}}^{\mathrm{u}}, \quad \forall j \in \mathcal{U},\ \beta > 1, \tag{36l}$$

$$R_k^{\mathrm{d}}(\mathbf{Q},\hat{\mathbf{B}},\mathcal{W},\mathcal{P},t_i) \ge \mathrm{R}_{\mathrm{req}}^{\mathrm{d}}, \quad \forall k \in \mathcal{K}/\mathcal{D}, \tag{36m}$$

$$R_k^{\mathrm{d}}(\mathbf{Q},\hat{\mathbf{B}},\mathcal{W},\mathcal{P},t_2) \ge \beta\mathrm{R}_{\mathrm{req}}^{\mathrm{d}}, \quad \forall k \in \mathcal{D},\ \beta > 1. \tag{36n}$$

(36b) and (36c) restrict two sub-slots in one schedule unit and two types of service, respectively. (36d) means each user can be only assigned one subcarrier at most. (36e) and (36f) imply that only the scheduled users could be silenced subject to identifiers in (36e). (36g) and (36h) determine the work modes of each BS antenna. $p_{\max}$ and $P_{\max}$ are the maximum powers for the user and BS, respectively. Thus, (36i) and (36j) are each the maximum power constraint for DL and UL users. (36k) and (36m) are the quality of service (QoS) constraints for the unmuted UL and DL users at sub-slot 1 or 2, while (36l) and (36n) are the QoS constraints for the resumed UL and DL users at sub-slot 2, which were once muted at sub-slot 1. The compensation coefficient β is used to remedy the performance loss for the muted users.

We fulfill the integration of the abovementioned three elements through (36a). By solving the optimization problem, we can acquire a maximum SE with the optimal UL/DL power allocation, which also considers MM and the assignment of BS antennas and subcarriers.

V. ALGORITHM DESCRIPTION

Apart from the binary constraints, the object function (36a) and the constraints (36k)-(36n) are all nonconvex. Hence, this is a non-deterministic polynomial hard (NP-hard) optimization problem [37]. Furthermore, binary variables $\mathbf{Q}$ and $\hat{\mathbf{B}}$ with coupled UL and DL power allocation make the traditional solution even more impractical. Considering $\mathbf{Q}$ and $\hat{\mathbf{B}}$ are discrete variables while $\mathcal{W}$ and $\mathcal{P}$ are continuous, based on different variable types, we mainly present two hierarchical methods to solve the problem of (36) in this section. The hierarchical method intends to split the problem into two subproblems. We can go through each subproblem by looping to solve the initial problem ultimately.

A. HYBRID GREEDY-CONVEX METHOD

A practicable method for the continuous variables related subproblem is to construct an approximate function that is easier to solve than the original NP-hard problem. Several approximation algorithms, such as successive convex approximation and majorization-minimization, are used to address this issue [38], [39]. Considering the nature of SE equations, we apply another approximate method called the FOTA in this paper.

First, we reformulate the problem of (36) based on the fixed $\mathbf{Q}$ and $\hat{\mathbf{B}}$ to realize problem decomposition as

$$\max_{\{\mathcal{W},\mathcal{P}\}} \mathrm{R}(\mathcal{W},\mathcal{P}) \tag{37a}$$

$$s.t.\ (36\mathrm{b}), (36\mathrm{i}), (36\mathrm{j}), \tag{37b}$$

$$R_j^{\mathrm{u}}(\mathcal{W},\mathcal{P},t_i) \ge \mathrm{R}_{\mathrm{req}}^{\mathrm{u}}, \quad \forall j \in \mathcal{J}/\mathcal{U}, \tag{37c}$$

$$R_j^{\mathrm{u}}(\mathcal{W},\mathcal{P},t_2) \ge \beta\mathrm{R}_{\mathrm{req}}^{\mathrm{u}}, \quad \forall j \in \mathcal{U},\ \beta > 1, \tag{37d}$$

$$R_k^{\mathrm{d}}(\mathcal{W},\mathcal{P},t_i) \ge \mathrm{R}_{\mathrm{req}}^{\mathrm{d}}, \quad \forall k \in \mathcal{K}/\mathcal{D}, \tag{37e}$$

$$R_k^{\mathrm{d}}(\mathcal{W},\mathcal{P},t_2) \ge \beta\mathrm{R}_{\mathrm{req}}^{\mathrm{d}}, \quad \forall k \in \mathcal{D},\ \beta > 1, \tag{37f}$$

where the expansion of (37a) through the logarithmic property is written as

$$\mathrm{R}(\mathcal{W},\mathcal{P}) = \mathrm{R}_1(\mathcal{W},\mathcal{P}) - \mathrm{R}_2(\mathcal{W},\mathcal{P}), \tag{38}$$

where

$$\mathrm{R}_1(\mathcal{W},\mathcal{P}) = \sum_{i=1}^{2}\left(\sum_{j\in\mathcal{J}}\log\det\left(\mathbf{J}_j^{\mathrm{u}}(t_i) + p_j(t_i)\mathbf{A}_j^{\mathrm{u}}(t_i)\right) + \sum_{k\in\mathcal{K}}\log\left(\mathrm{K}_k^{\mathrm{d}}(t_i) + \hat{b}_k^{\mathrm{d}}(t_i)\left(\tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i)\right)^H \boldsymbol{w}_k(t_i)\left(\boldsymbol{w}_k(t_i)\right)^H \tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i)\right)\right), \tag{39}$$

$$\mathrm{R}_2(\mathcal{W},\mathcal{P}) = \sum_{i=1}^{2}\left(\sum_{j\in\mathcal{J}}\log\det\left(\mathbf{J}_j^{\mathrm{u}}(t_i)\right) + \sum_{k\in\mathcal{K}}\log\left(\mathrm{K}_k^{\mathrm{d}}(t_i)\right)\right). \tag{40}$$

On the right hand side of (39) and (40), several newly defined expressions are represented as follows

$$\mathbf{J}_j^{\mathrm{u}}(t_i) = \mathbf{\Phi}_{\mathrm{UU}}(t_i) + \mathbf{\Phi}_{\mathrm{SI}}(t_i) + \mathbf{C}_j^{\mathrm{u}}(t_i), \tag{41}$$

$$\mathrm{K}_k^{\mathrm{d}}(t_i) = \phi_{\mathrm{DD}}(t_i) + \phi_{\mathrm{UD}}(t_i) + \mathrm{D}_k^{\mathrm{d}}(t_i), \tag{42}$$

$$\mathbf{A}_j^{\mathrm{u}}(t_i) = \hat{b}_j^{\mathrm{u}}(t_i) \left(\tilde{\boldsymbol{h}}_j^{\mathrm{u}}(t_i)\right)^H \tilde{\boldsymbol{h}}_j^{\mathrm{u}}(t_i), \tag{43}$$

where $\mathbf{C}_j^{\mathrm{u}}(t_i)$ and $\mathrm{D}_k^{\mathrm{d}}(t_i)$ in (41) and (42) are defined as

$$\mathbf{C}_j^{\mathrm{u}}(t_i) = \sigma_{\mathrm{u},j}^2 \hat{b}_j^{\mathrm{u}}(t_i) \left(\mathbf{q}^{\mathrm{u}}(t_i)\right)^T \mathbf{q}^{\mathrm{u}}(t_i), \tag{44}$$

$$\mathrm{D}_k^{\mathrm{d}}(t_i) = \hat{b}_k^{\mathrm{d}}(t_i)\sigma_{\mathrm{d},k}^2, \tag{45}$$

respectively.

From (39) and (40), we can see that both $\mathrm{R}_1(\mathcal{W}, \mathcal{P})$ and $\mathrm{R}_2(\mathcal{W}, \mathcal{P})$ are concave logarithmic functions. As the formation of (40) is more straightforward than (39), we only need to analyze (40) mathematically.

