PAPER

Power Allocation with QoS and Max-Min Fairness Constraints for Downlink MIMO-NOMA System*

Jia SHAO†a), Cong LI†b), *and* Taotao YAN†c), *Nonmembers*

SUMMARY Non-orthogonal multipe access based multiple-input multiple-output system (MIMO-NOMA) has been widely used in improving user's achievable rate of millimeter wave (mmWave) communication. To meet different requirements of each user in multi-user beams, this paper proposes a power allocation algorithm to satisfy the quality of service (QoS) of head user while maximizing the minimum rate of edge users from the perspective of max-min fairness. Suppose that the user who is closest to the base station (BS) is the head user and the other users are the edge users in each beam in this paper. Then, an optimization problem model of max-min fairness criterion is developed under the constraints of users' minimum rate requirements and the total transmitting power of the BS. The bisection method and Karush-Kuhn-Tucher (KKT) conditions are used to solve this complex non-convex problem, and simulation results show that both the minimum achievable rates of edge users and the average rate of all users are greatly improved significantly compared with the traditional MIMO-NOMA, which only consider max-min fairness of users.

key words: *MIMO-NOMA, max-min fairness, power allocation, quality of service*

1. Introduction

Millmeter wave (mmWave) communication is considered as one of the key technologies of the fifth generation (5G) wireless communication. The abundant spectrum of mmWave band (30–300 GHz) provides great potential to meet the requirements of high data rate and low transmission delay. Furthermore, since the wavelength of mmWave is small, plenty of antennas can be deployed in a relatively limited space. This enables the integration of mmWave communications with the multi-antenna techniques to acquire large beamforming gains, which constitutes the concept of mmWave massive multiple-input multiple-output (MIMO) [1]. However, the large number of radio frequency (RF) chains in MIMO may cause huge energy consumption. Therefore, this paper adopts hybrid precoding to reduce the number of RF chains. Nevertheless, usually the number of supported users can not exceed that of the RF chains simultaneously, although the reduction of the number of RF chains brings benefits, it also introduces a serious problem of limited connections. To address this issue, the number of connected users can be improved by leveraging non-orthogonal multiple access (NOMA) [2], [3].

In mmWave massive non-orthogonal multiple access based multiple-input multiple-output system (MIMO-NOMA), the existing researches based on power allocation mainly consider from two aspects: one is to meet user's quality of service (QoS) requirements, and the other is to meet the fairness criteria. The QoS requirement is equivalent to the minimal achievable rate in this paper. In reference [4], the power allocation of maximizing sum rate was proposed under the constraints of total power and user's minimum rate requirement, so as to ensure the QoS of users. However, only one beam was considered in this scheme. The multiple beams case considered in [5], which proposed a power allocation optimization problem to study the power allocation in mmWave NOMA system and maximize the sum rate under the QoS and total power constraints. [6] maximized the minimum achievable rate in downlink NOMA through the optimal power allocation algorithm from the fairness perspective, but only two users were considered, and the multi-user case was studied in [7].

Each of the researches above considered fairness or QoS for all users in each beam of the system. Moreover, most of the researches were considered to improve the sum rate, which may cause serious problems in some cases, as NOMA tends to group users with very different channel qualities, maximizing sum rate will result in the case that head users occupy most of the system resources, which may cause an unbearable loss of edge users' achievable rates. Both the minimum rate requirement of head uesrs and max-min fairness criterion of edge users were considered in [8], but only single beam was studied.

Suppose that the user who is closest to the base station (BS) is the head user and the other users are the edge users in each beam in this paper. Considering the business requirements of different users in each beam, unlike conventional method like [9], which considered fairness for all users, this paper considers the fairness of edge users under the QoS of head users, so as to improve the performance of the system. Then, the bisection method and Karush-Kuhn-Tucher (KKT) conditions are used to solve the problem.

Notation: Uppercase boldface letter and lowercase boldface letter represent matrix and vector, respectively. $\mathbf{X}^T$ and $\mathbf{X}^H$ denote the transpose and conjugate transpose. $E(\cdot)$ denotes the expectation. $\|.\|$ denotes the l_2-norm. Diag $\left(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K}\right)$ denotes a diagonal matrix of size

Manuscript received April 20, 2023.
Manuscript revised July 10, 2023.
Manuscript publicized September 6, 2023.

†The authors are with the Anhui University of Technology, Maanshan 243000, China.

*This work was supported by The Open Research Fund of National Engineering Research Center for Agro-Ecological Big Date Analysis & Application, Anhui University, China (AE202202).

a) E-mail: 1689553685@qq.com
b) E-mail: licong@ahut.edu.cn (Corresponding author)
c) E-mail: 1091607179@qq.com

DOI: 10.1587/transcom.2023EBP3065

$K \times K$ whose diagonal elements are set as $\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K}$. Let $\mathbf{T}_1 = \left[t_{1,1}, t_{1,2}\right]$ and $\mathbf{T}_2 = \left[t_{2,1}, t_{2,2}\right]$, then diag$(\mathbf{T}_1, \mathbf{T}_2)$ $= \begin{bmatrix} t_{1,1} & t_{1,2} & & \\ & & t_{2,1} & t_{2,2} \end{bmatrix}$. $\mathcal{CN}(0,1)$ denotes the circular symmetric complex Gaussian distribution with mean 0 and variance 1. $U(a, b)$ represents a uniform distribution between a and b. And $|X|$ denotes the number of elements in set X.

