Deep Reinforcement Learning for Distributed Dynamic MISO Downlink-Beamforming Coordination

Jungang Ge, *Student Member, IEEE*, Ying-Chang Liang, *Fellow, IEEE*,
Jingon Joung, *Senior Member, IEEE*, and Sumei Sun, *Fellow, IEEE*

***Abstract*—We consider a homogeneous cellular network where a multi-antenna base station (BS) in each cell transmits messages to its intended user over a common frequency band. To improve the system capacity of this multi-cell multi-input single-output (MISO) interference channel, one of the state-of-the-art algorithms, namely, downlink-beamforming coordination, allows all BSs to cooperate with one another to mitigate the effect of inter-cell interference. However, most existing algorithms are suboptimal and impractical in a dynamic wireless environment, due to the high computational complexity and the overhead involved in collecting global channel state information (CSI). In this study, we exploit deep reinforcement learning (DRL) and propose a distributed *dynamic downlink-beamforming coordination* (DDBC) method with partial observability of the CSI. Each BS is able to train its own deep Q-network and employs appropriate beamformer depending on its environment, which is observed through a designed limited-information exchange protocol. The simulation results show that the proposed DRL-based DDBC method, with a considerably lower system overhead, achieves a system capacity that is very close to that of the fractional programming algorithm with global and instantaneous CSI measurements. In addition, this work demonstrates the potential of utilizing DRL to solve DDBC problems in a more practical manner.**

***Index Terms*—Downlink-beamforming coordination, multi-input single-output (MISO) interference channel, deep reinforcement learning, interference mitigation.**

Manuscript received January 1, 2020; revised April 9, 2020 and June 8, 2020; accepted June 17, 2020. Date of publication June 24, 2020; date of current version October 16, 2020. This work was supported by the National Natural Science Foundation of China under Grants 61631005 and U1801261, the National Key R&D Program of China under Grant 2018YFB1801105, the Key Areas of Research and Development Program of Guangdong Province, China, under Grant 2018B010114001, the Central Universities under Grant ZYGX2019Z022, the Programme of Introducing Talents of Discipline to Universities under Grant B20064, and in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (2018R1A4A1023826 & 2019R1A2C1084168). The associate editor coordinating the review of this article and approving it for publication was M. Vaezi. *(Corresponding author: Ying-Chang Liang.)*

Jungang Ge and Ying-Chang Liang are with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China (e-mail: gejungang@std.uestc.edu.cn; liangyc@ieee.org).

Jingon Joung is with the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, South Korea (e-mail: jgjoung@cau.ac.kr).

Sumei Sun is with the Institute for Infocomm Research, Singapore 138632 (e-mail: sunsm@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2020.3004524

I. Introduction

Conventional mobile communication systems are usually designed with a cellular architecture, which provides an effective way to mitigate inter-cell interference by letting the neighboring base stations (BSs) in different cells operate over orthogonal frequency bands. With the exponential increase in wireless data traffic during the past decades, the frequency reuse constraint in a cellular network has been relaxed to one; i.e., the entire frequency band is shared by all the cells. However, this may cause severe inter-cell interference, which degrades the system capacity. Hence, the interference should be managed appropriately, and multi-cell coordination is regarded as one of the promising interference management techniques.

On the other hand, downlink beamforming technique has also attracted considerable attention from the industry and academia because it improves the performance of multiple antenna in downlink transmissions [1]. The downlink-beamforming coordination problem in cellular communication systems has been formulated from various viewpoints. Some techniques that allow the user data to be jointly processed by several interfering BSs have been proposed to improve the system capacity by exploiting inter-cell interference [2]–[4]. In addition, many interference management techniques have been developed with a practical constraint, i.e., the single-user detection mechanism [5]–[11]. This means that intra-cell and inter-cell interferences can be mitigated or even utilized to enhance the system capacity through various beamforming coordination techniques.

In general, the system capacity of a cellular communication system is represented by the sum of the achievable rate of all users, i.e., the sum rate. Evidently, the sum-rate maximization problem is nondeterministic polynomial (NP)-hard and nonconvex; hence, it is difficult to obtain the optimal solution. The rate-profile approach is an effective method to maximize the system capacity when a specific rate-profile vector is given [6]. The global maximum system capacity can be obtained by exhausting all possible rate-profile vectors, which is still an NP-hard problem. A number of suboptimal approaches have been developed to deal with this problem based on several optimization techniques, such as the fractional programming (FP) algorithm [11], weighted minimum-mean-square-error algorithm [12], and branch-and-bound

0090-6778 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

algorithm [13]. These algorithms are state-of-the-art approaches that achieve near-optimal performance; however, all of them have to be fed up-to-date global channel state information (CSI). Furthermore, they are usually designed with a centralized structure; i.e., a central controller is required to collect the global instantaneous CSI and then compute the near-optimal beamformers for all the BSs in the network. Therefore, these approaches are usually associated with a non-negligible delay due to the cascade procedure of collecting the CSI, computing beamformers, and sending beamformers to the corresponding BSs, and they may be applied to a static or quasi-static environment in various ways. However, in a dynamic wireless environment, which is a more practical scenario, their effectiveness severely reduced. Therefore, a more practical suboptimal approach with a low computation complexity and less requirement for CSI or other information needs to be developed urgently.

Reinforcement learning (RL) has been shown to be an emerging efficient technique to address various problems in many areas of communication and networking, such as the Internet of Things, heterogeneous networks (HetNets), and unmanned aerial vehicle networks [14]. In these networks, the network entities (e.g., BSs) are supposed to make their own decisions locally to optimize network performance, and an RL technique is usually employed to generate an optimal policy while making the decisions [15]–[19]. An RL-based enhanced inter-cell interference coordination approach was investigated for HetNets [15]. In [16] and [18], RL-based spectrum access approaches were studied for secondary users in cognitive radio networks. An RL technique was developed for an intelligent BS selection algorithm in a millimeter-wave HetNet [17] and applied to self-organized wireless networks to deal with user scheduling problems [19].

RL techniques have been found to be effective in many decision-making scenarios with moderate-size state and action spaces. In reality, however, especially in the wireless environment, there are typically enormous state-action pairs, for which it is hard to find the optimal policy between the agent and the environment within a finite number of interaction steps. Recently, with the significant advancements in deep learning (DL), a promising technique that combines DL and RL, namely, *deep reinforcement learning* (DRL), was developed, and it has shown great advantages in solving decision-making problems with large state-action spaces. DRL, which was proposed in [20], has attracted considerable attention from researchers in various fields, including wireless communications, in recent years [21]–[26]. A DRL technique was employed to find the optimal anti-jamming policy for the secondary user in cognitive radio networks in [21]. A DRL-based decentralized resource allocation technique was developed for vehicle-to-vehicle communications in [22]. In [23], a DRL-based intelligent coexistence algorithm was proposed as a solution to the spectrum sharing problem between a license-assisted access long-term evolution system and a WiFi system. The authors of [24] investigated the application of DRL for selecting a modulation coding scheme in a cognitive HetNet. In [25], two types of DRL-based intelligent user association schemes were proposed in symbiotic radio networks. In [26], the authors applied DRL techniques to the transmit power control problem and proposed a distributively executed dynamic power allocation scheme.

In this work, we propose a distributed *dynamic downlink-beamforming coordination* (DDBC) method based on the DRL technique, which allows all BSs to compute the most appropriate beamformers from their local observations of the dynamic environment in real time. In our setup, each BS is equipped with multiple antennas, and each user equipment (UE) is equipped with a single antenna. Because the UEs usually implement single-user detection, i.e., each UE decodes only the messages from its associated BS, the signals received from other BSs are treated as interference. Hence, the scenario considered in this study can be modeled as a multi-input single-output (MISO) interference channel (IC), termed MISO-IC. The main contributions and advantages of our proposed DRL-based distributed DDBC method are summarized as follows:

- We use an intuitive approach to propose a general framework for downlink data transmission in the dynamic cellular communication systems. The designed data transmission framework consists of two phases: i) the first phase is a preparing phase for the subsequent data transmission, and ii) the second phase is for data transmission.
- To overcome the disadvantages of the centralized approaches in the dynamic wireless environment, we propose a *distributed* DRL-based approach. To be specific, both the training process and execution process are carried out in a distributed manner. To adapt to the distributed nature of our proposed approach, in the first phase of our designed data transmission framework, we also propose a limited-information exchange protocol between neighboring BSs.
- The downlink beamforming coordination problem is a continuous vector optimization problem. The actions in DQL technique, however, are usually the discrete values. To tackle with this problem, we firstly decompose the beamformer into two parts, namely, the transmit power and the beam direction. Then, we propose to exploit the codebook technique to discretize the beam directions while the transmit power is discretized with a set of available discrete power levels.
- In our designed approach, all the local observations at the BSs are scalar values excluding the complex channel matrix. This means that the channel training process is not necessary for the proposed approach, and it can be applied to both time-division-duplex and frequency-division-duplex communication systems; however, the overhead of acquiring the channel matrix in frequency-division-duplex systems is very large.

The rest of this paper is organized as follows: Section II illustrates the system model. Section III formulates a DDBC problem, deduces a particular FP algorithm for the formulated problem, and introduces a designed limited-information exchange protocol. In Section IV, the basics of DRL are introduced, and a practical DRL-based DDBC approach is proposed. In Section V, we present the results of simulations

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

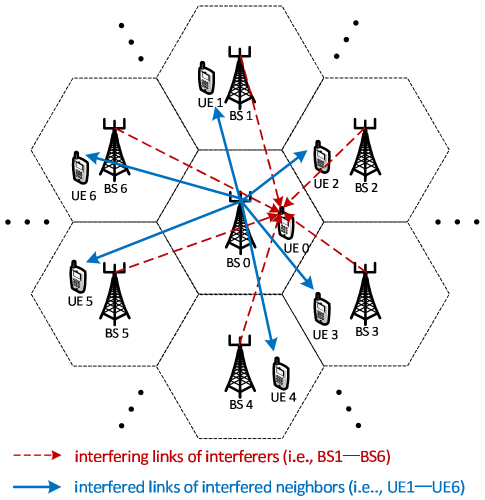

UE 1
BS 1
UE 6
BS 6
UE 2
BS 2
BS 0
UE 0
UE 5
BS 5
UE 3
BS 3
BS 4
UE 4
interfering links of interferers (i.e., BS1—BS6)
interfered links of interfered neighbors (i.e.,, UE1—UE6)

Fig. 1. Considered multi-cell MISO-IC model.

conducted to evaluate the effectiveness of our proposed algorithm. Finally, Section VI concludes the paper.

Notations: The notations used in this paper are as follows. To represent a column vector and matrix, bold lowercase and uppercase symbols are used (e.g., $\mathbf{x}$ and $\mathbf{X}$), respectively. $\mathbf{X}^{\dagger}$ and $\mathbf{X}^{-1}$ denote the Hermitian transposition and inverse of $\mathbf{X}$, respectively. $|x|$ represents the absolute value of x; $|\mathcal{X}|$ denotes the number of elements in set $\mathcal{X}$, i.e., the cardinality of $\mathcal{X}$; and $\|\cdot\|$ represents a Euclidean norm operator. $\mathbf{I}_x$ is an x-dimensional identity matrix. $\Re\{\cdot\}$ denotes the operation to obtain the real part. The set of real numbers is denoted by $\mathbb{R}$, while the set of the complex numbers is denoted by $\mathbb{C}$. Further, $\mathbf{x} \sim \mathcal{CN}(0, \mathbf{\Sigma})$ indicates that $\mathbf{x}$ is a complex Gaussian vector with zero mean and covariance matrix $\mathbf{\Sigma}$.

II. System Model

In this section, we introduce the multi-cell MISO-IC signal model and the channel model of the homogeneous cellular network considered in this paper.

