


# On the Max-Min Fairness of Beam-space MIMO-NOMA

Ruicheng Jiao , *Student Member, IEEE*, and Linglong Dai , *Senior Member, IEEE*

**Abstract**—Non-orthogonal multiple access (NOMA) based beam-space multiple-input multiple-output (MIMO) is a MIMO-NOMA scheme for millimeter-wave (mmWave) communications to improve the number of connections with increased sum rate. However, most of existing works only aim at maximizing the sum rate, which may cause an unbearable rate loss to weak users. To guarantee the rate performance for all served users, we maximize the minimal rate of the system from the max-min fairness perspective, where the NOMA users in the same beam share the same digital precoding vector, i.e., the beam-specific digital precoding, is adopted. The challenge is that, both the inter-beam and intra-beam interferences exist in the system, which makes the minimal rate maximization problem non-convex and thus hard to solve. To cope with this challenge, we propose an alternating optimization method to optimize the power allocation for each user and the digital precoding vector for each beam. Moreover, we break the commonly adopted *beam-specific* digital precoding scheme by using the *user-specific* digital precoding, i.e., each user is assigned with a unique digital precoding vector, to further improve the max-min rate. This can be achieved by the proposed two-stage optimization method, where the user-specific digital precoding vectors are firstly designed, and then the power allocation for all users is finetuned. Simulation results verify the proposed two methods. Moreover, the two-stage optimization method for the user-specific digital precoding outperforms the alternating optimization method for the beam-specific digital precoding, since the former can provide more degrees of freedom for designing the digital precoding vectors.

**Index Terms**—Beam-space MIMO-NOMA, max-min fairness, power allocation, beam-specific digital precoding, user-specific digital precoding.

## I. INTRODUCTION

WITH the rapid development of the applications such as virtual reality (VR), video gaming, etc., smart terminal users are expecting an unprecedented increase of wireless data rate by orders of magnitude, which is also required by the fifth generation (5G) wireless communications and beyond [1]. To meet this demanding requirement, millimeter-wave (mmWave)

Manuscript received February 9, 2020; revised June 30, 2020; accepted August 5, 2020. Date of publication August 21, 2020; date of current version September 10, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. A. Tölli. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1805005 and in part by the National Natural Science Foundation of China for Outstanding Young Scholars under Grant 61722109. (Corresponding author: Linglong Dai.)

The authors are with Beijing National Research Center for Information Science and Technology (BNRist) as well as the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: jjaors16@mails.tsinghua.edu.cn; daili@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TSP.2020.3018559

massive multiple-input multiple-output (MIMO) has attracted extensive investigation from both the academia and the industry, since it utilizes a much wider bandwidth and can simultaneously improve the spectrum efficiency [2]. However, the energy consumption and hardware complexity hindered the practical deployment of mmWave massive MIMO, which is mainly caused by the requirement of a great number of energy-hungry radio frequency (RF) chains [3]. To this end, several solutions have been proposed to reduce the number of RF chains, e.g., the hybrid precoding, the beam-space MIMO, etc. [3]–[6]. Nevertheless, since usually the number of simultaneously supported users can not exceed that of the RF chains, the reduced number of RF chains will limit the number of connections. Thus, it is difficult for mmWave massive MIMO to support massive users, e.g., users at a concert or a sports game [7].

To deal with this problem, non-orthogonal multiple access (NOMA) can be introduced into mmWave massive MIMO to serve more users than RF chains [7]. NOMA is also a prominent technology for 5G and beyond, which is shown to outperform the traditional orthogonal multiple access (OMA) schemes in several aspects, e.g., spectrum efficiency, outage probability, etc. [8]–[10]. Particularly, NOMA superposes the multi-user signals using the same time/frequency resources at the transmitter, and decodes each user's signal via successive interference cancellation (SIC) at the receiver. In this way, the number of simultaneously supported users can be increased at the cost of introducing controllable inter-user interferences. Therefore, the integration of NOMA and mmWave massive MIMO is promising to largely increase the number of connections, which has attracted increasing research interests recently [11]–[15].

### A. Prior Works

For mmWave massive MIMO with the beam-space MIMO structure, multiple users can be served in each beam using NOMA, which is defined as beam-space MIMO-NOMA in [11]. In this way, the number of supported users within the same time/frequency resource block can exceed that of the RF chains. Particularly, the sum rate was maximized by jointly considering the beam selection, digital precoding, and power allocation. The use of NOMA in the hybrid precoding-based mmWave massive MIMO system was investigated in [12], where the sum rate was maximized by jointly optimizing user scheduling and power allocation. Specifically, a branch and bound method was proposed to obtain the optimal solution within a desirable accuracy at first, which was followed by a suboptimal optimization method

based on matching theory and successive convex optimization. Similarly, how to improve the sum rate for a NOMA-based mmWave massive MIMO system was also considered in [13], where the digital precoding vectors were calculated by the modified zero forcing method, and the analog precoding vectors were optimized by the particle swarm algorithm. Moreover, a multi-beam NOMA scheme for the mmWave massive MIMO system was studied in [14], where the sum rate was maximized by the joint optimization of user grouping, antenna allocation, and power allocation. In addition, NOMA was introduced into a mmWave massive MIMO system with simultaneous wireless information and power transfer (SWIPT) [15], where the sum rate was improved by designing both the power allocation for NOMA users and the power splitting factors for SWIPT.

However, all the existing works mentioned above only aim to maximize the sum rate, which may cause serious problems in some cases. As NOMA tends to group users with very different channel qualities [8], maximizing sum rate will lead to the case that strong users occupy most of the system resources, which may cause an unbearable loss to the achievable rates for weak users. In the worst case, the weak users may not be able to get any communication resources [16]. In light of this, the performance metric of user fairness should be considered to ensure the rate performance for all users, which is however seldom studied in existing works. Note that [17] has recently studied the max-min fairness problem in a NOMA-based mmWave massive MIMO system, but only the single-beam scenario where the base station (BS) was equipped with one RF chain was studied, while the more general multi-beam scenario with multiple RF chains was not considered.

### B. Our Contributions

To this end, we study the multi-beam beamspace MIMO-NOMA system with multiple RF chains from the max-min fairness perspective. Specifically, our contributions can be summarized as follows:

- 1) To guarantee the achievable rates for all users, we investigate the max-min fairness problem in the multi-beam beamspace MIMO-NOMA system with both inter-beam and intra-beam interferences. To the best of our knowledge, this is the first work to discuss this problem in the literature.
- 2) We formulate a minimal rate maximization problem for all NOMA users in the system, where the beam-specific digital precoding is considered, i.e., NOMA users in the same beam share the same digital precoding vector. Because of the existence of the inter-beam and intra-beam interferences, the formulated problem is different from the single-beam counterpart, and it is more difficult to solve. In order to mitigate the interferences and improve the max-min rate, we design the power allocation parameters as well as the beam-specific digital precoding vectors. Particularly, we partition the variables into the power allocation block and the digital precoding block, and an alternating optimization method is proposed to optimize them. The effectiveness of the proposed method is verified by simulation results.

- 3) Furthermore, we break the commonly adopted *beam-specific* digital precoding scheme [11], [13]–[15], and investigate the minimal rate maximization problem with the *user-specific* digital precoding, where each user is assigned with a unique precoding vector. Correspondingly, we propose an optimization method consisting of two-stages to further improve the max-min rate. To be more specific, the digital precoding vectors are designed at first by using a bisection-based semidefinite relaxation method, and then the power allocation parameters are finetuned by utilizing a bisection-based linear programming method. Simulation results<sup>1</sup> show that the two-stage optimization method for the user-specific digital precoding can provide higher max-min rate than the alternating optimization method for the beam-specific digital precoding. This is because the former can provide more degrees of freedom for the design of digital precoding vectors.

### C. Organizations and Notations

*Organizations:* The remainder of the paper is organized as follows. Section II describes the basics of beamspace MIMO at first, and then introduces the system model of beamspace MIMO-NOMA with beam-specific digital precoding. Based on the above system model, a minimal rate maximization problem is formulated in Section III, and the alternating optimization method is also proposed. In Section IV, the minimal rate maximization problem with user-specific digital precoding is formulated, which can be solved by the proposed two-stage optimization method. Simulations are carried out in Section V, and finally conclusions are drawn in Section VI.

*Notations:* The upper-case and lower-case boldface letters are used to denote matrices and vectors, respectively.  $\text{rank}(\cdot)$ ,  $\text{Tr}(\cdot)$ ,  $(\cdot)^\dagger$ ,  $(\cdot)^{-1}$ , and  $(\cdot)^H$  denote the rank, trace, Moore-Penrose inversion, inversion, conjugate transpose of the matrix, respectively.  $\|\cdot\|$  denotes the  $\ell_2$ -norm, and  $\mathbb{E}(\cdot)$  denotes the expectation.  $\mathbf{I}_K$  denotes the  $K \times K$  identity matrix, and  $\text{diag}(p_1, p_2, \dots, p_K)$  denotes a diagonal matrix of size  $K \times K$  whose diagonal elements are set as  $p_1, p_2, \dots, p_K$ .  $\mathbb{C}^{K \times 1}$  denotes the set for the  $K$ -dimension complex vector, and  $|\mathcal{B}|$  denotes the number of elements in set  $\mathcal{B}$ . Finally,  $\mathcal{CN}(\mathbf{a}, \mathbf{B})$  denotes the circular symmetric complex Gaussian distribution with mean  $\mathbf{a}$  and covariance matrix  $\mathbf{B}$ .

## II. SYSTEM MODEL

In this section, we will firstly introduce the basics of beamspace MIMO, and then describe the detailed system model of beamspace MIMO-NOMA.

### A. Beamspace MIMO

Beamspace MIMO is an efficient solution to lower the energy consumption and hardware complexity in mmWave massive MIMO systems [4], [5], the basic idea of which is to deploy a well-designed lens antenna array at the BS transforming the spatial channel into the sparse beamspace channel. In this way,

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

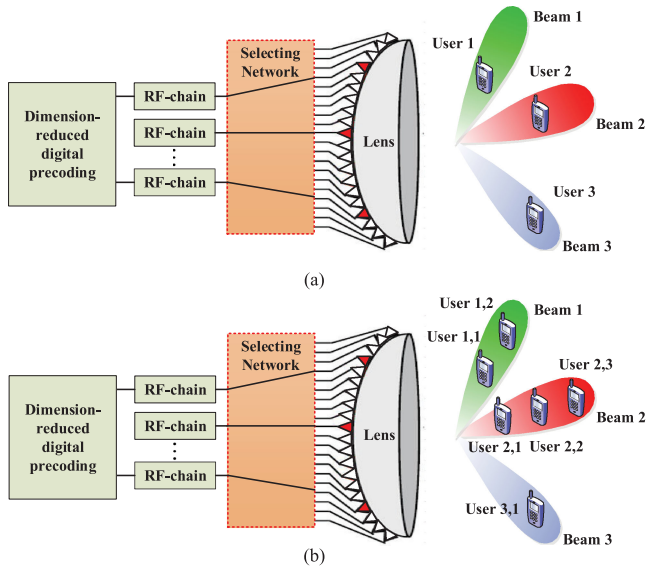


Fig. 1. System models of beamspace structures: (a) beamspace MIMO; (b) beamspace MIMO-NOMA.

only a small number of the beams are needed to serve users without obvious performance loss, and thus the number of required RF chains can be reduced [4]–[6].

