Channel Estimation and Transmission Strategy for Hybrid MmWave NOMA Systems

Dian Fan, Feifei Gao, *Senior Member, IEEE*, Gongpu Wang, Zhangdui Zhong, *Senior Member, IEEE*, Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—This paper presents a novel channel estimation and transmission strategy for millimeter wave (mmWave) non-orthogonal multiple access (NOMA) communication system with hybrid architecture. We first propose a general iterative index detection-based channel estimation algorithm (IDCEA) that can obtain both direction of arrival (DOA) and channel gain of each channel path. We then design an enhanced hybrid precoding scheme from the angle domain viewpoint to reduce the inter-beam interferences. Next we investigate the multi-user scheduling and power allocation with the objective of maximizing the overall achievable rate. The problem turns to be non-convex and then we decompose it into two sub-problems which separately consider user scheduling and power allocation. The former is solved by a novel algorithm based on the many-to-one two sided matching theory while the latter is solved by an iterative optimization algorithm. Simulation results show that the proposed channel estimation and user scheduling can be better than traditional methods. Finally, numerical examples are provided to corroborate the proposed studies.**

***Index Terms*—Millimeter wave (mmWave), non-orthogonal multiple access (NOMA), channel estimation, user scheduling, power allocation.**

I. Introduction

With the increasing demand for radio spectrum resources, the underutilized millimeter wave (mmWave) band (between 30 GHz and 300 GHz) has received broad attention due to its wide bandwidth and higher spectral efficiency [1]–[3]. Since the transceiver can compensate for the relatively high propagation loss using the beam gain provided by the large-scale antenna array, mmWave combined with large-scale antenna array becomes a core supporting technology in the fifth generation (5G) communication systems [4].

In mmWave massive multiple input multiple output (MIMO) system, a hybrid architecture has been proposed to reduce the hardware complexity and energy consumption [5]–[11]. In order to avoid interference between users, each radio frequency (RF) chain can only support one user at the same time-frequency resources. Thus, the maximum number of users that can be served is equal to the number of RF chains. When the number of users increases, the signal for different users cannot be separated by linear operation. Nevertheless, non-orthogonal multiple access (NOMA) technology can break this fundamental limit by performing superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver [12]–[21].

Specifically, NOMA technology can simultaneously support multiple users on the same time-frequency space resource, and can convert different channel gain between users into multiplexed gain via superposition coding, which is essentially different from the traditional beamspace MIMO. Therefore, the number of supported users can be larger than the number of RF chains in mmWave NOMA system, and the downlink achievable sum rate will also be significantly increased.

The combination of spatial MIMO and NOMA has been extensively studied in [22]–[26]. In [22], the authors presented a beamforming design that minimizes transmission power, where a multi-antenna base station (BS) transmits two single-antenna users using NOMA technology. The authors in [23] introduced an effective transmission scheme that ensure user fairness, in which a multi-antenna BS transmits multiple single antenna users using NOMA technology. Furthermore, in [24]–[26], the user clustering and power allocation scheme was developed to optimize the sum rate and user fairness of MIMO-NOMA systems. However, these MIMO-NOMA techniques are all focus low frequency, and cannot be used for mmWave communication, where channel sparsity, the uncertainty of the number of conflicting users, etc., also need to be considered.

It is highly desirable to use NOMA in mmWave due to the following advantages: 1) The channel of different users in the same direction are heavily correlated in mmWave. This special channel characteristics of mmWave is very suitable to apply NOMA technology; 2) Large-scale antenna array can provide highly directional beams in mmWave. The directional beams results in larger beamforming gains and smaller inter-beam interference, where NOMA transmissions can be applied on each beam; 3) The user overload can be increased using NOMA into mmWave communication, which can improve spectral efficiency.

It is recognized that the full benefits of mmWave with NOMA technology heavily rely on the accurate channel state information (CSI) estimation, which is also regarded as one

Manuscript received September 4, 2018; revised December 29, 2018; accepted February 8, 2019. This work was supported in part by the National Key R&D Program of China (2016YFE0200900), by Major projects of Beijing Municipal Science and Technology Commission under Grant No. Z181100003218010, by the National Natural Science Foundation of China under Grant {61831013, 61771274, 61531011}, and by Beijing Municipal Natural Science Foundation under Grant (4182030, L182042).

Dian Fan, Gongpu Wang and Zhangdui Zhong are with School of Computer and Information Technology, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, P. R. China. (e-mail: fandian@bjtu.edu.cn, gpwang@bjtu.edu.cn, and zhdzhong@bjtu.edu.cn).

Feifei Gao is with Institute for Artificial Intelligence, Tsinghua University (THUAI), State Key Lab of Intelligent Technologies and SystemsTsinghua University, Beijing National Research Center for Information Science and Technology (BNRist), and Department of Automation, Tsinghua University, Beijing, P.R. China (e-mail: feifeigao@ieee.org).

Arumugam Nallanathan is with School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom (e-mail: a.nallanathan@qmul.ac.uk).

of the main challenges for mmWave NOMA system. Specific channel estimation techniques for mmWave system have been proposed based on compressive sensing (CS), discrete Fourier transform (DFT), and channel covariance matrix [27]–[30]. However the channel estimation methods introduced in [27]–[30] are mainly based on the on-grid approach, which always suffer from the performance loss due to the leakage of energy over some DFT bins. Hence, many researchers also designed the off-grid channel estimation approaches [28], [29].

In this paper, we study the channel estimation and transmission strategy for hybrid mmWave NOMA system. First, we propose a general iterative index detection-based channel estimation algorithm (IDCEA) for mmWave NOMA systems from angle domain viewpoint that can obtain both direction of arrivals (DOAs) and channel gain of each channel path. Then an enhanced hybrid precoding scheme for the mmWave NOMA system is designed to reduce the inter-beam interference with DOA information. After that, we formulate a sum rate maximization problem for the downlink mmWave NOMA systems subject to the user scheduling and power allocation strategy. The problem turns to be non-convex and then we decompose the original optimization into two subproblems as user scheduling and power allocation. For the user scheduling, we develop a matching theory based user scheduling algorithm (MTBUSA) to maximize the achievable sum rate. To reduce the number of iterations, we also utilize the DOAs of different users to design a heuristic initialization user scheduling algorithm. Next, we develop an iterative optimization algorithm to realize the dynamic power allocation under the transmit power constraint.

The rest of the paper is organized as follows. In section II, the transmission model and channel model of mmWave NOMA system with hybrid precoding are described. In section III, we present an iterative IDCEA. The design of hybrid precoding, user scheduling and power allocation are provided in the section IV. Simulation results are then presented in Section V and conclusions are drawn in Section VI.

Notations: Small and upper bold-face letters donate column vectors and matrices, respectively; the superscripts $(\cdot)^H$, $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^{-1}$, $(\cdot)^\dagger$ stand for the conjugate-transpose, transpose, conjugate, inverse, pseudo-inverse of a matrix, respectively; $[\mathbf{A}]_{ij}$ is the (i,j)th entry of $\mathbf{A}$; Diag$\{\mathbf{a}\}$ denotes a diagonal matrix with the diagonal element constructed from $\mathbf{a}$, while Diag$\{\mathbf{A}\}$ denotes a vector whose elements are extracted from the diagonal components of $\mathbf{A}$; $\mathbf{A}(:,i)_{i\in\mathcal{C}}$ denotes the submatrix of $\mathbf{A}$ that consists of the ith column of $\mathbf{A}$ for all $i \in \mathcal{C}$; $\mathcal{R}\{\mathbf{A}\}$ denotes the real part of $\mathbf{A}$; $\mathcal{S}\{\mathbf{A}\}$ denotes the Imaginary part of $\mathbf{A}$; $\dot{g}(x)$ and $\ddot{g}(x)$ denote the first order derivative and the second order derivative of x; $\mathbb{E}\{\cdot\}$ denotes the statistical expectation, and $\|\mathbf{h}\|^2$ is the Euclidean norm of $\mathbf{h}$.

II. System Model

We consider a multiuser mmWave massive MIMO system, where one BS with a hybrid structure simultaneously serve K single antenna users in a small region, as shown in Fig. 1. The BS is composed of M antennas in the form of uniform linear array (ULA) and has M_{RF} RF chains. We denote the displacement between the adjacent antennas in BS as d. Moreover, the BS applies a complex valued based-band digital beamformer $\mathbf{F}_{BB} \in \mathbb{C}^{M_{RF}\times M_{RF}}$, followed by an analog beamformer $\mathbf{F}_{RF} \in \mathbb{C}^{M\times M_{RF}}$.

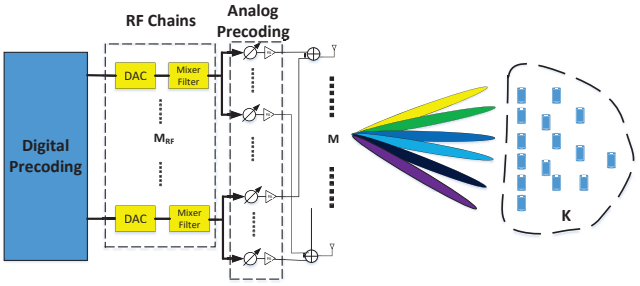

Fig. 1. Simplified hybrid mmWave NOMA system.

A. Downlink Transmission Model

When $K > M_{RF}$, the signal from different users cannot be separated by linear operation in traditional hybrid mmWave communication systems. Nevertheless, we here use NOMA technology to allow all users to transmit simultaneously.

Denote $\mathcal{S}_m$, $m = 1, 2, \cdots, M_{RF}$ as the set of users scheduled on the mth beam to perform NOMA, and $\cup_{m=1}^{M_{RF}} \mathcal{S}_m = \{1, 2, \cdots, K\}$. We also assume that each user is served by a single beam, namely, $\mathcal{S}_m \cap \mathcal{S}_n = \emptyset, m \neq n$. Hence, user u_k^m represents the kth user in the mth beam. For simplicity, the mth beamofrming vector after digital and analog beamforming is denoted as $\mathbf{w}_m = [\mathbf{F}_{RF}\mathbf{F}_{BB}]_{:,m}$.

During downlink transmission, the received signal of u_k^m can be expressed as

$$\begin{aligned} y_k^m =& \mathbf{h}_k^{mH} \sum_{i=1}^{M_{RF}} \sum_{j=1}^{|\mathcal{S}_i|} \mathbf{w}_i \sqrt{p_{j,i}} s_{j,i} + n_k \\ =& \underbrace{\mathbf{h}_k^{mH} \mathbf{w}_m \sqrt{p_{k,m}} s_{k,m}}_{\text{Desired signal}} + \underbrace{\mathbf{h}_k^{mH} \mathbf{w}_m \sum_{j\neq k}^{|\mathcal{S}_m|} \sqrt{p_{j,m}} s_{j,m}}_{\text{Intra-beam interferences}} \\ &+ \underbrace{\mathbf{h}_k^{mH} \sum_{i\neq m}^{M_{RF}} \sum_{j=1}^{|\mathcal{S}_i|} \mathbf{w}_i \sqrt{p_{j,i}} s_{j,i}}_{\text{Inter-beam interference}} + n_k^m, \end{aligned} \tag{1}$$

where $\mathbf{h}_k^m$ means the $M\times 1$ channel vector between the BS and u_k^m, $p_{j,i}$ and $s_{j,i}$ represent the transmit power and downlink signal for u_j^i, $n_k^m \sim \mathcal{CN}(0, \sigma_n^2)$ is additive white Gaussian noise (AWGN) for u_k^m, and σ_n^2 is the unit noise covariance.

