Deep Reinforcement Learning Based Beam Selection for Hybrid Beamforming and User Grouping in Massive MIMO-NOMA System

IRFAN AHMED, (Senior Member, IEEE), MUHAMMAD KHALIL SHAHID, (Member, IEEE), AND TARIG FAISAL

Department of Electrical Engineering, Higher Colleges of Technology, Abu Dhabi, United Arab Emirates

Corresponding author: Irfan Ahmed (iahmed3@hct.ac.ae)

This work was supported in part by the HCT Research Grants from the Higher Colleges of Technology (HCT), Abu Dhabi, United Arab Emirates, under Grant 2215 and Grant 2307.

ABSTRACT

This paper presents a deep reinforcement learning-based beam-user selection and hybrid beamforming design for the multiuser massive multiple-input multiple-output (MIMO) non-orthogonal multiple access (NOMA) downlink systems. The conventional hybrid beamforming in massive MIMO provides multiple directional beams, but each beam serves only one user. The integration of NOMA with the massive MIMO enables power domain multiplexing within a beam, hence increasing the system capacity. In this paper, we first design a channel gain and correlation-based users grouping algorithm per beam, and then using the deep reinforcement learning-based beam selection, a beamspace orthogonal analog precoder is obtained. The deep Q-network consists of a main network and target network with Adam optimizer. Finally, optimal power is allocated to the users in each beam. Simulation results show that at transmit SNR of 10 dB, the proposed scheme provides a 42% increase in sum-rate and energy efficiency performance as compared to the state-of-the-art K-means users’ grouping and Stable Matching-based beam selection NOMA scheme.

INDEX TERMS

Reinforcement learning, hybrid beamforming, millimeter wave, massive MIMO, non-orthogonal multiple access.

I. INTRODUCTION

The current and future demand for wireless and mobile data can only be met by the ultra-high-speed beyond-fifth-generation (B5G) wireless networks. The 5G/B5G technology uses a high-frequency millimeter wave (mmWave) band ranging from 30 GHz to 300 GHz with massive multiple-input-multiple-output (MIMO) systems [1]. The hybrid beamforming overcomes the high power consumption in the radio frequency (RF) chain, and non-orthogonal multiple access (NOMA) enables multiplexing users within each beam. However, the mmWave frequencies suffer from high pathloss and have low penetration power. These shortfalls are compensated by the massive MIMO technique. Usually, massive MIMO is deployed at the base-station to get the benefits of massive MIMO precoding in the downlink and combining in the uplink. The conventional MIMO techniques use one RF chain (analog-to-digital converter (ADC), digital-to-analog converter (DAC), mixer, data converters), which is infeasible in massive MIMO scenario because of the large capital cost (CAPEX) and the operational cost (OPEX). In order to reduce the number of RF chains and hence, the power consumption, researchers take advantage of the mmWave channel’s sparsity and split the beamforming into two stages: analog beamforming (AB) and digital beamforming (DB). This method is called hybrid beamforming (HBF). In the downlink, the data streams are first processed by the digital beamformer then, and then the pre-processed signal is fed to the phase shifter network in the analog (or RF) domain. Finally, the
symbol vector is transmitted from a large antenna array at mmWave frequencies. In this way, mmWave massive MIMO system becomes the key enabler of next-generation 5G communications.

The mmWave massive MIMO system can produce large number of narrow directional beams and can serve many users in beam domain without interference, for example, discrete Fourier transform (DFT) based orthogonal beams. Most of the current research on beam selection focuses on one-beam one-user model or at least one beam per user [2], [3], [4]. However, there is always a probability that a particular beam is the strongest beam for more than one user, even if there are a large number of beams. For $B$ beams and $K$ users, the probability of a user sharing the same beam is $P = 1 - \frac{B^B}{(B-K)^K}$ [5]. For $B = 128$ and $K = 16$, $P = 62.4\%$.

In orthogonal multiple access (OMA) MIMO networks, the multiplexing gain and the system capacity reduce when channel correlation increases. However, when the users’ channels are highly correlated, the use of NOMA-MIMO can provide the optimal dirty paper coding (DPC) performance [6]. Therefore, NOMA has an inherent optimal match with the highly directive and correlated mmWave MIMO channel characteristics.

There is a lot of research work on the design of hybrid beamforming and NOMA in the mmWave massive MIMO system. In [7], a beamspace channel-based massive MIMO-NOMA scheme is presented. It designs a zero-forcing (ZF) precoder by taking a stronger user’s channel per beam. Then, a dynamic power allocation scheme is devised to maximize the sum-rate subjected to total power and per user minimum rate constraints. However, this paper does not discuss the beam-user selection. The paper [8] introduces multi-beam beamspace NOMA-MIMO, where NOMA is applied across multiple beams using a power splitter connected to one RF chain. In the beam-user pairing, each user selects the beam which has the maximum gain for that user. The inter-group interference is controlled by the singular value decomposition (SVD) based on equivalent channel ZF precoding [7]. The NOMA power allocation is based on the minimization of the total power consumption. This scheme has inferior energy efficiency (EE) and spectral efficiency (SE) with higher transmit power as compared with the single beam NOMA. In [9], beam selection has been performed by using an intelligent technique, like particle swarm optimization (PSO) and a correlation-based scheme on the beamspace channel. Then, target rate-based and PSO-based power allocation are carried out for NOMA users. The proposed beam selection methods only utilize the beam strength and beam correlation without considering the users’ channels’ gains with the assumption that the number of users ($K$) is greater than the number of RF chains ($N_{RF}$). A two-user per beam NOMA for joint power allocation and hybrid beamforming design is proposed in [10]. It decomposes the problem into two sub-problems: a power allocation problem and the hybrid beamforming design problem with constant modulus constraint due to the phase shifter implementation of the analog beamformer.

The authors compare the proposed solution with the time division multiple access (TDMA) and show the efficacy of the proposed solution. It does not show the performance gap between the proposed suboptimal solution and the optimal solution. Another two-user per beam NOMA is proposed in [11]. Each beam serves two users who have maximum channel gain difference and channel covariance. It uses a finite resolution analog beamformer which minimizes the difference with the angle of the channel vector. Finally, the digital beamformer is formed by using ZF precoding based on the stronger user equivalent channel. But this paper does not consider the massive MIMO system and utilizes only 8 antennas. In the extended work [12], the authors compare the EE and SE performance of fully-connected and sub-connected architectures. Though optimal power allocation has been derived for multiple users [13], [14] but this is for a single channel. Multi-channel NOMA optimal power allocation for two users per channel is given in [15].

In most of the previous work, NOMA techniques within the same beam are realized by different combinations of $K$-means, intra-beam inter-user correlation, and channel gains [9], [16], [17]. However, we propose to use the upper bound on the correlation for the selections of users’ groups, e.g., among larger channel gain users, a user with a correlation less than the threshold is selected as primary user or grouphead. In most of the previous work, analog beamformers have been realized by discrete Fourier transform (DFT), conjugate transpose channel vectors, eigenvalue decomposition (EVD), and singular value decomposition (SVD) based methods followed by ZF or minimum mean squared error (MMSE) based baseband precoders.

In this paper, we propose a novel users’ channel gain and channel correlation-based users grouping and deep reinforcement learning based beam selection and analog precoder design to efficiently reduce the inter-user interference in the analog domain in the mmWave massive MIMO channel. It provides the precoder with orthogonal column vectors and reduces the effective channel dimension to help design a low-dimensional regularized zero-forcing (RZF) digital precoder. Finally, we consider a general case of multi-beam and multi-user in the beam for power allocation and obtain an optimal solution.

The rest of the paper is organized as follows. The system, signal, and channel models are described in Section II. The problem is formulated in Section III. Section IV presents the proposed users grouping and various beam-user selection methods for hybrid precoder design in a multiuser massive MIMO system. Deep reinforcement learning-based beam selection for grouped users is given in Section V. Power allocation solution is presented in Section VI. Simulation results are given in Section VII, followed by the conclusions in Section VIII.

Notations: Bold upper/lower case letters denote vectors and matrices, respectively. The notations $X^T$, $X^H$, $X^*$, and $X^H$
denote the pseudo-inverse, transpose, conjugate, and conjugate transpose of a matrix $X$. The element in the $i$th row and $j$th column of $X$ is denoted by $[X]_{i,j}$. For a set $A$, $\text{Card}(A)$ is the cardinality of set $A$. For the quick reference, the list of main symbols used in this paper is given in Table 1.

