IEEE ICCC 2023 1570914545

Leveraging Deep Reinforcement Learning for Geolocation-based MIMO Transmission in FD-RAN

Zongxi Liu[1], Kai Yu[1,2], Tianqi Zhang[1], Jingbo Liu[2,3], Jiacheng Chen[2], Haibo Zhou[1], and Xuemin (Sherman) Shen[4]
1. School of Electronic Science and Engineering, Nanjing University, Nanjing, China, 210023
2. Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen, China, 518000
3. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, 200240
4. Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada, N2L 3G1
Email: {zongxiliu, kaiyu, tianqizhang}@smail.nju.edu.cn, liujingbo@sjtu.edu.cn, chenjch02@pcl.ac.cn, haibozhou@nju.edu.cn, sshen@uwaterloo.ca

2023 IEEE/CIC International Conference on Communications in China (ICCC) | 979-8-3503-4538-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCC57788.2023.10233428

***Abstract*—In the fully-decoupled radio access network (FD-RAN), the base stations (BSs) are classified as control BSs, uplink BSs, and downlink BSs for better flexibility of the BS-user equipment (UE) association and elasticity of resource scheduling. However, the high feedback delay brought by the physically decoupling architecture renders conventional feedback-based MIMO transmission ineffective. Considering the strong correlation between user geolocation and the channel propagation environment, especially when a line-of-sight (LOS) path exists, we propose a geolocation-based MIMO transmission scheme that determines the user's precoder and modulation and coding scheme (MCS) without requiring the conventional feedback procedures. To capture the correlation, we utilize deep reinforcement learning (DRL) to perform the task of mapping from geolocation to transmission parameters. Based on the principles of Type I codebook, we further decompose the task into subtasks to reduce the action space. The extensive simulations on realistic ray-tracing channels demonstrate comparable performance to CSI feedback-based methods in 5G under ideal (zero feedback delay) conditions and better performance against 5G (with feedback delay) and DRL without task decomposition.**

***Index Terms*—MIMO, precoding, codebook, deep reinforcement learning, FD-RAN**

I. Introduction

Multiple-input multiple-output (MIMO) is acknowledged as the keystone technology of the fifth-generation mobile communication networks (5G). In 5G protocols, the MIMO transmission parameters, including precoder and modulation and coding scheme (MCS), are selected based on a feedback-based mechanism called the closed-loop spatial multiplexing transmission (CLSM) [1]. Take downlink as an example, user equipment (UE) estimates the channel with channel state information reference signal (CSI-RS) and selects the optimal transmission parameters in the format of channel state information (CSI). Then, the CSI is fed back to the base station (BS). However, the overhead of CSI-RS and CSI results in a degradation in spectral efficiency, restricting the improvement of 5G network performance.

Additionally, for the next generation mobile communication networks (6G), an original radio access network (RAN) architecture, namely fully-decoupled RAN (FD-RAN) has been proposed where the conventional BS is decoupled into control BS, uplink BS, and downlink BS [2]. Through hardware separation, the uplink and downlink transmission can be fully independent and elastically adjusted to satisfy UE's diverse and ever-changing service and coverage requirements [3]. However, FD-RAN also brings new challenges to downlink MIMO transmission. Because the uplink and downlink BS are isolated on hardware, feedback needs to be relayed through the control BS, resulting in high feedback delay, which has a negative impact on transmission performance. Although there have been a number of works interested in reducing feedback overhead and delay [4]–[6], the problem could not be solved essentially as long as the feedback-based mechanism exists.

To avoid the problems caused by feedback, geolocation-based transmission schemes without feedback have been studied recently [7], [8]. Intuitively, it is feasible to obtain the transmission parameters based on geolocation information especially when a line-of-sight (LOS) path exists because the direction and travel distance of the LOS path are strongly correlated with the transmission parameters selection. However, this correlation is difficult to be modeled explicitly. Besides, these works merely focused on beamforming with model-driven methods, which may not be suitable for all real-world scenarios. Currently, there still does not exist a geolocation-based transmission scheme based on a data-driven model.

In this work, we propose a geolocation-based MIMO transmission scheme in a downlink FD-RAN leveraging deep reinforcement learning (DRL) with task decomposition. Instead of conventional channel estimation and feedback procedures, we exploit DRL to fit the mapping from geolocation information to transmission parameter selection for situations where an LOS path exists. Furthermore, based on the design principles of Type I codebook, we decompose the task into subtasks so as to reduce the action space [1]. Our contributions are summarized as follows:

- We propose a geolocation-based MIMO transmission method that circumvents the conventional channel estimation and feedback procedure and its drawbacks. The proposed method is also applicable to the FD RAN architecture.
- We exploit DRL to uncover the correlation between geolocation and transmission parameters. We further decompose

the task into subtasks to reduce the action space of DRL.