Specifically, as shown in Fig. 1(a), the BS is equipped with a lens antenna array of  $N$  antennas and  $N_{\text{RF}}$  RF chains. This lens antenna array acts as a discrete fourier transformation matrix  $\mathbf{U} = [\mathbf{a}(\hat{\psi}_0), \mathbf{a}(\hat{\psi}_1), \dots, \mathbf{a}(\hat{\psi}_{N-1})]^H$ , where  $\hat{\psi}_n = (1/N)(n - (N-1)/2)$  is the predefined spatial direction,  $\mathbf{a}(\hat{\psi}_n) = \frac{1}{\sqrt{N}}[e^{-j2\pi\hat{\psi}_n m}]_{m \in \mathcal{J}(N)}$  denotes the array steering vector for that direction, and  $\mathcal{J}(N) = \{n - (N-1)/2, n = 0, 1, 2, \dots, N-1\}$  [4]. Denote the spatial channel vector and beamspace channel vector for user  $k$  as  $\mathbf{h}_k$  and  $\tilde{\mathbf{h}}_k$ , respectively, the spatial channel matrix  $\mathbf{H}$  for  $K$  served users can be transformed to the beamspace channel matrix  $\tilde{\mathbf{H}}$  as

$$\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \dots, \tilde{\mathbf{h}}_K] = \mathbf{U}\mathbf{H} = [\mathbf{U}\mathbf{h}_1, \dots, \mathbf{U}\mathbf{h}_K], \quad (1)$$

and the  $N$  orthogonal beams with predefined spatial directions  $\hat{\psi}_0, \hat{\psi}_1, \dots, \hat{\psi}_{N-1}$  correspond to the  $N$  rows of  $\tilde{\mathbf{H}}$  [4]–[6].

In this article, the widely-used Saleh-Valenzuela channel model for mmWave channel is considered [4]. The spatial channel vector  $\mathbf{h}_k$  for user  $k$  can be expressed as

$$\mathbf{h}_k = \Omega_k^{(0)} \mathbf{a}(\psi_k^{(0)}) + \sum_{l=1}^L \Omega_k^{(l)} \mathbf{a}(\psi_k^{(l)}), \quad (2)$$

where  $\Omega_k^{(0)}$  represents the complex gain and  $\mathbf{a}(\psi_k^{(0)})$  is the array steering vector for the line-of-sight (LoS) path,  $\Omega_k^{(l)}$  and  $\mathbf{a}(\psi_k^{(l)})$  denote the complex gain and steering vector for the  $l$ -th non-line-of-sight (NLoS) path, and  $L$  denotes the number of NLoS paths.  $\psi = \frac{d \sin(\theta)}{\lambda}$  is the spatial direction of the channel, where  $d$  is the antenna spacing,  $\theta$  denotes the physical direction of the corresponding path, and  $\lambda$  represents the signal wavelength.

Since only a limited number of dominant scatters exist in the mmWave channel, the number of NLoS paths  $L$  is usually much

smaller than the number of the BS antennas  $N$  [4]. As a result, the number of dominant elements in each beamspace channel vector  $\tilde{\mathbf{h}}_k$  is much less than its dimension, which indicates that  $\tilde{\mathbf{h}}_k$  is sparse. Based on this sparse nature, beamspace MIMO selects only a part of the  $N$  orthogonal beams by classical beam selection algorithms, e.g., the maximum magnitude-based beam selection [4], or the maximization of the signal-to-noise-plus-interference-ratio selection [5], etc., to serve all  $K$  users without obvious performance loss. Thus, the received signal after beam selection can be written as

$$\mathbf{y} = \tilde{\mathbf{H}}_r^H \mathbf{W}_r \sqrt{\mathbf{P}} \mathbf{s} + \mathbf{v}, \quad (3)$$

where  $\mathbf{s} \in \mathbb{C}^{K \times 1}$  denotes the transmitted signal for all  $K$  users with normalized power, i.e.,  $\mathbb{E}(\mathbf{s}\mathbf{s}^H) = \mathbf{I}_K$ ,  $\mathbf{P} = \text{diag}(p_1, p_2, \dots, p_K)$  is the diagonal power allocation matrix of size  $K \times K$ , and  $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_K)$  denotes the thermal noise. Note that the  $N$  rows of  $\tilde{\mathbf{H}}$  correspond to  $N$  orthogonal beams. Denote the set for the selected beams as  $\mathcal{B}$ , and the beamspace channel matrix after beam selection can be expressed as  $\tilde{\mathbf{H}}_r = \tilde{\mathbf{H}}(b, : )_{b \in \mathcal{B}}$ , where  $\tilde{\mathbf{H}}(b, : )_{b \in \mathcal{B}}$  represents the corresponding rows of the matrix  $\tilde{\mathbf{H}}$  with their indices in set  $\mathcal{B}$ . As usually one beam is generated by one RF chain, the number of the selected beams  $|\mathcal{B}|$  is equal to that of the RF chains  $N_{\text{RF}}$  [11]. Moreover,  $\mathbf{W}_r$  is the dimension-reduced digital precoding matrix with the row dimension equal to  $|\mathcal{B}| = N_{\text{RF}} < N$ . In this way, beamspace MIMO can reduce the number of RF chains, and thus lower the energy consumption as well as the hardware complexity in mmWave massive MIMO systems [4].

However, usually one beam can only support one user using the same time/frequency resource in beamspace MIMO [11]. Thus, the number of simultaneously supported users cannot exceed that of the RF chains, i.e.,  $K \leq N_{\text{RF}}$ . This indicates that the reduced number of RF chains will largely limit the number of simultaneously served users. To break this limit, NOMA can be integrated into the beamspace MIMO system, which will be introduced in detail in the next subsection.

## B. Beamspace MIMO-NOMA

As shown in Fig. 1(b), multiple users can be served by NOMA within each selected beam in the beamspace MIMO system, which constitutes beamspace MIMO-NOMA. In this way, the number of simultaneously supported users can be larger than that of the RF chains. Particularly, the classical maximum magnitude-based beam selection scheme [4] can also be adopted, where the strongest beam will be selected for each user, and the users served by the same beam naturally form a NOMA cluster. The power-domain signal superposition as well as the SIC are both carried out within this cluster. In the following paragraphs, we use the term “beam” to refer to the NOMA cluster.

Note that the adopted beam selection scheme is also beneficial to the max-min fairness, and the two reasons are listed below. Firstly, since the beamspace channels are sparse, selecting the strongest beam for each user can guarantee its rate performance without obvious performance loss. Therefore, the minimal rate performance is also guaranteed, which is beneficial to max-min fairness. Secondly, the beamspace channels of the users sharing



the same strongest beam have correlation with one another, while those in the different beams are poorly correlated [11]. This is beneficial to mitigate inter-beam interference, and thus also facilitates max-min fairness.

After beam selection, the dimension-reduced beamspace channel matrix can be expressed as  $\hat{\mathbf{H}} = \tilde{\mathbf{H}}(\mathcal{B}, \mathcal{U})$ , where  $\mathcal{B}$  is the set for the selected beams, and  $\mathcal{U}$  denotes the set for all the supported users with  $|\mathcal{U}| = K$ . Let  $S_m$  denotes the set of NOMA users served by the  $m$ -th beam,  $m = 1, 2, \dots, N_{\text{RF}}$ , and we have  $S_m \cap S_n = \emptyset$ ,  $\cup_{m=1}^{N_{\text{RF}}} S_m = \mathcal{U}$ . Moreover, the NOMA users served by the same beam will be labeled in the descending order of the channel quality:

$$\|\hat{\mathbf{h}}_{m,1}\|^2 \geq \|\hat{\mathbf{h}}_{m,2}\|^2 \geq \dots \geq \|\hat{\mathbf{h}}_{m,|S_m}|\|^2, \quad (4)$$

where  $\hat{\mathbf{h}}_{m,i}$  is the beamspace channel vector after beam selection for the  $i$ -th user in the  $m$ -th beam, and  $|S_m|$  denotes the number of the NOMA users in the  $m$ -th beam. As commonly adopted in relevant works [11], [13]–[15], we assign the same digital precoding vector to all NOMA users in the same beam, which is defined as *beam-specific* digital precoding in this paper. Since the users in the same beam are assigned the same precoding vector but allocated different power, the digital precoding vectors and the power allocation parameters are separately denoted. Therefore, the digital precoding vectors need to be normalized to avoid repeated calculation of the power allocation. Then, the received signal  $y_{m,i}$  at the  $i$ -th NOMA user in the  $m$ -th beam can be written as

$$y_{m,i} = \hat{\mathbf{h}}_{m,i}^H \sum_{k=1}^{|S_m|} \sqrt{p_{m,k}} \mathbf{w}_m s_{m,k} + \hat{\mathbf{h}}_{m,i}^H \sum_{j \neq m} \sum_{k=1}^{|S_j|} \sqrt{p_{j,k}} \mathbf{w}_j s_{j,k} + v_{m,i}, \quad (5)$$

where  $\mathbf{w}_m = \mathbf{W}_r(:, m)$  is the normalized  $N_{\text{RF}} \times 1$  digital precoding vector for all NOMA users in the  $m$ -th beam,  $p_{m,k}$  is the corresponding power allocation parameter,  $s_{m,k}$  is the transmitted signal with normalized power, and  $v_{m,i} \sim \mathcal{CN}(0, \sigma^2)$  denotes the thermal noise.

For signal detection, SIC will be conducted within each beam so that each user can eliminate the interferences from the users with worse channel qualities. To be more specific, based on the channel ordering  $\|\hat{\mathbf{h}}_{m,1}\|^2 \geq \|\hat{\mathbf{h}}_{m,2}\|^2 \geq \dots \geq \|\hat{\mathbf{h}}_{m,|S_m}|\|^2$ , the  $i$ -th user in the  $m$ -th beam will successively detect and remove from its received signal the  $n$ -th user's signal in the same beam for all  $n$  satisfying  $i < n \leq |S_m|$ . After that, the  $i$ -th user will decode its own signal. This SIC decoding order is widely adopted in relevant works concerning precoding design in NOMA systems [15], [17], [18]. It is worth pointing out that this SIC decoding order may not be optimal, and the system performance may be further improved by carefully designing SIC decoding order, but this is not the focus of this article.