The FOTA of (40) with multiple iterations will converge to $\mathrm{R}_2(\mathcal{W}, \mathcal{P})$ due to the function concavity [40]. Accordingly, we can acquire the approximate value of $\mathrm{R}_2(\mathcal{W}, \mathcal{P})$ by taking derivatives. To facilitate partial differentiation, we convert $\mathbf{J}_j^{\mathrm{u}}(t_i)$ and $\mathrm{K}_k^{\mathrm{d}}(t_i)$ in (40) to a formalization with only two direct variables as

$$\begin{aligned}\mathbf{J}_j^{\mathrm{u}}(t_i) = &\sum_{j' \in \mathcal{J}, j' \neq j} p_{j'}(t_i) \mathbf{A}_{j'}^{\mathrm{u}}(t_i) \\ &+ \sum_{k \in \mathcal{K}} (\boldsymbol{w}_k(t_i))^H \mathbf{E}_k^{\mathrm{d}}(t_i) \boldsymbol{w}_k(t_i) + \mathbf{C}_j^{\mathrm{u}}(t_i),\end{aligned} \tag{46}$$

$$\begin{aligned}\mathrm{K}_k^{\mathrm{d}}(t_i) = &\sum_{k' \in \mathcal{K}, k' \neq k} \hat{b}_{k'}^{\mathrm{d}}(t_i) \left(\tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i)\right)^H \boldsymbol{w}_{k'}(t_i) \left(\boldsymbol{w}_{k'}(t_i)\right)^H \tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i) \\ &+ \sum_{j \in \mathcal{J}} \mathrm{F}_j^{\mathrm{u}}(t_i) p_j(t_i) + \mathrm{D}_k^{\mathrm{d}}(t_i),\end{aligned} \tag{47}$$

where

$$\mathbf{E}_k^{\mathrm{d}}(t_i) = \hat{b}_k^{\mathrm{d}}(t_i) \tilde{\mathbf{H}}_{\mathrm{SI}}(t_i) \left(\tilde{\mathbf{H}}_{\mathrm{SI}}(t_i)\right)^H, \tag{48}$$

$$\mathrm{F}_j^{\mathrm{u}}(t_i) = g_{k,j}^2(t_i) \hat{b}_j^{\mathrm{u}}(t_i). \tag{49}$$

As a result, we calculate n iterations to obtain the FOTA of (40) as

$$\begin{aligned}\mathrm{R}_2(\mathcal{W}, \mathcal{P}) &\approx \mathrm{R}_2(\mathcal{W}^{(n)}, \mathcal{P}^{(n)}) + \mathrm{R}_2'(\mathcal{W}^{(n)})(\mathcal{W} - \mathcal{W}^{(n)}) \\ &\quad + \mathrm{R}_2'(\mathcal{P}^{(n)})(\mathcal{P} - \mathcal{P}^{(n)}) \\ &= \mathrm{R}_2^{(n)}(\mathcal{W}, \mathcal{P}),\end{aligned} \tag{50}$$

where

$$\begin{aligned}&\mathrm{R}_2'(\mathcal{W}^{(n)})(\mathcal{W} - \mathcal{W}^{(n)}) \\ &= \sum_{i=1}^{2} \Bigg(\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \mathrm{tr} \Bigg(2 \left(\left(\mathbf{J}_j^{\mathrm{u}}(t_i) \right)^{(n)} \right)^{-1} \mathbf{E}_k^{\mathrm{d}}(t_i) \left(\boldsymbol{w}_k(t_i)\right)^{(n)} \\ &\quad \cdot \left(\boldsymbol{w}_k(t_i) - \left(\boldsymbol{w}_k(t_i)\right)^{(n)} \right) \Bigg) \\ &\quad + \sum_{k \in \mathcal{K}} \sum_{k' \in \mathcal{K}, k' \neq k} \Bigg(2 \left(\mathrm{K}_k^{\mathrm{d}}(t_i) \right)^{-1} \hat{b}_{k'}^{\mathrm{d}}(t_i) \\ &\quad \cdot \left(\tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i) \right)^H \left(\boldsymbol{w}_{k'}(t_i)\right)^{(n)} \tilde{\boldsymbol{h}}_k^{\mathrm{d}}(t_i) \left(\boldsymbol{w}_{k'}(t_i) - \left(\boldsymbol{w}_{k'}(t_i)\right)^{(n)} \right) \Bigg) \Bigg),\end{aligned} \tag{51}$$

$$\begin{aligned}&\mathrm{R}_2'(\mathcal{P}^{(n)})(\mathcal{P} - \mathcal{P}^{(n)}) \\ &= \sum_{i=1}^{2} \Bigg(\sum_{j \in \mathcal{J}} \sum_{j' \in \mathcal{J}, j' \neq j} \mathrm{tr} \Bigg(\left(\left(\mathbf{J}_j^{\mathrm{u}}(t_i) \right)^{(n)} \right)^{-1} \mathbf{A}_{j'}^{\mathrm{u}}(t_i) \\ &\quad \cdot \left(p_{j'}(t_i) - \left(p_{j'}(t_i)\right)^{(n)} \right) \Bigg) \\ &\quad + \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} \left(\mathrm{K}_k^{\mathrm{d}}(t_i) \right)^{-1} \mathrm{F}_j^{\mathrm{u}}(t_i) \left(p_j(t_i) - \left(p_j(t_i)\right)^{(n)} \right) \Bigg).\end{aligned} \tag{52}$$

Obviously, (51) and (52) are affine functions with respect to $\mathcal{W}$ and $\mathcal{P}$, respectively. Hence, we transform (50) into an affine function approximately. Substituting (50) into (38), we acquire the concave object function of problem (37) accordingly.

Similarly, the nonconvex constraints (37c)-(37f) can be each decomposed with two logarithmic functions subtracted, so we approximately achieve the concave constraints with the assistance of the FOTA method.

Consequently, we convert the problem of (37) into an approximate convex optimization problem as

$$\max_{\{\mathcal{W}, \mathcal{P}\}} \mathrm{R}_1(\mathcal{W}, \mathcal{P}) - \mathrm{R}_2^{(n)}(\mathcal{W}, \mathcal{P}) \tag{53a}$$

$$s.t.\ (36b), (36i), (36j), \tag{53b}$$

$$R_{j,1}^{\mathrm{u}}(\mathcal{W}, \mathcal{P}, t_i) - R_{j,2}^{\mathrm{u}(n)}(\mathcal{W}, \mathcal{P}, t_i) \geq \mathrm{R}_{\mathrm{req}}^{\mathrm{u}}, \quad \forall j \in \mathcal{J}/\mathcal{U}, \tag{53c}$$

$$R_{j,1}^{\mathrm{u}}(\mathcal{W}, \mathcal{P}, t_2) - R_{j,2}^{\mathrm{u}(n)}(\mathcal{W}, \mathcal{P}, t_2) \geq \beta \mathrm{R}_{\mathrm{req}}^{\mathrm{u}}, \quad \forall j \in \mathcal{U}, \quad \beta > 1, \tag{53d}$$

$$R_{k,1}^{\mathrm{d}}(\mathcal{W}, \mathcal{P}, t_i) - R_{k,2}^{\mathrm{d}(n)}(\mathcal{W}, \mathcal{P}, t_i) \geq \mathrm{R}_{\mathrm{req}}^{\mathrm{d}}, \quad \forall k \in \mathcal{K}/\mathcal{D}, \tag{53e}$$

$$R_{k,1}^{\mathrm{d}}(\mathcal{W}, \mathcal{P}, t_2) - R_{k,2}^{\mathrm{d}(n)}(\mathcal{W}, \mathcal{P}, t_2) \geq \beta \mathrm{R}_{\mathrm{req}}^{\mathrm{d}}, \quad \forall k \in \mathcal{D}, \quad \beta > 1. \tag{53f}$$

With Matlab convex tool [41], we can work out the convex optimization problem and obtain the optimal solution for $\mathcal{W}$, $\mathcal{P}$, and the corresponding total SE.

Since the optimum solution of discrete variables $\mathbf{Q}$ and $\hat{\mathbf{B}}$ can not be acquired through differentiating regularly, a direct approach to choosing an appropriate configuration is an exhaustive search referring to [14]. Nevertheless, when problem parameters extend, global search, such as the exhaustive search [10], is incompetent due to the curse of dimensionality with two discrete variables. Accordingly, the greedy algorithm only searches several local optimums (namely, candidates) instead of the global optimum. Subsequently, it selects the best candidate from the candidate list to approximate the global optimum [42]. Since the greedy algorithm adopts a top-down structure, in which the backtracking is unnecessary, the efficiency is promoted to some extent compared with the exhaustive search. Thereby we apply the greedy algorithm to find a suboptimal configuration. In order to decrease the ergodic samples, we evenly pick up $i_{\mathbf{Q}}$ and $i_{\hat{\mathbf{B}}}$ samples through a sample rate from universal sets of $\mathbf{Q}$ and $\hat{\mathbf{B}}$, respectively,

where

$$i_{\mathbf{Q}} = \begin{cases} 2(2N)^{\mathrm{r_Q}}, & i_{\mathbf{Q}} < 2 \cdot 3^N, \\ 2 \cdot 3^N, & \text{otherwise}, \end{cases} \tag{54}$$