### II. System Model

We consider a downlink of a single cell multiuser massive MIMO-NOMA system. The base-station (BS) is located in the cell center and is equipped with $N$ antennas and $N_{RF}$ RF chains to serve $K$, single antenna users, as shown in Fig. 1. The MIMO-NOMA system has $N_{RF}$ RF chains, such that $N_{RF} \leq K < N$. At BS, each antenna is connected with all the RF chains through an independent phase shifter to form a fully-connected structure [1]. The user set $\mathcal{K}$ is defined as $\mathcal{K} = \{k | k \in \{1, 2, \ldots, K\}\}$.

Assuming $N_t$ the number of information symbols streams at the input of baseband precoder is equal to the number of users $K$, the symbol vector is

$$s = [s_1, s_2, \ldots, s_K]^T.$$  

(1)

The hybrid precoder can be written as,

$$F_B = F_AB F_{DB} \in \mathbb{C}^{N \times K}.$$  

(2)

With the $N_t = K$ is the number of information symbols streams at the input of the baseband precoder, upon receiving the data block, the BS first uses a low dimensional digital precoder $F_{DB} \in \mathbb{C}^{N_{RF} \times K}$ and then, an analog precoder $F_{AB} \in \mathbb{C}^{N \times N_{RF}}$ is used to create the $N \times 1$ transmit symbols as shown in Fig. 1. The digital precoder is given as

$$F_{DB} = [f_{DB,1}, f_{DB,2}, \ldots, f_{DB,K}] \in \mathbb{C}^{N_{RF} \times K},$$  

(3)

and the analog precoder is

$$F_{AB} = [f_{AB,1}, f_{AB,2}, \ldots, f_{AB,K}] \in \mathbb{C}^{N \times N_{RF}}.$$  

(4)

After passing through the digital and analog precoders, the transmitted symbol vector is $x \in \mathbb{C}^{N \times 1}$. It can be written as,

$$x = F_{AB} F_{DB} Ps \in \mathbb{C}^{N \times 1}$$  

(5)

$$= \sum_{i=1}^{K} F_{AB} F_{DB} s_i,$$  

(6)

where $x \sim \mathcal{CN}(0, I_N)$ is the normalized transmit signal vector with $\mathbb{E}[xx^H] = I_N$ or $\mathbb{E}[|s_i|^2] = 1$ and $P = \text{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K})$ is the diagonal matrix representing the power allocation to users.

The received signal $y_k$ at the user $k$ is given as

$$y_k = h_k^H x + z_k,$$

$$= h_k^H F_{AB} F_{DB} Ps + z_k,$$  

(7)

where $z_k \sim \mathcal{CN}(0, 1)$ is the AWGN noise and $h_k$ is the $\mathbb{C}^{N \times 1}$ channel vector between BS and user $k$. The received signal of each user can be combined to form the composite received signal as $y = [y_1, \ldots, y_K]^T$. We assume $N_{RF} < K$, for MIMO-NOMA implementation.

In order to describe the mmWave channel, we use the extended Saleh-Valenzuela model [18]. Due to the directional nature of mmWave propagation, we assume that the channel contributes limited propagation paths $L_p$ between the transmitter and receiver [19]. It is a geometrical channel model which describes the physical propagation between transmit and receive antenna array. At mmWave frequencies, the electromagnetic wave propagation is near optical line-of-sight (LOS) wave propagation. The $N \times 1$ MISO channel matrix can be written as

$$h_k = \sum_{l=0}^{L_p} a_{l}^k a(l)^{\dagger}_{k},$$  

(8)

where $j = \sqrt{-1}$, $a_{l}^k$ and $a(l)^{\dagger}_{k}$, $l = 1, \ldots, L_p$ represent the complex gains of the LOS, $l = 0$ path and non-line-of-sight (NLOS), $l = 1, \ldots, L_p$ paths with i.i.d. $\mathcal{CN}(0, 1)$, respectively. Moreover, $a$ is the array steering vector of the uniform linear array (ULA). The variable $a(l)^{\dagger}_{k}$ is the $l$th path’s azimuth angle (boresight angles in the array) of arrival [2] for user $k$ and it is uniformly distributed with $\phi(l)^{\dagger}_{k} \sim U(-\pi/2, \pi/2)$. The transmit steering vector is given by

$$a(l)^{\dagger}_{k} = \frac{1}{\sqrt{N}} \left[\exp(-j2\pi \lambda^2 d l)\right]_{l \in \mathcal{I}(N)},$$  

(9)

where $\mathcal{I}(N) = \{i - (N - 1)/2, i = 0, 1, \ldots, N - 1\}$ is a symmetric index set centered at zero [20], $\phi(l)^{\dagger}_{k} = \frac{\pi}{2} \sin(\theta(l)^{\dagger}_{k})$ is the array response angle, $\theta$ is the physical angle whose value ranges between $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$, $\lambda$ is the wavelength, and $d$ is the antenna spacing at the base-station. Finally, the overall

### Table 1. List of main symbols.

| Symbol | Description |
|--------|-------------|
| $N$    | Number of BS antennas |
| $N_{RF}$ | Number of RF chains at BS |
| $K$    | Number of users |
| $N_s$  | Number of data streams |
| $\mathcal{K}$ | Set of users |
| $\mathcal{B}$ | Set of selected beams |
| $K_n$  | Set of users in beam $n$ |
| $\chi$ | beam-user selection binary variable |
| $\mathcal{K}_{PU}$ | Set of primary users |
| $\mathcal{K}_{SU}$ | Set of secondary users |
| $\mathcal{C}$ | Correlation matrix |
| $\phi(l)^{\dagger}_{k}$ | Array response angle of users $k$ in path $l$ |
| $\theta(l)^{\dagger}_{k}$ | Physical angle of users $k$ in path $l$ |
| $\mathcal{B}$ | Beams index set |
| $\lambda$ | Carrier wavelength |
| $\rho$ | Transmit SNR |
| $\Gamma$ | SNR gap |
| $\Theta$ | Weights matrix of DNN |
| $S$ | State space |
| $A$ | Action space |
| $\nabla$ | Gradient operator |
channel matrix $\mathbf{H}$ of multiuser MIMO system is given as, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{C}^{N \times K}$.

### III. PROBLEM FORMULATION

In this section, we formulate the joint beam selection and hybrid beamforming design problem for the MIMO-NOMA downlink system to maximize the sum-rate performance. Let $K_n$ be the set of users in a beam $n$, where $n \in \mathcal{B}$ and $\mathcal{B}$ is the set of selected beams. No two beams serve the same user i.e., $\sum_{n \in \mathcal{B}} \text{Card}(K_n) = K$. After the NOMA users’ grouping, we get the channel matrix $\tilde{\mathbf{H}} \in \mathbb{C}^{N \times \text{Card}(\mathcal{B})}$, with Card$(\mathcal{B})$ number of users groups. The mmWave channel has a sparse structure with a limited number of channels with significant power. This mmWave massive MIMO spatial channel can be transformed into a beam domain channel $\mathbf{U}^H \tilde{\mathbf{H}} \in \mathbb{C}^{N \times \text{Card}(\mathcal{B})}$, where $\mathbf{U}$ is a unitary discrete Fourier transform (DFT) matrix of size $N$. Based on the beam domain channel $\mathbf{U}^H \tilde{\mathbf{H}}$, we select Card$(\mathcal{B})$ number of beams using the beam selection algorithms such that $\mathbf{U}^H \tilde{\mathbf{H}}(b,.)|_{b \in \mathcal{B}} \in \mathbb{C}^{\text{Card}(\mathcal{B}) \times \text{Card}(\mathcal{B})}$. This also gives us the DFT-based analog precoder implementation as $\mathbf{F}_{AB} = \mathbf{U}(b,.)|_{b \in \mathcal{B}}$ of size Card$(\mathcal{B}) \times N$. Finally, the equivalent channel matrix is obtained by stacking back the users’ channel columns within each group. The equivalent channel matrix for the $n^{th}$ beam is given by

$$\tilde{\mathbf{H}}_{eq,n} = [\tilde{\mathbf{h}}_{eq,1,n}, \ldots, \tilde{\mathbf{h}}_{eq,\text{Card}(K_n),n}] \in \mathbb{C}^{\text{Card}(\mathcal{B}) \times \text{Card}(K_n)},$$

where $\tilde{\mathbf{h}}_{eq,k,n} = h_{AB,n}^H \tilde{\mathbf{h}}_{k,n}$ is the equivalent reduced dimension channel vector for user $k \in K_n$ in the beam $n$. When using the user’s of highest channel gain per beam and stacking all selected beams together, we get the overall equivalent reduced dimension channel matrix of size Card$(\mathcal{B}) \times \text{Card}(\mathcal{B})$,

$$\tilde{\mathbf{H}}_{eq} = [\tilde{\mathbf{h}}_{eq,1}, \ldots, \tilde{\mathbf{h}}_{eq,\text{Card}(\mathcal{B})}].$$