While performing SIC to decode the  $n$ -th user's signal, the remaining signal at the  $i$ -th user in the  $m$ -th beam can be

expressed as

$$\hat{y}_{n,i}^m = \underbrace{\hat{\mathbf{h}}_{m,i}^H \sqrt{p_{m,n}} \mathbf{w}_m s_{m,n}}_{\text{desired signal}} + \underbrace{\hat{\mathbf{h}}_{m,i}^H \sum_{k=1}^{n-1} \sqrt{p_{m,k}} \mathbf{w}_m s_{m,k}}_{\text{intra-beam interferences}} + \underbrace{\hat{\mathbf{h}}_{m,i}^H \sum_{j \neq m} \sum_{k=1}^{|S_j|} \sqrt{p_{j,k}} \mathbf{w}_j s_{j,k}}_{\text{inter-beam interferences}} + \underbrace{v_{m,i}}_{\text{noise}}. \quad (6)$$

According to (6), the signal-to-interference-plus-noise-ratio (SINR) for decoding the  $n$ -th user's signal at the  $i$ -th user ( $n \geq i$ ) in the  $m$ -th beam can be expressed as

$$\gamma_{n,i}^m = \frac{|\hat{\mathbf{h}}_{m,i}^H (\sqrt{p_{m,n}} \mathbf{w}_m)|^2}{\xi_{n,i}^m}, \quad (7)$$

where

$$\xi_{n,i}^m = \sum_{k=1}^{n-1} |\hat{\mathbf{h}}_{m,i}^H (\sqrt{p_{m,k}} \mathbf{w}_m)|^2 + \sum_{j \neq m} \sum_{k=1}^{|S_j|} |\hat{\mathbf{h}}_{m,i}^H (\sqrt{p_{j,k}} \mathbf{w}_j)|^2 + \sigma^2. \quad (8)$$

Therefore, the achievable rate  $R_{n,i}^m$  for decoding the  $n$ -th user's signal at the  $i$ -th user in the  $m$ -th beam can be formulated as

$$R_{n,i}^m = \log_2 (1 + \gamma_{n,i}^m). \quad (9)$$

Note that from the above expressions (7)–(9), it is clear that the digital precoding vectors and the power allocation parameters appear together in the form of  $\sqrt{p_{m,n}} \mathbf{w}_m$ . Thus, the achievable rate can be viewed as the function of the products  $\{\sqrt{p_{m,n}} \mathbf{w}_m\}$ , i.e., the same products will yield same achievable rates.

To ensure successful SIC, the  $n$ -th user's signal should be detectable at each user with an index smaller than  $n$  in the same beam. Thus, the achievable rate  $R_n^m$  of the  $n$ -th user's signal in the  $m$ -th beam should be the minimum of all those rates:

$$R_n^m = \min_{i=1,2,\dots,n} R_{n,i}^m = \min_{i=1,2,\dots,n} \log_2 (1 + \gamma_{n,i}^m), \quad (10)$$

and the minimal rate of all NOMA users can be expressed as

$$R_{\min} = \min_{m,n} \{R_n^m\}. \quad (11)$$

As stated in [16], only maximizing the sum rate may cause an unbearable loss to the achievable rates for weak users. In order to guarantee the rate performance for both the strong users and weak users, we maximize the minimal rate of all NOMA users in this paper, which can also be termed as max-min fairness. In contrast to [19]–[21] which mainly consider maximizing the minimal rate/minimal SINR via power allocation, we formulate the minimal rate maximization problem by optimizing both power allocation and digital precoding. This is because the minimal rate  $R_{\min}$  largely depends on both the power allocation parameters and the digital precoding vectors, as shown by (7), (8), and (11). The corresponding optimization method will be proposed in the next section.

It is worth pointing out that the digital precoding parts of the beamspace MIMO structure and the hybrid precoding structure are almost the same [22]. Therefore, once the analog precoding matrix of the hybrid precoding system can be designed, the optimization method proposed in our paper can be extended to the hybrid precoding system to optimize its digital precoding and power allocation.

### III. ALTERNATING OPTIMIZATION OF POWER ALLOCATION AND BEAM-SPECIFIC DIGITAL PRECODING

In this section, we maximize the minimal rate  $R_{\min}$  in (11) in the beamspace MIMO-NOMA system via alternating optimization of power allocation and beam-specific digital precoding, i.e., users within each beam are assigned the same digital precoding vector. Specifically, we will firstly formulate the max-min fairness problem. Since the problem is non-convex and thus hard to solve, we then propose an alternating optimization method to obtain the solution. Finally, we will provide more insights concerning the optimality of the proposed algorithm.

#### A. Problem Formulation

To guarantee the achievable rates for all NOMA users, we maximize the minimal rate  $R_{\min}$  by optimizing both the power allocation parameters  $\{p_{m,n}\}$  and the beam-specific digital precoding vectors  $\{\mathbf{w}_m\}$ . Thus, the objective function can be expressed as

$$\max_{\{p_{m,n}, \mathbf{w}_m\}} \min_{m,n} \{R_n^m\}. \quad (12)$$

Given that the power allocation parameters must be non-negative, and that the total transmission power is limited, the corresponding power constraints can be expressed as

$$p_{m,n} \geq 0, \quad \forall m, n, \quad (13)$$

$$\sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \|\mathbf{w}_m\|^2 \leq P_{\max}, \quad (14)$$

where  $P_{\max}$  denotes the maximum transmit power at the BS. With the objective function and the constraints presented above, the max-min fairness problem can be expressed as follows

$$\begin{aligned} \mathcal{P}_1 : \quad & \max_{\{p_{m,n}, \mathbf{w}_m\}} \min_{m,n} \{R_n^m\}, \\ \text{s.t. } \quad & C_1 : p_{m,n} \geq 0, \quad \forall m, n, \\ & C_2 : \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \|\mathbf{w}_m\|^2 \leq P_{\max}. \end{aligned} \quad (15)$$

The problem  $\mathcal{P}_1$  in (15) is challenging, since the expression for the objective function is quite complicated. As shown in (10), the achievable rate  $R_n^m$  for each user is the minimum among a series of achievable rates. Moreover, as indicated by (6) and (8), both the inter-beam and intra-beam interferences exist in the system, which makes the optimization variables appear in both the nominators and denominators of the SINRs for users, as shown in (7) and (8). Therefore, the problem  $\mathcal{P}_1$  is non-convex. Additionally, the users in the same beam

are assigned the same precoding vector but allocated different power, and the power allocation parameters  $\{p_{m,n}\}$  as well as the beam-specific digital precoding vectors  $\{\mathbf{w}_m\}_{m=1}^{N_{\text{RF}}}$  entangle with each other in the form of  $p_{m,n} |\hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m|^2$ , which makes it difficult to simultaneously optimize them.

To overcome the difficulties mentioned above, we partition the optimization variables into the power allocation block  $\{p_{m,n}\}$  and the digital precoding block  $\{\mathbf{w}_m\}$ , and  $\{\mathbf{w}_m\}$  will be lifted to  $\{\mathbf{W}_m\}$  by semidefinite relaxation (SDR). Then, an alternating optimization framework is proposed to optimize the  $\{p_{m,n}\}$  and the  $\{\mathbf{W}_m\}$  in an iterative way, and eigenvalue decomposition will be adopted to obtain the precoding vectors from the lifted matrices after the iteration. In the following subsections, we will firstly introduce the optimization of the beam-specific digital precoding block as well as the power allocation block in Subsection III-B and III-C, respectively, and then present the complete optimization scheme in detail in Subsection III-D. Finally, we will provide more insights concerning the optimality of the proposed algorithm in Subsection III-E.

#### B. Beam-Specific Digital Precoding Optimization

In this subsection, we optimize the beam-specific digital precoding vectors  $\{\mathbf{w}_m\}$  for any given power allocation parameters  $\{p_{m,n}\}$  by lifting the precoding vectors into positive-semidefinite matrix variables  $\{\mathbf{W}_m\}$  using semidefinite relaxation (SDR). The corresponding optimization problem  $\mathcal{P}_{\text{beam}}$  is formulated as follows, which is reduced from the original problem  $\mathcal{P}_1$  in (15):

$$\begin{aligned} \mathcal{P}_{\text{beam}} : \quad & \max_{\{\mathbf{w}_m\}} \min_{m,n} \{R_n^m\}, \\ \text{s.t. } \quad & C_1 : \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \|\mathbf{w}_m\|^2 \leq P_{\max}, \end{aligned} \quad (16)$$

and the optimal value for the objective function of the problem  $\mathcal{P}_{\text{beam}}$  is denoted as  $r_{\text{beam}}^*$ . Then, we will propose a bisection-based SDR method to solve this problem.

In order to cope with the complicated objective function in (16), we will introduce an auxiliary variable  $r$  to simplify the objective function, and the problem  $\mathcal{P}_{\text{beam}}$  can be equivalently transformed to  $\mathcal{P}'_{\text{beam}}$  as follows:

$$\begin{aligned} \mathcal{P}'_{\text{beam}} : \quad & \max_{\{\mathbf{w}_m\}, r} r, \\ \text{s.t. } \quad & C_1 : \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \|\mathbf{w}_m\|^2 \leq P_{\max}, \\ & C_2 : R_{n,i}^m \geq r, \quad \forall m, n, i \leq n \leq |S_m|, \end{aligned} \quad (17)$$

where  $C_2$  denotes the rate constraints to ensure the successful SIC for NOMA users in a beam [12]. As mentioned before, for the successful use of NOMA in a beam, the  $n$ -th user's signal should be decoded by the  $i$ -th user in the same beam as long as  $i \leq n$ . Thus, the total number of constraints in  $C_2$  can be calculated as  $\sum_{m=1}^{N_{\text{RF}}} |S_m|(|S_m| + 1)/2$ , which is approximately in the order of  $K^2$ .

Note that in this new problem  $\mathcal{P}'_{\text{beam}}$ , the objective function and the constraint  $C_1$  are all convex, and the challenge mainly lies in the constraint  $C_2$ . Specifically, the constraint  $C_2$  in (17) can be expressed as

$$\begin{aligned} R_{n,i}^m &\geq r \\ \Leftrightarrow &\left( \sum_{k=1}^{n-1} p_{m,k} - \frac{p_{m,n}}{2^r - 1} \right) \left| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right|^2 \\ &+ \sum_{j \neq m} \sum_{k=1}^{|S_j|} p_{j,k} \left| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_j \right|^2 + \sigma^2 \leq 0. \end{aligned} \quad (18)$$

This constraint is challenging due to the following two reasons. Firstly, variable  $r$  appears in the denominator. Secondly, the parameter  $\left( \sum_{k=1}^{n-1} p_{m,k} - \frac{p_{m,n}}{2^r - 1} \right)$  of the norm term  $\left| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right|^2$  is negative to ensure feasibility. Therefore, this constraint is non-convex, and thus hard to deal with.

