Spectrum and Energy-Efficient Beamspace MIMO-NOMA for Millimeter-Wave Communications Using Lens Antenna Array

Bichai Wang, *Student Member, IEEE*, Linglong Dai, *Senior Member, IEEE*, Zhaocheng Wang, *Senior Member, IEEE*, Ning Ge, *Member, IEEE*, and Shidong Zhou, *Member, IEEE*

***Abstract*— The recent concept of beamspace multiple input multiple output (MIMO) can significantly reduce the number of required radio frequency (RF) chains in millimeter-wave (mmWave) massive MIMO systems without obvious performance loss. However, the fundamental limit of existing beamspace MIMO is that the number of supported users cannot be larger than the number of RF chains at the same time-frequency resources. To break this fundamental limit, in this paper, we propose a new spectrum and energy-efficient mmWave transmission scheme that integrates the concept of non-orthogonal multiple access (NOMA) with beamspace MIMO, i.e., beamspace MIMO-NOMA. By using NOMA in beamspace MIMO systems, the number of supported users can be larger than the number of RF chains at the same time-frequency resources. In particular, the achievable sum rate of the proposed beamspace MIMO-NOMA in a typical mmWave channel model is analyzed, which shows an obvious performance gain compared with the existing beamspace MIMO. Then, a precoding scheme based on the principle of zero forcing is designed to reduce the inter-beam interferences in the beamspace MIMO-NOMA system. Furthermore, to maximize the achievable sum rate, a dynamic power allocation is proposed by solving the joint power optimization problem, which not only includes the intra-beam power optimization, but also considers the inter-beam power optimization. Finally, an iterative optimization algorithm with low complexity is developed to realize the dynamic power allocation. Simulation results show that the proposed beamspace MIMO-NOMA can achieve higher spectrum and energy efficiency compared with the existing beamspace MIMO.**

***Index Terms*— Millimeter-wave, beamspace MIMO, NOMA, sum rate, precoding, power allocation.**

Manuscript received January 27, 2017; revised May 15, 2017; accepted May 23, 2017. Date of publication July 11, 2017; date of current version September 15, 2017. This work was supported in part by the National Key Basic Research Program of China under Grant 2013CB329203, in part by the National Natural Science Foundation of China under Grant 61571267 and Grant 61571270, and in part by the Royal Academy of Engineering through the U.K.–China Industry Academia Partnership Programme Scheme under Grant UK-CIAPP\49. *(Corresponding author: Linglong Dai.)*

The authors are with the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wbc15@mails.tsinghua.edu.cn; daill@tsinghua.edu.cn; zcwang@tsinghua.edu.cn; gening@tsinghua.edu.cn; zhousd@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2017.2725878

I. Introduction

With the rapid development of the Mobile Internet and the Internet of Things (IoT), challenging requirements for the 5th generation (5G) of wireless communication systems are expected to be satisfied, which are fuelled by the prediction that the global mobile data traffic will grow in the range of 10-100 times from 2020 to 2030. The emerging millimeter-wave (mmWave) communications, operating from 30-300 GHz, provide an opportunity to meet such explosive capacity demand for 5G [1]. In addition to orders-of-magnitude lager bandwidths, the smaller wavelengths at mmWave allow more antennas in a same physical space, which enables massive multiple input multiple output (MIMO) to provide more multiplexing gain and beamforming gain [2]–[8]. In fact, it has been demonstrated that mmWave massive MIMO can achieve orders-of-magnitude increase in system capacity [6].

However, it is difficult to realize mmWave massive MIMO in practice due to high transceiver complexity and energy consumption [5], [9]. Particularly, each antenna in MIMO systems usually requires one dedicated radio-frequency (RF) chain [10]. Therefore, the use of a very large number of antennas in mmWave massive MIMO systems leads to an equally large number of RF chains. Moreover, it is shown that RF components may consume up to 70% of the total transceiver energy consumption [5], [11]. As a result, the hardware cost and energy consumption caused by a large number of RF chains in mmWave massive MIMO systems become unaffordable in practice.