The ZF digital precoder is obtained by the pseudo-inverse of $\tilde{\mathbf{H}}_{eq}$ as $\mathbf{F}_{DB,eq} = \tilde{\mathbf{H}}_{eq}^H (\tilde{\mathbf{H}}_{eq}^H \tilde{\mathbf{H}}_{eq})^{-1}$. This can be represented as a stack of Card$(\mathcal{B})$ digital precoders for users, $\mathbf{F}_{DB,eq} = \{\mathbf{f}_{DB,eq,1}, \ldots, \mathbf{f}_{DB,eq,\text{Card}(\mathcal{B})}\}$. To meet the total power constraint, the normalized digital precoder is given as

$$\tilde{\mathbf{f}}_{DB,eq,i} = \frac{\mathbf{f}_{DB,eq,i}}{\|\mathbf{f}_{DB,eq,i}\|}, \quad \forall i \in \mathcal{B}. \tag{12}$$

Although, this digital precoding cannot completely eliminate the inter-beam interference because of the use of an equivalent channel of the stronger user in a beam. But in the sparse mmWave channel, if LOS exists (which is often the case in highly directive mmWave channel), then the sparse channel vectors of different users in the same beam are highly correlated [7]. Without loss of generality, we assume that $|\tilde{\mathbf{h}}_{eq,1,1}^H \tilde{\mathbf{F}}_{DB,eq,1}| \geq |\tilde{\mathbf{h}}_{eq,2,1}^H \tilde{\mathbf{F}}_{DB,eq,1}| \geq \ldots \geq |\tilde{\mathbf{h}}_{eq,\text{Card}(\mathcal{B}),1}^H \tilde{\mathbf{F}}_{DB,eq,1}|$ for $n = 1, \ldots, \text{Card}(\mathcal{B})$. The successive interference cancellation (SIC) at user with higher channel gain decode and removes the signals of weaker users. With this hybrid precoding, the received signal at user $k$ in the beam $n$ is given by

$$y_{k,n} = \tilde{\mathbf{h}}_{eq,k,n}^H \sum_{j=1}^{\text{Card}(\mathcal{B})} \sum_{i=1}^{\text{Card}(\mathcal{B})} \tilde{\mathbf{f}}_{DB,eq,j}^H \sqrt{p_{j,i}} s_{i,n} + z_{k,n}. \tag{13}$$

This equation can be decomposed into the following components, the desired signal, intra-beam interference, inter-beam interference, and the AWGN noise as follows:

$$y_{k,n} = \tilde{\mathbf{h}}_{eq,k,n}^H \tilde{\mathbf{f}}_{DB,eq,n} \sqrt{p_{k,n} s_{k,n}} \tag{14}$$

$$+ \tilde{\mathbf{h}}_{eq,k,n}^H \tilde{\mathbf{f}}_{DB,eq,n} \sum_{i=1}^{\text{Card}(\mathcal{B})} \sqrt{p_{i,n}} s_{i,n} + \sum_{j \neq k} \sum_{i=1}^{\text{Card}(\mathcal{B})} \tilde{\mathbf{f}}_{DB,eq,j}^H \sqrt{p_{j,i}} s_{i,j} + z_{k,n}.$$
Where information (CSI) is available at the BS, then the spectral efficiency (SE) for user $k$ in the beam $n$ is given as

$$R_{k,n} = \log_2(1 + \frac{\gamma_{k,n}}{\Gamma}),$$  \hspace{1cm} (16)$$

where $\Gamma$ is the SNR gap that relates the Shannon capacity and the received SNR by the employed modulation and coding scheme in the practical wireless channel. In case of M-ary quadrature amplitude modulation (M-QAM) and target bit error rate $Pe$, $\Gamma = -(2/3)\ln(5Pe)$ [21]. Our objective is to design a hybrid precoder to maximize the spectral efficiency (SE) for multiuser massive MIMO-NOMA systems. The sum-rate is given by

$$R_{sum} = \sum_{n=1}^{\text{Card}(B)} \sum_{k=1}^{\text{Card}(K_n)} \log_2(1 + \gamma_{k,n}).$$  \hspace{1cm} (17)$$

The optimization problem is given as $\mathcal{P}$:

$\mathcal{P}: \max_{\chi, \mathbf{F}_{AB}, \mathbf{P}} R_{sum}(\mathbf{H}, \mathbf{F}_{AB}, \chi, \mathbf{P})$ \hspace{1cm} (18)

s.t.

$\forall 1 \leq i \leq N, \forall 1 \leq j \leq \text{Card}(B), \hspace{1cm} (18a)$

$\chi_{k,n} = \text{Card}(K_n), \hspace{1cm} (18b)$

$\forall n \in \{1, \ldots, N\}, \hspace{1cm} (18c)$

$\sum_{n=1}^{N} \chi_{k,n} = 1, \forall k \in \{1, \ldots, K\}$ \hspace{1cm} (18d)

$\frac{\text{Card}(B)}{\text{Card}(K_n)} \leq K$ \hspace{1cm} (18e)

$\sum_{n=1}^{N} \sum_{k=1}^{K} p_{k,n} \leq P$ \hspace{1cm} (18f)

$\sum_{n=1}^{N} p_{1,n} < \sum_{n=1}^{N} p_{2,n} < \cdots < \sum_{n=1}^{N} p_{\text{Card}(K_n)}, \hspace{1cm} (18g)$

$R_{k,n} \geq \gamma_{k,n}^{\text{min}}, \forall n, k \hspace{1cm} (18h)$

$\sum_{n=1}^{N} \sum_{k=1}^{K} \chi_{k,n} \leq K, \hspace{1cm} (18i)$

where $\chi_{k,n}$ is the beam-user selection binary variable. $\chi_{k,n} = 1$ if user $k$ is selected in beam $n$, otherwise $\chi_{k,n} = 0$. The first constraint (18a) is a uni-modulus constraint for phase shifter implementation of the analog beamformer. The constraint (18b) ensures that the beam-user selection variable does not exceed the number of users in a particular beam. Constraint (18c) is used to allocate at most one beam to a user. (18d) restrict total users in all beams to $K$. Constraints (18e) and (18f) regulate the power allocation. The per user quality of service (QoS) is ensured by the constraint (18g). The beam-user selection variable can take only binary values (18h), and the constraint (18i) gives the upper bound on the beam-user selection.

The problem $\mathcal{P}$ is a mixed-integer programming (MIP) problem, which is NP-hard [22]. Even the optimal solution to the beam-users selection problem with $K \geq \text{Card}(B)$ can only be obtained by the exhaustive search method. This solution has very high computational complexity of $N^K$ [21]. The power allocation to users is coupled and depends on the power allocation to other users in a beam. The intra-beam interference is minimized by the SIC technique in each beam, and the inter-beam interference is minimized by the ZF precoding with the help of stronger users’ equivalent channels.

### IV. USER GROUPING AND BEAM-GROUP SELECTION

In a massive MIMO system with $N$ antennas at BS and $K$ single antenna users, an exhaustive beam search gives optimal beam-user pairing. In order to select $K$ optimal beams out of $N$ beams, the required number of searches is $\binom{N}{K}$. For example, if $N = 128$ and $K = 16$, then it requires $9.33 \times 10^{19}$ searches. When $K \neq N_{RF}$, the beam-user selection problem becomes an unbalanced assignment problem. In an unbalanced assignment problem, one set of the bipartite graph is larger than the other set, as shown in Fig. 2. One way to make the problem tractable is to transform the unbalanced assignment problem into a balanced assignment problem by adding $N - N_{RF}$ new vertices to the smaller set with edges of

$$\gamma_{k,n} = \frac{p_{k,n}||\mathbf{h}^H_{eq,k,n} \tilde{\mathbf{h}}_{DB,eq,n}||^2}{||\mathbf{h}^H_{eq,k,n} \tilde{\mathbf{h}}_{DB,eq,n}||^2 \sum_{i=1}^{k-1} p_{i,n} + \sum_{j \in B[n]} ||\mathbf{h}^H_{eq,k,n} \tilde{\mathbf{h}}_{DB,eq,j}||^2 \sum_{i=1}^{\text{Card}(B)} p_{i,j} + \sigma^2}. \hspace{1cm} (15)$$
cost zero. In this section, first, we form users’ grouping and then apply various beam-group selection algorithms.