Based on (1), the signal-to-interference-plus-noise ratio (SINR) of u_k^m can be expressed as

$$\text{SINR}_k^m = \frac{\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m}}{\xi + \sigma_n^2}, \tag{2}$$

where $\xi = \|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 \sum_{j\neq k}^{|\mathcal{S}_m|} p_{j,m} + \|\mathbf{h}_k^{mH} \sum_{i\neq m}^{M_{RF}} \mathbf{w}_i\|^2 \sum_{j=1}^{|\mathcal{S}_i|} p_{j,i}$. Therefore, the achievable

rate of u_k^m is given by

$$R_k^m = \log_2(1 + \text{SINR}_k^m), \tag{3}$$

and the overall downlink rate of the mmWave NOMA system is

$$R = \sum_{i=1}^{M_{RF}} \sum_{j=1}^{|\mathcal{S}_i|} R_j^i = \sum_{i=1}^{M_{RF}} \sum_{j=1}^{|\mathcal{S}_i|} \log_2(1 + \text{SINR}_j^i). \tag{4}$$

B. Channel Model

In mmWave communications, due to the high free space path loss, the number of dominant paths is very limited [31]–[33]. We use the Saleh-Valenzuela (SV) model to represent the channel, which can show the spatial correlation and sparse characteristics of mmWave communication system [34]–[36], as

$$\mathbf{h}_k^m = \underbrace{a_{1,k}^m \mathbf{a}_{1,k}(\theta_{1,k}^m)}_{\text{LOS path}} + \underbrace{\sum_{l=2}^{L} a_{l,k}^m \mathbf{a}_{l,k}(\theta_{l,k}^m)}_{\text{NOLS paths}}, \tag{5}$$

where L denotes the number of propagation paths, $a_{l,k}^m$ is the complex path gain of the lth channel path of u_k^m. Moreover, $\mathbf{a}_{l,k}(\theta_{l,k}^m)$ is the steering vector of the lth path of u_k^m, defined as

$$\mathbf{a}_{l,k}(\theta_{l,k}^m) = \frac{1}{\sqrt{M}}[1, e^{j2\pi df \sin\theta_{l,k}^m}, \cdots, e^{j2\pi df(M-1)\sin\theta_{l,k}^m}]^T, \tag{6}$$

where f is the carrier frequency, $\theta_{l,k}^m \in [-\frac{\pi}{2}, \frac{\pi}{2})$ is the DOA of the lth path of u_k^m. Note that $l = 1$ represents the line of sight (LOS) component, and $2 \le l \le L$ is the $L-1$ non-line of sight (NLOS) components. Defining $\omega_{l,k}^m = \sin\theta_{l,k}^m \in [-1, 1)$, the steering vector $\mathbf{a}_{l,k}(\theta_{l,k}^m)$ can be equivalently expressed as

$$\mathbf{a}_{l,k}(\omega_{l,k}^m) = \frac{1}{\sqrt{M}}[1, e^{j2\pi df\omega_{l,k}^m}, \cdots, e^{j2\pi df(M-1)\omega_{l,k}^m}]^T. \tag{7}$$

III. Index Detection-based Channel Estimation Algorithm

For time division duplexing (TDD) systems, downlink CSI could be obtained via the uplink channel estimation due to reciprocity. Thus, we focus on the uplink channel estimation to present the basic principle of the proposed IDCEA. During the uplink training stage, the received signal at the BS from u_k^m can be expressed as

$$\mathbf{Y}_k^{m,ul} = \mathbf{F}_{BB}^H \mathbf{F}_{RF}^H \sum_{l=1}^{L} a_{l,k}^m \mathbf{a}_{l,k}(\theta_{l,k}^m)(\mathbf{x}_k^m)^T + \mathbf{N}_k^m, \tag{8}$$

where $\mathbf{x}_k^m = [x_1, x_2, \cdots, x_\tau]^T$ represents the training sequence of u_k^m, $\tau \ge K$ is the length of training sequences, and $\mathbf{N}_k \sim \mathcal{CN}(0, \sigma_n^2\mathbf{I})$ is the AWGN noise matrix.

The BS does not have any prior knowledge before the channel is obtained, so the received signals from K users are inseparable from each other. Therefore, we must use orthogonal pilot sequences to distinguish different users. Define $\mathbf{X} = [\mathbf{x}_1^1, \cdots, \mathbf{x}_{|\mathcal{S}_1|}^1, \cdots \mathbf{x}_1^{M_{RF}}, \cdots, \mathbf{x}_{|\mathcal{S}_{M_{RF}}|}^{M_{RF}}]^T$ as the orthogonal training matrix, satisfied $\|\mathbf{x}_k^m\|^2 = 1$ and $(\mathbf{x}_l^n)^T\mathbf{x}_k^m = 0, l \ne k$ or $n \ne m$. Similar to (8), the received signal matrix $\mathbf{Y}^{ul}$ at the BS can be expressed as

$$\begin{aligned}\mathbf{Y}^{ul} =& \mathbf{F}_{BB}^H \mathbf{F}_{RF}^H \sum_{m=1}^{M_{RF}} \sum_{k=1}^{|\mathcal{S}_m|} \sum_{l=1}^{L} a_{l,k}^m \mathbf{a}_{l,k}(\omega_{l,k}^m)(\mathbf{x}_k^m)^T + \mathbf{N} \\ =& \mathbf{F}_{BB}^H \mathbf{F}_{RF}^H \mathbf{H}\mathbf{X} + \mathbf{N},\end{aligned} \tag{9}$$

where $\mathbf{H} = [\mathbf{h}_1^1, \cdots, \mathbf{h}_{|\mathcal{S}_1|}^1, \cdots \mathbf{h}_1^{M_{RF}}, \cdots, \mathbf{h}_{|\mathcal{S}_{M_{RF}}|}^{M_{RF}}]$, and $\mathbf{N} = \sum_{m=1}^{M_{RF}} \sum_{k=1}^{|\mathcal{S}_m|} \mathbf{N}_k^m$ is the sum noise of K users.

To estimate the channel of each user, we need to separate the received signal from each user first. Since the training sequences of different users are orthogonal, we can obtain the initial channel estimate

$$\begin{aligned}\bar{\mathbf{h}}_k^m =& ((\mathbf{F}_{RF}\mathbf{F}_{BB})(\mathbf{F}_{RF}\mathbf{F}_{BB})^H)^{-1}(\mathbf{F}_{RF}\mathbf{F}_{BB})\mathbf{Y}^{ul}\mathbf{x}_k^m \\ =& \mathbf{H}\mathbf{X}\mathbf{x}_k^m + \tilde{\mathbf{N}} = \mathbf{h}_k^m + \tilde{\mathbf{n}}_k^m \\ =& \sum_{l=1}^{L} a_{l,k}^m \mathbf{a}_{l,k}(\omega_{l,k}^m) + \tilde{\mathbf{n}}_k^m.\end{aligned} \tag{10}$$

where

$$\tilde{\mathbf{n}}_k^m = ((\mathbf{F}_{RF}\mathbf{F}_{BB})(\mathbf{F}_{RF}\mathbf{F}_{BB})^H)^{-1}(\mathbf{F}_{RF}\mathbf{F}_{BB})\mathbf{N}\mathbf{x}_k^m. \tag{11}$$

During the data transmission after uplink training stage, BS and users may not change their physical location in a short period of time, thus the DOA component of the channel can be assumed unchanged over several or even dozens of channel coherence times. When the channel needs to be updated, the BS only needs to re-estimate the remaining channel gain components [37]–[39]. Therefore, after obtaining the initial channel estimation, the next step is to extract the DOAs $\{\theta_k^m\}_{l=1}^L$ for each user.

Due to the sparse characteristics of the mmWave channel, the number of paths L is much less than the number of BS antennas M. Thus, we can treat (10) as the sparse DOA estimation problem, which can be formulated as the following convex optimization problem

$$\begin{aligned}\hat{\mathbf{q}}_k^m =& \arg\min_{\mathbf{q}_k^m} \frac{1}{2}\|\mathbf{A}(\boldsymbol{\vartheta})\mathbf{q}_k^m - \bar{\mathbf{h}}_k^m\|^2 + \lambda\|\mathbf{q}_k^m\|_1 \\ =& \arg\min_{\mathbf{q}_k^m} \frac{1}{2}\|\mathbf{q}_k^m - \mathbf{A}^H(\boldsymbol{\vartheta})\bar{\mathbf{h}}_k^m\|^2 + \lambda\|\mathbf{q}_k^m\|_1,\end{aligned} \tag{12}$$

where $\mathbf{q}_k^m$ is a vector consisting of different $a_{l,k}^m$, $\lambda \in \mathbb{R}_+$ is the regularization parameter determining the sparsity, namely, the number of non-zero elements in the estimated vector $\hat{\mathbf{q}}_k^m$, $\mathbf{A}(\boldsymbol{\vartheta})$ denotes the $M \times N$ complete dictionary matrix which is obtained by sampling the field-of-view angular directions $\boldsymbol{\vartheta} = [\vartheta_1, \vartheta_2, \cdots, \vartheta_N]^T$:

$$\begin{aligned}\mathbf{A}(\boldsymbol{\vartheta}) =& \frac{1}{\sqrt{M}}[\mathbf{a}(\vartheta_1), \mathbf{a}(\vartheta_2), \cdots, \mathbf{a}(\vartheta_N)] \\ =& \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & \cdots & 1 \\ e^{j2\pi df\vartheta_1} & \cdots & e^{j2\pi df\vartheta_N} \\ \vdots & \ddots & \cdots \\ e^{j2\pi df(M-1)\vartheta_1} & \cdots & e^{j2\pi df(M-1)\vartheta_N} \end{bmatrix},\end{aligned} \tag{13}$$

where $N > M$. Note that the whole field-of-view range is $[-1, 1)$, and are divided into N parts uniformly. Given a sparse estimate $\hat{\mathbf{q}}_k^m$, the DOA estimation problem reduces to

the identification of the support set, i.e., the indices of the non-zero elements in $\hat{\mathbf{q}}_k^m$. Thus, the relationship between ϑ_n and the corresponding index q_n can be expressed as

$$\vartheta_n = \frac{2(q_n-1)}{N} - 1, n = 1, 2, \cdots, N. \tag{14}$$

Now we focus on some properties of $\mathbf{A}^H(\boldsymbol{\vartheta})\bar{\mathbf{h}}_k^m = \mathbf{A}^H(\boldsymbol{\vartheta})\mathbf{h}_k^m + \mathbf{A}^H(\boldsymbol{\vartheta})\tilde{\mathbf{n}}_k^m$ in (12). In order to better illustrate the property of $\mathbf{A}^H(\boldsymbol{\vartheta})\mathbf{h}_k^m$, we first consider the case that the channel of u_k^m only have one path, i.e., $\mathbf{h}_k^m = a_{1,k}^m \mathbf{a}_{1,k}(\omega_{1,k}^m)$. Define $\tilde{\mathbf{h}}_k^m = \mathbf{A}^H(\boldsymbol{\vartheta})\mathbf{h}_k^m$ and the qth element of $\tilde{\mathbf{h}}_k^m$ can be computed as

$$\begin{aligned}[\tilde{\mathbf{h}}_k^m]_q =& \mathbf{A}^H(\boldsymbol{\vartheta})\mathbf{h}_k^m = \frac{1}{N}\sum_{n=0}^{N-1} a_{1,k}^m e^{-j2\pi df(n-1)(\vartheta_q - \omega_{1,k}^m)} \\ =& a_{1,k}^m e^{-j\frac{N-1}{2}2\pi df(\vartheta_q-\omega_{1,k}^m)} \cdot \frac{\sin[N\pi df(\vartheta_q-\omega_{1,k}^m)]}{N\sin[\pi df(\vartheta_q-\omega_{1,k}^m)]}.\end{aligned} \tag{15}$$