2. System Model

In this paper, a single cell downlink mmWave massive MIMO-NOMA system is studied and the system model is shown in Fig. 1. The BS which is equipped with N antennas and N_{RF} RF chains adopts a hybrid precoding structure to serve K single-antenna users, where $K > N_{RF}$. N_S data streams in baseband are precoded by a digital precoding matrix $\mathbf{D}$, then through the corresponding RF chain, the signal is delivered to N phase shifters to perform analog precoding.

In order to obtain higher multiplexing gain, this paper assumes that the number of data streams is the same as the number of RF chains, i.e. $N_S = N_{RF}$. Referring to [10], K users need to be divided into N_{RF} beams firstly, and each beam corresponds to an independent data stream. The users whose channels are highly correlated should be assigned to the same beam to make full use of the multiplexing gain, while the users whose channels are uncorrelated should be assigned to different beams to decrease the interference. Users who are in the same beam can implement successive interference cancellation (SIC) while the signals from different beams are considered interferences. The set of users in the n-th beam is denoted by S_n, $S_i \cap S_j = \emptyset (i \neq j)$, $\sum_{n=1}^{N_{RF}} |S_n| = K$. Since N_{RF} RF chains can support at most N_{RF} data streams, therefore, there should be at least one user in each beam to avoid the waste of RF chain resources, i.e. $|S_n| \geq 1$. Thus, the signal received by the m-th user in the n-th beam is:

$$y_{n,m} = \mathbf{h}_{n,m}^H \mathbf{ADPs} + v_{n,m} \tag{1}$$

where $\mathbf{h}_{n,m}$ represents the $N \times 1$ channel response vector between the BS and the m-th user in the n-th beam. $v_{n,i} \sim \mathcal{CN}(0, 1)$ denotes the Gaussian white noise. $\mathbf{s}$ is the $K \times 1$ vector of transmission signals, $\mathbf{s} = \left[s_{1,1}, \ldots, s_{1,|s_1|}, \ldots, s_{n,1}, \ldots, s_{n,|s_n|}\right]^T$, $E\left(\mathbf{ss}^H\right) = \mathbf{I}_K$. $\mathbf{P}$ is the $N_{RF} \times K$ power allocation matrix, $\mathbf{P} = \text{diag}(\sqrt{\mathbf{p}_1}, \sqrt{\mathbf{p}_2}, \ldots, \sqrt{\mathbf{p}_{N_{RF}}})$, where $\sqrt{\mathbf{p}_i} = \left[\sqrt{p_{i,1}}, \sqrt{p_{i,2}}, \ldots, \sqrt{p_{i,|S_i|}}\right]$, $1 \leq i \leq N_{RF}$. $\mathbf{D}^{N_{RF} \times N_S}$ is the digital precoding matrix. $\mathbf{A}^{N \times N_{RF}}$ is the analog precoding matrix with constant modulus (CM) constraint:

$$\left|[\mathbf{A}]_{i,j}\right| = \frac{1}{\sqrt{N}}, 1 \leq i \leq N, 1 \leq j \leq N_{RF} \tag{2}$$

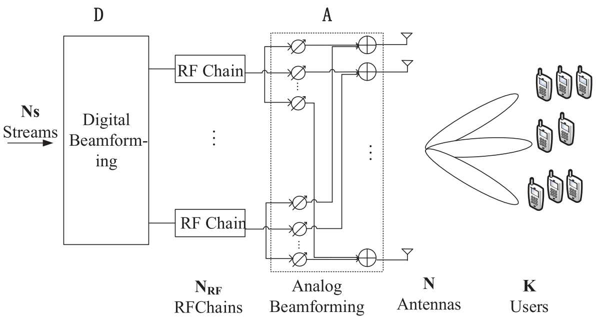

Fig. 1 System model of mmWave MIMO-NOMA.

Suppose that the BS uses a single antenna configuration with each user, and the channel state information (CSI) is known. In this paper, the widely-used Saleh-Valenzuela channel model for mmWave channel is considered [11], $\mathbf{h}_{n,m}$ can be expressed as

$$\mathbf{h}_{n,m} = \beta_{n,m}^{(0)} \cdot \boldsymbol{\alpha}\left(\theta_{n,m}^{(0)}\right) + \sum_{l=1}^{L} \beta_{n,m}^{(l)} \cdot \boldsymbol{\alpha}\left(\theta_{n,m}^{(l)}\right) \tag{3}$$

where $\beta_{n,m}^{(0)}$ represents the complex gain, $\theta_{n,m}^{(l)}$ is the spatial direction of the channel. $\boldsymbol{\alpha}\left(\theta_{n,m}^{(0)}\right)$ is the array steering vector for the line-of-sight (LoS) path, and $\boldsymbol{\alpha}(\cdot)$ is defined as $\boldsymbol{\alpha}(\theta) = \left[e^{j2\pi 0(d/\lambda)\theta}, e^{j2\pi(d/\lambda)\theta}, \cdots, e^{j2\pi(N-1)(d/\lambda)\theta}\right]^T$, which depends on the array geometry. d is the antenna spacing, and λ represents the signal wavelength. $\beta_{n,m}^{(l)} \boldsymbol{\alpha}(\theta_{n,m}^{(l)})$ represents the l-th non-line-of-sight (NLoS) part of user m in the n-beam, $1 \leq l \leq L$, L represents the total number of NLoS paths.