A. USERS GROUPING
We form Card(B) groups of K users such that \( \sum_{n=1}^{\text{Card}(B)} \text{Card}(K_n) = K \). The users’ grouping and beam-group based hybrid beamforming algorithm 1 is summarized below:

Step 1: In the first step, we find the channel-to-noise ratio (CNR) in line 2 with the noise power of unity. The correlation matrix is computed in line 4. Next line masks the \( C_{i,j} = 1, \forall i = j \) with zero values to eliminate the correlation with itself.

Step 2: Since there is one primary-user in each beam, there should be a minimum correlation between the primary-users to minimize the inter-beam interference. Let \( \epsilon_{\text{low}} \) be the upper threshold required to minimize the correlation between primary-users. The while loop forms a set of \( N_{\text{RF}} \) primary users. If the maximum correlation of a user is less than the \( \epsilon_{\text{low}} \), then that user is selected as the primary user (PU). After that, the second highest user is checked for \( \text{Corr} < \epsilon_{\text{low}} \) then, this second highest user is selected otherwise next in line user is tested for \( \text{Corr} < \epsilon_{\text{low}} \) and so on till Card(B) users are selected. Ideally \( \epsilon = 0 \) for no correlation.

Step 3: The for loop in line 13 finds the secondary users for primary users in each beam. The primary user makes the pair with a user in the set \( K \setminus K_{\text{PU}} \) with maximum correlation.

An RL-based analog precoder \( F_{AB} \) is designed with the help of primary/strong users’ channel-based channel matrix \( \tilde{H} \). Since the number of users is greater than the number of beams, Card(B) RF chains can only produce \( N_{\text{RF}} \) beams. Therefore, we find the equivalent channel with size of \( N \times \text{Card}(B) \). We use the strongest user-based equivalent channel [7] because, in each beam, the strongest user has to perform SIC to decode the signals of all the other users. The \( N \times \text{Card}(B) \) channel is given by

\[
\tilde{H} = [h_{st,1}, \ldots, h_{st, \text{Card}(B)}],
\]

where \( h_{st,n} \) is the channel vector of strongest user in the \( n^{th} \) group.

Though the use of the strongest user-based equivalent channel reduces the intra-group interference, there is significant inter-group interference. We propose an RL-based selected beam from the DFT-based orthogonal analog precoder to minimize inter-group interference. The detailed design of RL-based beam selection will be presented in the next section. In addition to RL, we use Hungarian, Gale-Shapley, Greedy search, and mutual information-based beam selections.

B. HUNGARIAN-BASED BEAM-GROUP SELECTION
When the unbalanced assignment problem of Card(B) user groups and \( N \) beams is transformed to the balanced assignment problem by adding \( N - \text{Card}(B) \) edges of cost zero. The balanced assignment problem can be solved by the well-known Hungarian method with polynomial complexity of \( O((N/\lambda)^3) \) [23].

C. GALE-SHAPLEY STABLE MATCHING FOR BEAM-GROUP SELECTION
The Gale-Shapley stable matching (or deferred acceptance) algorithm [2] finds stable match between two equal sets with an equal number of elements. Each element of a set has a preference order in the other set. It has the complexity of \( O((N/\lambda)^N) \).

D. GREEDY SEARCH-BASED BEAM-GROUP SELECTION
In this method, users selection with each beam within each beam combination is done iteratively on the basis of the maximum sum-rate [2]. In this way, each beam combination has a user combination set. After that, the beam combination which has the maximum sum-rate is chosen along with the user set. The computational complexity of this method is \( O(N(N - 1)K) \).

E. MUTUAL INFORMATION BASED BEAM-GROUP SELECTION
This method selects \( K \) out of \( N \) beams by maximum relevance and minimum redundancy principle [9]. Specifically, it calculates the correlation among beams. Between two beams with the highest correlation (max redundancy), the beam which has low energy (low relevance) is eliminated iteratively till \( K \) beams are left. The computational complexity of the mutual information-based method is \( O(N(N - 1)K) \).

F. DIGITAL BEAMFORMING
The reduced dimension digital precoder requires Card(B) RF chains, where \( \text{Card}(B) \leq N_{\text{RF}} < N \). The digital precoder is a ZF precoder to eliminate the inter-group interference by taking the strongest user channel as the equivalent channel gain of the group. Since the grouped users have a high correlation, therefore, the inter-group interference for weak channel gain users can be easily minimized [12]. Finally, the optimal power allocation has been done for multiple users using the technique in [13].

V. DEEP REINFORCEMENT LEARNING BASED BEAM SELECTION AND HYBRID BEAMFORMING DESIGN
In this section, we briefly describe reinforcement learning, then the deep reinforcement learning, deep Q-network, and hyperparameters for our beam selection problem are discussed.

A. REINFORCEMENT LEARNING
In the simplest form, reinforcement or Q learning is a learning of an entity called an agent by interaction with the environment as shown in Fig. 3. For a give state \( s_t \) agent performs an action \( a_t \) on the environment. In return, environment produces reward \( r_{t+1} \) and new state \( s_{t+1} \). Thus we get the sequence of state, action, reward as \( s_0, a_0, r_1, s_1, a_1, r_2, \ldots \). Reinforcement or Q-learning is a value iteration method based on...
Algorithm 1 User-Grouping and Selected Beam-Group-Based Hybrid Beamforming

1: Input: \( H \in \mathbb{C}^{N \times K} \), \( \epsilon_{low} \)

   \{User-Grouping\}

2: \( G = \| H \| \)

3: \( K_{\text{sort}} = \{ \| g_i \| \| g_i \| > \| g_{i+1} \| \forall k \in K \} \)

4: \( C = \{ C_{i,j} | C_{i,j} = \frac{\| g_i^k \|}{\| g_i \|} \forall i, j \in \{1, \ldots, K \} \} \)

5: \( C' = \text{complementOones}(C) \)

6: \( C'_{\text{max}} = \arg \max_k C' \forall k \in K \)

7: while \( i < N_{RF} \) do

8: \( i + + \)

9: if \( C'_{\text{max}}(i) < \epsilon_{low} \) then

10: \( K_{PU} = K_{PU} \cup i \)

11: end if

12: end while

13: for \( i = 1 \) to \( K - \text{Card}(K_{PU}) \) do

14: \( \hat{k} = \arg \max_k \left( \frac{\| h_i^{k} \|}{\| h_i \|} \right) \forall j \in K_{PU} \)

15: \( K_{SU} = K_{SU} \cup \hat{k} \)

16: \( K_{PU} \leftarrow K_{PU} \setminus \hat{k} \)

17: end for

\{Beam Selection and Precoder Design\}

18: \( H_b = \text{dft}(N) \times H \in \mathbb{C}^{N \times K} \)

19: \( \hat{H}_b = H_b(c, u) | u \in K_{PU} \in \mathbb{C}^{N \times \text{Card}(B)} \) (user selection)

20: \( H_{b,\text{eq}} = \hat{H}_b(b, :) | b \in B \in \mathbb{C}^{ \text{Card}(B) \times \text{Card}(B)} \) (beam selection)

21: \( F_{AB} = U(b, :) | b \in B \)

\{Digital Precoder Design\}

22: \( F_{DB} = \hat{H}_{b,\text{eq}} \left( \hat{H}_{b,\text{eq}}^H \hat{H}_{b,\text{eq}} + I_{\text{Card}(B)} \right)^{-1} \)

23: Output: \( K_{PU}, K_{SU}, F_{AB}, F_{DB} \)

FIGURE 3. Agent and Environment interfacing in reinforcement learning.

empirical data to discover the best policy. It works as a series of activities in order to increase the expected cumulative reward in a long-term period. And this long-term expected reward is known as Q-function, which is the sum of the discounted rewards received when an action is performed at the initial state. The objective of reinforcement learning is to find the sequence of the actions (a policy) that maximize the total reward. In general Q-function is the mapping of two variables, \( s \) and \( a \) as \( s = f(a) \). In literature, this mapping often expressed as \( s = \pi(a) \), where \( s \in S \) is a state from state space \( S \) and \( a \in A \) is an action from action space \( A \). Conventionally, this mapping is from a table called Q-table, in which rows belong to states and columns belong to actions. For a given state, an action is selected which maximizes the Q-value in the table. For example, in Fig. 4, for state \( s_2 \), the action \( a_1 \) will be selected because \( \arg \max_{a \in A} Q(s, a) \).

