

Enhanced User Grouping and Power Allocation for Hybrid mmWave MIMO-NOMA Systems

Jinle Zhu, Qiang Li, Zilong Liu, *Senior Member, IEEE*,
Hongyang Chen, *Senior Member, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

Abstract—Non-orthogonal multiple access (NOMA) and millimeter wave (mmWave) are two key enabling technologies for the fifth-generation (5G) mobile networks and beyond. In this paper, we consider uplink communications with a hybrid beamforming structure and focus on improving the spectral efficiency (SE) and energy efficiency (EE) of mmWave multiple-input multiple-output (MIMO)-NOMA systems with enhanced user grouping and power allocation. Exploiting the directionality feature of mmWave channels, we first propose a novel initial agglomerative nesting (AGNES) based user grouping algorithm by taking advantage of the channel correlations. It is noted that the optimization of the SE/EE is a challenging task due to the non-linear programming nature of the corresponding problem involving user grouping, beam selection, and power allocation. Our idea is to decompose the overall optimization problem into a mixed integer problem comprising of user grouping and beam selection only, followed by a continuous problem involving power allocation and digital beamforming design. To avoid the prohibitively high complexity of the brute-force search approach, we propose two suboptimal low-complexity user grouping and beam selection schemes, the direct AGNES (DIR-AGNES) scheme and the successive AGNES (SUC-AGNES) scheme. We also introduce the quadratic transform (QT) to recast the non-convex power allocation optimization problem into a convex one subject to a minimum required data rate of each user. The continuous problem is solved by iteratively optimizing the power and the digital beamforming. Extensive simulation results have shown that our proposed mmWave-NOMA design outperforms the conventional orthogonal multiple access (OMA) scenario and the state-of-art NOMA schemes.

Index Terms—MIMO, mmWave, NOMA, user grouping, beam selection, power allocation.

I. INTRODUCTION

Non-orthogonal multiple access (NOMA) is an emerging paradigm which can support massive connectivity envisaged in the fifth-generation (5G) networks and beyond [1], [2]. Conventional orthogonal multiple access (OMA) suffers from limited user capacity as multiple users are separated in orthogonal channels [3], [4]. Take uplink power-domain NOMA for

example: multiple users transmit their signals over the same time-frequency resources based on superposition coding [5]–[10]. The messages of the multiple users are decoded at the base station (BS) by leveraging the different allocated power levels with successive interference cancellation (SIC), yielding a higher network capacity without further resource cost.

Millimeter wave (mmWave) communication is another key enabling technology for next generation wireless networks [11]–[13]. The mmWave frequency band ranges from 30 GHz to 300 GHz, where the signals experience an orders-of-magnitude increase in free-space pathloss compared to that in the Sub-6 GHz band. To combat the substantial propagation attenuation in mmWave channels, large antenna arrays can be deployed to attain beamforming [14], [15]. Their short wave lengths also facilitate the use of large array in multiple-input multiple-output (MIMO) systems [16]. Fully digital beamforming (DBF) allows us to control both the phase and the amplitude of a signal. However, DBF requires a dedicated radio frequency (RF) chain for each antenna which could result in tremendous energy consumption and signal processing complexity in the massive MIMO systems equipped with large antenna array [17]. In contrast, analog beamforming (ABF) is attractive for its low complexity. ABF is usually performed through a phase shifter network which places constant modulus constraints on the elements of the ABF matrix [18], [19]. That being said, ABF does not support spatial multiplexing which limits the enhancement of system throughput. To strike a balance between energy consumption and system performance, hybrid beamforming (HBF) has been proposed [20]–[22], in which a small number of RF chains are connected with a large number of antennas for harvest higher amount of multiplexing gain with low-complexity hardware.

The NOMA based mmWave MIMO-HBF systems have been attracting increasing research attention due to the following features: 1) The highly directional channels of mmWave systems facilitate the use of NOMA transmission for multiple users sharing the same beam but with different distances to the BS; 2) Whilst the directional analog beams can enable us to perform NOMA over each beam, the digital beamforming can be designed to combat the inter-beam interference.

A. Prior Works

In contrast to the conventional MIMO-OMA schemes, MIMO-NOMA has shown its promising future in supporting massive connectivity and expanding system capacity. There have been numerous research attempts concerning the appli-

J. L. Zhu and Q. Li are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China (UESTC), e-mails: sohpia_zhujl@163.com; liqiang@uestc.edu.cn.

Z. L. Liu is with the School of Computer Science and Electronics Engineering, University of Essex, UK, e-mail: zilong.liu@essex.ac.uk.

H. Chen is with the Research Center for Intelligent Network, Zhejiang Lab, Hangzhou 311121, China, e-mail: dr.h.chen@ieee.org.

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ, 08544, USA, e-mail: poor@princeton.edu.

Corresponding author is Qiang Li (e-mail: liqiang@uestc.edu.cn).

This work was supported in part by National Key R&D Program of China (No.2018YFC0807101) & National Natural Science Foundation of China (No. 61831004).

cations of NOMA in mmWave communications. [23] provided an in-depth capacity analysis for the integrated NOMA mmWave-massive-MIMO systems. Theoretical analysis and results have validated the significant capacity improvements achieved by NOMA.

In mmWave MIMO-OMA systems, a “virtual sectorization” concept has been proposed in [26], where the users are grouped virtually in the digital baseband stage, followed by a channel-statistics-based analog beamforming scheme [27]. Considering the difficulty of acquiring channel state information (CSI) in realistic mmWave channels, [28] exploited the spatial division and multiplexing (JSDM) algorithm to study the user grouping problem. However, since multiple users are served by one beam in mmWave MIMO-NOMA systems rather than allocated with dedicated RF chain tunnels, user grouping in NOMA schemes is more complicated than that in the OMA systems. [9] investigated the user grouping and power allocation in both the downlink and uplink communication networks. However, as discussed in [24], [25], the user grouping strategy in [9] is based on the channel gain difference which may not be suitable for the mmWave communication systems. In [29], the scheduled users are selected based on a matching theory which can avoid the prohibitively high complexity in exhaustive search. [30] and [31] adopted the same user selection strategy where the two users in a pair have a high channel correlation but large channel gain difference. For the user grouping strategies which aim to serve all users in a system, [32] and [33] discussed the 2-user downlink and uplink mmWave-NOMA system, respectively, in which joint Tx-Rx beamforming and power allocation problems are addressed. However, the design freedom of these 2-user grouping strategies may be limited which could be a barrier for further enhancement of system performance. [34], [35] and [36] extended the user grouping work to K -user NOMA systems. In [34], a cluster-head selection algorithm is proposed to select one user for each beam at first. [35] performed the user grouping based on the K-means algorithm and designed the analog beamforming by a boundary-compressed particle swarm optimization algorithm. By assuming the users are physically clustered, [36] allocated the users with an machine learning framework building upon the K-means algorithm.

As mentioned above, the beamforming design can influence the performance of NOMA in mmWave-HBF systems. The authors in [24] proposed an angle-based user pairing strategy and analyzed the performance where beam misalignment at both the BS and the users is taken into account. In [25], the lower bound for the achievable rate and an upper bound for the sum rate gap expression between the perfectly aligned and misaligned are established. The simulation results validate that beam misalignment can significantly degrade the rate performance in MIMO-HBF-NOMA systems. The employment of ABF is considered in [30]–[33]. In [30] and [31], a predefined DFT codebook is used to perform beam sweeping. After the users are paired, each pair chooses its beam element based on the beam gain. Zero-forcing (ZF) beamforming is implemented at the baseband to combat the inter-cluster interference. [32] and [33] decomposed the formulated joint power control and beamforming problem into two sub-problems: one

for improving the power control and beam gain allocation, and the other for optimization of analog beamforming under a constant-modulus constraint. In [37], a new beamspace-NOMA framework was proposed, in which an equivalent channel hybrid beamforming scheme and an iterative power allocation algorithm are developed. Random beamforming is used to further reduce the feedback overhead in [38] and [29].

B. Motivations and Contributions

This paper is concerned with the setting of an uplink hybrid mmWave-NOMA communication system. We adopt the beam sweeping approach with a prior discrete Fourier transform (DFT) codebook known by the BS and the users. In principle, the channel gain and the beam gain will be used at the BS to determine the decoding order of multiple users clustered over a group. Thus, it is of vital importance to cluster multiple users into different groups (where each group consists of highly correlated users) to suppress the inter-group interference, whilst optimizing the power allocation to maximize the system throughputs. A major objective of this work is to look for enhanced user grouping, beam selection, and power allocation schemes for more efficient mmWave MIMO-HBF-NOMA.

Due to the combinatorial nature of the aforementioned three problems, it is challenging to attain a global optimum solution. The current state-of-the-art mostly advocates the idea of decomposing the entire optimization problem into three separate sub-problems and then sequentially addressing them one by one [29]–[31], [35]. Such a sequential optimization approach may lead to a solution which is far away from the global optimum due to the limited design freedom. Furthermore, the existing algorithms only reduce the inter-group interference at the digital beamforming stage (by implementing ZF beamforming). By considering the fact that there are limited number of propagation paths related to a few scatters in the mmWave communications, it is highly possible that different users share some common scatters. When some users in two (or more) different groups transmit their signals through common scatters, the angles of arrival (AoA) of these signals may be highly correlated, yielding a larger amount of inter-group interference as identical beam elements may be used by different groups. When ZF method is adopted in the baseband DBF, any two identical beam elements in the ABF can result in a rank-deficient DBF matrix. Even if these groups do not necessarily choose an identical beam element, the highly correlated beam patterns under this circumstance could still result in severe inter-group interference. Such a problem is referred to as the beam overlapping problem which has been illustrated in Fig. 1.