To address this challenging problem, a lot of studies have been done to reduce the hardware complexity and energy consumption. Particularly, the antenna selection technique has been considered to solve this problem [12]–[14]. However, an obvious performance loss will be introduced. Recently, the concept of beamspace MIMO has been proposed in the pioneering work [2] to significantly reduce the number of required RF chains in mmWave massive MIMO systems. By using the lens antenna array, which plays a role in realizing spatial discrete Fourier transformation [15], beamspace MIMO can transform the conventional spatial channel to the

beamspace channel to capture the channel sparsity at mmWave frequencies [9]. Accordingly, the dominant beams are selected according to the sparse beamspace channel to reduce the number of required RF chains. Moreover, by the use of lens antenna array, narrow beams can be preserved even with a reduced number of RF chains, which allows to significantly reduce the power required per beam and the inter-beam interferences [5]. Therefore, unlike the antenna selection technique, the performance of beamspace MIMO with beam selection is close-to-optimal [4], [5]. Nevertheless, a fundamental limit of beamspace MIMO that was explicitly or implicitly considered in all published papers on beamspace MIMO [2], [4], [5], [9], [15] is that, each RF chain can only support one user at the same time-frequency resources, so the maximum number of users that can be supported cannot exceed the number of RF chains. The reason is that, the degree of freedom (DoF) provided by the RF chains must be larger than or equal to the DoF required by users, otherwise signal for different users cannot be separated by linear operation.

In this paper, we aim to break this fundamental limit by proposing a spectrum and energy efficient mmWave transmission scheme that integrates the new concept of non-orthogonal multiple access (NOMA) with beamspace MIMO, i.e., beamspace MIMO-NOMA.[1] Particularly, NOMA has also been considered as a promising candidate for 5G to improve spectrum efficiency and connectivity density [16]–[23]. In contrast to the orthogonal multiple access (OMA) schemes relying on the time-, frequency-, code-domain or on their combinations, NOMA can be realized in a new domain, i.e., the power domain. By performing superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, multiple users can be simultaneously supported at the same time-frequency-space resources, and the channel gain difference among users can be translated into multiplexing gain by superposition coding. By integrating NOMA into beamspace MIMO, potential performance gain can be achieved. Specifically, The contributions of this paper can be summarized as follows.

1) We propose a new spectrum and energy efficient mmWave transmission scheme, i.e., beamspace MIMO-NOMA, that combines the advantages of NOMA and beamspace MIMO. To the best of our knowledge, this is the first work using NOMA as a potential multiple access scheme for beamspace MIMO in mmWave communications. Particularly, by using intra-beam superposition coding and SIC, more than one user can be simultaneously supported in one beam, which is essentially different from existing beamspace MIMO using one beam to only serve one user. Thus, the number of supported users can be larger than the number of RF chains at the same time-frequency resources in the proposed beamspace MIMO-NOMA scheme, and the achievable sum rate in a typical mmWave channel model can be also significantly improved. Note that although the combination of spatial MIMO and NOMA has been widely investigated [24]–[31], it only focused on the conventional MIMO systems rather than the mmWave massive MIMO systems. Therefore, the existing MIMO-NOMA schemes [24]–[31] have not considered the transmission characteristics in mmWave communications, e.g., the channel sparsity, as well as the uncertainty of the number of conflicting users in beamspace MIMO systems.
2) To reduce the inter-beam interferences in the proposed beamspace MIMO-NOMA system, the equivalent channel vector is determined for each beam to realize precoding based on the principle of zero-forcing (ZF). On the one hand, when the line-of-sight (LoS) component of users' channels is dominant, the high correlation [32] of users' beamspace channels in the same beam at mmWave frequencies is utilized to generate the equivalent channel vectors. Note that potential performance gain can be achieved by exploiting the high channel correlation in NOMA [22]. On the other hand, when the LoS component does not exist or the effect of non-line-of-sight (NLoS) components is significant, the channel correlation in the same beam may be not high enough. To this end, we also consider the singular value decomposition (SVD)-based equivalent channel, which exploits all of the beamspace channel vectors of users in the same beam.
3) In the proposed beamspace MIMO-NOMA scheme, in addition to intra-beam interferences caused by superposition coding, users also suffer from interferences from other beams. Thus, a direct combination of NOMA and beamspace MIMO is not able to guarantee the reliable performance in practice. In most existing MIMO-NOMA schemes, the fixed inter-beam power allocation is usually exploited without optimization, and intra-beam power optimization is considered only for two users. On the contrary, in the proposed beamspace MIMO-NOMA scheme, a dynamic power allocation scheme is realized to maximize the achievable sum rate with the transmitted power constraint by solving the joint power optimization problem, which not only includes the intra-beam power optimization, but also considers the inter-beam power optimization. Furthermore, an iterative optimization algorithm is developed to realize the dynamic power allocation, and the convergence as well as computational complexity of this algorithm are also analyzed.
4) We verify the performance of the proposed beamspace MIMO-NOMA scheme by simulations. The convergence of the developed iterative optimization algorithm for dynamic power allocation is validated, and it is shown that only 10 times of iteration are required to make it converged. Furthermore, we show that the proposed beamspace MIMO-NOMA can achieve higher spectrum and energy efficiency than that of existing beamspace MIMO systems, e.g., 25% energy efficiency gain can be achieved.

