Distributed Multi-Cell Multi-User MISO Downlink Beamforming via Deep Reinforcement Learning

JIA Haonan[1], HE Zhenqing[1], TAN Wanlong[1],
RUI Hua[2,3], LIN Wei[2,3]

(1. National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China)

DOI: 10.12142/ZTECOM.202204009

https: //kns.cnki.net/kcms/detail/34.1294.TN.20221124.0844.002.html, published online November 25, 2022

Manuscript received: 2022-02-24

Abstract: The sum rate maximization beamforming problem for a multi-cell multi-user multiple-input single-output interference channel (MISO-IC) system is considered. Conventionally, the centralized and distributed beamforming solutions to the MISO-IC system have high computational complexity and bear a heavy burden of channel state information exchange between base stations (BSs), which becomes even much worse in a large-scale antenna system. To address this, we propose a distributed deep reinforcement learning (DRL) based approach with limited information exchange. Specifically, the original beamforming problem is decomposed of the problems of beam direction design and power allocation and the costs of information exchange between BSs are significantly reduced. In particular, each BS is provided with an independent deep deterministic policy gradient network that can learn to choose the beam direction scheme and simultaneously allocate power to users. Simulation results illustrate that the proposed DRL-based approach has comparable sum rate performance with much less information exchange over the conventional distributed beamforming solutions.

Keywords: deep reinforcement learning; downlink beamforming; multiple-input single-output interference channel

Citation (IEEE Format): H. N. Jia, Z. Q. He, W. L. Tan, et al., "Distributed multi-cell multi-user miso downlink beamforming via deep reinforcement learning," *ZTE Communications*, vol. 20, no. 4, pp. 69 – 77, Dec. 2022. doi: 10.12142/ZTECOM.202204009.

1 Introduction

To meet the increasing wireless communication traffic demand, the frequency reuse factor of cellular systems is expected to be slackened as one, which indicates that all the cells operate in the same frequency band. However, the small frequency reuse factor brings an inter-cell interference problem, heavily degrading the achievable sum rate performance of the wireless system. Therefore, inter-cell interference should be managed carefully. A multi-cell multiple-input single-output (MISO) downlink beamforming technique with cooperation among the base stations (BSs) is introduced as a promising solution. The typical zero-forcing beamforming[1] works in a highly coordinated scenario where each piece of user equipment (UE) is served by all the BSs, which needs all the transmit data and channel state information (CSI) to be shared among the BSs. Nevertheless, it is impractical due to the heavy data-sharing burden[2]. A centralized solution[3] collects the global CSI and jointly design beamforming vectors based on fractional programming (FP). Although it can achieve near-optimal performance, it has high computational complexity and leads to unavoidable delays when collecting CSI and sending beamforming vectors, thereby making it impossible to be applied in a dynamic channel environment.

Many distributed schemes are proposed to reduce the computational cost of centralized solutions. In particular, an achievable rate region of the multi-cell MISO Gaussian interference channel (MISO-IC) system is analyzed in Ref. [4], which proves that the well-designed distributed schemes can reach the Pareto boundary. To reduce information sharing among BSs, a data-sharing-based distributed scheme is proposed in Ref. [5]. A fully distributed scheme with no CSI or data sharing is discussed in Ref. [6], which works well at a high signal-to-interference-plus-noise-ratio (SINR). However, these works assume that the BSs are capable of obtaining the instantaneous downlink CSI of the UE in other cells without

This work is supported by the joint research project with ZTE Corporation under Grant No. HC-CN-2020120002.

CSI sharing, which is also infeasible in a practical system.

Deep reinforcement learning (DRL) has shown great potential in decision-making problems. By converting the multi-cell downlink beamforming problem into a decision-making problem, several distributed approaches based on DRL are developed[7–8]. Particularly, a multi-agent deep Q-learning based approach is introduced in Ref. [7], in which each BS learns to make the decisions of beam direction and power for each UE based on the local CSI and the exchanged information among the BSs. However, because of the curse of dimensionality[9], the Q-learning based approach in Ref. [7] is almost impossible to be applied in the cases where there are multiple user devices in the same cell or the BSs are equipped with large-scale antennas, since the discrete action space expands exponentially with the number of user devices and antennas.

In this paper, we develop a distributed-training distributed-execution (DTDE) multi-agent deep deterministic policy gradient (MADDPG) based algorithm to maximize the instantaneous sum rate of the multicell multi-user MISO-IC system under the power constraint for each BS. Thanks to the features of DDPG, the policy network gives continuous value directly, which significantly reduces the dimension of actions. Our main contributions are summarized as follows:

1) A new distributed MADDPG-based scheme is proposed, capable of solving the instantaneous sum rate maximization problem when cells have multiple user devices and BSs are equipped with large-scale antennas. By decomposing the original beamforming problem into the beam direction design and power allocation problems, each BS as an agent can learn to choose beam direction and power allocation based on the wireless environment.

2) A new limited information exchange protocol is proposed for the distributed BSs to share information for beamforming design. Instead of sharing CSI directly, we choose the equivalent channel gains of UE, the received interference of UE, and the sum rate of UE in one cell as the shared information. Different from other DRL-based algorithms which only consider equivalent channel gains and the sum rate of UE, we consider the received interference (also known as the interference temperature) as the crucial information.

3) Extensive experiments are conducted to evaluate the efficiency and scalability of the proposed MADDPG approach by comparing the conventional distributed and centralized solutions. The simulation results show that the proposed MADDPG can reach the state-of-the-art sum rate performance with a much smaller amount of information sharing among BSs.

As far as we know, this is the first attempt to tackle the multi-cell MISO beamforming via MADDPG-based DRL. In contrast to the related work[7], this paper aims to solve the multi-cell sum rate maximization problem in the continuous action space by using the MADDPG method which is more flexible for different wireless environments and is easy for agents (e.g., BSs) to learn since the dimension of action space is much smaller than that of codebook space in Ref. [7].

In this paper, we use $\mathbb{C}^{m \times n}$ and $\mathbb{R}^{m \times n}$ to represent the spaces of the $m \times n$ dimensional complex number and real number, respectively. The superscripts "$*$", "T", and "H" denote the conjugate, the transpose, and the conjugate transpose, respectively. In addition, we use $\jmath \triangleq \sqrt{-1}$, $\mathbb{E}\{\cdot\}$, and $\|\cdot\|$ as the imaginary unit, the expectation operator, and the ℓ_2 norm, respectively.

