Max-Min Fairness for Beamspace MIMO-NOMA: From Single-Beam to Multi-Beam

Ruicheng Jiao, *Student Member, IEEE*, Linglong Dai, *Fellow, IEEE*, Wei Wang, *Member, IEEE*, Feng Lyu, *Member, IEEE*, Nan Cheng, *Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*— With the help of non-orthogonal multiple access (NOMA), the number of connections of the beamspace multiple-input multiple-output (MIMO) systems can be improved with enhanced sum-rate performance, which constitutes beamspace MIMO-NOMA. Thus, most relevant papers focus on improving the system sum rate, which may inflict unbearable rate loss to weak users. To ensure the achievable rates of weak users, we maximize and analyze the minimal rate of the system in the single-beam case as well as the multi-beam case, where two completely different phenomena are revealed. Particularly, in the single-beam case, the maximized minimal rate of the beamspace MIMO-NOMA always grows rapidly with the signal-to-noise-ratio (SNR), and is larger than that of the beamspace MIMO using orthogonal multiple access (beamspace MIMO-OMA). However, in the multi-beam case, the maximized minimal rate of the beamspace MIMO-NOMA grows slower and slower in the high-SNR region, where it is smaller than that of the beamspace MIMO-OMA. To explain this difference, it is disclosed that the intra-beam interference in the single-beam case is of *successive pattern*, which is proved to have no limit on the max-min rate. In contrast, the inter-beam interference in the multi-beam case is of *mutual pattern*, which is proved to restrict the max-min rate to a derived upper bound.**

***Index Terms*— Beamspace MIMO-NOMA, max-min rate, interference, successive pattern, mutual pattern.**

Manuscript received February 2, 2021; revised June 14, 2021; accepted July 7, 2021. Date of publication July 28, 2021; date of current version February 14, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1805005, in part by the National Natural Science Foundation of China under Grant 62031019, and in part by the European Commission through the H2020-Marie Skłodowska-Curie Actions-Innovative Training Networks (MSCA-ITN) META WIRELESS Research Project under Grant 956256. This article was presented in part at 2019 IEEE GLOBECOM, HI, USA, December 2019 [1]. The associate editor coordinating the review of this article and approving it for publication was L. Dasilva. *(Corresponding author: Linglong Dai.)*

Ruicheng Jiao and Linglong Dai are with the Beijing National Research Center for Information Science and Technology (BNRist) and the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: jiaors16@mails.tsinghua.edu.cn; daill@tsinghua.edu.cn).

Wei Wang is with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: wei_wang@nuaa.edu.cn).

Feng Lyu is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: fenglyu@csu.edu.cn).

Nan Cheng is with the State Key Lab. of ISN and the School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: dr.nan.cheng@ieee.org).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3098900.

Digital Object Identifier 10.1109/TWC.2021.3098900

I. Introduction

A NEW paradigm for future wireless communications has been defined by the millimeter-wave (mmWave) communications, which is very likely to achieve the 1000-fold increase in data rate due to the abundant spectrum resources at mmWave band [2]–[4]. 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 gain, which constitutes the concept of mmWave massive multiple-input multiple-output (MIMO) [5]–[7]. However, mmWave massive MIMO cannot directly adopt the widely used full-digital structure in traditional massive MIMO systems, where each antenna element is equipped with a radio-frequency (RF) chain. This is because the large number of RF chains may cause huge energy consumption [8]. To address this issue, several solutions have been proposed to achieve energy efficient and cost-effective mmWave communications by reducing the number of RF chains, such as hybrid precoding [9], [10], beamspace MIMO [11]–[13], etc.

However, although the reduction of the number of RF chains brings benefits mentioned above, it also introduces a serious problem of limited connections. Since one RF chain can simultaneously serve only one user in the traditional mmWave communication systems, the number of connected users will be largely constrained by the limited RF chains [14]. To address this issue, the number of connected users can be improved by leveraging non-orthogonal multiple access (NOMA). Constituting a prominent technology for the next-generation wireless communications [15]–[19], NOMA is known for advantages over traditional orthogonal multiple access (OMA) in terms of the connection number, sum capacity [15], etc. NOMA superimposes the multi-user signals in the power domain using the same resources, and implements successive interference cancellation (SIC) to detect multi-user signal. In this way, it can accommodate more users than RF chains [14], and thus the NOMA-based mmWave communications has acquired extensive attention [20]–[25].

A. Prior Works

For NOMA-based mmWave communications, the single-beam case has been investigated, where only one RF chain is deployed at the base station (BS). Particularly, [20] explored

NOMA in a hybrid precoding-based mmWave communication system, and maximized the sum rate of two single-antenna users via joint optimization of power allocation parameters and beamforming vectors. Furthermore, [21] considered a downlink mmWave communication scenario, where the BS and the users are equipped with multiple antennas and single RF chain. NOMA is integrated in this system to serve multiple users, where power allocation parameters, transmit beamforming, and receive beamforming are jointly optimized for sum-rate maximization. Besides, [22] investigated a NOMA-based mmWave communication system with hybrid precoding in the single-beam case, where the joint optimization of power allocation parameters and beamforming vectors is carried out for minimal rate maximization.

In the multi-beam case with the BS equipped with multiple RF chains, [23] firstly explored the beamspace MIMO-NOMA scheme, where NOMA is utilized in the beamspace MIMO based mmWave communication system to support multiple users in each beam. Moreover, power allocation is optimized to maximize its sum rate by the weighted minimum mean square error (WMMSE) algorithm. Similarly, [24] also aimed at improving the sum rate in a multi-beam mmWave communication system integrated with NOMA. To achieve this goal, antenna allocation, user grouping, and power allocation are jointly optimized. Additionally, [25] maximized the sum rate for a NOMA based mmWave communication system, where user clustering is designed by a Stackelberg game-based approach, and the optimal power allocation parameters are acquired in closed-form expressions.

It is worth pointing out that most of the above-mentioned works mainly consider to improve the sum rate, which is an important performance metric extensively explored in NOMA based mmWave communication systems. However, only focusing on the sum rate may incur substantial rate loss to weak users. This is because the system tends to allocate most of the communication resources to the strong users when the sum rate is maximized. In some extreme cases such as that considered in [26], the weak users even cannot be served. Therefore, in order to guarantee the achievable rates for weak users, the important performance metric of user fairness needs to be considered. Correspondingly, we maximize the minimal rate of the system from the perspective of max-min fairness. Note that some researchers tend to set a target data rate for all users to avoid the extreme case in [26]. Under this circumstance, studying the maximized minimal rate still makes sense, since it provides an upper bound for the target data rate.

B. Our Contributions

To ensure the rate performance for all users, we maximize and analyze the minimal rate of beamspace MIMO-NOMA considering the single-beam case as well as the multi-beam case. The corresponding theoretical analysis is challenging, because the interference patterns of the considered system are complicated, especially for the multi-beam case which suffers from both the intra-beam and inter-beam interferences. As a result, the minimal rate maximization problem is hard to solve, and it is difficult to analyze the effects of both the interferences on the maximized minimal rate (max-min rate). To this end, we utilize a bisection-based method for minimal rate maximization, disclose the different patterns of the intra-beam as well as the inter-beam interferences, and comprehensively analyze their different effects on the max-min rate. To the best of our knowledge, we are the first to carry out the max-min fairness analysis for this system in both of the two cases.[1] Our specific contributions are listed below.

1) Different from [22] and [27] which maximize the minimal rate of NOMA-based single-antenna/single-beam systems, we maximize the minimal rate of beamspace MIMO-NOMA in both the single-beam and multi-beam cases via power allocation. Moreover, we point out the difference between the two cases in terms of the inter-beam interference, and extend the single-beam property that the max-min rate is achieved when rates of all NOMA users are equal to one another to the multi-beam case.
2) We analyze the max-min rate of beamspace MIMO-NOMA by comparing it with that of the beamspace MIMO using orthogonal multiple access (beamspace MIMO-OMA) in both cases, and reveal that the comparison results in the two cases are completely different. To be more specific, in the single-beam case, beamspace MIMO-NOMA is always superior to beamspace MIMO-OMA in terms of the max-min rate, and the max-min rate of both systems grow rapidly with the SNR. However, in the multi-beam case, beamspace MIMO-NOMA is inferior to beamspace MIMO-OMA in the high-SNR region, where the max-min rate of the former system grows slower and slower with the increase of SNR.
3) To account for the difference in the two cases, we point out that the interfering patterns of the intra-beam and inter-beam interferences are completely different, and analyze their different effects on the max-min rate. Specifically, the intra-beam interference in the single-beam case is of successive pattern, and we prove that it will not limit the max-min rate. However, the inter-beam interference in the multi-beam case is of mutual pattern, which is proved to largely constrain the max-min rate, while the upper bound is also derived. Note that the above analysis does not rely on specific system design schemes, and can serve as a general framework for the max-min fairness analysis of beamspace MIMO-NOMA.

C. Organization and Notation

Organization: The remainder of the paper is organized as follows. Section II detailedly introduces the system model of beamspace MIMO-NOMA. Section III maximizes the minimal rate of all NOMA users in the single-beam case, which is shown to be always larger than that of the traditional beamspace MIMO-OMA. Moreover, Section IV maximizes the minimal rate of all NOMA users in the multi-beam case, and reveals that it is smaller than the maximized minimal

[1] Simulation codes are provided to reproduce the results presented in this paper: http://oa.ee.tsinghua.edu.cn/dailinglong/publications/publications.html.

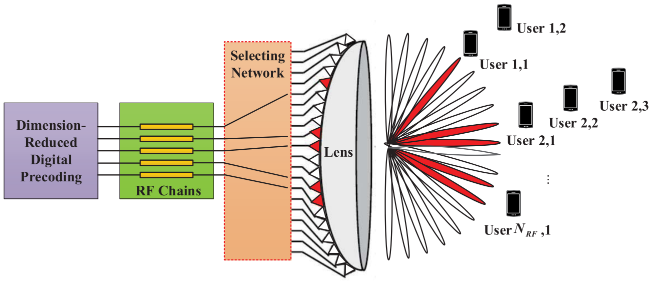

Fig. 1. System model of beamspace MIMO-NOMA.

rate of beamspace MIMO-OMA in the high-SNR region. To explain this difference, Section V points out the different interfering patterns of the intra-beam and inter-beam interferences, and analyze their different effects on the max-min rate. Finally, Section VI draws the conclusion.

Notation: Vectors and matrices are denoted by lower-case and upper-case boldface letters, respectively. Conjugate transpose, Moore-Penrose inversion, and inversion of the matrix are denoted by $(\cdot)^H$, $(\cdot)^\dagger$, and $(\cdot)^{-1}$, respectively. Besides, the ℓ_2-norm is denoted by $||\cdot||$, the expectation is denoted by $\mathbb{E}(\cdot)$, and the cardinality of set $\mathcal{B}$ is denoted by $|\mathcal{B}|$. Finally, the circular symmetric complex Gaussian distribution is denoted by $\mathcal{CN}(a, b)$, with its mean set to a and variance set to b.

