IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MOBILE MULTIMEDIA: METHODOLOGY AND APPLICATIONS

Received December 8, 2019, accepted December 23, 2019, date of publication January 1, 2020, date of current version January 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963468

A Machine Learning Approach for Beamforming in Ultra Dense Network Considering Selfish and Altruistic Strategy

CHANGYIN SUN, ZHAO SHI, AND FAN JIANG

Shaanxi Key Laboratory of Information Communication Network and Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Corresponding author: Fan Jiang (fjiangwbc@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61801382 and Grant 61871321, in part by the Key Project of Natural Science Foundation of Shaanxi Province under Project 2019JZ-06, and in part by the Key Industrial Chain Project of Shaanxi Province under Project 2019ZDLGY07-06.

ABSTRACT Coordinated beamforming is very efficient at managing interference in ultra dense network. However, the optimal strategy remains as a challenge task to obtain due to the coupled nature among densely and autonomously deployed cells. In this paper, the deep reinforcement learning is investigated for predicting coordinated beamforming strategy. Formulated as a sum-rate maximization problem, the optimal solution turns out as a balanced combination of selfish and altruistic beamforming. As the balancing coefficients depend on the beamforming vectors of all the cells, iterations are inevitable to get the final solution. To address this problem and improve efficiency, deep reinforcement learning (DL) is proposed to predict the balancing coefficients. Specifically, the agent, on behalf of a base station-user pair, will rely on Deep Q-network to learn the highly complex mapping between the balancing coefficients and signal-interference environment of each user. Subsequently, the beamforming vectors are obtained efficiently through the learned balancing coefficients. Due to the distinguished feature in exploration of the beamforming parameterization, the complexity problem brought by predicting the beamforming matrix directly is avoided. The performance of the proposed scheme is investigated by experiments with arguments regarding multiple input and multiple output configuration, shadow fading and state design. Simulation results indicate the facts that: 1) the theoretically infinite strategy space can be discretized with limited levels and granularity;2)it is feasible to approximate the complex mapping by Q-learning for wireless channel consisting both the large and small scale fading, 3) the balancing coefficients only concerns large scale fading, so the coordinated beamforming can be decomposed to two sub-problems with different time scales: parameterization at large time scales and instant beamforming based on balancing coefficients.

INDEX TERMS Deep learning, beamforming, ultra dense network (UDN), Q-learning, interference management.

I. INTRODUCTION

The ultra dense network (UDN) is the key enable technology for the future mobile communication system such as 5G and beyond to meet the exponential demand growth of data traffic [1] and mobile multimedia services [2], [3]. Within the architecture of UDN, a large amount of small cells are deployed autonomously around the served users, eventually, system performance in terms of capacity, coverage, and service efficiency can be improved due to the ever decreased distance between the transmitter and receiver. However, when cells density continues to increase, UDN will suffer from severe issues such as bad inter cells interference and complex mobility management problem. While the later problem can be circumvented via separation of control and data planes, interference management faces more challenges. To solve the problem, traditional means such as game theory, graph theory, and optimization method have been employed. Under the framework of optimization approach, coordination for UDN can be modeled as utilization maximum problem in

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

multi-cells, then scheme based on joint multi-dimensional resource allocation can be obtained by, for example, convex optimization in some case. Although coordination based on optimization such as beamforming or power control among cell cluster can be promising, the optimal solution is ordinarily difficult to get, because the denser and larger scale the UDN is, the more and stricter constraints are imposed on the network performances and resource usages of the optimum problem. As a compromise, some kind of sub-optimal strategy or iterative optimization approach is employed [1], [4].

Recently, deep learning (DL) has shown great potentials for improving the performance in communication system. At present, much attempts have been made to apply DL in areas of physical layer [5], resource allocation such as power control [6], [7], and beamforming [8]–[10]. For instance, [6] applied the full connected deep neural network (DNN) to approximate the weighted minimum mean square error (WMMSE) power allocation algorithm. Experiment results indicate that DNN can approximate WMMSE algorithm with high approximation accuracy and less computation complex. Reference [7] proposed to solve the transmit power control problem by using convolution neural network (CNN). The objective is to maximize spectrum efficiency (SE) and energy efficiency (EE). Simulation results show that the CNN-based power control method can achieve almost the same or even higher SE and EE than conventional power control scheme with much less computation time. Reference [8] dealt with transmitter beamforming based on an outage-based scheme, the proposed work attempts to cope with channel uncertainty at base station (BS), however, only simple scenarios assumptions of point-to-point and single group multicasting were adopted. In [9], a coordinated beamforming scheme based on neural network model for mm Wave BS was proposed. The DL is supplied with the OFDM omni-received sequences from the coordinated BSs in uplink to predict the RF beamforming codeword in downlink. To meet the real time applications requirements, [10] explores the beamforing structure of uplink-downlink duality, and relies on a CNN network to predict the virtual uplink power allocation, from which the final beamforming matrix can be obtained. In contract to approaches in [8]–[10] prevent the neural network from predicting directly the beamforming matrix, as a result, complexity is reduced.

Despite the progress made in DL based coordinate beamforming, it is noticed that such efforts mainly focus on approaches regarding only single cell scenario, but relatively speaking, approach for multi-cells is scarce in literature. In our opinion, scheme for multi-cells is more beneficial and practical for UDN. Motivated by this finding, we propose to use deep reinforcement learning to perform the multi-cells coordinated beamforming in this paper.

Traditionally, multi-cells beamforming can be in the form of coordination and cooperation. In the beamforming coordination such as Cooperative Beamforming(CB) in Coordinated Multi-Point downlink transmission(COMP) [14] in LTE-Advanced, the cells will share control information each other firstly, then apply classical beamforming algorithm such as maximal ratio combining (MRC) [14], zero forcing (ZF) [11], and block diagonalization (BD) [12], [13] to reduce inter cell interference. In the beamforming cooperation such as Joint Processing(JP) in COMP [14], the cells will share both control and data information each other, then beamforming algorithm such as virtual SINR (VSINR) in [15] can be employed. Moreover, aiming at maximizing different system objective, such as spectrum efficiency (SE) [16], energy efficiency (EE) [17], or content distribution efficiency [18], the multi-cells beamforming can be formulated as optimum problems while guaranteeing quality of experience (QoE). Based on the formulated model, optimal or sub-optimal solutions can be obtained by convex optimization or heuristic method. For 5G heterogeneous network with massive MIMO(multiple input and multiple output), new beamforming structure appeared such as hybrid beamforming (HBF) proposed in [19]. HBF involves beamforming at analog and digital stages concurrently in order to reduce the complexity and cost due to the large number of antennas.

Distinguished from the traditional or single cell approaches, we focus coordinated beamforming for multi cells UDN based on DL approach. Specifically, the beamforming is formulated as a system sum-rate maximum problem. As the problem is non-convex, it is solved by optimum condition with standard theory of Lagrange duality. The obtained optimum beamforming vectors for multi cells couple with each other, moreover, each vector depends on coefficients balancing the selfish and altruistic strategy. To refrain from iterations in calculating the coefficients, we propose to use DL to learn the balancing coefficients which parameterize the beamforming vectors. More precisely, the Deep Q-network is applied to approximate the mapping between the strategy equilibrium and observations of each user, including desired channel state information and interference levels towards other users. Accordingly, each serving cell, on behalf of its scheduled user, acts as an agent, and by deep learning, the agent will get the best strategy in an offline style. Based on the learned strategy and observed information, the agent will automatically search and react with a balanced beamforming strategy between selfish and altruistic. Moreover, as balancing coefficients only concern large scale channel fading, the corresponding beamforming can be decomposed to two sub-problems with different time scales: 1)large time scales parameterization at different levels for MIMO configurations; 2) instant beamforming based on balancing coefficients. Eventually, the proposed scheme avoids direct estimate of complex beamforming matrix.

The contributions of this work can be summarized as follows.

1) We propose to use deep reinforcement learning to determine the balancing coefficients of selfish and altruistic strategy in coordinated beamforming.
2) The mapping function between the observations of each user and the balancing coefficients which determine the beamforming can be learned in large time

scales by a Deep Q-network. Based on the learned mapping function, discretized coefficients are properly selected for instant beamforming according to the environments of base station and user pair.