To deal with the variable  $r$  appearing in the denominator, we construct a bisection framework to firstly set  $r$  to some  $r_0$  in advance, and then solve the corresponding feasibility problem to examine whether the max-min rate can achieve  $r_0$ . In this way, the constant  $r_0$  will appear at the denominator of the corresponding constraint, and will not affect its convexity. To be more specific, we utilize the total consumed power as the indicator for feasibility, and minimize this power while constraining that all users' rates are larger than  $r_0$ . If the minimized power consumption is smaller than the maximum transmission power  $P_{\text{max}}$ , then  $r_0$  can be achieved by the max-min rate, otherwise  $r_0$  cannot be achieved. The formulated problem  $\hat{\mathcal{P}}_{\text{beam}}$  for the examination of  $r_0$  can be expressed as follows:

$$\begin{aligned} \hat{\mathcal{P}}_{\text{beam}} : &\min_{\{\mathbf{w}_m\}} \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \|\mathbf{w}_m\|^2, \\ \text{s.t. } C_1 : &R_{n,i}^m \geq r_0, \forall m, n, i \leq n \leq |S_m|. \end{aligned} \quad (19)$$

Once we can determine  $r_0$  is achievable or not by solving a series of problems  $\hat{\mathcal{P}}_{\text{beam}}$ , a bisection procedure can be carried out to acquire  $r_{\text{beam}}^*$ . However, the problem  $\hat{\mathcal{P}}_{\text{beam}}$  in (19) is still hard to solve, which is because the new constraint  $C_1$  in (19) is still non-convex:

$$\begin{aligned} R_{n,i}^m &\geq r_0 \\ \Leftrightarrow &\left( \sum_{k=1}^{n-1} p_{m,k} - \frac{p_{m,n}}{2^{r_0} - 1} \right) \left| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right|^2 \\ &+ \sum_{j \neq m} \sum_{k=1}^{|S_j|} p_{j,k} \left| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_j \right|^2 + \sigma^2 \leq 0, \end{aligned} \quad (20)$$

since the parameter  $\left( \sum_{k=1}^{n-1} p_{m,k} - \frac{p_{m,n}}{2^{r_0} - 1} \right)$  is negative. To address this issue, the SDR method can be used to rewrite the mathematical terms concerning digital precoding vectors  $\{\mathbf{w}_m\}$

and beamspace channels  $\{\hat{\mathbf{h}}_{m,i}\}$  in the form of matrix trace:

$$\|\mathbf{w}_m\|_2^2 = \text{Tr}(\mathbf{w}_m \mathbf{w}_m^H) = \text{Tr}(\mathbf{W}_m), \quad (21)$$

$$\begin{aligned} \left| \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \right|^2 &= \text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m \mathbf{w}_m^H) \\ &= \text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_m), \end{aligned} \quad (22)$$

where  $\mathbf{w}_m$  is lifted to  $\mathbf{W}_m = \mathbf{w}_m \mathbf{w}_m^H$ . Then, the problem  $\hat{\mathcal{P}}_{\text{beam}}$  in (19) can be equivalently rewritten as

$$\begin{aligned} \tilde{\mathcal{P}}_{\text{beam}} : &\min_{\{\mathbf{W}_m\}} \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \text{Tr}(\mathbf{W}_m), \\ \text{s.t. } C_1 : &f_{n,i}^m \leq 0, \forall m, n, i \leq n \leq |S_m|, \\ C_2 : &\mathbf{W}_m \in \mathcal{H}_{N_{\text{RF}}}^+, \forall m, \\ C_3 : &\text{rank}(\mathbf{W}_m) = 1, \forall m, \end{aligned} \quad (23)$$

where  $\mathcal{H}_{N_{\text{RF}}}^+$  denotes the set for all Hermitian positive semidefinite matrices of size  $N_{\text{RF}} \times N_{\text{RF}}$ , and

$$\begin{aligned} f_{n,i}^m &\leq 0 \\ \Leftrightarrow &\left( \sum_{k=1}^{n-1} p_{m,k} - \frac{p_{m,n}}{2^{r_0} - 1} \right) \text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_m) \\ &+ \sum_{j \neq m} \sum_{k=1}^{|S_j|} p_{j,k} \text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_j) + \sigma^2 \leq 0. \end{aligned} \quad (24)$$

We can find from (24) that  $C_1$  now becomes a convex constraint. Since  $\mathbf{W}_m = \mathbf{w}_m \mathbf{w}_m^H$ ,  $\mathbf{W}_m$  must be a Hermitian positive semidefinite matrix with rank one, which accounts for the constraints  $C_2$  and  $C_3$  in (23).

After the manipulations mentioned above, all the objective function and constraints are convex, except for the rank constraint  $C_3$  in (23). To be more specific, the problem  $\tilde{\mathcal{P}}_{\text{beam}}$  in (23) is NP-hard [23], and thus approximation/relaxation methods must be introduced to make it solvable in polynomial time. Therefore, the SDR method relaxes the rank constraint and brings the following new problem  $\tilde{\mathcal{P}}_{\text{beam}}^r$ :

$$\begin{aligned} \tilde{\mathcal{P}}_{\text{beam}}^r : &\min_{\{\mathbf{W}_m\}} \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \text{Tr}(\mathbf{W}_m), \\ \text{s.t. } C_1 : &f_{n,i}^m \leq 0, \forall m, n, i \leq n \leq |S_m|, \\ C_2 : &\mathbf{W}_m \in \mathcal{H}_{N_{\text{RF}}}^+, \forall m, \end{aligned} \quad (25)$$

which has become convex and is ready to be solved by classical optimization methods. Let  $P^*$  and  $\{\mathbf{W}_m^*\}$  denote the optimal value and the solution for problem  $\tilde{\mathcal{P}}_{\text{beam}}^r$  in (25), we have  $r_{\text{beam}}^* \geq r_0$  when  $P^* \leq P_{\text{max}}$ , and  $r_{\text{beam}}^* < r_0$  when  $P^* > P_{\text{max}}$ . Thus, a bisection-based SDR method is proposed to obtain  $r_{\text{beam}}^*$  and  $\{\mathbf{W}_m^*\}$ , respectively, which is presented in Algorithm 1. Based on the obtained matrices  $\{\mathbf{W}_m^*\}$ , the power allocation parameters will be further optimized in each iteration of the alternating optimization framework, which will

---

**Algorithm 1: Bisection-Based SDR.**


---

**Input:** Lower bound  $r_L$ , upper bound  $r_H$ , beamspace channels  $\{\hat{\mathbf{h}}_{m,n}\}$ , power allocation parameters  $\{p_{m,n}\}$ , noise power  $\sigma^2$ , total power  $P_{\max}$ , desirable accuracy  $\epsilon$ .

**Output:** Lifted matrices  $\{\mathbf{W}_m^*\}$ , and the max-min rate  $r_{\text{beam}}^*$ .

- 1: **while**  $r_H - r_L > \epsilon$  **do**
  - 2:   Set  $r_0 = (r_H + r_L)/2$ , solve the optimization problem  $\hat{\mathcal{P}}_{\text{beam}}^r$  to obtain the matrices  $\{\mathbf{W}_m\}$ .
  - 3:   **if**  $\sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \text{Tr}(\mathbf{W}_m) \leq P_{\max}$  **then**
  - 4:     Set  $\{\mathbf{W}_m^*\} = \{\mathbf{W}_m\}$ ,  $r_L = r_0$ ,  $r_{\text{beam}}^* = r_0$ .
  - 5:   **else**
  - 6:     Set  $r_H = r_0$ .
  - 7:   **end if**
  - 8: **end while**
- 

be introduced in the next subsection. Note that we optimize the lifted matrices  $\{\mathbf{W}_m\}$ , not the precoding vectors  $\{\mathbf{w}_m\}$ . Therefore, after the alternating optimization framework, an eigenvalue decomposition-based rank one approximation method will be utilized to calculate the corresponding digital precoding vectors from the lifted matrix variables [23], [24].

Let  $M$  denotes the total number of the linear inequality constraints for the problem  $\hat{\mathcal{P}}_{\text{beam}}^r$ , which is approximately in the order of  $K^2$ . As a result, the worst case computational complexity of solving problem  $\hat{\mathcal{P}}_{\text{beam}}^r$  can be calculated as  $\mathcal{O}(N_{\text{RF}}^{3.5} N_{\text{RF}}^{6.5} + M N_{\text{RF}}^{1.5} N_{\text{RF}}^{2.5}) \sim \mathcal{O}(N_{\text{RF}}^{10} + K^2 N_{\text{RF}}^4)$  [23], [24]. Thus, the worst case complexity of the proposed algorithm is  $\mathcal{O}(\log(1/\epsilon)(N_{\text{RF}}^{10} + K^2 N_{\text{RF}}^4))$ , where  $\epsilon$  is the desired accuracy.

### C. Power Allocation Optimization

After solving the problem  $\mathcal{P}_{\text{beam}}$  in Subsection III-B, we now optimize the power allocation parameters  $\{p_{m,n}\}$  for any given lifted matrices  $\{\mathbf{W}_m\}$  by a bisection-based linear programming method. Utilizing the matrix trace forms presented in (21) and (22), the corresponding power allocation problem  $\mathcal{P}_{\text{power}}$  is the reduced problem from the original problem  $\mathcal{P}_1$  in (15), which is presented as follows:

$$\begin{aligned} \mathcal{P}_{\text{power}} : & \max_{\{p_{m,n}\}} \min_{m,n} \{R_n^m\}, \\ \text{s.t. } & C_1 : p_{m,n} \geq 0, \forall m, n, \\ & C_2 : \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \text{Tr}(\mathbf{W}_m) \leq P_{\max}. \end{aligned} \quad (26)$$

The maximal value for the objective function of the above optimization problem  $\mathcal{P}_{\text{power}}$  is denoted as  $r_{\text{power}}^*$ .

Similar to the solution of the optimization problem  $\mathcal{P}_{\text{beam}}$  in Subsection III-B above, we will introduce an auxiliary variable  $r$  to simplify the object function in (26), set its value to a certain  $r_0$ , and construct a new optimization problem  $\hat{\mathcal{P}}_{\text{power}}$  as follows

to examine whether  $r_0$  is achievable:

$$\begin{aligned} \hat{\mathcal{P}}_{\text{power}} : & \min_{\{p_{m,n}\}} \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \text{Tr}(\mathbf{W}_m), \\ \text{s.t. } & C_1 : p_{m,n} \geq 0, \forall m, n, \\ & C_2 : R_{n,i}^m \geq r_0, \forall m, n, i \leq n \leq |S_m|. \end{aligned} \quad (27)$$

Once we can determine  $r_0$  is achievable or not by solving a series of  $\hat{\mathcal{P}}_{\text{power}}$ , a bisection procedure similar to Algorithm 1 in the previous Subsection III-B can be carried out to acquire  $r_{\text{power}}^*$  and  $\{p_{m,n}^*\}$ , respectively. Note that for the above problem  $\hat{\mathcal{P}}_{\text{power}}$ , the objective function and the constraint  $C_1$  are already linear, and the constraint  $C_2$  denotes the rate constraints to ensure the successful SIC for NOMA users in a beam [12], which can be reformulated in the following form using the matrix trace

$$\begin{aligned} & R_{n,i}^m \geq r_0 \\ \Leftrightarrow & \log_2 \left( 1 + \frac{|\hat{\mathbf{h}}_{m,i}^H \mathbf{w}_m|^2 p_{m,n}}{\xi_{n,i}^m} \right) \geq r_0 \\ \Leftrightarrow & \text{Tr} \left( \hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_m \right) \left( \sum_{k=1}^{n-1} p_{m,k} - \frac{p_{m,n}}{2^{r_0} - 1} \right) \\ & + \sum_{j \neq m} \text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_j) \sum_{k=1}^{|S_j|} p_{j,k} + \sigma^2 \leq 0. \end{aligned} \quad (28)$$

As shown in (28), the two terms  $\text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_m)$  and  $1/(2^{r_0} - 1)$  are constants, which makes the constraints in  $C_2$  linear. Therefore, the examination problem  $\hat{\mathcal{P}}_{\text{power}}$  can be readily solved by classical linear programming methods. Denote the minimal value for the objective function of the problem  $\hat{\mathcal{P}}_{\text{power}}$  as  $P^*$ , which is the minimized power consumption given that all the users' rates are larger than  $r_0$ . As mentioned above, If  $P^* \leq P_{\max}$ , we have  $r_{\text{power}}^* \geq r_0$ , and  $r_{\text{power}}^* < r_0$  when  $P^* > P_{\max}$ . In this way, the optimal value  $r_{\text{power}}^*$  for the optimization problem  $\mathcal{P}_{\text{power}}$  in (26) within a desirable accuracy  $\epsilon$  as well as the corresponding power allocation parameters  $\{p_{m,n}^*\}$  can be obtained by solving a series of linear programming problems, which is presented in Algorithm 2.