When N tends to infinity, the elements of $\tilde{\mathbf{h}}_k^m$ can be obtained as

$$[\tilde{\mathbf{h}}_k^m]_q = \begin{cases} a_{l,k}^m & \text{if } \vartheta_q = \omega_{1,k}^m \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

It is shown that $\tilde{\mathbf{h}}_k^m$ only has one non-zero element in the index q, while other elements are all zero. In other words, all power of $\tilde{\mathbf{h}}_k^m$ is concentrated on the index q. Thus, the channel can be equivalent recovered by

$$\mathbf{h}_k^m = [\mathbf{A}(\boldsymbol{\vartheta})]_{:,q}[\tilde{\mathbf{h}}_k^m]_q = \mathbf{a}(\vartheta_q)[\tilde{\mathbf{h}}_k^m]_q. \tag{17}$$

In practice, the N is large but not infinite. Note that there only have N on-grid angle in the whole angle range. Under this circumstances, there can be two cases: 1) $\omega_{1,k}^m = \vartheta_q, q \in \{1, 2, \cdots, N\}$, namely, $\sin\theta_{1,k}^m$ is equal to one on-grid angle ϑ_q exactly. In this special case, the elements distribution of $\tilde{\mathbf{h}}_k^m$ also shown in (16), and all power of $\tilde{\mathbf{h}}_k^m$ is concentrated on a separated single index q; 2) In more general case, $\omega_{1,k}^m \neq \vartheta_q, q \in \{1, 2, \cdots, N\}$, namely, $\sin\theta_{1,k}^m$ does not match any on-grid angle ϑ_q. From (15), we obtain that all the elements of $\tilde{\mathbf{h}}_k^m$ are non-zero, i.e., $[\tilde{\mathbf{h}}_k^m]_q \neq 0, q \in \{1, 2, \cdots, N\}$. Nevertheless, we can find the element containing the maximum power from $\tilde{\mathbf{h}}_k^m$, which is

$$\begin{aligned}\hat{q}_k^m =& \arg\max_{q\in\{1,2,\cdots,N\}} \|[\tilde{\mathbf{h}}_k^m]_q\|^2 \\ =& \arg\min_{q\in\{1,2,\cdots,N\}} |\vartheta_q - \omega_{1,k}^m|.\end{aligned} \tag{18}$$

Based on (18), we learn that the larger $|\vartheta_q - \omega_{1,k}^m|$ is, the less the power $[\tilde{\mathbf{h}}_k^m]_q$ is. Therefore, the power of $\tilde{\mathbf{h}}_k^m$ will leak from the central point q_k^m to its nearby elements, where q_k^m can be obtained by (18). The element that contains the largest power of $\mathbf{A}^H(\boldsymbol{\vartheta})\bar{\mathbf{h}}_k^m$ can be obtained from

$$q_{1,k}^m = \arg\max_{q_{1,k}^m\in\{1,2,\cdots,M\}} |[\mathbf{A}^H(\boldsymbol{\vartheta})\bar{\mathbf{h}}_k^m]_{q_{1,k}^m}|^2. \tag{19}$$

Thus, the coarsely estimated DOA can be obtained from

$$\hat{\theta}_{1,k}^m = \arcsin\omega_{1,k}^m = \arcsin\left(\frac{2(\hat{q}_{1,k}^m - 1)}{M} - 1\right), \tag{20}$$

and the coarsely estimated path gain is given by

$$\hat{a}_{1,k}^m = \frac{\mathbf{a}^H(\vartheta_{q_{1,k}^m})\bar{\mathbf{h}}_k^m}{\|\mathbf{a}^H(\vartheta_{q_{1,k}^m})\|^2}. \tag{21}$$

After obtain the coarsely estimated DOA, our goal is to find the more accurate group of channel component $(a_{1,k}^m, \omega_{1,k}^m)$ that satisfy

$$\begin{aligned}(\hat{a}_{1,k}^m, \hat{\omega}_{1,k}^m) =& \arg\min_{(a_{1,k}^m,\omega_{1,k}^m)} \|\bar{\mathbf{h}}_k^m - a_{1,k}^m \mathbf{a}_{1,k}(\omega_{1,k}^m)\|^2 \\ =& \arg\max_{(a_{1,k}^m,\omega_{1,k}^m)} |a_{1,k}^m|^2 2\mathcal{R}\{a_{1,k}^m (\bar{\mathbf{h}}_k^m)^H \mathbf{a}_{1,k}(\omega_{1,k}^m) \\ & - \|\mathbf{a}_{1,k}(\omega_{1,k}^m)\|^2\} \\ =& \arg\max_{(a_{1,k}^m,\omega_{1,k}^m)} x(a_{1,k}^m, \omega_{1,k}^m),\end{aligned} \tag{22}$$

where $x(a_{1,k}^m, \omega_{1,k}^m) = |a_{1,k}^m|^2 2\mathcal{R}\{a_{1,k}^m (\bar{\mathbf{h}}_k^m)^H \mathbf{a}_{1,k}(\omega_{1,k}^m) - \|\mathbf{a}_{1,k}(\omega_{1,k}^m)\|^2$ denotes the cost function of $(a_{1,k}^m, \omega_{1,k}^m)$.

Since $\omega_{1,k}^m \in [-1, 1)$ is a monotonically increasing function, we can use the following equation to refine the coarsely estimated $\hat{\omega}_{1,k}^m$, which is shown as (23).

Note that we only carry the refinement function (23) if and only if $\ddot{x}(\hat{a}_{1,k}^m, \hat{\omega}_{1,k}^m) < 0$, namely, the function (22) is locally concave in the range of $|\vartheta_{q_{1,k}^m} - \omega_{1,k}^m| < \frac{1}{2N}$. After refine $\hat{\omega}_{1,k}^{m\,\prime}$, the estimated DOA $\hat{\theta}_{1,k}^{m\,\prime}$ is obtained by $\hat{\theta}_{1,k}^{m\,\prime} = \arcsin(\hat{\omega}_{1,k}^{m\,\prime})$, and the path gain is also updated from (21).

Remark 1. *The authors in [9] introduced a simple effective angle rotation based DOA estimation method for a massive ULA through DFT and then searches for accurate estimates within a very small region. However, the accuracy of angle rotation based estimation depends on the searching grid and it is not possible to obtain the true angle with a limited searching grid. Therefore, we propose a novel DOA estimation method* (23) *that can eliminate the impact of the resolution of search grid.*

In multiple paths case, we assume the channel of kth user in the mth beam has L paths, i.e., $\mathbf{h}_k^m = \sum_{l=1}^L a_{l,k}^m \mathbf{a}_{l,k}(\omega_{l,k}^m) = \sum_{l=1}^L \mathbf{h}_{l,k}^m$, where $\mathbf{h}_{l,k}^m$ is the lth channel component of channel $\mathbf{h}_k^m$. Then we have

$$\begin{aligned}&(\mathbf{A}^H(\boldsymbol{\vartheta})\mathbf{h}_{i,k}^m)^H \mathbf{A}^H(\boldsymbol{\vartheta})\mathbf{h}_{j,k}^m \\ =& \mathbf{h}_{i,k}^{m\,H}\mathbf{A}(\boldsymbol{\vartheta})\mathbf{A}^H(\boldsymbol{\vartheta})\mathbf{h}_{j,k}^m \\ =& a_{i,k}^m a_{j,k}^m \mathbf{a}^H(\omega_{i,k}^m)\mathbf{a}(\omega_{j,k}^m) \\ =& a_{i,k}^m a_{j,k}^m e^{-j\frac{M-1}{2}2\pi df(\omega_{i,k}^m - \omega_{j,k}^m)} \cdot \frac{\sin[M\pi df(\omega_{i,k}^m - \omega_{j,k}^m)]}{M\sin[\pi df(\omega_{i,k}^m - \omega_{j,k}^m)]}.\end{aligned} \tag{24}$$

Based on (24), we can get

$$\begin{aligned}0 \leq& \lim_{M\to\infty} \frac{a_{i,k}^m a_{j,k}^m e^{-j\frac{M-1}{2}2\pi df(\omega_{i,k}^m - \omega_{j,k}^m)} \sin[M\pi df(\omega_{i,k}^m - \omega_{j,k}^m)]}{M\sin[\pi df(\omega_{i,k}^m - \omega_{j,k}^m)]} \\ \leq& \left|\frac{a_{i,k}^m a_{j,k}^m e^{-j\frac{M-1}{2}2\pi df(\omega_{i,k}^m - \omega_{j,k}^m)}}{M\sin\pi Mdf(\omega_{i,k}^m - \omega_{j,k}^m)}\right| = 0.\end{aligned} \tag{25}$$

Note that the different path components of u_k^m channel can be treated as approximately orthogonal when the number of

$$\hat{\omega}_{1,k}^{m\,\prime}=\hat{\omega}_{1,k}^{m}-\frac{\dot{x}(\hat{a}_{1,k}^{m},\hat{\omega})}{\ddot{x}(\hat{a}_{1,k}^{m},\hat{\omega})}=\hat{\omega}_{1,k}^{m}-\frac{2\mathcal{R}\left\{(a_{1,k}^{m}(\bar{\mathbf{h}}_{k}^{m})^{H}-|a_{1,k}^{m}|^{2}\mathbf{a}^{H}(\omega))(\partial\mathbf{a}(\omega)/\partial\omega)\right\}}{2\mathcal{R}\left\{(a_{1,k}^{m}(\bar{\mathbf{h}}_{k}^{m})^{H}-|a_{1,k}^{m}|^{2}\mathbf{a}^{H}(\omega))(\partial^{2}\mathbf{a}(\omega)/\partial\omega^{2})\right\}-|a_{1,k}^{m}|^{2}\|\partial\mathbf{a}(\omega)/\partial\omega\|^{2}}. \tag{23}$$

antennas in BS is large. Therefore, we can use the iterative IDCEA to obtain the independent channel parameters $\theta_{l,k}^{m}$ and $a_{l,k}^{m}$ for $l=1,2,\cdots,L, k=1,2,\cdots,K$. The basic principles of IDCEA are as follows: The DOA and path gain of the strongest channel component can be estimated firstly. After that, the estimated strongest channel component can be removed from the total channel and then the DOA and path gain of the second strongest channel component can be estimated. After the Lth iteration, the DOAs and paths gains for all L channel components have been estimated.