• We conduct extensive simulations based on realistic ray-tracing channels and 5G-compatible simulators and demonstrate that our method can achieve comparable performance with feedback-based 5G, and outperform 5G when feedback delay exists.

The remainder of the paper is structured as follows. Section II introduces the system model and the principles for selecting optimal transmission parameters. Section III describes our proposed scheme. Section IV shows simulation results and analysis. At last, we make our conclusion in Section V.

II. System Model

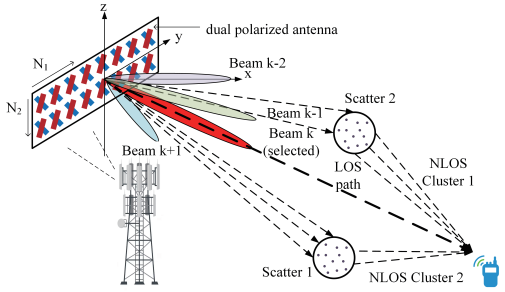

Fig. 1. Transmission scene of a MIMO OFDM downlink transceiver with a frequency selective fading channel.

A. Channel Model

As shown in Fig.1, we consider a downlink MIMO OFDM transceiver system operating in frequency division duplex (FDD) mode, where a frequency selective fading channel that follows the clustered delay line (CDL) model exists between the BS and the UE. The transmitting antenna is a dual-polarized uniform planar array (UPA) with N_T physical antenna ports. In each polarization, the antenna array is composed of N_1 and N_2 antenna ports in the horizontal and vertical dimensions, respectively, and is mapped to $N_T/2$ antenna ports. Therefore, the BS transmitter uses $N_T = 2N_1N_2$ ports to transmit data in two polarizations. The UE has a single-polarized uniform linear array (ULA) with N_R physical antenna ports considering the miniaturization of UE antennas. Based on the CDL model, The frequency selective fading characteristics of the channel for each link between t-th BS antenna ports and r-th UE antenna ports are described by N_c clusters featured by different time delays. Simultaneously, each cluster consists of N_p subpaths having similar angles of arrival (AoA) and angle of departure (AoD). The time-variant channel impulse responses of the link between r-th receiver antenna port and t-th transmitter antenna port at sampling time instant m and sampling time-delay instant η can be expressed as:

$$\begin{aligned} h_{r,t}(m,\eta) &= \sum_{c=1}^{N_c}\sum_{p=1}^{N_p} h_{r,t,c,p}(m,\eta) \\ &= \sum_{c=1}^{N_c}\sum_{p=1}^{N_p} \alpha_{c,p}(m\Delta\xi)\mathrm{e}^{\mathrm{j}\varphi_{c,p}(m\Delta\xi)}\delta\left((\eta-\eta_c)\Delta\tau\right), \end{aligned} \tag{1}$$

where $\alpha_{c,p}$ and $\varphi_{c,p}$ are the gain and phase of the p-th subpath in the c-th cluster respectively, while η_c is the index of the discrete delay of the c-th cluster. $\Delta\xi$ and $\Delta\tau$ represent the sampling interval of the time and time delay respectively and δ is the Dirac delta function. Although the channel is modeled through a power delay profile (PDP) in the time-delay domain, the OFDM technique actually typically describes the channel in the frequency domain rather than the time-delay domain due to the capacity of transforming the frequency selective fading channel into a series of narrowband flat fading channel. And the transfer function of the time-variant channel with N_f subcarriers that captures the frequency domain features of the channel on the k-th subcarrier $H_{r,t}(m,k)$ can be computed based on discrete Fourier transform (DFT):

$$\begin{aligned} H_{r,t}(m,k) &= \mathrm{DFT}[h_{r,t}(m,\eta)] \\ &= \sum_{c=0}^{N_c}\sum_{p=0}^{N_p} h_{r,t,c,p}(m,\eta)\mathrm{e}^{-\mathrm{j}\frac{2kn_c\pi}{N_f}} \end{aligned} \tag{2}$$

B. Downlink Data Transmission

we assume that the channel between the BS and the UE at sampling time instant m is a 3D channel matrix $\mathbf{H}(m) = \left[\begin{array}{cccc}\mathbf{H}_1(m) & \mathbf{H}_2(m) & \cdots & \mathbf{H}_{N_f}(m)\end{array}\right] \in \mathbb{C}^{N_f\times N_R\times N_T}$ where $\mathbf{H}_k(m) = (H_{r,t}(m,k))_{N_R\times N_T}$. The received signal on the k-th subcarrier at sampling time instant m $\mathbf{y}_k(m) \in \mathbb{C}^{N_R\times 1}$ can be expressed as:

$$\mathbf{y}_k(m) = \mathbf{H}_k(m)\mathbf{W}_{\boldsymbol{i}}\mathbf{x}_k(m) + \mathbf{n}_k(m), k \in \{1,\ldots,N_f\}, \tag{3}$$