The worst case complexity for solving the problem  $\hat{\mathcal{P}}_{\text{power}}$  using interior point method can be calculated as  $\mathcal{O}(K^{3.5} + M K^{1.5}) \sim \mathcal{O}(K^{3.5})$  [23], [24]. Thus, the worst case complexity for Algorithm 2 is  $\mathcal{O}(\log(1/\epsilon) K^{3.5})$ .

### D. Alternating Optimization Framework

Based on the two bisection-based methods introduced in the above Subsection III-B and III-C, we now present the complete alternating optimization algorithm for the optimization of the power allocation and the beam-specific digital precoding, as shown in Algorithm 3. As mentioned above, the optimization variables are partitioned into two blocks, i.e.,  $\{p_{m,n}\}$  and  $\{\mathbf{w}_m\}$ , and  $\{\mathbf{w}_m\}$  are lifted to  $\{\mathbf{W}_m\}$  by semidefinite relaxation. Then,  $\{p_{m,n}\}$  and  $\{\mathbf{W}_m\}$  are alternatively optimized by keeping the



**Algorithm 2:** Bisection-Based Linear Programming.

**Input:** Lower bound  $r_L$ , upper bound  $r_H$ , beamspace channels  $\{\hat{\mathbf{h}}_{m,n}\}$ , lifted matrices  $\{\mathbf{W}_m\}$ , noise power  $\sigma^2$ , total power  $P_{\max}$ , desirable accuracy  $\epsilon$ .

**Output:** max-min rate  $r_{\text{power}}^*$ , corresponding power allocation parameters  $\{p_{m,n}^*\}$ .

- 1: **while**  $r_H - r_L > \epsilon$  **do**
- 2:   Set  $r_0 = (r_H + r_L)/2$ , solve problem  $\hat{\mathcal{P}}_{\text{power}}$  to obtain the power allocation parameters  $\{p_{m,n}\}$ .
- 3:   **if**  $\sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} p_{m,n} \text{Tr}(\mathbf{W}_m) \leq P_{\max}$  **then**
- 4:     Set  $r_L = r_0$ ,  $r_{\text{power}}^* = r_0$ ,  $\{p_{m,n}^*\} = \{p_{m,n}\}$ .
- 5:   **else**
- 6:     Set  $r_H = r_0$ .
- 7:   **end if**
- 8: **end while**

**Algorithm 3:** Proposed Alternating Optimization Algorithm With Beam-Specific Digital Precoding.

**Input:** Beamspace channels  $\hat{\mathbf{h}}_{m,n}$ , noise power  $\sigma^2$ , total power  $P_{\max}$ , desirable accuracy  $\epsilon$ .

**Output:** max-min rate  $r^*$ , power allocation parameters  $\{p_{m,n}^*\}$ , beam-specific digital precoding vectors  $\{\mathbf{w}_m^*\}$ .

- 1: **Init.**  $r_L = 0$ ,  $r_H = \log_2(1 + P_{\max} h_{\min}/\sigma^2)$ , where  $h_{\min}$  is the minimal value among all  $\|\hat{\mathbf{h}}_{m,n}\|^2$ . Calculate  $\{\mathbf{w}_m^{(0)}\}$  by SVD-based zero forcing [11], [14], and obtain the corresponding lifted matrices  $\{\mathbf{W}_m^{(0)} = \mathbf{w}_m^{(0)} (\mathbf{w}_m^{(0)})^H\}$ . Set  $r_{\text{beam}}^{(0)} = 0$ , set  $t = 0$ .
- 2: **while true do**
- 3:   Solve problem  $\mathcal{P}_{\text{power}}$  for given  $\{\mathbf{W}_m^{(t)}\}$  using Algorithm 2. Denote the max-min rate and the power allocation parameters as  $r_{\text{power}}^{(t+1)}$  and  $\{p_{m,n}^{(t+1)}\}$ , respectively.
- 4:   Solve problem  $\mathcal{P}_{\text{beam}}$  for given  $\{p_{m,n}^{(t+1)}\}$  using Algorithm 1. Denote the max-min rate and the corresponding lifted matrices as  $r_{\text{beam}}^{(t+1)}$  and  $\{\mathbf{W}_m^{(t+1)}\}$ , respectively.
- 5:   **if**  $r_{\text{beam}}^{(t+1)} - r_{\text{beam}}^{(t)} \leq \epsilon$  **then**
- 6:     **Break.**
- 7:   **end if**
- 8:   Set  $r_L = r_{\text{beam}}^{(t+1)}$ .
- 9:    $t = t + 1$ .
- 10: **end while**
- 11: Set  $\{\mathbf{W}_m^*\} = \{\mathbf{W}_m^{(t)}\}$ .
- 12: **for**  $1 \leq m \leq N_{\text{RF}}$  **do**
- 13:   Get the eigenvalue decomposition of  $\mathbf{W}_m^*$ , and set  $\mathbf{w}_m^*$  to be the normalized eigenvector with the largest eigenvalue.
- 14: **end for**
- 15: Solve problem  $\mathcal{P}_{\text{power}}$  for given  $\{\mathbf{w}_m^* (\mathbf{w}_m^*)^H\}$  using Algorithm 2. Denote the max-min rate and the power allocation parameters as  $r^*$  and  $\{p_{m,n}^*\}$ , respectively.

other one fixed, until the incremental increase of the max-min rate is less than a threshold  $\epsilon$ . Particularly, in an arbitrary  $(t+1)$ -th iteration, since  $\{p_{m,n}^{(t+1)}\}$  and  $\{\mathbf{W}_m^{(t+1)}\}$  are all optimized solutions, we will have  $r_{\text{beam}}^{(t+1)} \geq r_{\text{power}}^{(t+1)} \geq r_{\text{beam}}^{(t)}$ . In this sense, we can obtain a monotonically non-decreasing sequence of the max-min rates  $\{r_{\text{beam}}^{(t)}\}$ . Due to the limited transmit power, the max-min rate is upper-bounded, and thus the alternative procedure will converge. Note that for each iteration, we optimize the lifted matrices  $\{\mathbf{W}_m\}$ , not the precoding vectors  $\{\mathbf{w}_m\}$ , and thus the semidefinite relaxation will not affect the convergence.

Moreover, since we optimize the lifted matrices in each iteration, we can only obtain optimized matrix variables  $\{\mathbf{W}_m^*\}$  by the alternative procedure. Therefore, eigenvalue decomposition will be carried out to  $\{\mathbf{W}_m^*\}$  after the alternative procedure to calculate the corresponding precoding vectors [23], [24]. Particularly, for those  $\{\mathbf{W}_j^*\}$  whose ranks are larger than one, we can obtain the eigenvalue decomposition of them, i.e.,  $\mathbf{W}_j^* = \sum_n \lambda_{j,n} \mathbf{q}_{j,n} \mathbf{q}_{j,n}^H$ , where  $\lambda_{j,1} \geq \lambda_{j,2} \geq \dots \geq 0$ . Then, the corresponding digital precoding vector  $\mathbf{w}_j^*$  is set as the normalized eigenvector with the largest eigenvalue, i.e.,  $\mathbf{q}_{j,1}$ . For those  $\{\mathbf{W}_i^*\}$  satisfying  $\text{rank}(\mathbf{W}_i^*) = 1$ , we have  $\mathbf{W}_i^* = \lambda_i \mathbf{q}_i \mathbf{q}_i^H$ , and the corresponding digital precoding vector  $\mathbf{w}_i^*$  is just  $\mathbf{q}_i$ . Based on the digital precoding vectors  $\{\mathbf{w}_m^*\}$  calculated above, the power allocation will be optimized for the last time to obtain the final max-min rate  $r^*$  and the corresponding power allocation parameters  $\{p_{m,n}^*\}$ . Note that by setting the  $\mathbf{w}_m^*$  to be the normalized eigenvector, it seems that we ignored the corresponding maximum eigenvalue. However, since the achievable rate is determined by the products  $\{\sqrt{p_{m,n}} \mathbf{w}_m\}$ , the seeming loss of ignoring the maximum eigenvalue can also be compensated by the power allocation optimization after the eigenvalue decomposition.

In addition, for the initialization of the bisection procedure, the lower bound  $r_L$  is set as 0. Since both the inter-beam and intra-beam interferences exist in the system, the minimal rate must be smaller than that of allocating the maximum power to the weakest user, and thus the upper bound  $r_H$  is set as  $\log_2(1 + P_{\max} h_{\min}/\sigma^2)$ . The computational complexity of the alternating optimization algorithm mainly comes from the iteration part. Assume the iteration times is equal to  $T$ , based on the worst case complexity analysis in the previous subsections, the overall complexity in the worst case can be calculated as  $\mathcal{O}(T \log(1/\epsilon)(K^{3.5} + N_{\text{RF}}^{10} + K^2 N_{\text{RF}}^4))$ . Defining the overloading factor  $\eta$  as  $\eta = K/N_{\text{RF}} > 1$  [8], the overall worst case complexity can also be expressed as  $\mathcal{O}(T \log(1/\epsilon)(\eta^{3.5} N_{\text{RF}}^{3.5} + N_{\text{RF}}^{10} + \eta^2 N_{\text{RF}}^6))$ .

**E. Optimality Analysis**

Based on the above discussions, although the convergence of the proposed Algorithm 3 can be guaranteed, its optimality cannot be guaranteed. Due to the relaxation of the rank-one constraints, the converged solution of the alternative procedure may not be feasible, i.e., the lifted matrices  $\{\mathbf{W}_m^*\}$  may not satisfy rank-one constraints. Therefore, the final solution after



carrying out eigenvalue decomposition to the  $\{\mathbf{W}_m^*\}$  are not guaranteed to be optimal.

To elaborate a little further, we will provide more insights on why the relaxed rank-one constraints are not guaranteed to be satisfied. For the system considered in our paper, to ensure successful SIC, the  $n$ -th user's signal also needs to be decoded by the  $i$ -th user in the same beam for  $i < n$ . Therefore, it is similar to the multicast communication scenario, where each signal needs to be decoded by a group of the users [25]. As a result, the strategies of the unicast beamforming problem cannot be directly applied to it. Besides, it is widely accepted that the lifted matrix precoders of the multicast beamforming problem returned by SDR are not guaranteed to be of unit rank [25]. Since both the intra-beam and inter-beam interferences exist in the considered beamspace MIMO-NOMA system, our problem is even more complicated than the multicast beamforming problem. Due to the reasons mentioned above, the relaxed rank-one constraints are not guaranteed to be satisfied.

#### IV. TWO-STAGE OPTIMIZATION OF POWER ALLOCATION AND USER-SPECIFIC DIGITAL PRECODING

In this section, we break the commonly adopted beam-specific digital precoding scheme [11], [13]–[15] and further improve the max-min rate by user-specific digital precoding, i.e., each user is assigned with a unique digital precoding vector. Correspondingly, we will formulate the max-min fairness problem in this case, and propose a two-stage optimization method to firstly design the precoding vectors and then finetune the power allocation. Finally, the optimality of the proposed algorithm will be analyzed.