2 System Model

We consider a wireless cellular downlink system of N cells, in each of which there is a multi-antenna transmitter (e.g., a BS) with M antennas to serve K single-antenna receivers (e.g., UE). We use $\boldsymbol{N} = \{1,\cdots,N\}$ to denote the set of all BSs. We assume that all UE in this system shares the same frequency band, thereby leading to both intra-cell and inter-cell interference with each UE. As a result, the received signal of the k-th UE in the n-th cell at time t can be expressed as:

$$y_{n,k}(t) = \underbrace{\boldsymbol{h}^{\mathrm{T}}_{n,n,k}(t)\boldsymbol{w}_{n,k}(t)x_{n,k}(t)}_{\text{desired signal}} + \underbrace{\sum_{j=1,j\neq k}^{K} \boldsymbol{h}^{\mathrm{T}}_{n,n,k}(t)\boldsymbol{w}_{n,j}(t)x_{n,j}(t)}_{\text{intra-cell interference}} + \underbrace{\sum_{i=1,i\neq n}^{N} \sum_{j=1}^{K} \boldsymbol{h}^{\mathrm{T}}_{i,n,k}(t)\boldsymbol{w}_{i,j}(t)x_{i,j}(t)}_{\text{inter-cell interference}} + z_{n,k}(t), \tag{1}$$

where $\boldsymbol{h}_{i,n,k}(t) \in \mathbb{C}^{M}$ denotes the downlink channel vector from the BS in the i-th cell to the k-th UE in the n-th cell, $\mathbf{w}_{i,n,k}(t) \in \mathbb{C}^{M}$ denotes the beamforming vector for the j-th UE in the i-th cell, $x_{i,j}$ denotes the transmitted symbol to the j-th UE in the i-th cell, and $z_{i,j}(t) \sim \mathcal{CN}\left(0,\sigma^2_{n,k}\right)$ denotes the additive noise with $\sigma^2_{n,k}$ being the noise power. Under the single user detection mechanism, the instantaneous SINR and achievable rate of the k-th UE in the n-th cell are given by:

$$\gamma_{n,k}(t) = \frac{\left|\boldsymbol{h}^{\mathrm{T}}_{n,n,k}(t)\boldsymbol{w}_{n,k}(t)\right|^2}{\beta^{\text{intra}}_{n,k}(t) + \beta^{\text{inter}}_{n,k}(t) + \sigma^2_{n,k}}, \tag{2a}$$

$$R_{n,k}(t) = \log_2\left(1 + \gamma_{n,k}(t)\right), \tag{2b}$$

where $\beta^{\text{intra}}_{n,k}(t) = \sum_{j=1,j\neq k}^{K} |\boldsymbol{h}^{\mathrm{T}}_{n,n,k}(t)\boldsymbol{w}_{n,j}(t)|^2$ and $\beta^{\text{inter}}_{n,k}(t) = \sum_{i=1,i\neq n}^{N} \sum_{j=1}^{K} |\boldsymbol{h}^{\mathrm{T}}_{i,n,k}(t)\boldsymbol{w}_{i,j}(t)|^2$ represent the intra-cell and inter-cell interferences.

We assume that the BS in each cell is equipped with the uniform rectangular array (URA) structure with $M = M_x M_y$ antenna elements,[1] where M_x and M_y denote the horizontal and vertical scales of the URA, respectively. According to the ray-based channel modeling[10], the dynamic URA channel response of $\boldsymbol{h}_{n,j,k}(t)$ with L-paths for the n-th BS to the k-th UE in the cell j can be expressed as:

$$\boldsymbol{h}_{n,j,k}(t) = \sqrt{\kappa_{n,j,k}} \sum_{l=1}^{L} g_{n,j,k,l}(t) \bar{\boldsymbol{a}}\left(u_{n,j,k,l}\right) \otimes \bar{\boldsymbol{b}}\left(v_{n,j,k,l}\right), \tag{3}$$

where $\kappa_{n,j,k}$ is the large-scale fading factor related to the path loss and shadowing and $g_{n,j,k,l}(t)$ is the dynamic small-scale Rayleigh fading factor. The steering vectors $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{b}}$ of URA in Eq. (3) are given by:

$$\bar{\boldsymbol{a}}\left(u_{n,j,k,l}\right) = \left[1, e^{-u_{n,j,k,l}}, \cdots, e^{-\left(M_x - 1\right)u_{n,j,k,l}}\right]^{\top}, \tag{4a}$$

$$\bar{\boldsymbol{b}}\left(v_{n,j,k,l}\right) = \left[1, e^{-v_{n,j,k,l}}, \cdots, e^{-\left(M_y - 1\right)v_{n,j,k,l}}\right]^{\top}, \tag{4b}$$

where $u_{n,j,k,l} = \frac{2\pi d_1}{\lambda} \sin\left(\theta_{n,j,k,l}\right) \cos\left(\phi_{n,j,k,l}\right)$, $v_{n,j,k,l} = \frac{2\pi d_2}{\lambda} \cos\left(\theta_{n,j,k,l}\right)$, $d_1 = d_2 = \frac{\lambda}{2}$, and λ is the signal wave length. The elevation angle-of-departure (AoD) $\theta_{n,j,k,l}$ and azimuth AoD $\phi_{n,j,k,l}$ of each path are given by:

$$\theta_{n,j,k,l} \sim \mathcal{U}\left(\theta_{n,j,k} - \frac{\Delta}{2}, \theta_{n,j,k} + \frac{\Delta}{2}\right), \tag{5a}$$

$$\phi_{n,j,k,l} \sim \mathcal{U}\left(\phi_{n,j,k} - \frac{\Delta}{2}, \phi_{n,j,k} + \frac{\Delta}{2}\right), \tag{5b}$$

where $\theta_{n,j,k}$ and $\phi_{n,j,k}$ are the nominal AoD between the cell n and the UE k in cell j, respectively, and Δ is the associated angular perturbation.

To characterize the small-scale fading dynamics of the time varying channel, we utilize the following first-order Gauss-Markov process[11].

$$g_{n,j,k,l}(t+1) = \sqrt{\rho}\, g_{n,j,k,l}(t) + \sqrt{1-\rho}\, \delta_{n,j,k,l}(t), \tag{6}$$

where ρ denotes the fading correlation coefficient between any two consecutive time slots and $\delta_{n,j,k,l}(t) \sim \mathcal{CN}(0,1)$.

3 Problem Statement

Considering the channel variation, we aim to solve the following instantaneous achievable sum-rate maximization problem at the time slot t:

$$\begin{aligned} &\max_{\boldsymbol{w}_{n,k}(t)} R_{\text{sum}}(t) = \sum_{n=1}^{N} \sum_{k=1}^{K} R_{n,k}(t) \\ &\text{s.t.} \sum_{k=1}^{K} \left\| \boldsymbol{w}_{n,k}(t) \right\|^2 \leqslant P_{\max}, \forall n, \end{aligned} \tag{7}$$

where $P_{\max}$ denotes the maximum transmit power for each BS. Unfortunately, Problem (7) is generally NP-hard even with global CSI, and finding its globally optimal solution requires exponential running time[3]. Conventionally, several centralized methods are proposed to find a sub-optimal solution. All centralized algorithms assume that there is a central node to collect global CSI from all BSs, and then the central node computes and returns beamformers of all BSs. However, it is hard to obtain the global CSI for all BSs. Moreover, due to the dynamics of channels, the beamformers are already outdated when the BSs obtain the returns. Therefore, it is more reasonable to apply a distributed approach. However, information sharing design between the BSs is also a problem for the distributed methods. Generally, the BSs communicate with other BSs through the backhaul links. Conventional beamforming methods need the BSs to exchange global or cross-talk CSI, which is an unacceptable burden for the rate-limited backhaul links. Therefore, the amount of shared information for beamforming should be limited, and we try to seek a sub-optimal distributed solution with limited information exchange between the BSs in different cells.