To make a difference, we observe that the optimization problem consists of an integer problem involving user grouping and beam selection and a continuous problem posed by power allocation. Our key idea is to solve the integer problem and the continuous problem separately. For the joint user grouping and beam allocation integer problem, exhaustive search is infeasible due to high computational complexity. In order to reduce the computational complexity whilst maintaining the

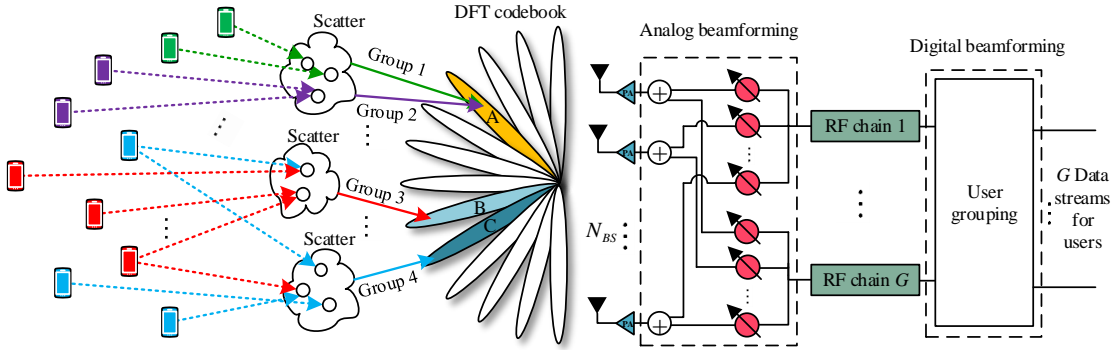


Fig. 1. System model of uplink mmWave MIMO-HBF-NOMA communications and illustration of the beam overlapping problem: 1) The green group (Group 1) and the purple group (Group 2) share the same scatterer and hence select the same beam pattern A, and thus the digital beamforming cannot separate the signals of Group 1 and Group 2; 2) The red group (Group 3) and the blue group (Group 4) choose the two highly correlated beams (B and C), which may produce significant interference to each other.

design freedom of the joint problem, we propose two low-complexity joint user grouping and beam selection schemes which are capable of circumventing the beam overlapping problem in different ways.

In view of the characteristics of mmWave channels, channel correlation¹ is proved to be a major criteria for user grouping in mmWave-NOMA such as the K-means user grouping algorithms in [35] and [36]. However, it is known that the performance of the K-means algorithm is heavily dependent on its initial value. Since the initial cluster-head in each group is randomly selected in [35] and [36], an improper initial user could significantly affect the system performance. Aiming for creating a group in a spontaneous way, we propose a novel user grouping strategy based on the agglomerative nesting (AGNES) clustering algorithm.

For the continuous problem on power allocation, we consider to optimize the system spectral efficiency (SE) and energy efficiency (EE) with the aid of quadratic transform (QT) [39]. Perfect channel state information (CSI) is assumed to be known at the BS and the users. The main contributions of this paper are summarized as follows:

- We develop a joint optimization framework for the mmWave MIMO-HBF-NOMA system, where joint user grouping and beam selection scheme and power allocation strategy are introduced to combat the overlapping beam problem whilst improving the system performance.
- Aiming to mitigate the inter-group interference in both digital and analog beam stage, we first propose an initial user grouping algorithm based on the AGNES clustering algorithm, and then develop two schemes for the integer problem which are the direct AGNES (DIR-AGNES) scheme and successive AGNES (SUC-AGNES) scheme.
- We first formulate out the maximization problems of SE and EE, then recast them with the aid of QT for solvable problems. We devise an iterative approach to obtain the optimal power allocation strategy and the digital beamforming design. Extensive simulation results under the generic mmWave channels verify the validity of

our proposed schemes over the state-of-the-art schemes under various typical parameter settings.

C. Organization and notations

The remainder of this paper is organized as follows. In Section II, we describe the system model and formulate the problems for the mmWave-NOMA communications. Our proposed AGNES user grouping algorithm is introduced in Section III. In Section IV, two user grouping and beam selection schemes are proposed. Section V introduces the power allocation algorithm. In Section VI, we summarize the proposed algorithms and analyze the computational complexity. In Section VII, simulation results are given to demonstrate the performance. Finally, Section VIII concludes this paper.

Notation: The following notations will be used throughout this paper: Upper-case and lower-case boldface letters denote matrices and vectors, respectively; $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and the Hermitian transpose of a matrix or a vector. \mathcal{S} denotes a set; $|\cdot|$ denotes the absolute value of a scalar or the cardinality of a set; $\|\cdot\|_2$ denotes the Frobenius norm of a vector or a matrix. $\mathbb{C}^{M \times N}$ denotes the set of all $M \times N$ matrices with complex entries. $\mathbb{E}\{\cdot\}$ denotes the expectation operation.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider an uplink mmWave MIMO-NOMA transmission scenario, where a BS communicates with K users. The user set is denoted as $\mathcal{U} = \{U_1, U_2, \dots, U_K\}$. Before the implementation of user grouping, the k^{th} user is denoted as U_k . The BS is equipped with N_{BS} antennas and N_{rf} RF chains and each user is installed with a single antenna² ($K > N_{rf}$). G data streams can be supported by the BS. To obtain a higher multiplexing gain, we assume that the number of the RF chain N_{rf} is equal to the number of the data streams G , i.e., $N_{rf} = G$. The hybrid structure of the BS and the system model of this paper are illustrated in Fig. 1.

¹The channel correlation in this paper only involves the angle correlation and has no relationship with the channel gain.

²The considered model can be easily generalized to the case where the users have different antennas, which will be specified in Section VI.

To perform NOMA, the K users are divided into G clusters served by the G RF chains with each cluster mapping to a dedicated data stream. The detailed user grouping and power allocation process will be described in the following sections. After the user grouping, the g -th user set is denoted as $\mathcal{S}_g = \{U_{g,1}, U_{g,2}, \dots, U_{g,|\mathcal{S}_g}|\}$ where $U_{g,u}$ presents the u -th user in the g -th cluster. In our considered scenario, all users communicate simultaneously and each user is served by one single cluster only. Thus, we have $\sum_{g=1}^G |\mathcal{S}_g| = K$ and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for $i \neq j$. $U_{g,u}$ transmits its signal $x_{g,u}$ with the allocated power $P_{g,u}$. The BS will receive the signals from all users which can be presented as

$$\mathbf{r} = \sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} \sqrt{P_{g,u}} \mathbf{h}_{g,u} x_{g,u} + \mathbf{n}, \quad (1)$$

where $\mathbf{h}_{g,u} \in \mathbb{C}^{N_{BS} \times 1}$ denotes the channel matrix of $U_{g,u}$. The transmitted signal $x_{g,u}$ satisfies $\mathbb{E}\{|x_{g,u}|^2\} = 1$. $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I})$ denotes the $N_{BS} \times 1$ Gaussian noise vector corrupting the received signals. After receiving the signals from the users, the BS applies an $N_{BS} \times G$ analog RF combiner $\mathbf{F}_{RF} = [\mathbf{f}_1^{RF}, \mathbf{f}_2^{RF}, \dots, \mathbf{f}_G^{RF}]$. The analog beamformer is realized by a phase shifter network. Thus, each element of \mathbf{F}_{RF} is constrained by the constant modulus (CM) value, i.e., $|\mathbf{F}_{RF}^{(i,j)}|^2 = \frac{1}{N_{BS}}$. In our scenario, to reduce the hardware processing complexity and the feedback overhead, we consider the beam sweeping approach, in which each column of \mathbf{F}_{RF} is chosen from a predefined DFT codebook \mathcal{F} [22]. The DFT codebook \mathcal{F} is formed by N_{beam} bases where each base is an array response vector given by

$$\mathbf{a}(N, \xi) = \frac{1}{\sqrt{N}} \left[1, e^{j \frac{2\pi d}{\lambda} \cos(\xi)}, \dots, e^{j \frac{(N-1)2\pi d}{\lambda} \cos(\xi)} \right]^T, \quad (2)$$

where λ is the wavelength and $d = \frac{\lambda}{2}$ denotes the antenna spacing. We discretize the angle ξ into N_{beam} levels over $[0, 2\pi)$. The DFT codebook is expressed as

$$\mathcal{F} = [\mathbf{a}(N_{BS}, \varrho_1), \mathbf{a}(N_{BS}, \varrho_2), \dots, \mathbf{a}(N_{BS}, \varrho_{N_{beam}})],$$

with $\varrho_i = \frac{2\pi(i-1)}{N_{beam}}$. Then, the BS implements a $G \times G$ digital combiner $\mathbf{F}_{BB} = [\mathbf{f}_1^{BB}, \mathbf{f}_2^{BB}, \dots, \mathbf{f}_G^{BB}]$ to process the baseband signals. The processed received signal at the BS is given by

$$\mathbf{y} = \mathbf{F}_{BB}^H \mathbf{F}_{RF}^H \sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} \sqrt{P_{g,u}} \mathbf{h}_{g,u} x_{g,u} + \mathbf{F}_{BB}^H \mathbf{F}_{RF}^H \mathbf{n}. \quad (3)$$

Due to the inefficiency of diffraction as propagation process, the number of significant multipath components may be reduced and spatial selectivity are limited. The small number of the multipath components (MPC) leads to high directionality and spatial sparsity in the angle domain. We use the widely adopted double directional channel model [26] as the considered mmWave channel model. In this channel model, the uplink channel matrix of $U_{g,u}$, $\mathbf{h}_{g,u} \in \mathbb{C}^{N_{BS} \times 1}$, is assumed to be a sum of the contributions of the scattering propagation paths as

$$\mathbf{h}_{g,u} = \sqrt{\frac{N_{BS}}{L_{g,u}}} \sum_{l=1}^{L_{g,u}} \alpha_{g,u,l} \mathbf{a}_{BS}(N_{BS}, \theta_{g,u}), \quad (4)$$

where $L_{g,u}$ is the number of the propagation paths of $U_{g,u}$ and $\alpha_{g,u,l}$ denotes the channel gain of the l^{th} path which is independently and identically Gaussian distributed with zero mean and variance of 1. Uniform linear arrays (ULAs) are applied at the BS and $\mathbf{a}_{BS}(N_{BS}, \theta_{g,u})$ is the normalized receive array response vectors with AoA $\theta_{g,u} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. In this paper, we discuss over the flat-fading mmWave channel, but it is noted that our proposed algorithms may also work in frequency-selective channels with second-order statistics CSI, such as in [26] and [27].