in which $\mathbf{x}_k(m) \in \mathbb{C}^{L\times 1}$ is the transmit symbol vector with L layers normalized to unit power. $\mathbf{n}_k(m) \in \mathbb{C}^{N_R\times 1} \sim \mathcal{CN}\left(0,\sigma_n^2\cdot\mathbf{I}\right)$ is the additive zero means complex-valued white Gaussian noise with variance σ^2. $\mathbf{W}_{\boldsymbol{i}} \in \mathbb{C}^{N_T\times L}$ is the wideband precoder both in the frequency and time domain, which means that over the whole bandwidth and subframe duration, the same precoder will be used on every resource element. Here $\boldsymbol{i} \in \mathbb{C}^{1\times 3}$ is the index of the precoder in the codebook which will be introduced in Section II-C.

After the received signal is obtained, the receiver will equalize it through an equalizer given by $\mathbf{F}_k(m) \in \mathbb{C}^{L\times N_R}$ which can map the signal from N_R ports to L layers. Typically, $\mathbf{F}_k(m)$ is a linear equalizer satisfying the minimum mean square error (MMSE) design criterion. The output of the equalizer is:

$$\begin{aligned} \mathbf{r}_k(m) &= \mathbf{F}_k(m)\mathbf{y}_k(m) \\ &= \mathbf{F}_k(m)\mathbf{H}_k(m)\mathbf{W}_{\boldsymbol{i}}\mathbf{x}_k(m) + \mathbf{F}_k(m)\mathbf{n}_k(m) \\ &= \mathbf{K}_{k,\boldsymbol{i}}(m)\mathbf{x}_k(m) + \mathbf{F}_k(m)\mathbf{n}_k(m), \end{aligned} \tag{4}$$

where $\mathbf{K}_{k,\boldsymbol{i}}(m)$ is the effective channel using $\mathbf{W}_{\boldsymbol{i}}$ as precoder.

C. Codebook-based Transmission Parameters Selection

For the Type-I codebook, the precoder is selected from a beams candidate set consisting of beams pointing in different

directions. The basis of the codebook $\mathbf{v}_{q_d,n_d} \in \mathbb{C}^{1\times N_d}$ is computed by:

$$\begin{aligned}\mathbf{v}_{q_d,n_d} &= \mathbf{r}_{q_d} \times \mathbf{v}_{n_d} \\ &= \mathrm{diag}\left(\begin{bmatrix} 1 & \mathrm{e}^{\mathrm{j}\frac{2\pi q_d}{N_d O_d}} & \dots & \mathrm{e}^{\mathrm{j}\frac{2\pi q_d (N_d-1)}{N_d O_d}} \end{bmatrix}\right) \\ &\times \begin{bmatrix} 1 \\ \mathrm{e}^{\mathrm{j}\frac{2\pi n_d}{N_d}} \\ \vdots \\ \mathrm{e}^{\mathrm{j}\frac{2\pi n_d (N_d-1)}{N_d O_d}} \end{bmatrix} = \begin{bmatrix} 1 \\ \mathrm{e}^{\mathrm{j}\frac{2\pi (n_d O_d+q_d)}{N_d O_d}} \\ \vdots \\ \mathrm{e}^{\mathrm{j}\frac{2\pi (n_d O_d+q_d)(N_d-1)}{N_d O_d}} \end{bmatrix},\end{aligned} \tag{5}$$

where $d \in \{1,2\}$ corresponds to the horizontal and vertical dimensions of the UPA antenna, respectively. $\mathbf{v}_{n_d} \in \mathbb{C}^{1\times N_d}$ is the DFT orthogonal basis and $\mathbf{r}_{q_d} \in \mathbb{C}^{N_d\times N_d}$ is a rotation factor to refine the beam granularity by O_d oversampling. And the beam candidate set is obtained by:

$$\begin{aligned}&\mathcal{B} = \{\mathbf{b}_{l_1,l_2}\} = \{\mathbf{v}_{q_1,n_1} \otimes \mathbf{v}_{q_2,n_2}\}, \\ &q_1 = 0,1,\dots,O_1-1 \quad n_1 = 0,1,\dots,N_1-1 \\ &q_2 = 0,1,\dots,O_2-1 \quad n_2 = 0,1,\dots,N_2-1,\end{aligned} \tag{6}$$

where, $\otimes$ is the Kronecker product, $l_1 = n_1 O_1 + q_1$ and $l_2 = n_2 O_2 + q_2$; the beam candidate set contains $N_1 O_1 N_2 O_2$ beams vector $\mathbf{b}_{l_1,l_2} \in \mathbb{C}^{1\times N_1 N_2}$ totally. Due to the beam candidate set satisfies the matrix form of DFT, multiplying $\mathbf{H}_k$ with the codebook $\mathcal{B}$ can be viewed as the projection onto the space defined by the beam vector in the spatial domain, the method for selecting the optimal beam is:

$$\begin{aligned}&\mathbf{b}_{i_{11},i_{12}} = \arg\max_{\mathbf{b}_{l_1,l_2}\in\mathcal{B}} \ \mathrm{Pr}_{l_1,l_2} \\ &\text{s.t.} \quad \mathrm{Pr}_{l_1,l_2} = \sum_{k=1}^{N_f} \left(\left\|\mathbf{H}_k^1 \mathbf{b}_{l_1,l_2}\right\|_2 + \left\|\mathbf{H}_k^2 \mathbf{b}_{l_1,l_2}\right\|_2\right),\end{aligned} \tag{7}$$

where Pr_{l_1,l_2} represents the projection value of the $\mathbf{b}_{l_1,l_2}$ on $\mathbf{H}_k$. And $\mathbf{H}_k^1$ and $\mathbf{H}_k^2$ represent the channel element for the ports belonging to polarization 1 and 2 of matrix $\mathbf{H}_k(m)$, respectively. Once the optimal beam selected, the optimal precoder is in the form of:

$$\mathbf{W}_{\boldsymbol{i}} = \begin{bmatrix} \mathbf{b}_{i_{11},i_{12}} \\ \phi_{i_2} \mathbf{b}_{i_{11},i_{12}} \end{bmatrix}, \tag{8}$$

where $\boldsymbol{i} = [i_{11} \quad i_{12} \quad i_2]$ and $\phi_{i_2} = \mathrm{e}^{\mathrm{j}\frac{i_2}{4}} (i_2 \in \{0,1,\dots,7\})$ is the phase difference between the two polarization in order to adapt to the impact of the environment for different polarization. The optimal i_2 is to maximize the effective mutual information over the entire bandwidth and subframe duration. Here, we estimate the effective mutual information on k-th subcarrier at sampling time instant m as:

$$\begin{aligned}\mathbf{I}_{k,m}(\mathbf{W}_{\boldsymbol{i}}) &= \log_2\left(1 + \mathrm{SNR}_{\mathrm{eff}}(m,k,\boldsymbol{i})\right) \\ &= \log_2\left(1 + \frac{|\mathbf{K}_{k,\boldsymbol{i}}(m)|^2}{\sigma_n^2 \sum_{j=1}^{N_R} |\mathbf{F}_k^j(m)|^2}\right),\end{aligned} \tag{9}$$

where $\mathrm{SNR}_{\mathrm{eff}}(m,k,\boldsymbol{i})$ is the effective signal-to-noise ratio (SNR) when using $\mathbf{W}_{\boldsymbol{i}}$ as precoder on k-th subcarrier at sampling time instant m. And $\mathbf{F}_k^j(m)$ is the j-th element of the vector $\mathbf{F}_k(m)$ in (4). Based on the effective mutual information, the optimal precoder is:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}_{\boldsymbol{i}}} \sum_{k=1}^{N_f} \mathbf{I}_{k,m}(\mathbf{W}_{\boldsymbol{i}}). \tag{10}$$

After the precoder is selected, the effective SNR calculated based on the optimal precoder $\mathrm{SNR}_{\mathrm{eff}}^*(m,k)$ can serve as a basis to select optimal MCS, namely channel quality indicator (CQI) value, to maximize system performance. Here the Effective SINR Mapping method is used [9], where effective SNR over the entire bandwidth is averaged to an equivalent SNR of an equivalent single-input-single-output additive white Gaussian noise (AWGN) channel. And the performance of the equivalent SISO AWGN channel is positively correlated with the OFDM system. Mathematically, the equivalent SNR $\mathrm{SNR}_{\mathrm{eq}}(m)$ can be expressed as:

$$\mathrm{SNR}_{\mathrm{eq}}(m) = \beta \Upsilon^{-1}\left(\frac{1}{N_f}\sum_{k=1}^{N_f} \Upsilon\left(\frac{\mathrm{SNR}_{\mathrm{eff}}^*(m,k)}{\beta}\right)\right), \tag{11}$$

where Υ is the Bit Interleaved Coded Modulation (BICM) capacity and β is the fixed calibration value which is one-to-one corresponding with the MCS. Through calibration, the block error rate (BLER) for the equivalent AWGN channel will match that of the OFDM system channel approximately. The meaning of the optimal CQI value CQI^* is to select the maximum achievable MCS while ensuring a BLER ≤ 0.1.

III. Geolocation-Based MIMO Transmission Scheme with DRL

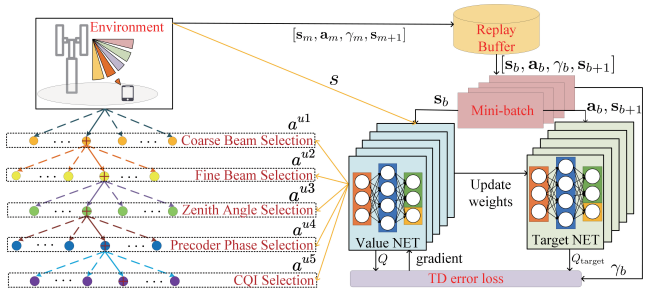

Fig. 2. DRL framework with task decomposition.