From the perspective of the BS n, the beamformer $\boldsymbol{w}_{n,k}$ can be expressed as:

$$\boldsymbol{w}_{n,k}(t) = \sqrt{P_{n,k}(t)}\, \overline{\boldsymbol{w}}_{n,k}(t), \tag{8}$$

where $P_{n,k}(t) = \left\| \boldsymbol{w}_{n,k}(t) \right\|^2$ denotes the transmit power of the BS n to user k and $\overline{\boldsymbol{w}}_{n,k}(t)$ denotes the corresponding normalized beamformer, which represents the direction of the transmit beam. Note that once the beam direction is fixed, the beam power allocation only needs the equivalent channel and interference information, which significantly reduces the cost of the information exchange[3].

The typical beam direction solutions include the virtual SINR[12] and the weighted minimum-mean-square-error (WMMSE) [13]. However, these solutions are closely coupled with power allocation strategies, which cannot be easily adopted to guide the beam direction design. According to Ref. [14], given global CSI in the multi-cell scenario, the optimal beamformer can be expressed as a linear combination of the conventional zero-forcing (ZF) and maximum ratio transmission (MRT). However, it is difficult to obtain the instantaneous global CSI. This inspires us to apply the available ZF and MRT to give heuristic solutions to our proposed approach based on only local CSI[5]. Specifically, the ZF and the MRT solutions are given by:

[1] We assume the URA model here for simplicity. Nevertheless, the proposed scheme can be applicable to arbitrary array geometry.

$$\bar{\boldsymbol{w}}^{\mathrm{ZF}}_{n,k} = \frac{\left(\boldsymbol{H}^{\mathrm{H}}_{n}\left(\boldsymbol{H}_{n}\boldsymbol{H}^{\mathrm{H}}_{n}\right)^{-1}\right)^{\mathrm{T}}_{[:,k]}}{\left\|\left(\boldsymbol{H}^{\mathrm{H}}_{n}\left(\boldsymbol{H}_{n}\boldsymbol{H}^{\mathrm{H}}_{n}\right)^{-1}\right)^{\mathrm{T}}_{[:,k]}\right\|}, \tag{9a}$$

$$\bar{\boldsymbol{w}}^{\mathrm{MRT}}_{n,k} = \frac{\boldsymbol{h}^{*}_{n,n,k}}{\|\boldsymbol{h}_{n,n,k}\|}, \tag{9b}$$

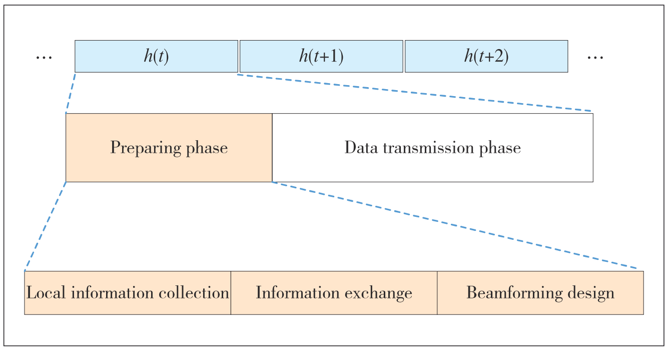
where $\boldsymbol{H}_n \in \mathbb{C}^{K \times M}$ denotes the downlink channel of all K users in the n-th cell. Note that the MRT works well at low SNR regions[15], especially in the case that most UE is at the edge of the cell, since it only focuses on the maximization of the received signal power. In contrast, ZF tries to minimize the received interference for the UE, which makes it outperform the MRT at high SNR regions where it is dominated by intra-cell interference. Therefore, we introduce the DRLbased method to choose an appropriate approach according to the dynamic wireless communication environment, which can be viewed as a typical decision-making problem. On the other hand, the DRL-based approaches, e.g., deep Q-learning[17], have been introduced to solve the power allocation task. However, the conventional deep Q-learning approach can only output discrete power levels, which may make training intractable with the increase of the action dimension. This motivates us to apply the deep deterministic policy gradient (DDPG) approach, which will be introduced in the following section, to tackle the challenging beam direction and power allocation tasks for each BS.

4 Proposed Limited Information Exchange Protocol

In principle, all the BSs share information through the backhaul links between BSs. However, it is an unaffordable burden for the backhaul links to transmit the global CSI among all BSs, especially when the BSs are equipped with large-scale antennas. Therefore, we develop a limited information exchange protocol, in which BSs only need to share a small amount of equivalent channel gain and interference information rather than the global CSI.

Assuming a flat and block-fading downlink channel, we propose a framework for the downlink data transmission process as shown in Fig. 1. In this framework, the channels are invariant during one time slot. Each time slot is divided into two phases. The first phase is a preparation phase for the BSs to collect local information, information exchange and beamforming design. The second phase is the downlink data transmission phase. Conventionally, the BSs only estimate downlink channels in the local information collection phase. To be specific, the BSs send reference symbols to UE first, then the UE estimates the downlink channel according to the reference symbols, and finally give the local CSI back to the corresponding BSs.

In our proposed protocol, in the local information collection phase, the UE needs to give back not only the local CSI but also the received interference from the other BSs. Let us take the UE k in cell n as an example. The BSs need to send orthogonal reference symbols to UE k, so that UE k can estimate the downlink local CSI $\boldsymbol{h}_{n,n,k}(t)$ and the received interference $c_{i,n,k}(t-1) = \sum_{j=1}^{K} |\boldsymbol{h}^{\mathrm{T}}_{i,n,k}(t)\boldsymbol{w}_{i,j}(t-1)|^2, \forall i \in N, i \neq n$ from the other BSs before the BSs update their beamformers.



▲ **Figure 1. Framework of the designed downlink data transmission process**

During the information exchange phase, the BSs first calculate the equivalent channel gain of each UE based on the local CSI and the previous beamformer. Meanwhile, the achievable rate of each UE can also be obtained according to Eq. (2b). Then the BSs concatenate the equivalent channel gains, the achievable sum rate of the served UE, and the interference information together, and then send them to the other BSs. Fig. 2 shows the information exchange process of the BS n. In cell n, the BS n collects the feedback information from all UE and calculates the equivalent channel gain of UE as $\bar{\boldsymbol{h}}_{n,k}(t) = |\boldsymbol{h}^{\mathrm{T}}_{n,n,k}(t)\boldsymbol{w}_{n,k}(t-1)|$. Besides, the BS n computes the achievable rate of each UE according to Eq. (2b) and obtains the sum rate $R_n(t-1)$ of UE in the cell n at time slot t-1. Then the BS n concatenates these information as the set $\left\{\boldsymbol{C}_n(t-1), \bar{\boldsymbol{h}}_n(t), R_n(t-1)\right\}$, where $\boldsymbol{C}_n(t-1) \in \mathbb{R}^{K \times (N-1)}$ is a matrix formed by concatenating the interference vectors of all UE in cell n and $\bar{\boldsymbol{h}}_n \in \mathbb{R}^K$ is a column vector composed of the equivalent channel of all UE in cell n. Note that in this information protocol, the BSs do not need to share the global or cross-talk CSI and the amount of exchanged information is only related to the number of cells and UE. Although other information exchange protocols try to further cut down the cost of information shared by the only exchange between adjacent cells[7–8], the sum-rate performance cannot be guaranteed when the interference generated by nonadjacent cells becomes nonnegligible. Therefore, we design the BSs to exchange information with all the other BSs.