3) The performance of the proposed scheme is simulated and evaluated by experiments with arguments regarding MIMO configuration, shadow fading and state design options. By simulation results, we find that the balance coefficients depend on the large scale fading of the channels, the simulation results also confirm the feasibility and effectiveness of the proposed method.

The paper is organized as follows. Section II introduces the system model. In Section III, the optimization problem and its resulting solution are presented. Section IV is devoted to the Deep Learning design, and Section V presents simulation results. Finally, Section VI concludes the paper.

Notation: The following notations are used: $||\mathbf{X}||_2$: 2-norm of $\mathbf{X}$, $\mathbf{X}^H$: Hermitian transpose of $\mathbf{X}$, $|\mathcal{X}|$: cardinal number of set $\mathcal{X}$, $\Pi_{\mathbf{X}}^{\perp} = \mathbf{I} - \mathbf{X}^H(\mathbf{X}\mathbf{X}^H)^{-1}\mathbf{X}$, $\Pi_{\mathbf{X}} = \mathbf{X}^H(\mathbf{X}\mathbf{X}^H)^{-1}\mathbf{X}$.

II. SYSTEM MODEL FOR BEAMFORMING IN UDN SYSTEM

We consider downlink UDN consisting of $\mathcal{N}$ cells, each with one base station(BS) and one served user(UE). For convenience of presentation, the same index is assumed for both the BS and its served user. Further, each BS has N_t antennas and each UE has N_r antennas, while for multiple input and single output (MISO) configuration, each UE has one antenna. We assume frequency reuse one in the system, so there will be severe inter cell interference across the cells. To combat inter cell interference, beamforming coordination across N_c cells in a cluster is sought to suppress the interference.

We denote by $\mathbf{H}_{ki} \in \mathbb{C}^{N_r \times N_t}$ the channel from BS i to UE k, both large and small scale fading components are included. The large fading includes path loss and shadowing. The small fading is modeled as independent identically distributed complex Gaussian random process with zero mean and unit variance. The signal to noise plus interference ratio of UE k with noise power of σ_k^2 is

$$\gamma_k = \frac{|\mathbf{v}_k^H \mathbf{H}_{kk} \mathbf{w}_k|^2 P}{\sum_{j \neq k}^{N_c} |\mathbf{v}_k^H \mathbf{H}_{kj} \mathbf{w}_j|^2 P + \sigma_k^2}, \tag{1}$$

where $\mathbf{w}_k \in \mathbb{C}^{N_t \times 1}$ is the BS transmitting beamforming vector and $\mathbf{v}_k \in \mathbb{C}^{N_r \times 1}$ is the UE receiving beamforming vector, P is the transmit power of BS.

Assume the goal is to maximize the sum-rate of the system with constraint $\|\mathbf{w}_k\|_2^2 = 1$, then the beamformimg vectors can be obtained by solving the optimization problem:

$$\begin{aligned} &\underset{\mathbf{V},\mathbf{W}}{\text{Max}} \sum_{k=1}^{N_c} R_k \\ &s.t. \ \|\mathbf{w}_k\|_2^2 = 1, \end{aligned} \tag{2}$$

where R_k is the achievable rata of user k

$$R_k = \log(1 + \gamma_k(\mathbf{V}, \mathbf{W})), \tag{3}$$

$\mathbf{V}$ and $\mathbf{W}$ are matrices of $\mathbf{v}_k$ and $\mathbf{w}_k$ respectively.

Problem (2) is well known to be NP hard, to circumvent the obstacle, we resort to approach harnessing standard theory of Lagrange duality, and get the final solution through characterizing the optimum condition of Lagrange function.

III. COORDINATE BEAMFORMING ALGORITHM BASED ON BALANCED STRATEGY

In this section, the beamforming algorithm based on the optimum condition with standard theory of Lagrange duality is first outlined. Then based on the structure of the obtained solution, different levels of parameterization are presented for MIMO, MISO, and special case of 2×1 MISO configurations respectively.

A. THE COORDINATE BEAMFORMING ALGORITHM

First, let define the Lagrangian function as follows:

$$L(\boldsymbol{\mu}, \mathbf{W}) = \sum_{k=1}^{N_c} R_k - \sum_{k=1}^{N_c} \mu_k (1 - ||\mathbf{w}_k||_2^2), \tag{4}$$

where μ_k is the Lagrangian multiply of the constraint for $\mathbf{w}_k$ in (2), $\boldsymbol{\mu}$ is the vector consisting of μ_k s.

Then, if we take the partial derivative of the Lagrangian function with respect to $\mathbf{w}_k$ first, then we can express the Karush-Kuhn-Tucker (KKT) condition $\frac{\partial}{\partial \mathbf{w}_k^H} L(\boldsymbol{\mu}, \mathbf{W}) = 0$ as follows [4],

$$\begin{aligned} &\frac{1/\ln 2}{\sum_{j=1}^{N_c} |\mathbf{v}_k^H \mathbf{H}_{kj} \mathbf{w}_j|^2 P + \sigma_k^2} \mathbf{H}_{kk}^H \mathbf{v}_k \mathbf{v}_k^H \mathbf{H}_{kk} \mathbf{w}_k P - \mu_k \mathbf{w}_k \\ &= \sum_{j \neq k}^{N_c} \frac{\gamma_j / \ln 2}{\sum_{j=1}^{N_c} |\mathbf{v}_j^H \mathbf{H}_{ji} \mathbf{w}_i|^2 P + \sigma_j^2} \mathbf{H}_{jk}^H \mathbf{v}_j \mathbf{v}_j^H \mathbf{H}_{jk} \mathbf{w}_k P. \end{aligned} \tag{5}$$

By rearranging the terms in (5), (5) can be further expressed as an eigenvector problem as follows.

$$(\mathbf{E}_k - \sum_{j \neq k}^{N_c} \lambda_{jk} \mathbf{B}_{jk}) \mathbf{w}_k = \mu_k \mathbf{w}_k, \tag{6}$$

where $\mathbf{E}_k = \omega_k^{-1} \mathbf{H}_{kk}^H \mathbf{v}_k \mathbf{v}_k^H \mathbf{H}_{kk}$, $\mathbf{B}_{jk} = \omega_k^{-1} \mathbf{H}_{jk}^H \mathbf{v}_j \mathbf{v}_j^H \mathbf{H}_{jk}$, and $\omega_k = \ln 2 (\sum_{j=1}^{N_c} \left|\mathbf{v}_k^H \mathbf{H}_{kj} \mathbf{w}_j\right|^2 + \sigma_k^2 / P)$.

The coefficient λ_{jk} in (6) is defined as:

$$\lambda_{jk} = \frac{z_j}{\sum_{i=1, i \neq j}^{N_c} L_j^i + \sigma_j^2} \frac{\omega_k}{\omega_j}, \tag{7}$$

where $z_j = |\mathbf{v}_j^H \mathbf{H}_{jj} \mathbf{w}_j|^2 P$ is the received signal of UE j from BS j, $L_j^i = |\mathbf{v}_j^H \mathbf{H}_{ji} \mathbf{w}_i|^2 P$ is the signal leakage from BSs i to UE j.

Once coefficients λ_{jk} are given, (6) is an eigenvector problem, the optimal beamforming vector $\mathbf{w}_k$ is obtained by

$$\mathbf{w}_k = \mathrm{V}_G^{Max}(\mathbf{E}_k - \sum_{j \neq k}^{N_c} \lambda_{jk} \mathbf{B}_{jk}) \tag{8}$$

where $\mathrm{V}_G^{Max}(\mathbf{X})$ is the eigenvector of matrix $\mathbf{X}$ with the largest eigenvalue.

TABLE 1. Algorithm1: Iterative beamforming for MIMO.