In this section, we investigate MIMO transmission with only geolocation information. Foremost, we explain the mechanism behind our scheme. For the precoder, the optimal beam is selected based on (7). As analyzed in Section II-C, the projection is always maximized along the direction of the LOS path because the power of the LOS path is several times higher than the sum of the powers of the other NLOS paths based on the Rician channel model. As for MCS, its selection is closely related to the relative distance between the BS and the UE, because the relative distance and pathloss are positively correlated, determining the MCS selection further. Hence, it is feasible to select the optimal transmission parameters based

on geolocation information. The challenge of this task is that the correlation between UE's geolocation information and the optimal transmission parameters is extremely complex, thereby we seek for DRL.

A. State-Action-Reward Construction

Fig.2 shows the complete DRL framework with task decomposition. In our framework, the agent interacts with the environment in a discrete step manner. For each sampling time instant m, the agent obtains state $\mathbf{s}_m$ and takes the action $\mathbf{a}_m$. The state $\mathbf{s}_m = \left[s_m^0, s_m^1, s_m^2\right]$ is composed of the geolocation information between UE and BS given by $[x(m), y(m)]$ and the transmitting power $\mathrm{P_{BS}}(m)$, which is given by:

$$\left[s_m^0, s_m^1, s_m^2\right] = \left[x(m), y(m), \mathrm{P_{BS}}(m)\right]. \tag{12}$$

The action $\mathbf{a}_m = \left[a_m^0, a_m^1, a_m^2, a_m^3\right]$ is designed to select the optimal transmission parameters. The precoder will be selected as $\mathbf{W}_{[a_m^0, a_m^1, a_m^2]}$ as shown in (8) and the CQI value will be a_m^3. The objective of the agent is to select the precoder and MCS consistent with the optimal selection, respectively. Hence, the reward function also consists of two independent components: $\gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{pmi}}(m)$ and $\gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{cqi}}(m)$. The $\gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{pmi}}(m)$ evaluates the precoder selection as:

$$\gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{pmi}}(m) = \frac{\sum_{k=1}^{N_f} \mathbf{I}_{k,m}(\mathbf{W}_{[a_m^0, a_m^1, a_m^2]} - \mathbf{W}^*)}{\sum_{k=1}^{N_f} \mathbf{I}_{k,m}(\mathbf{W}^*)}. \tag{13}$$

The agent will receive a penalty if it does not take the optimal action, and the closer action is to optimal, the smaller penalty it receives. The $\gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{cqi}}$ evaluates the CQI selection as:

$$\gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{cqi}}(m) = \begin{cases} \frac{a_m^3 - \mathrm{CQI}^*}{\mathrm{CQI}^*}, & a_m^3 \le \mathrm{CQI}^* \\ -1, & a_m^3 > \mathrm{CQI}^* \end{cases}. \tag{14}$$

We set the maximum penalty for the condition $a_m^3 > \mathrm{CQI}^*$ because selecting an overly aggressive MCS will result in high bit error rates (BER) and zero throughputs. And the reward function $\gamma_{\mathbf{s},\mathbf{a}}(m)$ is designed as follows:

$$\gamma_{\mathbf{s},\mathbf{a}}(m) = \gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{pmi}}(m) + \gamma_{\mathbf{s},\mathbf{a}}^{\mathrm{cqi}}(m). \tag{15}$$

B. D3QN Network Design

We utilize the Dueling Double Deep Q-Network (D3QN) algorithm as the basic framework of our proposed scheme. The D3QN algorithm extends the original DQN algorithm [10] by incorporating a dueling architecture for the neural network model used by the agent [11]. The neural network is split into two parts: one part predicts the value of the state, and the other part predicts the advantages of each action. The Q-value is then calculated by combining these two parts:

$$\begin{aligned} \mathrm{Q}_\theta(\mathbf{s}_m, \mathbf{a}_m) = {} & \mathrm{V}_\theta(\mathbf{s}_m, \mathbf{a}_m) + \mathrm{A}_\theta(\mathbf{s}_m, \mathbf{a}_m) \\ & - \frac{1}{|\mathcal{A}|} \sum_{\mathbf{a}'_m} \mathrm{A}_\theta\left(\mathbf{s}_m, \mathbf{a}'_m\right), \end{aligned} \tag{16}$$

where $\mathrm{Q}_\theta(\mathbf{s}_m, \mathbf{a}_m)$, $\mathrm{V}_\theta(\mathbf{s}_m, \mathbf{a}_m)$ and $\mathrm{A}_\theta(\mathbf{s}_m, \mathbf{a}_m)$ is the output for the D3QN, state prediction branch and the advantage prediction branch with weight θ, respectively and $|\mathcal{A}|$ is the size of the action space. Additionally, D3QN also uses a double deep Q-learning approach [12], where the best actions are selected based on the training network while the Q-values are updated using the target network. The trick can reduce the overestimation bias commonly associated with Q-learning algorithms and the target Q-value can be computed as:

$$\mathrm{Q}_{\mathrm{target}}(m) = \gamma_{\mathbf{s},\mathbf{a}}(m) + \zeta \mathrm{Q}_{\theta^-}(\mathbf{s}_{m+1}, \arg\max_{\mathbf{a}'} \mathrm{Q}_\theta(\mathbf{s}_{m+1}, \mathbf{a}')), \tag{17}$$

where θ and θ^- are the weight of the value network and the target network, and ζ is the discount factor. And in the training process, for a batch of data $\{[\mathbf{s}_b, \mathbf{a}_b, \gamma_b, \mathbf{s}_{b+1}]\}_{N_b}$ of size N_b, the aim of the optimizer is to minimize the error between the $\mathrm{Q}_{\mathrm{target}}(b)$ and the $\mathrm{Q}_\theta(\mathbf{s}_b, \mathbf{a}_b)$ through the loss function $\mathrm{L}(\theta)$:

$$\mathrm{L}(\theta) = \frac{1}{2N_b} \sum_{b=1}^{N_b} ||\mathrm{Q}_\theta\left(\mathbf{s}_b, \mathbf{a}_b\right) - \mathrm{Q}_{\mathrm{target}}(b)||_2. \tag{18}$$

C. DRL with Task Decomposition

During the training process as shown in Algorithm 1, the replay buffer memorizes the results of each interaction between the agent and the environment and updates the value network parameters using the mini-batch stochastic gradient descent algorithm with learning rate κ. The parameters of the target network are only synchronized with those of the value network every m_{fixed} step. The action is obtained through $\mathrm{Q}_{\mathbf{s},\mathbf{a}}$ and the ϵ-greedy strategy to make a balance between exploration and exploitation. The core of the algorithm lies in the low dimensionality of the geolocation information that serves as the state. Even if the neural networks have amazing fitting capabilities, it is difficult to deal with tasks with high-dimensional actions and low-dimensional states. Therefore, we decompose the task into multiple subtasks with smaller action spaces, and each subtask is handled by an agent. Obviously, the selection of beam, phase, and CQI are independent of each other, so they can be decomposed into three subtasks. In beam selection, l_1 and l_2 in $\mathbf{b}_{l_1,l_2}$ determine the azimuth and zenith angles of the beams, so this can also be decomposed into two subtasks. In addition, for the azimuth angle selection task with a larger action space, we use a coarse-beam + fine-beam method for selection to reduce the action space. Therefore, there are a total of five subtasks with agent $\mathbf{G}_{ui}$ and action $\mathbf{a}_m^{ui}$ ($ui \in \{1, 2, \ldots, 5\}$): coarse-beam selection, fine-beam selection, zenith angle selection, precoder phase selection, and CQI value selection. In this section, simulation results are given to validate the feasibility of the geolocation-based MIMO transmission scheme leveraging DRL with task decomposition. We generate the channel between the BS and the UE through the DeepMIMO framework [13], which is a large-scale MIMO dataset based on accurate Remcom 3D ray-tracing. Here, the channel is generated through MATLAB API nrCDLChannel and the ray-tracing result from DeepMIMO. The detailed parameters of the channel generation can be seen

Algorithm 1: DRL with Task Decomposition

Input: transmitting power, the position of the UE $\mathbf{s}_m$
Output: index of the precoder and the CQI value $\mathbf{a}_m$

Initialize relay buffer R_{ui} for every agent;
for *episode* $e = 1, 2, \ldots, E$ **do**
  Obtain the initial state $\mathbf{s}_1$;
  for *time* $m = 1, 2, \ldots, M$ **do**
    foreach *agent* $\mathbf{G}_{ui}$ **do**
      Obtain action $\mathbf{a}_m^{ui}$ using ϵ-greedy strategy;
    end
    Obtain reward $\gamma_{\mathbf{s},\mathbf{a}}(m)$ based on (15);
    Update the state to $\mathbf{s}_{m+1}$;
    Store $[\mathbf{s}_m, \mathbf{a}_m, \gamma_m, \mathbf{s}_{m+1}]$ in the replay buffer R_{ui};
    if $N_b \leq m$ **then**
      Sample $\{[\mathbf{s}_b, \mathbf{a}_b, \gamma_b, \mathbf{s}_{b+1}]\}_{N_b}$ from R_{ui};
    end
    Compute $\mathrm{Q}_{target}(m)$ based on (17);
    foreach *agent* $\mathbf{G}_{ui}$ **do**
      Update weights to minimize $\mathrm{L}(\theta)$ in (18);
      if $\mathrm{mod}(m, m_{\mathrm{fixed}}) == 0$ **then**
        Update the weight of the target network
      end
    end
  end
end

in the TABLE I where the candidate BS and UE indices are carefully selected to ensure all the channels have LOS path and form a candidate area in the shape of a 100-meter square.