A. Multi-Cell MISO-IC Signal Model

As shown in Fig. 1, we consider a downlink cellular network of K cells, in each of which there is a direct link consisting of a multi-antenna transmitter (i.e., the BS) and a single-antenna receiver (i.e., the UE). In each cell, multiple users are actually supported by the utilization of orthogonal frequency bands, and thus, there is no intra-cell interference. However, each user shares the same frequency band with the users in other cells, and inter-cell interference is present. Thus, by focusing on one frequency band, the system can be modeled as a multi-cell MISO-IC. Without loss of generality, we assume that all the BSs are equipped with a uniform linear array having N ($N \geq 1$) antenna elements. Hence, the received signal of the kth receiver at time t is written as follows ($k, j \in \mathcal{K} = \{1, \ldots, K\}$):

$$y_k(t) = \mathbf{h}^{\dagger}_{k,k}(t)\mathbf{w}_k(t)x_k(t) + \sum_{j \neq k} \mathbf{h}^{\dagger}_{j,k}(t)\mathbf{w}_j(t)x_j(t) + z_k(t), \tag{1}$$

where $\mathbf{w}_k(t) \in \mathbb{C}^{N\times 1}$ denotes the beamformer of BS k; $\mathbf{h}_{k,k}(t) \in \mathbb{C}^{N\times 1}$ denotes the direct downlink channel between UE k and its associated BS k; $\mathbf{h}_{j,k}(t) \in \mathbb{C}^{N\times 1}$ denotes the cross-link channel between UE k and BS j; $x_k(t)$ denotes the symbol transmitted by BS k; and $z_k(t) \sim \mathcal{CN}(0, \sigma^2)$ denotes the additive complex Gaussian noise at UE k. Under the assumption that each UE decodes only the message from its associated BS (i.e., the single-user detection mechanism), signals from the other BSs are treated as the interference. The instantaneous signal-to-interference-noise-ratio (SINR) and the achievable rate of UE k at time t are then written, respectively, as follows:

$$\psi_k(\mathbf{W}(t)) = \frac{\left|\mathbf{h}^{\dagger}_{k,k}(t)\mathbf{w}_k(t)\right|^2}{\sum\limits_{j\neq k}\left|\mathbf{h}^{\dagger}_{j,k}(t)\mathbf{w}_j(t)\right|^2 + \sigma^2}, \tag{2a}$$

$$C_k(\mathbf{W}(t)) = \log\left(1 + \psi_k(\mathbf{W}(t))\right), \tag{2b}$$

where $\mathbf{W}(t) = \left[\mathbf{w}_1(t)\ \mathbf{w}_2(t) \cdots \mathbf{w}_K(t)\right] \in \mathbb{C}^{N\times K}$.

B. Channel Model

We assume a flat-and-block fading channel model to express the dynamics of downlink MISO channels. To be specific, the downlink channel vector between BS j and UE k in time slot t is modeled as follows ($\forall j, k \in \mathcal{K}$):

$$\mathbf{h}_{j,k}(t) = \sqrt{\frac{\beta_{j,k}}{L}}\mathbf{A}\left(N, \theta_{j,k}, \Delta\right)\mathbf{g}_{j,k}(t) \in \mathbb{C}^{N\times 1}, \tag{3}$$

where $\beta_{j,k}$ denotes the large-scale fading coefficient consisting of the path loss and shadowing from BS j to UE k, which remains the same over many time slots; L denotes the total number of multi-paths between BS j and UE k; and $\mathbf{g}_{j,k}(t) = [g_{j,k}(t,1), g_{j,k}(t,2), \cdots, g_{j,k}(t,L)]^T \in \mathbb{C}^{L\times 1}$ denotes the Rayleigh fading vector composed of the small scale fading coefficients of all the multi-paths. Here, $\mathbf{A}(N, \theta_{j,k}, \Delta) \in \mathbb{C}^{N\times L}$ denotes the array response matrix composed of the steering vectors of all multi-paths, given by

$$\mathbf{A}(N, \theta_{j,k}, \Delta) = \left[\mathbf{a}_1(N, \phi_1) \cdots \mathbf{a}_L(N, \phi_L)\right] \in \mathbb{C}^{N\times L}, \tag{4}$$

where $\theta_{j,k}$ is called the nominal direction of departure (DoD) of the downlink channel between BS j and UE k; Δ is a small angular range, referred to as the angular spread [27]; and $\mathbf{a}_l(N, \phi_l) \in \mathbb{C}^{N\times 1}$, $\forall l \in \mathcal{L} = \{1, \ldots, L\}$, denotes the array response vector of the lth multi-path. In (4), $\mathbf{a}_l(N, \phi_l)$ is given by

$$\mathbf{a}_l(N, \phi_l) = \left[1, e^{j2\pi\frac{d}{\lambda}\cos\phi_2}, \cdots, e^{j2\pi\frac{d}{\lambda}(N-1)\cos\phi_l}\right]^T, \tag{5}$$

where λ denotes the wavelength of the downlink carrier wave; d denotes the antenna spacing, which is usually set to $\lambda/2$; and ϕ_l is the DoD of the lth multi-path. Here, we consider that the DoDs of all the multi-paths are uniformly distributed in Δ as

$$\phi_l \sim \mathcal{U}\left(\theta_{j,k} - \frac{\Delta}{2}, \theta_{j,k} + \frac{\Delta}{2}\right). \tag{6}$$

The Rayleigh fading vector satisfies the complex Gaussian distribution $\mathcal{CN}(0, \mathbf{I}_L)$, and it remains constant in each time

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

slot and varies between successive time slots according to the first-order Gaussian–Markov process [28] given by

$$\mathbf{g}_{j,k}(t+1) = \rho \mathbf{g}_{j,k}(t) + \sqrt{1-\rho^2}\mathbf{e}_{j,k}(t), \tag{7}$$

where $\mathbf{e}_{j,k}(t) \sim \mathcal{CN}(0, \mathbf{I}_L)$ is the white Gaussian driving noise, which is independent of the channel, and ρ denotes the correlation coefficient of the Rayleigh fading vector between adjacent time slots. Here, $\mathbf{g}_{j,k}(0) \sim \mathcal{CN}(0, \mathbf{I}_L)$.

III. Distributed Approach for Multi-Cell DDBC Problem

In this section, the DDBC problem is formulated with the multi-cell MISO-IC model introduced in Section II. We then investigate the conventional approach to solving the DDBC problem, i.e., an FP algorithm, and propose a limited-information exchange protocol to obtain a practical solution to the DDBC problem under a dynamic wireless channel environment.

A. DDBC Problem Formulation

We consider a sum-rate (i.e., system capacity) maximization problem in the cellular network. To be specific, by considering the variation of the channel, the DDBC problem in time slot t can be formulated as follows:

$$\max_{\mathbf{W}(t)} \sum_{k=1}^{K} C_k(\mathbf{W}(t)) \tag{8a}$$

$$\text{s.t. } 0 \leq \|\mathbf{w}_k(t)\|^2 \leq p_{\max}, \ \forall k \in \mathcal{K}, \tag{8b}$$

where $p_{\max}$ denotes the available maximum transmit power budget of each BS. Because problem (8a) is obviously non-convex and NP-hard, it is challenging to find the optimal solution by directly solving (8a).

B. FP Algorithm to Solve a DDBC Problem

In [11], [29], Shen and Yu proposed an iterative near-optimal approach based on FP to solve the IC problems. Here, a Lagrangian dual reformulation approach is utilized to take the SINR with fractional terms such as (2a) out of the logarithms and then optimize the reformulated objective function, which only contains fractions, in an iterative manner. We deduce a particular approach based on the FP algorithm for the MISO-IC scenario considered in this study. To avoid clutter, by omitting the time slot index t in (8a) and denoting $\psi_k(\mathbf{W}(t))$ in (2a) by ψ_k, we represent the DDBC problem in an arbitrary time slot and rewrite (8a) according to a multidimensional extension of [11, Proposition 2] as follows:

$$\max_{\mathbf{W},\boldsymbol{\psi}} f \triangleq \sum_{k=1}^{K} \left(\log(1+\psi_k) - \psi_k + \frac{(1+\psi_k)\left|\mathbf{h}_{k,k}^{\dagger}\mathbf{w}_k\right|^2}{\sum\limits_{j\in\mathcal{K}} \left|\mathbf{h}_{j,k}^{\dagger}\mathbf{w}_j\right|^2 + \sigma^2} \right), \tag{9}$$

where $\boldsymbol{\psi} = [\psi_1, \psi_2, \cdots, \psi_K]^T$ is an auxiliary-variable vector. Note that the denominator of the fraction in (9) contains the noise, interferences, and also desired signal, which is different from the achievable rate defined in (2a). Here, (9) is an equivalent transformation of (8a) according to *Lagrangian Dual Transform* [29, Theorem 3].

For fixed $\mathbf{W}$, the optimal ψ_k can be then obtained from the first-order optimality, i.e., $\frac{\partial f}{\partial \psi_k} = 0$, as

$$\psi_k^* = \frac{\left|\mathbf{h}_{k,k}^{\dagger}\mathbf{w}_k\right|^2}{\sum_{j\neq k}\left|\mathbf{h}_{j,k}^{\dagger}\mathbf{w}_j\right|^2 + \sigma^2}, \quad \forall k \in \mathcal{K}. \tag{10}$$

Hence, (9) is equivalent to (8a), and it is then reformulated by using a quadratic transform [11, Theorem 1] to

$$\max_{\mathbf{W},\boldsymbol{\psi},\mathbf{y}} \sum_{k=1}^{K} \left(2\sqrt{1+\psi_k}\Re\{y_k \mathbf{w}_k^{\dagger}\mathbf{h}_{k,k}\} \frac{(1+\psi_k)|y_k|^2}{\sum\limits_{j\in\mathcal{K}}\left|\mathbf{h}_{j,k}^{\dagger}\mathbf{w}_j\right|^2 + \sigma^2} \right), \tag{11}$$

where $\mathbf{y} = [y_1, y_2, \cdots, y_K]^T$ is a complex-valued auxiliary-variable vector.

With $\boldsymbol{\psi}$ and $\mathbf{W}$ fixed, the optimal $\mathbf{y}$ can be obtained as

$$y_k^* = \frac{\sqrt{1+\psi_k}\mathbf{h}_{k,k}^{\dagger}\mathbf{w}_k}{\sum_{j\in\mathcal{K}}\left|\mathbf{h}_{j,k}^{\dagger}\mathbf{w}_j\right|^2 + \sigma^2}. \tag{12}$$

Likewise, with $\mathbf{y}$ and $\boldsymbol{\psi}$ fixed, the optimal beamformer is given by

$$\mathbf{w}_k^* = \left(\sum_{j=1}^{K} \mathbf{h}_{k,j} y_j y_j^{\dagger} \mathbf{h}_{k,j}^{\dagger} + \eta_k \mathbf{I}_N \right)^{-1} \sqrt{1+\psi_k}\mathbf{h}_{k,k} y_k, \tag{13}$$

where η_k is a dual variable introduced for the power constraint, optimally determined as (due to complementary slackness)

$$\eta_k^* = \min\left\{ \eta_k \geq 0 : \|\mathbf{w}_k(\eta_k)\|^2 \leq p_{\max} \right\}. \tag{14}$$

Note that the optimal η_k in (14) can be determined efficiently by a bisection search method.

The FP algorithm deduced specifically for MISO-IC is able to solve problem (11) by iteratively obtaining the optimal $\mathbf{y}$, $\boldsymbol{\psi}$, and $\mathbf{W}$. The pseudocode is shown in Algorithm 1, where we follow the update order in [30, Algorithm 1].

Algorithm 1 Pseudocode of an FP Algorithm for a DDBC Problem

```
Initialize $\mathbf{W}, \boldsymbol{\psi}$ to feasible values.
repeat
    Update $\mathbf{y}$ by (12) for fixed $(\mathbf{W}, \boldsymbol{\psi})$.
    Update $\mathbf{W}$ by (13) for fixed $(\mathbf{y}, \boldsymbol{\psi})$.
    Update $\boldsymbol{\psi}$ by (10) for fixed $(\mathbf{W}, \mathbf{y})$.
until numerical convergence of the objective
return $\mathbf{W}$
```

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

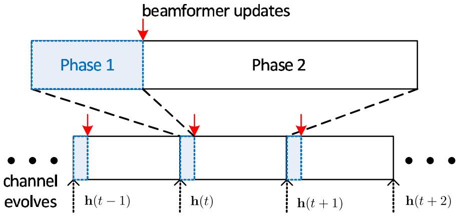


Fig. 2. Designed downlink data transmission framework.