Algorithm1:Iterative Beamforming for solving (8)
1:Initialization phase: Initialize $\mathbf{w}_k, \forall k$;
2:**Repeat**
3:Determing $\mathbf{v}_k, \forall k$ by maximum SINR beam combiner in (29);
4:Caculate λ_{jk} with updated $\mathbf{E}_k$ and $\mathbf{B}_k$;
5:Solve $\mathbf{w}_k$ by (8) , $\forall k$;
6: Continue until converges of $\mathbf{w}_k$ or predetermined number of iterations is reached;
7:**End of Repeat**

B. PARAMETERIZATION OF THE BEAMFORMING FOR MIMO BY BALANCING COEFFICIENTS

It is noted from (8) that the beamforming vector $\mathbf{w}_k$ is determined by coefficients $\lambda_{jk}, j \neq k$. Given the determined coefficients, $\mathbf{w}_k$ can be obtained directly by eigenvalue computation. However, (7) indicates that λ_{jk} depends on both $\mathbf{w}_k$ of user k and $\mathbf{w}_j, j \neq k$ of other users. In other word, the optimal bamforming vectors $\mathbf{w}_k, \forall k$ are coupled each other in nature. To address this issue, $\mathbf{w}_k$ is updated *iteratively* by (8) with other $\mathbf{w}_j, j \neq k$ fixed till the convergent results are reached. The iteration based beamforming algorithm is summarized in Algorithm 1 in Table 1.

From (8), it is also noted that coefficient λ_{jk} plays a balancing role in the beamforming coordination strategy. Specifically, for case $\lambda_{jk} = 0$, the beamforming vector corresponds to the traditional maximum-ratio transmission (MRT) scheme, which selfishly increases the desired signal of user k with *none* consideration of interference to other users $j \neq k$. On the other side, for case $\lambda_{jk} = \infty$, the beamforming effort is devoted to decrease the interference to users in other cells. The later corresponds to the classical Zero Forcing Beamforming(ZF) scheme. So λ_{jk} in (8) will guarantee the system sum-rate by balancing the selfish and altruistic strategy.

C. PARAMETERIZATION OF THE BEAMFORMING FOR MISO

For MISO configuration which can be regarded as a simplified case of optimization problem in (2), the beam combiner $\mathbf{v}_k = 1, \forall k$ in this case. As a result, the optimal beamforming vector $\mathbf{w}_k$ in (6) can be rewritten as follows.

$$\begin{aligned}\mathbf{w}_k &= \mu_k^{-1}(\mathbf{E}_k - \sum\nolimits_{j=1,j\neq k}^{N_c} \lambda_{jk}\mathbf{B}_{jk})\mathbf{w}_k \\ &= (\mu_k\omega_k)^{-1}(\mathbf{h}_{kk}^H\mathbf{h}_{kk}\mathbf{w}_k - \sum\nolimits_{j=1,j\neq k}^{N_c} \lambda_{jk}\mathbf{h}_{jk}^H\mathbf{h}_{jk}\mathbf{w}_k) \\ &= \sum\nolimits_{j=1}^{N_c} \zeta_{jk}\mathbf{h}_{jk}^H, \end{aligned} \tag{9}$$

where $\zeta_{jk} = (\mu_k\omega_k)^{-1}\lambda_{jk}\mathbf{h}_{jk}\mathbf{w}_k$ $j \neq k$, and $\zeta_{kk} = (\mu_k\omega_k)^{-1}\mathbf{h}_{kk}\mathbf{w}_k$ are complex numbers. $\mathbf{h}_{jk} \in \mathbb{C}^{1\times N_t}$ is the channel vector from base station k to user j.

Equation (9) indicates that the optimal beamforming vector $\mathbf{w}_k$ in MISO configuration can be parameterized by complex numbers ζ_{jk}, that is, the weighted combination of $\mathbf{h}_{jk} \in \mathbb{C}^{1\times N_t}$ by ζ_{jk}. Actually, (9) has been pointed in [21] for parameterization of the Pareto boundary. The Pareto boundary [21] is the outer boundary of the achievable rata region $\mathcal{R}$ defined as

$$\mathcal{R} = \cup_{|\mathbf{w}_k|=1,\forall k}(R_1 \cdots, R_k, \cdots, R_{N_c}). \tag{10}$$

Based on the definition of the Pareto region, the outer boundary of $\mathcal{R}$ corresponds to optimum points where one user's rate cannot be increased without decreasing the rate of other users.

D. PARAMETERIZATION OF THE BEAMFORMING FOR 2×1 MISO

For special case of 2×1 MISO configuration, the coefficients

$$\begin{aligned}\lambda_{21} = &\frac{\mathbf{w}_1^H\mathbf{h}_{11}^H\mathbf{h}_{11}\mathbf{w}_1 P}{\mathbf{w}_2^H\mathbf{h}_{22}^H\mathbf{h}_{22}\mathbf{w}_2 P + \mathbf{w}_1^H\mathbf{h}_{21}^H\mathbf{h}_{21}\mathbf{w}_1 P + \sigma_2^2} \\ &\times \frac{\mathbf{w}_1^H\mathbf{h}_{11}^H\mathbf{h}_{11}\mathbf{w}_1 P + \mathbf{w}_2^H\mathbf{h}_{12}^H\mathbf{h}_{12}\mathbf{w}_2 P + \sigma_1^2}{\mathbf{w}_1^H\mathbf{h}_{21}^H\mathbf{h}_{21}\mathbf{w}_1 P + \sigma_2^2}, \end{aligned} \tag{11}$$

and

$$\begin{aligned}\lambda_{12} = &\frac{\mathbf{w}_2^H\mathbf{h}_{22}^H\mathbf{h}_{22}\mathbf{w}_2 P}{\mathbf{w}_1^H\mathbf{h}_{11}^H\mathbf{h}_{11}\mathbf{w}_1 P + \mathbf{w}_2^H\mathbf{h}_{12}^H\mathbf{h}_{12}\mathbf{w}_2 P + \sigma_1^2} \\ &\times \frac{\mathbf{w}_2^H\mathbf{h}_{22}^H\mathbf{h}_{22}\mathbf{w}_2 P + \mathbf{w}_1^H\mathbf{h}_{21}^H\mathbf{h}_{21}\mathbf{w}_1 P + \sigma_2^2}{\mathbf{w}_2^H\mathbf{h}_{12}^H\mathbf{h}_{12}\mathbf{w}_2 P + \sigma_1^2}. \end{aligned} \tag{12}$$

Considering that $\prod_{\mathbf{h}_{21}^H} + \prod_{\mathbf{h}_{21}^H}^{\perp} = \mathbf{I}$ and $\gamma\mathbf{h}_{21}^H = \prod_{\mathbf{h}_{21}^H}\mathbf{h}_{11}^H$, where γ is a complex number, the optimal beamforming vectors $\mathbf{w}_k, k = 1, 2$ can be expressed as

$$\mathbf{w}_k = \lambda_k\mathbf{w}_k^{MRT} + (1-\lambda_k)\mathbf{w}_k^{ZF}, \tag{13}$$

where $\mathbf{w}_k^{MRT}$ and $\mathbf{w}_k^{ZF}$ are the MRT and ZF beamforming vectors respectively. $\lambda_k, k = 1, 2$ are real value parameters in the range of $0 \leq \lambda_k \leq 1$.

Take $k = 1$ as an example,

$$\mathbf{w}_1^{MRT} = \frac{\mathbf{h}_{11}^H}{||\mathbf{h}_{11}||}, \tag{14}$$

and

$$\mathbf{w}_1^{ZF} = \frac{\Pi_{\mathbf{h}_{21}^H}^{\perp}\mathbf{h}_{11}^H}{||\Pi_{\mathbf{h}_{21}^H}^{\perp}\mathbf{h}_{11}^H||} \tag{15}$$

Equation (13) has been proved differently in [21] and [22] respectively. It indicates that the optimal beamforming vector is a balanced combination of MRT and ZF vectors in the beamforming vectors space for special case of a two users MISO system.

Now, if we look back and make an observation of (8) and (9), we will find that the beamforming vector obtained from (13) is at the highest level in term of parameterization with balancing coefficients, the parameterization level ascends from MIMO to two users MISO configuration.

Based on the observation made above, we conclude that the balancing coefficients are determinant of coordination strategy for beamforming algorithm. However, to get the balancing coefficients such as (7), iteration among different cells is inevitable, which will increase complexity in implement and introduce additional delay in signaling.

To remedy the issue brought by iteration, the beamforimg problem (8) is decomposed to two sub-problems with different time scales:

1) Determine balancing coefficients at large time scales;

2) Instant beamforming based on balancing coefficients.

In next section, deep reinforcement learning is introduced as an alternative approach to approximate the complex mapping function in (8) and (13).