IV. Simulation Results and Analysis

A. Training Details

We train the neural network in Python3.8 and Pytorch 1.13.0 and the hardware of the environment is 12th Gen Intel(R) Core(TM)i7-12700H in Windows 11. Each neural network has three fully-connected hidden layers with size 128×1024×128 and each layer has a ReLU activation function except the output layer. All the hyperparameters we used can be seen in the TABLE I. Additionally, we mentioned that the network will not rake the optimal actions if we just set the reward function as (15) shown because the penalty is close to zero

TABLE I
Simulation Parameters

Parameter	Value
Channel Scenario	O1_3p5
Candidate BS Index	5, 8
Candidate UE Index	BS5:1213-1702, 3784-3852 rows BS8:1523-2013, 3853-3997 rows
K-factor in Rician Model	13.3 (CDL-D channel in)
Cross-polarization Power Ratio	11 (CDL-D channel in)
BS Antenna Configuration	8×2 ports with two polarization
UE Antenna Configuration	1 port
Subcarrier Spacing	15 kHz
Number of Subcarriers	72
κ, ζ, ϵ, E, M, m_{fixed}	0.002, 0.98, 0.01, 2000, 100, 10
Frame Structure	FDD
Center Frequency	3.5GHz
Equalizer Type	MMSE
Encoder Type	LDPC
Decoder Type	Piece Wise Linear Min-sum

if the action is adjacent to the optimal selection. Hence we set the reward for the optimal selection $r_{\mathbf{s},\mathbf{a}}$ to 10 instead of 0, so that the network is more motivated to converge to the optimal selection. Next, we compare the effectiveness of our proposed method with the original D3QN algorithm without subtasks decomposition. The original D3QN has three fully-connected hidden layers with size 256×1024×256 to ensure that the size of the model parameters is of the same order of magnitude as the sum of that of all agents in our proposed algorithm. The comparison result can be seen in Fig.3, and it shows that reducing action space is vital to improving the performance of the DRL when the input is low-dimensional.

B. Transmission Performance

We verify the transmission performance of our proposed scheme including BER and throughput based on the Vienna 5G Link Level Simulator v1.2 [14]. The simulation parameters can be seen in TABLE I. And we take the ideal CLSM scheme, CLSM scheme with 5 ms delay, and the geolocation-based MIMO transmission scheme using the original D3QN as the benchmarks. We make the comparison in terms of two aspects:

1) BER of the transmission: Fig.4 and Fig.5 illustrate the performance of our proposed scheme in terms of BER. We perform transmission simulations for all links between candidate BSs and UEs to cover the entire candidate area and

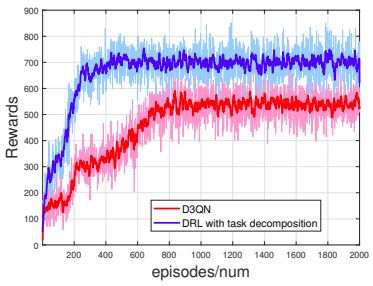

Fig. 3. Reward variation during training process.

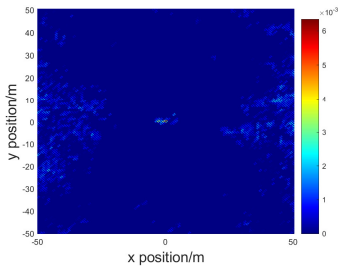

Fig. 4. The BER gap heatmap of the candidate area between the ideal CLSM scheme and our proposed scheme.

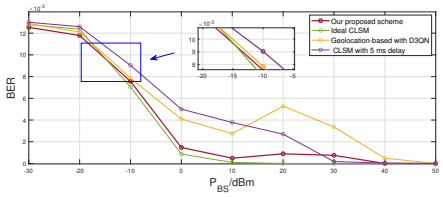

Fig. 5. BER-P_{BS} curve of our proposed scheme versus different benchmarks.

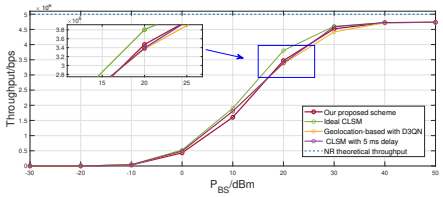

Fig. 6. Throughput-P_{BS} curve of our proposed scheme versus different benchmarks.

obtain the BER gap heatmap of the candidate area between the ideal CLSM scheme and our proposed scheme when the transmitting power P_{BS} was set to 0 dBm as shown in Fig.4. It can be seen that the performance of our proposed scheme is almost identical to that of the ideal CLSM scheme, except for the center. It is affected by the rapid change in the optimal beam selection when the UE is close to the BS. Furthermore, we sweep the P_{BS} from -30 dBm to 50 dBm to obtain the BER-P_{BS} curve with the transmitting power changing as shown in Fig.5. Here, the BER performance of our proposed scheme is 18.8% better than the CLSM scheme with a 5 ms delay and just 4.3% worse than the ideal CLSM scheme. For geolocation-based transmission schemes, abnormal BER increase occurs at high transmitting power levels due to the wrong selection of the CQI value. The abnormal increase is particularly evident in the D3QN scheme, which reflects the superiority of task decomposition to reduce action space.