B. Problem Formulation

Assume that we have finished user grouping and hybrid beamforming design for the groups, according to the uplink MIMO-NOMA technique [9], each user in a group suffers from the intra-group interference and the inter-group interference. The channel gains and beam gains of the users in the same group are key to decide the decoding order at the BS when implementing SIC. Without loss of generality, we sort the users with their channel and beam gains, i.e., $\|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{g,1}\|_2 \geq \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{g,2}\|_2 \geq \dots \geq \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{g,|\mathcal{S}_g}|\|_2$ for $g = 1, \dots, G$. The SIC decoding is performed to decode the strongest users' signals first by viewing the signals of the other users as interference. Assuming a perfect decoding, the receiver then recovers the signals of the strongest user which will be subtracted from the received signals when we decode the the remaining relatively weak users. This process is successively performed until all the users are decoded. After applying the SIC decoding rule, the signal to interference plus noise power ratio (SINR) of $U_{g,u}$ is given by

$$SINR_{g,u} = \frac{\|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{g,u}\|_2^2 P_{g,u}}{I_{g,u}^{intra} + I_{g,u}^{inter} + \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H\|_2^2 \sigma^2}. \quad (5)$$

The numerator of (5) represents the desired signal gain. The first term $I_{g,u}^{intra}$ in the denominator denotes the intra-group interference as

$$I_{g,u}^{intra} = \sum_{v=u+1}^{|\mathcal{S}_g|} \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{g,v}\|_2^2 P_{g,v}, \quad (6)$$

whereas the second term $I_{g,u}^{inter}$ denotes the inter-group interference as

$$I_{g,u}^{inter} = \sum_{q \neq g} \sum_{v=1}^{|\mathcal{S}_q|} \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{q,v}\|_2^2 P_{q,v}. \quad (7)$$

Thus, the average achievable data rate of $U_{g,u}$ can be expressed as

$$R_{g,u} = \log_2(1 + SINR_{g,u}). \quad (8)$$

In this paper, we aim for optimizing SE and EE, respectively. The sum data rate of the system is given by

$$SE = \sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} R_{g,u}. \quad (9)$$

$$\mathcal{L}(\mathcal{S}_{i,j}, \mathcal{S}_q) = \frac{(|\mathcal{S}_i| + |\mathcal{S}_q|)\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q) + (|\mathcal{S}_j| + |\mathcal{S}_q|)\mathcal{L}(\mathcal{S}_j, \mathcal{S}_q) - |\mathcal{S}_q|\mathcal{L}(\mathcal{S}_i, \mathcal{S}_j)}{|\mathcal{S}_i| + |\mathcal{S}_j| + |\mathcal{S}_q|}. \quad (15)$$

The EE of the system is given by

$$EE = \frac{R_{sum}}{\xi P_{sum} + P_C}, \quad (10)$$

where ξ denotes a constant of the inefficiency of the PA and P_C denotes the fixed power consumption of the system [30].

Note that the SE or the EE performance of the system is determined by user grouping, hybrid beamforming, and power allocation strategy, which is challenging to analyze. The main idea of this paper is to divide each coupled problem into an integer problem and a continuous problem respectively. We first consider the joint user grouping and beam selection integer problem. Before proceeding to this, we first derive a novel initial user grouping algorithm.

III. AGGLOMERATIVE NESTING USER GROUPING

In a mmWave MIMO-HBF-NOMA communication system, the users in the same group obtain their beam gain by the same beam pattern while different groups are distinguished by different beams. Having this in mind, we propose an intuitive algorithm where the users with a high channel correlation are clustered to a same group to achieve a high beam gain and the users whose channels are weakly correlated are allocated to different groups to suppress the interference. Besides, in contrary to the K-means algorithms in [35] and [36], the proposed user grouping algorithm enable the users to form clusters spontaneously at once without iteration process.

We use the AGNES algorithm to perform the user grouping. The AGNES hierarchical clustering is a tree structure which is able to form the groups spontaneously for high intra-group similarity and low inter-group similarity. The ‘‘similarity’’ of the users is referred to as the AoA similarity in the angle domain rather than the geographical distance in our scenario. The angle similarity can be measured by the channel correlation value \mathcal{L} . Due to the high spacial directivity of the mmWave channel, the similarity between U_k and U_l is defined as the channel correlation [36]:

$$\mathcal{L}(k, l) = \frac{|\mathbf{h}_k \mathbf{h}_l^H|}{|\mathbf{h}_k| |\mathbf{h}_l|}, \quad (11)$$

where \mathbf{h}_i is the channel vector of the i -th user ($i = 1, 2, \dots, K$).

Remark 1: The hierarchical clustering algorithm can recursively partition the users in either a top-down or bottom-up fashion. In our previous work [40], we consider the bottom-up fashion because the number of users in [40] is not significantly larger than the number of RF chains.

We take the bottom-up fashion as an example to introduce the AGNES user grouping algorithm with which the top-down fashion can be derived by the opposite process. Each user initially belongs to a group of its own, then the groups are successively merged into new groups based on the predefined criteria until the desired group number (G) is reached. The criteria for generating new groups depends on the linkage

method [41]. Typical linkage methods include: single linkage, complete linkage, average linkage, ward linkage and centroid linkage which are explained as follows.

We introduce the linkage methods in an inductive manner [42] because only two user groups are merged at each step. Suppose that $\mathcal{S}_{i,j}$ is the user group merged from \mathcal{S}_i and \mathcal{S}_j , namely, $\mathcal{S}_{i,j} \triangleq \mathcal{S}_i \cup \mathcal{S}_j$. Let \mathcal{S}_q be one of the remaining groups except for \mathcal{S}_i and \mathcal{S}_j . The single linkage between $\mathcal{S}_{i,j}$ and \mathcal{S}_q is given by

$$\mathcal{L}(\mathcal{S}_{i,j}, \mathcal{S}_q) = \min\{\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q), \mathcal{L}(\mathcal{S}_j, \mathcal{S}_q)\}, \quad (12)$$

which denotes the minimum distance between $\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q)$ and $\mathcal{L}(\mathcal{S}_j, \mathcal{S}_q)$. $\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q)$ and $\mathcal{L}(\mathcal{S}_j, \mathcal{S}_q)$ are obtained from the previous calculation in a same manner.

Average linkage distance is the average of the pair distances which can be written as

$$\mathcal{L}(\mathcal{S}_{i,j}, \mathcal{S}_q) = \frac{|\mathcal{S}_i|\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q) + |\mathcal{S}_j|\mathcal{L}(\mathcal{S}_j, \mathcal{S}_q)}{|\mathcal{S}_{i,j}|}. \quad (13)$$

Complete linkage distance is the maximum value of the distances $\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q)$ and $\mathcal{L}(\mathcal{S}_j, \mathcal{S}_q)$ which is expressed as

$$\mathcal{L}(\mathcal{S}_{i,j}, \mathcal{S}_q) = \max\{\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q), \mathcal{L}(\mathcal{S}_j, \mathcal{S}_q)\}. \quad (14)$$

Ward linkage is the weighted distance between the groups which is shown in (15).

Centroid linkage is defined by the virtual centroid of the two groups which is calculated by

$$\mathcal{L}(\mathcal{S}_{i,j}, \mathcal{S}_q) = \|C_{\mathcal{S}_{i,j}} - C_{\mathcal{S}_q}\|_2^2, \quad (16)$$

where $C_{\mathcal{S}_i}$ is the centroid of the group \mathcal{S}_i .

These common linkage methods can be unified calculated by Lance-Williams formulation [42] which is explained with Table I and (17). Thus, we have

$$\mathcal{L}(\mathcal{S}_{i,j}, \mathcal{S}_q) = \Delta_i \mathcal{L}(\mathcal{S}_i, \mathcal{S}_q) + \Delta_j \mathcal{L}(\mathcal{S}_j, \mathcal{S}_q) + \Lambda \mathcal{L}(\mathcal{S}_i, \mathcal{S}_j) + \Upsilon |\mathcal{L}(\mathcal{S}_i, \mathcal{S}_q) - \mathcal{L}(\mathcal{S}_j, \mathcal{S}_q)|. \quad (17)$$

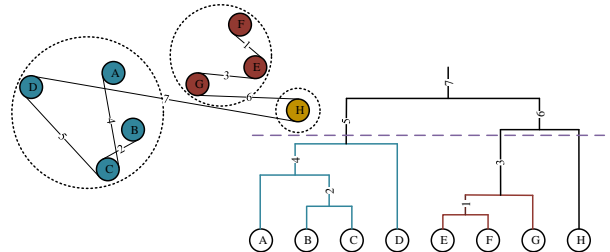


Fig. 2. The dendrogram of the proposed hierarchical clustering user grouping method.