In the proposed IDCEA, after a group parameters of channel component $(\hat{a}_{i-1,k}^{m},\hat{\omega}_{i-1,k}^{m})$ is estimated in the $(i-1)$th iteration, we should remove the corresponding component from $\bar{\mathbf{h}}_{k}$. Namely, in the ith iteration, the residual received signal is given by

$$\bar{\mathbf{h}}_{k}^{m}(i)=\bar{\mathbf{h}}_{k}^{m}-\sum_{l=1}^{i-1}\hat{a}_{l,k}^{m}\hat{\mathbf{a}}_{l,k}(\hat{\omega}_{l,k}^{m}), \tag{26}$$

where $\{(\hat{a}_{l,k}^{m},\hat{\omega}_{l,k}^{m})\}_{l=1}^{i-1}$ are the estimated parameters in the previous $(i-1)$ iterations.

In practice mmWave communication system, the BS does not know the number of propagation paths. Thus, we need to set a threshold to stop the iteration. One feasible method is to check the power of the residual $\bar{\mathbf{h}}_{k}^{m}(i)$. If the residual power is less than the total noise power, i.e.,

$$\|\bar{\mathbf{h}}_{k}^{m}(i)\|^{2}<\kappa=\mathbb{E}\{\|\tilde{\mathbf{n}}_{k}^{m}\|^{2}\}=M\sigma_{n}^{2}, \tag{27}$$

the estimated algorithm will be stopped and the estimated number of paths $\hat{L}$ is equal to the iteration times i. After DOAs of all paths have been detected from (23), we can use the classical least square (LS) algorithm to re-estimate the path gains of channel. From (10), we have

$$\bar{\mathbf{h}}_{k}^{m}=\sum_{l=1}^{L}a_{l,k}^{m}\mathbf{a}_{l,k}(\omega_{l,k}^{m})+\tilde{\mathbf{n}}_{k}^{m}=\mathbf{A}_{k}^{m}\mathbf{g}_{k}^{m}+\tilde{\mathbf{n}}_{k}^{m}, \tag{28}$$

where $\mathbf{A}_{k}^{m}=[\mathbf{a}_{1,k}(\omega_{1,k}^{m}),\mathbf{a}_{2,k}(\omega_{2,k}^{m}),\cdots,\mathbf{a}_{L,k}(\omega_{L,k}^{m})]$, and $\mathbf{g}_{k}^{m}=[a_{1,k}^{m},a_{2,k}^{m},\cdots,a_{L,k}^{m}]^{T}$. Note that the steering matrix $\mathbf{A}_{k}^{m}$ are known at the BS after completing the DOA estimate. Then the BS can refine the channel gains as

$$\hat{\mathbf{g}}_{k}^{m}=\mathbf{A}_{k}^{m\dagger}\bar{\mathbf{h}}_{k}^{m}+\mathbf{A}_{k}^{m\dagger}\tilde{\mathbf{n}}_{k}^{m}. \tag{29}$$

With the DOA information from (23) and the channel gains information from (29), we can obtain the uplink channel estimation for all users as

$$\hat{\mathbf{h}}_{k}^{m}=\sum_{l=1}^{L}\hat{a}_{l,k}^{m}\hat{\mathbf{a}}_{l,k}(\theta_{l,k}^{m}). \tag{30}$$

Based on the discussion so far, the proposed IDCEA can be summarized in **Algorithm** 1. Note that we use the iteration method instead of directly estimating the channel $\mathbf{h}_{k}$ since the power of each index will be influenced by all path components. Thus we cannot find the L indices directly at the same time which contain L largest powers. Especially when L is large and the DOAs of all paths are very close. Using the iterative procedure, we can obtain the largest power index in each iteration and remove it for the next path component estimation.

Algorithm 1 Index detection-based channel estimation algorithm

Input: Received vector: $\mathbf{Y}^{ul}$ in (9); Digital beamforming matrix: $\mathbf{F}_{BB}$; Analog beamforming matrix: $\mathbf{F}_{RF}$; Training sequence: $\mathbf{x}_{k}^{m}$; Stoping threshold: κ;

1. Initialization: $i=1$; $\bar{\mathbf{h}}_{k}^{m}(i)=\bar{\mathbf{h}}_{k}$ from (10);

2. Detect the index $q_{i,k}^{m}$ of $\mathbf{A}^{H}(\boldsymbol{\vartheta})\bar{\mathbf{h}}_{k}^{m}(i)$, which contain the largest power, as $q_{i,k}^{m}=\arg\max_{q_{i,k}^{m}\in\{1,2,\cdots,M\}}\|[\mathbf{A}^{H}(\boldsymbol{\vartheta})\bar{\mathbf{h}}_{k}^{m}(i)]_{q_{i,k}^{m}}\|^{2}$;

3. Calculate the coarsely estimated $\omega_{i,k}^{m}=\frac{2(q_{i,k}^{m}-1)}{M}-1$, and the coarsely estimated path gain according to (21);

4. Refine $(a_{i,k}^{m},\omega_{i,k}^{m})$ from (23) and (21);

5. Remove the influence of the estimated path component and update $\bar{\mathbf{h}}_{k}^{m}(i+1)$ as $\bar{\mathbf{h}}_{k}^{m}(i+1)=\bar{\mathbf{h}}_{k}^{m}-\sum_{l=1}^{i}\hat{a}_{l,k}^{m}\hat{\mathbf{a}}_{l,k}(\hat{\theta}_{l,k}^{m})$,

6. If the residual power $\|\bar{\mathbf{h}}_{k}^{m}(i+1)\|^{2}>\kappa$, update $i=i+1$ and go back to **Step 2**.

7. Update all path gains according to (29).

Output: Estimated channel $\hat{\mathbf{h}}_{k}^{m}$ of u_{k}^{m} from (30).

IV. Transmission strategy in mmWave NOMA systems

A. Hybrid Beamforming Design

The traditional linear zero-forcing (ZF) beamforming can be applied to do the classical hybrid mmWave massive MIMO systems, where each beam only serve one user at the same time-frequency resource such that there does not exist the inter-beam interference. However, in the considered mmWave NOMA system, the number of users is larger than the number of beams that provided by limited RF chains. In this scenario, the ZF beamforming cannot be used directly due to the pseudo-inverse of the matrix of size $M_{RF}\times K$ does not exist.

To address this problem, we use a representative channel for each beam to design the analog and digital precoding matrices. Note that the channel of users with similar DOA information are highly correlated. In other words, the directions of all channel vectors of different users in the same beam are considered the same. We assume that $\|\mathbf{h}_{1}^{m}\|^{2}\geq\|\mathbf{h}_{2}^{m}\|^{2}\geq\cdots,\|\mathbf{h}_{|\mathcal{S}_{m}|}^{m}\|^{2}$, for $m=1,2,\cdots,M_{RF}$. Therefore, we use the channel vector which has the largest channel power as the representative channel vector in each beam. Then, the representative channel matrix can be given by $\bar{\mathbf{H}}=[\mathbf{h}_{1}^{1},\mathbf{h}_{1}^{2},\cdots,\mathbf{h}_{1}^{M_{RF}}]$.

$$\text{SINR}_{i\to k}^{m}=\frac{\|\mathbf{h}_k^{m\,H}\mathbf{w}_m\|^2 p_{k,m}}{\|\mathbf{h}_k^{m\,H}\mathbf{w}_m\|^2\sum_{j>k}^{|\mathcal{S}_m|}p_{j,m}+\|\mathbf{h}_k^{m\,H}\sum_{l\neq m}^{M_{RF}}\mathbf{w}_l\|^2\sum_{j=1}^{|\mathcal{S}_l|}p_{j,l}+\sigma_n^2}, \tag{35}$$

$$\text{SINR}_{i\to i}^{m}=\frac{\|\mathbf{h}_i^{m\,H}\mathbf{w}_m\|^2 p_{i,m}}{\|\mathbf{h}_i^{m\,H}\mathbf{w}_m\|^2\sum_{j>i}^{|\mathcal{S}_m|}p_{j,m}+\|\mathbf{h}_i^{m\,H}\sum_{l\neq m}^{M_{RF}}\mathbf{w}_l\|^2\sum_{j=1}^{|\mathcal{S}_l|}p_{j,l}+\sigma_n^2}. \tag{36}$$

From the previous discussion, the analog precoding matrix can be immediately obtained as

$$\mathbf{F}_{RF}=\mathbf{A}(\boldsymbol{\vartheta})(:,i)_{i\in\mathcal{C}}, \tag{31}$$

where the set $\mathcal{C}=\{q_1^1,q_1^2,\cdots,q_1^{M_{RF}}\}$ is the index set of all beams, and q_1^m is the index of $\bar{\mathbf{h}}_1^m$ that contains the maximum power. Each column of $\mathbf{F}_{RF}$ represents the selected spatial support of the representative channel. Note that BS can only separate the whole spatial space into M beams, and each RF chain corresponds to a single beam.

Similar to the conventional digital precoding problem, $\mathbf{F}_{BB}$ can be generated via the ZF precoding by

$$\mathbf{F}_{BB}=(\bar{\mathbf{H}}^H\mathbf{F}_{RF})^{\dagger}=(\bar{\mathbf{H}}^H\mathbf{F}_{RF})((\bar{\mathbf{H}}^H\mathbf{F}_{RF})^H(\bar{\mathbf{H}}^H\mathbf{F}_{RF}))^{-1}. \tag{32}$$

Hence, the equivalent precoding matrix can be obtained by

$$\begin{aligned}\mathbf{W}=&[\mathbf{w}_1',\mathbf{w}_2',\cdots,\mathbf{w}_{M_{RF}}']\\ =&\mathbf{A}(\boldsymbol{\vartheta})(:,i)_{i\in\mathcal{C}}(\bar{\mathbf{H}}^H\mathbf{F}_{RF})((\bar{\mathbf{H}}^H\mathbf{F}_{RF})^H(\bar{\mathbf{H}}^H\mathbf{F}_{RF}))^{-1}.\end{aligned} \tag{33}$$

The equivalent precoding vector of the mth beam can be normalized as

$$\mathbf{w}_m=\frac{\mathbf{w}_m'}{\|\mathbf{w}_m'\|^2}. \tag{34}$$

It is worth noting that different beams are mutually orthogonal, namely, $\mathbf{w}_m^H\mathbf{w}_n=0$.

After obtaining the equivalent precoding vectors, the task is to use the SIC to optimize user grouping and power allocation in the mmWave NOMA systems, which will be discussed next.

B. Problem Formulation

Since each beam serves multiple users in the mmWave NOMA systems, each user in the same beam should perform SIC in a successive order to remove the intra-beam interference. The SINR of u_k^m after decode u_i^m can be written as (35), while the SINR of user i can be written as (36). In order to guarantee the success of SIC, we need to ensure $\text{SINR}_{i\to k}^m\geq\text{SINR}_{i\to i}^m$. Without loss of generality, we assume that $\|\mathbf{h}_1^m\mathbf{w}_m\|^2,\geq\|\mathbf{h}_2^m\mathbf{w}_m\|^2\geq\cdots,\geq\|\mathbf{h}_{|\mathcal{S}_m|}^m\mathbf{w}_m\|^2$ for $m=1,2,\cdots,M_{RF}$. Therefore, for any two users $i<k$ in the mth beam, user k can detect the ith user and then remove the detected signal from its received signal.