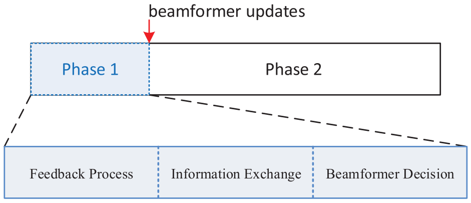


Fig. 3. Detailed timing information of phase 1 in the proposed DRL-based approach.

C. Proposed Limited-Information Exchange Protocol

First of all, we design a downlink data transmission framework as shown in Fig. 2 for a DDBC problem. In the framework, each time slot is divided into two phases. The first phase (phase 1) is a preparing phase for the subsequent data transmission while the second phase (phase 2) is for the downlink data transmission. All the BSs update their own beamformers at the end of phase 1. To be specific, in the centralized approaches, the cascade procedure of collecting global CSI, computing beamformers, and sending beamformers to the corresponding BSs is supposed to be carried out within phase 1.

In practical networks, the downlink channels, including the direct links and interference channels, are usually measured by BS's serving UE. Then BS can acquire the measurements through the feedback from its serving UE. In centralized approaches, it is often assumed that the global instantaneous CSI is collected through a backhaul network which connects the BSs and the central controller, resulting in large signalling overhead. The central controller sends the applicable beamformers to the BSs only when the iterative algorithm converges, causing additional non-negligible delay due to the high computation complexity of the centralized approaches. In short, the global CSI collection and the high computational complexity of the centralized optimization-based approaches lead to high system overhead to make the centralized approaches practically infeasible in a dynamic environment.

In modern cellular networks, BSs are usually allowed to exchange information with each other through predefined interfaces. For example, in LTE networks, eNodeBs (evolved-NodeBs, i.e., the base stations in LTE networks) can exchange information with each other through $X2$ interface. The signaling overhead of information exchange among BSs is considerably lower than that of the backhaul network in centralized approaches. To this end, we develop a distributed approach with a designed limited-information exchange protocol, in which a BS can exchange information with a limited number of neighboring BSs.

Inspired by the idea in [26], [31], we first define two limited sets for each direct link: the *interferers* and the *interfered neighbors*. Taking the kth direct link comprising BS k and UE k as an example, its interferers refer to the neighbor links whose BS interferes with the kth UE, while its interfered neighbors refer to the neighbor links whose UE is interfered by BS k, e.g., BSs 1–6 as the interferers for BS0-UE0 direct link and UEs 1–6 as the interfered neighbors of BS0, as illustrated in Fig. 1. The neighbor links in the sets are sorted by the received signal strength (RSS) at the receiver that corresponds with cross-links, and two dynamic thresholds are determined to limit the cardinality of the two sets. Hence, for the kth direct link, its interferers and interfered neighbors in time slot t can be defined, respectively, as follows:

$$\mathcal{I}_k(t) = \{j \in \mathcal{K},\ j \neq k,\ \left|\mathbf{h}_{j,k}^{\dagger}(t-1)\mathbf{w}_j(t-1)\right|^2 \geq \xi_{\mathcal{I}_k(t)}\},$$
$$\mathcal{O}_k(t) = \{i \in \mathcal{K},\ i \neq k,\ \left|\mathbf{h}_{k,i}^{\dagger}(t-1)\mathbf{w}_k(t-1)\right|^2 \geq \xi_{\mathcal{O}_k(t)}\}, \tag{15}$$

where $\xi_{\mathcal{I}_k(t)}$ and $\xi_{\mathcal{O}_k(t)}$ are the dynamic thresholds for satisfying that $|\mathcal{I}_k(t)| = U$ and $|\mathcal{O}_k(t)| = U$, respectively.

As shown in Fig. 3, phase 1 of each time slot is split to three sub-phases in our proposed DRL-based approach, the first two sub-phases are for the designed limited-information exchange process while the third sub-phase is for the beamformer-decision process. The feedback process between the UE and its associated BS is carried out within the first sub-phase. With the feedback of UE k, BS k is able to obtain the RSS at UE k at time slot t, i.e., $|\mathbf{h}_{k,k}^{\dagger}(t)\mathbf{w}_k(t-1)|^2$, and the total received interference-plus-noise at UE k, i.e., $\sum_{j\neq k}|\mathbf{h}_{j,k}^{\dagger}(t)\mathbf{w}_j(t-1)|^2 + \sigma^2$, before the BSs update their beamformers. In the first sub-phase, BS k can also estimate the RSS from each interferer $j \in \mathcal{I}_k(t)$, i.e., $|\mathbf{h}_{j,k}^{\dagger}(t)\mathbf{w}_j(t-1)|^2$. The second sub-phase is for the information exchange process among BSs. Defining a normalized beamformer as $\bar{\mathbf{w}}_k(t) \triangleq \mathbf{w}_k(t)/\|\mathbf{w}_k(t)\|$ for simplicity, BS k sends its own measurements comprising the following information to its interferer $j \in \mathcal{I}_k(t)$ in return for information about $\mathbf{w}_j(t-1)$ and $C_j(\mathbf{W}(t-1))$:

- The achievable rate of direct link k in time slot $t-1$, i.e., $C_k(\mathbf{W}(t-1))$.
- The equivalent channel gain of direct link k in time slot $t-1$, i.e., $\left|\mathbf{h}_{k,k}^{\dagger}(t-1)\bar{\mathbf{w}}_k(t-1)\right|^2$.
- The received interference power from interferer $j \in \mathcal{I}_k(t)$ in time slot $t-1$, i.e., $\left|\mathbf{h}_{j,k}^{\dagger}(t-1)\mathbf{w}_j(t-1)\right|^2$.
- The total interference-plus-noise power of UE k in time slot $t-1$, i.e., $\sum_{l\neq k}\left|\mathbf{h}_{l,k}^{\dagger}(t-1)\mathbf{w}_l(t-1)\right|^2 + \sigma^2$.

The details of the designed limited-information exchange process in time slot t are depicted in Fig. 4. Note that while BS k exchanges its measurements with its interferers, its interfered neighbors also exchange their measurements with BS k because BS k plays the role of an interferer to its interfered neighbors. Moreover, $\bar{\mathbf{w}}_k(t)$ and $\mathbf{w}_k(t)$ can be expressed in simpler forms to further reduce the overhead of

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

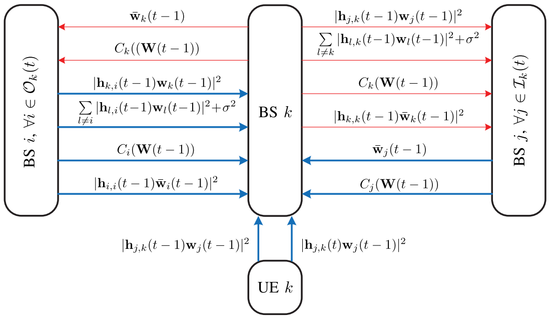

$\bar{\mathbf{w}}_k(t-1)$
$C_k((\mathbf{W}(t-1))$
$|\mathbf{h}_{k,i}(t-1)\mathbf{w}_k(t-1)|^2$
$\sum_{l\neq i} |\mathbf{h}_{l,i}(t-1)\mathbf{w}_l(t-1)|^2+\sigma^2$
$C_i(\mathbf{W}(t-1))$
$|\mathbf{h}_{i,i}(t-1)\bar{\mathbf{w}}_i(t-1)|^2$
BS i, $\forall i \in \mathcal{O}_k(t)$
BS k
$|\mathbf{h}_{j,k}(t-1)\mathbf{w}_j(t-1)|^2$
$\sum_{l\neq k} |\mathbf{h}_{l,k}(t-1)\mathbf{w}_l(t-1)|^2+\sigma^2$
$C_k(\mathbf{W}(t-1))$
$|\mathbf{h}_{k,k}(t-1)\bar{\mathbf{w}}_k(t-1)|^2$
$\bar{\mathbf{w}}_j(t-1)$
$C_j(\mathbf{W}(t-1))$
BS j, $\forall j \in \mathcal{I}_k(t)$
$|\mathbf{h}_{j,k}(t-1)\mathbf{w}_j(t-1)|^2$
$|\mathbf{h}_{j,k}(t)\mathbf{w}_j(t-1)|^2$
UE k

Fig. 4. Designed limited-information exchange protocol in time slot t. BSs are able to share their historical measurements and other information with their interferers and interfered neighbors.

information exchange in the design of our proposed approach which is described later.

Our proposed DRL-based scheme significantly reduces the system overhead. Each BS needs to acquire at most $6U$ messages from the neighboring BSs during the information exchange process, while the centralized approaches need to collect K^2 N-dimensional complex-valued vectors. In addition, as shown in Fig. 4, the information exchanged among BSs in time slot t is actually the channel measurements obtained in the previous time slot $t-1$. Thus, the information exchange process among BSs can be alternatively carried out after the channels are measured in time slot $t-1$. However, in the centralized approaches, a central controller needs to collect the K^2 N-dimensional complex-valued vectors of the instantaneous channels within phase 1 at time slot t, which is much more challenging to realize in practice. Furthermore, our DRL-based scheme significantly reduces the computational complexity compared with the centralized approaches, as we verify by simulation in Section V.

IV. Proposed Distributed DRL-Based Approach for a DDBC Problem

In this section, we first present an overview of deep Q-learning. Then we propose a distributed DRL-based approach for the original problem (8a) accordingly.

A. Brief Overview of Deep Q-Learning

RL is usually identified as an efficient approach to deal with MDP problems, which are defined as $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$ in general, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{T}$, and $\mathcal{R}$ are the sets of states, actions, transition probabilities, and rewards, respectively, and γ is a discount factor. To be specific, at time step t, the agent (i.e., the decision maker) in state $s \in \mathcal{S}$ takes an action $a \in \mathcal{A}$ according to policy $\pi(a|s)$, receives a reward $r = \mathcal{R}(s, a) \in \mathbb{R}$, and then transfers to the next state $s' \in \mathcal{S}$ with probability $\mathcal{T}(s, a, s')$ in time step $t+1$. Formally, $\mathcal{T}(s, a, s')$ is the transition probability from state s to state s' after executing action a; i.e., $\Pr(s'|s, a)$. The tuple $e = \langle s, a, r, s' \rangle$ is called an experience of the agent. The discount factor, $\gamma \in [0, 1)$, indicates the degree to which the future reward is taken into consideration for the present decision.

The optimal policy $\pi^*(a|s)$ to solve the MDP problem can be regarded as a mapping from states to actions that maximize the long-term cumulative discount reward defined as

$$R_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} \mathcal{R}(s_{t+\tau+1}, a_{t+\tau+1}), \tag{16}$$

where $s_{t+\tau+1}$ and $a_{t+\tau+1}$ are the involved values at time slot $t+\tau+1$ corresponding to s and a, respectively. It can be acquired by using dynamic programming (DP) methods, e.g., a value iteration method. The DP methods, however, usually require knowledge of environment dynamics, i.e., $\mathcal{T}(s, a, s')$, which is impractical in fading channel environments due to their uncertain variation.