We choose the complete linkage in order to make the users of the same group enjoy higher correlation in the angle domain so that they can be better covered by a same beam to obtain beamforming gain. Assume that there is a user set containing P users $\mathcal{V} = \{U_1, U_2, \dots, U_P\}$ which needs to be divided into

TABLE I
LINKAGE METHOD PARAMETER

Linkage method	Δ_i	Δ_j	Λ	Υ
Single linkage	1/2	1/2	0	-1/2
Complete linkage	1/2	1/2	0	1/2
Average linkage	$\frac{ S_i }{ S_i + S_j }$	$\frac{ S_j }{ S_i + S_j }$	0	0
Centroid linkage	$\frac{ S_i }{ S_i + S_j }$	$\frac{ S_j }{ S_i + S_j }$	$\frac{- S_i S_j }{ S_i + S_j }$	0
Ward linkage	$\frac{ S_i + S_q }{ S_i + S_j + S_q }$	$\frac{ S_j + S_q }{ S_i + S_j + S_q }$	$\frac{- S_q }{ S_i + S_j + S_q }$	0

N groups stored in \mathcal{C} ($P > N$). As we take the bottom-up approach, so the P users initially form P groups. Then, the similarity of any two users is calculated by (11). Two groups are merged into a new group based on the complete linkage at each time. The number of the current groups ind keeps decreasing with the merging process. When the desired group number is reached ($ind = N$), the user grouping procedure is completed. The dendrogram of the proposed initial hierarchical clustering user grouping method is illustrated in Fig. 2 and the AGNES user grouping algorithm is summarized in **Algorithm 1**. The simulation result of the different linkage methods in Fig. 3 also proves our analysis that the complete linkage can achieve the best performance.

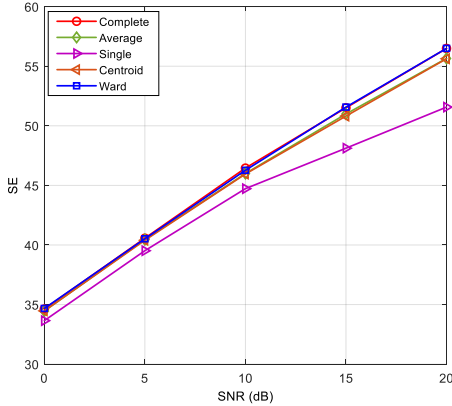


Fig. 3. System performance comparisons of different chain rule methods. $K = 7, G = 4, L = 6, P_{max} = 24\text{mW}, P_{tot} = 2\text{mW}$.

Algorithm 1 AGNES User Grouping algorithm: $\mathcal{C} = \text{user_group}(\mathcal{V}, N)$

Inputs: User set $\mathcal{V} = \{U_1, U_2, \dots, U_P\}$, desired number of group N ;

Outputs: User grouping strategy $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_N\}$.

Initialization: Initial single user groups $\mathcal{C}_p = \{U_p\}, ind = P$.

- 1: Calculate the correlation \mathcal{L} in \mathcal{V} by (11);
- 2: **repeat**
- 3: Search for two groups with the maximal similarity by the complete linkage method;
- 4: Merge the groups with the maximal similarity;
- 5: $ind \leftarrow ind - 1$;
- 6: **until** $ind = N$

IV. TWO PROPOSED USER GROUPING AND BEAM SELECTION SCHEMES

In this section, we aim to solve the integer problem formed by user grouping and beam selection. The optimal solution of the integer problem can be obtained by exhaustively searching over all user grouping and beam selection combinations which is infeasible due to prohibitively high complexity. A number of works such as [30], [31] adopt the sequential two-stage method to reduce the complexity, where the users are grouped at first and then the BS chooses the beams according to the fixed group strategy. Nevertheless, as mentioned before, this method leads to a beam overlapping problem if we consider to serve all users simultaneously in the system. For low complexity and to tackle the beam overlapping problem, two new user grouping and beam selection schemes are developed.

A. DIR-AGNES user grouping and beam selection scheme

To solve the beam overlapping problem, the most natural way is to delete the chosen beam element from the codebook.

At first, we perform user grouping with **Algorithm 1**. K users are allocated to G groups as $\mathcal{C} = \cup_{g=1}^G \mathcal{C}_g = \mathcal{U}$. The user grouping strategy is obtained as $\mathcal{S} = \hat{\mathcal{C}}$. The g -th group chooses its desired beam element $\tilde{\mathbf{f}}_g^{RF}$ from the predefined codebook \mathcal{F} based on the beam gain as

$$\{\tilde{\mathbf{f}}_g^{RF}\} = \arg \max_{\tilde{\mathbf{f}}_g^{RF} \in \mathcal{F}} \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\tilde{\mathbf{f}}_g^{RF} \right)^H \mathbf{h}_{g,u} \right|^2, g = 1, \dots, G, \quad (18)$$

and the obtained beamforming gain of the g -th group is

$$\zeta_g = \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\tilde{\mathbf{f}}_g^{RF} \right)^H \mathbf{h}_{g,u} \right|^2, g = 1, \dots, G. \quad (19)$$

Next, we choose the group index g^* corresponding to the group with the largest group beamforming gain as

$$g^* = \arg \max_g \zeta_g. \quad (20)$$

The group \mathcal{S}_{g^*} has the priority to choose its beam element. We assign the corresponding analog beam $\mathbf{f}_i^{RF} = \tilde{\mathbf{f}}_{g^*}^{RF}$ for \mathcal{S}_{g^*} where i starts from 1. Since $\tilde{\mathbf{f}}_{g^*}^{RF}$ has been chosen for \mathcal{S}_{g^*} , the codebook \mathcal{F} is updated as $\mathcal{F} \leftarrow \mathcal{F} - \{\tilde{\mathbf{f}}_{g^*}^{RF}\}$. For the other groups $\mathcal{C} \leftarrow \mathcal{C} - \mathcal{C}_{g^*}$, they need to choose their desired beams from the new codebook and decide the priori group index in the same manner. Since we directly delete the chosen beam elements from the codebook, this algorithm is called the direct AGNES (DIR-AGNES) user grouping and beam selection scheme which is summarized in **Algorithm 2**.

Algorithm 2 DIR-AGNES Joint User Grouping and Beam Selection Procedure

Inputs: Desired cluster number G , user channels \mathbf{h}_k for $k = 1, 2, \dots, K$;

Outputs: User grouping strategy $\Pi = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_G\}$, $\mathbf{F}_{RF} = [\mathbf{f}_1^{RF}, \mathbf{f}_2^{RF}, \dots, \mathbf{f}_G^{RF}]$.

- 1: Form groups $\mathcal{C} = user_group(\mathcal{U}, G)$ with **Algorithm 1**.
 - 2: **for** $i = 1 : G$ **do**
 - 3: $\{\tilde{\mathbf{f}}_g^{RF}\} = \arg \max_{\hat{\mathbf{f}}_g^{RF} \in \mathcal{F}} \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\hat{\mathbf{f}}_g^{RF} \right)^H \mathbf{h}_{g,u} \right|^2$ for all \mathcal{C}_g ;
 - 4: $\zeta_g = \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\tilde{\mathbf{f}}_g^{RF} \right)^H \mathbf{h}_{g,u} \right|^2$;
 - 5: $g^* = \arg \max_g \zeta_g$;
 - 6: $\mathcal{F} \leftarrow \mathcal{F} - \left\{ \tilde{\mathbf{f}}_{g^*}^{RF} \right\}$, $\mathcal{C} \leftarrow \mathcal{C} - \mathcal{C}_{g^*}$, $\mathcal{S}_i = \mathcal{C}_{g^*}$;
 - 7: $\mathbf{f}_i^{RF} = \tilde{\mathbf{f}}_{g^*}^{RF}$;
 - 8: **end for**
-

B. SUC-AGNES user grouping and beam selection scheme

The proposed DIR-AGNES procedure can effectively solve the beam overlapping problem. However, this user grouping procedure only avoids allocating the same beam to different groups, which may still lead to severe inter-group interference from the predefined beams. Thus, in this subsection, we further propose a successive AGNES (SUC-AGNES) joint user grouping and beam selection scheme. This proposed scheme exploits the multi-path feature of mmWave communication channels to select the beam, which can actively mitigate the interference from the defined beam elements.