II. System Model

The single-cell downlink beamspace MIMO-NOMA system is elaborated in this section. To provide a clearer description, its basic principles are summarized below. Firstly, utilizing the intrinsic sparsity of mmWave channel, the spatial channel can be transformed to the sparse beamspace channel by an N-antenna lens antenna array deployed at the BS. Then, based on the sparse beamspace channels, the BS is able to select part of the beams to serve all users, which largely reduces the number of RF chains N_{RF}. Finally, NOMA is adopted in each beam to serve multiple users, which is shown in Fig. 1.

Particularly, the function of the lens antenna array can be mathematically described as an $N \times N$ discrete Fourier transformation matrix $\mathbf{U}$:

$$\mathbf{U} = [\mathbf{a}(\hat{\psi}_0), \mathbf{a}(\hat{\psi}_1), \cdots, \mathbf{a}(\hat{\psi}_{N-1})]^H, \tag{1}$$

where the antenna response vectors $\mathbf{a}(\hat{\psi}_n)^H = \frac{1}{\sqrt{N}}[e^{-j2\pi\hat{\psi}_n m}]^H_{m\in\mathcal{J}(N)}$ constitute the rows of $\mathbf{U}$ with predefined spatial directions $\hat{\psi}_n = \dfrac{1}{N}(n - \dfrac{N-1}{2})$, and $\mathcal{J}(N) = \{n - (N-1)/2, n = 0, 1, 2, \ldots, N-1\}$. With this discrete Fourier transformation matrix $\mathbf{U}$, the spatial channel matrix $\mathbf{H}$ for all K users can be transformed into the beamspace channel matrix $\tilde{\mathbf{H}}$:

$$\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \cdots, \tilde{\mathbf{h}}_K] = [\mathbf{U}\mathbf{h}_1, \mathbf{U}\mathbf{h}_2, \cdots, \mathbf{U}\mathbf{h}_K], \tag{2}$$

where the N-dimension vectors $\tilde{\mathbf{h}}_k$ and $\mathbf{h}_k$ denote the beamspace channel and the spatial channel for the k-th user, respectively.

Furthermore, in mmWave propagation environments, the number of dominant scatters L is usually much smaller than the number of BS antennas N, which indicates that only a few multi-path components exist [28]. This can be characterized by the Saleh-Valenzuela channel model, where the spatial channel for the k-th user is expressed as

$$\mathbf{h}_k = \sum_{l=0}^{L} \beta_k^{(l)} \mathbf{a}(\psi_k^{(l)}). \tag{3}$$

In the above equation (3), the l-th multi-path component is denoted by $\beta_k^{(l)}\mathbf{a}(\psi_k^{(l)})$, where $\beta_k^{(l)}$ represents the complex channel gain, and $\mathbf{a}(\psi_k^{(l)})$ is the antenna response vector for the spatial direction $\psi_k^{(l)}$. Besides, the line-of-sight (LoS) component is denoted by $\beta_k^{(0)}\mathbf{a}(\psi_k^{(0)})$, and the non-line-of-sight (NLoS) components are denoted by $\beta_k^{(l)}\mathbf{a}(\psi_k^{(l)})$, with l set to $1, 2, \cdots, L$. Based on the spatial channel vector $\mathbf{h}_k$ given by (3), the corresponding beamspace channel vector can be calculated as $\tilde{\mathbf{h}}_k = \mathbf{U}\mathbf{h}_k$, and the N elements of $\tilde{\mathbf{h}}_k$ correspond to the N orthogonal beams. Moreover, the number of dominant elements in $\tilde{\mathbf{h}}_k$ is proved to be approximately in the order of L, which makes $\tilde{\mathbf{h}}_k$ sparse since L is much smaller than its dimension N [11]. Utilizing this sparse characteristic, we adopt the classic maximum-magnitude based beam selection method [12] to select a small part of the beams for the accommodation of all users without sacrificing the system performance. Since usually one RF chain generates one beam, beam selection mentioned above can largely reduce the number of RF chains, which brings the advantages of energy-efficiency and cost-effectiveness.

However, reducing the number of RF chains also causes the problem of limited connections. Utilizing the same resource block, one beam can only accommodate one user. Since the number of RF chains is limited, the number of connections will also be constrained. To break this limit, NOMA is adopted so that each beam can support multiple users. Specifically, users served by the same beam will be regarded as a NOMA group, where intra-beam superposition coding in power domain is conducted to transmit multi-user signals, and SIC is carried out within the beam for signal detection [23]. Thus, the transmitted signal at the BS can be expressed as

$$x = \sum_{m'=1}^{N_{\mathrm{RF}}} x_{m'} = \sum_{m'=1}^{N_{\mathrm{RF}}} \sum_{n'=1}^{|S_{m'}|} \sqrt{p_{m',n'}}\mathbf{w}_{m'} s_{m',n'}, \tag{4}$$

where $x_{m'}$ denotes the superposed signal for each beam, $S_{m'}$ is the set of users in the m'-th beam, and $|S_{m'}|$ denotes its cardinality. $p_{m',n'}$ is the power allocated to the corresponding user, $\mathbf{w}_{m'}$ is the normalized digital precoding vector for the m'-th beam with $||\mathbf{w}_{m'}|| = 1$, and $s_{m',n'}$ is the normalized transmitted signal with $\mathbb{E}(|s_{m',n'}|^2) = 1$. Then, the received signal $y_{m,n}$ of the n-th user in the m-th beam can be expressed as

$$\begin{aligned} y_{m,n} &= \hat{\mathbf{h}}_{m,n}^H x + v_{m,n} \\ &= \hat{\mathbf{h}}_{m,n}^H \sum_{n'=1}^{|S_m|} \sqrt{p_{m,n'}}\mathbf{w}_m s_{m,n'} \\ &\quad + \hat{\mathbf{h}}_{m,n}^H \sum_{m'\neq m} \sum_{n'=1}^{|S_{m'}|} \sqrt{p_{m',n'}}\mathbf{w}_{m'} s_{m',n'} + v_{m,n}, \end{aligned} \tag{5}$$

where $\hat{\mathbf{h}}_{m,n}^H$ denotes the corresponding beamspace channel vector of size $N_{\mathrm{RF}} \times 1$ after beam selection, and $v_{m,n} \sim \mathcal{CN}(0, \sigma^2)$ denotes the thermal noise. Note that the digital precoding vectors $\{\mathbf{w}_m\}$ are the same for the users served by the same beam. They are calculated by the SVD-based zero-forcing method proposed in [23], which exploits the beamspace channel vectors of all the NOMA users in each beam. With the digital precoding vectors, the NOMA users in the same beam are sorted in the descending order of their effective channel qualities calculated by multiplying channel vectors and digital precoding vectors, i.e., $||\hat{\mathbf{h}}_{m,1}^H \mathbf{w}_m||^2 > ||\hat{\mathbf{h}}_{m,2}^H \mathbf{w}_m||^2 > \cdots > ||\hat{\mathbf{h}}_{m,|S_m|}^H \mathbf{w}_m||^2$.

To decode multi-user signal, SIC is adopted by the users within each beam. Particularly, while carrying out SIC, the i-th user in the m-th beam ($i < n$) will firstly decode the signals for the users with worse effective channel qualities, i.e., the users in the same beam with indices larger than i, including the n-th user. Then, the i-th user will successively remove them from its received signal, in the order of $|S_m|$, $|S_m| - 1$, $\cdots$, $i + 1$, until its own signal can be detected. Therefore, when the i-th user is about to decode the n-th user's signal, the remaining signal $\hat{y}_{m,i(n)}$ is given by

$$\begin{aligned}\hat{y}_{m,i(n)} = &\underbrace{\hat{\mathbf{h}}_{m,i}^H \sqrt{p_{m,n}} \mathbf{w}_m s_{m,n}}_{\text{signal to be decoded}} \\ &\underbrace{+ \hat{\mathbf{h}}_{m,i}^H \sum_{n'=1}^{n-1} \sqrt{p_{m,n'}} \mathbf{w}_m s_{m,n'}}_{\text{intra-beam interference}} \\ &\underbrace{+ \hat{\mathbf{h}}_{m,i}^H \sum_{m' \neq m} \sum_{n'=1}^{|S_{m'}|} \sqrt{p_{m',n'}} \mathbf{w}_{m'} s_{m',n'}}_{\text{inter-beam interference}} + v_{m,i}. \end{aligned} \tag{6}$$

In equation (6), the first term denotes the n-th user's signal to be decoded by the i-th user, the second term represents the intra-beam interference from users in the m-th beam with better effective channel qualities (their indices are smaller than n), and the third term denotes the inter-beam interference from users served by other beams. Based on (6), the signal-to-interference-plus-noise-ratio (SINR) for the i-th user to decode the n-th user's signal in the m-th beam is given by

$$\gamma_{n,i}^m = \frac{\left\| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right\|_2^2 p_{m,n}}{\xi_{n,i}^m + \sigma^2}, \tag{7}$$

where $\xi_{n,i}^m$ denotes the power of the intra and inter-beam interferences:

$$\begin{aligned}\xi_{n,i}^m = &\sum_{n'=1}^{n-1} \left\| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right\|_2^2 p_{m,n'} \\ &+ \sum_{m' \neq m} \sum_{n'=1}^{|S_{m'}|} \left\| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_{m'} \right\|_2^2 p_{m',n'}. \end{aligned} \tag{8}$$

Thus, the corresponding achievable rate can be expressed as $R_{n,i}^m = \log_2 \left(1 + \gamma_{n,i}^m\right)$.

Moreover, in the m-th beam, SIC requires successful detection of the n-th user's signal by the i-th user in the same beam for all $i < n$ [27], [29], [30]. Therefore, the corresponding achievable rate R_n^m is the minimum among a series of rates:

$$R_n^m = \min_{i=1,2,\cdots,n} R_{n,i}^m = \min_{i=1,2,\cdots,n} \log_2 \left(1 + \gamma_{n,i}^m\right). \tag{9}$$

The minimal rate of all NOMA users is given by:

$$R_{\min} = \min_{m,n} \{R_n^m\}. \tag{10}$$

Based on the equations (7) and (8), it is worth pointing out that the nominator of the SINR is the desired signal power, while the denominator includes the power for the intra-beam and inter-beam interferences, both of which are mainly affected by the power allocation parameters $\{p_k\}$. This indicates that the rates of the NOMA users are mainly determined by power allocation. As a result, power allocation problems will be formulated to maximize the minimal rate in both the single-beam case and the multi-beam case. Moreover, by formulating the power allocation problems, we are improving the nominator and reducing the denominator of the corresponding SINRs. This is to say, power allocation can increase the desired signal power and mitigate the inter-beam and intra-beam interferences. Accordingly, two bisection-based methods are proposed in our paper for power allocation, and they are able to suppress the inter-beam and intra-beam interferences, which are presented in Sections III and IV, respectively.