The well-known Q-learning algorithm is model-free, i.e., it is useful even without knowledge of environment dynamics. It can constantly improve the policy through a trial-and-error mechanism, i.e., by observing the received rewards after interacting with the environment. In Q-learning, a Q-value (action value) function is associated with a certain policy. The Q-value function associated with policy $\pi(a|s)$ is given by

$$Q_{\pi}(s, a) = \mathbb{E}\left[R_t | s_t = s, a_t = a\right]. \tag{17}$$

The Q-value function in (17) can be computed iteratively by the Bellman equation, given by

$$Q_{\pi}(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \left(\mathcal{T}\left(s, a, s'\right) \sum_{a' \in \mathcal{A}} Q_{\pi}\left(s', a'\right) \right), \tag{18}$$

where $\pi(s', a')$ denotes the probability of executing action a' while the current state is s'; and $\mathcal{R}(s, a) = r$ is the immediate reward of taking action a at state s. As shown in (18), $Q_{\pi}(s, a)$ can be calculated by exhaustively evaluating all the possible state-action pairs, i.e., $\forall (s', a') \in \mathcal{S} \times \mathcal{A}$. However, it is evidently infeasible. [32] has shown that the Q-learning algorithm can efficiently converge to the action value function (17) in an iterative manner. Here, $\pi^*(s, a) = 1$ if the action a is optimal in state s. Thus, the Q-value function associated with the optimal policy $\pi^*(a|s)$ is given by

$$Q^{\pi^*}(s, a) = \mathcal{R}(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') \max_{a'} Q^{\pi^*}(s', a'). \tag{19}$$

In Q-learning, the Q-value function $Q_{\pi}(s, a)$ is represented by a Q-value table $q(s, a)$, namely, Q-table. At the beginning of the Q-learning algorithm, an $|\mathcal{S}| \times |\mathcal{A}|$ Q-table is randomly initialized. The agent then initiates the trial-and-error process by constantly taking actions according to an ϵ-greedy policy. The ϵ-greedy policy is an efficient approach to solving the exploration-exploitation dilemma. It allows the agent to take the action a^* that has the maximum Q-value for a given state s in the table with exploration probability $1-\epsilon$ or to execute a random action $a \in \mathcal{A}$ with exploitation probability ϵ to guarantee that the algorithm converges to a near-optimal (even the optimal) solution without getting stuck at a local optimal point. After the agent completes experience e_t, the Q-value table is updated according to

$$q(s, a) \leftarrow (1-\alpha) q(s, a) + \alpha \left(r + \gamma \max_{a'} q(s', a') \right), \tag{20}$$

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

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

In regard to the MDP problems with a large state space, the classical Q-learning algorithm becomes infeasible due to the enormous complexity of exhausting the massive state-action pairs and the required large storage to save the huge Q-table. This issue is addressed by introducing the deep neural network (DNN), called the deep Q-network (DQN). The DQN is exploited to approximate the mapping from state s to action a. In general, we use $\boldsymbol{\theta}$ to denote the weights of the DNN, therefore, the function of the DQN can be expressed as $q(s,a;\boldsymbol{\theta})$. The optimal policy $\pi^*(a|s)$ corresponds to a particular set of weights $\boldsymbol{\theta}^*$ of the DQN. Further, deep Q-learning exploits two techniques to accelerate convergence while guaranteeing its stability [20]. One is *quasi-static target network*, which allows the agent to establish another DQN with weights $\boldsymbol{\theta}^-$, namely the target DQN. The target DQN is supposed to generate the target Q-value, which is utilized to construct the loss function in the training procedure. The other is *experience replay*, which allows the agent to train its DQN using the historical experiences (e.g. $\langle s,a,r,s'\rangle$). In each training step, M_b experiences are sampled from the experience pool $\mathcal{M}$, which is a first-input-first-output queue composed of M_m experiences.

The training process aims to minimize the prediction error between the trained DQN and the target DQN. In training step t, the prediction error is defined as a loss function given by

$$L(\boldsymbol{\theta}) = \frac{1}{2M_b} \sum_{\langle s,a,r,s'\rangle \in \mathcal{D}} \left(r' - q(s,a;\boldsymbol{\theta})\right)^2, \qquad (21)$$

where the target value of reward $r' = r + \gamma \max_{a'} q(s',a';\boldsymbol{\theta}^-)$.

Finally, the stochastic gradient descent optimizer is utilized to minimize the loss function in (21). After computing with the mini-batch $\mathcal{D}$, the optimizer returns a set of gradients shown in (22) to update the weights of the trained DQN through the back-propagation (BP) technique:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{M_b} \sum_{\langle s,a,r,s'\rangle \in \mathcal{D}} \left(r' - q(s,a;\boldsymbol{\theta})\right) \nabla q(s,a;\boldsymbol{\theta}). \qquad (22)$$

It is shown that the trained DQN converges quickly with a set of good parameters [33].

B. Proposed Distributed DRL-Based Approach

In our setup, each BS can determine its beamformers, i.e., each BS is an independent agent. Hence, the DDBC problem can be modeled as a multi-agent reinforcement learning problem. Several schemes are available to deal with such problems. In one scheme, all the agents are treated as a single-integrated agent. The integrated agent collects states from all agents and makes the joint decision that contains actions for all agents, i.e., a *centralized-training-centralized-executing* (CTCE) scheme [25]. The other scheme is a *centralized-training-distributed-executing* (CTDE) scheme [26], in which a multi-agent learning system can be divided into a centralized training unit and distributed executing agents. Here, all the agents upload their experiences to a central unit for training the public DQN, and the central training unit broadcasts the weights of the public DQN to all agents. Each agent then makes a decision using the local DQN with the weights of the public DQN according to the local observations of the environment. However, no settled approach for multi-agent learning problems has been found so far, in spite of various empirical approaches with good performance.

In this paper, we adopt a distributed scheme, called the *distributed-training-distributed-executing* (DTDE) scheme, which is illustrated in Fig. 5. Here, the solid lines denote the online decision-making process, while the dotted lines refer to the offline training procedure. In the online decision-making process, agent k (BS k) takes action a_k according to its current state s_k, which is obtained from the designed limited-information exchange protocol, and its policy $\pi_k(a|s)$, which is determined by the output of the train DQN and ϵ-greedy policy. In regard to the offline training stage, agent k takes out a mini-batch $\mathcal{D}_k$ that consists of M_b experiences from the experience pool, and calculates the loss function according to (21) using the experience $\langle s_k,a_k,r_k,s'_k\rangle$ from $\mathcal{D}_k$, i.e., input (r_k,s'_k) to the target DQN in return for r' and input (s_k,a_k) to the trained DQN in return for $q(s_k,a_k;\boldsymbol{\theta}_k)$. BS k can then use an optimizer, such as a stochastic gradient descent optimizer, to minimize the loss function and update the weights of the trained DQN through a BP method. After every T_{step} training step, the weights of the target DQN are updated with the new weights of the trained DQN. The pseudocode of the proposed distributed DRL-based DTDE scheme for DDBC is shown in Algorithm 2.

1) Actions of Distributed DRL: As described in Section IV-A, the available actions of the agent in a DRL algorithm are usually a set of discrete real values. The transmit beamformer $\mathbf{w}_k(t)$ in the original problem (8a), however, is a continuous complex vector, $\forall k \in \mathcal{K}$. Hence, we propose a simple yet effective method to deal with continuous complex-valued beamformers.

First, we decompose the beamformer into two parts as

$$\mathbf{w}_k(t) = \sqrt{p_k(t)}\bar{\mathbf{w}}_k(t), \qquad (23)$$

where $p_k(t) = \|\mathbf{w}_k(t)\|^2$ denotes the transmit power of BS k in time slot t, such that $0 \le p_k(t) \le p_{\max}$, and the normalized beamformer (also called a code in the following) $\bar{\mathbf{w}}_k(t)$ indicates the direction of the transmit beam. Therefore, we can use a combination of $p_k(t)$ and $\bar{\mathbf{w}}_k(t)$ to represent the beamformer of BS k in time slot t. Note that the directions are usually represented by the degrees of angles, and the directions are represented in $[0,2\pi)$, i.e., the periodic property for angles is ignored.

We then take Q_{pow} values between 0 and $p_{\max}$ uniformly for the available transmit power levels of each BS, and define set $\mathcal{P}$ with the selected power levels. We also consider codebook $\mathcal{C}$ composed of Q_{code} code vectors, $\mathbf{c}_q \in \mathbb{C}^{N\times 1}$, covering an arbitrary direction in $[0,2\pi)$ for $\bar{\mathbf{w}}_k(t)$, where $q \in \{0,1,\ldots,Q_{\text{code}}-1\}$. At the third sub-phase in phase 1 of each time slot, BS k is able to determine its beamformer by simply picking power level $p_k(t)$ and code $\mathbf{c}_k(t)$ from $\mathcal{P}$ and $\mathcal{C}$, respectively. Therefore, the actions available at each BS

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

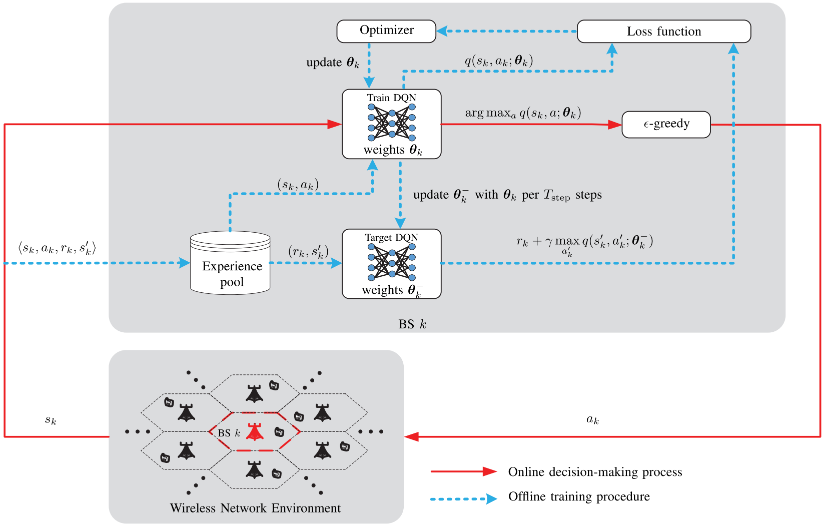

Optimizer
Loss function
update $\boldsymbol{\theta}_k$
$q(s_k, a_k; \boldsymbol{\theta}_k)$
Train DQN
weights $\boldsymbol{\theta}_k$
$\arg\max_a q(s_k, a; \boldsymbol{\theta}_k)$
ϵ-greedy
(s_k, a_k)
update $\boldsymbol{\theta}_k^-$ with $\boldsymbol{\theta}_k$ per T_{step} steps
$\langle s_k, a_k, r_k, s_k' \rangle$
Experience pool
(r_k, s_k')
Target DQN
weights $\boldsymbol{\theta}_k^-$
$r_k + \gamma \max_{a_k'} q(s_k', a_k'; \boldsymbol{\theta}_k^-)$
BS k
s_k
a_k
BS k
Wireless Network Environment
Online decision-making process
Offline training procedure

Fig. 5. Illustration of the proposed distributed DRL-based DTDE scheme in the considered multi-agent system.

Algorithm 2 Pseudocode of the Proposed Distributed DRL-Based DTDE Scheme for DDBC

```
Establish a pair of DQNs (a trained DQN with weights $\boldsymbol{\theta}_k$ and a target DQN with weights $\boldsymbol{\theta}_k^-$) and an empty experience pool $\mathcal{M}_k$ for BS $k$, $\forall k \in \mathcal{K}$.
Initialize the trained DQN with random weights; set $\boldsymbol{\theta}_k^- = \boldsymbol{\theta}_k$, $\forall k \in \mathcal{K}$.
In time slot $t$ ($t \leq M_b$), each agent takes action randomly and stores the corresponding experience $\langle s, a, r, s' \rangle$ in its experience pool.
repeat
    Agent $k$ observes its state $s_k$ in time slot $t$, $\forall k \in \mathcal{K}$.
    In time slot $t$ ($t > M_b$), agent $k$ chooses an action $a_k$ according to $\epsilon$-greedy policy: agent $k$ chooses an action $a_k = \arg\max_{a \in \mathcal{A}} q(s_k, a; \boldsymbol{\theta}_k)$ with probability $(1 - \epsilon)$, or randomly chooses an action $a_k \in \mathcal{A}$ with probability $\epsilon$, $\forall k \in \mathcal{K}$.
    Agent $k$ executes the chosen action $a_k$, then gets an immediate reward $r_k = \mathcal{R}(s_k, a_k)$, $\forall k \in \mathcal{K}$.
    Agent $k$ observes a new state $s_k'$ in time slot $t+1$, $\forall k \in \mathcal{K}$.
    Agent $k$ saves its new experience $\langle s_k, a_k, r_k, s_k' \rangle$ into its own experience pool $\mathcal{M}_k$, $\forall k \in \mathcal{K}$.
    Agent $k$ samples a mini-batch consisting of $M_b$ experiences from its experience pool $\mathcal{M}_k$, $\forall k \in \mathcal{K}$.
    Agent $k$ updates the weights $\boldsymbol{\theta}_k$ of its trained DQN using BP technique with (21) and (22), $\forall k \in \mathcal{K}$.
    Agent $k$ updates $\boldsymbol{\theta}_k^-$ with $\boldsymbol{\theta}_k$ every $T_{\text{step}}$ time slots, $\forall k \in \mathcal{K}$.
until convergence.
```

are simply (omitting indices k and t) represented as follows:

$$\mathcal{A} = \{(p, \mathbf{c}),\ p \in \mathcal{P},\ \mathbf{c} \in \mathcal{C}\}, \tag{24}$$

where $\mathcal{P} = \left\{0, \frac{1}{Q_{\text{pow}}-1} p_{\max},\ \frac{2}{Q_{\text{pow}}-1} p_{\max},\ \cdots,\ p_{\max}\right\}$ and $\mathcal{C} = \{\mathbf{c}_0,\ \mathbf{c}_1,\ \cdots,\ \mathbf{c}_{Q_{\text{code}}-1}\}$.