In the initialization, we allocate the K users to G groups by **Algorithm 1** as $\mathcal{C} = \cup_{g=1}^G \mathcal{C}_g = \mathcal{U}$. Let $\mathbf{q}_{g,u} = \mathbf{h}_{g,u}$ be the auxiliary channel variable. The G groups choose their desired beams from the predefined codebook \mathcal{F} by

$$\{\tilde{\mathbf{f}}_g^{RF}\} = \arg \max_{\hat{\mathbf{f}}_g^{RF} \in \mathcal{F}} \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\hat{\mathbf{f}}_g^{RF} \right)^H \mathbf{q}_{g,u} \right|^2, g = 1, \dots, G, \quad (21)$$

and the obtained beamforming gain of each group by

$$\zeta_g = \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\tilde{\mathbf{f}}_g^{RF} \right)^H \mathbf{q}_{g,u} \right|^2, g = 1, \dots, G. \quad (22)$$

We choose the group g^* with the largest group beamforming gain by (20). Then, the corresponding analog beam is assigned to be $\mathbf{f}_i^{RF} = \tilde{\mathbf{f}}_{g^*}^{RF}$ for the group g^* and record the chosen users in the group g^* , i.e., $\mathcal{S}_i = \mathcal{C}_{g^*}$. \mathcal{S}_i stored the chosen users and i starts from 1. For the other unchosen users $\mathcal{C} = \mathcal{C} - \mathcal{C}_{g^*}$, we aim to choose the analog beam appropriately to actively avoid the interference from the users who have been chosen. To achieve this, we remove the component of the previous determined analog beam from the unchosen users' channels by a Gram-Schmidt based procedure. Let $\mathbf{b}_i \triangleq \tilde{\mathbf{f}}_{g^*}^{RF}$ be the determined analog beam for the group \mathcal{S}_i . The component of the previous determined beam is removed from \mathbf{b}_i by

$$\mathbf{b}_i \leftarrow \mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{b}_j^H \mathbf{b}_i \mathbf{b}_j, \mathbf{b}_i = \mathbf{b}_i / \|\mathbf{b}_i\|_2. \quad (23)$$

The channels of the remaining unchosen users $U_m \in \mathcal{C}$ are updated by an OMP fashion [22]:

$$\mathbf{q}_m \leftarrow \left(\mathbf{I}_{N_{BS}} - \mathbf{b}_i \mathbf{b}_i^H \right) \mathbf{q}_m, \quad (24)$$

By this way, the unchosen users can select the paths which is less correlated with the channel paths of the chosen users. Then, the user grouping and beam selection can be finished by the above scheme recursively. The whole SUC-AGNES joint user grouping and beam selection scheme is summarized in **Algorithm 3**.

Algorithm 3 SUC-AGNES Joint User Grouping and Beam Selection Procedure

Inputs: Desired cluster number G , user channels \mathbf{h}_k for $k = 1, 2, \dots, K$;

Outputs: User grouping strategy $\Pi = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_G\}$, $\mathbf{F}_{RF} = [\mathbf{f}_1^{RF}, \mathbf{f}_2^{RF}, \dots, \mathbf{f}_G^{RF}]$;

Initialization: $\mathbf{q}_k = \mathbf{h}_k$ for $k = 1, 2, \dots, K$.

- 1: Form initial groups $\mathcal{C} = user_group(\mathcal{U}, G)$ by **Algorithm 1**.
 - 2: **for** $i = 1 : G$ **do**
 - 3: $\{\tilde{\mathbf{f}}_g^{RF}\} = \arg \max_{\hat{\mathbf{f}}_g^{RF} \in \mathcal{F}} \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\hat{\mathbf{f}}_g^{RF} \right)^H \mathbf{q}_{g,u} \right|^2$ for all \mathcal{C}_g ;
 - 4: $\zeta_g = \sum_{u=1}^{|\mathcal{C}_g|} \left| \left(\tilde{\mathbf{f}}_g^{RF} \right)^H \mathbf{q}_{g,u} \right|^2$;
 - 5: $g^* = \arg \max_g \zeta_g$;
 - 6: $\mathcal{C} \leftarrow \mathcal{C} - \mathcal{C}_{g^*}$, $\mathcal{S}_i = \mathcal{C}_{g^*}$;
 - 7: $\mathbf{f}_i^{RF} = \tilde{\mathbf{f}}_{g^*}^{RF}$;
 - 8: $\mathbf{b}_i \triangleq \tilde{\mathbf{f}}_{g^*}^{RF}$;
 - 9: for $i > 1$, $\mathbf{b}_i \leftarrow \mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{b}_j^H \mathbf{b}_i \mathbf{b}_j$, $\mathbf{b}_i \leftarrow \mathbf{b}_i / \|\mathbf{b}_i\|$;
 - 10: $\mathbf{q}_m = \left(\mathbf{I}_{N_{BS}} - \mathbf{b}_i \mathbf{b}_i^H \right) \mathbf{q}_m$ for $U_m \in \mathcal{C}$;
 - 11: Regroup the remaining users $\mathcal{C} = user_group(\mathcal{C}, G - i)$;
 - 12: **end for**
-

After determining all user groups and the analog beamformer \mathbf{F}_{RF} by **Algorithm 2** or **Algorithm 3**, we sort the users in each group with their channel and analog beam gains, i.e., $\|\mathbf{F}_{RF}^H \mathbf{h}_{g,1}\|_2^2 \geq \|\mathbf{F}_{RF}^H \mathbf{h}_{g,2}\|_2^2 \geq \dots \geq \|\mathbf{F}_{RF}^H \mathbf{h}_{g,|\mathcal{S}_g}|\|_2^2$ for $g = 1, \dots, G$. Assuming we have obtained the power allocation strategy, we choose the user with the largest gain as the beam centroid in this group and design the digital beamformer with the effective channels of the strongest users of all groups [37]. The effective channel to design the digital beamformer is written as $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \dots, \tilde{\mathbf{h}}_G]$ where $\tilde{\mathbf{h}}_g = \sqrt{P_{g,1}} \mathbf{F}_{RF}^H \mathbf{h}_{g,1}$. The digital combiner is presented as

$$\mathbf{F}_{BB} = \tilde{\mathbf{H}} (\tilde{\mathbf{H}}^H \tilde{\mathbf{H}})^{-1}. \quad (25)$$

Each column of the digital beamformer is further normalized to satisfy the unit power constraint for the HBF beamformer below

$$\mathbf{f}_g^{BB} = \frac{\mathbf{f}_g^{BB}}{\|\mathbf{F}_{RF}^H \mathbf{f}_g^{BB}\|_2}, g = 1, \dots, G. \quad (26)$$

$$Q_{SE}(\mathbf{P}, \mathbf{m}) = \sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} \log \left(1 + 2m_{g,u} \sqrt{d_g(g,u)P_{g,u}} - m_{g,u}^2 \left(\sum_{U_{q,v} \in \Omega_{g,u}} d_g(q,v)P_{q,v} + \sigma^2 \right) \right). \quad (31)$$

V. POWER ALLOCATION

After solving the integer problem (user grouping and beam selection), we now consider the continuous problem. To optimize the power for the users in an mmWave MIMO-HBF-NOMA system, SE and EE are two widely used criteria for system evaluation. We formulate the power optimization problems for SE and EE respectively, which are both non-convex. We first introduce QT method to tackle the SE maximization problem, and then derive nested QT (NQT) for EE maximization.

A. Spectrum Efficiency

First, we take SE as our optimization objective function which has been formulated in **P1**:

$$\mathbf{P1}: \max_{\{P_{g,u}\}} SE \quad (27)$$

$$s.t. \text{C1} : P_{g,u} \leq P_{g,u}^{\max}, \quad \forall g = 1, 2, \dots, G, u = 1, 2, \dots, |\mathcal{S}_g|, \quad (27a)$$

$$\text{C2} : R_{g,u} \geq R_{g,u}^{\min}, \quad \forall g = 1, 2, \dots, G, u = 1, 2, \dots, |\mathcal{S}_g|, \quad (27b)$$

$$\text{C3} : \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{g,u}\|_2^2 P_{g,u} - \sum_{r=u+1}^{|\mathcal{S}_g|} \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{g,r}\|_2^2 P_{g,r} \geq P_{tol}, \quad \forall g = 1, 2, \dots, G, u = 1, 2, \dots, |\mathcal{S}_g| - 1, \quad (27c)$$

where C1 is the transmitted power constraint with $P_{g,u}^{\max}$ being the maximum transmitted power for $U_{g,u}$. C2 is the data rate constraint with $R_{g,u}^{\min}$ being the minimum data rate to satisfy the Quality of Service (QoS) requirement. Since in NOMA communication systems, the signals of the different users in the same cluster are distinguished by power divergence, C3 is imposed to guarantee that the final power gain of the users has enough gap to ensure the success of the SIC decoding. P_{tol} is the minimum power difference required to distinguish the desired decoded signal and the remaining non-decoded interference code in a cluster.

Constraints in C2 are specified as follow:

$$\text{C2} : d_g(g,u)P_{g,u} - \left(2^{R_{g,u}^{\min}} - 1 \right) \left(\sum_{U_{q,v} \in \Omega_{g,u}} d_g(q,v)P_{q,v} + \sigma^2 \right) \geq 0, \quad (28)$$

where $d_g(q,v) = \|(\mathbf{f}_g^{BB})^H \mathbf{F}_{RF}^H \mathbf{h}_{q,v}\|^2$ denotes the combination consisting of the beam gain and the channel gain from $U_{q,v}$ to the g^{th} data stream and $\Omega_{g,u}$ is the user set containing the users weaker than $U_{g,u}$ in the g^{th} group and the users in other groups, i.e., $\Omega_{g,u} = \{U_{g,u+1}, \dots, U_{g,|\mathcal{S}_g|}\} \bigcup_{q \neq g} \mathcal{S}_q$.