In the considered mmWave NOMA system, the BS needs to maximize the sum data rates while guaranteeing the total power constraint and the quality of Service (QOS) constraints for each user. To this end, the BS needs to optimally design the user scheduling scheme and the power allocation coefficients, i.e., $\mathcal{S}_m$ and $p_{k,m}$ for $m=\{1,2,\cdots,M_{RF}\}$, $k=\{1,2,\cdots,|\mathcal{S}_m|\}$. Therefore, the optimization problem can be formulated as

$$\max_{\mathcal{S},\rho}\sum_{m=1}^{M_{RF}}\sum_{k=1}^{\mathcal{S}_m}R_{k\to k}^m \tag{37}$$

$$\text{s. t. } p_{k,m}\geq 0,\qquad\sum_{m=1}^{M_{RF}}\sum_{k=1}^{\mathcal{S}_m}p_{k,m}\leq P, \tag{38}$$

$$\sum_{m=1}^{M_{RF}}|\mathcal{S}_m|=K,\qquad\mathcal{S}_i\cap\mathcal{S}_j=\emptyset, \tag{39}$$

$$\text{SINR}_{k\to k}^m\geq\text{SINR}_{min}, \tag{40}$$

where $\mathcal{S}=\{\mathcal{S}_1,\mathcal{S}_2,\cdots,\mathcal{S}_{M_{RF}}\}$ represents the optimal user scheduling, $\rho=\{p_{k,m}\},m=\{1,2,\cdots,M_{RF}\}$, and $k=\{1,2,\cdots,|\mathcal{S}_m|\}$ are the optimal power allocation coefficients. The constraint (38) indicates that the power of each user must be positive and satisfy total transmit power constraint P; (39) is the user scheduling constraint that each user could only be served by one beam, and all users are served simultaneously; (40) is the QOS constraint that the SINR of each user should be larger than the minimum SINR.

Note that problem (37) is an NP hard problem, which is difficult to solve directly. Then, we decompose the original optimization problem into two-step optimization which separately consider user scheduling and power allocation. The former can be solved by a many-to-one MTBUSA while the latter can be solved by an iterative optimization algorithm.

C. Matching Theory Based User Scheduling Algorithm

For the user scheduling sub-problem, the power allocation coefficients are assumed to be fixed in each beam. Thus, the optimal sub-problem of user scheduling can be formulated as

$$\max_{\mathcal{S}}\sum_{m=1}^{M_{RF}}\sum_{k=1}^{\mathcal{S}_m}R_{k\to k}^m \tag{41}$$

$$\text{s. t. }\sum_{m=1}^{M_{RF}}|\mathcal{S}_m|=K,\qquad\mathcal{S}_i\cap\mathcal{S}_j=\emptyset. \tag{42}$$

Note that the formulated sub-problem is also non-convex due to the existence of the interference term in the objective function. A viable way is to perform an exhaustively search. However, the complexity of the exhaustive search method increases exponentially with the number of users and the number of beams, which is impractical.

Fortunately, the beams and users can be treated as two sets of independent players and interact with each other to maximize the sum rate, because each user is only served by one beam and all users must be served simultaneously by all beams. Hence, user scheduling can be considered as a two-sided many-to-one matching process between the sets of users and the sets of beams. It is well known that matching theory provides mathematically low-complexity and tractable solutions for the combinatorial problem of matching players in two distinct sets [41]–[43]. Next, we formulate the user scheduling sub-problem as a many-to-one matching problem and propose a novel MTBUSA.

Let us first define some notations for the matching model between the set of users and the set of beams.

Definition 1. *Define a function $f(k)$ from the set of users $\mathcal{S}$ to the set of beams $\mathcal{M}$ by $f(k) = m$ and a function $f^{-1}(m)$ from the set of beams $\mathcal{M}$ to the set of users $\mathcal{S}$ by $f^{-1}(m) = \mathcal{S}_m$ in the many-to-one matching model. Then we have: 1) $|f(k)| = 1, \forall k \in \mathcal{S}$; 2) $f(k) = m, k \in \mathcal{S}_m$, for $\forall m \in \mathcal{M}$, where $|\mathcal{S}_m|$ is a positive quota which indicates the number of users that can be matched with the mth beam; 3) $|f^{-1}(m)| = |\mathcal{S}_m| \geq 1$, i.e., beam can support multiple users at the same time-frequency resources; 4) If $m = f(k)$, then we have $k \in f^{-1}(m)$.*

From (41), the goal of user scheduling is to maximize the overall achievable rate. So we treat the achievable sum rate as the preference value for users and beams. Thus, the preference value of the kth user in the mth beam (k, m) can be obtained from (3) and (36), while the preference value of the mth beam can be expressed as

$$R^m = \sum_{k \in \mathcal{S}_m} R^m_{k \to k}. \tag{43}$$

Due to the achievable rate of each user is affected by the inter-beam interference and the intra-beam interference, the preference value of users not only depends on the beam that it matches with and the set of users that are matched to the same beam, but also depends on the set of users that are matched to the other beams. Similarly, the preference value of beams both depends on the users that it serve and the users that the other beams support. Therefore, the proposed matching model exists peer effects, where each user has a dynamic preference list over the opposite set of users. Note that this is different from the conventional matching model in which users have fixed preference lists.

Define (k, m) as a matching state between the user k and the beam m, and the relationship among the preference value for the kth user in any beams under different matching states as

$$(k, f(k)) \succ_k (k, f'(k)) \Leftrightarrow R^m_{k \to k} > R^{m'}_{k \to k}, \tag{44}$$

where $f(k) = m$ represents the user k matching the beam m, $f'(k) = m'$ represents the user k matching the beam m', and $\succ_k$ denotes the preference value ordering of users. It is worth noting that if user k can achieve a higher data rate on beam m than beam m', then user k prefers beam m to beam m'.

Similarly, the relationship of the preference value for the beam m under different matching states can be expressed as

$$\begin{aligned} &(f^{-1}(m), m) \succ_m (f'^{-1}(m), m) \\ \Leftrightarrow &\sum_{k \in \mathcal{S}_m} R^m_{k \to k} > \sum_{k \in \mathcal{S'}_m} R^m_{k \to k}, \end{aligned} \tag{45}$$

where $f^{-1}(m) = \mathcal{S}_m$ and $f^{-1}(m') = \mathcal{S'}_m$ are two matching set of users, and $\succ_m$ denotes the preference value ordering of beams. It is worth noting that if beam m can achieve a higher data rate from the set of users $\mathcal{S}_m$, then the beam m prefers the set of users $\mathcal{S}_m$ to the set of users $\mathcal{S'}_m$.

To better describe the exchange stability, we first define the concept of matching swap between user k and k' as follows:

$$\Omega^{k'}_k = \{\Omega / \{(k, m), (k', m')\} \cup \{(k, m'), (k', m)\}\}, \tag{46}$$

where Ω represents the matching pair of users and beams, and (k, m) means user k matches beam m. Note that user k and user k' switch beams while keeping other users and beams matching unchanged. Furthermore, based on the concept of matching swap, the conditions for the swap are defined in the following.

Definition 2. *If the matching pair satisfies the following condition:*
1) $R^{m'}_{k \to k} \geq R^m_{k \to k}$ and $R^m_{k' \to k'} \geq R^{m'}_{k' \to k'}$;
2) $\sum_{x \in \mathcal{S}_m / \{k\} \cup \{k'\}} R^m_{x \to x} > \sum_{x \in \mathcal{S}_m} R^m_{x \to x}$ and $\sum_{x \in \mathcal{S}_{m'} / \{k'\} \cup \{k\}} R^{m'}_{x \to x} > \sum_{x \in \mathcal{S}_{m'}} R^{m'}_{x \to x}$;
then the matching swap between user k and user k' can be approved and (k, k') is defined as a swap pair.

The two conditions of the above definition imply that the achievable data rate of any user should not be decreased after the swap operation between the swap pair (k, k'). Moreover, the achievable data rate of beams are also increased after the swap operation. Therefore, these two conditions can avoid cyclical swap operation between different matching states where the preference value of all other users are indifferent. It is worth noting that the sum rate of the system will increases after each swap operation is completed.

The total number of candidate swap pairs is closely related to the number of users and the number of beams. Every two users matching different beams can be arranged as a candidate swap pair. The BS will check all swap pairs whether they satisfy the swap conditions by exchanging their matched beams. After a series of swap operations, the states will be stable and there will not exist a feasible swap pair. In this stable state, the user and beam matching scheme is the user scheduling solution that we are searching for.

Next we propose a MTBUSA between users and beams based on multiple swap operations. The details of the proposing algorithm are shown in **Algorithm** 2. The input of the MTBUSA is the initial user scheduling state, which will be shown in **Algorithm** 3. The main process of the proposed algorithm is the swap operation between different users, where each user searches for all other users that match different beams to determine whether there exist an operational swap pair. The final matching state will output until all swap operation done and there are no swap pairs.

Algorithm 2 Matching theory based user scheduling algorithm

Input: Initial user scheduling state: Ω, $\Omega' = \Omega$
Output: Final optimal user scheduling state: Ω'.
Swap Operations:
1 Repeat
2. **For** $\forall k \in K$
3. **For** $\forall k'$ do not match the same beam with user k, i.e., $\forall k' \in \mathcal{S}/\mathcal{S}_m$, where $k \in \mathcal{S}_m$
4. **If** (k, k') is a swap pair
5. $\Omega' = \{\Omega'/\{(k,m),(k',m')\} \cup \{(k,m'),(k',m)\}\}$, where $m' = f(k')$.
6. **End If**
7. **End For**
8. **End For**
9. **Until** no swap pairs in the current matching state Ω'.

Note that the complexity of the proposed MTBUSA mainly lies in the times of swap operation. If the initial user scheduling state can already reach a high achievable sum rate, then it will need less swap operations to achieve the stable status. Therefore, we can design a heuristic initialization user scheduling algorithm to reduce the complexity of MTBUSA, which can provide a good initial matching state.

Since all users are randomly distributed within a small region, more than one users could have the same or adjacent DOAs. We here provide a greedy initial user scheduling algorithm based on the index set $\mathcal{C}$ of the selected beams. Without loss of generality, each selected beam has a one-to-one correspondence with a specific angle and this particular correspondence can be obtained from (20). For each user, we need to find the matching beam that maximizes the achievable rate. From (15), it is worth pointing out that the equivalent channel gain after hybrid precoding is inversely proportional to $|\vartheta_q - \omega_k|$. Thus, the user k should match beam m if

$$m = \arg\min_{q_m \in \mathcal{C}} |\vartheta_{q_m} - \omega_k|. \tag{47}$$

The detailed steps of initial user scheduling can be found in **Algorithm** 3.