III. Max-Min Fairness Analysis for Single-Beam Beamspace MIMO-NOMA

The maximized minimal rate of beamspace MIMO-NOMA in the single-beam case is analyzed in this section, where only one RF chain is deployed at the BS. Particularly, the minimal rate of all NOMA users is maximized by formulating and solving a single-beam power allocation problem, which is shown to be always larger than that of the traditional beamspace MIMO-OMA.

A. Problem Formulation

For the single-beam beamspace MIMO-NOMA system, the BS is equipped with only one RF chain, which indicates that only one beam can be generated at a time. Therefore, for the ease of notation, we will omit the index for the beam in this section. Moreover, the system is only influenced by intra-beam interference, and the beamspace channel $\hat{h}_n$ for the n-th user after beam selection is a scalar. Simplified from (7), the i-th user can decode the n-th user's signal with an SINR $\gamma_{n,i}$ given by:

$$\gamma_{n,i} = \frac{|\hat{h}_i|^2 p_n}{\sum\limits_{k=1}^{n-1} |\hat{h}_i|^2 p_k + \sigma^2}, \tag{11}$$

where p_k is the power allocated to the k-th user.

Given that in the single-beam case, the order of effective channel quality reduces to $|\hat{h}_1|^2 > |\hat{h}_2|^2 > \cdots > |\hat{h}_K|^2$. Therefore, for any $i < j < n$, the SINR for the i-th user to decode the signal for the n-th user is larger than that of the j-th user to decode the same signal, which is shown by:

$$\gamma_{n,i} - \gamma_{n,j} = \frac{p_n \sigma^2 (|\hat{h}_i|^2 - |\hat{h}_j|^2)}{(\sum\limits_{k=1}^{n-1} |\hat{h}_i|^2 p_k + \sigma^2)(\sum\limits_{k=1}^{n-1} |\hat{h}_j|^2 p_k + \sigma^2)} > 0. \tag{12}$$

Note that $\gamma_{n,i} - \gamma_{n,j} > 0$ is due to the channel ordering $|\hat{h}_i|^2 - |\hat{h}_j|^2 > 0$, which is independent of the specific value of the power coefficients $\{p_k\}$. Thus, we can conclude that equation (12) always holds, which has nothing to do with the specific value of power coefficients. Based on equation (12), for the n-th user, we have $\gamma_{n,1} > \gamma_{n,2} > \cdots > \gamma_{n,n}$, and we know before power allocation that its achievable rate R_n is actually determined by the rate of decoding its signal at itself:

$$\begin{aligned} R_n &= \min_{i=1,2,\cdots,n} R_{n,i} \\ &= \min_{i=1,2,\cdots,n} \log_2(1+\gamma_{n,i}) \\ &= \log_2(1+\gamma_{n,n}) \\ &= R_{n,n}. \end{aligned} \tag{13}$$

Thus, to maximize the minimal rate, the objective function of the single-beam power allocation problem is formulated by

$$\max_{\{p_n\}} \min_{n} \{R_{n,n}\}, \tag{14}$$

and we only need to consider the following two constraints:

$$p_n \geq 0, \ \forall 1 \leq n \leq K, \tag{15}$$

$$\sum_{n=1}^{K} p_n \leq P_{\max}, \tag{16}$$

where (15) ensures that each user is assigned non-negative power, and (16) sets the maximum transmit power $P_{\max}$. Then, the corresponding single-beam power allocation problem $\mathcal{P}_1$ is given by

$$\begin{aligned} \mathcal{P}_1 : &\max_{\{p_n\}} \min_{n} \{R_{n,n}\}, \\ &\text{s.t. } C_1 : p_n \geq 0, \quad \forall 1 \leq n \leq K, \\ &\qquad C_2 : \sum_{n=1}^{K} p_n \leq P_{\max}, \end{aligned} \tag{17}$$

where the max-min rate is denoted by r^*.

The above problem $\mathcal{P}_1$ is challenging, because the achievable rate $R_{n,n}$ is a non-convex function due to the intra-beam interference, as shown by (11) and (13). To cope with this challenge, we will introduce a bisection-based method in the next subsection to solve this problem.

B. Minimal Rate Maximization

Algorithm 1 Bisection-Based Single-Beam Power Allocation Algorithm

```
Input: beamspace channels $\{\hat{h}_n\}$, noise power $\sigma^2$, total power
    $P_{\max}$, desirable accuracy $\epsilon$.
Output: Optimal power allocation parameters $\{p_n^*\}$, max-min
    rate $r^*$.
Set lower bound $t_L = 0$.
Set upper bound $t_H = \log_2(1 + P_{\max} h_{\max}/\sigma^2)$, where
    $h_{\max}$ is the maximal value among all $|\hat{h}_n|^2$.
while $t_H - t_L > \epsilon$ do
    Set $t_0 = (t_H + t_L)/2$, solve problem $\hat{\mathcal{P}}_1$ via equation (20)
    to obtain $\{p_n\}$.
    if $\sum_{n=1}^{K} p_n \leq P_{\max}$ then
        Set $t_L = t_0$, $r^* = t_0$, $\{p_n^*\} = \{p_n\}$.
    else
        Set $t_H = t_0$.
    end if
end while
```

In this subsection, the max-min rate r^* is acquired by utilizing a bisection procedure. To deal with the non-convex objective function, an auxiliary variable t is introduced to equivalently convert the original single-beam power allocation problem $\mathcal{P}_1$ into a new problem $\mathcal{P}_1'$:

$$\begin{aligned} \mathcal{P}_1' : &\max_{\{p_n\}} \ t, \\ &\text{s.t. } C_1 : p_n \geq 0, \quad \forall 1 \leq n \leq K, \\ &\qquad C_2 : \sum_{n=1}^{K} p_n \leq P_{\max}, \\ &\qquad C_3 : R_{n,n} \geq t, \quad \forall 1 \leq n \leq K. \end{aligned} \tag{18}$$

The above converted problem $\mathcal{P}_1'$ resembles the power allocation problem in [27], and the bisection-based method proposed in [27] can also be utilized to solve $\mathcal{P}_1'$. Note that solving this single-beam power allocation problem is not our contribution. Our contributions mainly lie in the revealing of the different interfering patterns of the intra-beam and inter-beam interferences, as well as the analysis of their different effects on the max-min rate in the single-beam and multi-beam cases. These will be detailedly presented in the following sections.

To be more specific, we can set the value of t to a certain t_0 in a bisection manner, and examine whether t_0 can be achieved by the max-min rate r^*, which is detailedly described in **Algorithm 1**. For the examination of t_0, the achievable rates for all NOMA users are firstly constrained to be larger than t_0, and the total power consumption $\sum_{n=1}^{K} p_n$ is minimized to see whether it is smaller than the maximum transmit power $P_{\max}$. If the minimized power consumption is smaller than $P_{\max}$, then t_0 can be achieved, otherwise it cannot be achieved. This brings the following optimization problem $\hat{\mathcal{P}}_1$:

$$\begin{aligned} \hat{\mathcal{P}}_1 : &\min_{\{p_n\}} \ \sum_{n=1}^{K} p_n, \\ &\text{s.t. } C_1 : p_n \geq 0, \quad \forall 1 \leq n \leq K, \\ &\qquad C_2 : R_{n,n} \geq t_0, \quad \forall 1 \leq n \leq K. \end{aligned} \tag{19}$$

According to **Proposition 2** in [27], the minimized power consumption of the problem $\hat{\mathcal{P}}_1$ is achieved by satisfying all the constraints in C_2 with equality, which means that the rates for all NOMA users will be equal. Since the max-min rate is obtained by solving a series of problems $\hat{\mathcal{P}}_1$, it will be achieved when the rates for all NOMA users are equal to one another [27]. Based on this important finding, the optimal solution of $\hat{\mathcal{P}}_1$ can be obtained by solving the following set of equations:

$$\begin{aligned} & R_{n,n} = t_0 \\ \Leftrightarrow\ & \log_2\left(1+\frac{|\hat{h}_n|^2 p_n}{\sum\limits_{k=1}^{n-1}|\hat{h}_n|^2 p_k+\sigma^2}\right) = t_0 \\ \Leftrightarrow\ & p_n = (2^{r_0}-1)\left(\sum_{k=1}^{n-1} p_k + \frac{\sigma^2}{|\hat{h}_n|^2}\right), \end{aligned} \tag{20}$$

and we have $p_1 = (2^{r_0}-1)\frac{\sigma^2}{|\hat{h}_n|^2}$. Based on equations (20), we can obtain the power allocation parameters $p_2, p_3, \cdots, p_K$ successively from p_1, each with linear complexity $\mathcal{O}(K)$ [27]. Since there are altogether K power allocation parameters, the total complexity for solving $\hat{\mathcal{P}}_1$ is $\mathcal{O}(K^2)$, and the optimal power consumption for problem $\hat{\mathcal{P}}_1$ can be calculated as $P^* = \sum_{n=1}^{K} p_n$. By comparing P^* with $P_{\max}$, we can examine whether t_0 can be achieved as mentioned before, and the bisection-based method is shown in **Algorithm 1**, where the lower bound is set to 0, and the upper bound is set to the rate of utilizing full power to serve the strongest user.

According to [31], the bisection-based method will converge to the optimal solution of the original problem P_1 within a desirable accuracy ϵ with linear convergence, which also demonstrates the effectiveness of the method, and the number of required iterations is approximately $\mathcal{O}(\log(\frac{1}{\epsilon}))$ [31]. Since in each iteration the computation complexity to solve $\hat{\mathcal{P}}_1$ for examining t_0 is $\mathcal{O}(K^2)$ [27], the overall complexity of the method can be calculated as $\mathcal{O}(\log(\frac{1}{\epsilon})K^2)$. In order to further analyze single-beam beamspace MIMO-NOMA in terms of its max-min rate, we will carry out simulations in the next subsection, and show that its max-min rate always grows rapidly with respect to SNR.

C. Simulations and Analysis

In this subsection, simulations are carried out for the max-min rate of the beamspace MIMO-NOMA and the beamspace MIMO-OMA in the single-beam case. For the fairness of comparison, the same beam selection scheme is adopted in the two systems. Particularly, beamspace MIMO-OMA adopts the classic OMA scheme of time division multiple access (TDMA) to serve multiple users. Since only one RF chain is deployed at the BS, only one user can be served in each time slot, and the number of time slots is equal to the number of the users. To maximize the minimal rate, the user will be served with full power in each time slot [22]. For both of the single-beam systems, $N = 64$ antennas and $N_{\mathrm{RF}} = 1$ RF chain are deployed at the BS, while 1 LoS component and $L = 5$ NLoS components are assumed for the mmWave channel. The complex gain $\beta_k^{(0)}$ of the LoS component for an arbitrary k-th user follows complex Gaussian distribution $\mathcal{CN}(0, 10)$, and that of the NLoS component $\beta_k^{(l)}$ follows $\mathcal{CN}(0, 1)$, with l varying from 1 to L [23]. The spatial directions $\psi_k^{(0)}$ and $\psi_k^{(l)}$ for the LoS and NLoS components follow uniform distribution within $[-\frac{1}{2}, \frac{1}{2}]$, respectively [23]. Additionally, the simulation results are obtained using CVX [32].