The method described in [12] is used to design analog precoding, and the corresponding equivalent channel is:

$$\hat{\mathbf{H}} = [\hat{\mathbf{h}}_{1,1}, \hat{\mathbf{h}}_{2,1}, \ldots, \hat{\mathbf{h}}_{N_{RF},1}] \tag{4}$$

where $\hat{\mathbf{h}}_{i,1}^H = \mathbf{h}_{i,1}^H \mathbf{A}, \leq i \leq N_{RF}$. Then, the digital precoding matrix can be generated by:

$$\hat{\mathbf{D}} = [\hat{\mathbf{d}}_1, \hat{\mathbf{d}}_2, \ldots, \hat{\mathbf{d}}_{N_{RF}}] = \hat{\mathbf{H}}\left(\hat{\mathbf{H}}^H \hat{\mathbf{H}}\right)^{-1} \tag{5}$$

After normalizing, the digital precoding vector for the n-th beam can be written as:

$$\mathbf{d}_n = \frac{\hat{\mathbf{d}}_n}{\left\|\mathbf{A}\hat{\mathbf{d}}_n\right\|} \tag{6}$$

$\mathbf{d}_n$ is the $N_{RF} \times 1$ digital precoding vector, and we have $\|\mathbf{Ad}_n\| = 1$, for $n = 1, 2, \ldots, N_{RF}$.

This paper assumes the perfect SIC case, users of the same beam are sorted in descending order of channel quality. Take the n-th beam, for example:

$$\left\|\hat{\mathbf{h}}_{n,1}^H \mathbf{d}_n\right\|^2 \geq \left\|\hat{\mathbf{h}}_{n,2}^H \mathbf{d}_n\right\|^2 \geq \cdots \geq \left\|\hat{\mathbf{h}}_{n,|s_n|}^H \mathbf{d}_n\right\|^2 \tag{7}$$

$\hat{\mathbf{h}}_{n,i}$ represents the equivalent channel vector of the i-th user in the n-th beam. The optimal order of SIC for decoding in the downlink is in the increasing order of the equivalent channel gain [13]. More specifically, for a 2-user NOMA case with $\left\|\hat{\mathbf{h}}_{n,m}^H \mathbf{d}_n\right\|^2 \geq \left\|\hat{\mathbf{h}}_{n,j}^H \mathbf{d}_n\right\|^2$, the j-th user does not perform interference cancellation since it comes first in the decoding order; whereas, user m first decodes the signal of user j and subtracts its component from received signal before decoding its own signal. It is worth pointing out that may be the SIC decoding order can be further improved by carefully optimizing the beam selection and precoding matrix to solve some problems in power allocation, but this is not the focus of this paper. Therefore, the signal received by the user m in the n-th beam $\hat{y}_{n,m}$ can be expressed as:

$$\begin{aligned}\hat{y}_{n,m} = &\underbrace{\hat{\mathbf{h}}_{n,m}^{H}\sqrt{p_{n,m}}\mathbf{d}_n s_{n,m}}_{\text{desired signal}} + \underbrace{\hat{\mathbf{h}}_{n,m}^{H}\sum_{i=1}^{m-1}\sqrt{p_{n,i}}\mathbf{d}_n s_{n,i}}_{\text{intra-beam interferences}} \\ &+ \underbrace{\hat{\mathbf{h}}_{n,m}^{H}\sum_{j\neq n}\sum_{k=1}^{|S_j|}\sqrt{p_{j,k}}\mathbf{d}_j s_{j,k}}_{\text{inter-beam interferences}} + \underbrace{v_{n,m}}_{\text{noise}}\end{aligned} \tag{8}$$

where $p_{n,k}$ is the corresponding power allocation parameter. $s_{n,k}$ denotes the transmitted signal. According to (8), the signal-tointerference-plus-noise-ratio (SINR) of the user m in the n-th beam can be expressed as:

$$\gamma_{n,m} = \frac{\left|\hat{\mathbf{h}}_{n,m}^{H}\mathbf{d}_n\right|^2 p_{n,m}}{\left|\hat{\mathbf{h}}_{n,m}^{H}\mathbf{d}_n\right|^2 \sum\limits_{i=1}^{m-1} p_{n,i} + \sum\limits_{j\neq n}\sum\limits_{k=1}^{|S_j|}\left|\hat{\mathbf{h}}_{n,m}^{H}\mathbf{d}_j\right|^2 p_{j,k} + \sigma^2} \tag{9}$$