The values inside the table are called Q-values or long-term reward values, \( Q: S \times A \rightarrow \mathbb{R} \). Initially, all values are taken as zero or any other random values, and then these Q-values are iteratively obtained by the application of the Bellman equation:

\[
Q^{new}(s_t, a_t) = Q^{old}(s_t, a_t) + \alpha (r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q^{old}(s_t, a_t)),
\]

where \( \alpha \) is the learning rate (0 \( < \alpha \leq 1 \)) and \( \gamma \in [0, 1] \) is the discount factor. When \( s_{t+1} \) reaches the last or terminal state, an episode of the training algorithm ends. The learning rate \( \alpha = 0 \) corresponds to ‘noting learning new’ or exploiting the old knowledge, whereas \( \alpha = 1 \) means pure exploration. Generally, at the start of training, we set \( \alpha = 1 \) and then gradually decrease its value up to 0. The discount factor \( \gamma \) ensures less weight for future rewards. The solution of control problems using the Bellman equation is called dynamic programming. In the case of the discrete domain, the Bellman equation solution is called Markov Decision Processes. The computational complexity grows exponentially with the dimensions of state or action, which is known as the “curse of dimensionality”.

In order to solve the “curse of dimensionality” problem, deep reinforcement learning plays its role. Deep neural networks (DNNs) use a function approximator to replace the Q-table-based state and action mapping function. The DNN takes the state as input and gives action as output. But the training of this DNN requires samples and target data as in any other DNN. As we know, in neural network (NN), the loss or error function \( L = \sum (y_i - \hat{y}_i)^2 \) minimizes over the weights of the network (i.e., differentiate w.r.t weights and set to zero). Here, \( y_i \) is the target value of \( i^{th} \) sample and \( \hat{y}_i \) is the predicted value. In the Bellman equation, the term \( r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) \) is the target value and
Q_{old}(s_t, a_t) is the predicted value. In a neural network, the loss function is used to adjust the weights of NN. If we use the same NN for getting the predicted value at one time stamp and the target value at the next time stamp, we cannot minimize the difference between the two values. Hence, we use two separate NNs, Q-network and target network for predicted Q-value and target Q-value, respectively. The target network is a clone of the Q-network. The weights of the target network are updated from the weights of Q-network after certain time steps. This target network update frequency is also a hyperparameter which needs to be set according to the requirements.

A Q-value function is an estimation of how good it is to perform a given action in a given state [24, A3.5].

The Q-function is given by the below equation:

\[
Q\pi(s, a) = r(s, a) + \gamma \sum_{s' \in S} \sum_{a' \in A} P_{ss'}^{a} Q\pi(s', a')
\]  

(21)

In this equation, \(a\) represents the action, \(s\) is state, and \(\gamma \in [0, 1]\) is the discount factor which determines the weight of the future rewards, \(P_{ss'}^{a}\) is the probability of the state transition from state \(s\) to state \(s'\) when action \(a\) is taken. After an agent executes the action at the state the new state-action pair becomes \((s', a')\). Moreover, the aim of the agent is to maximize the long-term cumulative reward \(Q_\pi(s, a)\) by finding the optimal policy \(\pi^*(s)\). And to do that we can write the previous equation as:

\[
Q_{\pi^*}(s, a) = r(s, a) + \gamma \sum_{s' \in S} \sum_{a' \in A} P_{ss'}^{a} Q_{\pi^*}(s', a')
\]  

(22)

In addition, the optimal policy is given as:

\[
\pi^*(s) = \arg \max_{a \in A} \{Q_{\pi^*}(s, a)\}, \forall s \in S
\]  

(23)

However, these equations cannot be calculated directly because the agent isn’t aware of the state transition probability. Hence, Q-Learning Algorithm is used to deal with this problem. It constructs a Q-table with Q-values \(Q(s, a)\) as elements, and in order to select an action for each state, the agent adopts an \(\epsilon\)-greedy algorithm and updates each element in the Q-table by using the equation below:

\[
Q(s, a) \leftarrow (1 - \alpha) + \alpha \left[ r(s, a) + \gamma \max_{a' \in A} Q(s', a') \right]
\]  

(24)

where \(\alpha\) represents the learning rate. However, the Q-learning algorithm performance relies on the space size of the state-action. It is easier for the agent to find the optimal action policy when the state-action space is small. But, when the size of the state-action space becomes larger, the performance of the Q-learning algorithm becomes limited because the agent may not explore all the state-action space.

### B. DEEP REINFORCEMENT LEARNING

As mentioned above, the agent can be replaced by a deep Q-learning network in order to deal with the drawback of the Q-learning algorithm. The Deep Q-Learning uses a small DQN instead of the large Q-table and stores only the weights of the DQN in local memory. In other words, the deep Q-network uses the input and output method, the input is the state-action pair, and the output is the Q-values. In consequence, the enhancement of the Q-function \(Q(s, a)\) in the Q-learning algorithm is equivalent to the enhancement of the set of weights in the deep neural network (DNN) \(Q(s, a; \Theta)\). To stabilize the learning, deep reinforcement learning adopts two specialized DQN: a Q-network \(Q(s, a; \Theta_Q)\) and a target network \(Q(s, a; \Theta^Q)\) along with an experience replay memory as shown in Fig. 6. Instead of training the DQN with one experience, the agent samples a group of batches of a random size from the replay memory for batch training of the Q-network. The weights \(\Theta_Q\) of the target network are updated periodically with the weights \(\Theta_Q\) of the Q-network according to a predefined hyperparameter ‘update frequency’. Mathematically, the loss function is given by:

\[
L(\Theta) = [y - Q(s, a; \Theta)]^2
\]  

(25)

where \(y\) represents the output (target value) of the target DQN, which given by:

\[
y = r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \Theta^-)
\]  

(26)

The gradient descent method is used to minimize the loss and get the corresponding weights of the Q-network

\[
\Theta \leftarrow \Theta - [y - Q(s, a; \Theta)]\nabla Q(s, a; \Theta),
\]  

(27)

where \(\nabla\) is the gradient operator.

1) ACTOR-CRITIC

In deep reinforcement learning, a DNN that implements the policy \(\pi\) is called an actor. An actor \(\pi(s; \Theta_\pi)\) selects an action deterministically/stochastically based on the input state \(s\) without consulting a value function. During the training, the actor tunes the weights to maximize the Q-value. The critic \(Q(s, a; \Theta_Q)\) is a DNN which estimates the long-term expected reward (Q-value) for a given state \(s\) and a given discrete action \(a\). For a low-dimensional action space, the critic can be used stand-alone. But for the larger dimension action space, actor-critic architecture is more efficient because in this case, the critic gets only one input action at one time stamp. The actor and critic approximators could be a DNN, basis function, or a lookup table. In all cases, the critic is used to learn the policy parameters weights as shown in Fig. 5. The critic network Q-value is compared with the reward to calculate the loss function, which is, then, used to update the weights of the critic as well as the actor.

We use a model-free and off-policy deep Q-Network (DQN) based agent. It consists of critics only and provides a value-based output. During the training, the DQN explores the action space with a given exploration probability epsilon (\(\epsilon\)). At each time step, either it selects a random action with probability (\(\epsilon\)) or follows the value function to determine the action with probability \((1 - \epsilon)\). The weights of DQN are updated after every mini-batch samples. These mini-batches are taken randomly from the experience buffer. The DQN
consists of a Q-network and a target network, as shown in Fig. 5 and Fig. 6.

A complete picture of actor-critic with Q-network and target network is shown in Fig. 6.

2) STATE AND OBSERVATION

The state space is derived from the channel state information $\mathbf{H} \in \mathbb{C}^{N \times K}$. The state space $\mathcal{S} \in \mathbb{R}^{2 \times N \times K}$ consists of i) $N \times K$ absolute squared real values of channel matrix and ii) $N \times K$ binary selection matrix $\mathbf{X}$. The binary element $x_{n,k} = 1$ if user $k$ is selected in beam $n$. At any time $t$, the state is represented by a $N \times 2$ tensor with $N \times 1$ user channel vector absolute squared values and $N \times 1$ binary tensor to indicate the selected beam in the previous time step.