Denoting a codebook matrix by $\mathbf{C} = [\mathbf{c}_0\ \cdots\ \mathbf{c}_{Q_{\text{code}}-1}] \in \mathbb{C}^{N \times Q_{\text{code}}}$, each column of $\mathbf{C}$ is a code that specifies a beam direction. Here, we employ a codebook matrix in [34], where a sufficient number of codes can be designed such that $Q_{\text{code}} \geq N$. Noting that the (n, q)th element of $\mathbf{C}$, denoted by $\mathbf{C}[n, q]$, refers to the phase shift of the nth antenna element in the qth code, the codebook matrix is designed as follows:

$$\mathbf{C}[n, q] = \frac{1}{\sqrt{N}} \exp\left(j \frac{2\pi}{S} \left\lfloor \frac{n \mathrm{mod}\left(q + \frac{Q_{\text{code}}}{2}, Q_{\text{code}}\right)}{Q_{\text{code}}/S} \right\rfloor\right), \tag{25}$$

where S denotes the number of available phase values for each antenna element, and $\lfloor \cdot \rfloor$ and $\mathrm{mod}(\cdot)$ represent the floor and modulo operations, respectively. In this paper, we set $S = 16$ as the default value for accuracy. The beam patterns when $N = 3$ and $S = 16$ are shown in Figs. 6(a), (b), (c), and (d), for $Q_{\text{code}} = 4$, 6, 8, and 10, respectively. As can be seen, the BS can determine its beamformer by selecting a discrete transmit power level and a code associated with a specified beam direction independently. Hence, the total number of available actions is $Q = Q_{\text{pow}} Q_{\text{code}}$, and the total number of output ports in the DQN is the same as Q.

2) States of Distributed DRL: As described in Section IV-A, the states of an agent are supposed to include the representative features extracted from observations of the environment, which are obtained through the designed limited-information exchange protocol. Using an intuitive approach, we organize the state of BS k in time slot t, i.e., $s_k(t)$, from three pieces of information defined as follows:

- *Local Information*: Local information consists of seven elements. The first two elements are the action taken by BS k in time slot $t-1$, namely, the transmit power $p_k(t-1)$ and the index of the selected code (i.e., a normalized beamformer). The code index is denoted by $I_k(t-1) = q$ when $\mathbf{c}_q$ is selected. Instead of $\mathbf{c}_q$, its index is used during the information exchange process to reduce the network overhead. The third element is the achievable rate of direct link k in time slot $t-1$, i.e., $C_k(\mathbf{W}(t-1))$. The fourth and fifth elements are

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

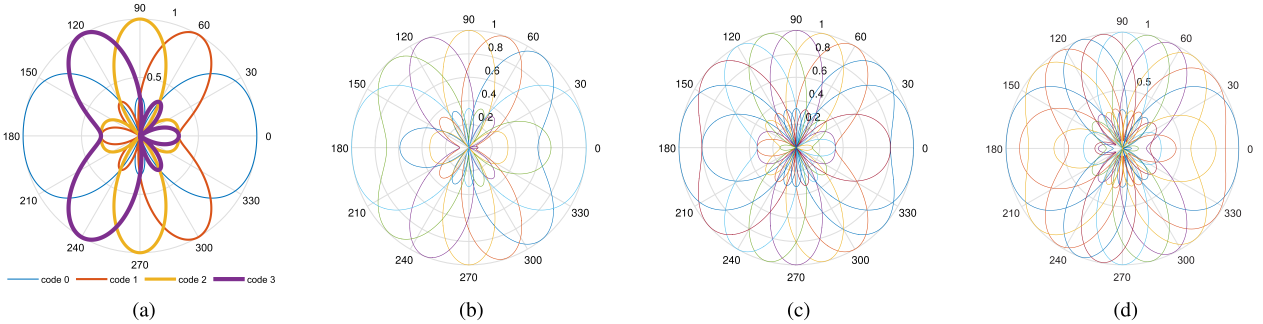


Fig. 6. Set of beam directions when $N = 3$ and $S = 16$: (a) $Q_{\text{code}} = 4$ (b) $Q_{\text{code}} = 6$ (c) $Q_{\text{code}} = 8$ (d) $Q_{\text{code}} = 10$.

the equivalent channel gains, which are obtained by dividing the received signal power at UE k by the transmit power at BS k, namely, $\left|\mathbf{h}^{\dagger}_{k,k}(t-1)\bar{\mathbf{w}}_k(t-2)\right|^2$ and $\left|\mathbf{h}^{\dagger}_{k,k}(t)\bar{\mathbf{w}}_k(t-1)\right|^2$. The sixth and seventh elements are the total interference-plus-noise power at UE k: $\sum_{j\neq k} p_j(t-1)\left|\mathbf{h}^{\dagger}_{j,k}(t)\bar{\mathbf{w}}_j(t-1)\right|^2 + \sigma^2$ and $\sum_{j\neq k} p_j(t-2)\left|\mathbf{h}^{\dagger}_{j,k}(t-1)\bar{\mathbf{w}}_j(t-2)\right|^2 + \sigma^2$. Therefore, seven input ports of the DQN are reserved for the local information.

- *Interferers' Information*: When direct link k is regarded as an *interfered* link, agent k is able to learn about the changes in its environment by observing the exchanged information from its interferers in the current and past slots, i.e., the interference information of interferer j, where $j \in \mathcal{I}_k(t-\tau), \tau = 0, 1, \ldots, t$. Considering the channels conforming to the first-order Gaussian Markov process, the current CSI at t is determined statistically by only the CSI in time slot $t-1$. Hence, we take only the interference information of interferers in time slots t and $t-1$ into consideration, namely, $\mathcal{I}_k(t)$ and $\mathcal{I}_k(t-1)$. For each interferer $j \in \mathcal{I}_k(t)$, four input ports are reserved for information on i) which interferer causes, ii) how much interference power to the considered direct link k using, iii) which normalized beamformer, and iv) the interferer's contribution to the objective function (8a), namely, i) j, ii) $p_j(t-1)\left|\mathbf{h}^{\dagger}_{j,k}(t-1)\bar{\mathbf{w}}_j(t-1)\right|^2$, iii) $I_j(t-1)$, and iv) $C_j(\mathbf{W}(t-1))$, respectively. Likewise, four input ports are also reserved for each interferer $j' \in \mathcal{I}_k(t-1)$, i.e., j', $p_{j'}(t-2)\left|\mathbf{h}^{\dagger}_{j',k}(t-2)\bar{\mathbf{w}}_{j'}(t-2)\right|^2$, $I_{j'}(t-2)$, and $C_{j'}(\mathbf{W}(t-2))$. Because there are U interferers, i.e., $|\mathcal{I}_k(t)| = U$ as defined in Section III-C, in all, the $8U$-input ports are reserved for interferers' information.
- *Interfered Neighbors' Information*: When direct link k is regarded as an *interferer* link, agent k is supposed to measure its effect on its interfered neighbors after taking an action. However, when agent k chooses a beamformer $\mathbf{w}_k(t-1)$ such that $\|\mathbf{w}_k(t-1)\|^2 = 0$, i.e., when BS k is inactive in time slot $t-1$, the set of its interfered neighbors is useless because no interference occurs from BS k. To deal with this problem, we use t'_k to denote the last time slot in which BS k is active. Note that if $t'_k < t-1$, the interfered neighbor $i \in \mathcal{O}_k(t'_k)$ has no knowledge of $\left|\mathbf{h}^{\dagger}_{k,i}(t-1)\bar{\mathbf{w}}_k(t-1)\right|^2$: yet agent i is still able to send its local information to agent k. Hence, three input ports are reserved for each interfered neighbor $i \in \mathcal{O}_k(t'_k)$, which are $\left|\mathbf{h}^{\dagger}_{i,i}(t-1)\bar{\mathbf{w}}_i(t-1)\right|^2$, $C_i(\mathbf{W}(t-1))$, and $\frac{p_k(t'_k)|\mathbf{h}^{\dagger}_{k,i}(t'_k)\bar{\mathbf{w}}_j(t'_k)|^2}{\sum_{j\neq i} p_j(t-1)|\mathbf{h}_{j,i}(t-1)\bar{\mathbf{w}}_j(t-1)|^2+\sigma^2}$. Therefore, in all, the $3U$-input ports are reserved for the interfered neighbors as $|\mathcal{O}_k(t)| = U$.

3) Reward Function of Distributed DRL: In the considered MISO-IC scenario, one agent can maximize its achievable rate by choosing the best action (e.g., the combination of most advantageous direction and maximum transmit power) whereas other agents do not update their beamformers. At the same time, that agent also introduces high interference to its interfered neighbors while it tries to maximize its own achievable rate, and then causes a decrease of the objective function in (8a). Thus, the maximization of the achievable rate of one agent does not maximize the objective function in (8a). A well-designed reward function should help the agent maximize the objective function in (8a), and minimize the negative effect of its action on the other agents. To this end, we design the reward function as follows:

$$r_k(t) = C_k(\mathbf{W}(t)) - P_k(\mathbf{W}(t)), \tag{26}$$

where the first term is the achievable rate of agent k, and the second term is a penalty for BS k. In other words, the reward function can be considered as a net gain of the action's impact on the objective function in (8a). Similar reward functions are found for DRL-based schemes [25], [26] and non-cooperative game theoretical algorithms [35]. To be specific, the penalty on BS k is defined as the sum of the achievable rate losses of the interfered neighbors $j \in \mathcal{O}_k(t+1)$, which are interfered by BS k, as follows:

$$P_k(\mathbf{W}(t)) = \sum_{j\in\mathcal{O}_k(t+1)} \log\left(1+\frac{p_j(t)|\mathbf{h}^{\dagger}_{j,j}\bar{\mathbf{w}}_j(t)|^2}{\sum_{i\neq k,j} p_i(t)|\mathbf{h}^{\dagger}_{i,j}\bar{\mathbf{w}}_i(t)|^2+\sigma^2}\right) - C_j(\mathbf{W}(t)). \tag{27}$$

The reward $r_k(t)$ in (26) is calculated in time slot $t+1$, i.e., after agent k executes an action $a_k(t)$ in time slot t. On the other hand, $\sum_{i\neq k,j} p_i(t)\left|\mathbf{h}^{\dagger}_{i,j}\bar{\mathbf{w}}_i(t)\right|^2$ in the penalty function

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

can be calculated by subtracting the interference of BS k from the total received interference-plus-noise power at UE j.

C. Discussions About the Non-Stationarity of the Proposed Distributed DRL-Based DTDE Scheme

Actually, both conventional RL techniques (e.g., Q-learning) and the emerging DRL techniques (e.g., deep Q-learning) are designed for the single-agent cases with stationary environments. However, the environment dynamics are not stationary anymore for the multi-agent reinforcement learning (MARL) problems due to the coupled actions of the other learning agents [36].