Algorithm 3 Initial user scheduling algorithm

Input: DOA information for each user $\theta_1, \theta_2, \cdots, \theta_K$, the index set $\mathcal{C}$ of the selected beams
Output: Initial user scheduling state: Ω.
1 Initialize $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_{M_{RF}}\}$, $\mathcal{S}_m = \emptyset$, for $m = 1, 2, \cdots, M_{RF}$.
2. **For** $\forall k \in K$
3. Calculate the optimal matched beam from $m = \arg\min_{q_m \in \mathcal{C}} |\vartheta_{q_m} - \omega_k|$.
4. Update the set of users of beam m as $\mathcal{S}_m = \mathcal{S}_m \cup \{k\}$.
5. **End For**

D. Power Allocation

In this subsection, we solve the sub-problem of power allocation for a given user scheduling scheme. By substituting (36) into (40), the power allocation problem can be formulated as

$$\max_{\rho} \sum_{m=1}^{M_{RF}} \sum_{k=1}^{\mathcal{S}_m} R_{k\to k}^m \tag{48}$$

$$\text{s. t. } p_{k,m} \geq 0, \quad \sum_{m=1}^{M_{RF}} \sum_{k=1}^{\mathcal{S}_m} p_{k,m} \leq P,$$

$$\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m} - \text{SINR}_{min}\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 \sum_{j>k}^{|\mathcal{S}_m|} p_{j,m}$$

$$- \text{SINR}_{min}\|\mathbf{h}_k^{mH} \sum_{l\neq m}^{M_{RF}} \mathbf{w}_l\|^2 \sum_{j=1}^{|\mathcal{S}_l|} p_{j,l} \geq \sigma_n^2 \text{SINR}_{min}.$$

Since the achievable sum rate is affected by the inter-beam interference and the intra-beam interference, the power allocation problem is non convex. Therefore, it is rather difficult to obtain the closed-form of the global optimal power assignment with affordable complexity.

Then, we replace the objective function with an equivalent formula, as

$$R_{k\to k}^m = \log(1 + \text{SINR}_{k\to k}^m) = -\log(1 + \text{SINR}_{k\to k}^m)^{-1}. \tag{49}$$

According to the extension of the Sherman-Morrison-Woodbury formula [44]

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I} + \mathbf{CDA}^{-1}\mathbf{B})^{-1}\mathbf{CDA}^{-1}, \tag{50}$$

we have equition (51).

Define a convex function $g(x) = \frac{1}{\ln 2} - \frac{x}{x_0 \ln 2} + \log_2 x$, where $x > 0$ and x_0 is the minimum value of x. Then we have $\text{argmax}_{x>0}\, g(x) = -\log_2 x_0^{-1}$. Therefore, we can define a new variable $c_{k.m} > 0$, and rewrite (49) as

$$R_{k\to k}^m = \underset{c_{i,m}>0}{\text{argmax}} \frac{1}{\ln 2} - \frac{c_{k,m}}{(1 + \text{SINR}_{k\to k}^m)\ln 2} + \log_2 c_{k,m}. \tag{52}$$

Substituting (52) into (48), the objective function for the power allocation problem can be transformed into

$$\max_{\rho} \sum_{m=1}^{M_{RF}} \sum_{k=1}^{\mathcal{S}_m} \underset{c_{i,m}>0}{\text{argmax}} \left(\frac{1}{\ln 2} \right.$$

$$\left. - \frac{c_{k,m}}{(1 + \text{SINR}_{k\to k}^m)\ln 2} + \log_2 c_{k,m} \right) \tag{53}$$

$$\text{s. t. } p_{k,m} \geq 0, \quad \sum_{m=1}^{M_{RF}} \sum_{k=1}^{\mathcal{S}_m} p_{k,m} \leq P,$$

$$\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m} - \text{SINR}_{min}\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 \sum_{j>k}^{|\mathcal{S}_m|} p_{j,m}$$

$$- \text{SINR}_{min}\|\mathbf{h}_k^{mH} \sum_{l\neq m}^{M_{RF}} \mathbf{w}_l\|^2 \sum_{j=1}^{|\mathcal{S}_l|} p_{j,l} \geq \sigma_n^2 \text{SINR}_{min}.$$

Then, we propose an iterative algorithm to solve the optimization problem (53). When we get the optimal power allocation solution $\rho^{(t)}$ in the tth iteration, the corresponding

$$(1+\text{SINR}_{k\to k}^{m})^{-1} = 1 - \frac{\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m}}{\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m} + \|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 \sum_{j>k}^{|\mathcal{S}_m|} p_{j,m} + \|\mathbf{h}_k^{mH} \sum_{l\neq m}^{M_{RF}} \mathbf{w}_l\|^2 \sum_{j=1}^{|\mathcal{S}_l|} p_{j,l} + \sigma_n^2}. \tag{51}$$

$$(1+\text{SINR}_{k\to k}^{m})^{-1,(t+1)} = 1 - \frac{\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m}^{(t)}}{\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m}^{(t)} + \|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 \sum_{j>k}^{|\mathcal{S}_m|} p_{j,m}^{(t)} + \|\mathbf{h}_k^{mH} \sum_{l\neq m}^{M_{RF}} \mathbf{w}_l\|^2 \sum_{j=1}^{|\mathcal{S}_l|} p_{j,l}^{(t)} + \sigma_n^2}. \tag{54}$$

$(1+\text{SINR}_{k\to k}^{m})^{-1}$ in the $(t+1)$th iteration can be expressed as (54).

Then the optimal $c_{k,m}^{(t+1)}$ in the $(t+1)$th iteration is given by

$$c_{k,m}^{(t+1)} = (1+\text{SINR}_{k\to k}^{m})^{(t+1)}. \tag{55}$$

Thus, we can rewrite the equivalent objective function in (53), as

$$\min_{\rho} \sum_{m=1}^{M_{RF}} \sum_{k=1}^{\mathcal{S}_m} c_{k,m}^{(t+1)} (1+\text{SINR}_{k\to k}^{m})^{-1,(t+1)} \tag{56}$$

$$\text{s. t. } p_{k,m} \geq 0, \qquad \sum_{m=1}^{M_{RF}} \sum_{k=1}^{\mathcal{S}_m} p_{k,m} \leq P,$$

$$\|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 p_{k,m} - \text{SINR}_{min} \|\mathbf{h}_k^{mH}\mathbf{w}_m\|^2 \sum_{j>k}^{|\mathcal{S}_m|} p_{j,m}$$

$$- \text{SINR}_{min} \|\mathbf{h}_k^{mH} \sum_{l\neq m}^{M_{RF}} \mathbf{w}_l\|^2 \sum_{j=1}^{|\mathcal{S}_l|} p_{j,l} \geq \sigma_n^2 \text{SINR}_{min}.$$

It is worth noting that problem (56) is convex and it can be solved by a standard convex tool such as CVX [45], [46].

Since (52) and (56) are convex, the obtained $c_{k,m}^{(t+1)}$, $p_{k,m}^{(t+1)}$ are the optimal solutions in the $(t+1)$th iteration. Therefore, with the constraint of the maximum transmitted power P, the iteratively updating of $c_{k,m}$ and $p_{k,m}$ will increase or maintain the value of the objective function stable. Thus, we will get a monotonically non-decreasing sequence of the objective value with an upper bound, which can be the global maximum achievable sum rate. As a result, the proposed iterative optimization algorithm for power allocation will converge to a stationary solution to the problem (53).

E. Joint User Scheduling and Power Allocation

Based on the original optimization problem (37), the sum rate are jointly influenced by user scheduling scheme and power allocation coefficients. Therefore, we propose two approaches that jointly consider user scheduling and power allocation to solve (37).

In the first approach, the power allocation coefficient is updated after each swap operation in MTBUSA. The computational complexity of this method is extremely high due to the need of iterative power allocation increases exponentially with the number of swap operations. However, the advantage of this algorithm is that the timely updated power allocation coefficient can achieve a maximum sum rate after each swap operation.

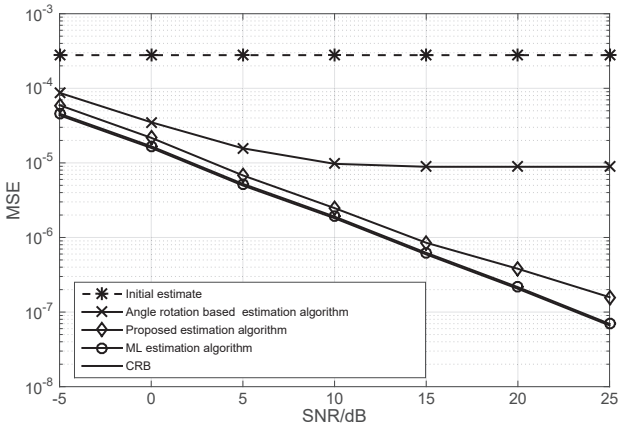

Fig. 2. Comparison of MSE performances of the initial estimation, angle rotation based estimation, the proposed estimation, ML estimation and CRB.

Due to the high complexity of the first approach, we then propose a low-complexity approach in which we first solve the user scheduling problem based on a fixed initial power allocation coefficients. After completing user scheduling, the BS allocates power coefficients using the proposed power allocation algorithm. In this low-complexity way, the power allocation only needs to be executed once. Since the MTBUSA is not sensitive to power, this low complexity algorithm would not cause a essential loss to the overall system performance.

V. Simulation Results

In this section, we show the effectiveness of the proposed algorithms through numerical examples. Specifically, we consider a typical TDD mmWave massive MIMO system, where the ULA at the BS has $M = 128$ antennas of $d = \lambda/2$ and $M_{RF} = 16$ RF chains to communicate with $K = 40$ single antenna users. We use (5) to model the mmWave channels, with the number of paths $L = 3$ (one LOS component and two NLOS components). The DOAs of different users are uniformly distributed in $[-\frac{\pi}{3}, \frac{\pi}{3})$. The mmWave NOMA system is assumed to operate at 30GHz carrier frequency.

In the first example, Fig. 2 plots the MSE performances of DOA estimation as a function of SNR for initial estimation, angle rotation based estimation [9], our proposed estimation method, maximum likelihood (ML) theoretical bound and Cramér-Rao bound (CRB) [47]. Specifically, the searching grid size of angle rotation based estimation algorithm is $\frac{\pi}{5M}$. It can be seen that our proposed DOA estimation method

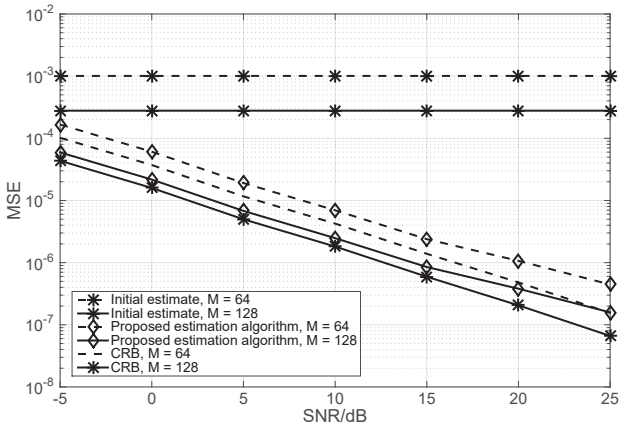

Fig. 3. The MSE performance comparison of the initial estimation, the proposed DOA estimation, and CRB, with $M = 64, M = 128$, respectively.