5 MADDPG-Based Approach for Distributed Multi-Cell Multi-User MISO Beamforming

In this section, we introduce a DTDE MADDPG-based scheme for the MISO-IC system, as illustrated in Fig. 3, where each BS acts as a trainable agent. In the following, we take BS n as an example to elaborate on the online deci-

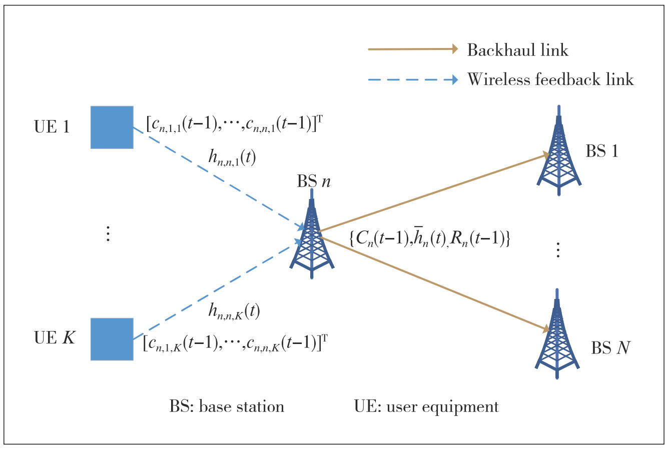

Backhaul link
Wireless feedback link
UE 1
$[c_{n,1,1}(t-1),\cdots,c_{n,n,1}(t-1)]^{\mathrm{T}}$
$h_{n,n,1}(t)$
BS n
BS 1
$\{C_n(t-1),\bar{h}_n(t),R_n(t-1)\}$
⋮
⋮
$h_{n,n,K}(t)$
UE K
$[c_{n,1,K}(t-1),\cdots,c_{n,n,K}(t-1)]^{\mathrm{T}}$
BS N
BS: base station
UE: user equipment

Figure 2. Illustration of information exchange process for the BS n

sion and offline training processes in detail.

5.1 Online Decision Process

In the online decision process, BS n observes the states from the wireless communication environment and takes actions based on the online policy network.

At the time slot t, the BS n observes the wireless communication environment and collects the state vector $\boldsymbol{s}_n(t)$. To be specific, during the online decision process in the DRL method at the time slot t, BSs firstly collect the information from UE and exchange information with each other according to the proposed information exchange protocol. With the received information from other BSs, the BS n can form the state vector $\boldsymbol{s}_n(t)$. The action $\boldsymbol{a}_n(t)$ is taken by the online policy network based on the observed state. Note that all the distributed agents take actions simultaneously, which means that none of them has instantaneous information about other BSs. To make the DDPG fully explore the action space, the output of the online policy network is added with action noise $\boldsymbol{n}_a \in \mathcal{N}\left(0,\sigma_a^2\right)$. With the training process moving on, the action noise decreases to zero gradually. With the action vector $\boldsymbol{a}_n(t)$ decided, the beamformers $\{\boldsymbol{w}_{n,k}\}$ can be formed and utilized for downlink data transmission. The reward $r(t)$ and the next state vector $\boldsymbol{s}_n(t+1)$ can be obtained in the next time slot $t+1$ through the proposed information exchange protocol. Meanwhile, the transition $\{\boldsymbol{s}_n(t),\ \boldsymbol{a}_n(t),\ r(t),\ \boldsymbol{s}_n(t+1)\}$ is stored in the memory replay buffer. The action vector $\boldsymbol{a}_n(t)$ is designed as

$$\boldsymbol{a}_n(t) = [\,\boldsymbol{p}_n^{\mathrm{T}}(t), P_{n,\mathrm{sum}}(t), D_n(t)\,]^{\mathrm{T}}. \tag{10}$$

In the action vector $\boldsymbol{a}_n(t)$, $\boldsymbol{p}_n \in \mathbb{R}^K$, $\sum_{k=1}^{K} p_{n,k} = 1$ denotes the normalized allocated power levels for UE and $P_{n,\,\mathrm{sum}} \in (0,1]$ denotes the normalized total transmit power of the cell n. Then the real transmit power for user k can be expressed as $P_{n,k}(t) = P_{\max} P_{\mathrm{n,sum}}(t) p_{n,k}(t)$. $D_n \in \{0,1\}$ is a Boolean value that denotes the selected beam direction solution in Eq. (8). When D_n=0, the BS n chooses ZF as the beam direction solution; when $D_n = 1$, the BS n chooses MRT. With the selected beam direction $D_n(t)$ and power strategy $P_{n,k}(t)$, the beamformer for UE k becomes

$$\boldsymbol{w}_{n,k}(t) = \sqrt{P_{n,k}(t)}\,\bar{\boldsymbol{w}}_{n,k}^{D_n(t)}(t). \tag{11}$$

The state vector $\boldsymbol{s}_n(t)$ is given by

$$\boldsymbol{s}_n(t) = \left[\mathrm{vec}\left(\overline{\boldsymbol{H}}(t)\right), \mathrm{vec}\left(\overline{\boldsymbol{C}}_n(t-1)\right)\right], \tag{12}$$

where $\overline{\boldsymbol{H}}(t) \in \mathbb{R}^{N \times K}$ denotes the equivalent channel gain of all UE, $\overline{\boldsymbol{H}}(t)_{[i,j]} = \left|\boldsymbol{h}_{i,i,j}^{\mathrm{T}}(t)\boldsymbol{w}_{i,j}(t-1)\right|$; $\overline{\boldsymbol{C}}_n(t-1) \in \mathbb{R}^{(N-1)\times K}$ denotes generated interference from cell n to the UE in the other cells, and $\overline{\boldsymbol{C}}_n(t-1)_{[i,j]} = \sum_{k=1}^{K}\left|\boldsymbol{h}_{n,i,j}^{\mathrm{T}}(t-1)\boldsymbol{w}_{n,k}(t-1)\right|$. For convenience, we use vec(·) to convert the matrix into a row vector by concatenating its rows. Note that all the elements in the state vector $\boldsymbol{s}_n(t)$ can be obtained through the above proposed information exchange protocol . The equivalent channel gain $\overline{\boldsymbol{H}}(t)$ contains the knowledge of receiving signal power of all UE and the generated interference $\overline{\boldsymbol{C}}_n(t-1)$ can lead the agent n to adjust actions to reduce the inter-cell interference to other cells. Since our goal is to maximize the achievable