and

Algorithm 1 Hybrid Greedy-Convex Method

1: Set $I = 1$.
Initialize ζ, $I_{\max}$, ω_{th}.
2: **Repeat** each candidate member in the list:
3: Pick up $i_{\mathbf{Q}}$ and $i_{\hat{\mathbf{B}}}$ samples from the universal set stochastically, following the greedy search rule.
4: Set i= 1.
5: **Repeat** to identify a candidate member:
6: Choose one candidate member $\mathbf{Q}_{i,I}$, $\hat{\mathbf{B}}_{i,I}$ from the samples.
7: Set $n = 0$.
Generate the initial value of $\mathcal{W}^{(0)}$, $\mathcal{P}^{(0)}$.
8: **Repeat** each FOTA using (50):
9: Set $n = n + 1$.
10: **Until** $\left|\mathrm{R}_2^{(n)}(\mathcal{W}, \mathcal{P})_i - \mathrm{R}_2^{(n-1)}(\mathcal{W}, \mathcal{P})_i\right| < \zeta$ or reach the max iteration.
11: Solve (53) with $\mathrm{R}_2^{(n)}(\mathcal{W}, \mathcal{P})_i$ to acquire the approximate maximum total SE: $\mathrm{R}(\mathbf{Q}_{i,I}, \hat{\mathbf{B}}_{i,I}, \mathcal{W}^*, \mathcal{P}^*)$.
12: If $i = 1$, set $\mathrm{R}_I^* = \mathrm{R}(\mathbf{Q}_{i,I}, \hat{\mathbf{B}}_{i,I}, \mathcal{W}^*, \mathcal{P}^*)$.
13: Else if
$\left|\mathrm{R}(\mathbf{Q}_{i,I}, \hat{\mathbf{B}}_{i,I}, \mathcal{W}^*, \mathcal{P}^*) - \mathrm{R}(\mathbf{Q}_{i-1,I}, \hat{\mathbf{B}}_{i-1,I}, \mathcal{W}^*, \mathcal{P}^*)\right| > \zeta$
Set $\mathrm{R}_I^* = \mathrm{R}(\mathbf{Q}_{i,I}, \hat{\mathbf{B}}_{i,I}, \mathcal{W}^*, \mathcal{P}^*)$, $\omega = 0$.
Else $\omega = \omega + 1$
14: Set $i = i + 1$.
15: **Until** reach the max iteration $i_{\mathbf{Q}} i_{\hat{\mathbf{B}}}$ or $\omega > \omega_{\mathrm{th}}$.
16: If $I > 1$
If $\left|\mathrm{R}_I^* - \mathrm{R}_{I-1}^*\right| > \zeta$
$\omega = 0$.
Else $\omega = \omega + 1$.
17: Set $I = I + 1$.
18: **Until** reach the max iteration $I_{\max}$ or $\omega > \omega_{\mathrm{th}}$.
19: Choose the maximum candidate member as R^* from list.
20: Return R^*.

$$i_{\hat{\mathbf{B}}} = \begin{cases} (2ZM)^{\mathrm{r}_{\hat{\mathbf{B}}}} + (2Z^2M)^{\mathrm{r}_{\hat{\mathbf{B}}}}, & i_{\hat{\mathbf{B}}} < \left(2^{J+K}+1\right) M^{J+K-M}, \\ \left(2^{J+K}+1\right) M^{J+K-M}, & \text{otherwise}. \end{cases} \tag{55}$$

$\mathrm{r_Q}$ and $\mathrm{r}_{\hat{\mathbf{B}}}$ are the resolution ratios of sampling for $\mathbf{Q}$ and $\hat{\mathbf{B}}$, respectively. 3^N, M^{J+K-M}, and 2^{J+K} are each the number of combinations for BS antenna, subcarrier, and muted user assignment. $2N$ and $2ZM(2Z^2M)$ are each product of rows and columns of $\mathbf{Q}$ and $\hat{\mathbf{B}}$.

The greedy selection rule should follow two steps: 1) Initiate the configuration of all emitting/receiving antennas shared and all scheduled users unmuted; 2) Gradually decrease the share level of antennas and increase the muted users in a random process. Besides, we set a tolerance threshold ω_{th} to accelerate seeking the sub-optimal candidate.

Through the outer loop (namely, greedy method) and inner loop (namely, convex method) updates, a relatively optimum solution can be acquired. Accordingly, the hybrid greedy-convex method is summarized in Algorithm 1.

Nevertheless, the greedy method could be easily trapped in the local optimum for nonconvex problems even though it explores the last candidate. This is due to the fact that there are limited candidates in the list. Since the RL technique has a significant advantage in tackling a vast amount of data, we will adopt the DRL technique based on the previous proposal.

B. HYBRID DRL-CONVEX METHOD

It is known that Markov Decision Process (MDP) is a tuple that includes four elements as sets of current states s_t, next states s_{t+1}, actions a_t, and rewards r_{t+1}, where the t means the time step. In our devised system model, the work mode for BS antennas, the assignment of subcarriers, and the user MM are all handled by FD BS. We take BS as an agent for this reason. Because the set of actions is finite, we use discrete variables $\mathbf{Q}$ and $\hat{\mathbf{B}}$ as action $a = (a^q, a^{\hat{b}})$, in which a^q represents the BS antenna assignment, $a^{\hat{b}}$ denotes the joint of subcarrier assignment and MM. Considering that BS-agent adopts two sets of actions for each sub-slot in a schedule unit, we mainly focus on the agent behavior at sub-slot 1. This is because the action at sub-slot 2 is pared-down owing to no muting orders imported compared with sub-slot 1. Therefore analyzing realization at sub-slot 1 can reasonably cover the following implementation at sub-slot 2.

The action space can be written as

$$\mathcal{A} = \{(a^q, a^{\hat{b}})_1, (a^q, a^{\hat{b}})_2 \ldots, (a^q, a^{\hat{b}})_{A_1A_2}\}, \tag{56}$$

where A_1 and A_2 are the total number of combinations of a^q and $a^{\hat{b}}$, respectively.

The determined action at time step t from $\mathcal{A}$ will act on the constrained multi-user interference model and the FOTA algorithm, namely, the environment. Subsequently, the environment outputs the SINR of each scheduled user and total SE at time step $t + 1$, which are treated as the next state s_{t+1} and reward r_{t+1}, respectively. The state and the state space are recorded as

$$s_{t+1} = (\gamma_1^{\mathrm{u}}, \gamma_2^{\mathrm{u}}, \ldots, \gamma_J^{\mathrm{u}}, \gamma_1^{\mathrm{d}}, \gamma_2^{\mathrm{d}}, \ldots, \gamma_K^{\mathrm{d}})_{t+1} \tag{57}$$

and

$$\mathcal{S} = \{s_1, s_2, \ldots, s_{A_1A_2}\}, \tag{58}$$

while the output reward is denoted as

$$r_{t+1} = \left\{ \sum_{j \in \mathcal{J}} R_j^{\mathrm{u}} + \sum_{k \in \mathcal{K}} R_k^{\mathrm{d}} \,\middle|\, t+1 \right\}. \tag{59}$$

It is evident that once the BS-agent chooses a specific action at step t, each scheduled user will transit from the current state s_t to the next state s_{t+1} that is calculated based on the determined action, and thus the BS-agent is rewarded in the meantime. Correspondingly, the state transition probability is $P(s_{t+1}|s_t, a_t)$. With the new state and benefit, BS-agent will adapt its policy via trial-and-error and repeatedly make a new round of decisions. Note that the learning process of BS-agent

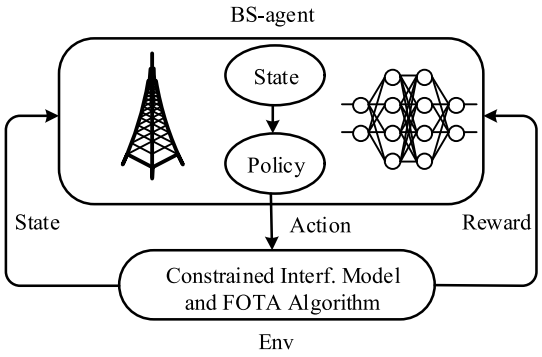

FIGURE 3. Interaction of learning BS-agent and environment.

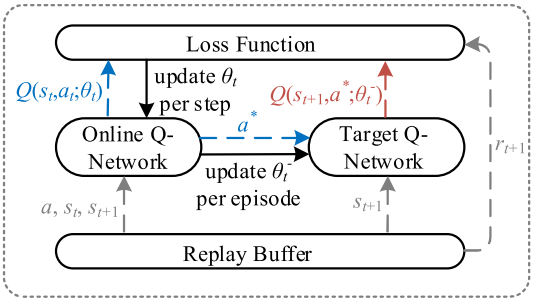

FIGURE 4. Flow of DDQN.

is directed by a reward that follows constraints of (36) in the environment.

The interaction between BS-agent and environment is visualized in Fig. 3 and deemed an MDP, which is a discrete decision problem on the time sequence.

In an MDP, the state value is determined by Behrman optimal equation [43]. Therefore an optimal reward table (i.e., Q table) will be acquired from the state and action values. Each element of the Q table is a necessary return from MDP and is written as

$$Q(s_t, a_t) = \mathbb{E}_\pi \left[\sum_{i=t}^{T} \gamma^{i-t} r_{i+1} | s_t, a_t \right], \tag{60}$$

where γ denotes the discount factor to the future step reward. If γ is set to 1 (0), the agent concentrates on the long-term (short-term) step reward [44]. The above flow is called Q learning, which is suitable for solving nonconvex problems with discrete variables.