The rest of this paper is organized as follows. The system model of the proposed beamspace MIMO-NOMA system is

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

introduced in Section II. Section III analyzes the achievable sum rate of the proposed beamspace MIMO-NOMA system, and Section IV introduces the precoding scheme based on the principle of ZF. In Section V, a dynamic power allocation scheme is proposed to maximize the achievable sum rate. Simulation results are provided in Section VI. Finally, conclusions are drawn in Section VII.

Notation: We use upper-case and lower-case boldface letters to denote matrices and vectors, respectively; $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$, $(\cdot)^\dagger$, tr$(\cdot)$, and $\|\cdot\|_p$ denote the transpose, conjugate transpose, matrix inversion, Moore-Penrose matrix inversion, the trace of the matrix, and l_p norm operation, respectively. diag$\{\mathbf{p}\}$ denotes the diagonal matrix whose diagonal elements consist of the elements in the vector $\mathbf{p}$. E$\{\cdot\}$ denotes the expectation. $|\Gamma|$ denotes the number of elements in set Γ. $\mathbf{A}(i,:)_{i\in\Gamma}$ denotes the submatrix of $\mathbf{A}$ that consists of the ith row of $\mathbf{A}$ for all $i \in \Gamma$. We use the notation $\mathcal{CN}(\mathbf{m},\mathbf{R})$ to denote the complex Gaussian distribution with mean $\mathbf{m}$ and covariance $\mathbf{R}$. Finally, $\mathbf{I}_N$ is the $N \times N$ identity matrix.

II. System Model

In this paper, we consider a single-cell downlink mmWave communication system, where the base station (BS) is equipped with N antennas and N_{RF} RF chains, and K single-antenna users are simultaneously served by the BS [4], [5]. The system model of existing beamspace MIMO will be introduced at first in this section, and then the proposed beamspace MIMO-NOMA scheme will be presented in detail.

A. Beamspace MIMO

In traditional MIMO systems as shown in Fig. 1 (a), the received signal vector $\mathbf{y} = [y_1, y_2, \cdots, y_K]^T$ can be expressed as

$$\mathbf{y} = \mathbf{H}^H\mathbf{W}\mathbf{P}\mathbf{s} + \mathbf{v}, \tag{1}$$

where $\mathbf{s} = [s_1, s_2, \cdots, s_K]^T$ is the $K \times 1$ transmitted signal vector for all K users with normalized power $\mathrm{E}\left(\mathbf{s}\mathbf{s}^H\right) = \mathbf{I}_K$, $\mathbf{P} = \mathrm{diag}\{\mathbf{p}\}$ includes the transmitted power for all K users where $\mathbf{p} = \left[\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_K}\right]$ satisfies $\sum_{k=1}^{K} p_k \le P$ (the maximum transmitted power at the BS), $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K]$ is the $N \times K$ precoding matrix with $\|\mathbf{w}_k\|_2 = 1$ for $k = 1, 2, \cdots, K$, and $\mathbf{v}$ is the noise vector following the distribution $\mathcal{CN}\left(0, \sigma^2\mathbf{I}_K\right)$. Finally, $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_K]$ of size $N \times K$ is the channel matrix, where $\mathbf{h}_k$ of size $N \times 1$ denotes the spatial channel vector between the BS and the kth user. Particularly, in this paper, we consider the widely used Saleh-Valenzuela channel model for mmWave communications [2], [4], [5], [9], [15], so $\mathbf{h}_k$ can be represented as