IV. DEEP REINFORCEMENT LEARNING BASED BEAMFORMING

Reinforcement learning (RL) is widely used in machine learning area. Under the infrastructure of RL, the agent will interact with environment, and during this process, it will acquire the best action strategy through learning from the exploration and exploitation of the environment. In the sense of RL learning, both activities are based on the experiences of the agent, however, exploration are based on experience which the agent has never been come across previously in the space of state-action pairs, on the other hand, exploitation is based on the experiences so far.

In the literature, the RL can be classified to three categories: (1) critic only; (2) actor only; (3) critic and actor. Q-learning, which falls into the critic only category, is a widely used algorithm in RL for the agent to learn the best action strategy. The action strategy, also called policy, is a sequence of actions in each upcoming time instants of an episode:

$$\pi = \{a_t = \varphi_t(\mathrm{s}_t), t = 1, 2, \cdots\}, \tag{16}$$

where $\varphi_t(\mathrm{s}_t)$ is the mapping function from the state to action. In Q-learning, a Q-function is defined as $Q(s_t, a_t)$, which reflects the value of action a_t in state s_t. So an optimal Q-function $Q^*(s_t, a_t)$ means that the agent will get the maximum expected rewards when it takes the action a_t in state s_t following the optimal policy π^*. The objective of the Q-learning is to find the best policy π^* which achieves the maximum expected rewards. In this paper, the mapping function $\varphi_t(\mathrm{s}_t)$ in Q-learning is utilized to map the complex relationship in (8) and (13) between balancing coefficients and signal-interference environment.

For multi agents learning process, distributed or coordinated Q-learning can be assumed. In distributed Q-learning, each agent learns independently without sharing of policy information each other, and the target agent regards other agents as part of the environment. On the contrary, in coordinated Q-learning, portion of the policy information will be shared among agents taking part in the coordination, and the convergence time can be reduced by providing the learned policy to a new agent for initialization. However, extra signal overhead is required for coordinated Q-learning scheme. Moreover, states and actions among the agents should be updated with synchronization, otherwise, oscillation and instability will occur in the system. Based on this consideration and the additional observation in (9) that *only* partial observations of the channels $\mathbf{h}_{jk} \in \mathbb{C}^{1\times N_t}$ are required to calculate the optimal beamforming vector $\mathbf{w}_k$, in this paper, distributed Q-learning is assumed. Specifically, each *BS* in the cluster is considered as an agent, which will interact with the environment. The environment, on the other hand, corresponds to the UDN which excludes the target agent. For deep reinforcement learning framework formulation, the key elements are detailed in the following sub-sections.

A. STATE SPACE

Under the deep reinforcement learning framework, the state will characterize the environment the agent faces. The agent interacts with environment through action and reward based on the observed state. To parameterize the optimal beamforming vector, the state is defined differently for MIMO and MISO configuration, as each configuration possesses different parameterization level.

MIMO Configuration:

The state space $\mathcal{S}^k$ of the agent k is defined as:

$$S^k = \{\mathbf{z}, \mathbf{L}^k\}, \tag{17}$$

where $\mathbf{z}$ is the signal strength vector $\{z^k = \log(||\mathbf{H}_{kk}||_2), k = 1, \cdots, N_c\}$, $\mathbf{L}^k = \{log\left(||\mathbf{H}_{jk}||_2\right), j \neq k\}$ is the signal leakage vector of agent k towards other UEs. Since the channel amplitude varies across several orders in magnitude, the logarithmic form in the state definition is beneficial for quick convergence of the DQN network during training process.

MISO configuration:

The state space of each agent $\mathcal{S}^k$ is defined as:

$$\mathcal{S}^k = \{z^k, \mathbf{L}^k\}. \tag{18}$$

In (18), only signal strength of desired user k is considered compared with (17). $\mathbf{L}^k = \{\log\left(||\mathbf{h}_{jk}||_2\right), j \neq k\}$ is defined similarly as in MIMO configuration. The reason for this design comes from the observation of (9) that the beamforming vector of user k is the weighted combination of $\mathbf{h}_{jk} \in \mathbb{C}^{1\times N_t}, j = 1 \cdots N_c$. Based on this observation, the logarithmic form of channels from the BS k to all the users are taken as elements of the state space. On the contrary, for MIMO configuration, as balance coefficient in (7) depends on the receiver combiner vectors, so the signal strength of other users in the environment are also considered in (17) to reflect the effect of the receiver combiner. The rationale of the different design for different MIMO configuration relies on the fact that higher level of parameterization exists in beamforming structure in MISO, so less information is required in the state space to determine the best solution.

B. ACTION SET

At each instant t, the agent in state s_t will take an action a_t from the action set $\mathcal{A}$. In this paper, the action set $\mathcal{A}$ is a discretized balancing coefficients for the parameterization in the structure of the beamforming vector. As different MIMO configuration possesses different level of parameterization, different action space is defined accordingly.

MIMO configuration:

Theoretically, the action space for (7) seems infinite, ranging from $\lambda_{jk} = 0$ to $\lambda_{jk} = \infty$. So it is a challenge to discretize the large space with limited levels and granularity while trade off the performance loss against the complexity. To cope with this issue, we discretize the balancing coefficient based on its distribution probability by examining the statistics from experiment result in Section V. Based on the statistics, the action space is defined as:

$$a^k = \left\{ a_j^k, j \neq k \right\} \tag{19}$$

where

$$a_j^k = \{0, 1, \cdots, |\mathcal{A}| - 1\} . \tag{20}$$

Correspondingly, $\lambda_{jk} = 10^{\log(\lambda_{jk}^{\min}) + a_j^k \delta}$, and

$$\delta = (\log|\lambda_{jk}^{\max}| - \log|\lambda_{jk}^{\min}|)/|\mathcal{A}|. \tag{21}$$

$\lambda_{jk}^{\max}$ and $\lambda_{jk}^{\min}$ are the maximum and minimum value of the truncated distribution in λ_{jk}.

2×1 *MISO Configuration:*

For special case of 2×1 MISO configuration, the action space is a discretized balancing coefficients of $\lambda_k, k = 1, 2$ in (13). As the value of $\lambda_k, k = 1, 2$ is in the range of $0 \leq \lambda_k \leq 1, k = 1, 2$, even discretization based on a fixed granularity $\Delta = 1/35$ is assumed with 36 levels in total:

$$a^k = \{1, \cdots, 36, for\ \lambda_k = 0,\ \Delta, \cdots, 35\Delta\} \tag{22}$$

The selected action a_t by an agent in state s_t is based on a decision policy π, which will be learned by the agent from the reward fuction defined in the next sub section.

C. REWARD FUNCTION

The reward function will reflect the objective of the UDN system, the agent will receive a reward from environment as the degree of satisfaction to the action. Based on the objective in (2), the reward is defined as the system sum-rate, which is expressed as:

$$r = \sum\nolimits_{k=1}^{N_c} R_k. \tag{23}$$

With deep learning, the agent will get the maximum return if it follows the optimal decision policy based on the observation of the environment. In deep learning, the return is calculated as the expected cumulative discount rewards defined as $G_t = \mathrm{E}[\sum_{n=0}^{\infty} \beta^n r_{t+n}]$, where β is the discount rate, and $\mathrm{E}[\bullet]$ is the expectation.

To get the optimal decision policy, we adopt deep Q-learning scheme. The core concept of Q-learning is the Q-function $Q^t(s_t, a_t)$, which is defined as the ultimate expected action value the agent will receive if it follows a policy π depending on the state thereafter. The optimal decision policy follows the optimal Q-function actually. In practice, the optimal Q-function is obtained with Bellman equation by iterative updates as follows:

$$Q^{t+1}(s_t, a_t) = \mathrm{E}_{s_{t+1}}[r_t + \beta \max_{a_{t+1}} Q^t(s_{t+1}, a_{t+1}) | s_t, a_t]. \tag{24}$$

The remarkable features of Q-learning is that it is model free, which means that it can get the optimal Q-function by (24) without knowledge of transition probability from one state to others. It has been shown that the iterative updates are guaranteed to converge to the optimal Q-function when $t \to \infty$.

In this paper, the Q-function is utilized to approximate relations between wireless environment states and balancing coefficients in (8), and (13). Once the optimal Q-function is obtained, the agent will select the best λ_{jk} based on the wireless states for the instant coordinate beamforming across the cells.