Therefore, the achievable rate of the user m in the n-th beam can be formulated as:

$$R_{n,m} = \log_2\left(1 + \gamma_{n,m}\right) \tag{10}$$

The users in each beam are numbered according to their distances from the BS. Assume that the nearest user to the BS is "1". Next, they are divided into two categories. Take the j-th beam, for example: $m_1 \in k_1 = \left\{2, 3, \cdots, \left|S_j\right|\right\}$, i.e. the edge users mentioned above. $m_2 \in k_2 = \{1\}$, i.e. the head user. Then the optimization problem model is established as follows:

$$\begin{aligned}&\max_{p_{n,m}}\ \min_{m_1\in k_1}\ R_{n,m_1} \\ &s.t.\ C_1 : \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} p_{n,m} \leq P, \forall m \in k_1 \cup k_2 \\ &\qquad C_2 : R_{n,m_2} \geq \tilde{R}_{n,m_2}, \forall m_2 \in k_2 \\ &\qquad C_3 : p_{n,m} \geq 0\end{aligned} \tag{11}$$

where $\tilde{R}_{n,m_2}$ represents the minimum achievable rate satisfied by the head user in the n-th beam.

3. Power Allocation Algorithm

As for the power allocation problem, this paper mainly applies KKT conditions to the case of multiple RF chains. Different from [9] in dealing with the problem, the results are obtained through rigorous mathematical derivation in combination with bisection method in this paper. The detailed processes are as follows.

To solve complex objective functions, let $\min\limits_{m_1\in k_1} R_{n,m_1} = \gamma$, then (11) is transformed into:

$$\begin{aligned}&\max_{p_{n,m}} \gamma \\ &s.t.\ C_1 : \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} p_{n,m} \leq P, \forall m \in k_1 \cup k_2 \\ &\qquad C_2 : R_{n,m_1} \geq \gamma, \forall m_1 \in k_1 \\ &\qquad C_3 : R_{n,m_2} \geq \tilde{R}_{n,m_2}, \forall m_2 \in k_2 \\ &\qquad C_4 : p_{n,m} \geq 0\end{aligned} \tag{12}$$

Suppose that γ is an arbitrary constant value, (12) can be further transformed into the following optimization problem with the condition C_1 of (12) as the target function to find the minimum total power of the system, i.e.:

$$\begin{aligned}&\min_{p_{n,m}} \sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|} p_{n,m} \\ &s.t.\ C_2 : R_{n,m_2} \geq \tilde{R}_{n,m_2}, \forall m_2 \in k_2 \\ &\qquad C_3 : p_{n,m} \geq 0 \\ &\qquad C_4 : R_{n,m_1} \geq \gamma, \forall m_1 \in k_1\end{aligned} \tag{13}$$

And γ will be solved later. Problem (13) is a linear convex optimization problem to find the minimum value, which can be solved by using KKT conditions. According to (10), $R_{n,m_1} \geq \gamma$ can be equivalent to:

$$\begin{aligned}\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 p_{n,m_1} \geq \left(2^{\gamma} - 1\right)&\left(\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 \sum_{k=2}^{m_1-1} p_{n,k}\right. \\ &\left.+ \sum_{j\neq n}\sum_{i=1}^{|S_j|}\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_j\right|^2 p_{j,i} + \sigma^2\right)\end{aligned} \tag{14}$$

$R_{n.m_2} \geq \tilde{R}_{n,m_2}$ can be equivalent to:

$$\begin{aligned}&\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_n\right|^2 \cdot p_{n,m_2} \geq \\ &\left(2^{\tilde{R}_{n,m_2}} - 1\right)\left(\sum_{j\neq n}\sum_{k=1}^{|S_j|}\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_j\right|^2 p_{j,k} + \sigma^2\right)\end{aligned} \tag{15}$$

Let:

$$\begin{aligned}G_{n,m_1} = \left(2^{\gamma} - 1\right)&\left(\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 \sum_{k=2}^{m_1-1} p_{n,k}\right. \\ &\left.+ \sum_{j\neq n}\sum_{i=1}^{|S_j|}\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_j\right|^2 p_{j,i} + \sigma^2\right)\end{aligned}$$

$$G_{n,m_2} = \left(2^{\tilde{R}_{n,m_2}} - 1\right)\left(\sum_{j\neq n}\sum_{k=1}^{|S_j|}\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_j\right|^2 p_{j,k} + \sigma^2\right)$$

To facilitate the use of KKT conditions to solve problem (13), it is converted into the following Lagrangian function equivalently.

$$L\left(p,\lambda,\mu,\varphi\right)=\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}p_{n,m}+\sum_{m_2\in k_2}\mu_{n,m_2}\left(G_{n,m_2}-\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_n\right|^2\cdot p_{n,m_2}\right)+\sum_{\mathrm{m}_1\in k_1}\lambda_{n,m_1}\left(G_{n,m_1}-\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 p_{n,m_1}\right)-\varphi_{n,m}p_{n,m} \tag{16}$$

where λ_{j,m_1}, $\mu_{\mathrm{j,m_2}}$, $\varphi_{n,m}$ are the Lagrange multiplier. (13) shall meet the following KKT conditions:

1) $$\frac{\partial L}{\partial p_{n,m_1}}=1-\varphi_{n,m_1}-\lambda_{n,m_1}\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2+\sum_{i=m_1+1}^{|S_n|}\lambda_{n,i}\left(2^{\gamma}-1\right)\left|\hat{\mathbf{h}}_{n,i}^{H}\cdot\mathbf{d}_n\right|^2+\sum_{j\neq n}\sum_{i=1}^{|S_j|}\lambda_{j,i}\left(2^{\gamma}-1\right)\left|\hat{\mathbf{h}}_{j,i}^{H}\cdot\mathbf{d}_n\right|^2=0$$