##### A. Problem Formulation

Unlike the beam-specific digital precoding, the  $n$ -th user in the  $m$ -th beam will be assigned a unique digital precoding vector  $\mathbf{w}_{m,n}$  for the user-specific digital precoding, where  $\|\mathbf{w}_{m,n}\|^2$  denotes the power allocated to this user. Other than that, the beam selection scheme and the SIC decoding order are the same as that in beam-specific precoding scheme. Thus, this SIC decoding order is also not optimal, and its optimal design can be left for future work.

As a result, the expression for the achievable rate  $\hat{R}_n^m$  of the  $n$ -th user in the  $m$ -th beam is similar to (10) in Section II:

$$\hat{R}_n^m = \min_{i=1,2,\dots,n} \hat{R}_{n,i}^m = \min_{i=1,2,\dots,n} \log_2(1 + \hat{\gamma}_{n,i}^m), \quad (29)$$

where  $\hat{R}_{n,i}^m$  and  $\hat{\gamma}_{n,i}^m$  denote the achievable rate and SINR of the  $n$ -th user's signal decoded by the  $i$ -th user in the  $m$ -th beam, respectively. Similar to (7), the SINRs for the user-specific digital precoding can be expressed as:

$$\hat{\gamma}_{n,i}^m = \frac{|\hat{\mathbf{h}}_{m,i}^H \mathbf{w}_{m,n}|^2}{\hat{\xi}_{n,i}^m}, \quad (30)$$

where

$$\hat{\xi}_{n,i}^m = \sum_{k=1}^{n-1} |\hat{\mathbf{h}}_{m,i}^H \mathbf{w}_{m,k}|^2 + \sum_{\substack{j \neq m \\ k=1}}^{|S_j|} |\hat{\mathbf{h}}_{m,i}^H \mathbf{w}_{j,k}|^2 + \sigma^2. \quad (31)$$

Then, the minimal rate maximization problem  $\mathcal{P}_2$  can be formulated as:

$$\begin{aligned} \mathcal{P}_2 : & \max_{\{\mathbf{w}_{m,n}\}} \min_{m,n} \{\hat{R}_n^m\}, \\ \text{s.t. } & C_1 : \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} \|\mathbf{w}_{m,n}\|^2 \leq P_{\text{max}}. \end{aligned} \quad (32)$$

Since each NOMA user is assigned a unique precoding vector, the power allocated to each user can be directly expressed by the squared norm of the corresponding precoding vector. As a result, by optimizing the digital precoding vectors in the above problem  $\mathcal{P}_2$ , the power allocation parameters are simultaneously optimized. Therefore, the proposed optimization framework for this problem only consists of two stages, and it is not iterative, which will be introduced in the next subsection.

##### B. Two-Stage Optimization Framework

The proposed optimization method consists of two stages, which is shown in Algorithm 4. Firstly, the digital precoding vectors  $\{\mathbf{w}_{m,n}\}$  are optimized by a bisection-based SDR method. Then, the power allocation parameters for every user are finetuned by a bisection-based linear programming method. Since the optimization framework only consists of two stages, convergence is not an issue here.

In the first stage, we solve the problem  $\mathcal{P}_2$  in (32) via optimization of the user-specific digital precoding vectors. Specifically, we formulate the following optimization problem  $\hat{\mathcal{P}}_{\text{user}}$  to examine whether a certain  $r_0$  is achievable:

$$\begin{aligned} \hat{\mathcal{P}}_{\text{user}} : & \min_{\{\mathbf{w}_{m,n}\}} \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} \|\mathbf{w}_{m,n}\|^2, \\ \text{s.t. } & C_1 : \hat{R}_{n,i}^m \geq r_0, \forall m, n, i \leq n \leq |S_m|, \end{aligned} \quad (33)$$

and the SDR method is adopted to solve the above problem for the examination of  $r_0$ . In this way, a bisection procedure similar to that in the previous Section III can be proposed to obtain the max-min rate  $r^*$  and the corresponding digital precoding vectors  $\{\mathbf{w}_{m,n}^*\}$ . Once again, the mathematical terms concerning digital precoding vectors will be equivalently rewritten using matrix trace. With  $\mathbf{w}_{m,n} \mathbf{w}_{m,n}^H$  denoted by  $\mathbf{W}_{m,n}$ , the following new problem  $\tilde{\mathcal{P}}_{\text{user}}^r$  after relaxation of the rank-one constraints is formulated as:

$$\begin{aligned} \tilde{\mathcal{P}}_{\text{user}}^r : & \min_{\{\mathbf{W}_{m,n}\}} \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} \text{Tr}(\mathbf{W}_{m,n}), \\ \text{s.t. } & C_1 : g_{n,i}^m \leq 0, \forall m, n, i \leq n \leq |S_m|, \\ & C_2 : \mathbf{W}_{m,n} \in \mathcal{H}_{N_{\text{RF}}}^+, \forall m, n, \end{aligned} \quad (34)$$

---

**Algorithm 4:** Proposed Two-Stage Optimization Algorithm with User Specific Digital Precoding.

---

**Input:** Beamspace channels  $\hat{\mathbf{h}}_{m,n}$ , noise power  $\sigma^2$ , total power  $P_{\max}$ , desirable accuracy  $\epsilon$ .

**Output:** max-min rate  $r^*$ , user-specific digital precoding vectors  $\{\mathbf{w}_{m,n}^*\}$ .

1: **Init.**  $r_L = 0$ ,  $r_H = \log_2(1 + P_{\max}h_{\min}/\sigma^2)$ , where  $h_{\min}$  is the minimal value among all  $\|\hat{\mathbf{h}}_{m,n}\|^2$ .

**Stage one:**

2: **while**  $r_H - r_L > \epsilon$  **do**

3: Set  $r_0 = (r_H + r_L)/2$ , solve the optimization problem  $\hat{\mathcal{P}}_{\text{user}}^r$  to obtain the matrices  $\{\mathbf{W}_{m,n}\}$ .

4: **if**  $\sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} \text{Tr}(\mathbf{W}_{m,n}) \leq P_{\max}$  **then**

5: Set  $\{\mathbf{W}_{m,n}^*\} = \{\mathbf{W}_{m,n}\}$ ,  $r_L = r_0$ ,  $r^* = r_0$ .

6: **else**

7: Set  $r_H = r_0$ .

8: **end if**

9: **end while**

10: Set  $\mathbf{w}_{m,n}^*$  as the normalized eigenvector of  $\mathbf{W}_{m,n}^*$  with maximum eigenvalue.

11: Set  $r_H = \log_2(1 + P_{\max}h_{\min}/\sigma^2)$ .

**Stage two:**

12: **while**  $r_H - r_L > \epsilon$  **do**

13: Set  $r_0 = (r_H + r_L)/2$ , solve the optimization problem  $\hat{\mathcal{P}}_{\text{scale}}$  to obtain the scaling factors  $\{\beta_{m,n}\}$ .

14: **if**  $\sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} \beta_{m,n} \|\mathbf{w}_{m,n}^*\|^2 \leq P_{\max}$  **then**

15: Set  $\{\beta_{m,n}^*\} = \{\beta_{m,n}\}$ ,  $r_L = r_0$ ,  $r^* = r_0$ .

16: **else**

17: Set  $r_H = r_0$ .

18: **end if**

19: **end while**

20: Set  $\mathbf{w}_{m,n}^* = \sqrt{\beta_{m,n}^*} \mathbf{w}_{m,n}^*$ .

---

where

$$\begin{aligned} & g_{n,i}^m \leq 0 \\ \Leftrightarrow & \sum_{k=1}^{n-1} \text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_{m,k}) - \frac{\text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_{m,n})}{2^{r_0} - 1} \\ & + \sum_{j \neq m} \sum_{k=1}^{|S_j|} \text{Tr}(\hat{\mathbf{h}}_{m,i} \hat{\mathbf{h}}_{m,i}^H \mathbf{W}_{j,k}) + \sigma^2 \leq 0. \end{aligned} \quad (35)$$

Thus, the problem  $\hat{\mathcal{P}}_{\text{user}}^r$  becomes convex, and is ready to be solved for examination of  $r_0$ . Denote the optimal value of the objective function and the solution of the problem  $\hat{\mathcal{P}}_{\text{user}}^r$  as  $P^*$  and  $\{\mathbf{W}_{m,n}^*\}$ , we have  $r^* \geq r_0$  when  $P^* \leq P_{\max}$ , and  $r^* < r_0$  when  $P^* > P_{\max}$ , based on which the proposed bisection-based SDR method is shown in stage one of Algorithm 4. Because of the relaxation of the rank-one constraints, the ranks of the obtained matrices  $\mathbf{W}_{m,n}^*$  may not be one. Therefore, eigenvalue decomposition is carried out for  $\mathbf{W}_{m,n}^*$ , and  $\mathbf{w}_{m,n}^*$  is set as the normalized eigenvector with the maximum eigenvalue. In case that the rank-one constraints are not satisfied, the  $\{\mathbf{w}_{m,n}^*\}$

obtained in the first stage may not be optimal, and the scaling of the vectors  $\{\mathbf{w}_{m,n}^*\}$  in the second stage is introduced to further improve the performance [23], [24].

To be more specific, based on the obtained power allocation  $\{\|\mathbf{w}_{m,n}^*\|^2\}$  in the first stage, the second stage finetunes it by optimizing the scaling factors  $\{\beta_{m,n}\}$  via bisection-based linear programming method, and the final solution will be  $\{\sqrt{\beta_{m,n}} \mathbf{w}_{m,n}^*\}$ . Different from that of the beam-specific precoding scheme, since the power allocation can be readily expressed by the squared norm of the corresponding precoding vector, the variables  $\{\beta_{m,n}\}$  are not power allocation parameters. They only serve as scaling factors to be attached to the corresponding precoding vectors in the form of  $\{\sqrt{\beta_{m,n}} \mathbf{w}_{m,n}^*\}$ . Since they only alter the norms of the precoding vectors without changing their directions, they are used to finetune the power allocation, i.e., the power allocation will change from  $\{\|\mathbf{w}_{m,n}^*\|^2\}$  to  $\{\beta_{m,n} \|\mathbf{w}_{m,n}^*\|^2\}$  after the second stage.

With the  $\{\mathbf{w}_{m,n}^*\}$  given in the first stage, the following problem is formulated to optimize the scaling factors in the second stage:

$$\mathcal{P}_{\text{scale}} : \max_{\{\beta_{m,n}\}} \min_{m,n} \{\hat{R}_n^m\},$$

$$\text{s.t. } C_1 : \beta_{m,n} \geq 0, \forall m, n,$$

$$C_2 : \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} \beta_{m,n} \|\mathbf{w}_{m,n}^*\|^2 \leq P_{\max}. \quad (36)$$

To be more specific, we will bound a certain  $r_0$  by the bisection procedure, and formulate the following problem to examine whether  $r_0$  can be achieved:

$$\hat{\mathcal{P}}_{\text{scale}} : \min_{\{\beta_{m,n}\}} \sum_{m=1}^{N_{\text{RF}}} \sum_{n=1}^{|S_m|} \beta_{m,n} \|\mathbf{w}_{m,n}^*\|^2,$$

$$\text{s.t. } C_1 : \beta_{m,n} \geq 0, \forall m, n,$$

$$C_2 : \hat{R}_{n,i}^m \geq r_0, \forall m, n, i \leq n \leq |S_m|. \quad (37)$$

The above problem  $\hat{\mathcal{P}}_{\text{scale}}$  is also a linear programming problem, which is very similar to the problem  $\hat{\mathcal{P}}_{\text{power}}$  in (27), except that the optimization variables have changed from  $\{p_{m,n}\}$  to  $\{\beta_{m,n}\}$ . Thus, the optimal solution  $\{\beta_{m,n}^*\}$  can also be calculated by a bisection-based linear programming method, as shown in stage two of Algorithm 4. Finally, the proposed user-specific digital precoding vectors are equal to  $\{\sqrt{\beta_{m,n}^*} \mathbf{w}_{m,n}^*\}$ .