3) ACTION

The action space $\mathcal{A}$ is discrete and it consists of $N \times 1$ beam selection tensors, $\mathcal{A} = \{1, 2, \ldots, N\}$. At any time $t$, only one beam is selected by the agent. In an episode, there are total of $\text{card}(\mathcal{B})$ time steps, so total of $\text{card}(\mathcal{B})$ beams are selected out of $N$ beams. Once we have trained policy $\pi^*(s)$, action can be found as

$$a_t = \max_{a} Q_{\pi^*}(s_t, a; \Theta)$$

4) REWARD

The reward function consists of two components, i) the channel element-based information rate $I(t)$ as a reward, and ii) the penalty $\vartheta(t)$ to avoid the same beam selection within an episode. Mathematically, these functions are written as

$$I(t) = \log_2(1 + |h_{n,k}|^2),$$

and

$$\vartheta(t) = w[(a_{t-1} == a_{t-2}) \lor (a_{t-1} == a_{t-3}) \lor \ldots \lor(a_{t-1} == a_{t-\text{card}(\mathcal{B})})]$$

where $w$ is the weight parameter for the penalty. Thus, the total reward is given by

$$r_t = I(t) - \vartheta(t).$$

Fig. 7 shows the reward function implementation.

C. DRL AGENT HYPERPARAMETERS FOR BEAM-GROUP SELECTION

1) EXPERIENCE REPLAY BUFFER

During the training of DQN, we calculate target value as in (20) and the loss function in (25). These expressions require $(s_t, a_t, s_{t+1}, r_t)$ information tuple $(s_t, a_t, s_{t+1}, r_t)$. This information is stored in the experience buffer. The agent takes a random mini-batch from the experience buffer for an episode of training.

2) EPSILON GREEDY EXPLORATION ($0 \leq \epsilon \leq 1$)

In order to incorporate a suitable trade-off between exploration and exploitation of action space, we use epsilon greedy hyperparameter ($0 \leq \epsilon \leq 1$). The agent opts exploration (i.e., random selection of action) with probability $\epsilon$ and exploitation (i.e., determining the action by (28)) with probability $1 - \epsilon$.

3) DISCOUNT FACTOR ($0 \leq \gamma \leq 1$)

The agent’s goal is to maximize the expected cumulative reward within an episode. The expected reward at time $t$ is the reward at time $t$ and all the future rewards till the terminal time step. Since the future reward has less weight as compared to the present reward, therefore, the agent tries to select actions so that the sum of discounted rewards over the future is maximized [24].

$$Q_t = r_t + \sum_{i=1}^{N_m} \gamma^i Q_{t+i},$$

where $\gamma$ is the discount factor, and $N_m$ is the number of time steps from the present time to the terminal time step. If $\gamma = 0$, the agent is concerned with maximizing the immediate reward irrespective of the future rewards.

4) LEARNING RATE ($0 < \alpha < 1$)

Learning rate ($\alpha$) is used to control the step size during the learning of neural network weights. A too small value of $\alpha$ increases the training time, and a too large value results in a suboptimal trained network.

VI. POWER ALLOCATION

There are two levels for power allocation: power allocation for beams and the power allocation to users within each beam. Since each beam contains multiple users and not necessarily an equal number of users, the total power is divided among beams according to $P_n = P_{\text{tot}} \frac{\text{Card}(K_n)}{K}$.

Within a beam, a user performs the SIC decoding of all the low channel gain users to subtract those users’ signals and treats the higher channel gain users’ signals as intra-beam interference.

Though optimal power allocation has been derived for multiple users [13], [14] but for a single channel. Multi-channel NOMA optimal power allocation for two users per channel is given in [15]. We consider a general case of $\text{Card}(\mathcal{B})$ beams and $\text{Card}(K_n)$ users in the beam $n$, where $n = 1, \ldots, \text{Card}(\mathcal{B})$. The power allocation problem is given as

$$\mathcal{P}1 : \max_{\mathbf{P}} \sum_{n=1}^{\text{Card}(\mathcal{B})} \sum_{k \in K_n} \sum_{p_k \leq P} R_{k,n}(p_{k,n})$$

s.t. $\sum_{k=1}^{N} \sum_{n=1}^{\text{Card}(K_n)} p_{k,n} \leq P$ (33a)

$$P_{1,n} < P_{2,n} < \ldots < P\text{Card}(K_n),$$

$\forall n \in \mathcal{B}$ (33b)

$$R_{k,n} \geq R_{k,n}^{\text{min}}, \quad \forall n \in \mathcal{B}, k \in K_n.$$ (33c)

The optimization problem $\mathcal{P}1$ is non-convex problem due to inter-user interference. First, we decompose the opt-
mization problem into $N$ sub-problems and then maximize each sub-problem independently. This is equivalent to the maximization of the original problem $\mathcal{P}1$. We use transformation of optimization variables for the beam $n$.
as, \( q_{k,n} = \sum_{j=1}^{k} p_{j,n} \), where \( k = 1, \ldots, \text{Card}(K_n) \) and \( n = 1, \ldots, \text{Card}(B) \). This results in the following
\[
p_{k,n} = \begin{cases} q_{k,n} & k = 1 \\ q_{k,n} - q_{k-1,n} & k = 2, \ldots, \text{Card}(K_n) \end{cases}.
\]
(34)
Therefore, the sum-rate of user \( k \) in the beam \( n \) can be written as
\[
R_{k,n} = \begin{cases} \log(1 + q_{k,n} \Gamma_{k,n}) & k = 1 \\ \log(1 + q_{k,n} \Gamma_{k,n}) - \log(1 + q_{k-1,n} \Gamma_{k,n}) & k = 2, \ldots, \text{Card}(K_n) \end{cases}
\]
(35)
where \( \Gamma_{k,n} \) is the channel to noise power ratio. After a simple rearrange, (35) can be written as in (36), shown at the bottom of the page.

The transformed optimization sub-problem for beam \( n \) is given by
\[
P^2: \text{maximize} \sum_{k \in K_n} R_{k,n}(q_{k,n}) \quad \text{s.t.} \quad \begin{align}
& q_{\text{Card}(K_n),n} \leq P_n \\
& 2 - R_{1,n} - 1 \leq q_{1,n} \leq q_{2,n} - q_{1,n} \leq \ldots \\
& \leq q_{\text{Card}(K_n),n} - q_{\text{Card}(K_n)-1,n} \\
& q_{k-1,n} \leq q_{k,n} \Gamma_{k,n} - 2^{R_{k,n}} + 1 \\
& k = 1, \ldots, \text{Card}(K_n).
\end{align}
\]
(37)
\[
(37a)
(37b)
(37c)
\]
The transformed problem \( P^2 \) has a concave objective function, hence, it can be solved by standard convex optimization techniques [13]. With the condition \( R_{\text{min}}^{\text{nom}} \geq 1 \) for \( k = 2, \ldots, \text{Card}(K_n) \) the optimal solution is given by
\[
q_{k,n}^* = \begin{cases} q_{k,n} \frac{\Gamma_{k,n} - 2^{-R_{k,n}} + 1}{2^{R_{k,n}} \Gamma_{k,n}} & k = 1, \ldots, \text{Card}(K_n) - 1 \\ P & k = \text{Card}(K_n) \end{cases}
\]
(38)
Therefore, the optimal solution to the parent problem \( P^1 \) is,
\[
p_{k,n}^* = \begin{cases} q_{k,n}^* & k = 1 \\ \frac{2^{R_{k,n}} \Gamma_{k,n} - 2^{-R_{k,n}} + 1}{2^{R_{k,n}} \Gamma_{k,n}} & k = 2, \ldots, \text{Card}(K_n) \end{cases}
\]
(39)
\[
R_{k,n} = \begin{cases} \log(1 + q_{k,n} \Gamma_{k,n}) & k = 1, \ldots, \text{Card}(K_n) - 1 \\ \log(1 + q_{\text{Card}(K_n),n} \Gamma_{\text{Card}(K_n),n}) & k = \text{Card}(K_n) \end{cases}
\]
(36)

\section{VII. SIMULATION RESULTS}
This section evaluates the performance of the proposed users grouping and RL-based beam selection for MIMO-NOMA using MATLAB. The proposed scheme is compared with the recent beam selection and power allocation scheme MI-based DFT-NOMA in [9], Gale-Shapley based Stable-matching and greedy-based beam selection in [2], and Hungarian-based beam selection in [23]. To implement NOMA with the beam selections schemes of [2] and [23] we use K-means based NOMA users grouping.