However, for traditional Q learning, BS will maintain a $A_1A_2 \times A_1A_2$ size Q table, which will cause excessive memory occupation. On account of Q learning, the deep Q-learning network (DQN) exploits a deep neural network to estimate the Q value instead of the lookup table [45], thereby avoiding the case that the dimension of the Q table is too large to be looked up. DDQN is an improvement of DQN, which contains two Q-networks: an online Q-network for action selection and a target Q-network for action evaluation. It evades overfitting when selection and estimation are processed in the same DQN [46]. Given the above superiority in DDQN, we propose another hierarchical solution of the hybrid DRL-convex method, where DDQN is an alternative to the greedy algorithm.

Fig. 3 presents a macroscopic perspective of the interaction process, while we will introduce DDQN in a microscopic view to show the training process of the BS-agent.

In DDQN, the online Q-network and the target Q-network are represented as $Q(s_t, a_t; \theta_t)$ and $Q(s_t, a_t; \theta_t^-)$, respectively. θ_t^- and θ_t are the weighting factors on the to-do lists of training.

BS-agent initializes the online Q values for each action and state. The action a_t (namely, $\mathbf{Q}$ and $\hat{\mathbf{B}}$ at time step t) with the maximum Q value is selected, where

$$a_t = \arg\max_a Q(s_t, a; \theta_t). \tag{61}$$

Therefore, the related reward r_{t+1} (the total SE) and new state s_{t+1} (i.e., the UL/DL SINR of each user) will be obtained through interaction with the environment. s_t, a_t, r_{t+1}, and s_{t+1} together constitute a tuple that is stored in a replay buffer R.

For the learning process, the tuples are randomly picked out in batch from the replay buffer R. BS-agent determines the next action through the online Q-network with the max operation based on s_{t+1}. The determined action is put into the target Q-network to acquire the true value as

$$y_t = r_{t+1} + \gamma Q\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta_t); \theta_t^-\right). \tag{62}$$

Then, BS-agent calculates the loss function through the mean squared error between the true values and the prediction values for the tuples as

$$L(\theta_t) = \mathbb{E}\left[(y_t - Q(s_t, a_t; \theta_t))^2\right]. \tag{63}$$

Later, weighting factors θ_t are updated in each step through backpropagation based on the gradient descent method as

$$\theta_{t+1} \leftarrow \theta_t + v \cdot \nabla_{\theta_t} L(\theta_t), \tag{64}$$

where v means the learning rate, which decides how much degree of deviation to learn. The gradient descent is defined as

$$\nabla_{\theta_t} L(\theta_t) = \mathbb{E}\left[(y_t - Q(s_t, a_t; \theta_t))\, \nabla_{\theta_t} Q(s_t, a_t; \theta_t)\right]. \tag{65}$$

The Q value is also updated in each time step:

$$Q(s_t, a_t; \theta_t) \leftarrow Q(s_t, a_t; \theta_t) + v \cdot (y_t - Q(s_t, a_t; \theta_t)). \tag{66}$$

Comparatively, weighting factors θ_t^- are only copied once in an episode through Polyak averaging method and represented as

$$\theta_t^- \leftarrow \rho\theta_t + (1-\rho)\theta_t^-, \tag{67}$$

where ρ is a hyperparameter, decides the soft update ability.

Fig. 4 shows the training process in one step or episode. All the above episodes (called one epoch) training acts on sub-slot 1, where BS-agent only considers partial constraints (36k) and (36m) in the environment. For

sub-slot 2, BS-agent should switch to the reduced action space (i.e., the muting orders are excluded.) and retrain the network with constraints (36k)- (36n). Since the environment in each learning process varies, the outputs of the prediction from the neural network in each sub-slot differ due to different environment interactions. As a consequence, two DDQNs should be trained separately, one DDQN for sub-slot 1 and the other for sub-slot 2, to maintain two groups of weighting factors. Considering that the training process at sub-slot 2 is similar to that at sub-slot 1, the procedure is not described in detail. In conclusion, in the training process, one iteration of training is the equivalent of a single pass for a time step. It is the same as one interaction with the environment, such as the involved sub-slot 1 or 2, through the given combined action at the corresponding time step.

Algorithm 2 Hybrid DRL-Convex Method

1: Set $D = 1$.
2: **Repeat** each DDQN:
3: Set $i = 1$.
Initialize $\theta_{t,D}$, $\theta^{-}_{t,D}$, $\mathcal{A}_D$.
4: **Repeat** each episode:
Set $t = 1$.
Initialize state s_1.
5: **Repeat** each step:
Use (61) to select $a_t = (a^q, a^{\hat{b}})$.
6: Set $n = 0$.
Generate the initial value of $\mathcal{W}^{(0)}$, $\mathcal{P}^{(0)}$.
7: **Repeat** each FOTA using (50):
8: Set $n = n+1$.
9: **Until** $\left|R^{(n)}_{2,D}(\mathcal{W},\mathcal{P})_i - R^{(n-1)}_{2,D}(\mathcal{W},\mathcal{P})_i\right| < \zeta$ or reach the max iteration.
10: Solve (53) with $R^{(n)}_{2,D}(\mathcal{W},\mathcal{P})_i$ to acquire the approximate maximum total SE: $R_D(\mathcal{W},\mathcal{P})_i$.
11: $R_D(\mathcal{W},\mathcal{P})_i$ assigned to r_{t+1}.
12: Store tuple $(s_t, a_t, r_{t+1}, s_{t+1})$ in R.
13: Pick up batch b from R.
14: Select a_{t+1} with maximum online Q value from b.
15: Use (63) to calculate the loss function.
16: Use (64) and (66) to update $\theta_{t,D}$ and Q value, respectively.
17: Set $t = t + 1$.
18: **Until** reach the max steps.
19: Use (67) to update $\theta^{-}_{t,D}$.
20: Set $i = i + 1$.
21: **Until** reach the max episodes.
22: Return $\theta_{t,D}$, $\theta^{-}_{t,D}$, r^{*}_{t+1}.
23: $D = D + 1$.
24: If $D = 2$, output muted users information.
25: **Until** $D > 2$.

Fig. 5 shows the information transmission between the two DDQNs, where the muted users information exchanges after DDQN_1 ends the training for the purpose of compensation to the muted users. At the end of a session, DDQNs will acquire convergent weighting factors for each neural net. With the trained neural nets, BS-agent has finally grasped a skill from the environment to select a suboptimal action. Meanwhile, the real-time performance is guaranteed as the samples can be trained off policy.

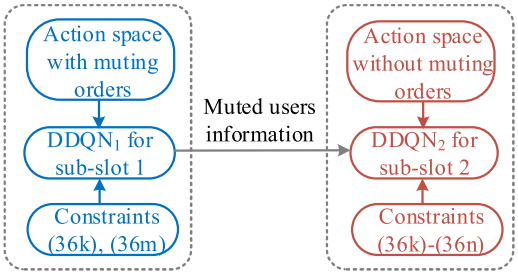

FIGURE 5. Information transmission between two DDQNs.

The hybrid DRL-convex method is summarized in Algorithm 2, where $\theta_{t,D}(\theta^{-}_{t,D})$ represents the weighting factors in the online (target) Q-network of DDQN_D ($D \in \{1, 2\}$), and the italic R_D indicates SE at sub-slot D. $\mathcal{A}_D$ stands for the action space at sub-slot D.

C. COMPLEXITY ANALYSIS

For expanded parameters, we select $2(2N)^{\mathrm{r_Q}}$ and $(2ZM)^{\mathrm{r}_{\hat{\mathbf{B}}}} + (2Z^2M)^{\mathrm{r}_{\hat{\mathbf{B}}}}$ samples from the universal sets in the greedy algorithm. Additionally, the time complexity of the FOTA method is $\mathcal{O}(nMJ(J+K))$. To sum up, the time complexity of the hybrid greedy-convex method (HGC) is $\mathcal{O}(nI_{\max}MJ(2(2N)^{\mathrm{r_Q}} + (2ZM)^{\mathrm{r}_{\hat{\mathbf{B}}}} + (2Z^2M)^{\mathrm{r}_{\hat{\mathbf{B}}}})(J+K))$. It is obvious that $\mathrm{r_Q}$ and $\mathrm{r}_{\hat{\mathbf{B}}}$ mainly determine the exponential computational complexity [47]. To evaluate the effectiveness of HGC, we take the hybrid exhaustive-convex method (HEC) as a baseline. Since exhaustive search is incompetent to traverse all combinations in the expanded parameters, to realize the method's feasibility, we keep the same iterations for exhaustive search and greedy algorithm for the sake of fairness. In this regard, the time complexity of HGC equals to that of HEC.