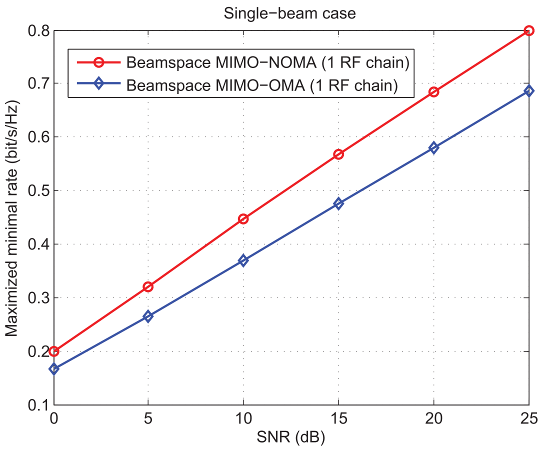

Fig. 2. Max-min rate with respect to the SNR in the single-beam case.

Fig. 2 presents the max-min rates of the two single-beam systems serving 16 users with respect to the SNR, which is defined as $\log_{10}(P_{\max}/\sigma^2)$. It shows that the max-min rates of both systems increase rapidly with the SNR, and the growing slopes of the two corresponding curves almost remain unchanged. This is because single-beam beamspace MIMO-OMA is interference-free, and also because intra-beam interference of the single beam beamspace MIMO-NOMA has no restriction on its max-min rate, which will be detailedly analyzed in Section V. As a result, in the single-beam case, beamspace MIMO-NOMA is superior to beamspace MIMO-OMA in terms of the max-min rate in the whole SNR region, which is because NOMA can efficiently utilize the limited communication resources.

Besides, Fig. 3 depicts the relationship between the two max-min rates and the number of served users, with the SNR set to 18 dB. From this figure, we can see that when the number of served users increases, both of the max-min rates decrease due to the limited maximum transmit power. In addition, since NOMA is adopted, the max-min rate of the beamspace MIMO-NOMA is always larger than that of the beamspace MIMO-OMA in the single-beam case.

IV. Max-Min Fairness Analysis for Multi-Beam Beamspace MIMO-NOMA

In this section, we will conduct the analysis for the multi-beam case, where multiple RF chains are deployed at the BS. Particularly, the difference between the single-beam case and the multi-beam case in terms of the inter-beam interference is pointed out, and a more complicated multi-beam power allocation problem is formulated to maximize the minimal rate, which is also solved by a bisection approach. Moreover, our analysis reveal that beamspace MIMO-NOMA is not always superior to the beamspace MIMO-OMA in the multi-beam case, especially when the inter-beam interference is relatively

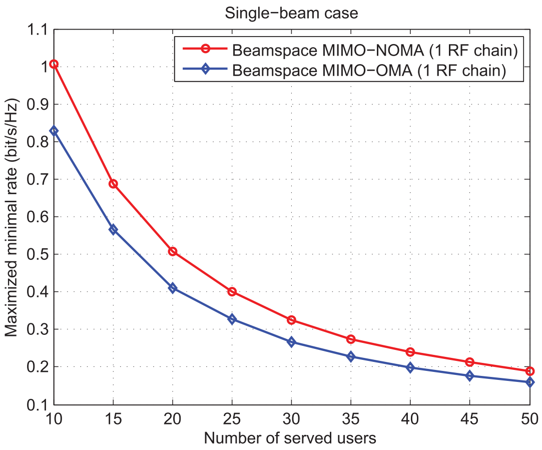

Fig. 3. Max-min rate with respect to the number of served users in the single-beam case.

severe, which is owing to the high SNR or the large number of served users.

A. Problem Formulation

The minimal rate of beamspace MIMO-NOMA in the multi-beam case is maximized by formulating a multi-beam power allocation problem. Note that within each beam, the SIC decoding order is based on the order of effective channel qualities, i.e., $||\hat{\mathbf{h}}_{m,1}^H\mathbf{w}_m||^2 > ||\hat{\mathbf{h}}_{m,2}^H\mathbf{w}_m||^2 > \cdots > ||\hat{\mathbf{h}}_{m,|S_m|}^H\mathbf{w}_m||^2$. According to equations (9)-(10), to ensure successful SIC, the achievable rate R_n^m for the n-th user in the m-th beam is the minimum among a series of rates:

$$R_n^m = \min_{i=1,2,\cdots,n} R_{n,i}^m = \min_{i=1,2,\cdots,n} \log_2\left(1+\gamma_{n,i}^m\right), \quad (21)$$

where $R_{n,i}^m$ is the rate for the i-th user to decode the n-th user's signal, and the corresponding SINR $\gamma_{n,i}^m$ can be expressed as:

$$\gamma_{n,i}^m = \frac{\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_m\right\|_2^2 p_{m,n}}{\xi_{n,i}^m+\sigma^2}. \quad (22)$$

In the above equation (22), $\xi_{n,i}^m$ denotes the power of the intra and inter-beam interferences:

$$\xi_{n,i}^m = \underbrace{\sum_{n'=1}^{n-1}\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_m\right\|_2^2 p_{m,n'}}_{\text{intra}-\text{beam interference}} + \underbrace{\sum_{m'\neq m}\sum_{n'=1}^{|S_{m'}|}\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_{m'}\right\|_2^2 p_{m',n'}}_{\text{inter}-\text{beam interference}}. \quad (23)$$

It is worth pointing out that due to the inter-beam interference term presented in equation (23), we cannot decide whether $\gamma_{n,i}^m$ is larger than $\gamma_{n,j}^m$ $(i<j)$ or not before power allocation:

$$\begin{aligned}
&\gamma_{n,i}^m-\gamma_{n,j}^m\\
&=\frac{p_{m,n}\left[\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_m\right\|_2^2\left(\xi_{n,j}^m+\sigma^2\right)-\left\|\hat{\mathbf{h}}_{m,j}^H\mathbf{w}_m\right\|_2^2\left(\xi_{n,i}^m+\sigma^2\right)\right]}{(\xi_{n,i}^m+\sigma^2)(\xi_{n,j}^m+\sigma^2)}\\
&=\frac{p_{m,n}\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_m\right\|_2^2\left\|\hat{\mathbf{h}}_{m,j}^H\mathbf{w}_m\right\|_2^2}{(\xi_{n,i}^m+\sigma^2)(\xi_{n,j}^m+\sigma^2)}\\
&\times\left(\frac{\sum_{m'\neq m}\sum_{n'}p_{m',n'}\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_{m'}\right\|_2^2+\sigma^2}{\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_m\right\|_2^2}\right.\\
&\left.-\frac{\sum_{m'\neq m}\sum_{n'}p_{m',n'}\left\|\hat{\mathbf{h}}_{m,j}^H\mathbf{w}_{m'}\right\|_2^2+\sigma^2}{\left\|\hat{\mathbf{h}}_{m,j}^H\mathbf{w}_m\right\|_2^2}\right). \quad (24)
\end{aligned}$$

Note that the term

$$\left(\frac{\sum_{m'\neq m}\sum_{n'}p_{m',n'}\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_{m'}\right\|_2^2+\sigma^2}{\left\|\hat{\mathbf{h}}_{m,i}^H\mathbf{w}_m\right\|_2^2}-\frac{\sum_{m'\neq m}\sum_{n'}p_{m',n'}\left\|\hat{\mathbf{h}}_{m,j}^H\mathbf{w}_{m'}\right\|_2^2+\sigma^2}{\left\|\hat{\mathbf{h}}_{m,j}^H\mathbf{w}_m\right\|_2^2}\right)$$

in equation (24) is manipulated from the inter-beam interference term in equation (23), and its sign cannot be determined without the knowledge of power allocation. Therefore, unlike the single-beam case, it is unknown which user decides the rate for the n-th user in the m-th beam before power allocation, which is mainly due to the inter-beam interference. This is a significant difference between the single-beam case and the multi-beam case. Based on the above discussions, the objective function to maximize the minimal rate can be written as

$$\max_{\{p_{m,n}\}}\min_{m,n}\{R_n^m\} = \max_{\{p_{m,n}\}}\min_{\substack{m,n,\\1\le i\le n}}\{R_{n,i}^m\}. \quad (25)$$

Besides, the two power constraints are similar to (15) and (16) in Section III-A. Therefore, the formulated multi-beam power allocation problem is given by

$$\begin{aligned}
\mathcal{P}_2:\ &\max_{\{p_{m,n}\}}\min_{\substack{m,n,\\1\le i\le n}}\{R_{n,i}^m\},\\
&\text{s.t. } C_1: p_{m,n}\ge 0,\quad \forall m,n,\\
&\qquad C_2: \sum_{m=1}^{N_{\text{RF}}}\sum_{n=1}^{|S_m|}||\mathbf{w}_m||^2 p_{m,n}\le P_{\max}, \quad (26)
\end{aligned}$$

and the maximized minimal rate is denoted as r^*.

Similar to its single-beam counterpart, the multi-beam power allocation problem $\mathcal{P}_2$ is also non-convex. Moreover, due to the inter-beam interference, the achievable rate for each user has to be expressed by the minimum of a series of rates, which makes the multi-beam problem more complex. In order to solve this non-convex optimization problem, a bisection approach will be introduced in detail in Section IV-B.