D. DEEP Q-NETWORK

In case the number of discrete states and action spaces are small, Q-table can be used for the Q-function iterative updates. As the name suggests, Q-table is a table with dimensions of $|\mathcal{A}|$ rows and $|\mathcal{S}|$ columns. The content of the Q-table corresponds to the value of a specific state-action pairs. Once the Q-table is obtained from training stage, the optimal policy follows the output of the Q-table. But for the DL approach investigated in this paper, the dimension of action spaces is $N_c \times (N_c - 1) \times |\mathcal{A}|$ for (19), so the action space scales with N_c^2. Moreover, the state value in (17) is continues, apparently, Q-learning based on Q-table is not affordable, so we resort to Deep Q-network (DQN) in [23].

In DQN, deep neural network with weights $\{\theta\}$ is employed to approximate the Q-function, and the weights $\{\theta\}$ are obtained by training algorithm with data samples in the offline training stage. At the online testing stage, the trained DQN will output the value of Q-function given the state as the input. To update parameter weights $\{\theta\}$, the DQN use a mean-squared error as loss function, and the loss function is defined as:

$$Loss_t(\theta) = \sum\nolimits_{(s_t, a_t) \in D} (y - Q(s_t, a_t, \theta))2, \tag{25}$$

where D is the data sample set for training, y is the target value defined as

$$y = r_t + \max_{a_t \in A} Q_{old}(s_t, a_t, \theta^-), \tag{26}$$

$\{\theta^-\}$ are the parameter weights in previous iteration, Q_{old} is generated by target network for producing target value. With loss function, the parameter weights $\{\theta\}$ are updated as follows:

$$\theta_{t+1} = \theta_t + \alpha \nabla_{\theta_t} Loss_t \tag{27}$$

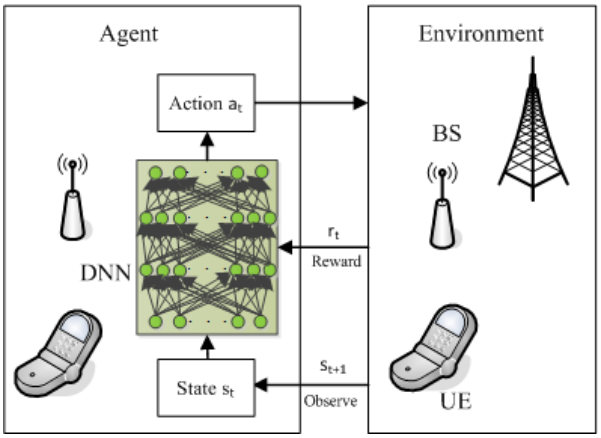

FIGURE 1. Principle of the proposed beamforing based on DQN.

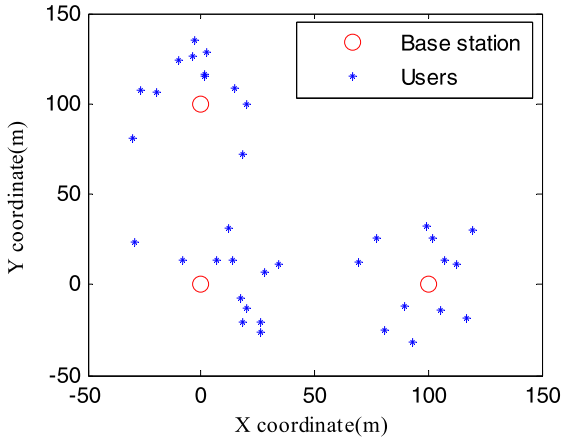

FIGURE 2. Distribution of base stations and users.

where $\nabla_{\theta_t} Loss_t$ is the loss function gradient with respect to parameter weights $\{\theta\}$,$\alpha$is the step size.

E. DEEP Q-NETWORK BASED BEAMFORMING

The Deep Q-network based beamforming is achieved by training the DQN agent to get the balancing coefficients with the experiences during the interactions with the environment. The principle is showed in Fig.1.

To train the agent, the emulator is firstly constructed which is a UDN with BSs and served UEs distributed under the coverage of their serving base station. Based on the distance between the BS and UE pair, the channel can be generated with the path loss model assumed. The training process will consist of fixed number of episodes. During each episode, the training proceeds with the following procedures: 1) one of the BS and UE pairs is selected randomly as agent, and other pairs in the UDN act as the environment; 2) the channel of each pair is generated independently; 3) a random action is selected for the agent from the action sets; 4) the beamforming vectors for users in the environment are initialized with classical algorithm such as ZF; 4) the beamforming vectors for agent is calculated based on the selected action; 5) the beamforming vectors for other users are calculated with the balancing coefficient in (7) for MIMO configuration; 6) the observation and reward are produced and fed to the DQN by the environment; 7) based on the observation and reward, the DQN determines the action according to the δ-greedy algorithm; 8) the DQN records the Quadruple $\mathbf{e}^t = \{S_t^k, a_t^k, r_t, S_{t+1}^k\}$ in the memory pool $\mathbf{D}$, and updates the DQN parameters with random samples from $\mathbf{D}$.

The proposed training algorithms are outlined in algoritm 2, algorithm 3, and algorithm 4 for MIMO, MISO and 2×1 MISO configuration respectively in the following tables.

MIMO Training
MISO training
2×1 *MIMO Training*

V. SIMULATION RESULTS

In this section, the simulation layout is first described, afterwards, the simulation results will be presented.

TABLE 2. Algorithm2: Training DQN for MIMO beamforming.

Algorithm2:Training DQN based beamforming for MIMO configuration
1:Initialization phase: Initialize $\{\theta\}$,the Quadruple $\mathbf{e}^t = \{\mathcal{S}_t^k, a_t^k, r_t, \mathcal{S}_{t+1}^k\}$,and the memory pool $\mathbf{D}$
2:**Repeat** for each episode
3: Initialize the UDN: a) random the position distribution of users; b)calculate path loss and shadowing for each BS-User pair;
4: Initialize $\mathbf{w}_k, \forall k$ with ZF algorithm;
5: Determining $\mathbf{v}_k, \forall k$ by maximum SINR beam combiner in (29);
6:Caculate state $\mathcal{S}_t^k$ for user k when it is the agent;
7: **Repeat** for each iteration in **algorithm 1**
8: **IF** user k is not the agent
9:Caculate λ_{jk} with updated $\mathbf{E}_k$ and $\mathbf{B}_k$;
10:**Else**
11: Feed $\mathcal{S}_t^k$ to DQN and determine Q-value with the parameter $\{\theta\}$, chose an action according to the δ-greedy algorithm.
12:**End of IF**
13:Solve $\mathbf{w}_k$ by(8);
14:Until converges of $\mathbf{w}_k$ or predetermined number of iterations is reached;
15:**End of Repeat** for iteration in **algorithm 1**
16:Update state $\mathcal{S}_{t+1}^k$ for user k who is agent;
17:Calculate the system sum-rate and the reward ;
18:Push $\mathbf{e}^t$ into the memory pool $\mathbf{D}$;
19:Random sample $\mathbf{D}$ and calculate $\{\theta\}$;
20: Update $\{\theta\}^-$ to $\{\theta\}$ at every target update instant.
21: **End of Repeat** for each episode

A. SIMULATION LAYOUT

A UDN network with cell radius of 50 meters is simulated. The cell number per cluster is 3. The layout of the base stations and users are shown in Fig.2. The transmit power of BS is 30dBm. For MIMO configuration, each BS has 4 antennas, and each UE has 2 antennas. For MISO case, one antenna for each user is assumed. The UE is dropped

TABLE 3. Algorithm3: Training DQN for MISO beamforming.