$$\mathbf{h}_k = \beta_k^{(0)}\mathbf{a}\left(\theta_k^{(0)}\right) + \sum_{l=1}^{L}\beta_k^{(l)}\mathbf{a}\left(\theta_k^{(l)}\right), \tag{2}$$

where $\beta_k^{(0)}\mathbf{a}(\theta_k^{(0)})$ is the LoS component of the kth user, in which $\beta_k^{(0)}$ denotes the complex gain and $\mathbf{a}(\theta_k^{(0)})$ represents the spatial direction. $\beta_k^{(l)}\mathbf{a}(\theta_k^{(l)})$ for $1 \le l \le L$ is the lth NLoS component of the kth user, where L is the total number of NLoS components. $\mathbf{a}(\theta)$ is the $N \times 1$ array steering vector. Note that at mmWave frequencies, the amplitudes $\{|\beta_k^{(l)}|\}_{l=1}^{N}$ of NLoS components are typically 5 to 10 dB weaker than the amplitude $|\beta_k^{(0)}|$ of the LoS component [4], [33].

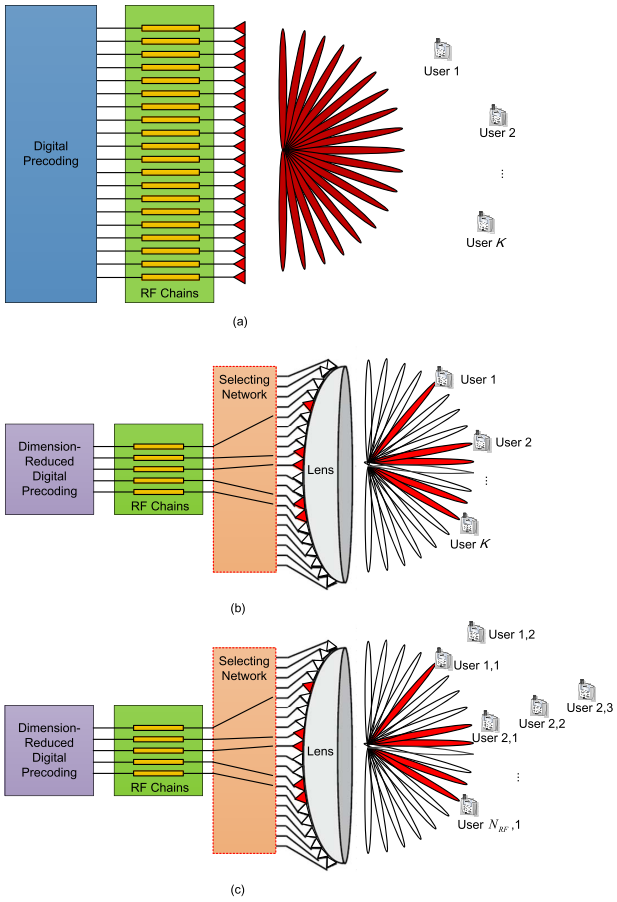

Fig. 1. System models of MIMO architectures: (a) traditional MIMO; (b) beamspace MIMO; (c) the proposed beamspace MIMO-NOMA.

For the typical uniform linear array (ULA) [5], the array steering vector $\mathbf{a}(\theta)$ can be expressed as

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N}}\left[e^{-j2\pi\theta m}\right]_{m\in J(N)}, \tag{3}$$

where $J(N) = \{i - (N-1)/2, i = 0, 1, \cdots, N-1\}$ is a symmetric set of indices centered around zero. The spatial direction is defined as $\theta = \frac{d}{\lambda}\sin(\phi)$, where ϕ is the physical direction satisfying $-\frac{\pi}{2} \le \phi \le \frac{\pi}{2}$, λ is the signal wavelength, and d is the antenna spacing.