In addition, the computational complexity of the two-stage optimization algorithm mainly comes from the optimization of the digital precoding vectors, the finetuning of the power allocation, and the eigenvalue decomposition. For the optimization of the digital precoding vectors, the worst case complexity can be calculated as  $\mathcal{O}(\log(1/\epsilon)(K^{3.5} N_{\text{RF}}^{6.5} + MK^{1.5} N_{\text{RF}}^{2.5})) \sim \mathcal{O}(\log(1/\epsilon)(K^{3.5} N_{\text{RF}}^{6.5}))$ . For the second stage, the worst case complexity is  $\mathcal{O}(\log(1/\epsilon) K^{3.5})$  [23], [24]. Since the complexity for the eigenvalue decomposition is no more than  $\mathcal{O}(KN_{\text{RF}}^3)$ , the total complexity for the two-stage optimization algorithm is  $\mathcal{O}(\log(1/\epsilon)(K^{3.5} N_{\text{RF}}^{6.5})) \sim \mathcal{O}(\log(1/\epsilon)(\eta^{3.5} N_{\text{RF}}^{10}))$  in the worst case, where  $\eta$  is the overloading factor defined to

describe the capability of NOMA systems serving multiple users.

Note that the complexity of the alternating algorithm for the beam-specific precoding scheme is  $\mathcal{O}(T \log(1/\epsilon)(\eta^{3.5} N_{\text{RF}}^{3.5} + N_{\text{RF}}^{10} + \eta^2 N_{\text{RF}}^6))$ , where the highest order for  $\eta$  and  $N_{\text{RF}}$  are  $\eta^{3.5}$  and  $N_{\text{RF}}^{10}$ , and they are in two separate terms. However, the complexity of the two-stage algorithm for the user-specific precoding scheme is  $\mathcal{O}(\log(1/\epsilon)(\eta^{3.5} N_{\text{RF}}^{10}))$ , where the two highest-order terms multiply with each other in the form of  $\eta^{3.5} N_{\text{RF}}^{10}$ . Due to the multiplying of the two highest-order terms, the complexity of the user-specific precoding scheme is higher than that of the beam-specific precoding scheme, especially when the number of the RF chains  $N_{\text{RF}}$  or the overloading factor  $\eta$  is large. In this sense, the beam-specific precoding scheme would be beneficial in terms of the complexity, when the overloading factor or the number of RF chains are relatively large.

### C. Optimality Analysis

It is worth pointing out that the optimality can not be guaranteed, and the main reason is the relaxation of the rank-one constraints in the first stage. To be more specific, since the rank-one constraints are not guaranteed to be satisfied by the optimized lifted matrices  $\{\mathbf{W}_{m,n}^*\}$  in the first stage, the  $\{\mathbf{w}_{m,n}^*\}$  obtained by eigenvalue decomposition are not guaranteed to be optimal. In the second stage, the finetune of the power allocation is based on the precoding vectors obtained in the first stage, and thus cannot fully compensate the loss introduced by the first stage. Therefore, it cannot be guaranteed to reach the optimal point, either. Moreover, the reason why the rank-one constraints are not guaranteed to be satisfied is the same as that in Subsection III-E.

## V. SIMULATION RESULTS

Simulations are carried out to verify the effectiveness of the proposed two optimization methods for beamspace MIMO-NOMA, where the maximum magnitude-based beam selection scheme is utilized. In this section, we focus on the minimal rate performance of five schemes, which are beamspace MIMO-NOMA with user-specific digital precoding (denoted as “user-specific beamspace MIMO-NOMA”), beamspace MIMO-NOMA with beam-specific digital precoding (denoted as “beam-specific beamspace MIMO-NOMA”), beamspace MIMO-NOMA with SVD-based zero-forcing (denoted as “SVD-based beamspace MIMO-NOMA”), beamspace MIMO-NOMA with strongest user-based zero forcing (denoted as “strongest user-based beamspace MIMO-NOMA”), and TDMA, respectively. For the first four beamspace MIMO-NOMA schemes, the differences only lie in the design of the power allocation parameters and digital precoding vectors. Particularly, the first two beamspace MIMO-NOMA schemes utilize the user-specific and beam-specific optimization methods proposed in this paper, respectively. The third and the fourth beamspace MIMO-NOMA schemes adopt the traditional digital precoding methods, i.e., SVD-based zero-forcing [11] and strongest user-based zero forcing [13], respectively, and they optimize the power allocation parameters via the

bisection-based linear programming method proposed in Section III. TDMA is also considered here as a benchmark for comparison, where the classical zero forcing precoding is adopted to eliminate interferences, and power allocation is optimized in each time slot [14]. In addition, to investigate the sum-rate performance of the two proposed algorithms, we also utilize the minorization-maximization algorithm proposed in [26] to maximize the sum rate of the beamspace MIMO-NOMA system under user-specific precoding scheme, which is denoted as “sum rate benchmark for beamspace MIMO-NOMA”.

Based on the four beamspace MIMO-NOMA schemes introduced above, the BS could select the deployed schemes according to the communication scenario for practical implementation. Since the third and the fourth schemes only optimize the power allocation parameters, their complexity will be lower than the first two schemes, and their performance in terms of the max-min rate will be inferior to them, which will be demonstrated by the following simulation results. Therefore, if the channel coherence time is long enough, then the BS can carry out the first two schemes to optimize both the power allocation parameters and the digital precoding vectors for enhanced performance. Otherwise, if the channel changes rapidly, then the simplified third and fourth schemes of only designing the power allocation parameters can be adopted.

As for the simulation settings, all the schemes share the same simulation parameters. To be more specific, the BS is deployed with 64 antennas and 8 RF chains. Note that 64 is a typical number for the large-scale antenna arrays adopted in a series of papers concerning mmWave communications [15], [17], [27]. Moreover, the strongest user-based zero forcing scheme is used as a benchmark, and the corresponding reference [13] also sets the number of antenna elements to 64. In accordance with the relevant works, and also for the fairness of comparison, we also set the number of transmit antennas to 64 in our paper. Besides, the channel parameters for an arbitrary user  $k$  satisfy: 1)  $\Omega_k^{(0)} \sim \mathcal{CN}(0, 10)$ ,  $\Omega_k^{(l)} \sim \mathcal{CN}(0, 1)$ , for  $l = \{1, \dots, L\}$ ; 2)  $\psi_k^{(0)}$  and  $\psi_k^{(l)}$  follow the uniform distribution within  $[-\frac{1}{2}, \frac{1}{2}]$  [11]. The desirable accuracy  $\epsilon$  is set as  $10^{-4}$ , while the SNR is defined as  $\log_{10}(P_{\text{max}}/\sigma^2)$ . The simulation results for those schemes under LoS as well as NLoS propagation environments are presented in this section, which are acquired using CVX and averaged among 1000 random realizations.

### A. Simulations With LoS Path

In this subsection, we assume that there are one LoS component and  $L = 5$  NLoS components. Fig. 2 presents the minimal rate comparison between the considered five schemes with respect to the SNR, where the number of served users is set to 16. Compared with the traditional design methods, where digital precoding vectors are calculated by the modified zero forcing [11], [13], Fig. 2 shows that the proposed two optimization methods can largely improve the minimal rate. For instance, when the SNR is 15 dB, the first and second beamspace MIMO-NOMA scheme yield a performance gain of 15% and 7% compared to the third scheme, respectively. The performance gain comes from optimizing both the power allocation and the

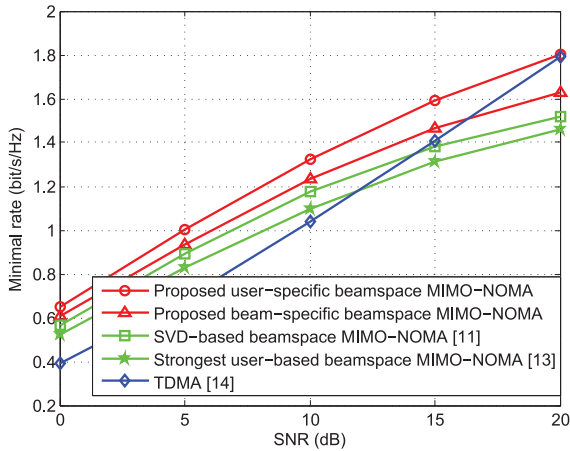


Fig. 2. Minimal rate comparison with respect to the SNR with LoS path.

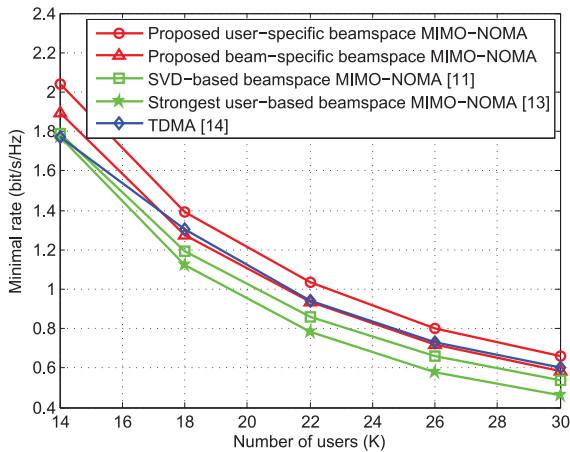


Fig. 3. Minimal rate comparison with respect to the number of served users with LoS path.

digital precoding vectors. Moreover, since the user-specific digital precoding provides more degrees of freedom for optimization than the beam-specific digital precoding, the first beamspace MIMO-NOMA scheme outperforms the second one, but suffers from higher worst case complexity as discussed in Section IV-B.

Besides, with the increased SNR, the performance of all beamspace MIMO-NOMA schemes will firstly be better and then worse than TDMA in terms of the minimal rate, which is clear from the intersection points of the performance curves in Fig. 2. This is because beamspace MIMO-NOMA system is rank deficient, where the number of RF chains is limited compared to that of the supported users. As a result, the inter-beam interference cannot be completely eliminated, while TDMA is free of it. Therefore, with the increase of the SNR, the inter-beam interference will become more and more severe, which leads to the intersection points mentioned above. Fortunately, utilizing the proposed optimization methods, the interference can be well suppressed, and the intersection points are shifted to the right, which indicates that the region where beamspace MIMO-NOMA outperforms beamspace MIMO with TDMA can be enlarged by the proposed methods.