Algorithm3:Training DQN based beamforming for MISO configuration
1:Initialization phase: Initialize $\{\theta\}$,the Quadruple $\mathbf{e}^t=\{\mathcal{S}_t^k,a_t^k,r_t,\mathcal{S}_{t+1}^k\}$,and the memory pool $\mathbf{D}$
2:**Repeat** for each episode
3: Initialize the UDN: a) random position distribution of users; b)calculate path loss and shadowing for each BS-User pair;
4: Initialize $\mathbf{w}_k,\forall k$ with ZF algorithm;
5: **IF** user k is the agent, calculate state $\mathcal{S}_t^k$;
6: Feed $\mathcal{S}^k$ to DQN and determine Q-value with the parameter $\{\theta\}$, chose an action according to the δ-greedy algorithm.
7: **Else IF** user k is not the agent
8: **Repeat** for each iteration of **algorithm 1**
9:Caculate λ_{jk} with updated $\mathbf{E}_k$ and $\mathbf{B}_k$;
10:Solve $\mathbf{w}_k$ by(8);
11:Until converges of $\mathbf{w}_k$ or predetermined number of iterations is reached;
12:**End of Repeat** for iteration in **algorithm 1**
13: **End of IF**
14:Update state $\mathcal{S}_{t+1}^k$ for user k who is agent;
15:Calculate the system sum-rate and obtain the reward ;
16:Push $\mathbf{e}^t$ into the memory pool $\mathbf{D}$;
17:Random sample $\mathbf{D}$ and calculate $\{\theta\}$;
18: Update $\{\theta\}^-$ to $\{\theta\}$ at every target update instant.
19: **End of Repeat** for each episode

TABLE 4. Algorithm4: Training DQN for 2 × 1 MIMO beamforming.

Algorithm4:Training DQN based beamforming for 2×1 MISO configuration
1:Initialization phase: Initialize $\{\theta\}$,the Quadruple $\mathbf{e}^t=\{S_t^{\ k},a_t^k,r_t,S_{t+1}^{\ k}\}$,and the memory pool $\mathbf{D}$
2:**Repeat** for each episode
3: Initialize the UDN: a) random the position distribution of users; b)calculate path loss and shadowing for each BS-User pair;
4: Initialize $\mathbf{w}_k,\forall k$;
5: **IF** user k is agent, calculate state $S_t^{\ k}$;
6: Feed S^k to DQN and determine Q-value with the parameter $\{\theta\}$, chose an action according to the δ-greedy algorithm.
7: **Else IF** user k who is not the agent
9:Caculate λ_k with (31);
10: **End of IF**
11:Solve $\mathbf{w}_k\rightarrow\mathbf{w}_k=\lambda_k\mathbf{w}_k^{MRT}+(1-\lambda_k)\mathbf{w}_k^{ZF}$;
12:Update state $S_{t+1}^{\ k}$ for user k who is agent;
13:Calculate the system sum-rate and obtain the reward ;
14:Push $\mathbf{e}^t$ into the memory pool $\mathbf{D}$;
15:Random sample $\mathbf{D}$ and calculate $\{\theta\}$;
16: Update $\{\theta\}^-$ to $\{\theta\}$ at every target update instant.
17: **End of Repeat** for each episode

uniformly in each cell, and the path loss model is given by

$$PL = 140.7 + 36.7 \times \log_{10}(R) \tag{28}$$

in dB, where R is distance in km.

The DQN is a fully connected neural network with five layers. Three hidden layers are configured, corresponding to 100, 60, and 30 neurons respectively. The Relu activation function is adopted. To update parameters, adaptive moment estimation method (Adam) [24] is used, and the beginning learning rate is 0.01. The DQN was first trained in offline stage with observation received from the simulated UDN, then, in the deployment stage, the DQN generates the coordinate coefficient λ_{jk} for beamforming, and the system throughput is collected for comparison.

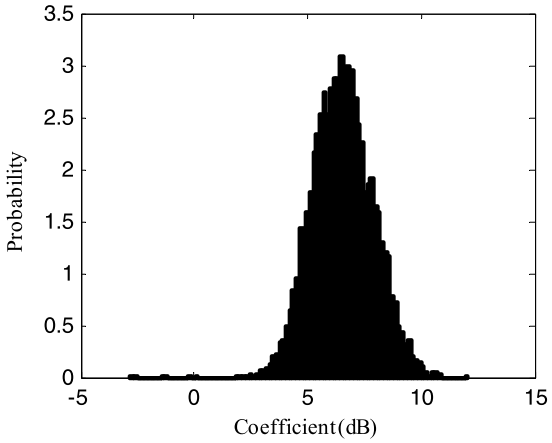

FIGURE 3. Distribution of the balancing coefficient.

The proposed method is compared with two reference beamforming algorithms. One is the SLNR in [20], and the other is the MRT algorithm. For MIMO configuration, maximum SINR beam combiner is assumed to maximize the achieved data rates. The beam combiner is defined as follows.

$$\mathbf{v}_k = \frac{\mathbf{C}_k^{-1}\mathbf{H}_{kk}\mathbf{w}_k}{|\mathbf{C}_k^{-1}\mathbf{H}_{kk}\mathbf{w}_k|}, \tag{29}$$

where

$$\mathbf{C}_k = \sum\nolimits_{i=1,i\neq k}^{N_c} \mathbf{H}_{ki}^H\mathbf{w}_i\mathbf{w}_i^H\mathbf{H}_{ki} + \mathbf{I}\sigma_k^2 \tag{30}$$

is the covariance matrix of received interference and noise at receiver side of user k.

B. DISTRIBUTION OF THE BALANCING COEFFICIENTS

To perform simulation based on Deep Q-learning, the first thing is to determine the discretizing granularity of the coefficients λ_{jk}. In this paper, we determine the granularity by analyzing the distribution of the coefficients with experiment. Fig.3 shows the probability distribution of the coefficients λ_{jk} when the aforementioned layout of UND is assumed. From Fig.3, it is evident that over 95% of the coefficients samples are within the range [4, 9] in dB. Based on this observation and a compromise between performance and complexity, the even discretization scheme with six levels in (20) is adopted in the simulation. The $\lambda_{jk}^{\max}$ and $\lambda_{jk}^{\min}$ are set according to $\log_{10}(\lambda_{jk}^{\max}) = 9$ and $\log_{10}(\lambda_{jk}^{\min}) = 4$ respectively.

For training and evaluation of the proposed method, 20000 channel samples are generated, and 16000 samples

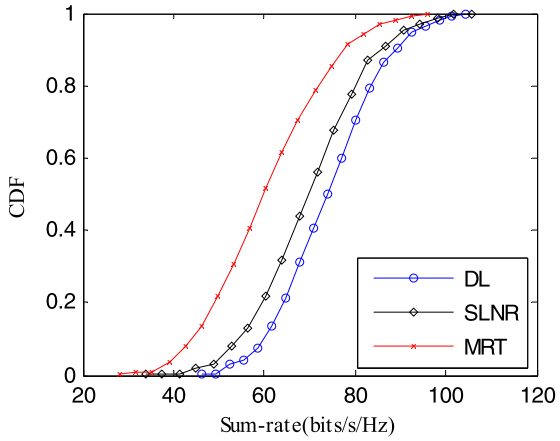

FIGURE 4. Sum-rate performance of different methods for 4 × 2 MIMO.

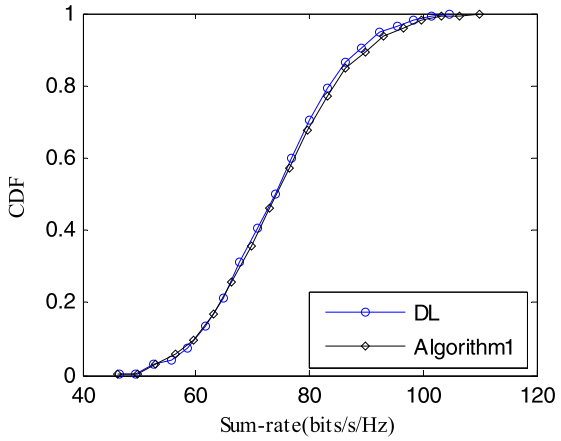

FIGURE 5. Sum-rate performance of different methods for 4 × 2 MIMO.

are employed for training of the DQN. At the training phase, one user is scheduled for transmission per cell, and multiple BS-UE pairs are formed for coordinated beamforming. The agent is selected with random from the BS-UE pairs, and other BS-UE pairs constitute the environment. During the training process, the beamforing vector of the agent is determined by the action provided by the DQN, while the beamforming vectors of BS-UE pairs in the environment are calculated with classical algorithm or algorithm1 in Tab.1. Based on the coordinated beamforming vectors, the corresponding states and the rewards, which are the sum-rate of the system, are calculated. To train the DQN network, both the states and the rewards are fed to the DQN as inputs, and the outputs are the actions of the agent. At the evaluation phase, 4000 samples are used. The predicted balance coefficients by the trained DQN were applied to substitute the values originally provided by (7).