Constraint C3 is specified as follow:

$$\text{C3} : d_g(g,u)P_{g,u} - \sum_{r=u+1}^{|\mathcal{S}_g|} d_g(g,r)P_{g,r} \geq P_{tol}. \quad (29)$$

Directly solving the non-convex **P1** is difficult, because the objective function $R_{g,u}$ is a fractional structure. By observing that **P1** is a sum-of-ratio problem, we consider to address it by the QT algorithm in [39]. According to *Corollary 2* in [39], **P1** is equivalent to

$$\mathbf{P2}: \max_{\mathbf{P}, \mathbf{m}} Q_{SE}(\mathbf{P}, \mathbf{m}) \quad (30)$$

$$s.t. \text{C1} : P_{g,u} \leq P_{g,u}^{\max}, \quad (30a)$$

$$\text{C2} : d_g(g,u)P_{g,u} - \left(2^{R_{g,u}^{\min}} - 1 \right) \left(\sum_{U_{q,v} \in \Omega_{g,u}} d_g(q,v)P_{q,v} + \sigma^2 \right) \geq 0, \quad (30b)$$

$$\text{C3} : d_g(g,u)P_{g,u} - \sum_{v=u+1}^{|\mathcal{S}_g|} d_g(g,v)P_{g,v} \geq P_{tol}, \quad (30c)$$

where $Q_{SE}(\mathbf{P}, \mathbf{m})$ is the new objective function given by (31). $\mathbf{m} \in \mathbb{R}$ is the auxiliary variable collection $\{m_{g,u}\}$.

We propose to optimize the primal variable \mathbf{P} and the auxiliary variable \mathbf{m} iteratively. When \mathbf{P} is fixed, the optimal $\{m_{g,u}\}$ is updated in a closed form as

$$m_{g,u}^* = \frac{\sqrt{d_g(g,u)P_{g,u}}}{\sum_{U_{q,v} \in \Omega_{g,u}} d_g(q,v)P_{q,v} + \sigma^2}. \quad (32)$$

When $\{m_{g,u}\}$ is fixed, the objective function $Q_{SE}(\mathbf{P}, \mathbf{m})$ is convex with respect to \mathbf{P} because the formulation in $\log(\cdot)$ function is concave and $\log(\cdot)$ function is nondecreasing and concave. This allows us to use an optimization method to obtain \mathbf{P} . When \mathbf{m} and \mathbf{P} both achieve their optimal values, the objective function $Q_{SE}(\mathbf{P}, \mathbf{m})$ obtains its maximum. The process to allocate the power for maximizing SE is shown in **Algorithm 4**. This algorithm is essentially a block coordinate ascent algorithm which can converge to a stationary point due to the concave-convex form. The details of the proof of the convergence can be found in [39].

B. Energy Efficiency

When we consider EE as the optimization objective target, the problem is presented as

$$\mathbf{P3} : \max_{\{P_{g,u}\}} EE \quad (33)$$

$$s.t. (30a), (30b), (30c).$$

$$\mathcal{Q}_{EE}(\mathbf{P}, n, \mathbf{w}) = 2n \left(\sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} \log \left(1 + 2w_{g,u} \sqrt{d_g(g, u) P_{g,u}} - w_{g,u}^2 \left(\sum_{U_{q,v} \in \Omega_{g,u}} d_g(q, v) P_{q,v} + \sigma^2 \right) \right) \right)^{\frac{1}{2}} - n^2 \left(\xi \sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} P_{g,u} + P_C \right). \quad (36)$$

Algorithm 4 Power Allocation for Maximizing SE

Inputs: User grouping strategy Π , \mathbf{F}_{BB} , \mathbf{F}_{RF} , $\mathbf{h}_{g,u}$.

Outputs: Power Allocation $\{P_{g,u}\}$;

- 1: **repeat**
 - 2: Update $m_{g,u}^*$ by (32);
 - 3: Update $P_{g,u}$ by solving the convex optimization problem **P2** for fixed \mathbf{m} ;
 - 4: **until** \mathcal{Q}_{SE} converges.
-

P3 is a non-convex problem because the objective function in a ratio form is non-convex and the sum rate in the numerator is also non-convex, which has been analyzed in **P1**. According to [39], we can treat the numerator as an inner multiple-ratio problem nested in the outer single-ratio energy efficiency problem. Thus, we introduce the NQT algorithm to deal with the nested ratio problem. First, we recast the outer ratio problem as

$$\mathbf{P4}: \max_{\{\mathbf{P}, n\}} 2n \left(\sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} R_{g,u} \right)^{\frac{1}{2}} - n^2 \left(\xi \sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} P_{g,u} + P_C \right) \quad (34)$$

s.t. (30a), (30b), (30c).

where n is the auxiliary variable for the single-ratio problem. For the inner multiple-ratio problem in $R_{g,u}$, we apply the quadratic transform again to the SINR term inside the $R_{g,u}$ and further recast **P4** as

$$\mathbf{P5}: \max_{\{\mathbf{P}, n, \mathbf{w}\}} \mathcal{Q}_{EE}(\mathbf{P}, n, \mathbf{w}) \quad (35)$$

s.t. (30a), (30b), (30c).

where $\mathcal{Q}_{EE}(\mathbf{P}, n, \mathbf{w})$ is a new objective function after two quadratic transform given by (36). $\{w_{g,u}\}$ are the auxiliary variables of the fractional programming from (34) to (35). We update $w_{g,u}$ as

$$w_{g,u}^* = \frac{\sqrt{d_g(g, u) P_{g,u}}}{\sum_{U_{q,v} \in \Omega_{g,u}} d_g(q, v) P_{q,v} + \sigma^2}. \quad (37)$$

After the update of $w_{g,u}$, the optimal n is updated as

$$n^* = \frac{\sqrt{\sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} R_{g,u}}}{\xi \sum_{g=1}^G \sum_{u=1}^{|\mathcal{S}_g|} P_{g,u} + P_C}. \quad (38)$$

Similar to the scheme proposed in subsection A, we iteratively update \mathbf{P} , n and \mathbf{w} until convergence to obtain the power allocation for maximizing EE. The power allocation algorithm for maximizing EE is shown in **Algorithm 5**.

Algorithm 5 Power Allocation for Maximizing EE

Inputs: User grouping strategy Π , \mathbf{F}_{BB} , \mathbf{F}_{RF} , $\mathbf{h}_{g,u}$.

Outputs: Power Allocation $\{P_{g,u}\}$;

- 1: **repeat**
 - 2: Update $w_{g,u}^*$ by (37);
 - 3: Update n^* by (38);
 - 4: Update $P_{g,u}$ by solving the convex optimization problem **P5** for fixed \mathbf{w} and n ;
 - 5: **until** \mathcal{Q}_{EE} converges
-

VI. ALGORITHM SUMMARY AND COMPUTATIONAL COMPLEXITY ANALYZE

A. Algorithm Summary

In Section III-V, we have proposed an initial user grouping algorithm, two user grouping and beam selection schemes, the digital beamforming algorithm and the power allocation algorithm. The overall algorithm for the mmWave MIMO-HBF-NOMA system is shown in **Algorithm 6**. T is the predefined maximum iteration number. We emphasize that our design idea is to divide the original problem into an integer problem and a continuous problem. The integer problem is solved at one time while the continuous problem is addressed in an iterative manner. The optimization of the digital beamforming matrix and the power of the users are both aimed at increasing the SE(EE) of the system. During the iteration, the power allocation might change the order of the users since their equivalent channel gains are changed. If the strongest user $U_{g,1}$ is different from the last iteration, the digital beamforming will be different too. Thus, the iteration might keep incessant flipping. Considering this, we set a maximum iteration number. In contrast to the conventional three-step approach as shown in Fig. 4 (a), our proposed two schemes give rise to more flexibility at resource allocation as well as low complexity. Moreover, the beam overlapping problem is solved by updating the codebook in two ways, respectively. The proposed SUC-AGNES scheme expands the idea of inter-group interference cancellation to analog domain which outperforms the traditional inter-group interference cancellation approach in only digital domain.

It should be noted that although the users in this paper are equipped with single antenna, our proposed user grouping can be easily extended to the multi-antenna situations to

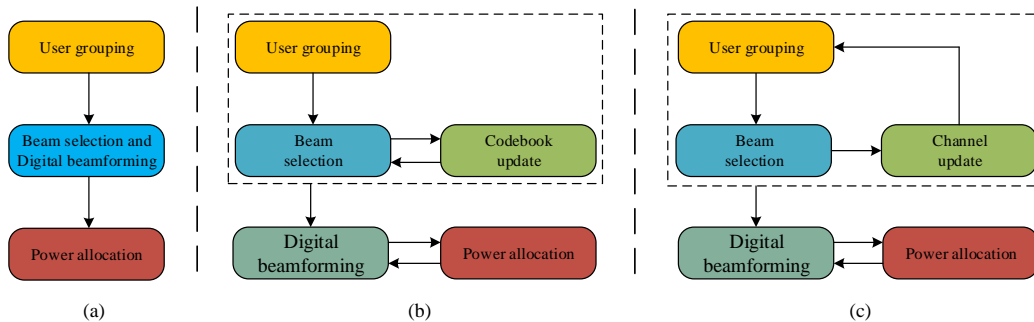


Fig. 4. Comparison of the HBF-NOMA design procedure: (a) the traditional HBF-NOMA design, (b) the proposed DIR-AGNES HBF-NOMA design, (c) the proposed SUC-AGNES HBF-NOMA design.

Algorithm 6 User grouping, beam selection and power allocation for mmWave-HBF-NOMA system

Inputs: $\mathbf{h}_k, G, \mathcal{F}, T$.

Outputs: $\Pi, \mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{P}$.