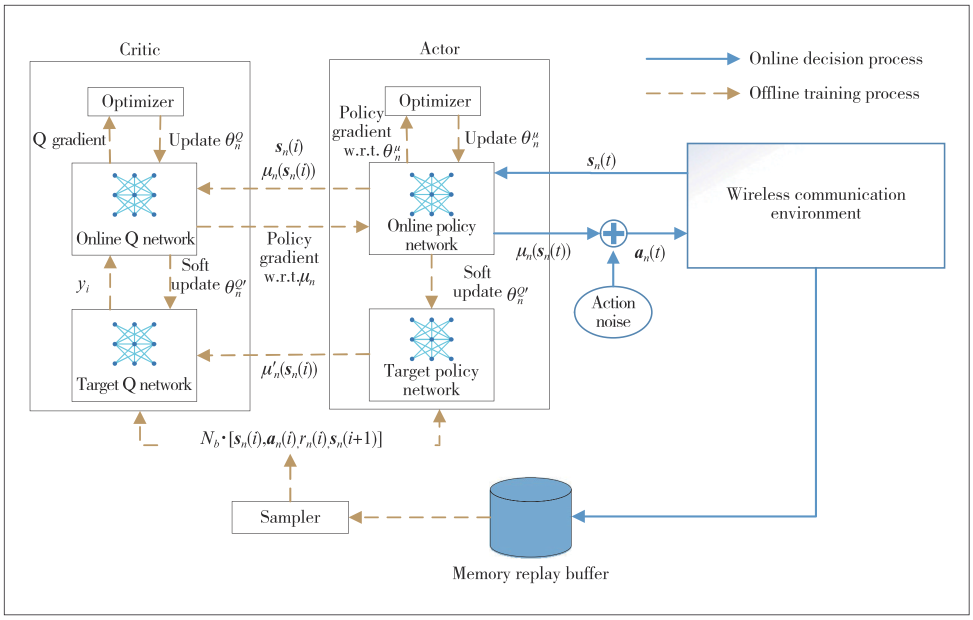


Figure 3. Illustration of the MADDPG-based scheme for multi-cell multi-user multiple-input single-output interference channel (MISO-IC) system

sum rate, we thus set the reward $r(t)=R_{\text{sum}}(t-1)$, which can be calculated based on the shared local information according to the limited information exchange protocol in Section 4.

5.2 Offline Training Process

In the offline training process, the sampler first randomly samples a batch of transition data $\{\boldsymbol{s}_n(i), \boldsymbol{a}_n(i), r(i), \boldsymbol{s}_n(i+1)\}$ from the memory replay buffer for training. By inputting the training transition i into the two target networks, the output of the target Q-network y_i can be expressed as:

$$y_i = r(i) + \eta Q'_n\left(\mathbf{s}_n(i+1), \mu'_n\left(\mathbf{s}_n(i+1) | \theta'_n\right) | \theta_n^{Q'}\right), \tag{13}$$

where η denotes the discount factor, and $\theta_n^{\mu'}$ and $\theta_n^{Q'}$ represent the network parameters of the target policy network μ' and Q-network Q', respectively. The Q-value is defined as the expectation of the future reward that can be obtained from the given state-action pair $\{\mathbf{s}_n(i), \mathbf{a}_n(i)\}$ when applying the strategy μ[18–19]. The Bellman equation of the Q-value can be expressed as:

$$Q^{\mu}(\mathbf{s}(i), \mathbf{a}(i)) = \mathbb{E}\left[r(i) + \eta Q^{\mu}(\mathbf{s}(i+1), \mathbf{a}(i+1))\right], \tag{14}$$

where the Q-value of the state-action pair$\{\mathbf{s}_n(i), \mathbf{a}_n(i)\}$ is composed of an instantaneous reward $r(i)$ and the Q-value of the next state-action pair$\{\mathbf{s}_n(i+1), \mathbf{a}_n(i+1)\}$. Note that the output of the target Q-network y_i is actually the estimated Q-value of the state-action pair $\{\mathbf{s}_n(i), \mathbf{a}_n(i)\}$.

According to the deterministic policy gradient theorem[19], the gradients of the online Q-network and policy network are:

$$\nabla_{\theta_n^Q} = \frac{1}{N_b} \frac{\left[\partial \sum_{i=1}^{N_b} \left(y_i - Q_n\left(\mathbf{s}_n(i), \boldsymbol{a}_n(i) | \theta_n^Q\right)\right)^2\right]}{\partial \theta_n^Q}, \tag{15a}$$

$$\nabla_{\theta_n^{\mu}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \left[\nabla_a Q_n\left(\mathbf{s}_n(i), \boldsymbol{a}_n(i)\right) \nabla_{\theta} \mu_{\theta}\left(\mathbf{s}_n(i)\right)\right], \tag{15b}$$

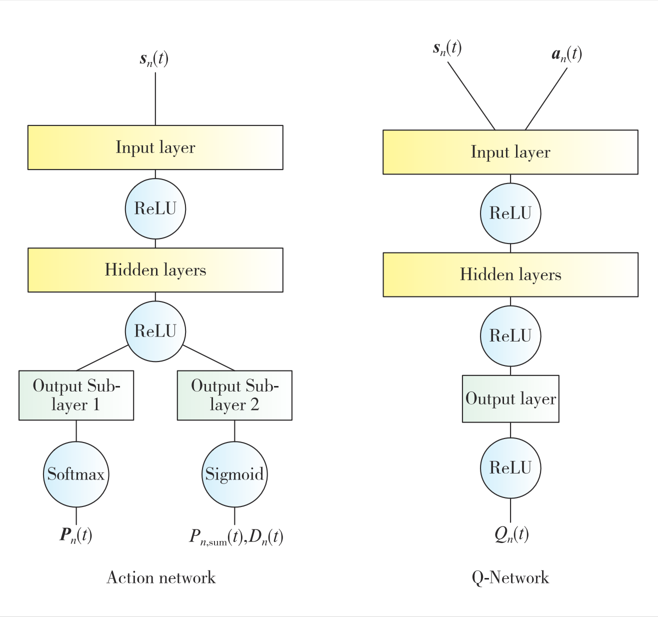
where N_b is the batch size of the sampled training data. The parameters in the online networks are updated by the optimizer. For the target networks, the parameters are softly updated as

$$\theta_n^{Q'} \leftarrow \tau \theta_n^Q + (1-\tau)\theta_n^{Q'}, \tag{16a}$$

$$\theta_n^{\mu'} \leftarrow \tau \theta_n^{\mu} + (1-\tau)\theta_n^{\mu'}, \tag{16b}$$

where $\tau \in (0,1)$ is the soft update factor.