2) $$\frac{\partial L}{\partial p_{n,m_2}}=1-\varphi_{n,m_2}-\mu_{n,m_2}\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_n\right|^2+\sum_{j\neq n}\sum_{i=1}^{|S_j|}\mu_{j,i}\left(2^{\tilde{R}_{n,m_2}}-1\right)\left|\hat{\mathbf{h}}_{j,i}^{H}\cdot\mathbf{d}_n\right|^2=0$$

3) $\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 p_{n,m_1}\geq G_{n,m_1}$

4) $\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_n\right|^2\cdot p_{n,m_2}\geq G_{n,m_2}$

5) $\lambda_{n,m_1}\left(G_{n,m_1}-\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 p_{n,m_1}\right)=0$

6) $\mu_{n,m_2}\left(G_{n,m_2}-\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_n\right|^2\cdot p_{n,m_2}\right)=0$

7) $\varphi_{n,m}p_{n,m}=0\left(\varphi_{n,m_1}p_{n,m_1}=0;\varphi_{n,m_2}p_{n,m_2}\right)$

8) $\lambda_{n,m}\geq 0,\mu_{n,m}\geq 0,\varphi_{n,m}\geq 0,p_{n,m}\geq 0$

To $\forall\ \gamma>0$ and $\sigma^2>0$, that $G_{n,m_1}>0$ strictly, so $p_{n,m_1}>0$ in 3), then it can be obtained $\varphi_{n,m_1}=0$ from 7), and it can be futher obtained from 1):

$$\lambda_{n,m_1}\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2=1+\sum_{i=m_1+1}^{|S_n|}\lambda_{n,i}\left(2^{\gamma}-1\right)\left|\hat{\mathbf{h}}_{n,i}^{H}\cdot\mathbf{d}_n\right|^2+\sum_{j\neq n}\sum_{i=1}^{|S_j|}\lambda_{j,i}\left(2^{\gamma}-1\right)\left|\hat{\mathbf{h}}_{j,i}^{H}\cdot\mathbf{d}_n\right|^2>0$$

so, $\lambda_{n,m_1}>0$, then combined with 5), it can be obtained:

$$G_{n,m_1}-\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 p_{n,m_1}=0$$

similarly:

$$G_{n,m_2}-\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_n\right|^2\cdot p_{n,m_2}=0$$

Thus, the optimal solution of problem (13) can be obtained as follows:

$$p_{n,m_1}=G_{n,m_1}/\left|\hat{\mathbf{h}}_{n,m_1}^{H}\cdot\mathbf{d}_n\right|^2 \tag{17}$$

$$p_{n,m_2}=G_{n,m_2}/\left|\hat{\mathbf{h}}_{n,m_2}^{H}\cdot\mathbf{d}_n\right|^2 \tag{18}$$

Through the above solution, the optimal solution of the linear programming problem (13) is obtained, but it is obtained under the assumption that γ is known. However, problem (12) can be solved by choosing an appropriate γ. Because the minimum total power of the system is monotonically increasing with γ, the bisection method can be used to obtained an appropriate γ, then the original non-convex problem (11) can be solved combined the solution result of (13). The steps of obtaining γ by the bisection method are as Algorithm 1.

Thus, this paper obtains the power allocation of users in each beam in the downlink mmWave MIMO-NOMA system. As a result, the complexity of the bisection and KKT condition are $O\left(\log\left({}^{1}\!/_{\varepsilon}\right)\right)$ and $O\left(K\right)$ respectively. The optimal solution of problem (11) is obtained by bisection and KKT condition, so the complexity is $O\left(\log\left({}^{1}\!/_{\varepsilon}\right)K\right)$.

Algorithm 1 Bisection Procedure

```
Input : Lower bound $\gamma_{min}$, upper bound $\gamma_{max}$, equivalent channels $\hat{\mathbf{h}}_{n,m}$, noise power $\sigma^2$, total power $P$, desirable accuracy $\varepsilon$.
Output : max-min $\gamma^*$, corresponding power allocation parameters $p^*_{n,m}$.
Determine the initial interval, $\gamma_{min}=0, \gamma_{max}=\log_2\left(1+\frac{P\cdot\mathbf{h}_{max}}{\sigma^2}\right)$;
While $\gamma_{max}-\gamma_{min}>\varepsilon$, do
    Set $\gamma_0=(\gamma_{min}+\gamma_{max})/2$, substituting it into (17),
    to obtain the optimal solution $p_{n,m}$.
    if $\sum_{n=1}^{N_{RF}}\sum_{m=1}^{|S_n|}p_{n,m}\leq P$ and $p_{n,m}\geq 0$, then
        Set $\gamma_{min}=\gamma_0, \gamma^*=\gamma_0, p^*_{n,m}=p_{n,m}$.
    else
        Set $\gamma_{max}=\gamma_0$;
    end if
end while
```

4. Computer Simulations

In this section, the simulation results are provided to verify the performance of our proposed scheme. As for the simulation settings, all the schemes share the same simulation conditions. The channel parameters for an arbitrary user m in the n-th beam and the main parameters of the BS are set as shown in the Table 1.