Initialization: $\{P_{g,u}\} = P_{\max}, ITE = 1$;

- 1: Perform joint user grouping and beam selection by **Algorithm 1** and **Algorithm 2**(**Algorithm 3**);
- 2: **repeat**
- 3: Calculate the digital beamformer \mathbf{F}_{BB} by (25) and (26).
- 4: Allocate power for maximizing SE(EE) by **Algorithm 4**(**5**);
- 5: $ITE \leftarrow ITE + 1$;
- 6: **until** $SE(EE)$ converges or $ITE > T$

obtain a low-complexity sub-optimal design. In the mmWave MIMO-HBF-NOMA systems where the users are with multiple antennas, the users can select their beam elements from the user DFT codebook in the beam sweeping step as $\{\mathbf{w}_{g,u}^{RF}\} = \arg \max_{\hat{\mathbf{w}}_{g,u}^{RF} \in \mathcal{W}} |\mathbf{H}_{g,u} \hat{\mathbf{w}}_{g,u}^{RF}|^2$, where $\mathbf{w}_{g,u}^{RF}$ denotes the analog beamforming vector of $U_{g,u}$, $\mathbf{H}_{g,u}$ denotes the channel matrix of $U_{g,u}$ and \mathcal{W} denotes the user DFT codebook, respectively. The beam sweeping procedure is implemented before the user grouping step, i.e., **Step 1** in **Algorithm 2** and **Step 1** and **Step 11** in **Algorithm 2**. The equivalent channel of $U_{g,u}$ is correspondingly revised to be $\left(\mathbf{f}_g^{BB} \right)^H \mathbf{F}_{RF}^H \mathbf{H}_{g,u} \mathbf{w}_{g,u}^{RF}$.

B. Computational complexity

Next, for the integer problem, we analyze the computation complexity of the two proposed user grouping and beam selection schemes. For the DIR-AGNES algorithm, the average computational complexity of the AGNES user grouping in **Algorithm 1** is $\mathcal{O}(K^2)$. The average computational complexity of the beam selection process is $\mathcal{O}(GKN_{BS}N_{beam})$. The total average computational complexity is $\mathcal{O}(K^2 + GKN_{BS}N_{beam})$. For the SUC-AGNES algorithm, the average computational complexity of the user grouping part is $\mathcal{O}(GK^2)$. The average computational complexity of the beam selection and channel updating is $\mathcal{O}(GKN_{BS}N_{beam} + GKN_{BS})$. The total average computational complexity is $\mathcal{O}(GK^2 + GKN_{BS}N_{beam} + GKN_{BS})$. The user grouping and the beam selection problem can be optimally solved by

the exhaustive search, but its computational complexity is prohibitively high, which is given by

$$\mathcal{O} \left(\left(G^K + \sum_{i=1}^{G-1} (-1)^i \binom{G}{i} (G-i)^K \right) \binom{N_{beam}}{G} G! K N_{BS} \right). \quad (39)$$

It can be seen that the proposed schemes significantly reduce the computational complexity from the exponentially level to the linear level.

VII. SIMULATION RESULTS

In this section, we present the simulation results to evaluate the performance of the proposed algorithms. We assume that the BS is equipped with $N_{BS} = 64$ antennas. Each user has one single antenna. The path gain of $U_{g,u}$ is set as: (1) $\alpha_{g,u,l} \sim \mathcal{CN}(0, 1)$ for $l = 1, \dots, L$; (2) $\theta_{g,u,l}$ for $l = 1, \dots, L$ are uniformly distributed within $[-\frac{\pi}{2}, \frac{\pi}{2}]$. We generally set the number of paths to be $L = 6$. The QoS minimum rate constraint for each user is $R_{\min} = 0.01$ bps/Hz [30]. The simulation results are obtained by 3000 Monte Carlo simulations. The parameter of the system is set to be $\xi = 1/0.38$ [30], $P_C = 100$ mW. The resolution of the DFT codebook is set to be $N_{beam} = N_{BS}$. The maximum iteration number of the continuous problem is set to be $T = 20$.

In the simulation, we compare several algorithms which are explained as follows:

- **Proposed DIR-AGNES:** The user grouping and beam selection are performed by **Algorithm 2**, the power is allocated by **Algorithm 4** for SE maximization and **Algorithm 5** for EE maximization.
- **Proposed SUC-AGNES:** The user grouping and beam selection are performed by **Algorithm 3**, the power is allocated by **Algorithm 4** for SE maximization and **Algorithm 5** for EE maximization.
- **Fully digital:** The user grouping strategy is the same as **SUC-AGNES** while the BS takes the fully digital beamformer rather than the hybrid beamformers. The digital beamforming matrix is obtained by ZF algorithm. The power is allocated by **Algorithm 4** for SE maximization and **Algorithm 5** for EE maximization.
- **K-means:** The user grouping and beam selection is performed by **Algorithm 3** with the initial user grouping algorithm changing to the K-means algorithm in [35].

The digital beamforming matrix is obtained by ZF algorithm. The power is allocated by **Algorithm 4** for SE maximization and **Algorithm 5** for EE maximization.

- **Algorithm in [9]:** The user grouping algorithm is based on the channel gain difference in [9]. The beam selection is performed by **Algorithm 2**. The digital beamforming matrix is obtained by ZF algorithm. The power is allocated by **Algorithm 4** for SE maximization and **Algorithm 5** for EE maximization.
- **OMA:** The uplink mmWave OMA transmission is performed via a ZF digital precoder and a power allocation design without intra-group interference terms and constraint C3. We allocate a user to at most one time slot in this TDMA system.

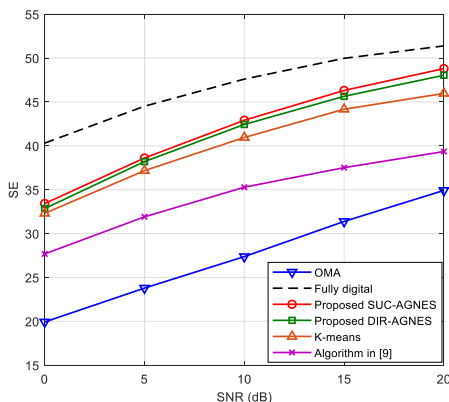


Fig. 5. SE versus SNR of the different algorithms. $K = 9$, $P_{max} = 24$ mW, $P_{tol} = 2$ mW.

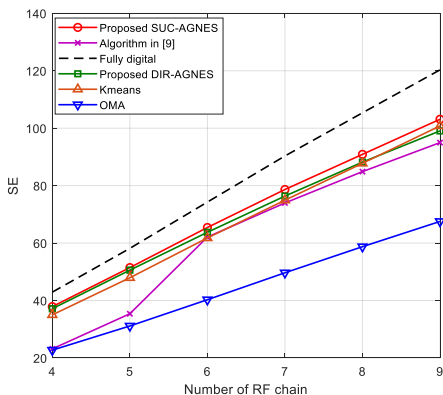


Fig. 6. SE versus the number of the RF chains of the different algorithms. $K = 12$, $SNR = 10$ dB, $P_{max} = 24$ mW, $P_{tol} = 1$ mW.

A. SE

We consider the normalized bandwidth in which the sum rate can be defined as the SE. Fig. 5 shows the SE versus SNR of the different algorithms. The number of users is $K = 9$, maximum power of each user is $P_{max} = 24$ mW and the interval to guarantee the decoding process is $P_{tol} = 2$ mW. The figure shows that our proposed DIR-AGNES and SUC-AGNES scheme outperform the traditional OMA communication and some previous algorithms. The SUC-AGNES scheme

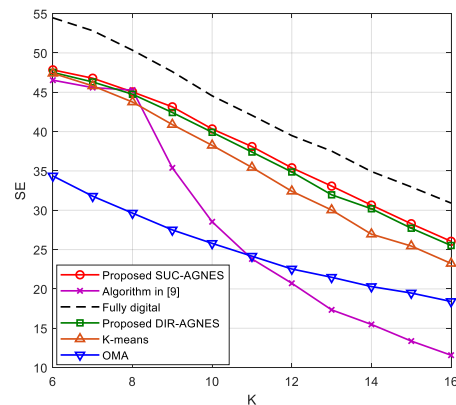


Fig. 7. SE versus the number of the users of the different algorithms. $G = 4$, $SNR = 10$ dB, $P_{max} = 24$ mW, $P_{tol} = 2$ mW.

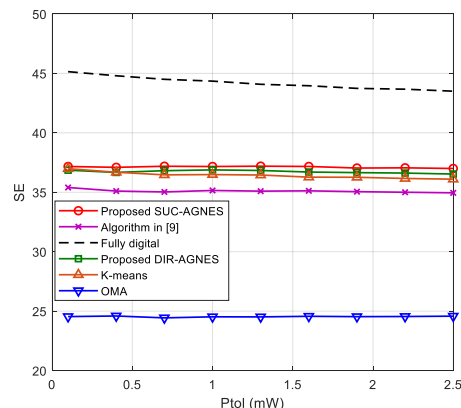


Fig. 8. SE versus the number of the users of the different algorithms. $K = 7$, $SNR = 3$ dB, $P_{max} = 20$ mW.

is able to achieve a better performance compared to the DIR-AGNES. This result is attributed to the update process in the SUC-AGNES scheme. The inter-group interference of the defined groups is actively avoided after renewing the remaining users' channels.

Fig. 6 shows the SE of the different algorithms under the fixed number of users. As the number of RF chains increases, the SUC-AGNES scheme shows salient advantage over the DIR-AGNES scheme. The performance of the algorithm in [9] improves from $G = 6$. At this point, the number of user is twice the number of the RF chains. It implies that the algorithm in [9] is more suitable for the scenario in which there are less than two users in one group on average. Our proposed schemes show stable superiority regardless of the number of the RF chains.