The basic structures of the Q and policy networks, as shown in Fig. 4, are similar, which both include the fully connected input layer, hidden layers and the output layer. To reduce the computational complexity, we design two hidden layers for the Q and policy networks. The number of neurons in the input layer is the same as the dimension of the input vector. Hence, the scale of the input layer in the policy network is the same as the length of the input vector $\boldsymbol{s}_n(t)$. On one hand, the input vector for the Q-network is the concatenating of $\boldsymbol{s}_n(t)$ and $\mu_n(t)$. We apply ReLU as the activation function due to its simplicity. Note that the output layer of the policy network consists of two sub-layers that apply the softmax and sigmoid activation function for $\boldsymbol{p}_n(t)$ and $[P_{\text{n,sum}}(t), D_n(t)]$, respectively. On the other hand, the output of the Q-network is a real value denoting the Q-value of the corresponding state-action pair.



▲ **Figure 4. Structures of the action and Q-networks**

For clarification, we summarize the overall procedure for distributed multi-cell multi-user beamforming in Algorithm 1, referred to as the MADDPG algorithm. The proposed MADDPG algorithm requires the perfect CSI resulting from the proposed information exchange protocol. However, imperfect CSI may lead to the shifting of the objective function in Eq. (7), resulting in significant performance degradation of the current approach. A possible solution to the imperfect CSI is redesigning the reward function based on appropriate historical information, which will be addressed in our future work.

Algorithm 1: MADDPG Algorithm

```
Randomly initialize the weights of critic $\theta^Q$ and actor $\theta^{\mu}$ for all agents.
Initialize the weights of target networks as $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^{\mu}$ for all agents.
Initialize replay memory buffer for all agents.
repeat
   Agent $n$ observes the state $\boldsymbol{s}_n(t)$ in time slot $t$, $\forall n \in \boldsymbol{N}$.
   Agent $n$ selects an action $\boldsymbol{a}_n(t) = \mu\left(\boldsymbol{s}_n(t) | \theta^{\mu}\right) + \boldsymbol{n}_a$ accord-
```

```
ing to the current policy network output and exploration noise, $\forall n \in \boldsymbol{N}$.
7: Agent $n$ takes an action $\boldsymbol{a}_n(t)$, obtains a reward $r(t)$ and observe a new state $\boldsymbol{s}_n(t+1), \forall n \in \boldsymbol{N}$.
8: Agent $n$ stores the new transition $\{\boldsymbol{s}_n(t), \boldsymbol{a}_n(t), r(t), \boldsymbol{s}_n(t+1)\}$ into memory buffer, $\forall n \in \boldsymbol{N}$.
9: Agent $n$ samples a random batch of $N_b$ transitions $\{\boldsymbol{s}_n(i), \boldsymbol{a}_n(i), r(i), \boldsymbol{s}_n(i+1)\}$ from memory buffer, $\forall n \in \boldsymbol{N}$.
10: Agent $n$ calculates $y_i$ according to Eq. (12), $\forall n \in \boldsymbol{N}$.
11: Agent $n$ updates the online critic network $\theta_n^Q$ according to Eq. (14a), $\forall n \in \boldsymbol{N}$.
12: Agent $n$ updates online actor network $\theta_n^\mu$ according to Eq. (14b), $\forall n \in \boldsymbol{N}$.
13: Agent $n$ updates target networks according to Eqs. (15a) and (15b), respectively, $\forall n \in \boldsymbol{N}$.
14: until a termination criterion is reached
15: End
```

5.3 Information Exchange Analysis

We list the required information and the information exchange of different schemes in Table 1. Note that the fractional programming[3] (FP) and the zero-gradient[6] (ZG) need to exchange much more instantaneous CSI than that of MADDPG while the MADDPG only needs to exchange real values of previous information of the wireless environment. Moreover, thanks to the local CSI beam direction design, our proposed MADDPG based scheme does not rely on the number of antennas M and requires much less information exchange than those of FP and ZG, and is therefore suitable for the case of a large number of antennas.

5.4 Computational Complexity Analysis

The computational complexity of the proposed MADDPG algorithm mainly comes from the network computation and the beamformer formation. In Algorithm 1, the agent n firstly initializes the weights of the networks. We denote the number of hidden layers as L_H and the number of neurons in each hidden layer as N_H, and the complexity of the network initialization is $\mathcal{O}(L_H \mathrm{N}_H)$. In the repeat steps, we assume the number of repetitions as N_r, the complexity of Step 6 can be expressed as $\mathcal{O}(NKH + L_H N_H^2)$, which consists of the linear multiplication in hidden layers. In Step 7, the agent n needs to compute the normalized beamformers according to Eqs. (9a) and (9b). The complexity of ZF and MRT is $\mathcal{O}(M^3)$ and $\mathcal{O}(MK)$ and ZF has higher complexity due to matrix inversion operations. Then in Step 10, the target networks need to calculate y_i for each sample I and the complexity of Step 10 is $\mathcal{O}(N_b NKH + N_b L_H N_H^2)$. For the parameter update in Steps 11 – 13, the complexity is also $\mathcal{O}(N_b NKH + N_b L_H N_H^2)$ according to the error back propagation algorithm. Hence, the total computational complexity of the whole MADDPG algorithm, including the online decision and offline training processes, is given as $\mathcal{O}(N_r(N_b NKH + N_b L_H N_H^2 + M^3))$.

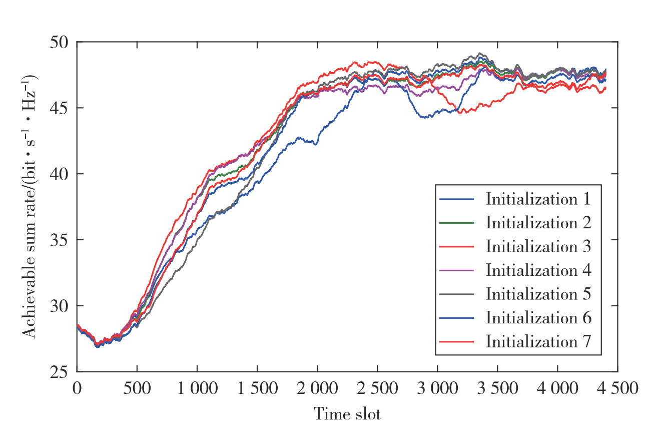
5.5 Convergence Discussion

The convergence behavior of the DRL algorithm, including the proposed MADDPG algorithm, depends on many factors such as the dynamics of the environment, the power of action noise σ_a^2 and the size of the memory replay buffer N_b. Up to now, the theoretical convergence analysis of the the DRL based algorithm is still an open problem and is generally based on empirical attempt. For example, when the action noise σ_a^2 is small, the MADDPG algorithm can converge faster. However, the performance of the MADDPG algorithm will degrade since the agents cannot explore the whole action space. Therefore, we need a large number of simulations to choose appropriate network hyper-parameters for achieving fast convergence and good performance.