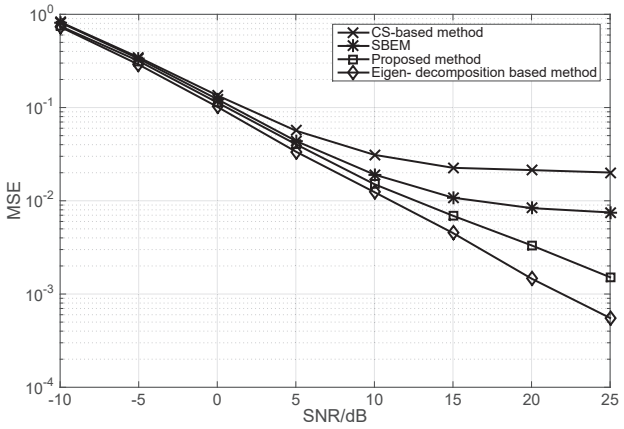

Fig. 4. The channel estimation MSE performances of the proposed method, eigen-decomposition method, SBEM method, and the CS-based method.

outperforms the initial estimation and angle rotation based estimation algorithms, but is slightly worse than ML estimator and CRB. It is also seen from Fig. 2 that the initial estimation remains constant in all SNR region since the Gaussian noise will keep the same level after multiplying the dictionary matrix in different SNRs. The angle rotation based estimation algorithm meets an error floor with the increasing of SNR due to the fixed resolution of searching operation.

Fig. 3 plots the MSE performances of DOA estimation as a function of SNR for various ULA sizes. We assume that the total transmit power for each BS antenna is constrained as a constant. It is clearly seen from Fig. 3 that increasing the number of BS antennas improves the DOA estimation accuracy for all estimation algorithms due to the improved spatial signatures accuracy. It is also seen from Fig. 3 that the proposed DOA estimation method outperforms the initial estimation in any SNR region and is slightly worse than the corresponding CRB.

Fig. 4 compares the MSE performances of the proposed channel estimation method, the CS-based method [8], spatial basis expansion model (SBEM) method [27], the eigen-decomposition based method [30]. It can be seen that the eigen-decomposition based method performs the best MSE performance since it takes full advantage of the channel statistics and utilizes the exact eigen-direction to recover the channel. Nevertheless, it is impractical to obtain accurate channel covariance matrix. The proposed one is slightly worse than the eigen-decomposition based method. It is important to mention that the algorithms except eigen-decomposition based method directly handle the instantaneous channel estimation.

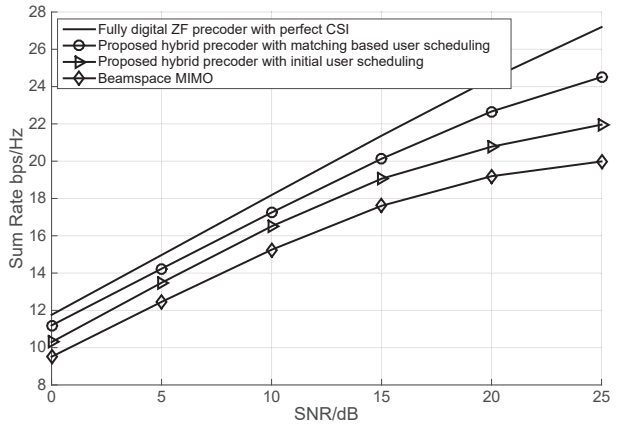

Fig. 5. Comparison of sum rate over different methods as a function of SNR.

Fig. 5 plots the downlink achievable sum rate over different methods, where the proposed hybrid precoding with MTBUSA, the proposed hybrid precoding with initial user scheduling, beamspace MIMO [8] and fully digital ZF precoding with prefect CSI are displayed for comparison. To make the comparison fair, the overall transmit power is set as the same for all methods. It can be seen from Fig. 5 that the proposed hybrid precoding is much better than beamspace MIMO method. By utilizing the MTBUSA instead of the initial user scheduling, we can achieve better performance, especially in the high SNR region. It is worth pointing out that the gap between the sum rate of fully digital ZF with perfect CSI and the proposed MTBUSA grows larger as SNR increases. The reason is that channel estimation replaces noise as the main factor affecting sum rate.

Fig. 6 plots the downlink achievable sum rate over the proposed method and other methods, as a function of the number of BS antennas for SNR$=$ 20dB. It can be seen that with the increasing of the number of BS antennas, the performances of all methods become better, especially when the number of BS antennas is small. The sum rate achieved by the proposed hybrid precoding and MTBUSA greatly outperforms that of the beamspace MIMO method, but is slightly worse than that of the fully digital ZF with perfect CSI. Moreover, our results clearly demonstrate the effectiveness of the proposed MTBUSA.

Fig. 7 illustrates the sum rate versus the SNR over different user scheduling and power allocation methods. To validate the effectiveness of the proposed MTBUSA and power allocation algorithms, we compare the proposed MTBUSA with fixed power scheme, the initial user scheduling with proposed power

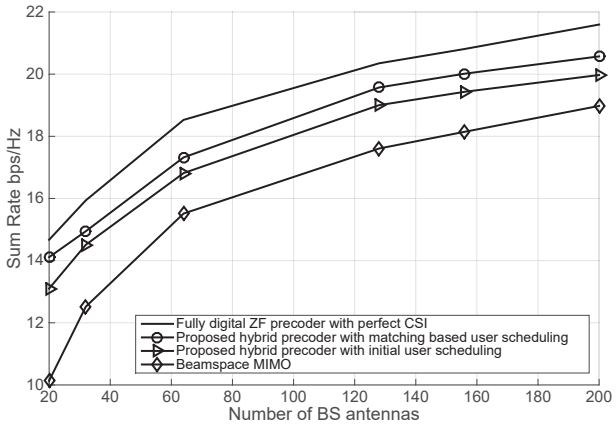

Fig. 6. Comparison of sum rate over different methods as a function of the number of BS antennas.

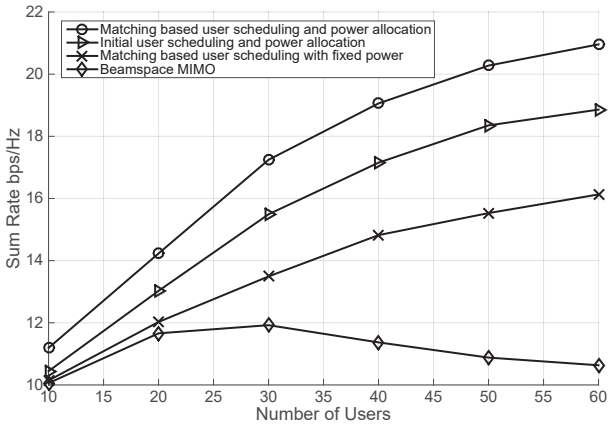

Fig. 8. Comparison of sum rate over different user scheduling and power allocation methods as a function of the number of users.

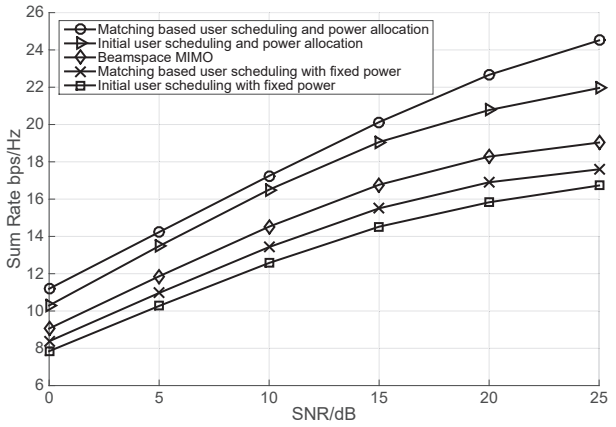

Fig. 7. Comparison of sum rate over different user scheduling and power allocation methods as a function of SNR.

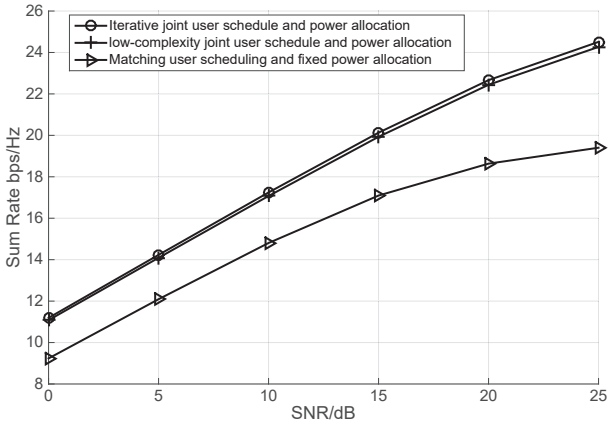

Fig. 9. Comparison of sum rate over different joint user scheduling and power allocation methods versus different SNR.

allocation, and the initial user scheduling with fix power scheme. In fixed power scheme, we assume that each user is assigned the same power. It can be seen from Fig. 7 that the fixed power results in poor sum rate. Specifically, it is worse than beamspace MIMO since the NOMA will fail when users are allocated the same power. It can be observed that the proposed MTBUSA can enhance the sum rate compared to the initial user scheduling algorithm.

Fig. 8 illustrates the sum rate versus the total number of users over different user scheduling and power allocation methods where SNR is set as 10dB. We can see from Fig. 8 that with the increasing of the number of users, the performance of beamspace MIMO becomes better first and then deteriorates. The reason is that increased the total number of users will raise the number of users served by the same beam. As a result, the beamspace MIMO will suffer from severe performance loss. Nevertheless, the proposed mmWave NOMA transmission strategy can still have a better performance while the number of users increases due to the use of NOMA technology. It can also be seen from Fig. 8 that the proposed MTBUSA and power allocation outperform other methods.

Fig. 9 plots the sum rate versus the SNR over different joint user scheduling and power allocation methods. It can be observed that the fixed power allocation algorithm achieves substantially lower performance than two proposed joint user scheduling and power allocation algorithms. Besides, the two proposed algorithms have similar sum rate. Therefore, when considering the computational complexity, low-complexity joint user scheduling and power allocation approach is more applicable in practice.

VI. Conclusion

This paper investigated the problem of channel estimation, hybrid precoding, user scheduling, and power allocation for mmWave NOMA system with hybrid architecture. By utilizing the special structural characteristics of mmWave channel, we proposed a general iterative IDCEA both estimating DOA and channel gain for each channel path. Then we proposed an angle domain hybrid precoding scheme to reduce the inter-beam interferences. With the objective of maximizing the system achievable sum rate, a non-convex problem that

jointly optimizes user scheduling and power allocation was formulated with the interference constraints of different users. To solve this non-convex problem, we decompose it into two sub-problems, i.e., user scheduling and power allocation sub-problems. A novel MTBUSA was designed for solving the user scheduling sub-problem. With the user scheduling results, an iterative optimization algorithm was developed to realize the dynamic power allocation. Simulation results show that the proposed DOA estimation and channel estimation outperform a better MSE performance compared with conventional methods. Moreover, the sum rate of the proposed MTBUSA and power allocation schemes also outperform the conventional mmWave beamspace MIMO system.