Up to now, how to perfectly deal with the non-stationary issue perfectly is still an open problem despite various approaches that are proposed to tackle it. One of the approaches that tackle the non-stationary issue is allowing agents to communicate with one another [37]. Thus, agents can share their observations, behaviors, and intensions to guarantee the stability of the training process. In other words, agents can learn the impact of their coupled actions on the dynamic environment. In addition, an agent can also acquire the knowledge of other agents' policy and predict their behaviors through the information exchanged from other agents [38].

The design of a reward function is also critical to solve the non-stationary issue in MARL. In the research field of game theory, a payoff function is usually designed with the net gain that is the difference between the original profit and the incurred cost [35]. In [35], the authors considered a sum-rate maximization problem in an ad hoc network and proposed to set the cost as a payment to the other interfered users. The similar design of a reward function was also considered for MARL in [25] and [26], providing good performance.

In our proposed DRL-based DTDE approach, the non-stationary issue is addressed with information exchange among BSs and the design of the reward function. In the designed limited-information exchange protocol, BSs are supposed to share the information of their historical actions, channel measurements, and achievable rates. In addition, we design a reward function which is similar to that in [25] and [26]. From the view point of game theory, the designed reward function can be interpreted as a net gain of the positive effects (profit) and negative effects (cost) on the objective function.

V. Simulation Results

In this section, we evaluate the average achievable rate, i.e., the system capacity divided by the number of UEs as (8a)/K, of the proposed distributed DRL-based DTDE scheme for DDBC through computer simulation. For comparison (as a benchmark), we evaluate an FP algorithm with perfect global CSI, i.e., the near-optimal approach described in Section III-B. In addition, we conduct a comparison with the performance of a greedy algorithm to validate the effectiveness of the proposed DRL-based approach in learning the trade-off between maximizing the achievable rate of an agent and reducing the interference it causes to other agents. For the greedy approach, it is also assumed that each BS knows the accurate downlink channel vector and that it maximizes only the throughput of its intended UE. In other words, each greedy BS transmits messages to its intended UE with its optimal code $\mathbf{c} = \max_{\mathbf{c}_q \in \mathcal{C}} |\mathbf{h}_{k,k}^{\dagger}\mathbf{c}_q|^2$ and the maximum power budget $p_{\max}$.

A. Simulation Setup

To begin with, we consider a homogeneous cellular network with 19 hexagonal cells, in which BS 0 is located at the center, BSs 1–6 are located in the first tier, and BSs 7–18 are located in the second tier. Each BS with three antennas is located at the center of the corresponding cell. The cell radius (i.e., half of the BS-to-BS distance) is set to 200 m. In addition, we define a small region of radius 10 meters where no UE exists. Each intended UE is randomly located in each cell area between the inner region boundary and the cell boundary. The maximum transmit power budget at each BS, i.e., $p_{\max}$, is set to 38 dBm. The path loss between BS k and UE j depends on the distance $d_{j,k}$ in kilometers between them as $\beta_{j,k} = 120.9 + 37.6 \log_{10} d_{j,k} \text{ dB}$. Furthermore, we set the log-normal shadowing standard deviation to be 8 dB and the additive white Gaussian noise power as $\sigma^2 = -114 \text{ dBm}$. The total number of multi-paths L is 4 and the angular spread Δ is 3°. The nominal DoD of the channels is set as the azimuth of the UE to the BS, because the locations of the UEs are initialized randomly. The time slot interval is set to 20 ms, while the correlation coefficient ρ between successive time slots is set to 0.64.

The hyperparameters adopted in the DQNs of the BSs are as follows.[1] The *universal approximation theorem* [39] states a neural network with a single hidden layer is sufficient to represent any function. However, the layer are supposed to be large enough that may fail to learn and a neural network with two (or more) hidden layers shows to be more efficient [40]. Based on some preliminary simulations, in our setup, each agent trains a DQN with an input layer, an output layer and two fully-connected hidden layers. The total number of input ports is identical to the number of state elements of an agent as illustrated in Section IV-B.2. The cardinality constraint on the number of neighbors is set as $U = 5$. The total number of input ports is 62 (i.e., $7 + 8U + 3U$). The numbers of neurons of the two hidden layers are 64 and 32, respectively. In the preliminary simulation, we set the number of available power levels and the codebook size as $Q_{\text{pow}} = 5$ and $Q_{\text{code}} = 4$, respectively, and therefore, the total number of output ports is $Q = Q_{\text{pow}} Q_{\text{code}} = 20$. In particular, the rectifier linear unit (ReLU) activation function is employed in each hidden layer. We also adopt an adaptive ϵ-greedy algorithm, in which the exploration probability ϵ at time t is reduced gradually as $\epsilon(t) = \max\{\epsilon_{min},\ (1 - \lambda_\epsilon)\epsilon(t-1)\}$, where the initial exploration probability $\epsilon(0) = 0.6$, minimum exploration probability $\epsilon_{min} = 0.01$, and $\lambda_\epsilon = 1e^{-4}$. The size of mini-batch M_b is set to 32 and the size of the experience pool at each agent M_m is set to 500. The discount factor is set as $\gamma = 0.5$ with $T_{\text{step}} = 100$, which means that the weights of the target DQN are updated with the weights of the trained DQN at each agent per 100 time slots. In addition,

[1] For more details, please refer to our simulation codes and results at https://github.com/JungangGe/DRL_for_DDBC.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

TABLE I

COMPARISON OF REQUIRED INFORMATION FOR VARIOUS DDBC METHODS

Required Information	At BS k	At UE k	Scheme
FP (near-optimal)	$\mathbf{h}_{j,k}(t),\ \forall j \in \mathcal{K}$	$\mathbf{h}_{j,k}(t),\ \forall j \in \mathcal{K}$	Centralized
DRL (proposed)	$C_k(t),\ p_k(t),$ $I_k(t-1),$ $\lvert\mathbf{h}^{\dagger}_{k,k}(t-1)\bar{\mathbf{w}}_k(t-1)\rvert^2$	$\lvert\mathbf{h}^{\dagger}_{j,k}(t-1)\bar{\mathbf{w}}_j(t-1)\rvert^2$, where $j \in \mathcal{I}_k(t)$, $\lvert\mathbf{h}^{\dagger}_{j,k}(t)\bar{\mathbf{w}}_j(t-1)\rvert^2$, where $j \in \mathcal{I}_k(t)$, $\sum_{k' \in \mathcal{K}, k' \neq k} \lvert\mathbf{h}^{\dagger}_{k',k}(t-1)\bar{\mathbf{w}}_{k'}(t-1)\rvert^2 + \sigma^2$	Distributed
Greedy	$\mathbf{h}_{k,k}(t)$	$\mathbf{h}_{k,k}(t)$	Fully distributed
Random	None	None	Fully distributed

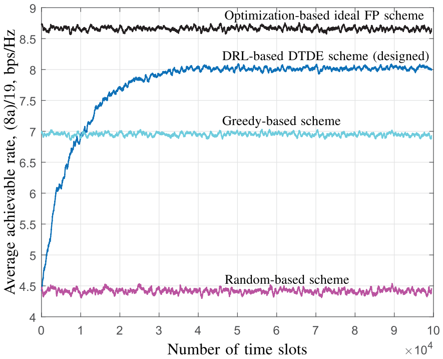


Fig. 7. Average achievable rate of different approaches in a dynamic scenario. Each value is a moving average over the previous 500 time slots.

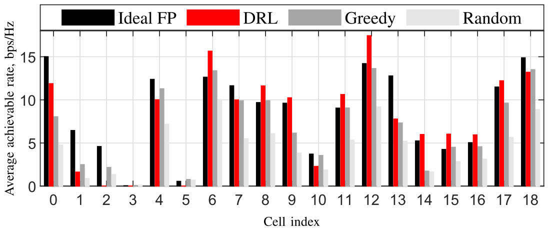


Fig. 8. Average achievable rates of each cell for different approaches in a dynamic scenario.

we employ an `RMSProp` optimizer with an adaptive learning rate to update $\boldsymbol{\theta}$ and set its initial learning rate as $\alpha(0) = 5e^{-4}$.

B. Preliminary Evaluation of the Proposed Distributed DRL-Based DTDE Scheme for DDBC

In the preliminary simulation, we validate the functionality of our proposed distributed DRL-based DTDE scheme by comparing it with that of the other three benchmark schemes, as shown in Table I. For the proposed DRL-based DTDE scheme, five levels of power and four codes are used, namely, $Q_{\text{pow}} = 5$ and $Q_{\text{code}} = 4$. The benchmark schemes are as follows:

- Optimization-based ideal FP scheme: Each agent obtains the action from Algorithm 1 with the instantaneous and perfect global CSIs.
- Greedy-based scheme: Each agent obtains the action by following a greedy approach.
- Random-based scheme: Each agent randomly chooses an action $a \in \mathcal{A}$ for all time slots.

As shown in Fig. 7, in a dynamic wireless environment, the FP algorithm fed with the instantaneous global CSI provides the largest average achievable rate. The designed greedy policy achieves a relatively good performance. The random policy provides the worst performance, as expected. From the simulation results, it can be seen that the average achievable rate of the proposed DRL-based DTDE scheme improves gradually with the training process, i.e., the decision-making policy is improved as the weights of the trained DQN are updated constantly. In the training process, the outcome of the DRL-based approach starts to exceed the greedy policy approximately at the end of $10,000$ time slots and finally converges to a fairly stable situation in approximately $40,000$ time slots. This indicates that the proposed DRL-based DTDE scheme has learned the trade-off between maximizing the achievable rate of an agent and maximizing the overall system capacity. Hence, the proposed DRL-based DTDE scheme can achieve approximately $90\ \%$ of the system capacity of the FP algorithm, using only the limited-information acquired from itself and its neighbors. Moreover, we consider the achievable rates of each cell, which provides useful details for evaluating the performance. We calculate the average achievable rates of each cell over $10,000$ time slots and the corresponding results are shown in Fig. 8. Since we consider a sum-rate maximization problem, it is observed that, the ideal FP approach and the proposed DRL-based DTDE scheme achieve relatively low average achievable rates of some cells due to their poor direct-link channel conditions, whereas they achieve high average achievable rates for the cells with good channel conditions. Higher average rates can be achieved by the proposed DRL-based scheme compared to the greedy and random based methods for 12 cells (i.e., 0, 6-9, 11-17), from 19 cells. It can also be observed that the proposed DRL-based approach outperforms the FP approach for 9 cells (i.e., 6, 8, 9, 11, 12, 14-17). This indicates that, for the considered sum-rate maximization problem in (8a), the proposed DRL-based scheme obtains sub-optimal solutions that are different of the sub-optimal solutions obtained by the FP approach.

In addition, the time required to obtain the solution of the proposed DRL-based scheme and that of the FP approach in the simulations can be compared to evaluate the computational complexity of the two schemes. Our simulation programs are executed on a personal computer with a single CPU (Intel Core i7-8700). Since the proposed DRL-based scheme allows each BS to determine its beamformer synchronously

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

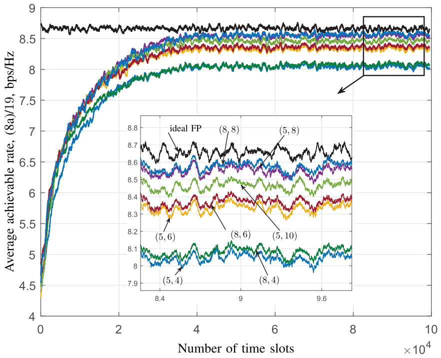


Fig. 9. Average achievable rate of the proposed DRL-based DTDE scheme with different sets of available actions in a dynamic scenario. Each value is a moving average over the previous 500 time slots.

in a local manner, the time required to obtain the solution of the proposed DRL-based scheme is the required time to make an online decision for an agent. In our simulations, the average time required to make an online decision for an agent is approximately 0.5 ms. However, the average time required to obtain the solution of the FP approach is approximately 8500 ms, which is much greater than that of the proposed DRL-based approach. In the centralized approaches, the central controller can send the applicable beamformers to the BSs only when the iterative algorithm converges. Hence, the large delay induced by waiting for the convergence of the FP approach makes it infeasible in a dynamic wireless environment. The proposed DRL-based scheme inducing a small delay, however, efficiently is applicable to the dynamic environment.