As the training process for the DRL method relates to many factors (e.g., the kernel size, size of the feature map, and number of channels for input and output), it is tough to provide an accuracy complexity. But from the point of each iteration view, the time complexity depends on the number of episodes (namely E) and steps (namely T). Thus, the time complexity of the hybrid DRL-convex method (HDC) is $\mathcal{O}(2nETT_{\text{iter}}MJ(J+K))$, where T_{iter} indicates the complexity in one iteration.

Compared with the exponential computational complexity of HGC/HEC, the complexity of HDC is much less in the scenario of high dimensions. Moreover, when DDQN_1 is trained well, it is unnecessary to retrain the DDQN_1 unless there are significant changes in the interference model. Accordingly, the time complexity drops to $\mathcal{O}(nETT_{\text{iter}}MJ(J+K))$. Since

TABLE 1. Simulation parameters.

Notation	Simulation value	description
σ^2_{SI}	-100 dB	SI power ratio of pre-SIC to post-SIC
$\sigma^2_{\mathrm{u},j}$	-107 dBm	noise power at UL user j
$\sigma^2_{\mathrm{d},k}$	-107 dBm	noise power at DL user k
a	0 dB	Rician factor
$p_{\max}$	22 dBm	UL/DL maximum transmit power for each user
$P_{\max}$	45 dBm	maximum transmit power for BS
$\mathrm{R}^{\mathrm{u}}_{\mathrm{req}}$	10 bps/Hz	QoS requirement for UL user
$\mathrm{R}^{\mathrm{d}}_{\mathrm{req}}$	10 bps/Hz	QoS requirement for DL user
β	1.4	compensation coefficient
$I_{\max}$	10^5	maximum candidate members
ω_{th}	500	tolerance threshold
$\mathrm{r}_{\mathbf{Q}}$	3	resolution ratio of sampling for $\mathbf{Q}$
$\mathrm{r}_{\hat{\mathbf{B}}}$	2	resolution ratio of sampling for $\hat{\mathbf{B}}$
ζ	0.001	convergence factor
γ	0.99	discount factor
v	0.001	learning rate
ρ	0.005	Polyak averaging
b	512	batch size
R	2.5×10^7	replay buffer

parameter dimensionality is highly related to the number of neurons and hidden layers in the deep neural network, we tend to apply a sophisticated network to cover the complex parameters and improve performance.

In the following section, we additionally introduce HDC/HGC/HEC without MM for comparison. The related complexity is $\mathcal{O}(nETT_{\mathrm{iter}}MJ(J+K))$ and $\mathcal{O}(nI_{\max}MJ((2N)^{\mathrm{r}_{\mathbf{Q}}}+(2ZM)^{\mathrm{r}_{\hat{\mathbf{B}}}})(J+K))$ for HDC and HGC (HEC) each. Although the complexity of HGC (HEC) without MM has decreased more than that of HDC without MM, that of HGC (HEC) without MM is still higher.

VI. PERFORMANCE EVALUATION

A. SIMULATION PARAMETERS

In this section, we illustrate multiple numerical results to evaluate the performance of our proposal. We assume Z users uniformly spread in a square with a side of 50 m, where the FD BS is located in the center. The FD BS is equipped with N smart antennas, while each user only provides one HD antenna. To simplify the experiment, we suppose all users are scheduled by BS (i.e., $\alpha^{\chi}_z(t_1)=1,2,\alpha^{\chi}_z(t_2)=2,\forall z\in\mathcal{Z}$), and half of the stochastic scheduled users receive $\mathrm{Sch}^d_z(t_1)=\mathrm{Sch}^d_z(t_2)=1$, at the same time the other half get $\mathrm{Sch}^{\mathrm{u}}_z(t_1)=\mathrm{Sch}^{\mathrm{u}}_z(t_2)=1$. M mutual orthogonal subcarriers are reused in the network. The detailed parameters information of the interference model refers to Table 1.

For the DRL method, we train the proposed DDQNs by using Python 3.6, TensorFlow-gpu 1.14, and Keras 2.1.6 for 5000 episodes with 2500 steps each. Each DDQN has three fully connected layers by applied dropout. There are 1024, 512, and 256 neurons, followed by the Relu activation functions in each hidden layer. The hyperparameters and other parameters also refer to Table 1.

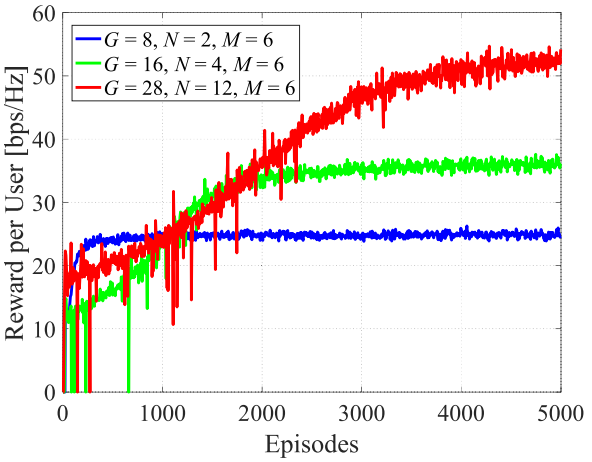

FIGURE 6. Training process of HDC with different parameter configurations at sub-slot 1.

B. RESULTS

1) ANALYSIS OF CONVERGENCE SPEED

Because there is no training process in HGC/HEC and both DDQNs are nearly the same training process, we only present HDC at sub-slot 1 in Fig. 6.

From Fig. 6, we can see that the reward per user for each configuration gradually increases and eventually converges to a relatively steady maximum value. When the reward stops growing, it indicates that the neural network has been trained well. Moreover, the more complicated parameter will incur a lower training speed and a more fluctuating final reward. It is clear that the algorithm convergence is bound up with the dimensionality of parameters. For instance, a neural network with $G=8, N=2, M=6$ undergoes 300 episodes to train stabilized, while 4000 episodes are required to train a neural network with $G=28, N=12, M=6$.

In the following subsections, we will further emphasize the superiority of the HDC algorithm with MM for various parameters in detail.

Note that the mentioned HDC, HGC, or HEC algorithm in the previous sections embodies MM by default. In the following subsections, the case of MM not included is regarded as the reference, so we will stress the condition of whether MM is introduced or not elaborately.

2) ANALYSIS OF THE NUMBER OF SCHEDULED USERS

From Fig. 7, since no JU appears at $G=8$, the total SE in a schedule unit under this case is served as a baseline for other cases of different numbers of scheduled users to compare with, thus highlighting the total SE gain. It can be seen that the increases of both the number of scheduled users and total SE gain are asymmetrical. For example, the total SE gain

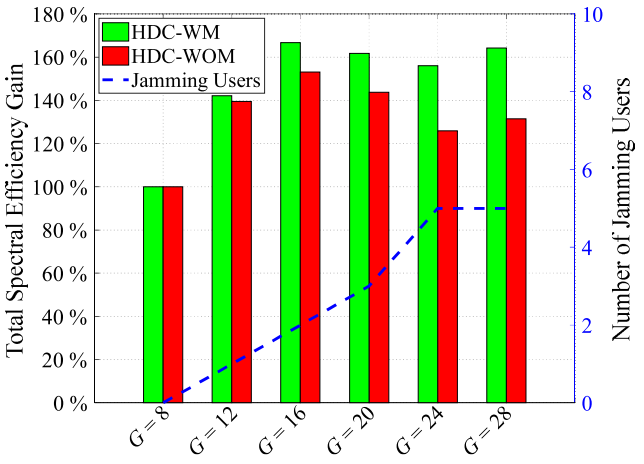

FIGURE 7. Emergence of JUs and performance comparisons at different numbers of scheduled users with $N = 4$, $M = 6$.

of HDC without MM (namely, HDC-WOM) at $G = 28$ is only 131.5%, while the number of scheduled users rises to 350% compared with that at $G = 8$. On the other hand, from $G = 8$ to 16, more scheduled users located in the current cell will make total SE growth. However, this upward tendency stops at $G = 20$. The above two observations suggest that the total SE is seriously restricted by the tight resource situation, where the spectrum resources are insufficient to maintain the scheduled users. So mutual interference becomes the villain of the performance exacerbation.

Compared with HDC-WOM, the HDC with MM (HDC-WM) shows a relatively robust performance advantage. The reason is that muting the JUs helps alleviate the spectrum resource competition, such as five JUs having been muted at $G = 28$. This action will bring HDC-WM more incremental gains. Meanwhile, it illuminates that the appeared probability of JUs increases as the number of scheduled users mounts. Consequently, the proposed MM successfully seeks the tradeoff between total SE and total scheduled users by muting several JUs.