To test the convergence behavior of the proposed MADDPG approach, we give an experimental result in Fig. 5, which illustrates the achievable sum rate versus the time slot under different initializations of network weights. The simulation settings are the same as that of Fig. 6 in Section 6. There are 7 simulation curves with different initial network weights in the same environment and all weights are randomly initialized following the standard Gaussian distribution. The simulation result shows that the different network initialization will basically converge to a similar performance around 4 000 time slots. This indicates that the proposed MADDPG method is insensitive to different network initialization.

6 Simulation Results

This section conducts numerical experiments to corroborate the performance of the proposed MADDPG algorithm in a wireless cellular system with (N, K)=(19, 4) and (M_x, M_y)=(8, 4). The distance between the centers of each hexagonal cell is 500 m

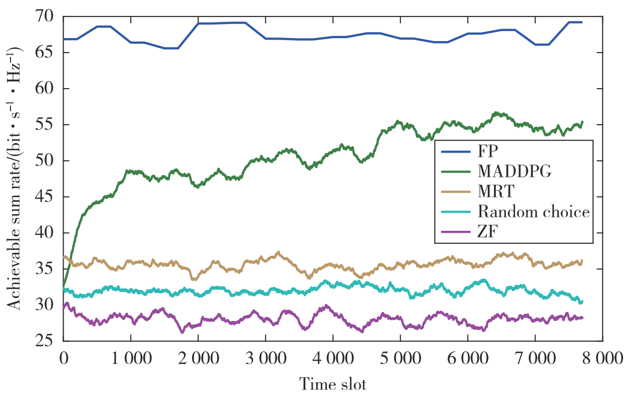


▲**Figure 5. Convergence behavior of the proposed multi-agent deep deterministic policy gradient (MADDPG) algorithm under different initialization of network weights**

▼Table 1. Comparison of the information exchange

Schemes	Required Information	Information Exchange
MADDPG	$\overline{\mathbf{H}}(t),\overline{\mathbf{C}}_n(t-1),R(t-1)$	$\mathcal{O}(NK)$
FP[3]	$\boldsymbol{h}_{i,j,k}(t),\forall i,j,k$	$\mathcal{O}(MNK)$
ZG[6]	$\boldsymbol{h}_{i,j,k}(t),\forall j,k$ for the BS i	$\mathcal{O}(MNK)$
MRT/ZF[5]	$\boldsymbol{h}_{i,i,k}(t),\forall k$	0

FP: fractional programming MADDPG: multi-agent deep deterministic policy gradient
MRT: maximum ratio transmission ZF: zero-forcing
ZG: zero-gradient



FP: fractional programming
MADDPG: multi-agent deep deterministic policy gradient
MRT: maximum ratio transmission
ZF: zero-forcing

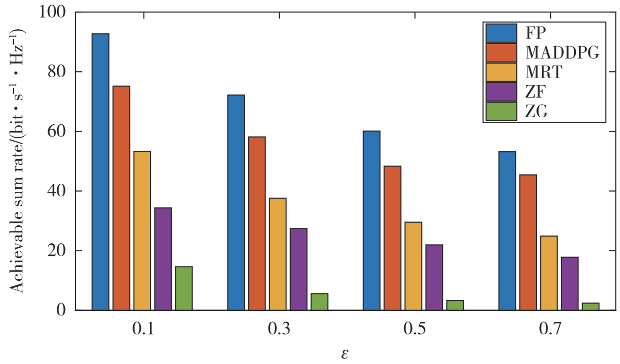
▲Figure 6. Average achievable rate of various schemes versus the number of time slots, where each point is a moving average over the previous 300-time slots with the UE distribution factor ϵ= 0.3

and the radius of each cell is r_{cell}=290 m. The first cell is located at the center, cells 2 – 7 are located at the first tier, and cells 8 – 19 are located in the second tier. We only count the achievable rates of cells 1 – 7 since the cells at the second tier suffer no interference from the outer cells. We define an inner region with radius r_{inner} where active UE does not exist. We define the UE distribution factor as $\epsilon = r_{\text{inner}}/r_{\text{cell}}$, where the factor ϵ determines how far the UE is from the BS. The received noise power σ_n^2 is set to 10^{-4} mW. The carrier frequency is f_c = 3.5 GHz. The large-scale fading factor is set as κ = 28+22 lg dis$_{3\text{D}}$+20 lg f_c/dB, where dis$_{3\text{D}}$ is the 3D distance between the UE and the BS. The total number of multipath L is 5 and the angular perturbation of each path Δ is 5°. The maximum transmit power of the BSs $P_{\max}$ and the time correlation coefficient ρ are set to 10^5 mW and 0.8, respectively. The action noise is initialized as $\sigma_a^2 = 1$.

The MADDPG scheme is deployed with the PyTorch framework and the hyper-parameters are set as follows. Both policy and Q-network parameters are designed as $L_H = 2$ fully-connected hidden layers with N_H = 200 neurons. The discount factor η is set to 0.6 and the soft update factor τ is equal to 0.01. The size of the memory replay buffer is set to 2 000 and the size of the sampled batch N_b is set to 64. Furthermore, we choose the Adam optimizer to update parameters with the learning rate being 10^{-3}.

Fig. 6 depicts the average achievable rate of various schemes versus the number of time slots. A random choice scheme means that each agent takes a random action in each time slot. For the ZF and MRT[5], the power allocation strategy is $P_{n,k} = P_{\max}\left\|\boldsymbol{h}_{n,k}\right\|^2 / \sum_i \left\|\boldsymbol{h}_{n,i}\right\|^2$. ZF and MRT exhibit the worst performance due to no extra information of the whole wireless environment. We see that the MADDPG based scheme learns to gradually improve the achievable sum rate of the system with the training process, as each agent updates the policy network to learn a better action policy for the system sum-rate maximization. The MADDPG converges to a fairly stable situation in 7 000 time slots and the variance is reasonable since the channels are dynamic. The MADDPG can achieve approximately 85% of the sum-rate performance of the FP algorithm, by using local CSI and limited information exchange only. It is worth noting that although the centralized FP algorithm has the largest achievable sum rate, it has a very high computational cost so we have to simulate the FP scheme every 500 time slots. Besides, FP[3] needs a large amount of instantaneous global CSI, which is unattainable in practical systems.

In Fig. 7, we evaluate the average achievable rate of different schemes vs the UE distribution factor. As the UE distribution factor increases, the average received power of UE can be reduced and the inter-cell interference problem becomes worse. The ZG algorithm, which is derived under high SINR assumption, has the worst performance under the 19 cells scenarios. The FP algorithm with global instantaneous CSI has undoubtedly the best performance in all scenarios. While as the users are getting closer to the cell edge, the performance gap between the FP and our proposed MADDPG is shrinking.