C. Performance Evaluation of the Proposed DRL-Based DTDE Scheme

In Fig. 7, it is also observed that an almost constant gap exists between the average achievable rate of an ideal FP algorithm and the saturated average achievable rate of the proposed DRL-based DTDE scheme. It should be noted that an FP algorithm can provide any of the infinite number of beam patterns for any arbitrary direction with any of the continuous transmit power levels, whereas the proposed DRL-based DTDE scheme provides the designed beam patterns for a finite number of directions with the selected power with the designed power levels. Thus, we use seven different sets of available actions to evaluate the effects of the performance of the proposed DRL-based DTDE scheme. With all the other hyperparameters fixed, we generate seven sets of available actions with different combinations of Q_{pow} and Q_{code}, denoted by $(Q_{\text{pow}}, Q_{\text{code}})$, namely, $(5,4)$, $(5,6)$, $(5,8)$, $(5,10)$, $(8,4)$, $(8,6)$, and $(8,8)$. The corresponding beam patterns of $Q_{\text{code}} = 4$, $Q_{\text{code}} = 6$, $Q_{\text{code}} = 8$, and $Q_{\text{code}} = 10$ are shown in Figs. 6(a), (b), (c), and (d), respectively. Note that the number of available power levels and beam directions increases as Q_{pow} and Q_{code} increase, respectively.

In Fig. 9, it can be seen that, while the number of available beam directions is fixed, the proposed DRL-based schemes with different number of power levels achieve almost the same performance, e.g., the performance with (8,4) is similar to that with (5,4). However, while the number of available power levels is fixed, the DRL-based scheme achieves much better performance with more available beam directions, i.e., $(5,6)$, $(5,8)$, and $(5,10)$, compared to that with $(5,4)$. From the results, we can surmise that increasing Q_{code} is more effective on the average achievable rate, i.e., system capacity, improvement than increasing Q_{pow}, which indicates that the number of available beam directions plays a more critical role in the proposed DRL-based DTDE scheme. Furthermore, it is observed that the proposed DRL-based DTDE scheme with appropriate Q_{pow} and Q_{code} (i.e., $(5,8)$) achieves an average achievable rate that is very close to that of the FP algorithm. Here, it should be noted that just increasing Q_{pow} and Q_{code} does not guarantee performance improvement because the optimal policy in the increased action space is more difficult to be learned through a trial-and-error mechanism. That is why $(5,10)$ does not provide better performance than $(5,8)$.

In addition, we evaluate the performance of the proposed DRL-based approach with various cardinality constraints on the two neighbor sets defined in Section III-C, i.e., parameter U. As introduced in Section III-C, the increase of U means that the BS needs to exchange information with more neighboring BSs in the designed limited-information exchange process. Thus, the impact of increasing U can be described from two points of view. Larger overhead is induced since more signallings are involved in the information exchange process. However, BS can possibly acquire more valuable observations by exchanging information with more neighboring BSs, which are helpful for making better decisions. Here, while the set of available actions is fixed as $(Q_{\text{pow}}, Q_{\text{code}}) = (5,8)$, we use five different values for U, i.e., $U = 3$, 4, 5, 6, and 7, to evaluate its impact on the performance of the proposed DRL-based scheme.

The corresponding simulation results are shown in Fig. 10. In general, the performance of the proposed scheme becomes better with the increase of U. The DRL-based scheme with $U = 3$ achieves relatively poor performance, the scheme with $U = 4$ realizes a moderate performance, whereas the schemes with $U = 5$, 6, and 7 similarly achieve the closest (best) performance to the performance of an ideal FP. We eventually choose $U = 5$ for the tradeoff between the performance and the overhead of the information exchange.

D. Comparison of CTDE and DTDE Schemes

Two DRL-based schemes, namely, CTDE and DTDE, are compared with respect to the average achievable rate. Since the complexity of the CTCE scheme is prohibitively large, e.g., 20^{19} possible joint actions need to be learned when $(Q_{\text{pow}}, Q_{\text{code}}) = (5,4)$ for 19 BSs, it is inapplicable to the multi-agent learning system and omitted in the comparison. The DRL-based CTDE scheme has been found to be advantageous in some scenarios [26]. However, as shown in Fig. 11, the DRL-based CTDE scheme may not be suitable for the DDBC problem in a dynamic wireless channel environment. This can be explained by the fact that the DRL-based CTDE scheme finds the common features among all the agents.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

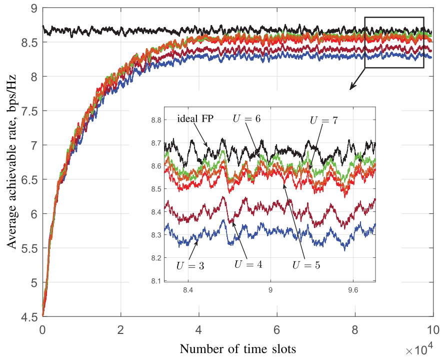


Fig. 10. Average achievable rate evaluation results for different cardinality constraints on the number of neighbors. Each value is a moving average over the previous 500 time slots.

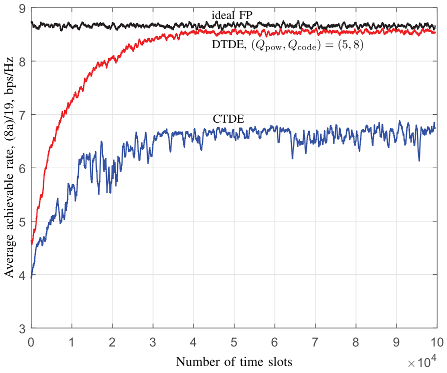


Fig. 11. Average achievable rate of the two DRL-based schemes, namely, CTDE and DTDE, in a dynamic scenario. Each value is a moving average over the previous 500 time slots.

It should be noted that the proposed DRL-based DTDE scheme is much better at finding characteristics that are different among the cells because it allows each agent (i.e., BS) to train a specific DQN. We may infer that the DRL-based CTDE scheme with the public DQN is more suitable for multi-agent scenarios, where all the agents make decisions under a similar situation so that an agent could benefit from the experiences of the other agents, such as the transmit power control problem in cellular networks [26]. In our considered scenario, however, each BS selects a beam pattern for its intended UE, and the optimal beam pattern is relevant to the DoD of the downlink channels, which differs considerably among the cells. Then an agent could be confused by the experiences of the other agents during the offline training procedure in the CTDE approach; hence, the proposed DTDE scheme outperforms the CTDE scheme under the considered dynamic wireless channel environment.

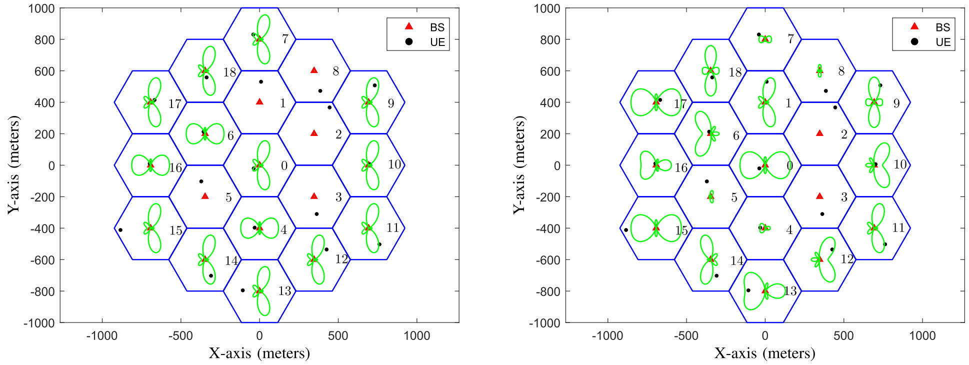
To further justify the above inference that the CTDE scheme does not perform well unlike the DTDE scheme, the two snapshots of CTDE and DTDE schemes in a same time slot are shown in Figs. 12(a) and (b), respectively, after they get saturated. In particular, we plot the locations of the BSs and UEs with red triangles and black dots, respectively. The blue hexagons denote the boundary of the cells. In addition, the green curve in each cell refers the transmit beamforming pattern adopted by the corresponding BS, where the length of the beam represents the transmit power of the BS. As shown in Fig. 12(a), after the CTDE scheme get saturated, most of the BSs tend to select a similar transmit beamforming pattern (e.g., the patterns of BSs 0, 7, 9–15, 17, and 18) because they perform actions using the DQN with the same weights. In other words, a BS is misled by the other BSs when it makes its own decisions because the public DQN are trained with the experiences of all BSs. It should also be noted that five BSs (i.e., BSs 1, 2, 3, 5, and 8) are turned off in the snapshot. This can be explained by the fact that we consider a sum-rate maximization objective; therefore, so in a time slot, if the contribution of a BS to the objective is less than the sum of the achievable rate loss of its interfered neighbors, i.e., the reward function (26) is negative, the BS tends to not transmit messages in the time slot. On the other hand, the snapshot of the DTDE scheme shown in Fig. 12(b) shows that only two BSs (i.e., BSs 2 and 3) are turned off. Furthermore, the BSs that are turned on in the DTDE scheme generate various shapes of beamforming patterns, so that they can improve the achievable rate of its intended UE as well as reduce the interference caused to other cells.

E. Discussions About the Feasibility of Model Retraining in the Proposed DRL-Based Scheme

In nature, the DRL-based scheme is a learning-based approach, in which all the agents can steadily improve their decision-making policies through their interactions with the environment. Thus, it is inevitable that the DRL-based scheme usually needs a period to reach a stationary point, i.e, the convergence. However, if the served UE in a single cell is changed by scheduling, the DRL-based beamforming coordination process should be restarted and the DQNs needs to be retrained. It is impractical to retrain the DQNs from scratch, i.e., re-initialize all the parameters and then retrain the DQNs following the procedure shown in Algorithm 2, whenever the network encounters a rescheduling. Moreover, it may be helpful if the training process is continued so that the weights in the previous well-trained DQNs can be utilized to retrain the new model. Hence, we evaluate the feasibility of the retrain that continuously reuses the previous training process when the network encounters a rescheduling.

We assume that the network does not reschedule the users until the DRL-based scheme can reach the first convergence through training from scratch. After this, when the network encounters a rescheduling after the first convergence, the DQNs can continue training following the previous training process. Here, we re-initialize the optimizers with the initial learning rate $\alpha(0) = 5e^{-4}$ in the retraining process since the learning rate of the optimizers decreased a lot in the previous training process and the re-initialized optimizers can help the agents modify their policies quickly. In addition, the rescheduling procedure in the network can be represented by a random change on the network topology (i.e., the users are regenerated randomly) in the simulation. We consider to set $(Q_{\rm pow}, Q_{\rm code}) = (5, 8)$ and $U = 5$ in the simulation, the corresponding simulation results are shown in Fig. 13. From the

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.



(a)

(b)

Fig. 12. Two snapshots of downlink-beamforming coordination after saturation of the DRL-based DTDE scheme and DRL-based CTDE scheme, respectively: (a) DRL-based CTDE scheme (b) DRL-based DTDE scheme.

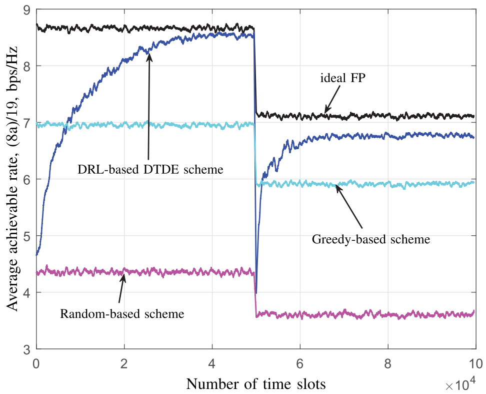


Fig. 13. Average achievable rate of different approaches in a dynamic scenario including a network rescheduling procedure, which happens at the 50, 000th time slot. Each value is a moving average over the previous 500 time slots.

results, it can be seen that the performance of all schemes change abruptly at the 50, 000th time slot because the network is rescheduled and the achievable network capacity is related to the scheduling scheme. As we continue the training process with re-initialized optimizers, the performance of DRL-based DTDE scheme can exceed that of the greedy-based scheme quickly and converges after approximately 10, 000 time slots. The convergence time in the retraining process is much shorter than that in the previous training process. With the quick convergence, the performance gain still exists compared other non-iterative schemes (e.g., greedy-based scheme) and the delay required for the retraining process is feasible. However, the performance degradation from the ideal FP increases and the retraining time is inevitable; therefore, the proposed DRL-based DTDE scheme is attractive especially for the scenarios where the users are not rescheduled very frequently.