For the convenience of performance comparison, we adopt the average total SE per scheduled user instead of the total SE as a performance metric below. Compared with Fig. 7, Fig. 8 presents the relationship between the performance and the number of scheduled users from another point of view. Because of the scarce spectrum resources, the performance of each algorithm is decreasing monotonously, correlated with raising the number of scheduled users. As expected, the HDC-WM outperforms the HDC-WOM. To be specific, the performance distinctness becomes more evident (from 0 to 25.8% gain) as the number of scheduled users increases. The reason is that the number of JUs determines the MM's marginal increment level (see Fig. 7).

Generally, both the HDC algorithms get better performance than the HGC/HEC algorithms since HDC has a more robust convergence owing to lower complexity. The gap

TABLE 2. Performance comparison.

Algorithm	Average spectral efficiency per user [bps/Hz]
	$G = 28$, $N = 12$, $M = 6$
HDC-WM	95.908
HDC-WOM	88.623
HGC-WM	36.227
HGC-WOM	63.352
HEC-WM	-
HEC-WOM	41.39

between the HGC with MM (namely, HGC-WM) and the HGC without MM (namely, HGC-WOM) is inconspicuous, which is 3 bps/Hz at most, so does the discrepancy between the HEC with MM (i.e., HEC-WM) and the HEC without MM (i.e., HEC-WOM).

For the greedy algorithm in HGC, limited candidates lead the trap in local optimum, thus decreasing the advantage of MM. Noticeably, the HGC-WM is outperformed by the HGC-WOM at $G = 20$ and even can not meet the minimum QoS for all users at $G = 28$. It is caused by the higher complexity of HGC-WM than that of HGC-WOM as the dimensionality expands. To be specific, the extra complexity of HGC-WM compared with HGC-WOM is $\mathcal{O}(nI_{\max}MJ((2N)^{r_Q} + (2Z^2M)^{r_{\hat{B}}})(J + K))$. In our experiment settings, the excess part can be rewritten as $\mathcal{O}(nI_{\max}MJG((2N)^{r_Q} + (2G^2M)^{r_{\hat{B}}}))$, which has an influential role rather than profits brought by MM in the case of large parameters. Since HEC has an identical complexity to HGC, the handling ability of HEC is similar to HGC. It also proves that MM in HGC or HEC is no longer privileged in terms of complicated parameters. It is noteworthy that, for a learning method, the complexity between HDC-WM and HDC-WOM is nearly the same. It only depends on T_{iter}, which can be ignored in a learning process. Thus the advantage of MM can be well displayed in HDC.

For HEC, The performance of HEC is worse than that of HGC. Although the complexity is equal between them, the exhaustive search in HEC directly seeks the global optimum with randomness and blindness, which is a lack of efficiency compared with adopting a top-down design in HGC. Therefore, the HEC can not cope with $G = 24$.

3) ANALYSIS OF THE NUMBER OF ANTENNAS

Fig. 9 shows that the performance is positively correlated with the number of antennas at BS. That is to say, configuring large antennas is a straightforward means of promoting spatial diversity gain. The HDC-WM achieves 6 bps/Hz marginal increment compared with the HDC-WOM, and the performance gap between them becomes virtually static along with the number of antennas. Combining Fig. 8, an alternate view points out that the MM performance gain mainly depends on the number of scheduled users.

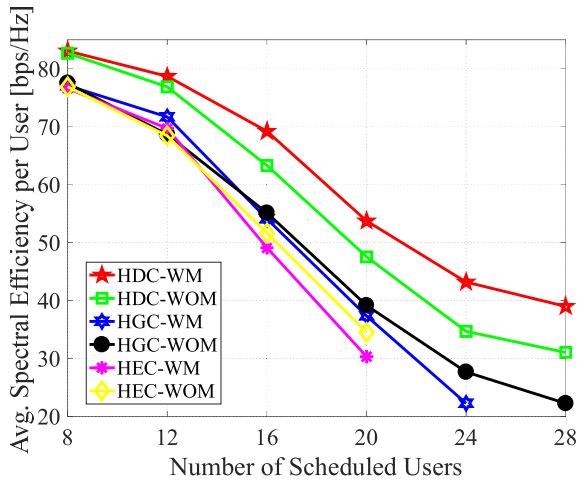

FIGURE 8. Performance comparisons at different numbers of scheduled users with $N = 4$, $M = 6$.

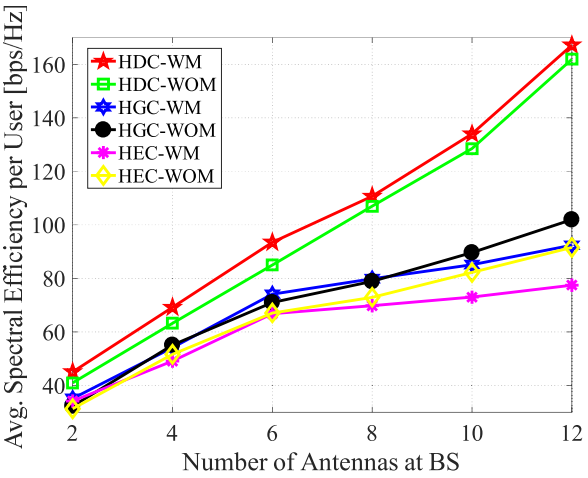

FIGURE 9. Performance comparisons at different numbers of antennas at BS with $G = 16$, $M = 6$.

The margin between HGC-WM and HGC-WOM is obscure at $N = 2/4/6/8$. Similarly to Fig. 8, too much calculation causes the HGC performance degradation in the case of large antennas, so HGC represents a moderative performance growth. Additionally, at $N = 12$, the HGC-WM is surpassed by HGC-WOM over 11 bps/Hz, stemming from the superiority in MM nullified by extra complexity. It is worth noting that the disparity between HEC-WM and HEC-WOM is more conspicuous than that between HGC-WM and HGC-WOM at $N = 12$. The result illustrates that when tackling large parameters, the HEC-WM is inferior to HGC-WM.

The performance distinctness between HDC and HGC/HEC increases with the number of antennas. The minimum gap between them is less than 10 bps/Hz at $N = 2$, while the peak discrepancy is over 60 bps/Hz at $N = 12$. The result demonstrates that HDC has superiority in tackling multiple antennas.

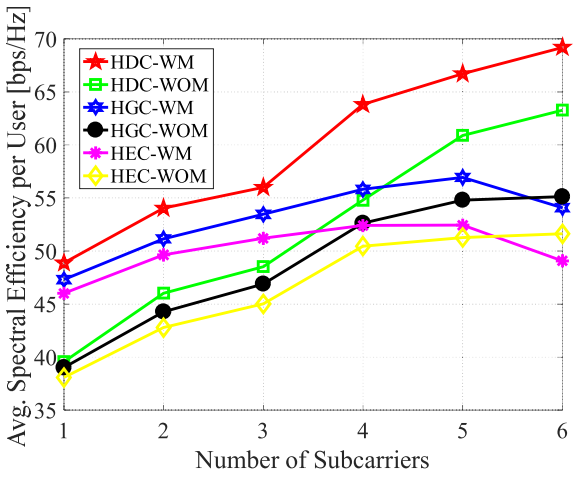

FIGURE 10. Performance comparisons at different numbers of subcarriers with $G = 16$, $N = 4$.

4) ANALYSIS OF THE NUMBER OF SUBCARRIERS

Fig. 10 depicts that the relative abundant subcarriers in the current network will provoke HDC into a better SE. It is obvious that more spectrum resources will decrease all kinds of interference. Nevertheless, the discrepancy between HDC-WM and HDC-WOM is narrowed with the number of subcarriers increasing. The reason is that the influence of JUs to the entirety has declined, owing to the relatively adequate spectrum resources. To be specific, the gap is 9 bps/Hz at $M = 1$, while it has a 3 bps/Hz drop at $M = 6$. Notice that our proposed MM is more appropriate in scenarios of scarce spectrum resources.

Although HGC for the two algorithms also displays a growing performance trend at $M = 1$ to 4, the ascending trend is suppressed or reversed at $M = 5, 6$. This is due to the fact that, for HGC, the profit derived from spectrum resources is less than the loss of calculation at the related configurations. Significantly, the phenomena that HDC-WOM begins to outperform HGC-WM at $M = 5$, emphasizes the conclusion in Fig. 8.

For HEC, the exasperate performance trend is more acute when more subcarriers are assigned. It proves that the traditional exhaustive search can be not fully applied to the scenario of two discrete variables.

From Fig. 7-10, we can see that the algorithm performance is highly correlated with the size of parameters, such as the number of scheduled users, antennas, and subcarriers. To fully evaluate the performance comparison between the proposed and traditional algorithms, we set the upper bound of parameters as $G = 28$, $N = 12$, and $M = 6$ in our simulations. The numerical results are presented in Table 2. It validates that only the proposed HDC with MM can gain better performance than that without MM in the case of extended parameters. Although the proposed HGC is worse than HDC, it shows relative robustness due to adopting a top-down structure rather than HEC.