The schemes in this section mainly include the following two: 1. Power allocation algorithm that satisfies QoS of head users and max-min fairness for edge users in each beam, i.e. the scheme this paper proposed. 2. A zero-forcing beamspace MIMO-NOMA max-min fairness power allocation algorithm [9]. Different from Scheme 2, where all users have the same rate, the head users have different rates with edge users in Scheme 1. To facilitate the comparison of the two schemes, the average rate of all users in Scheme 1 is given in the simulation figures, i.e.

Table 1 Main parameters.

Parameters	**Value**
Number of antennas at BS	$N = 64$
Number of antennas at each user	1
Number of RF chains	$N_{RF} = 4$
Number of date streams	$N_s = 4$
Desirable accuracy	$\varepsilon = 10^{-4}$
Number of LoS paths	1
Number of NLoS paths	5
Complex gain of LoS	$\beta^{(0)}_{n,m} \sim \mathcal{CN}(0, 0.8)$
Complex gain of NLoS	$\beta^{(l)}_{n,m} \sim \mathcal{CN}(0, 0.25)$
Spatial direction	$\theta_{n,m} \sim U(-0.5, 0.5)$

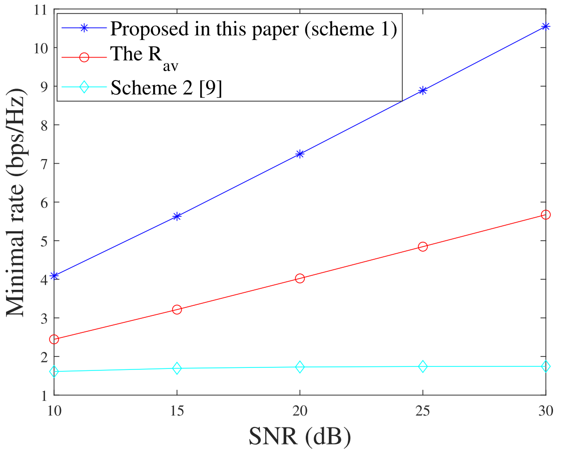

Fig. 2 Min. rate vs. SNR (N_{RF} = 4, K = 8, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.8 bps/Hz).

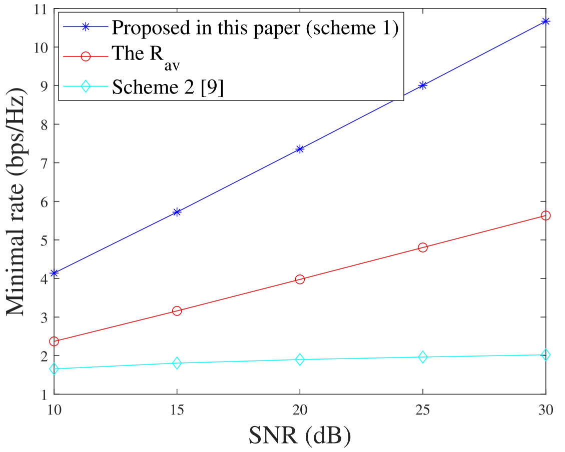

Fig. 3 Min. rate vs. SNR (N_{RF} = 4, K = 8, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.6 bps/Hz).

$R_{av} = \left[\tilde{R}_{\mathrm{n,m_2}} \cdot N_{RF} + R_{n,m_1} \cdot (K - N_{RF})\right]/K$, where $\tilde{R}_{\mathrm{n,m_2}}$ is the rate of head user and R_{n,m_1} is the rate of edge user. As mentioned before, each beam has at least one user, so the number of head users is the same as the N_{RF}.

Figures 2 to 4 show the minimum achievable rates of the two schemes with respect to the signal-to-noise-ratios (SNR), which is defined as $\log_{10}\left(P/\sigma^2\right)$. As can be seen from the three figures, when $\tilde{R}_{\mathrm{n,m_2}}$ changes from 0.8 bps/Hz to 0.4 bps/Hz under the QoS requirement, the R_{av} improves slightly and is higher than the rate in Scheme 2, which also shows the feasibility of the proposed scheme in this paper. Specifically, to show the superior performance of the proposed scheme, take Fig. 3 for example, when SNR is 10 dB, the average rate is improved by 43% compared with Scheme 2, and the rate performance is improved more significantly with the increase of SNR, because the fairness of all users is studied in Scheme 2, while Scheme 1 divides users into two categories for research. Compared with the rate of Scheme 2, the interference in Scheme 1 is relatively easy to deal with, so, the difference of interference processing level becomes more obvious with the increase of SNR. In order to compare to the situation when more users are served, the result when K = 12 is shown in Fig. 5, we can see that the performance gains for users are more significant compared to Fig. 3. This also proves the superiority of the scheme this paper proposed when the number of RF chains is fixed and the number of users is within a certain range.