\subsection{A. SIMULATION SCENARIO}
The simulation scenario consists of a single cell with \( K = 16 \) users. The base-station is equipped with \( N = 64 \) antennas and \( N_{RF} = 8 \) RF chains. In the proposed beam selection scheme, we use a maximum of two NOMA users [25]. The ULA has \( \lambda/2 \) inter-element distance, where \( \lambda \) is the transmission wavelength. We use limited multipath mmWave channel model with \( L_p = 5 \) [4], [26]. The transmit SNR is defined as \( \rho = \frac{P}{\sigma^2} \) and the value used is shown in each plot. The low correlation threshold \( \epsilon_{\text{low}} = 0.5 \) to select the groupheads. The QoS requirement is set as target minimum rate \( R_{\text{min}}^{\text{nom}} = 2 \) bits/s/Hz for all users. List of hyperparameters used for DQN is given in Table 2. In the simulation, the results are obtained by averaging over 50 channel realizations.

\subsection{B. LEARNING PERFORMANCE OF THE DEEP Q-NETWORK}
The DQN agent has two inputs (State, Reward) and one output (Action). These inputs are provided by the environment, and action is executed on the environment. We first

\begin{table}[h]
\centering
\caption{DQN hyperparameters values.}
\begin{tabular}{|l|l|}
\hline
Hyperparameters & Value \\
\hline Optimizer of DQN & Adam \\
Sample time & 0.025 sec \\
Learning rate (\( \alpha \)) & 0.01 \\
Discount factor (\( \gamma \)) & 0.99 \\
Batch size & 64 \\
Experience buffer & 20,000 \\
Initial Epsilon (\( \epsilon \)) & 1 \\
Epsilon decay & 0.3 \\
Epsilon minimum & 0.01 \\
Max Episodes & 500 \\
Max Episode length & 8 (For 8 users’ groups) \\
Average Window length & 50 (For average reward) \\
Stopping criteria & Average steps \\
Execution environment & GPU \\
Target smoothing factor & 0.001 \\
Target update frequency & 1 \\
\hline
\end{tabular}
\end{table}
investigate the learning performance of the agent. Fig. 8 shows the episode reward graph. At the start, all weights of DQN are randomly initialized, and the agent starts learning with $\epsilon = 1$ i.e., it takes action randomly and then gradually decreases the exploration according to the epsilon decay policy as shown in Fig. 9. As the training progresses, the agent learns and updates its weights. After 500 episodes, the average reward becomes 6.9529, with the last episode reward of 6.3936. The learning curve is approximately stable after 270 episodes.

Fig. 10 presents a graphical view of the agent inputs and output for an entire length of an episode. Since we choose a sampling time of 0.025 sec, therefore, the x-axis shows 0 to 0.2 sec for an episode length of 8 (corresponds to number of users’ groups). The top-left graph is the action output of the agent. It shows that beam 3 is selected for users’ group 1, beam 14 is selected for users’ group 2, and so on. The bottom-left graph is for the reward input at each time step within an episode. The right-side graph shows the input of $64 \times 8$ absolute squared values of the channel matrix. The legend on the top shows the beam selected for users’ groups. For users’ group 1 (i.e., time step 1 between 0-0.025 sec), beam 3 is selected, and so on. One can notice that beam 3 has less power than beams 14 and 56, but RL agent selects the beam based on the cumulative expected reward.

The effect of various hyperparameters on the learning performance is shown in table 3. We train the agent with 500 episodes and record the average reward over all episodes.

We can see as the learning rate decreases, the average reward increases but at the cost of more learning/training time. Also, the average reward depends more on the present and near future values, as indicated by the higher reward with lower discount factors. The experience buffer length between 10,000 to 50,000 does not affect the reward significantly. When we vary the epsilon decay factor, 0.3 gives the largest reward.

C. PERFORMANCE ANALYSIS OF USER GROUPING AND BEAM SELECTION DESIGN

Fig. 11 depicts the sum-rate performance with various values of transmit SNR. In order to assure the QoS, the minimum
The sum-rate performance has been investigated in Fig. 13, with the increasing number of users. When $N_{RF} = 8$, $N = 64$, and transmit SNR = 10 dB, rate is set as $R_{\text{min}}^\text{pr} = 2$ b/s/Hz. The proposed user grouping and RL-based beam-group selection scheme outperform all the other schemes. Specifically, at 10 dB SNR, the proposed scheme gives 42% better performance than the Stable Matching-based MIMO-NOMA. It can be seen that RL-based-DFT-NOMA becomes more beneficial at low transmit power per beam due to less interference within a beam. We also simulate the RL-beam selection with K-means-based users grouping, whose performance is similar to the Stable Matching-based NOMA. The K-means algorithm takes the number of groups as input but number of users within a group varies depending on the Euclidian distance. Therefore, in K-means grouping, there could be a different number of users. In our proposed users’ grouping algorithm, the maximum number of users within a group is two. This limits the intra-beam interference. The Stable Matching-based beam selection is an exhaustive search-based optimal scheme with very high computational complexity, a low complexity but the suboptimal version is used in greedy-based DFT-NOMA. The MI-based DFT-NOMA scheme first selects the beams by using the maximum relevance and minimum redundancy principle. Then all users calculate their best strongest beam. If a beam is strongest for two or more users, that beam is assigned to those users. Due to first selection of the beams without considering the users’ channel badly affects the performance of this scheme as shown in the graph.

The energy efficiency comparison is shown in Fig. 12. The less interference is due to the maximum of two users in a group. The proposed RL-based DFT-NOMA exhibits superior performance at low SNR values. At 10 dB SNR, we get 42% better EE as compared to the Stable Matching-based scheme. The RL-means DFT-NOMA also performs well for low SNR values, and the performance gap between two RL-based schemes is around 0.0255 bits/s/Joule. It is obvious that transmission power increases the interference and decreases the SE, hence, decreasing the EE as well.

In Fig. 13, the sum-rate performance has been investigated with the increasing number of users. With $K \leq 8$ MIMO-NOMA is equivalent to the OMA because $N_{RF} = 8$, and each beam serves at most one user. Again, three schemes, RL-based DFT NOMA, RL-Kmeans-based DFT NOMA, and Stable Matching-based DFT NOMA, are competitive in this figure. It can be seen that as the number of users
increases, the performance gap of RL-based DFT NOMA with the competitors increases. Due to the K-means-based users grouping, other schemes perform better with less number of users. The mean and standard deviation across the users are (8.739,0.595), (7.833,1.402), and (7.298,1.047) for RL-based DFT-NOMA, RL-kmeans-based DFT-NOMA, and Stable Matching-based DFT-NOMA, respectively. The lowest performer is again MI-based DFT-NOMA. There is a reason for the very low performance of the MI-based DFT-ABF scheme; i.e., it does not cater for the minimum correlation between the selected beams, which induces severe interbeam interference. Fig. 14 depicts the sum-rate versus users with SNR value of 20 dB. It can be noticed that RL-based DFT-NOMA performance w.r.t. other schemes is much better when the number of users approaches 2 × Card(E) and the SNR is between 10-20 dB. In the proposed users grouping scheme, the two users in a group have a large channel gain difference, and the low channel gain user requires more transmission power, which causes more interference at high transmit SNR regimes. In the case of OMA (i.e., K = 8) and NOMA in a few beams (when K = 9,10), there is marginal or no performance gain with RL-based DFT-NOMA. The mean and standard deviation across the users are (12.08,0.97), (11.72,1.5), and (11.57,1.227) for RL-based DFT-NOMA, RL-kmeans-based DFT-NOMA, and Stable Matching-based DFT-NOMA, respectively. This shows less deviation of SE in RL-based DFT-NOMA even with the change in the number of users.

Finally, the sum-rate versus minimum target date-rate is shown in Fig. 15 and Fig. 16 for transmit SNR of 10 and 20 dB, respectively. Due to the suboptimal performance of Hunfarian-, Greedy- and MI-based schemes and to ensure the information-rate well above the $R_{\text{min}}$ for each user, their sum-rate remains at least 4 b/s/Hz lower than the other near optimal schemes. The RL-based DFT-NOMA sum-rate exhibits highest sum-rate performance but decreases more sharply with increasing QoS requirement in the form of $R_{\text{min}}$. This is, to ensure the higher $R_{\text{min}}$ for low channel gain users who requires more power which in turn deprives the power share of the group-head, hence, results in the substantiate decrease in the sum-rate.