B. Minimal Rate Maximization

Since we aim at maximizing the minimal rate of the multi-beam beamspace MIMO-NOMA system, the existing algorithms used to maximize the sum rate (such as WMMSE) cannot be applied to solve our problem. To this end, a bisection approach is utilized in this subsection to maximize the minimal rate for the system. To deal with the complicated objective function in (26), an auxiliary variable t is introduced to equivalently transform the original objective function into constraint C_3 of the following new problem $\mathcal{P}_2'$:

$$\begin{aligned}\mathcal{P}_2' : \max_{\{p_{m,n}\}} \ & t, \\ \text{s.t. } & C_1 : p_{m,n} \geq 0, \quad \forall m, n, \\ & C_2 : \sum_{m=1}^{N_{\mathrm{RF}}} \sum_{n=1}^{|S_m|} ||\mathbf{w}_m||^2 p_{m,n} \leq P_{\max}, \\ & C_3 : R_{n,i}^m \geq t, \quad \forall m, n, \ 1 \leq i \leq n. \end{aligned} \tag{27}$$

Inspired by **Algorithm 1** in Section III-B, we set the auxiliary variable t to some t_0 in a bisection manner, and construct the following power consumption minimization problem to examine whether t_0 is larger than r^* or not:

$$\begin{aligned}\hat{\mathcal{P}}_2 : \min_{\{p_{m,n}\}} & \sum_{m=1}^{N_{\mathrm{RF}}} \sum_{n=1}^{|S_m|} ||\mathbf{w}_m||^2 \ p_{m,n}, \\ \text{s.t. } & C_1 : p_{m,n} \geq 0, \quad \forall m, n, \\ & C_2 : R_{n,i}^m \geq t_0, \quad \forall m, n, \ 1 \leq i \leq n. \end{aligned} \tag{28}$$

As mentioned before, due to the inter-beam interference, it is unknown which user decides the rate for the n-th user in the m-th beam before power allocation. Therefore, unlike the single-beam case, equations cannot be derived from problem $\hat{\mathcal{P}}_2$, and thus closed-form expressions cannot be directly deduced. Fortunately, the objective function as well as the constraint C_1 of the problem $\hat{\mathcal{P}}_2$ are linear with respect to the power allocation parameters $\{p_{m,n}\}$, while the constraint C_2 can be shown to be linear by the following equation (29):

$$\begin{aligned} & R_{n,i}^m \geq t_0 \\ \Leftrightarrow & \left\| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right\|_2^2 \sum_{k=1}^{n-1} p_{m,k} + \sum_{j \neq m} \left\| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_j \right\|_2^2 \sum_{k=1}^{|S_j|} p_{j,k} \\ & - \frac{\left\| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right\|_2^2 p_{m,n}}{2^{t_0} - 1} + \sigma^2 \leq 0. \end{aligned} \tag{29}$$

Therefore, the problem $\hat{\mathcal{P}}_2$ can also be solved by classic linear programming methods. Denote the minimized power consumption for problem $\hat{\mathcal{P}}_2$ as P^*, we also have $r^* \geq t_0$ when $P^* \leq P_{\max}$, and $r^* < t_0$ when $P^* > P_{\max}$. Correspondingly, the bisection-based method is presented in **Algorithm 2**. Similar to its single-beam counterpart, the bisection-based method in the multi-beam case can also converge to the optimal solution within a desirable accuracy ϵ with linear convergence [31], which demonstrates its effectiveness, and the number of required iterations is also $\mathcal{O}(\log(\frac{1}{\epsilon}))$. Since in each iteration the computation complexity to solve $\hat{\mathcal{P}}_2$ for examining t_0 is $\mathcal{O}(K^{3.5})$ using classic linear programming methods [33], its complexity can be calculated as $\mathcal{O}(\log(\frac{1}{\epsilon})K^{3.5})$.

Algorithm 2 Bisection-Based Multi-Beam Power Allocation Algorithm

```
Input: Total power $P_{\max}$, beamspace channels $\{\hat{\mathbf{h}}_{m,n}\}$, noise power
    $\sigma^2$, precoding vectors $\{\mathbf{w}_m\}$ calculated by the SVD-based zero-
    forcing [23], desirable accuracy $\epsilon$.
Output: Optimal power allocation parameters $\{p_{m,n}^*\}$, max-min rate
    $r^*$.
Set lower bound $t_L = 0$.
Set upper bound $t_H = \log_2(1 + P_{\max} h_{\max} / \sigma^2)$, where $h_{\max}$ is
    the maximal value among all $||\hat{\mathbf{h}}_{m,n}||^2$.
while $t_H - t_L > \epsilon$ do
    Set $t_0 = (t_H + t_L)/2$, solve problem $\hat{\mathcal{P}}_2$ to obtain the power
    allocation parameters $\{p_{m,n}\}$.
    if $\sum_{m=1}^{N_{\mathrm{RF}}} \sum_{n=1}^{|S_m|} ||\mathbf{w}_m||^2 \ p_{m,n} \leq P_{\max}$ then
        Set $t_L = t_0$, $r^* = t_0$, $\{p_{m,n}^*\} = \{p_{m,n}\}$.
    else
        Set $t_H = t_0$.
    end if
end while
```

Based on the bisection procedure presented above, in the following paragraphs, we will prove that max-min rate is achieved when the rates of NOMA users are equal to one another for multi-beam systems. Particularly, for an arbitrary n-th user in the m-th beam, its signal must be decoded by the i-th user for all $i \leq n$, and thus its rate is decided by a certain j-th user for $j \leq n$. This j-th user is unknown to us as mentioned before, but it could be the first user, the second user, etc. Accordingly, the power consumption minimization problem $\hat{\mathcal{P}}_2$ can be decomposed into a set of problems $\mathcal{P}_\bigtriangledown$ by traversing all the possibilities. Since each beam can accommodate $|S_m|$ users, the total number of problems in set $\mathcal{P}_\bigtriangledown$ can be calculated by $\prod_{m=1}^{N_{\mathrm{RF}}}(|S_m|!)$. To be more specific, a certain power consumption minimization problem $\hat{\mathcal{P}}_3$ in set $\mathcal{P}_\bigtriangledown$ can be formulated as

$$\begin{aligned}\hat{\mathcal{P}}_3 : \min_{\{p_{m,n}\}} & \sum_{m=1}^{N_{\mathrm{RF}}} \sum_{n=1}^{|S_m|} ||\mathbf{w}_m||^2 \ p_{m,n}, \\ \text{s.t. } & C_1 : p_{m,n} \geq 0, \quad \forall m, n, \\ & C_2 : R_{n,j(m,n)}^m \geq t_0, \quad \forall m, n, \end{aligned} \tag{30}$$

where $R_{n,j(m,n)}^m$ indicates that the rate for the n-th user in the m-th beam is determined by the $j(m,n)$-th user in the same beam. We can denote the set of solutions for all the problems in set $\mathcal{P}_\bigtriangledown$ as set $\mathcal{S}_\bigtriangledown$, and the solution for problem $\hat{\mathcal{P}}_2$ must lie in set $\mathcal{S}_\bigtriangledown$. Then, we prove that for every problem in set $\mathcal{P}_\bigtriangledown$, its solution satisfies the property that the rates of NOMA users are equal to one another.

Lemma 1: For an arbitrary problem $\hat{\mathcal{P}}_3$ in set $\mathcal{P}_\bigtriangledown$, its solution satisfies the property that the rates of all NOMA users are equal to one another.

Proof: Since problem $\hat{\mathcal{P}}_3$ is decomposed from $\hat{\mathcal{P}}_2$, it is also convex, and KKT condition can be utilized to prove this lemma. Particularly, for an arbitrary n-th user in the m-th beam, the corresponding power allocation parameters satisfy the KKT conditions presented below:

$$\begin{aligned} & (2^{t_0} - 1) \left(\sum_{n'=n+1}^{|S_m|} \lambda_{m,n'} \left\| \hat{\mathbf{h}}_{m,j(m,n')}^H \mathbf{w}_m \right\|_2^2 + \right. \\ & \left. \sum_{m' \neq m, n'} \lambda_{m',n'} \left\| \hat{\mathbf{h}}_{m',j(m',n')}^H \mathbf{w}_m \right\|_2^2 \right) + ||\mathbf{w}_m||^2 \\ & = \lambda_{m,n} \left\| \hat{\mathbf{h}}_{m,j(m,n)}^H \mathbf{w}_m \right\|_2^2 + \mu_{m,n} \end{aligned} \tag{31}$$

$$(2^{t_0}-1)(\xi^m_{n,j(m,n)}+\sigma^2) \leq \left\|\hat{\mathbf{h}}^H_{m,j(m,n)}\mathbf{w}_m\right\|_2^2 p_{m,n} \tag{32}$$

$$\lambda_{m,n}\left((2^{t_0}-1)(\xi^m_{n,j(m,n)}+\sigma^2) - \left\|\hat{\mathbf{h}}^H_{m,j(m,n)}\mathbf{w}_m\right\|_2^2 p_{m,n}\right) = 0 \tag{33}$$

$$\mu_{m,n}p_{m,n} = 0 \tag{34}$$

$$p_{m,n} \geq 0, \lambda_{m,n} \geq 0, \mu_{m,n} \geq 0, \tag{35}$$

where

$$\xi^m_{n,j(m,n)} = \sum_{n'=1}^{n-1}\left\|\hat{\mathbf{h}}^H_{m,j(m,n)}\mathbf{w}_m\right\|_2^2 p_{m,n'} + \sum_{m'\neq m}\sum_{n'=1}^{|S_{m'}|}\left\|\hat{\mathbf{h}}^H_{m,j(m,n)}\mathbf{w}_{m'}\right\|_2^2 p_{m',n'}$$

denotes the power for intra-beam and inter-beam interferences. Since the left-hand-side of equation (32) is positive, the corresponding right-hand side is also positive, which indicates that $p_{m,n} > 0$. Thus, according to equation (34), we have $\mu_{m,n} = 0$. Moreover, since the left-hand-side of equation (IV-B) is positive, its right-hand-side must also be positive. Given that $\mu_{m,n} = 0$, we now have $\lambda_{m,n} > 0$. Based on the above discussions, and note that equation (IV-B) is the complementary slackness condition for constraint C_2 of problem $\hat{\mathcal{P}}_3$, we can now conclude that the solution must satisfy constraint C_2 with equality. In this way, the rates for NOMA users will all be equal to t_0, and this completes the proof. ■

With **lemma 1** proved above, now we can prove that the optimal solution for the original multi-beam power allocation problem $\mathcal{P}_2$ satisfies the property that the rates of all the NOMA users are equal to one another.

Theorem 1: The max-min rate of the original multi-beam power allocation problem $\mathcal{P}_2$ can be achieved when all the users are assigned the same rate.

Proof: Based on **lemma 1**, we can conclude that each solution in set $\mathcal{S}_\nabla$ satisfies the property that the rates of all NOMA users are equal to one another. Since the solution for problem $\hat{\mathcal{P}}_2$ lies in set $\mathcal{S}_\nabla$, the solution for $\hat{\mathcal{P}}_2$ will also satisfy this property. Finally, given that the max-min rate of the original multi-beam power allocation problem $\mathcal{P}_2$ is obtained by solving a series of problems $\hat{\mathcal{P}}_2$, it will be achieved when the rates for all NOMA users are equal to one another [27], which completes the proof. ■

Theorem 1 shows that when we optimize the sum rate of the system with the minimum transmission rate equal to the max-min rate, then the rate for all the NOMA users must be equal to the max-min rate. At this time, there will be no loss in terms of the sum rate between optimizing the minimal rate and optimizing the sum rate. Moreover, when the minimum transmission rate decreases from the max-min rate, the power allocated to the weak user will become smaller, and the optimized sum rate of the system will gradually become larger. Therefore, the sum rate loss between optimizing the minimal rate and optimizing the sum rate will also become larger.

In order to further analyze the performance of the multi-beam beamspace MIMO-NOMA in terms of the max-min rate, we will carry out simulations in the next subsection, and show that the growing slope of its max-min rate obviously decreases in the high-SNR region.