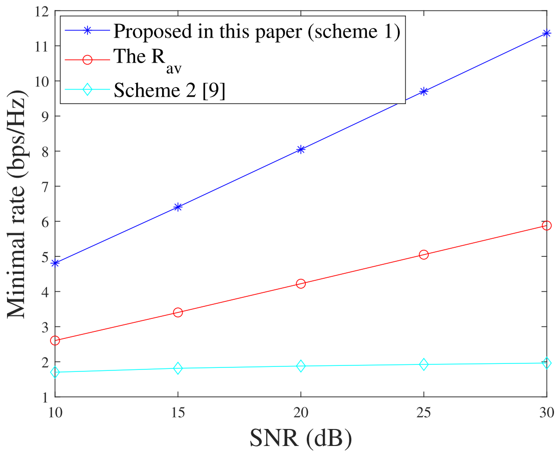

Fig. 4 Min. rate vs. SNR (N_{RF} = 4, K = 8, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.4 bps/Hz).

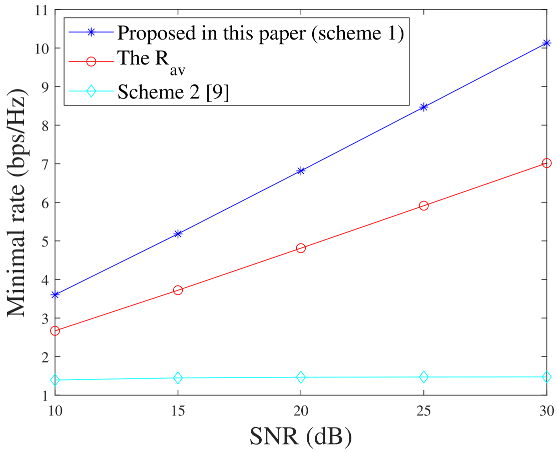

Fig. 5 Min. rate vs. SNR (N_{RF} = 4, K = 12, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.6 bps/Hz).

Figures 6 to 8 show the corresponding minimum achievable rate of each scheme with different number of users. It can be seen from the figures that the minimum achievable rate of each scheme decreases with the increase of the number of users because of the limitation of the total transmitted power

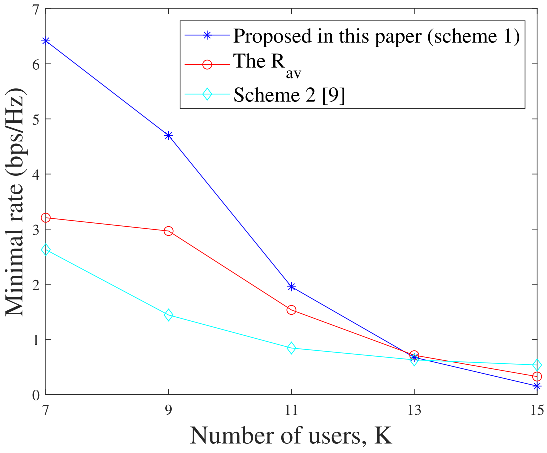

Fig. 6 Min. rate vs. K (N_{RF} = 4, SNR = 15 dB, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.8 bps/Hz).

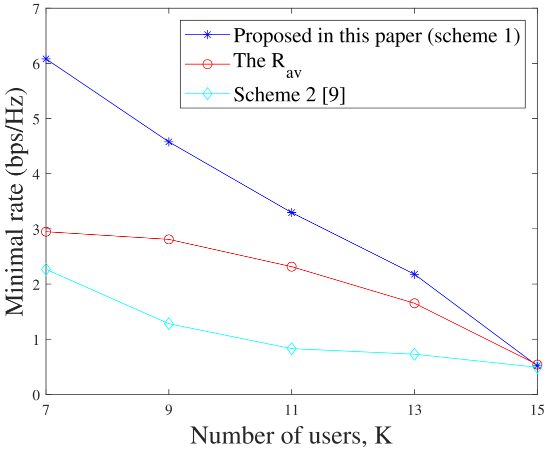

Fig. 7 Min. rate vs. K (N_{RF} = 4, SNR = 15 dB, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.6 bps/Hz).

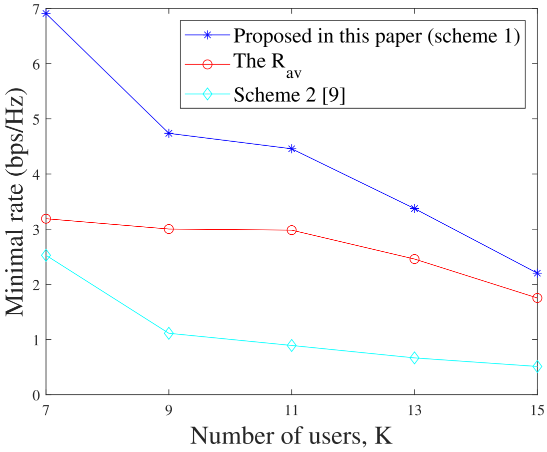

Fig. 8 Min. rate vs. K (N_{RF} = 4, SNR = 15 dB, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.4 bps/Hz).

of BS. When $\tilde{R}_{\mathrm{n,m_2}}$ changes from 0.8 bps/Hz to 0.4 bps/Hz under the QoS requirement, the R_{av} improves more significant compared with Scheme 2. However, when the total power is fixed and the number of users exceeds 13, the performance begins to decline, even worse than that of Scheme 2 in Fig. 6 due to the increase of edge users. Increasing the number of head user settings, like [8], may be able to improve this situation. Also, refer to [9], we show the result when N_{RF} = 8 in Fig. 9. It can be seen that the performance has improved significantly compared to Fig. 6. However, the increase in cost and complexity may make this approach less feasible. Much work remains to be done to solve this problem, which is worth exploring.