REFERENCES

[1] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Mattaiou, G. K. Karagiannidis, E. Bjornson, K. Yang, C. I, and A. Ghosh, "Millimeter Wave Communications for Future Mobile Networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[2] Q. Wu, G. Y. Li, W. Chen, D. Ng, and R. Schober, "An Overview of Sustainable Green 5G Networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.

[3] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 136–143, Dec. 2014.

[4] S. Han, C. L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid precoding analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[5] S Jin, X Liang, KK Wong, X Gao, and Q Zhu,"Ergodic rate analysis for multipair massive MIMO two-way relay networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1480–1491, Mar. 2015.

[6] L Fan, S Jin, CK Wen, and H Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Oct. 2015.

[7] S Jin, X Wang, Z Li, KK Wong, Y Huang, and X Tang, "On massive MIMO zero-forcing transceiver using time-shifted pilots," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 59–74, Jan. 2016.

[8] X. Rao and V. K. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, June 2014.

[9] R. Cao, B. Liu, F. Gao, and X. Zhang, "A low-complex one-snapshot DOA estimation algorithm with massive ULA," *IEEE commun. Lett.* vol. 21, no. 5, pp. 1071–1074, Jan. 2017.

[10] S Jin, MR McKay, C Zhong, and KK Wong, "Ergodic capacity analysis of amplify-and-forward MIMO dual-hop systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2204–2224, Apr. 2010.

[11] Q Zhang, S Jin, KK Wong, H Zhu, M Matthaiou, "Power scaling of uplink massive MIMO systems with arbitrary-rank channel means," *IEEE J. Sel. Topics Signal Process.*, vol.8, no. 5, pp. 966–981, May 2014.

[12] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor,"Cooperative Non-orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer",*IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953 Apr. 2016.

[13] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[14] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks, *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[15] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan and J. A. McCann,"Non-orthogonal Multiple Access in Large-Scale Heterogeneous Networks",*IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.

[16] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan and L. Hanzo"Non-orthogonal Multiple Access for 5G and Beyond",*Proceedings of the IEEE*; vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[17] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A Minorization-Maximization Method for Optimizing Sum Rate in the Downlink of Non-Orthogonal Multiple Access Systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[18] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393–1405, Aug. 2016.

[19] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[20] Z. Yang, Z. Ding, P. Fan, and N. AI-Dhahir, "A General Power Allocation Scheme to Guarantee Quality of Service in Downlink and Uplink NOMA Systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

[21] 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 Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[22] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, Feb. 2017.

[23] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016.

[24] 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.

[25] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum Allocation and Power Control for Non-Orthogonal Multiple Access in HetNets," *IEEE Trans. Wireless Commun.* vol. 16, no. 9, pp. 1502–1517, Sep. 2017.

[26] 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.

[27] H. Xie, F. Gao, S. Zhang, and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3170–3184, Apr. 2017.

[28] J. Dai, A. Liu, and V.K.N Lau, "FDD Massive MIMO Channel Estimation with Arbitrary 2D-Array Geometry," arXiv preprint arXiv:1711.06548, Nov. 2017.

[29] H.Tang, J.Wang, and L.He, "Off-Grid Sparse Bayesian Learning Based Channel Estimation for MmWave Massive MIMO Uplink," *IEEE Wireless Commun. Letter*, Jun. 2018.

[30] A.Adhikary, J.Nam, J.-Y.Ahn, and G.Caire, "Joint spatial division and multiplexingThe large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[31] T. Rappaport, Y. Qiao, J. Tamir, J. Murdock, and E. Ben-Dor, "Cellular broadband millimeter wave propagation and angle of arrival for adaptive beam steering systems," in *Proc. Radio and Wireless Symp. (RWS)*, Santa Clara, CA, USA, Jan. 2012, pp. 151–154.

[32] A. Sayeed and V. Raghavan, "Maximizing MIMO capacity in sparse multipath with reconfigurable antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 1, pp. 156–166, Jun. 2007.

[33] T. Rappaport, F. Gutierrez, E. Ben-Dor, J. Murdock, Y. Qiao, and J. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.

[34] R. Mendez-Rial, C. Rusu, A. Alkhateeb, N. Gonzalez-Prelcicy, and R. W. Heath Jr., "Channel estimation and hybrid combining for mmWave: phase shifters or switches?," in *Proc. Inf. Theory Appl. Workshop*, Feb. 2015, pp. 90–97.

[35] R. Mendez-Rial, C. Rusu, N. Gonzalez-Prelcicy, A. Alkhateeb, and R. W. Heath Jr., "Hybrid MIMO architectures for millimeter wave communications: phase shifters or switches?," *IEEE Access*, vol. 4, pp. 247–267, Jan 2016.

[36] J. Singh and S. Ramakrishna, "On the feasibility of codebook-based beamforming in millimeter wave systems with multiple antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2670–2683, May 2015.

[37] H. Xie, F. Gao, and S. Jin, "An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7313–7321, Nov. 2016.

[38] H. Xie, B. Wang, F. Gao, and S. Jin, "A full-space spectrum-sharing strategy for massive MIMO cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2537–2549, Oct. 2016.

[39] D. Fan, F. Gao, G. Wang, Z. Zhong, and A. Nallanathan, "Angle Domain Signal Processing aided Channel Estimation for Indoor 60GHz TDD/FDD Massive MIMO Systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1948–1961, Sep. 2017.
[40] C. Wang, F. Haider, X. Gao, H. You, and Y. Yang, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–130, Feb. 2014.
[41] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May 2015.
[42] A. E. Roth and M. A. O. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis.* Cambridge, U.K.: Cambridge Univ. Press, 1992.
[43] D. F. Manlove, *Algorithmics of Matching Under Preferences*, vol. 2. Singapore: World Scientific, 2013.
[44] J. R. Magnus and H. Neudecher, *Matrix Differential Calculus with Application in Statistics and Econometrics*. New York, NY, USA: Wiley, 1988.
[45] M. Grant and S. Boyd,"CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.
[46] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[47] P. Stoica and N. Arye, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. 37, no. 5, pp. 720–741, 1989.

Dian Fan received his B.Eng. degree from the School of Science, Beijing Jiaotong University (BJTU), Beijing, China, in 2014. He is currently pursuing the Ph.D. degree at the School of Computer and Information Technology, BJTU. He was a Visiting Ph.D. Student at the Department of Informatics at King's College London from October 2016 to March 2017, and at the School of Electronic Engineering and Computer Science at Queen Mary University of London from October 2017 to September 2018. His research interests include MIMO techniques, massive MIMO systems, millimeter wave systems and array signal processing.

Feifei Gao (M'09, SM'14) received the B.Eng. degree from Xian Jiaotong University, Xi'an, China in 2002, the M.Sc. degree from McMaster University, Hamilton, ON, Canada in 2004, and the Ph.D. degree from National University of Singapore, Singapore in 2007. He was a Research Fellow with the Institute for Infocomm Research (I2R), A*STAR, Singapore in 2008 and an Assistant Professor with the School of Engineering and Science, Jacobs University, Bremen, Germany from 2009 to 2010. In 2011, he joined the Department of Automation, Tsinghua University, Beijing, China, where he is currently an Associate Professor.

Prof. Gao's research areas include communication theory, signal processing for communications, array signal processing, and convex optimizations, with particular interests in MIMO techniques, multi-carrier communications, cooperative communication, and cognitive radio networks. He has authored/coauthored more than 120 refereed IEEE journal papers and more than 120 IEEE conference proceeding papers.

Prof. Gao has served as an Editor of IEEE Transactions on Wireless Communications, IEEE Signal Processing Letters, IEEE Communications Letters, IEEE Wireless Communications Letters, International Journal on Antennas and Propagations, and China Communications. He has also serves as the symposium co-chair for 2018 IEEE Vehicular Technology Conference Spring (VTC), 2015 IEEE Conference on Communications (ICC), 2014 IEEE Global Communications Conference (GLOBECOM), 2014 IEEE Vehicular Technology Conference Fall (VTC), as well as Technical Committee Members for many other IEEE conferences.

Gongpu Wang received the B.Eng. degree in communication engineering from Anhui University, Hefei, Anhui, China, in 2001, and the M.Sc. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2004. From 2004 to 2007, he was an assistant professor in School of Network Education, Beijing University of Posts and Telecommunications. He received Ph.D. degree from University of Alberta, Edmonton, Canada, in 2011. Currently, he is a full professor in School of Computer and Information Technology, Beijing Jiaotong University, China. His research interests include wireless communication, signal processing, artificial intelligence, and Internet of Things.

Zhangdui Zhong received the B.E. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 1983 and 1988, respectively. He is currently a Professor and an advisor of Ph.D. candidates with Beijing Jiaotong University, Beijing, China, where he is also a Chief Scientist of the State Key Laboratory of Rail Traffic Control and Safety. He is also a Director of the Innovative Research Team of the Ministry of Education, Beijing, and a Chief Scientist of the Ministry of Railways, Beijing. His interests include wireless communications for railways, control theory and techniques for railways, and GSM-R systems. His research has been widely used in railway engineering, such as at the Qinghai-Xizang railway, the Datong-Qinhuangdao Heavy Haul railway, and many high-speed railway lines in China. He is an Executive Council member of the Radio Association of China, Beijing, and a Deputy Director of Radio Association, Beijing. He has authored or co-authored seven books, five invention patents, and over 200 scientific research papers in his research area. He received the MaoYiSheng Scientific Award of China, the ZhanTianYou Railway Honorary Award of China, and the Top 10 Science/Technology Achievements Award of Chinese Universities.

Arumugam Nallanathan (S'97-M'00-SM'05-F'17) is Professor of Wireless Communications and Head of the Communication Systems Research (CSR) group in the School of Electronic Engineering and Computer Science at Queen Mary University of London since September 2017. He was with the Department of Informatics at Kings College London from December 2007 to August 2017, where he was Professor of Wireless Communications from April 2013 to August 2017 and a Visiting Professor from September 2017. He was an Assistant Professor in the Department of Electrical and Computer Engineering, National University of Singapore from August 2000 to December 2007. His research interests include 5G Wireless Networks, Internet of Things (IoT) and Molecular Communications. He published nearly 400 technical papers in scientific journals and international conferences. He is a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016 (ICC'2016) , IEEE Global Communications Conference 2017 (GLOBECOM'2017) and IEEE Vehicular Technology Conference 2018 (VTC'2018). He is an IEEE Distinguished Lecturer. He has been selected as a Web of Science Highly Cited Researcher in 2016.

He is an Editor for IEEE Transactions on Communications. He was an Editor for IEEE Transactions on Wireless Communications (2006-2011), IEEE Transactions on Vehicular Technology (2006-2017), IEEE Wireless Communications Letters and IEEE Signal Processing Letters. He served as the Chair for the Signal Processing and Communication Electronics Technical Committee of IEEE Communications Society and Technical Program Chair and member of Technical Program Committees in numerous IEEE conferences. He received the IEEE Communications Society SPCE outstanding service award 2012 and IEEE Communications Society RCC outstanding service award 2014.