The performance of the mmWave-NOMA scheme is related to the number of the users served by the BS simultaneously. We also investigate the relationship between the system performance and the system overload situation $\rho = K/G$. Similar to the results in Fig. 6, Fig. 7 shows the SE versus the number of the users K . It is more obvious that the algorithm in [9] is not effective in the scenario where $\rho > 2$. The performance of the K-means algorithm in [35] decreases rapidly when K increases as that the initial centroid of the beam in the K-means scheme is randomly chosen which cannot guarantee reasonable user grouping.

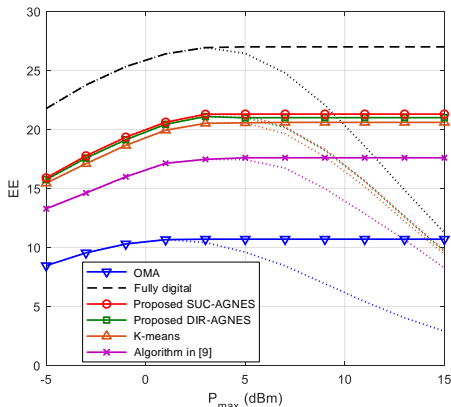


Fig. 9. EE versus the maximum power limit of the users of the different algorithms. $K = 9$, $SNR = 5$ dB, $P_{tol} = 0.5$ mW.

Fig. 8 illustrates the SE versus the required power interval P_{tol} . We set $K = 7$, $SNR = 3$ dB and $P_{max} = 20$ mW. The SE of the NOMA schemes generally decreases with the required power gap increasing. The OMA scheme is not influenced by the required power interval. The performance of the algorithm in [9] does not change much compared to the other NOMA schemes. This is because that the algorithm in [9] is based on the channel gain difference. The users with large channel gain difference are more prone to be grouped together which are not influenced by P_{tol} as much as the other NOMA schemes in the power allocation.

B. EE

Fig. 9 plots the EE versus the maximum power of the users P_{max} . It is observed that when the maximum power limit is low, the EE of the schemes increases as P_{max} increases. Then, after a certain threshold, the curve stops increasing after the peak. Further increase in power brings no improvement in EE. It means that allocating too much power on the users is not help from the perspective of EE. Moreover, we also provide the results in terms of SE, the SE performance even decreases when the power extend the peak point.

C. Convergence

The existing state-of-the-art mostly optimizes the transmitting power with Dinkelbach method [43]. In this paper, we propose to allocate the power with QT algorithm. In the simulation, we set $K = 9$, the maximum power of each user is $P_{max} = 24$ mW, $SNR = 10$ dB and $P_{tol} = 2$ mW. The optimization results of the two algorithms are almost the same as presented in Fig. 10, except for the convergence speed. In Fig. 11, we present the average convergence speed of the two algorithms. The minimum mean-squared error (MMSE) of the sum rate is defined as $\omega = \frac{R_{ite} - \hat{R}}{\hat{R}}$. R_{ite} is the sum rate of the ite -th iteration in the power optimization. \hat{R} is the optimal sum rate after the iteration. The QT algorithm has slight advantage in the first 20 iterations.

VIII. CONCLUSIONS

In this paper, we have considered the design of enhanced uplink mmWave-NOMA systems with a hybrid beamforming

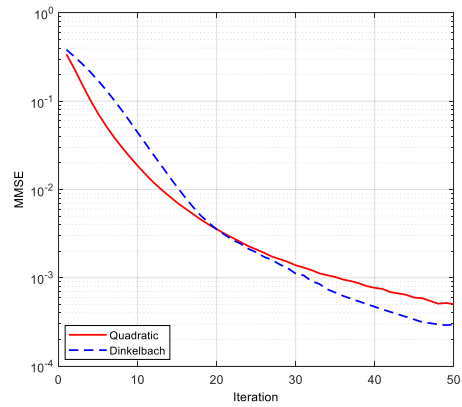


Fig. 10. SE versus the iteration number.

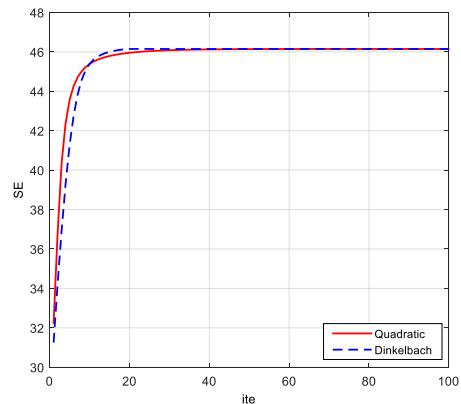


Fig. 11. Sum rate MMSE versus the iteration number.

structure. We have proposed a novel initial AGNES user grouping algorithm based on the channel correlation according to the feature of mmWave channels. The complete chain method is chosen to enable the users obtain better beam gain. Moreover, two user grouping and beam selection algorithms, the DIR-AGNES scheme and the SUC-AGNES scheme, are provided to combat the beam overlapping problem. The SUC-AGNES scheme further updates the users' channels to actively avoid the inter-group interference. The power allocation and the digital beamforming are iteratively optimized to further improve the system performance. The quadratic transform algorithm is introduced to allocate the power for each user. In the simulation, two system criteria are considered, i. e., SE and EE. Simulation results have shown that our proposed algorithms outperform the other designs in different system situation.

REFERENCES

- [1] V.W.S. Wong, R. Schober, D.W.K. Ng, and L. C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.
- [2] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. of IEEE PIMRC*, London, UK, Sept. 2013.
- [3] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. 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.
- [4] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

- [5] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.
- [6] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech Focus*, vol. 1, no. 2, Jun. 2017.
- [7] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.
- [8] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.
- [9] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation in non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.
- [10] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.
- [11] W. Roh, J. Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.
- [12] T. S. Rappaport, S. Sun, R. Mayuz, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: it will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.
- [13] M. Xiao, S. Mumtaz, Y. M. Huang, L. Dai, Y. H. Li, M. Matthaiou, G. K. Karagiannidis, E. Bjornson, K. Yang, I. Chih-Lin, A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.
- [14] E. Torkildson, C. Sheldon, U. Madhow, and M. Rodwell, "Millimeter-wave spatial multiplexing in an indoor environment," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2009, pp. 1–6.
- [15] E. Torkildson, B. Ananthasubramanian, U. Madhow, and M. Rodwell, "Millimeter-wave MIMO: wireless links at optical speeds," in *Proc. 2006 Allerton Conf. Commun., Control Comput.*
- [16] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Select. Top. Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.
- [17] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
- [18] Z. Xiao, T. He, P. Xia, and X. G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.
- [19] X. Gao, L. Dai, S. Han, C. L. I, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.
- [20] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE J. Select. Areas Commun.*, vol. 35, no. 7, pp. 1576–1590, Jul. 2017.
- [21] J. Zhu, Z. Wang, Q. Li, H. Chen and N. Ansari, "Mitigating intended jamming in mmWave MIMO by hybrid beamforming," *IEEE Wirel. Commu. Lett.*, vol. 8, no. 6, pp. 1617–1620, Dec. 2019.
- [22] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
- [23] D. Zhang, Z. Zhou, C. Xu, Y. Zhang, J. Rodriguez, and T. Sato, "Capacity analysis of NOMA with mmWave massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1606–1618, Jul. 2017.
- [24] Y. Zhou, V. W. S. Wong, and R. Schober, "Performance analysis of
- millimeter wave NOMA networks with beam misalignment," in *Proc. IEEE Int. Conf. Commun.*, May 2018, pp. 1–7.
- [25] M. A. Almasi, M. Vaezi, and H. Mehrpouyan, "Impact of beam misalignment on hybrid beamforming NOMA for mmWave communications," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4505–4518, Jun. 2019.
- [26] Z. Li, S. Han, A. F. Molisch, R. Wang, S. Sangodoyin, "Joint optimization of hybrid beamforming for multi-user massive MIMO downlink," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3600–3614, Jun. 2018.
- [27] Z. Li, S. Han, A. F. Molisch, R. Wang, S. Sangodoyin, "Joint optimization of hybrid beamforming for multi-user massive MIMO downlink," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4288–4303, Jul. 2017.
- [28] A. Adhikary, E. Al Safadi, M. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, "Joint spatial division and multiplexing for mm-Wave channels," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1239–1255, Jun. 2014.
- [29] J. Cui, Y. Liu, Z. Ding, P. Fan and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.
- [30] W. Hao, M. Zeng, Z. Chu and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wirel. Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec. 2017.
- [31] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3000–3004, Mar. 2019.
- [32] 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.
- [33] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. G. Xia, "Joint power control and beamforming for uplink non-orthogonal multiple access in 5G millimeter-wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6177–6189, Sep. 2018.
- [34] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.
- [35] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.
- [36] J. Cui, Z. Ding, P. Fan and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, Nov. 2018.
- [37] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "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.
- [38] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.
- [39] K. Shen and W. Yu, "Fractional programming for communication systems Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.
- [40] J. L. Zhu and Q. Li, "Flexible User Grouping for MIMO-NOMA Millimeter Wave Communication Systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020.
- [41] X. Sun, X. Gao, G. Y. Li and W. Han, "Agglomerative user clustering and cluster scheduling for FDD massive MIMO systems," *IEEE Access*, vol. 7, pp. 86522–86533, 2019.
- [42] F. Murtagh and P. Contreras, "Methods of hierarchical clustering," 2011; <http://arxiv.org/abs/1105.0121>.
- [43] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations Trends Commun. Inf. Theory*, vol. 11, no. 3, pp. 185–396, Jun. 2015.