The simulation results are illustrated below from Fig.4-10 for MIMO, MISO, and special case of 2 × 1 MISO configurations. In the simulation, the performances of the DL based algorithm are closely investigated considering the arguments listed below.

- MIMO configuration
- Reference algorithm
- Shadow fading effect
- Dominant factor determining agent state.

C. SIMULATION RESULTS COMPARED WITH CLASSICAL ALGORITHMS

MIMO Configuration Results: In MIMO case, for algoritm 1 to calculate the beamforimg vectors as well as balancing coefficients, iterations between transmitter and receiver sides are required as beamforming vectors at transmitter and receiver sides are coupled as indicated in (7) and (29). However, for case of the proposed DL scheme, the balancing coefficients are learned by DQN in an offline style, so in the online stage, they are provided by the DQN according to the instant states of the agent and keep fixed during the iterations between the transmitter and receiver. As a result, calculation of balancing coefficients is avoided during the iteration for the proposed scheme compared with algorithm1. Fig.4 illustrates the throughput results by the proposed and classical algorithms in terms of the Cumulative Distribution Function (CDF) for 4 × 2 MIMO. From Fig.4, we can see that the performance of MRT is the worst, as it assumes the selfish strategy, so it only enhances the desired signal without consideration of interference towards other UEs. Relatively, the SLNR is better than MRT, but worse than proposed deep learning based beamforming (DL indicated in figure), as the coefficient in SLNR is fixed but not the optimal value in (7), so the result is sub-optimal.

D. SIMULATION RESULTS COMPARED WITH ALGORITHM1

Next, the performance of the proposed DL based scheme is compared with algorithm1 in which the both beamforming vectors $\mathbf{w}_k$ and coefficients λ_{jk} are found iteratively. The number of iteration in algorithm 1 is fixed to 20. The CDF comparison results are shown in Fig.5.

As presented in Fig.5, the performances of the two schemes in terms of system sum-rate match well on the whole, however, there is still small difference between the CDF curves of the two schemes. The small difference mainly exists above the CDF value of 0.5, the difference indicates that the proposed scheme performs better for edge users than for center users compared with algorithm1. Typically, the cell center users and cell edge users correspond to 90th and 5th percentile users. The reasons for this difference can be elaborated from two aspects: 1) Performance loss due to discretization range of action space. From Fig.3, we notice that there are still samples outside the range of discretization. Due to the range limitation, more selfish strategy is not allowed for the center users to select. To overcome this problem, some more advanced discretization schemes such as nonlinear or non-uniform schemes may be required. 2) Absent of leakage components $\mathbf{L}^j, j \neq k$ from state space design in (7). Such absent can reduce the signaling overhead but to some extent at a cost of small performance loss.

MISO Configuration Result: In case of MISO for algoritm 1, iterations are still required to obtain the balancing coefficients, even though the beamforming vectors at the

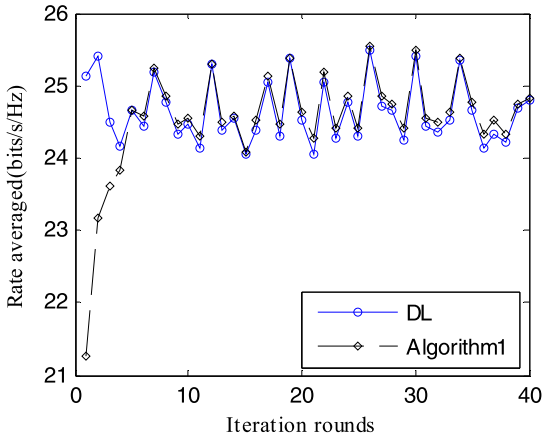

FIGURE 6. Convergence performance of different methods for 4 × 1 MISO.

transmitter and receiver sides are decoupled. However, for the proposed DL based algorithm, iteration for obtaining the balancing coefficients is avoided. The comparison results in Fig.6 are employed to demonstrate such difference in iteration. The results are averaged over 500 experiments. From Fig.6, it is obvious that no iteration is required for the proposed DL scheme, while iterations are required for algorithm1 to reach convergence of the balancing coefficients. The results verify the advantage of the proposed scheme in terms of efficiency for both the processing complexity and time delay.

E. SIMULATION RESULTS ON SHADOWING EFFECT

In perspective of (7), the balancing coefficient depends on both the large and small scale fading in channel. The large scale fading consists of path loss and shadowing effect. To investigate the influence of channel shadowing on the performance of the proposed scheme and the extent of such influence, the performances of the proposed DL based algorithm with and without shadowing effect are compared, moreover, the proposed method is also compared with algorithm1 for MIMO (MISO) and algorithm in [21] for 2 × 1 MISO configuration respectively. The shadowing fading is assumed as lognormal distributed with Standard Deviation (STD) of 5dB.

Fig.7 shows the simulation results in terms of CDF of system sum-rate for 4 × 2 MISO configuration with and without shadow fading, the last case is denoted with 'No' in the figure.

At a first glance of the Fig.7, we get the impression that performance differences due to shadow fading exist for both the proposed algorithm and algorithm1. More precisely, shadow fading will decrease the throughput of cell edge users while increase that of cell center users. But if we look deep at the details of the figure, we see the influence of shadow fading on DL agent is a little more than for algorithm1. The reason is that the shadow fading will introduce more uncertainty in the channel due to the random nature, as a result, there are more uncertainty in the distribution of the balancing coefficients, which is responsible for the performance difference in the cell center users. This phenomenon also exists for 4 × 1 MISO results in Fig.8.

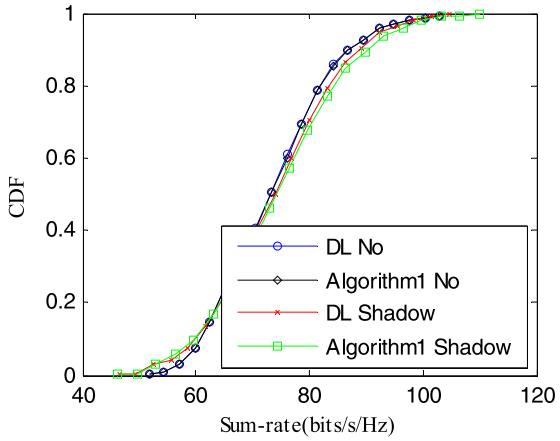

FIGURE 7. Sum-rate performance under shadowing for 4 × 2 MIMO.

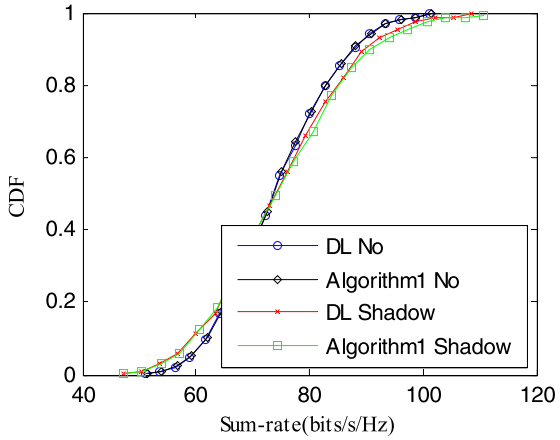

FIGURE 8. Sum-rate performance under shadowing for 4 × 1 MISO.

It should be pointed out that the results in Fig.8 are achieved at a much less cost compared with Fig.7, as they are obtained with reduced state information. This fact corroborates the assertion that beamforming for MISO possesses higher level of parameterization, so less information is required to train the agent, and signaling overhead is saved in practical implement.

For 2 × 1 MISO configuration, [21] has given the expression of $\lambda_k, k = 1, 2$ in (13) theoretically, for convenient reference, we reproduce it here.

$$\lambda_k = \sqrt{\zeta_k}, \quad k = 1, 2, \tag{31}$$

where ζ_k is expressed as

$$\zeta_k = \frac{a_k}{a_k + b_k(1 + c_k)^2}, \tag{32}$$

a_k, b_k, and c_k are defined as:

$$a_k = ||\Pi_{\mathbf{h}_{\bar{k}k}} \mathbf{h}_{11}||^2, \tag{33}$$

$$b_k = ||\Pi^{\perp}_{\mathbf{h}_{\bar{k}k}} \mathbf{h}_{kk}||^2, \tag{34}$$

$$c_k = \frac{p_k}{\sigma_k^2} ||\mathbf{h}_{\bar{k}k}||^2, \tag{35}$$

and $\bar{k}$ is the complement set of k for set k = {1, 2}.