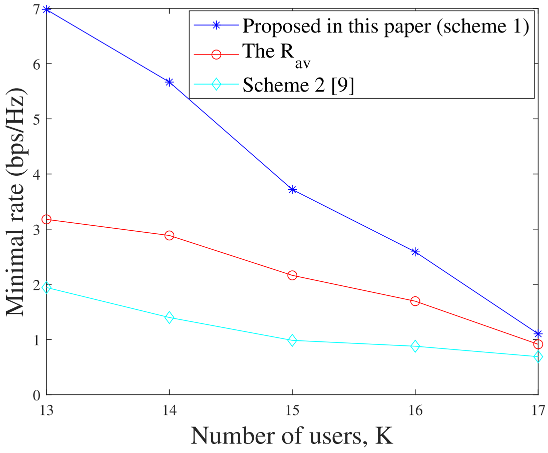

Fig. 9 Min. rate vs. K (N_{RF} = 8, SNR = 15 dB, $\tilde{R}_{\mathrm{n,m_2}}$ = 0.8 bps/Hz).

5. Conclusion

In this paper, the power allocation of the downlink mmWave MIMO-NOMA system is studied. In order to ensure the service requirements of different users in each multi-user beam, the minimum achievable rate of edge user is maximized from the perspective of max-min fairness while satisfying the QoS of head users. Based on the above considerations, the optimal power allocation algorithm is proposed, and the corresponding optimization problem model is established, then the problme is solved by bisection and KKT conditions. Finally, the proposed algorithm is verified by simulations, and the results show that both the minimum achievable rates of edge users and the average rates are greatly improved compared with the traditional MIMO-NOMA max-min fairness power allocation algorithm.

References

[1] S. Mumtaz, J. Rodriguez, and L. Dai, mmWave Massive MIMO: A Paradigm for 5G, Academic Press, New York, 2016.

[2] L. Qian, Y. Wu, N. Yu, F. Jiang, H. Zhou, and T.Q.S. Quek, “Learning driven NOMA assisted vehicular edge computing via underlay spectrum sharing,” IEEE Trans. Veh. Technol., vol.70, no.1, pp.977–992, Jan. 2021.

[3] Y. Wu, L.P. Qian, H. Mao, X. Yang, H. Zhou, and X. Shen, “Optimal power allocation and scheduling for non-orthogonal multiple access relay-assisted networks,” IEEE Trans. Mobile Comput., vol.17, no.11, pp.2591–2606, 1 Nov. 2018.

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

[5] W. Hao, F. Zhou, Z. Chu, P. Xiao, R. Tafazolli, and N. Al-Dhahir,

"Beam alignment for MIMO-NOMA millimeter wave communication systems," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), pp.1–6, 2019.

[6] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," IEEE Commun. Lett., vol.20, no.10, pp.2055–2058, Oct. 2016.

[7] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based power allocation and user set selection for downlink NOMA systems," 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, pp.1–6, 2016.

[8] X. Li, W. Ma, L. Luo, and F. Zhao, "Power allocation for NOMA system in downlink," Systems Engineering and Electronics, vol.40, no.7, pp.1595–1599, 2018.

[9] R. Jiao, L. Dai, W. Wang, F. Lyu, N. Cheng, and X. Shen, "Max-min fairness for beamspace MIMO-NOMA: From single-beam to multi-beam," IEEE Trans. Wireless Commun., vol.21, no.2, pp.739–752, Feb. 2022.

[10] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D.O. Wu, and X. -G. 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.

[11] R. Jiao and L. Dai, "On the max-min fairness of beamspace MIMO-NOMA," IEEE Trans. Signal Process., vol.68, pp.4919–4932, 2020.

[12] 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.

[13] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," IEICE Trans. Commun., vol.E98-B, no.3, pp.403–414, March 2015.

Jia Shao received his Bachelor's degree in civil engineering from Northeast Electric Power University, Jilin, China, in 2021. He is currently a graduate student in communication engineering at Anhui University of Technology, Anhui, China.

Cong Li received the B.S. degree from Northeast Forestry University, Harbin, China, in 2005 and the M.S. and Ph.D. degrees from Nagoya Institute of Technology, Nagoya, Japan, in 2010 and 2013 respectively, all in information and communications engineering. From April 2013 to October 2018, he worked as a research and development engineer at Fujitsu Group, Kawasaki, specializing in the research and development of communication systems. He is currently an associate professor at Anhui University of Technology, Anhui, China. His research interests include reconfigurable intelligent surfaces (RIS), massive MIMO communication, information theory, industrial Internet of Things, and machine learning for wireless communications.

Taotao Yan received his Bachelor's degree in communication engineering from Anhui University of Technology, Anhui, China, in 2020. He is currently a graduate student at this university.