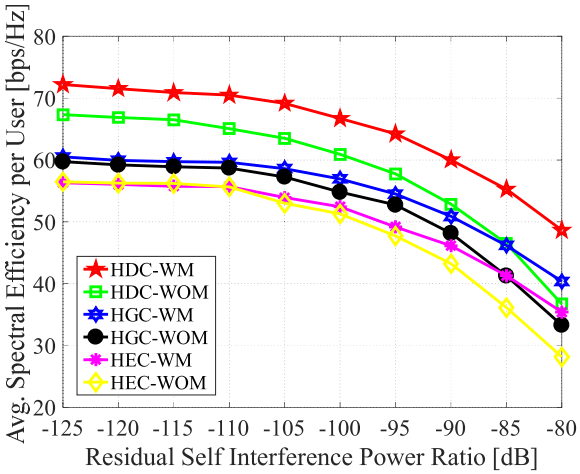

FIGURE 11. Performance comparisons at different residual self-interference power ratios with $G = 16$, $N = 4$, $M = 5$.

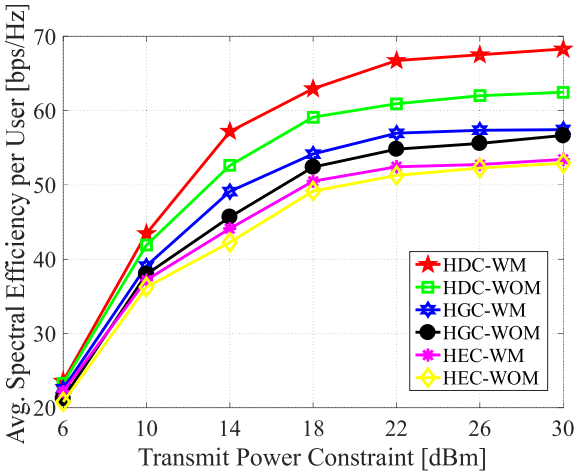

FIGURE 12. Performance comparisons at different uplink/downlink transmit power constraints with $G = 16$, $N = 4$, $M = 5$.

5) ANALYSIS OF RESIDUAL SELF-INTERFERENCE

The above analysis mainly focuses on the three factors, which are variable-related parameters in the objective function. In the following simulations, we will further analyze other influencing factors, such as the parameter σ_{SI}^2 of the interference model and parameters $p_{\max}$, β in the constraints.

Fig. 11 displays the average SE per user for different residual SI power ratios. It can be observed that the impact of residual SI on performance is almost weak when the ratio is less than -110 dB. Performance deteriorates drastically when the ratio exceeds -90 dB. It proves the essential of SIC technology and that the stronger residual SI can worsen the FD communication system. The performance gap between HDC-WM and HDC-WOM is 5 bps/Hz at first, which gradually increases as the ratio increases. Nevertheless, the growing trend is not obvious at the outset till the ratio is beyond -105 dB. On the other side, the performance enhancement between the two algorithms is 13 bps/Hz at -80 dB. The finding means that the strong residual SI would strengthen the JUs' adverse effects. As a result, applying MM will gain a prominent performance in the situation of stronger residual SI.

The performance tendency of HGC/HEC is similar to that of HDC. The main distinction is that the performance gain brought by MM in HGC/HEC is lower than that in HDC. It also results from the higher complexity of HGC-WM/HEC-WM than that of HGC-WOM/HEC-WOM, as we present in the complexity analysis.

6) ANALYSIS OF MAXIMUM TRANSMIT POWER

Fig. 12 represents the effect of different $p_{\max}$ settings on SE. As is known that the degree of interference depends on transmit power directly. The lower transmit power makes the interference more negligible. Irrespective of interference to some degree, the non-learning methods can easily find the best member without too much traversal. Meanwhile, the influence of JUs is minimal. Accordingly, there is little difference among algorithms at $p_{\max} = 6$ dBm.

Thereafter, the performance profit brought by increasing power still outweighs the cost from interference. Especially at $p_{\max} = 10$ dBm, the performance boost is remarkable. It means that the state of a low power level accompanied by a low interference floor condition will have a greater potential to improve performance. This circumstance terminates at $p_{\max} = 22$ dBm, where SE nearly reaches the peak for each algorithm. If the $p_{\max}$ continues increasing, interference will become the dominating factor to impact performance. For this reason, the transmit power of each user and BS should stop to elevate, so as to avoid performance degradation.

Note that the performance gap between HDC-WM and HDC-WOM gradually increases till convergence as increasing power budget. For instance, the gap between two HDCs is 0 and 7 bps/Hz at $p_{\max} = 6$ and 22 dBm, respectively. The reason is that the side effect of JUs at an upper power level is much greater than that at a comparatively low level. Correspondingly, there is a clear superiority with MM applied in the case of strong JUs. On the other hand, when interference becomes stronger, the exhaustive search and greedy method are tough to tackle MM, especially in the exhaustive search, so the advantage of MM is inconspicuous for them.

For the different types of algorithms (i.e., HEC, HGC, and HDC), their performance is mainly restricted to the parameter sizes of scheduled users, antennas, and subcarriers, while parameters of residual self-interference and maximum transmit power are irrelevant to complexity. As a result, the variation tendencies among different algorithms in Fig. 11 and Fig. 12 are more consistent than in Fig. 8, Fig. 9, and Fig. 10.

7) ANALYSIS OF COMPENSATION COEFFICIENT

In Fig. 13, we show the effect of each compensation coefficient on performance. Because compensation factors do not work on algorithms without MM, the algorithms maintain the same SE regardless of the coefficients. Note that each performance of algorithms with MM decreases monotonically with an increase in compensation coefficients. It demonstrates

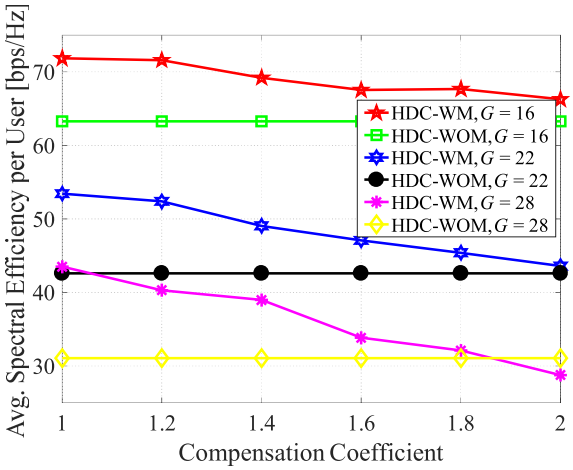

FIGURE 13. Performance comparisons at different compensation coefficients with $G = 16/22/28$, $N = 4$, $M = 6$.

that the remedy of muting users against interruption would intensify interference at sub-slot 2. That is, the means of compensation merely attends to the JUs at the expense of the entirety.

Since more scheduled users in the current cell would arouse more JUs, the decay of SE is more significant with increasing the compensation coefficient at a larger G. If we assume $G = 28$, the performance of HDC-WM is even worse than that of HDC-WOM at $\beta = 2$. In view of the principle of fairness, the value can not be higher than an upper limit. We set $\beta = 1.4$ to attempt to cover the interests of the majorities and the individuals.

VII. CONCLUSION

In this paper, a novel power allocation method in the FD multi-user MIMO system has been studied. Specifically, in this scheme, three major factors (such as smart antennas, scheduled users, and subcarriers) related to power allocation have been considered. To further improve performance from user level, we introduce MM by modifying the frame structure to alleviate interference related to JUs. After formulating the problem of the optimal power allocation method, we propose a hierarchical algorithm method concerning different types of variables. That is, the subproblem of continuous variables is solved through FOTA. Meanwhile, the other of discrete variables is addressed through the greedy algorithm based on the traditional exhaustive search. Considering the high computational complexity of the greedy algorithm when applied with extended parameters, we devise the jointed DRL method to obtain a better performance. The DRL method contains two DDQNs. One DDQN is used to train samples in sub-slot 1, and the other is applied at sub-slot 2. Simulation results reveal that HDC outperforms HGC/HEC in three main aspects. Meanwhile, compared with non-introduced MM, the case with introduced MM has achieved a performance enhancement due to degrading the side effect of JUs. In conclusion, our proposal offers a new way to improve SE in multi-user MIMO FD Systems. One possible extension of this work is to develop an improved DRL scheme to further optimize the performance in the scenario with massive users and antennas.

REFERENCES

[1] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[2] K. E. Kolodziej, B. T. Perry, and J. S. Herd, "In-band full-duplex technology: Techniques and systems survey," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 7, pp. 3025–3041, Jul. 2019.

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, 2019.