VI. Conclusion

In this study, we investigated the DDBC problem in MISO-IC and proposed a distributed DRL-based DDBC method to adapt the beamformers of all BSs in the cellular network to the distributed nature of our proposed approach. The proposed DDBC method is employed to maximize the system capacity through the proposed DRL-based DTDE scheme that allows for each BS to learn the dynamics of the environment from local observations and to select an appropriate beamformer with the transmit power level needed to enhance the system capacity. The simulation results show that with a much lower system overhead, i.e., the designed limited-information exchange protocol and the distributed nature, the proposed DRL-based DTDE scheme implements the DDBC method and achieves a system capacity that is very close to that of the ideal FP algorithm with global instantaneous CSI measurements.

References

[1] F. Rashid-Farrokhi, K. J. R. Liu, and L. Tassiulas, “Transmit beamforming and power control for cellular wireless systems,” *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1437–1450, Oct. 1998.

[2] D. Gesbert, S. Hanly, H. Huang, S. Shamai (Shitz), O. Simeone, and W. Yu, “Multi-cell MIMO cooperative networks: A new look at interference,” *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[3] R. Zakhour and D. Gesbert, “Distributed multicell-MISO precoding using the layered virtual SINR framework,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 8, pp. 2444–2448, Aug. 2010.

[4] H. Dai, A. F. Molisch, and H. V. Poor, “Downlink capacity of interference-limited MIMO systems with joint detection,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 2, pp. 442–453, Mar. 2004.

[5] H. Zhang and H. Dai, “Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks,” *EURASIP J. Wireless Commun. Netw.*, vol. 2004, no. 2, pp. 222–235, Dec. 2004.

[6] J. Qiu, R. Zhang, Z.-Q. Luo, and S. Cui, “Optimal distributed beamforming for MISO interference channels,” *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5638–5643, Nov. 2011.

[7] L.-N. Tran, M. F. Hanif, A. Tolli, and M. Juntti, “Fast converging algorithm for weighted sum rate maximization in multicell MISO downlink,” *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 872–875, Dec. 2012.

[8] H. Dahrouj and W. Yu, “Coordinated beamforming for the multicell multi-antenna wireless system,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May 2010.

[9] R. Zhang and S. Cui, “Cooperative interference management with MISO beamforming,” 2009, *arXiv:0910.2771*. [Online]. Available: http://arxiv.org/abs/0910.2771

[10] H. Pennanen, A. Tolli, and M. Latva-aho, “Decentralized robust beamforming for coordinated multi-cell MISO networks,” *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 334–338, Mar. 2014.

[11] K. Shen and W. Yu, “Fractional programming for communication systems—Part I: Power control and beamforming,” *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

[12] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[13] S. K. Joshi, P. C. Weeraddana, M. Codreanu, and M. Latva-aho, "Weighted sum-rate maximization for MISO downlink cellular networks via branch and bound," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 2090–2095, Apr. 2012.

[14] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[15] M. Simsek, M. Bennis, and A. Czylwik, "Dynamic inter-cell interference coordination in HetNets: A reinforcement learning approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Anaheim, CA, USA, Dec. 2012, pp. 5446–5450.

[16] N. Morozs, T. Clarke, and D. Grace, "Heuristically accelerated reinforcement learning for dynamic secondary spectrum sharing," *IEEE Access*, vol. 3, pp. 2771–2783, Dec. 2015.

[17] Y. Sun, G. Feng, S. Qin, Y.-C. Liang, and T.-S.-P. Yum, "The SMART handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 6, pp. 1456–1468, Jun. 2018.

[18] V. Raj, I. Dias, T. Tholeti, and S. Kalyani, "Spectrum access in cognitive radio using a two-stage reinforcement learning approach," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 20–34, Feb. 2018.

[19] M. Qiao, H. Zhao, L. Zhou, C. Zhu, and S. Huang, "Topology-transparent scheduling based on reinforcement learning in self-organized wireless networks," *IEEE Access*, vol. 6, pp. 20221–20230, Apr. 2018.

[20] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[21] G. Han, L. Xiao, and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2087–2091.

[22] H. Ye and G. Y. Li, "Deep reinforcement learning for resource allocation in V2 V communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[23] J. Tan, L. Zhang, Y.-C. Liang, and D. Niyato, "Deep reinforcement learning for the coexistence of LAA-LTE and WiFi systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.

[24] L. Zhang, J. Tan, Y.-C. Liang, G. Feng, and D. Niyato, "Deep reinforcement learning-based modulation and coding scheme selection in cognitive heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3281–3294, Jun. 2019.

[25] Q. Zhang, Y.-C. Liang, and H. V. Poor, "Intelligent user association for symbiotic radio networks using deep reinforcement learning," *IEEE Trans. Wireless Commun.*, early access, Apr. 7, 2020, doi: 10.1109/TWC.2020.2984758.

[26] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[27] Y.-C. Liang and F. P. S. Chin, "Downlink channel covariance matrix (DCCM) estimation and its applications in wireless DS-CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 2, pp. 222–232, Feb. 2001.

[28] M. Dong, L. Tong, and B. M. Sadler, "Optimal insertion of pilot symbols for transmissions over time-varying flat fading channels," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1403–1418, May 2004.

[29] K. Shen and W. Yu, "Fractional programming for communication systems—Part II: Uplink scheduling via matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, May 2018.

[30] K. Shen and W. Yu, "A coordinated uplink scheduling and power control algorithm for multicell networks," in *Proc. 49th Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2015, pp. 1305–1309.

[31] H. Zhang, L. Venturino, N. Prasad, P. Li, S. Rangarajan, and X. Wang, "Weighted sum-rate maximization in multi-cell networks via coordinated scheduling and discrete power control," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1214–1224, Jun. 2011.

[32] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Mach. Learn.*, vol. 38, no. 3, pp. 287–308, Mar. 2000.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[34] W. Zou, Z. Cui, B. Li, Z. Zhou, and Y. Hu, "Beamforming codebook design and performance evaluation for 60GHz wireless communication," in *Proc. 11th Int. Symp. Commun. Inf. Technol. (ISCIT)*, Hangzhou, China, Oct. 2011, pp. 30–35.

[35] J. Huang, R. A. Berry, and M. L. Honig, "Distributed interference compensation for wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 5, pp. 1074–1084, May 2006.

[36] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. Munoz de Cote, "A survey of learning in multiagent environments: Dealing with non-stationarity," 2017, *arXiv:1707.09183*. [Online]. Available: http://arxiv.org/abs/1707.09183

[37] G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht, "Dealing with non-stationarity in multi-agent deep reinforcement learning," 2019, *arXiv:1906.04737*. [Online]. Available: http://arxiv.org/abs/1906.04737

[38] A. Marinescu, I. Dusparic, and S. Clarke, "Prediction-based multi-agent reinforcement learning in inherently non-stationary environments," *ACM Trans. Auto. Adapt. Syst.*, vol. 12, no. 2, pp. 1–23, May 2017.

[39] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, Syst.*, vol. 2, no. 4, pp. 303–314, Dec. 1989.

[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

Jungang Ge (Student Member, IEEE) received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China (UESTC), China, in 2019, where he is currently pursuing the Ph.D. degree. His research interests include machine learning techniques and wireless communications.

Ying-Chang Liang (Fellow, IEEE) is currently a Professor with the University of Electronic Science and Technology of China, China, where he leads the Center for Intelligent Networking and Communications and serves as the Deputy Director of the Artificial Intelligence Research Institute. He was a Professor with The University of Sydney, Australia, a Principal Scientist, and a Technical Advisor with the Institute for Infocomm Research, Singapore, and a Visiting Scholar with Stanford University, USA. His research interests include wireless networking and communications, cognitive radio, symbiotic networks, dynamic spectrum access, the Internet of Things, artificial intelligence, and machine learning techniques.

Dr. Liang is a Foreign Member of Academia Europaea. He has been recognized by Thomson Reuters (now Clarivate Analytics) as a Highly Cited Researcher since 2014. He received the Prestigious Engineering Achievement Award from The Institution of Engineers, Singapore, in 2007, the Outstanding Contribution Appreciation Award from the IEEE Standards Association, in 2011, and the Recognition Award from the IEEE Communications Society Technical Committee on Cognitive Networks, in 2018. He was a recipient of numerous paper awards, including the IEEE Jack Neubauer Memorial Award in 2014 and the IEEE Communications Society APB Outstanding Paper Award in 2012. He was the Chair of the IEEE Communications Society Technical Committee on Cognitive Networks. He served as the TPC Chair and Executive Co-Chair of the IEEE Globecom'17. He is the Founding Editor-in-Chief of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS: COGNITIVE RADIO SERIES and the Key Founder and the Editor-in-Chief of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is also serving as an Associate Editor-in-Chief for *China Communications*. He served as a Guest/Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, *IEEE Signal Processing Magazine*, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORK. He was also an Associate Editor-in-Chief of *World Scientific Journal on Random Matrices: Theory and Applications*. He was a Distinguished Lecturer of the IEEE Communications Society and the IEEE Vehicular Technology Society.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.

Jingon Joung (Senior Member, IEEE) received the B.S. degree in radio communication engineering from Yonsei University, Seoul, South Korea, in 2001, and the M.S. and Ph.D. degrees in electrical engineering and computer science from KAIST, Daejeon, South Korea, in 2003 and 2007, respectively.

He was a Postdoctoral Fellow with KAIST, South Korea, and UCLA, CA, USA, in 2007 and 2008, respectively. He was a Scientist with the Institute for Infocomm Research (I^{2}R), Agency for Science, Technology, and Research (A*STAR), Singapore, from 2009 to 2015, and joined Chung-Ang University (CAU), Seoul, in 2016, as a Faculty Member. He is currently an Associate Professor with the School of Electrical and Electronics Engineering, CAU, where he is also the Principal Investigator of the Intelligent Wireless Systems Laboratory. His research interests include wireless communication signal processing, numerical analysis, algorithms, and machine learning.

Dr. Joung was a recipient of the First Prize of the Intel-ITRC Student Paper Contest in 2006. He was recognized as the Exemplary Reviewers of IEEE COMMUNICATIONS LETTERS in 2012 and IEEE WIRELESS COMMUNICATIONS LETTERS in 2012, 2013, 2014, and 2019. He served as the Guest Editor for IEEE ACCESS in 2016, the Editorial Board Member for *APSIPA Transactions on Signal and Information Processing* from 2014 to 2019, and a Guest Editor for *MDPI Electronics* in 2019. He is currently serving as the Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and *MDPI Sensors*.

Sumei Sun (Fellow, IEEE) is currently a Principal Scientist and the Head of the Communications and Networks Department, Institute for Infocomm Research (I2R), Singapore. She is also holding a joint appointment with the Singapore Institute of Technology, and an adjunct appointment with the National University of Singapore, as a Full Professor. Her current research interests are in cognitive communications and networks, next-generation wireless communications, and industrial Internet of Things. She is an Editor-in-Chief of IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, a member of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS STEERING COMMITTEE, and a Distinguished Speaker of the IEEE Vehicular Technology Society for the period of 2018–2021. She is also the Director of the IEEE Communications Society Asia Pacific Board and the Chapter Coordinator of Asia Pacific Region in the IEEE Vehicular Technologies Society, and a member of the IEEE Communications Society Globecom/ICC Management and Strategy Standing Committee.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on April 16,2022 at 12:09:24 UTC from IEEE Xplore. Restrictions apply.