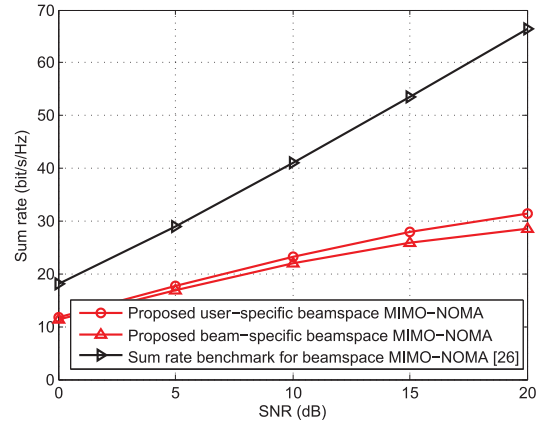


Fig. 4. Sum rate comparison with respect to the SNR with LoS path.

Meanwhile, Fig. 3 shows the minimal rate comparison of the five considered schemes with respect to the number of users, where the SNR is set as 15 dB. We can see from Fig. 3 that the minimal rate can be largely improved by the proposed two optimization methods. Moreover, the minimal rates of all schemes decrease with the number of users due to the total power constraint. Formerly, the third and fourth beamspace MIMO-NOMA schemes with traditional design methods are inferior to that of the TDMA scheme in the whole considered region. However, with the help of the carefully designed optimization methods in this paper, the minimal rate of the first beamspace MIMO-NOMA scheme can always be larger than that of the TDMA, while the minimal rate of the second beamspace MIMO-NOMA scheme is almost the same as that of the TDMA, which also demonstrates the effectiveness of our methods in Sections III and IV.

In addition, the sum rate performance of the two proposed algorithms with respect to the SNR are also investigated, which are compared with that of the benchmark algorithm proposed in [26], as shown by Fig. 4. It is clear from this figure that the sum rate of the two proposed algorithms are inferior to that of the minorization-maximization algorithm in [26], which are around 52% and 48% of the sum rate benchmark with the SNR set to 15 dB, respectively. This is mainly because the objective functions of the algorithms are different, i.e., we maximize the minimal rate while [26] maximizes the sum rate. Moreover, when carrying out sum-rate maximization using the algorithm in [26], the minimal rate of the beamspace MIMO-NOMA system is quite small. In fact, in order to maximize the sum rate, the BS tends to allocate most resource to the strong users, and thus the rate performance of the weak users will be deteriorated. Therefore, there is a trade-off between the minimal rate and the sum rate in the system, and the algorithm in [26] sacrifices the former for the latter. In contrast to the algorithm in [26], in order to guarantee the rate performance for all the NOMA users, we maximize the minimal rate while sacrificing the sum rate.

### B. Simulations Without LoS Path

We consider the NLoS transmission in this subsection, where only five NLoS paths are assumed, i.e.,  $L = 5$ , and the LoS path



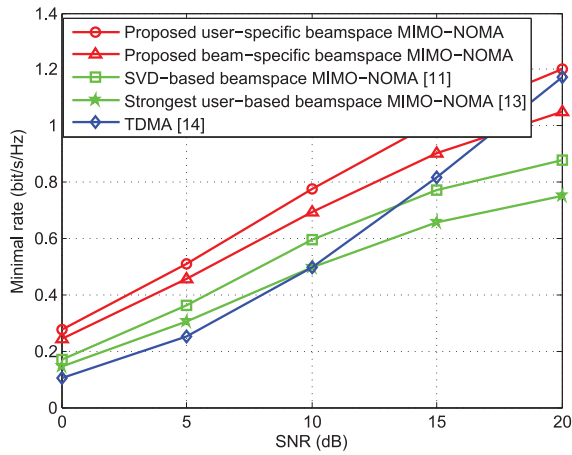


Fig. 5. Minimal rate comparison with respect to the SNR without LoS path.

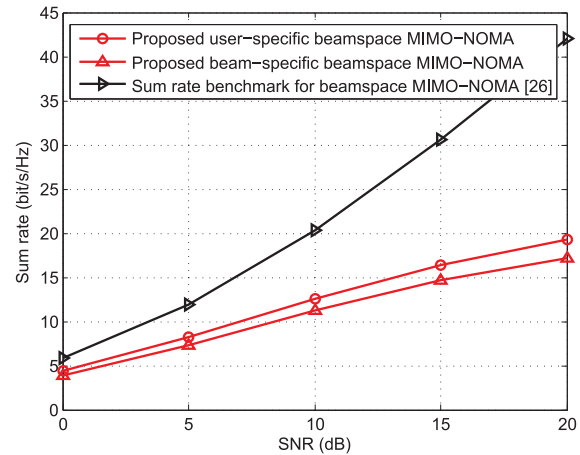


Fig. 7. Sum rate comparison with respect to the SNR without LoS path.

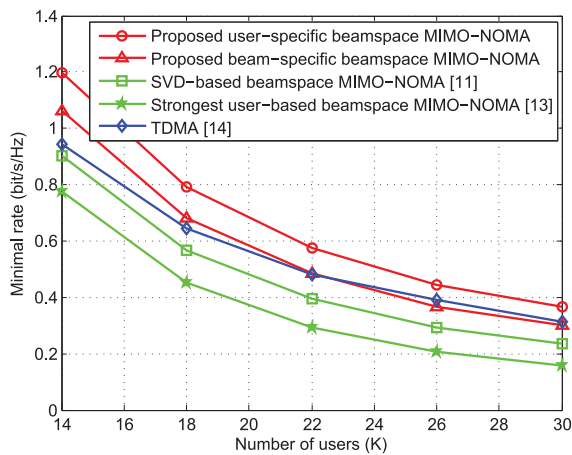


Fig. 6. Minimal rate comparison with respect to the number of served users without LoS path.

is absent. The minimal rate comparison of the five schemes with respect to the SNR is presented in Fig. 5, where the number of served users is set to 16. From this figure, we can see that compared with the traditional design method, the two proposed optimization methods can provide an even larger performance gain than that in Subsection V-A. Specifically, when the SNR is 15 dB, the first and second beamspace MIMO-NOMA schemes yield a performance gain of 31% and 16% compared to the third beamspace MIMO-NOMA scheme, respectively. Since the LoS path is absent, the beamspace channels in the same beam are not highly correlated [11]. As a result, the interferences turn out to be more severe than that when the LoS path exists, and thus the optimization methods can provide larger performance gain. Besides, the first beamspace MIMO-NOMA scheme is now superior to the TDMA scheme in the whole considered region because of the careful design of the proposed methods.

Meanwhile, Fig. 6 presents the minimal rate comparison of the five considered schemes with respect to the number of users under NLoS scenario, where the SNR is set as 15 dB. This figure shows that compared with the traditional design methods, the two proposed optimization methods can largely improve the

minimal rate of the beamspace MIMO-NOMA system. To be more specific, the third and fourth beamspace MIMO-NOMA schemes with modified zero forcing are inferior to that of the TDMA scheme in terms of the minimal rate. However, with the help of the carefully designed optimization methods, the minimal rate of the first beamspace MIMO-NOMA scheme can always be larger than that of the TDMA scheme, while the minimal rate of the second scheme is almost the same as that of the TDMA scheme, which again verifies the performance of our proposed methods.

In addition, the sum rate performance of the two proposed algorithms with respect to the SNR are also investigated when the LoS path does not exist, as shown by Fig. 7. This figure also shows that the sum rate of the two proposed algorithms are inferior to that of the minorization-maximization algorithm in [26], which are around 53% and 49% of the sum rate benchmark with the SNR set to 15 dB, respectively. The reason is the same as that in Subsection V-A: we maximize the minimal rate while [26] maximizes the sum rate, and the algorithm in [26] sacrifices the former for the latter.

## VI. CONCLUSION

In order to guarantee the achievable rate for every supported user, we have studied the multi-beam beamspace MIMO-NOMA system by formulating and solving the minimal rate maximization problem. Depending on the users in the same beam sharing the same digital precoding vector or not, two optimization solutions are proposed. Particularly, for beam-specific digital precoding where the same precoding vector is shared by all supported users in a beam, an alternating optimization method is proposed to design the power allocation parameters and the digital precoding vectors. Furthermore, we explore the user-specific digital precoding scheme, where different users in the same beam can use different digital precoding vectors, and a two-stage optimization method is proposed to firstly optimize the precoding vectors and then finetune the power allocation. Simulation results have verified the performance of our proposed methods. Moreover, we reveal that the two-stage optimization method for user-specific digital precoding outperforms the

alternating optimization method for the beam-specific digital precoding at the cost of higher worst case complexity, since the former provides more degrees of freedom for the optimization of digital precoding vectors in the beamspace MIMO-NOMA system.

## REFERENCES

- [1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
- [2] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.
- [3] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Select. Areas Commun.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.
- [4] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *Proc. IEEE Global Commun. Conf.*, Atlanta, USA, Dec. 2013, pp. 3679–3684.
- [5] P. Amadori and C. Masouros, "Low RF-complexity millimeter-wave beamspace-MIMO systems by beam selection," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2212–2222, Jun. 2015.
- [6] 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.
- [7] L. Zhu, Z. Xiao, X. Xia, and D. Oliver Wu, "Millimeter-wave communications with non-orthogonal multiple access for B5G/6G," *IEEE Access*, vol. 7, pp. 116123–116132, Aug. 2019.
- [8] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.
- [9] 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.
- [10] Y. Xu *et al.*, "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4874–4886, Sep. 2017.
- [11] 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. Select. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.
- [12] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.
- [13] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.
- [14] 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.
- [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] 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.
- [17] 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.
- [18] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Beamforming techniques for nonorthogonal multiple access in 5G cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9474–9487, Oct. 2018.
- [19] 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.
- [20] 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.
- [21] M. F. Hanif and Z. Ding, "Robust power allocation in MIMO-NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1541–1545, Dec. 2019.
- [22] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-complexity technologies to enable millimeter-wave MIMO with large antenna array for 5G wireless communications," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 211–217, Apr. 2018.
- [23] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.
- [24] E. Karipidis, N. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.
- [25] A. B. Gershman, N. D. Sidiropoulos, S. Shahbazpanahi, M. Bengtsson, and B. Ottersten, "Convex optimization-based beamforming," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 62–75, May 2010.
- [26] O. Lyons, M. F. Hanif, M. Juntti, and L. Tran, "Fast adaptive minorization-maximization procedure for beamforming design of downlink NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 8023–8027, Jul. 2020.
- [27] X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.



**Ruicheng Jiao** (Student Member, IEEE) received the B.S. degree in physics from Tsinghua University, Beijing, China, in 2016. He is currently working toward the Ph.D. degree with the Department of Electronic Engineering, Tsinghua University, Beijing, China. 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** (Senior Member, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2003, the M.S. degree (with the highest Hons.) from the China Academy of Telecommunications Technology, Beijing, China, in 2006, and the Ph.D. degree (with the highest Hons.) from Tsinghua University, Beijing, China, in 2011. From 2011 to 2013, he was a Postdoctoral 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. His current research interests include massive MIMO, millimeter-wave/THz communications, NOMA, reconfigurable intelligent surface (RIS), and machine learning for wireless communications. He has coauthored the book *MmWave Massive MIMO: A Paradigm for 5G* (Academic Press, 2016). He has authored or coauthored more than 60 IEEE journal papers and more than 40 IEEE conference papers. He also holds 16 granted patents. He has 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 has 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, and the IEEE ComSoc Leonard G. Abraham Prize in 2020. He is an Area Editor of IEEE COMMUNICATIONS LETTERS, and an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. Particularly, he is dedicated to reproducible research and has made a large amount of simulation code publicly available.