C. Simulations and Analysis

In this subsection, simulations are carried out for the max-min rate of the beamspace MIMO-NOMA and the beamspace MIMO-OMA in the multi-beam case. For the fairness of comparison, the same beam selection scheme is adopted in the two systems. Similar to that in Section III-C, multiple users in each beam of the beamspace MIMO-OMA are served by TDMA, and the number of time slots is equal to the maximum number of users within each beam [24]. Since the BS is equipped with multiple RF chains, multiple users will be served in each time slot. To maximize the minimal rate, for each time slot, zero-forcing is adopted to eliminate interference, and power allocation is optimized using the similar bisection procedure in this paper. Besides, the BS is equipped with $N = 64$ antennas, and the number of RF chains N_{RF} varies from 4 to 8, while the other simulation settings are the same as that in Section III-C.

Fig. 4 presents the max-min rates of the two multi-beam systems serving 16 users with respect to the SNR, where two scenarios are considered with the number of RF chains set to 4 and 8, respectively. It can be shown from the figure that the max-min rate with 8 RF chains is larger than the max-min rate with 4 RF chains. In fact, with the increase of the number of RF chains, the number of beams will also increase. Therefore, the number of users served by each beam will decrease, which mitigates the intra-beam interference. As a result, the max-min rate with more RF chains will be larger than that with less RF chains. Besides, this figure also shows that for both of the scenarios, with the increase of the SNR, the growing slope for the max-min rate of the beamspace MIMO-NOMA obviously decreases in the high-SNR region, while that of the beamspace MIMO-OMA almost remains unchanged. As a result, the max-min rate of the beamspace MIMO-NOMA becomes smaller than that of the beamspace MIMO-OMA in the high-SNR region, while the former is actually larger than the latter in the low-SNR region, which is shown by the intersection points of the two corresponding curves.

To explain this phenomenon, we would like to point out that beamspace MIMO-NOMA is rank-deficient, since the number of the RF chains N_{RF} is smaller than that of the served users K. Thus, the inter-beam interference of beamspace MIMO-NOMA in the multi-beam case cannot be completely eliminated by digital precoding, which largely constrains the growing slope of its max-min rate and makes it interference-limited in the high-SNR region. In contrast, beamspace MIMO-OMA is free of interference due to the TDMA scheme and the zero-forcing method. Therefore, the max-min rate of multi-beam beamspace MIMO-OMA will finally exceed that of the multi-beam beamspace MIMO-NOMA with sufficiently large SNR, which indicates that the intersection point always exists. Note that the intersection

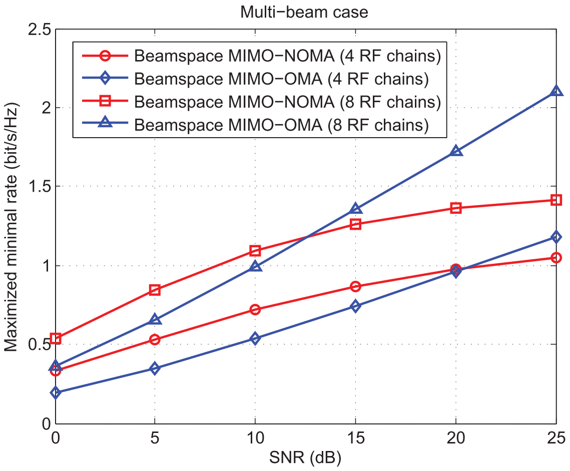

Fig. 4. Max-min rate with respect to the SNR in the multi-beam case.

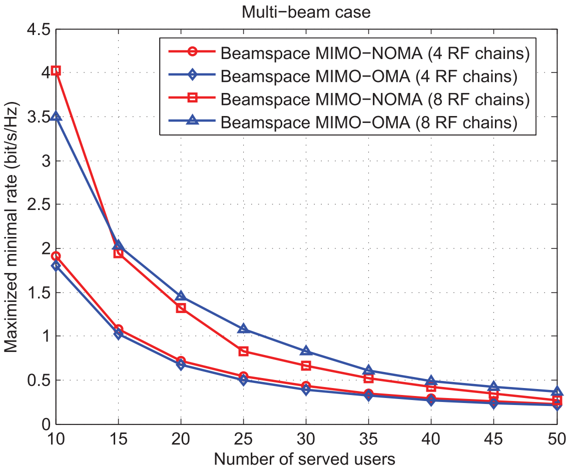

Fig. 5. Max-min rate with respect to the number of served users in the multi-beam case.

point of the two multi-beam systems with 4 RF chains appears at a larger SNR than that with 8 RF chains, and this is because the inter-beam interference of the former is less severe than the latter due to smaller number of beams.

With the SNR set to 18 dB, Fig. 5 depicts the max-min rates of the two systems against the number of served users with N_{RF} set to 4 and 8, respectively. It can be observed from the figure that since the inter-beam interference gets increasingly severe when the number of served users increases, beamspace MIMO-NOMA almost has no advantage over beamspace MIMO-OMA in terms of the max-min rate when serving a large number of users.

So far, we have conducted max-min fairness analysis for beamspace MIMO-NOMA in both the single-beam case and the multi-beam case. Specifically, although affected by the intra-beam interference, the max-min rate of the single-beam beamspace MIMO-NOMA is not interference-limited, and is always larger than that of the single-beam beamspace MIMO-OMA, as shown in Fig. 2. However, the max-min rate of the multi-beam beamspace MIMO-NOMA affected by the inter-beam interference is interference-limited, and is smaller than that of the multi-beam beamspace MIMO-OMA in high-SNR region, as shown in Fig. 4. To explain this difference, we will point out that the interfering patterns of intra-beam and inter-beam interferences are completely different in the next section.

V. Different Patterns of Intra-Beam and Inter-Beam Interferences

In this section, we will comprehensively analyze the different interfering patterns of the intra-beam and inter-beam interferences, and point out their different effects on the max-min rate. Particularly, the intra-beam interference is of successive pattern, which is proved to have no restriction on the max-min rate. However, the inter-beam interference is of mutual pattern, which is proved to restrict the max-min rate to a derived upper bound.

A. Successive Pattern of the Intra-Beam Interference

In the single-beam beamspace MIMO-NOMA system, the n-th user's achievable rate R_n is actually determined by the rate of decoding its signal by itself, as shown by equations (12) and (13):

$$R_n = \log_2\left(1+\frac{|\hat{h}_n|^2 p_n}{\sum\limits_{k=1}^{n-1}|\hat{h}_n|^2 p_k+\sigma^2}\right). \tag{36}$$

As shown by the denominator of the above equation (36), the successive pattern of the intra-beam interference means that the interference of the n-th user only comes from the users with better effective channel qualities, i.e., users whose indices are smaller than n. This is because SIC will successively decode and remove the signals of the users with worse effective channel qualities [27]. Therefore, assume that the n-th user is interfered by the k-th user for $k < n$, then the k-th user is not interfered by the n-th user, since the interference of the k-th user only comes from the users with indices smaller than k. This indicates that the users will not mutually interfere with one another in the single-beam case. As a result, we will show that the intra-beam interference will not limit the max-min rate by proving the following **Proposition 1**:

Proposition 1: In the single-beam beamspace MIMO-NOMA system, the max-min rate r^* will not be limited by the intra-beam interference, and can achieve any arbitrarily large value r_0 as long as the total transmit power P is sufficient.

Proof: We will prove this proposition by showing that the achievable rate of each NOMA user in the single-beam beamspace MIMO-NOMA system can achieve any arbitrarily large r_0 with sufficiently large P. Particularly, we plan to firstly set the achievable rates for all NOMA users equal to the arbitrarily large r_0, and examine whether this can be achieved by calculating the corresponding power allocation parameters $\{p_n\}$. Thus, the following set of equations are formulated:

$$R_n = r_0 \Leftrightarrow \left(\sum_{k=1}^{n-1} p_k - \frac{p_n}{2^{r_0}-1}\right)|\hat{h}_n|^2+\sigma^2 = 0. \tag{37}$$

From the above equation (37), we have

$$p_n = (2^{r_0} - 1)\left(\sum_{k=1}^{n-1} p_k + \frac{\sigma^2}{|\hat{h}_n|^2}\right). \tag{38}$$

Therefore, once the power allocation parameters $\{p_k\}_{k=1}^{n-1}$ are obtained, the power allocation parameter p_n can be calculated by the above equation (38). Note that p_1 can be directly calculated by $p_1 = (2^{r_0} - 1)\frac{\sigma^2}{|\hat{h}_n|^2}$, and thus we can obtain the other power allocation parameters $p_2, p_3, \cdots, p_K$ successively with (38). Moreover, it is straightforward to find that all the power allocation parameters are no less than zero. Since the corresponding power allocation parameters can be calculated, the achievable rate of each NOMA user can achieve any arbitrarily large r_0, and thus the max-min rate r^* of all NOMA users can achieve r_0, which completes the proof. ■

B. Mutual Pattern of the Inter-Beam Interference

In the multi-beam case, the inter-beam interference is of mutual pattern, which refers to the fact that different users served by different beams interfere with one another. To be more specific, take the n_1-th user in the m_1-th beam as an example, according to (8), the power of all the interferences for decoding its signal at itself is given by

$$\begin{aligned}\xi_{n_1,n_1}^{m_1} = &\sum_{k=1}^{n_1-1} \left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_1}\right\|_2^2 p_{m_1,k} \\ &+ \sum_{k=1}^{|S_{m_2}|} \left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_2}\right\|_2^2 p_{m_2,k} \\ &+ \sum_{j\neq m_1,m_2} \sum_{k=1}^{|S_j|} \left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_j\right\|_2^2 p_{j,k},\end{aligned} \tag{39}$$

where $m_2 \neq m_1$ is the index for another beam. Besides, the power of interference caused by users in the m_2-th beam is denoted by the second term of the right hand side in (39). This indicates that an arbitrary n_2-th user in the m_2-th beam will interfere with the n_1-th user in the m_1-th beam, and vice versa, which accounts for the mutual pattern of the inter-beam interference. Therefore, even if we only focus on these two arbitrary users mentioned above, the following **Proposition 2** shows that the inter-beam interference between them will largely constrain the max-min rate, and the corresponding upper bound is also derived:

Proposition 2: In the multi-beam beamspace MIMO-NOMA system, for any m_1, m_2, n_1, and n_2 satisfying

$$\begin{aligned}&m_1 \neq m_2, 1 \leq m_1, m_2 \leq N_{\mathrm{RF}}, \\ &1 \leq n_1 \leq |S_{m_1}|, 1 \leq n_2 \leq |S_{m_2}|,\end{aligned} \tag{40}$$

we have

$$r^* < \max\left\{B_{m_1,n_1}^{m_2,n_2}, B_{m_2,n_2}^{m_1,n_1}\right\}, \tag{41}$$

where

$$B_{m_2,n_2}^{m_1,n_1} = \log_2\left(1 + \frac{\left\|\hat{\mathbf{h}}_{m_2,n_2}^H \mathbf{w}_{m_2}\right\|_2^2}{\left\|\hat{\mathbf{h}}_{m_2,n_2}^H \mathbf{w}_{m_1}\right\|_2^2}\right), \tag{42}$$

and

$$B_{m_1,n_1}^{m_2,n_2} = \log_2\left(1 + \frac{\left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_1}\right\|_2^2}{\left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_2}\right\|_2^2}\right). \tag{43}$$

Proof: We will prove this proposition by showing that for any power allocation parameters $\{p_{m,n}\}$, the minimal rate of all NOMA users $r_{\min}$ satisfies $r_{\min} < \max\left\{B_{m_1,n_1}^{m_2,n_2}, B_{m_2,n_2}^{m_1,n_1}\right\}$.