[4] S. H. Chae and K. Lee, "Degrees of freedom of full-duplex cellular networks: Effect of self-interference," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4507–4518, Oct. 2017.

[5] A. T. Abusabah, L. Irio, R. Oliveira, and D. B. da Costa, "Approximate distributions of the residual self-interference power in multi-tap full-duplex systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 755–759, Apr. 2021.

[6] H. Zhao, U. De Silva, S. Pulipati, S. B. Venkatakrishnan, S. Bhardwaj, J. L. Volakis, S. Mandal, and A. Madanayake, "A broadband multistage self-interference canceller for full-duplex MIMO radios," *IEEE Trans. Microw. Theory Techn.*, vol. 69, no. 4, pp. 2253–2266, Apr. 2021.

[7] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: Challenges, solutions, and future research directions," *Proc. IEEE*, vol. 104, no. 7, pp. 1369–1409, Jul. 2016.

[8] M. H. N. Shaikh, V. A. Bohara, and A. Srivastava, "Performance enhancement in full-duplex AF relay system through smart antenna allocation," in *Proc. IEEE 3rd 5G World Forum (5GWF)*, Sep. 2020, pp. 303–308.

[9] K. Min, S. Park, Y. Jang, T. Kim, and S. Choi, "Antenna ratio for sum-rate maximization in full-duplex large-array base station with half-duplex multiantenna users," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10168–10173, Dec. 2016.

[10] H. Gao, Y. Su, S. Zhang, Y. Hou, and M. Jo, "Joint antenna selection and power allocation for secure co-time co-frequency full-duplex massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 655–665, Jan. 2021.

[11] K. Yang, H. Cui, L. Song, and Y. Li, "Efficient full-duplex relaying with joint antenna-relay selection and self-interference suppression," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3991–4005, Jul. 2015.

[12] X. Xia, P. Zhu, J. Li, H. Wu, D. Wang, and Y. Xin, "Joint optimization of spectral efficiency for cell-free massive MIMO with network-assisted full duplexing," *Sci. China Inf. Sci.*, vol. 64, no. 8, pp. 1–16, Aug. 2021.

[13] G. Liu, H. Ji, F. R. Yu, Y. Li, and R. Xie, "Energy-efficient resource allocation in full-duplex relaying networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 2400–2405.

[14] R. Aslani, M. Rasti, and A. Khalili, "Energy efficiency maximization via joint sub-carrier assignment and power control for OFDMA full duplex networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 11859–11872, Oct. 2019.

[15] Z. Liu and S. Feng, "Joint subcarrier assignment and power allocation for OFDMA full duplex distributed antenna systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11554–11564, Nov. 2021.

[16] F. J. Martin-Vega, M. C. Aguayo-Torres, G. Gomez, and M. Di Renzo, "Interference-aware muting for the uplink of heterogeneous cellular networks: A stochastic geometry approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[17] W. Xie, X. Xia, Y. Xu, K. Xu, and Y. Wang, "Massive MIMO full-duplex relaying with hardware impairments," *J. Commun. Netw.*, vol. 19, no. 4, pp. 351–362, Aug. 2017.

[18] Y. Liu, X. Xue, J. Zhang, X. Li, L. Dai, and S. Jin, "Multipair massive MIMO two-way full-duplex relay systems with hardware impairments," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.

[19] D. Li, D. Zhang, and G. Zhang, "Degrees of freedom for half-duplex and full-duplex multi-user cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2812–2827, Mar. 2020.

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

[21] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
[22] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2020.
[23] Q. Cheng, Z. Wei, and J. Yuan, "Deep reinforcement learning-based spectrum allocation and power management for IAB networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.
[24] A. Iqbal, M.-L. Tham, and Y. C. Chang, "Double deep Q-network-based energy-efficient resource allocation in cloud radio access network," *IEEE Access*, vol. 9, pp. 20440–20449, 2021.
[25] M. Lu, B. Zhou, Z. Bu, and Y. Zhao, "A learning approach towards power control in full-duplex underlay cognitive radio networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 2017–2022.
[26] Y. Al-Eryani, M. Akrout, and E. Hossain, "Antenna clustering for simultaneous wireless information and power transfer in a MIMO full-duplex system: A deep reinforcement learning-based design," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2331–2345, Apr. 2021.
[27] M. T. Mamaghani and Y. Hong, "Intelligent trajectory design for secure full- duplex MIMO-UAV relaying against active eavesdroppers: A model-free reinforcement learning approach," *IEEE Access*, vol. 9, pp. 4447–4465, 2021.
[28] S. Huang, Y. Ye, and M. Xiao, "Learning-based hybrid beamforming design for full-duplex millimeter wave systems," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 120–132, Mar. 2021.
[29] R. Wang, A. Yadav, E. A. Makled, O. A. Dobre, R. Zhao, and P. K. Varshney, "Optimal power allocation for full-duplex underwater relay networks with energy harvesting: A reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 223–227, Feb. 2020.
[30] C. Dai, K. Zhu, and E. Hossain, "Multi-agent deep reinforcement learning for full-duplex multi-UAV networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 2232–2237.
[31] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637–1652, Sep. 2014.
[32] J. M. B. D. Silva, H. Ghauch, G. Fodor, M. Skoglund, and C. Fischione, "Smart antenna assignment is essential in full-duplex communications," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3450–3466, May 2021.
[33] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layerprocedures*, document 3GPP TS 36.213 Version 15.4.0 Release 15, Sep. 2018.
[34] K. Lee, Y. Park, M. Na, H. Wang, and D. Hong, "Aligned reverse frame structure for interference mitigation in dynamic TDD systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6967–6978, Oct. 2017.
[35] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4296–4307, Dec. 2012.
[36] D. Nguyen, L.-N. Tran, P. Pirinen, and M. Latva-Aho, "On the spectral efficiency of full-duplex small cell wireless systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 4896–4910, Sep. 2014.
[37] E. Cela, *The Quadratic Assignment Problem: Theory and Algorithms*, vol. 1. Berlin, Germany: Springer, 2013.
[38] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, no. 1, pp. 29–51, Jul. 2010.
[39] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.
[40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[41] M. Grant and S. Boyd. (Jan. 2020). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.2*. [Online]. Available: http://cvxr.com/cvx
[42] B. Setho Kusuma Sakti, A. Fahmi, V. Sigit Widhi Prabowo, and D. Putra Setiawan, "Radio resource management for improving the spectral efficiency on D2D underlaying communications using a modified joint-greedy algorithm," in *Proc. 6th Int. Conf. Sci. Technol. (ICST)*, Sep. 2020, pp. 1–5.
[43] S. Peng, "Stochastic hamilton–Jacobi–Bellman equations," *SIAM J. Control Optim.*, vol. 30, no. 2, pp. 284–304, 1992.
[44] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[45] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, and J. Veness, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[46] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, Sep. 2016, pp. 2094–2100.
[47] E. Bender and S. Williamson, *Foundations of Combinatorics With Applications*. New York, NY, USA: Dover, 2006.

KUNBEI PAN received the B.Sc. degree from Hainan Tropical Ocean University, in 2012, and the M.Sc. degree from Hohai University, in 2015. He is currently pursuing the Ph.D. degree in communications and information systems with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences.

He worked as a Senior Communication Protocol Development Engineer at MediaTek Inc., from 2015 to 2021. The main work content involves wireless baseband chip research and development in 4G and 5G communication systems. His research interests include full-duplex wireless systems, satellite communication, and network intelligence.

BIN ZHOU received the B.Sc. and M.Sc. degrees in communications and information systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2004, respectively, and the Ph.D. degree in communications and information systems from the University of Chinese Academy of Sciences, Shanghai, China, in 2012.

He was with System Research, Nokia Technology Platform, Oulu, Finland, until October 2006. He is currently a Full Professor with the Shanghai Institute of Microsystem and Information Technology and the Key Laboratory of Wireless Sensor Network and Communications, Chinese Academy of Sciences, Shanghai. He is the first inventor of more than 50 Patent Cooperation Treaty or regional patents. His research interests include full-duplex wireless systems, mesh and relay networks, D2D communications, and signal processing. As the first author, he received the IEEE VTS Jack Neubauer Memorial Award for the Best System Paper, in 2016.

ZHIYONG BU received the B.Sc. and M.Sc. degrees from Sichuan University, Chengdu, China, in 1993 and 1996, respectively, and the Ph.D. degree in communications and information systems from the State Key Laboratory of Mobile Communication, Southeast University, Nanjing, China, in 1999.

He is currently a Full Professor and the Chief Engineer with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai. He is also the Director of the Key Laboratory of Wireless Sensor Network and Communications, Chinese Academy of Sciences. His research interests include communications, signal processing, and networking, with an emphasis on artificial intelligence, satellite communications, and 5G networks.