2) Throughput of the transmission: Fig.6 characterizes the throughput performance when the transmitting power is swept from -30 dBm to 50 dBm. The performance of our proposed scheme is very close to that of the CLSM scheme, with a gap of about 3% compared to the ideal CLSM scheme and a lead of 1% over the CLSM scheme with 5 ms delay as shown in Fig.6. Actually, the throughput of the feedback-free transmission scheme will significantly improve when considering that the overhead of the CSI-RS can save for transmitting more data.

V. Conclusion

In this paper, we have proposed a geolocation-based MIMO transmission scheme in a downlink FD-RAN leveraging DRL with task decomposition to avoid a negative impact on MIMO transmission of high feedback delay caused by the physically decoupling architecture. Our proposed scheme replaces conventional channel estimation and feedback procedure with DRL and decomposes the transmission parameters selection into multiple subtasks to reduce the action space of the agent. Simulation results based on realistic ray-tracing channels have demonstrated that the proposed task decomposition algorithm has similar transmission performance compared to the ideal CLSM scheme, and better performance than the CLSM scheme with a 5 ms delay, which means it is more suitable for the FD-RAN scenario. We will expand our work to multiple BS and UE scenarios for future work.

Acknowledgment

This work was supported in part by the National Natural Science Foundation Original Exploration Project of China under Grant 62250004, the Natural Science Foundation of China (NSFC) under Grant 62001259, 62271244, the Natural Science Fund for Distinguished Young Scholars of Jiangsu Province under Grant BK20220067, and the National R&D Program of China (2020YFB1807503).

References

[1] 3GPP, "NR; Physical layer procedures for data," Technical Specification (TS) 38.214, 3rd Generation Partnership Project (3GPP), 06 2022. Version 16.10.0.

[2] Q. Yu, H. Zhou, J. Chen, Y. Li, J. Jing, J. J. Zhao, B. Qian, and J. Wang, "A fully-decoupled ran architecture for 6g inspired by neurotransmission," *Journal of Communications and Information Networks*, vol. 4, no. 4, pp. 15–23, 2019.

[3] K. Yu, Q. Yu, Z. Tang, J. Zhao, B. Qian, Y. Xu, H. Zhou, and X. Shen, "Fully-decoupled radio access networks: A flexible downlink multi-connectivity and dynamic resource cooperation framework," *IEEE Transactions on Wireless Communications*, 2022.

[4] Z. Liu, S. Sun, Q. Gao, and H. Li, "CSI Feedback Based on Spatial and Frequency Domains Compression for 5G Multi-User Massive MIMO Systems," in *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 834–839.

[5] W. Wu, N. Cheng, N. Zhang, P. Yang, W. Zhuang, and X. Shen, "Fast mmwave beam alignment via correlated bandit learning," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5894–5908, 2019.

[6] Y. Liao, Y. Hua, and Y. Cai, "Deep learning based channel estimation algorithm for fast time-varying mimo-ofdm systems," *IEEE Communications Letters*, vol. 24, no. 3, pp. 572–576, 2020.

[7] A. Bletsas, A. Lippman, and J. N. Sahalos, "Simple, zero-feedback, distributed beamforming with unsynchronized carriers," *IEEE journal on selected areas in communications*, vol. 28, no. 7, pp. 1046–1054, 2010.

[8] S. Hanna, E. Krijestorac, and D. Cabric, "Destination-feedback free distributed transmit beamforming using guided directionality," *IEEE Transactions on Mobile Computing*, 2022.

[9] R. Sandanalakshmi, "Effective snr mapping for link error prediction in ofdm based systems," *IET-UK International Conference on Information and Communication Technology in Electrical Sciences*, pp. 684–687(3), January 2007.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *International conference on machine learning*, pp. 1995–2003, PMLR, 2016.

[12] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[13] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *Proc. of Information Theory and Applications Workshop (ITA)*, (San Diego, CA), pp. 1–8, Feb 2019.

[14] S. Pratschner, B. Tahir, L. Marijanovic, M. Mussbah, K. Kirev, R. Nissel, S. Schwarz, and M. Rupp, "Versatile mobile communications simulation: the Vienna 5G Link Level Simulator," vol. 2018, p. 226, Sept. 2018.