Firstly, since the overall minimum must be no larger than the local minimum, the minimal rate of all NOMA users must be no larger than the minimal rate of two arbitrary users. Therefore, according to (25), we have

$$\begin{aligned}r_{\min} &= \min_{m,n}\{R_n^m\} \\ &= \min_{\substack{m,n,\\1\leq i\leq n}} \{R_{n,i}^m\} \\ &\leq \min_{m,n}\{R_{n,n}^m\} \\ &\leq \min\{R_{n_1,n_1}^{m_1}, R_{n_2,n_2}^{m_2}\}.\end{aligned} \tag{44}$$

To elaborate a little further, $R_{n_1,n_1}^{m_1}$ is the achievable rate for user n_1 in beam m_1 to decode its own signal:

$$R_{n_1,n_1}^{m_1} = \log_2\left(1 + \frac{\left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_1}\right\|_2^2 p_{m_1,n_1}}{\xi_{n_1,n_1}^{m_1} + \sigma^2}\right). \tag{45}$$

For this user, its interference includes intra-beam interference within the same beam and inter-beam interference from all the users in every other beam. Therefore, the interference power must be larger than the power of the interference only coming from user n_2 in beam m_2:

$$\begin{aligned}\xi_{n_1,n_1}^{m_1} = &\sum_{k=1}^{n_1-1} \left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_1}\right\|_2^2 p_{m_1,k} \\ &+ \sum_{k=1}^{|S_{m_2}|} \left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_2}\right\|_2^2 p_{m_2,k} \\ &+ \sum_{j\neq m_1,m_2} \sum_{k=1}^{|S_j|} \left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_j\right\|_2^2 p_{j,k} \\ > &\left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_2}\right\|_2^2 p_{m_2,n_2}.\end{aligned} \tag{46}$$

Combining (45) and (46), it holds that

$$\begin{aligned}R_{n_1,n_1}^{m_1} &= \log_2\left(1 + \frac{\left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_1}\right\|_2^2 p_{m_1,n_1}}{\xi_{n_1,n_1}^{m_1} + \sigma^2}\right) \\ &< \log_2\left(1 + \frac{\left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_1}\right\|_2^2 p_{m_1,n_1}}{\left\|\hat{\mathbf{h}}_{m_1,n_1}^H \mathbf{w}_{m_2}\right\|_2^2 p_{m_2,n_2}}\right).\end{aligned} \tag{47}$$

Similarly, for user n_2 in beam m_2, it also holds that

$$R_{n_2,n_2}^{m_2} < \log_2\left(1 + \frac{\left\|\hat{\mathbf{h}}_{m_2,n_2}^H \mathbf{w}_{m_2}\right\|_2^2 p_{m_2,n_2}}{\left\|\hat{\mathbf{h}}_{m_2,n_2}^H \mathbf{w}_{m_1}\right\|_2^2 p_{m_1,n_1}}\right). \tag{48}$$

Based on (47), when $p_{m_1,n_1} \leq p_{m_2,n_2}$, we have

$$\begin{aligned}
&\min\{R^{m_1}_{n_1,n_1}, R^{m_2}_{n_2,n_2}\}\\
&\quad \leq R^{m_1}_{n_1,n_1}\\
&\quad < \log_2\left(1+\frac{\left\|\hat{\mathbf{h}}^H_{m_1,n_1}\mathbf{w}_{m_1}\right\|_2^2}{\left\|\hat{\mathbf{h}}^H_{m_1,n_1}\mathbf{w}_{m_2}\right\|_2^2}\right)\\
&\quad = B^{m_2,n_2}_{m_1,n_1},
\end{aligned} \tag{49}$$

and based on (48) it holds that

$$\begin{aligned}
&\min\{R^{m_1}_{n_1,n_1}, R^{m_2}_{n_2,n_2}\}\\
&\quad \leq R^{m_2}_{n_2,n_2}\\
&\quad < \log_2\left(1+\frac{\left\|\hat{\mathbf{h}}^H_{m_2,n_2}\mathbf{w}_{m_2}\right\|_2^2}{\left\|\hat{\mathbf{h}}^H_{m_2,n_2}\mathbf{w}_{m_1}\right\|_2^2}\right)\\
&\quad = B^{m_1,n_1}_{m_2,n_2},
\end{aligned} \tag{50}$$

when $p_{m_1,n_1} > p_{m_2,n_2}$. With the help of the equations (49) and (50) mentioned above, we can now claim that no matter $p_{m_1,n_1} \leq p_{m_2,n_2}$ or $p_{m_1,n_1} > p_{m_2,n_2}$, it always holds that

$$\min\{R^{m_1}_{n_1,n_1}, R^{m_2}_{n_2,n_2}\} < \max\{B^{m_2,n_2}_{m_1,n_1}, B^{m_1,n_1}_{m_2,n_2}\}. \tag{51}$$

According to (44) and (51), we have

$$\begin{aligned}
r_{\min} &\leq \min\{R^{m_1}_{n_1,n_1}, R^{m_2}_{n_2,n_2}\}\\
&< \max\{B^{m_2,n_2}_{m_1,n_1}, B^{m_1,n_1}_{m_2,n_2}\}
\end{aligned} \tag{52}$$

for any power allocation parameters $\{p_{m,n}\}$, and thus

$$r^* = \max_{\{p_{m,n}\}} r_{\min} < \max\{B^{m_2,n_2}_{m_1,n_1}, B^{m_1,n_1}_{m_2,n_2}\}. \tag{53}$$

This completes the proof. ■

It is worth pointing out that due to the mutual pattern of the inter-beam interference between user n_1 in beam m_1 and user n_2 in beam m_2, the two corresponding rates $R^{m_1}_{n_1,n_1}$ and $R^{m_2}_{n_2,n_2}$ are upper-bounded by $B^{m_2,n_2}_{m_1,n_1}$ and $B^{m_1,n_1}_{m_2,n_2}$, respectively, as shown by (49) and (50), which finally leads to the upper bound for the max-min rate r^*. Moreover, note that the two terms $\left\|\hat{\mathbf{h}}^H_{m_2,n_2}\mathbf{w}_{m_1}\right\|_2^2$ and $\left\|\hat{\mathbf{h}}^H_{m_1,n_1}\mathbf{w}_{m_2}\right\|_2^2$ are on the denominator of $B^{m_2,n_2}_{m_1,n_1}$ and $B^{m_1,n_1}_{m_2,n_2}$, respectively, and they denote the interference from the m_1-th beam to the m_2-th beam and vice versa. They show that the max-min rate is mainly restricted by the inter-beam interference. Specifically, when the inter-beam interference is more severe, the max-min rate will become more restricted, and the upper bound will become smaller. Besides, since the multi-beam beamspace MIMO-OMA is free of interference, its max-min rate will finally exceed that of the multi-beam beamspace MIMO-NOMA with sufficiently large SNR, which accounts for the intersection point in Fig. 4.

In addition, the upper-bound obtained in **Proposition 2** is only relevant to a pair of users served by two different beams. Thus, once two arbitrary users in different beams are selected, we can utilize **Proposition 2** to calculate one corresponding upper bound, and a set $\mathcal{U}$ of the upper bounds can be constructed in this way: $\mathcal{U} = \left\{\max\left\{B^{m_2,n_2}_{m_1,n_1}, B^{m_1,n_1}_{m_2,n_2}\right\} | m_1, m_2, n_1, n_2 \text{ satisfiying } (40)\right\}$.

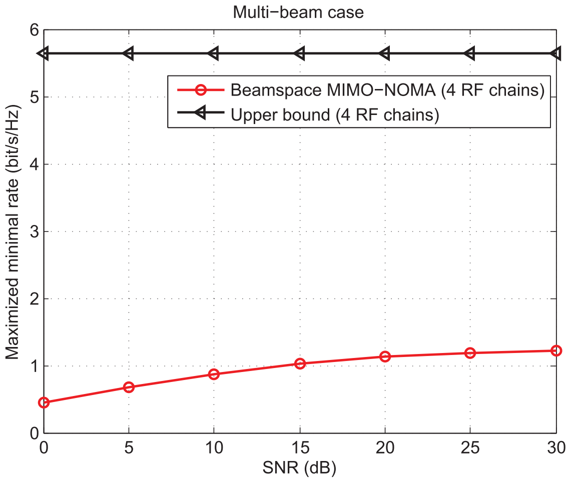

Fig. 6. Max-min rate and the corresponding upper bound with respect to the SNR in the multi-beam case with 4 RF chains.

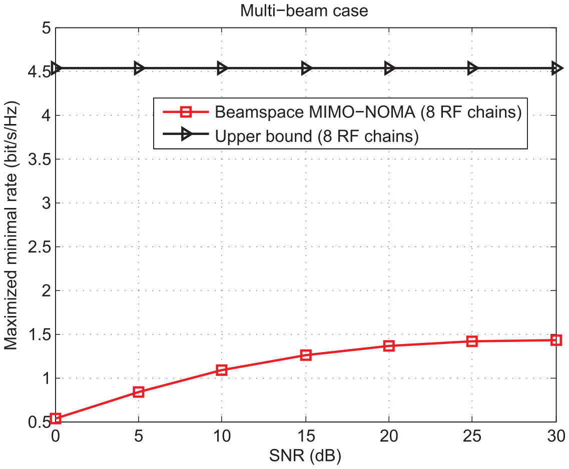

Fig. 7. Max-min rate and the corresponding upper bound with respect to the SNR in the multi-beam case with 8 RF chains.

With the set $\mathcal{U}$ constructed as mentioned above, a more tighter upper bound r_{bound} can be obtained by searching for the minimum in set $\mathcal{U}$:

$$\begin{aligned}
r_{\text{bound}} = \min&\left\{\max\left\{B^{m_2,n_2}_{m_1,n_1}, B^{m_1,n_1}_{m_2,n_2}\right\}\right\},\\
\text{s.t. } &\max\left\{B^{m_2,n_2}_{m_1,n_1}, B^{m_1,n_1}_{m_2,n_2}\right\} \in \mathcal{U}.
\end{aligned} \tag{54}$$

Using the formulas of combination, the number of elements in set $\mathcal{U}$ can be readily calculated as: $|\mathcal{U}| = C_K^2 - \sum_{m=1}^{N_{\text{RF}}} C^2_{|\mathcal{S}_m|}$. Therefore, the minimum in set $\mathcal{U}$ can always be obtained by searching among $|\mathcal{U}|$ elements. In the next subsection, we will verify the derived upper bound r_{bound} in (54) via simulations.