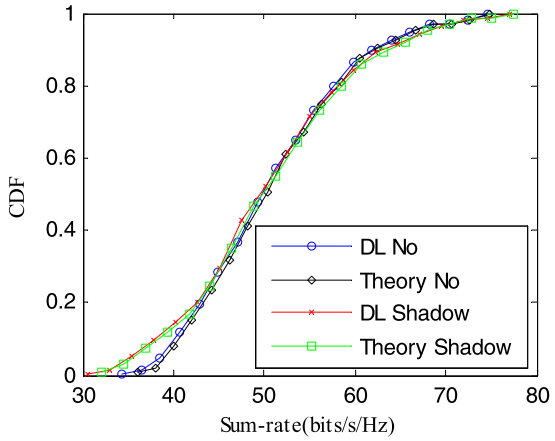

FIGURE 9. Sum-rate performance of different methods for 2 × 1 MISO.

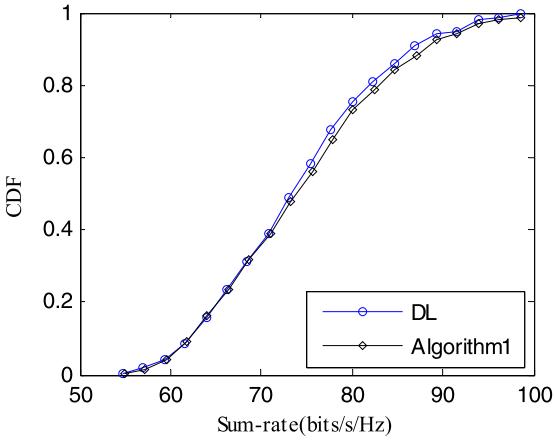

FIGURE 10. Sum-rate performance of different methods for 4 × 2 MIMO when only large scale channel fading is considered in calculating the balancing coefficients.

Fig.9 gives the simulation results in terms of system sum-rate CDF for 2×1 MISO. In the figure, DL based algorithm is compared with algorithm which calculates beamforming vectors by $\lambda_k, k = 1, 2$ in (31). The later is indicated by 'Theory' in Fig.9. From the figure, we conclude that the performance of the DL based algorithm in 2×1 MISO is the least influenced by shadow fading among the MIMO configurations considered. The DL based algorithm manifests the best due to the highest level of parameterization in 2×1 MISO configuration.

F. SIMULATING THE DEPENDENCE OF BALANCING COEFFICIENTS ON LARGE SCALE CHANNEL FADING

Considering that multi-cells beamforming in (8) depend on the wireless channel consisting of large and small scale fading with different time scales, we attempt to separate the original beamforming problem into two sub-problems: 1) predicting the balancing coefficients at large time scales; 2) instant beamforming based on obtained balancing coefficients. Beamforming with balancing coefficients predicted at large time scales can save signaling overheard for channel estimation in practice. To this end, we consider state space S^k design option based on large scale fading in this section. Specifically, the new state space $\mathcal{S}^k$ is predicted only by channels with large scale fading component. The predicted balancing coefficients are used for instant beamforming with MIMO channels consisting of both kinds of fading. To verify this attempt, simulations based on the new state space design are conducted. The simulation results are shown in Fig.10. It is surprising that the performance difference is rather small when state space designs with partial and full channel components are used to predict the balancing coefficients. This result is encouraging for implement: much signaling overheard can be saved for instant channel estimation, and more reliable estimate of the balancing coefficient based on large time scales can be guaranteed.

VI. CONCLUSION

In this paper, we present the parameterized beamforming structure of the coordinated beamforming considering balanced strategy in UDN. Based on analysis, the Deep Reinforcement Learning is proposed to predict the balancing coefficients which parameterize the final beamforming vectors. The proposed method was evaluated in simulated UDN system with different MIMO configurations. Experiment results demonstrate that strategy space for MIMO configuration can be discretized with limited levels and granularity although the range is infinite theoretically. Simulation results confirm the efficiency of the proposed scheme in terms of iteration. The results also reveal the fact that the learned DQN can predict the beamforming strategy only based on the large scale channel fading. The important aspect of this finding is the reduction of signaling overhead in implement.

REFERENCES

[1] M. Kamel, W. Hamouda, and A. Youssef, "Ultra–dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016, doi: 10.1109/comst.2016.2571730.
[2] D. Wu, Q. Liu, H. Wang, Q. Yang, and R. Wang, "Cache less for more: Exploiting cooperative video caching and delivery in D2D communications," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1788–1798, Jul. 2019.
[3] D. Wu, B. Liu, Q. Yang, and R. Wang, "Social-aware cooperative caching mechanism in mobile social networks," *J. Netw. Comput. Appl.*, vol. 149, Jan. 2020, Art. no. 102457, doi: 10.1016/j.jnca.2019.102457.
[4] J. Kim, H.-W. Lee, and S. Chong, "Virtual cell beamforming in cooperative networks," *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, pp. 1126–1138, Jun. 2014, doi: 10.1109/jsac.2014.2328392.
[5] T. O'shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017, doi: 10.1109/tccn.2017.2758370.
[6] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.
[7] W. Lee, M. Kim, and D.-H. Cho, "Deep power control: Transmit power control scheme based on convolutional neural network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1276–1279, Jun. 2018.
[8] Y. Shi, A. Konar, N. D. Sidiropoulos, X.-P. Mao, and Y.-T. Liu, "Learning to beamform for minimum outage," *IEEE Trans. Signal Process.*, vol. 66, no. 19, pp. 5180–5193, Oct. 2018.
[9] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly–mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37328–37348, 2018.
[10] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "Deep learning based beamforming neural networks in downlink MISO systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2019, pp. 1–5.

[11] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero–forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.
[12] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, and B. L. Evans, "Sum capacity of multiuser MIMO broadcast channels with block diagonalization," *IEEE Trans. Wireless Commun.*, vol. 6, no. 6, pp. 2040–2045, Jun. 2007.
[13] J. Zhang, R. Chen, J. Andrews, A. Ghosh, and R. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.
[14] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.
[15] 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, doi: 10.1109/twc.2010.061710.091648.
[16] T. E. Bogale and L. Vandendorpe, "Weighted sum rate optimization for downlink multiuser MIMO coordinated base station systems: Centralized and distributed algorithms," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1876–1889, Apr. 2012.
[17] O. Tervo, L.-N. Tran, and M. Juntti, "Optimal energy–efficient transmit beamforming for multi–user MISO downlink," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5574–5588, Oct. 2015.
[18] D. Wu, Q. Liu, H. Wang, D. Wu, and R. Wang, "Socially aware energy–efficient mobile edge collaboration for video distribution," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2197–2209, Oct. 2017.
[19] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.
[20] M. Sadek, A. Tarighat, and A. Sayed, "A leakage–based precoding scheme for downlink multi–user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1711–1721, May 2007, doi: 10.1109/twc.2007.360373.
[21] R. Zakhour and D. Gesbert, "Coordination on the MISO interference channel using the virtual SINR framework," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, Berlin, Germany, 2009, pp. 75–81.
[22] E. Jorswieck, E. Larsson, and D. Danev, "Complete Characterization of the Pareto Boundary for the MISO Interference Channel," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5292–5296, Oct. 2008.
[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

CHANGYIN SUN received the Ph.D. degree from Xidian University, in 2000. He is currently an Associate Professor with the School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications. His research interests include interference management and radio resource management in the evolved mobile communication systems.

ZHAO SHI received the B.E. degree in communication engineering from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2017, where he is currently pursuing the master's degree. His research interests include deep learning, optimization, and its application in power allocation and beamforming.

FAN JIANG received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree in circuits and systems from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010. She was a Visiting Scholar with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA, from 2016 to 2017. She is currently a Full Professor with the School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an. Her current research interests include wireless communication systems, such as device-to-device communications, fog computing, edge caching, heterogeneous networks, cooperative communications, and relay networks. She was a recipient of the Best Paper Award at the 2015 International Conference on Information and Communications Technologies.