VIII. CONCLUSION

In this paper, we propose a novel users’ grouping and reinforcement learning-based beam-user selection for a massive MIMO NOMA downlink system. We use channel correlation and channel gain information for intra-beam and interbeam user selections in the users grouping algorithm. After the users’ grouping, a deep Q-network selects the optimal beams as an action on the basis of CSI-based states and information rate-based reward function. Finally, multi-beam multiuser optimal power is allocated. It has been shown that the proposed RL-based DFT-NOMA outperforms the state-of-the-art Gale-Shapley based Stable Matching, Hungarian-, and MI-based MIMO-NOMA schemes. Specifically, we get 42% increase in SE and EE performance at 10 dB transmit SNR.

As an extension of this work, the performance of the DRL-based design can be examined by using proximal policy optimization (PPO) or trust region policy optimization (TRPO) agents. These agents are policy-based actor-critic agents.

REFERENCES

[1] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. Poorter, and I. Moerman, “A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives,” IEEE Commun. Surveys Tuts., vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.

[2] Z. Cheng, Z. Wei, and H. Yang, “Low-complexity joint user and beam selection for beamspace mmWave MIMO systems,” IEEE Commun. Lett., vol. 24, no. 9, pp. 2065–2069, Sep. 2020.

[3] R. Pal, A. K. Chaitanya, and K. V. Srinivas, “Low-complexity beam selection algorithms for millimeter wave beamspace MIMO systems,” IEEE Commun. Lett., vol. 23, no. 4, pp. 768–771, Apr. 2019.

[4] S. He, Y. Wu, D. W. K. Ng, and Y. Huang, “Joint optimization of analog beam and user scheduling for millimeter wave communications,” IEEE Commun. Lett., vol. 21, no. 3, pp. 2638–2641, Dec. 2017.

[5] X. Gao, L. Dai, Z. Chen, Z. Wang, and Z. Zhang, “Near-optimal beam selection for beamspace mmWave massive MIMO systems,” IEEE Commun. Lett., vol. 20, no. 5, pp. 1054–1057, May 2016.
[6] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, “A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[7] B. Wang and L. Dai, “Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[8] P. Liu, Y. Li, W. Cheng, X. Gao, and W. Zhang, “Multi-beam NOMA for millimeter-wave massive MIMO with lens antenna array,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11570–11583, Oct. 2020.

[9] Y.-T. Chiu and K.-H. Liu, “Beam selection and power allocation for massive connectivity in millimeter-wave NOMA systems,” *IEEE Access*, vol. 8, pp. 5368–5388, 2020.

[10] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X.-G. Xia, “Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2961–2974, May 2018.

[11] F. Xu, X. Yu, M. Li, and B. Wen, “Energy-efficient power allocation scheme for hybrid precoding mmWave-NOMA system with multi-user pairing,” in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Shanghai, China, May 2019, pp. 11–15.

[12] X. Yu, F. Xu, K. Yu, and X. Dang, “Power allocation for energy efficiency optimization in multi-user mmWave-NOMA system with hybrid precoding,” *IEEE Access*, vol. 7, pp. 109083–109093, 2019.

[13] Y. Huang, J. Wang, and J. Zhu, “Optimal power allocation for downlink NOMA systems,” in *Multiple Access Techniques for 5G Wireless Networks and Beyond*, V. Mojtaba, Z. Ding, and H. V. Poor, Eds. Springer, 2019, ch. 6, pp. 195–227.

[14] C.-L. Wang, J.-Y. Chen, and Y.-J. Chen, “Power allocation for a downlink non-orthogonal multiple access system,” *IEEE Wireless Lett.*, vol. 5, no. 5, pp. 532–535, Oct. 2016.

[15] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, “On optimal power allocation for downlink non-orthogonal multiple access systems,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[16] L. Pang, W. Wu, Y. Zhang, Y. Yuan, Y. Chen, A. Wang, and J. Li, “Joint power allocation and hybrid beamforming for downlink mmWave-NOMA systems,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10173–10184, Oct. 2021.

[17] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, “Non-orthogonal multiple access in a downlink multiuser beamforming system,” in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2013, pp. 1278–1283.

[18] F. Sohrabi and W. Yu, “Hybrid digital and analog beamforming design for large-scale MIMO systems,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 2929–2933.

[19] O. E. Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, “Low complexity precoding for large millimeter wave MIMO systems,” in *Proc. IEEE Int. Conf. Commun. (ICCU)*, Ottawa, ON, Canada, Jun. 2012, pp. 3724–3729.

[20] J. Brady, N. Behdad, and A. M. Sayeed, “Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements,” *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.

[21] I. Ahmed, H. Khammari, and A. Shahid, “Resource allocation for transmit hybrid beamforming in decoupled millimeter wave multiuser-MIMO downlink,” *IEEE Access*, vol. 5, pp. 170–182, 2017.

[22] M. Richard Karp, “Reducibility among combinatorial problems,” in *Complexity of Computer Computations*, Boston, MA, USA: Springer, 1972, pp. 85–103.

[23] I. Ahmed, M. K. Shahid, H. Khammari, and M. Masud, “Machine learning based beam selection with low complexity hybrid beamforming design for 5G massive MIMO systems,” *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 4, pp. 2160–2173, Dec. 2021.

[24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[25] J.-H. Tseng, Y.-F. Chen, and C.-L. Wang, “User selection and resource allocation algorithms for multicarrier NOMA systems on downlink beamforming,” *IEEE Access*, vol. 8, pp. 59211–59224, 2020.

[26] Z. Li, S. Han, S. Sangodoyin, R. Wang, and A. F. Molisch, “Joint optimization of hybrid beamforming for multi-user massive MIMO downlink,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3000–3014, Jun. 2018.

---

**IRFAN AHMED** (Senior Member, IEEE) received the B.E. degree in electrical engineering and the M.S. degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 1999 and 2003, respectively, and the Ph.D. degree in telecommunication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008. From 2011 to 2017, he has worked at Taif University as an Associate/Assistant Professor. He was a Postdoctoral Fellow with Qatar University, from April 2010 to March 2011, where he has worked on two research projects, wireless mesh networks with Purdue University, USA, and radio resource allocation for LTE with Qtel. He is currently working as an Associate Professor with the Higher Colleges of Technologies, United Arab Emirates. He has also been involved in the National ICT Pakistan funded research project “Design and development of MIMO and Cooperative MIMO Test-Bed” at Iqra University, Islamabad, Pakistan, from 2008 to 2010. His research interests include wireless LAN (WLAN) medium access control (MAC) protocol design and analysis, 5G communications, mmWave and massive MIMO communications, performance analysis of wireless channels, energy-constrained wireless networks, and radio resource allocation. He has served as the Session Chair for the IEEE Wireless Communications, Networking and Mobile Computing conference held in Shanghai, China, in September 2007, and IEEE ICC 2016. He is an active reviewer of IEEE, Springer, and Elsevier journals and conferences. He is an Associate Editor of IEEE Access journal.

**MUHAMMAD KHALIL SHAHID** (Member, IEEE) received the B.E. degree in electrical engineering from the University of Engineering and Technology, Lahore, in 1998, the M.S. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2005, and the Ph.D. degree in engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008. He is currently working as an Assistant Professor with the Higher Colleges of Technology. He has over 15 years of experience in telecommunication industry. He is an author of several journals and conference papers in the field of communications and information technology. He has worked on LTE MiFi Clouds, Hotspots, WiGig, USB Dongles, and Drive testing for CDMA/EVDO Network, for checking QoS parameters using NEMO Analyzer and Genex Probe. His current research interests include wireless communication, 5G communications, optical wireless communications, fiber optic systems and networks, optical transmission, optical fiber access networks, technology management, operational management, project management, and industrial organization.

**TARIQ FAISAL** received the master’s degree in mechatronics engineering from IJU University, in 2006, and the Ph.D. degree in signal processing from the University of Malaya, Malaysia, in 2011. He has been the Dean of academic operations at the Higher Colleges of Technology, since 2018. He has more than 20 years of academic and industry experience of which he worked as an Engineer, an Assistant Professor, the Programs Chair, the Head of Department, the Division Chair, and the Campus Director. His research interests include biomedical signal processing, intelligent systems, robotics, control, embedded system design, the IoTs, machine learning, and outcome-based education. He has been a reviewer for multiple journals including IEEE, Elsevier, Taylor & Francis, and Springer Nature. He is also a Charted Engineer as well a Senior Fellow of the Higher Education Academy.