C. Simulations and Analysis

We conduct simulations in this subsection to verify the derived upper bound for the max-min rate of the multi-beam beamspace MIMO-NOMA, while maintaining the same simulation settings as that in Section IV-C. Particularly, two scenarios with N_{RF} varying from 4 to 8 are considered. Fig. 6 presents the derived upper bound calculated by (54) and the max-min rate with respect to the SNR for $N_{\rm RF} = 4$, while those for $N_{\rm RF} = 8$ are shown in Fig. 7. Both of the two figures show that the max-min rate nearly stops growing in the high-SNR region, which demonstrates the effect of the inter-beam interference. Moreover, the max-min rate is always smaller than the derived upper bound, which verifies our theoretical analysis in Section V-B. Note that the upper bound is larger for smaller number of RF chains. This is because with smaller number of RF chains, the number of beams is also smaller, and the inter-beam interference will become less severe. As mentioned in Section V-B, when the inter-beam interference is less severe, the max-min rate will become less restricted, and the upper bound will become larger.

VI. Conclusion

In this paper, we have maximized the minimal rate of beamspace MIMO-NOMA in both the single-beam and multi-beam cases. Particularly, in the single-beam case, we find that the maximized minimal rate of beamspace MIMO-NOMA grows rapidly in the whole SNR region, while that in the multi-beam case grows slower and slower and becomes interference-limited in the high-SNR region. To explain this difference, we have comprehensively analyzed the different interfering patterns of the intra-beam and inter-beam interferences, pointed out their different effects on the maximized minimal rate, and derived the upper bound for the multi-beam case. More importantly, when considering max-min rate in the multi-beam case, we have demonstrated that beamspace MIMO-NOMA is superior to beamspace MIMO-OMA in the low SNR region, while the former becomes inferior to the latter in the high-SNR region. Note that the above analysis does not deny the application of NOMA, but points out that NOMA also has its weaknesses in spite of its known strengths in terms of the connection number, sum capacity, etc., which completes the analysis framework and provides a deeper understanding for NOMA.

References

[1] R. Jiao, L. Dai, W. Wang, F. Lyu, N. Cheng, and X. Shen, "Power allocation for multi-beam max-min fairness in millimeter-wave beamspace MIMO-NOMA," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Honolulu, HI, USA, Dec. 2019, pp. 1–6.

[2] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[3] W. Wu, N. Cheng, N. Zhang, P. Yang, K. Aldubaikhy, and X. Shen, "Performance analysis and enhancement of beamforming training in 802.11ad," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5293–5306, May 2020.

[4] K. Aldubaikhy, W. Wu, Q. Ye, and X. Shen, "Low-complexity user selection algorithms for multiuser transmissions in mmWave WLANs," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2397–2410, Apr. 2020.

[5] S. Mumtaz, J. Rodriquez, and L. Dai, *MmWave Massive MIMO: A Paradigm for 5G*. New York, NY, USA: Academic, 2016.

[6] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[7] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, Sep. 2014.

[8] R. W. Heath, Jr., N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[9] 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.

[10] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[11] Y. Zeng and R. Zhang, "Millimeter wave MIMO with lens antenna array: A new path division multiplexing paradigm," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1557–1571, Apr. 2016.

[12] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 3679–3684.

[13] X. Gao, L. Dai, S. Zhou, A. M. Sayeed, and L. Hanzo, "Wideband beamspace channel estimation for millimeter-wave MIMO systems relying on lens antenna arrays," *IEEE Trans. Signal Process.*, vol. 67, no. 18, pp. 4809–4824, Sep. 2019.

[14] L. Zhu, Z. Xiao, X.-G. Xia, and D. O. Wu, "Millimeter-wave communications with non-orthogonal multiple access for B5G/6G," *IEEE Access*, vol. 7, pp. 116123–116132, 2019.

[15] 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.

[16] 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.

[17] L. Qian, Y. Wu, F. Jiang, N. Yu, W. Lu, and B. Lin, "NOMA assisted multi-task multi-access mobile edge computing via deep reinforcement learning for industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5688–5698, Aug. 2021.

[18] Y. Wu *et al.*, "Optimal power allocation and scheduling for non-orthogonal multiple access relay-assisted networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 11, pp. 2591–2606, Nov. 2018.

[19] H. He, H. Shan, A. Huang, Q. Ye, and W. Zhuang, "Partial NOMA-based resource allocation for fairness in LTE-U system," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Honolulu, HI, USA, Dec. 2019, pp. 1–6.

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

[21] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Joint Tx-Rx beamforming and power allocation for 5G millimeter-wave non-orthogonal multiple access networks," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5114–5125, Jul. 2019.

[22] Z. Xiao, L. Zhu, Z. Gao, D. O. Wu, and X. Xia, "User fairness non-orthogonal multiple access (NOMA) for millimeter-wave communications with analog beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3411–3423, Jul. 2019.

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

[24] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmWave systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1705–1719, Feb. 2019.

[25] K. Wang, J. Cui, Z. Ding, and P. Fan, "Stackelberg game for user clustering and power allocation in millimeter wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2842–2857, May 2019.

[26] K. Senel, H. V. Cheng, E. Björnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 597–611, Jun. 2019.

[27] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[28] M. R. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[29] 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.

[30] Q. Zhang, Q. Li, and J. Qin, "Robust beamforming for nonorthogonal multiple-access systems in MISO channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10231–10236, Dec. 2016.

[31] E. Hildebrand, *Introduction to Numerical Analysis*. New York, NY, USA: Dover, 1987.

[32] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

[33] Y. Ye, *Interior Point Algorithms: Theory and Analysis*. New York, NY, USA: Wiley, 1997.

Ruicheng Jiao (Student Member, IEEE) received the B.S. degree in physics from Tsinghua University, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests include mmWave communications, non-orthogonal multiple access, and reconfigurable intelligent surface (RIS). He has received the IEEE COMMUNICATIONS LETTERS Exemplary Reviewer Award in 2017.

Linglong Dai (Fellow, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2003, the M.S. degree (Hons.) from the China Academy of Telecommunications Technology, Beijing, China, in 2006, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2011.

From 2011 to 2013, he was a Post-Doctoral Research Fellow with the Department of Electronic Engineering, Tsinghua University, where he was an Assistant Professor from 2013 to 2016 and has been an Associate Professor since 2016. He has coauthored the book *MmWave Massive MIMO: A Paradigm for 5G* (Academic Press, 2016). He has authored or coauthored over 60 IEEE journal articles and over 40 IEEE conference papers. He also holds 19 granted patents. His current research interests include massive MIMO, reconfigurable intelligent surface (RIS), millimeter-wave/terahertz communications, and machine learning for wireless communications. He received five IEEE best paper awards at the IEEE ICC 2013, the IEEE ICC 2014, the IEEE ICC 2017, the IEEE VTC 2017-Fall, and the IEEE ICC 2018. He also received the Tsinghua University Outstanding Ph.D. Graduate Award in 2011, the Beijing Excellent Doctoral Dissertation Award in 2012, the China National Excellent Doctoral Dissertation Nomination Award in 2013, the URSI Young Scientist Award in 2014, the IEEE TRANSACTIONS ON BROADCASTING Best Paper Award in 2015, the *Electronics Letters* Best Paper Award in 2016, the National Natural Science Foundation of China for Outstanding Young Scholars in 2017, the IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award in 2017, the IEEE ComSoc Asia–Pacific Outstanding Paper Award in 2018, the China Communications Best Paper Award in 2019, the IEEE ACCESS Best Multimedia Award in 2020, and the IEEE Communications Society Leonard G. Abraham Prize in 2020. He was listed as a Highly Cited Researcher by Clarivate in 2020 and 2021. He was elevated as an IEEE Fellow in 2022. He is currently an Area Editor of IEEE COMMUNICATIONS LETTERS, and an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. Particularly, he is also dedicated to reproducible research and has made a large amount of simulation code publicly available.

Wei Wang (Member, IEEE) received the B.Eng. degree in information countermeasure technology and the M.Eng. degree in signal and information processing from Xidian University in 2011 and 2014, respectively, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2018. From September 2018 to August 2019, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Professor with the Nanjing University of Aeronautics and Astronautics. His research interests include wireless communications, space-air-ground integrated networks, wireless security, and blockchain. He was awarded the IEEE Student Travel Grants for IEEE ICC 2017 and the Chinese Government Award for Outstanding Self-Financed Students Abroad.

Feng Lyu (Member, IEEE) received the B.S. degree in software engineering from Central South University, Changsha, China, in 2013, and the Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2018. From October 2016 to October 2017 and September 2018 to December 2019, he worked as a Post-Doctoral Fellow and a Visiting Ph.D. Student with BBCR Group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Professor with the School of Computer Science and Engineering, Central South University. His research interests include vehicular networks, beyond 5G networks, big data measurement and application design, and could/edge computing. He served as a TPC member for many international conferences. He is a member of the IEEE Computer Society, the IEEE Communication Society, and the IEEE Vehicular Technology Society. He was a recipient of the Best Paper Award of IEEE ICC 2019. He also serves as an Associate Editor for IEEE SYSTEMS JOURNAL and a Leading Guest Editor for *Peer-to-Peer Networking and Applications*.

Nan Cheng (Member, IEEE) received the B.S. and M.S. degrees from the Department of Electronics and Information Engineering, Tongji University, Shanghai, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, in 2016. He worked as a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, from 2017 to 2019. He is currently a Professor with State Key Lab. of ISN, and the School of Telecommunication Engineering, Xidian University, Shaanxi, China. His current research focuses on B5G/6G, space-air-ground integrated network, big data in vehicular networks, self-driving systems, performance analysis, MAC, opportunistic communication, and application of AI for vehicular networks.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is also a registered Professional Engineer in ON, Canada. His research focuses on network resource management, wireless network security, the Internet of Things, 5G and beyond, and vehicular *ad hoc*, and sensor networks. He is an Engineering Institute of Canada Fellow, the Canadian Academy of Engineering Fellow, the Royal Society of Canada Fellow, the Chinese Academy of Engineering Foreign Member, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and the IEEE Communications Society. He received the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021; the R. A. Fessenden Award in 2019 from IEEE, Canada; the Award of Merit from the Federation of Chinese Canadian Professionals, ON, Canada, in 2019; the Technical Recognition Award from Wireless Communications Technical Committee in 2019 and the AHSN Technical Committee in 2013; the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society; and the Education Award in 2017 and Joseph LoCicero Award in 2015 from the IEEE Communications Society (ComSoc). He also received the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2003 from the Province of ON, Canada. He served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom'16, IEEE Infocom'14, IEEE VTC'10 Fall, and IEEE Globecom'07, and the Chair for the IEEE ComSoc Technical Committee on Wireless Communications. He was the Vice President for Technical and Educational Activities, the Vice President for Publications, a Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a Member of IEEE Fellow Selection Committee of the ComSoc. He is also the President Elect of IEEE ComSoc. He served as the Editor-in-Chief of the IEEE INTERNET OF THINGS JOURNAL, IEEE NETWORK, and *IET Communications*.