FP: fractional programming
MADDPG: multi-agent deep deterministic policy gradient
MRT: maximum ratio transmission
ZF: zero-forcing
ZG: zero-gradient

▲Figure 7. Average achievable rate of various schemes versus the UE distribution factor ϵ

7 Conclusions

In this paper, we reflect on the instantaneous sum rate maximization problem in the multi-cell MISO interference channel scenario. We propose a MADDPG scheme, in which each BS learns to choose an appropriate beam direction solution and allocate power based on the local CSI and limited exchange information among the BSs. The simulation results show that the proposed MADDPG scheme can achieve a relatively high sum rate with much less information exchange than the conventional centralized and distributed solutions.

References

[1] SOMEKH O, SIMEONE O, BAR-NESS Y, et al. Cooperative multicell zero-forcing beamforming in cellular downlink channels [J]. IEEE transactions on information theory, 2009, 55(7): 3206 – 3219. DOI: 10.1109/TIT.2009.2021371

[2] HUANG Y M, ZHENG G, BENGTSSON M, et al. Distributed multicell beamforming with limited intercell coordination [J]. IEEE transactions on signal processing, 2011, 59(2): 728 – 738. DOI: 10.1109/TSP.2010.2089621

[3] SHEN K M, YU W. Fractional programming for communication systems—part I: power control and beamforming [J]. IEEE transactions on signal processing, 2018, 66(10): 2616 – 2630. DOI: 10.1109/TSP.2018.2812733

[4] ZHANG R, CUI S G. Cooperative interference management with MISO beamforming [J]. IEEE transactions on signal processing, 2010, 58(10): 5450 – 5458. DOI: 10.1109/TSP.2010.2056685

[5] BJÖRNSON E, ZAKHOUR R, GESBERT D, et al. Cooperative multicell precoding: rate region characterization and distributed strategies with instantaneous and statistical CSI [J]. IEEE transactions on signal processing, 2010, 58 (8): 4298 – 4310. DOI: 10.1109/TSP.2010.2049996

[6] PARK S H, PARK H, LEE I. Distributed beamforming techniques for weighted sum-rate maximization in MISO interference channels [J]. IEEE communications letters, 2010, 14(12): 1131 – 1133. DOI: 10.1109/LCOMM.2010.12.101635

[7] GE J G, LIANG Y C, JOUNG J, et al. Deep reinforcement learning for distributed dynamic MISO downlink-beamforming coordination [J]. IEEE transactions on communications, 2020, 68(10): 6070 – 6085. DOI: 10.1109/TCOMM.2020.3004524

[8] KHAN A A, ADVE R S. Centralized and distributed deep reinforcement learning methods for downlink sum-rate optimization [J]. IEEE transactions on wireless communications, 2020, 19(12): 8410 – 8426. DOI: 10.1109/TWC.2020.3022705

[9] INDYK P, MOTWANI R. Approximate nearest neighbors: towards removing the curse of dimensionality [C]//The Thirtieth Annual ACM Symposium on Theory of Computing. STOC, 1998: 604 – 613

[10] YING D W, VOOK F W, THOMAS T A, et al. Kronecker product correlation model and limited feedback codebook design in a 3D channel model [C]//Proceedings of 2014 IEEE International Conference on Communications. IEEE, 2014: 5865 – 5870. DOI: 10.1109/ICC.2014.6884258

[11] DONG M, TONG L, SADLER B M. Optimal insertion of pilot symbols for transmissions over time-varying flat fading channels [J]. IEEE transactions on signal processing, 2004, 52(5): 1403 – 1418. DOI: 10.1109/TSP.2004.826182

[12] SCHUBERT M, BOCHE H. Solution of the multiuser downlink beamforming problem with individual SINR constraints [J]. IEEE transactions on vehicular technology, 2004, 53(1): 18 – 28. DOI: 10.1109/TVT.2003.819629

[13] CHRISTENSEN S S, AGARWAL R, DE CARVALHO E, et al. Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design [J]. IEEE transactions on wireless communications, 2008, 7(12): 4792 – 4799. DOI: 10.1109/T-WC.2008.070851

[14] JORSWIECK E A, LARSSON E G, DANEV D. Complete characterization of the Pareto boundary for the MISO interference channel [J]. IEEE transactions on signal processing, 2008, 56(10): 5292 – 5296. DOI: 10.1109/TSP.2008.928095

[15] LIM Y G, CHAE C B, CAIRE G. Performance analysis of massive MIMO for cell-boundary users [J]. IEEE transactions on wireless communications, 2015, 14(12): 6827 – 6842. DOI: 10.1109/TWC.2015.2460751

[16] MENG F, CHEN P, WU L N, et al. Power allocation in multi-user cellular networks: deep reinforcement learning approaches [J]. IEEE transactions on wireless communications, 2020, 19(10): 6255 – 6267. DOI: 10.1109/TWC.2020.3001736

[17] HESTER T, VECERIK M, PIETQUIN O, et al. Deep Q-learning from demonstrations [EB/OL]. [2022-02-02]. https://arxiv. org/abs/1704.03732. DOI: 10.1609/aaai.v32i1.11757

[18] DONG S K, CHEN J R, LIU Y, et al. Reinforcement learning from algorithm model to industry innovation : a foundation stone of future artificial intelligence [J]. ZTE communications, 2019, 17(3): 31 – 41. DOI: 10.12142/ZTECOM.201903006

[19] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms [C]//Proceeding of International Conference on Machine Learning. ICML, 2014: 387 – 395

Biographies

JIA Haonan received his BS and MS degrees in communication engineering from the University of Electronic Science and Technology of China, in 2019 and 2022, respectively. His research interests focus on deep learning with application to wireless communications.

HE Zhenqing (zhenqinghe@uestc.edu.cn) received his PhD degree in communication and information system from the University of Electronic Science and Technology of China (UESTC) in 2017. Since 2018, he has been with the National Key Laboratory of Science and Technology on Communications, UESTC, where he is currently an associate professor. His main research interests include statistical signal processing, wireless communications, and machine learning. He was a recipient of the IEEE Communications Society Heinrich Hertz Prize Paper Award in 2022.

TAN Wanlong received his BS degree in communication engineering from Jilin University, China in 2020. He is currently pursuing his MS degree in communication engineering with the University of Electronic Science and Technology of China. His research interests include wireless communications and reconfigurable intelligent surface.

RUI Hua received his BS, MS and PhD degrees from Nanjing University of Aeronautics and Astronautics, China in 1999, 2002, and 2005, respectively. He currently works as a senior pre-research expert and the head of the 6G Future Wireless Lab in ZTE Corporation. He has been engaged in wireless communication product and new technology pre-research, including 3G/4G/WIFI/5G/6G network architecture and key technologies. His main research direction is the 6G wireless communication technology. He has published more than 20 invention patents and papers in related fields. He has been engaged in more than 10 industry technical standards and white papers including 3GPP 3G/4G/5G series standards and IEEE 802.11 series standards.

LIN Wei received her BS and MS degrees in communication and information system from Northwestern Polytechnical University, China in 2002 and 2005 respectively. At present, she works in ZTE Corporation as a senior algorithm engineer in the Algorithm Department. Her research interests include 6G wireless communication physical layer technology and wireless AI technology. She has applied for more than 20 invention patents in related fields.