As shown in Fig. 1 (a), in traditional MIMO systems, the number of required RF chains is equal to the number of BS antennas, i.e., $N_{\mathrm{RF}} = N$, which is usually large for mmWave massive MIMO systems, e.g., $N_{\mathrm{RF}} = N = 256$ [9], [34]. Therefore, the direct application of massive MIMO at mmWave frequencies is prohibitive due to high hardware cost and energy consumption caused by RF chains [5], [11], e.g., about 250 mW is consumed by each RF chain, and 64 W is required by a mmWave massive MIMO system with 256 antennas [6].

To address this issue, the concept of beamspace MIMO has been recently proposed, which can utilize lens antenna

array to significantly reduce the number of required RF chains in mmWave massive MIMO systems without obvious performance loss. As shown in Fig. 1. (b), by employing lens antenna array, the channel (2) in the spatial domain can be transformed to the beamspace channel [15]. Specifically, the mathematical function of the lens antenna array is to realize the spatial discrete Fourier transformation with the $N \times N$ transform matrix $\mathbf{U}$ [9], which contains the array steering vectors of N directions covering the entire space as follows:

$$\mathbf{U} = \left[\mathbf{a}\left(\bar{\theta}_1\right), \mathbf{a}\left(\bar{\theta}_2\right), \cdots, \mathbf{a}\left(\bar{\theta}_N\right)\right]^H, \tag{4}$$

where $\bar{\theta}_n = \frac{1}{N}\left(n - \frac{N+1}{2}\right)$ for $n = 1, 2, \cdots, N$ are the predefined spatial directions. Then, the received signal vector $\bar{\mathbf{y}}$ in beamspace MIMO systems can be represented as

$$\bar{\mathbf{y}} = \mathbf{H}^H \mathbf{U}^H \mathbf{W}\mathbf{P}\mathbf{s} + \mathbf{v} = \bar{\mathbf{H}}^H \mathbf{W}\mathbf{P}\mathbf{s} + \mathbf{v}, \tag{5}$$

where the beamspace channel matrix $\bar{\mathbf{H}}$ is defined as

$$\bar{\mathbf{H}} = \mathbf{U}\mathbf{H} = [\mathbf{U}\mathbf{h}_1, \mathbf{U}\mathbf{h}_2, \cdots, \mathbf{U}\mathbf{h}_K] = \left[\bar{\mathbf{h}}_1, \bar{\mathbf{h}}_2, \cdots, \bar{\mathbf{h}}_K\right], \tag{6}$$

where $\bar{\mathbf{h}}_k = \mathbf{U}\mathbf{h}_k$ is the beamspace channel vector between the BS and the kth user, which is the Fourier transformation of the spatial channel vector $\mathbf{h}_k$ in (2).

As for the beamspace channel matrix $\bar{\mathbf{H}}$ defined in (6), each row of $\bar{\mathbf{H}}$ corresponds to one beam, and all N rows correspond to N beams with spatial directions $\bar{\theta}_1, \bar{\theta}_2, \cdots, \bar{\theta}_N$, separately. In mmWave communications, since the number of dominant scatters is very limited, the number of NLoS components L is much smaller than the number of beams N [9]. Therefore, the number of dominant elements of each beamspace channel vector $\bar{\mathbf{h}}_k$ is much smaller than N, namely, the beamspace channel matrix $\bar{\mathbf{H}}$ has a sparse nature [15]. This sparse structure can be exploited to design dimension-reduced beamspace MIMO systems without obvious performance loss by beam selection [4], [5]. Specifically, according to the sparse beamspace channel matrix, only a small number of beams can be selected to simultaneously serve K users. Then, the received signal vector in (5) can be rewritten as

$$\bar{\mathbf{y}} = \bar{\mathbf{H}}_r^H \mathbf{W}_r \mathbf{P}\mathbf{s} + \mathbf{v}, \tag{7}$$

where $\bar{\mathbf{H}}_r = \bar{\mathbf{H}}(i, :)_{i \in \Gamma}$ of size $|\Gamma| \times K$ is the dimension-reduced beamspace channel matrix including selected beams, and Γ is the index set of selected beams. $\mathbf{W}_r$ of size $|\Gamma| \times K$ is the dimension-reduced precoding matrix. Since the row dimension of $\mathbf{W}_r$ is much smaller than N (the row dimension of the original precoding matrix $\mathbf{W}$), the number of required RF chains can be significantly reduced, and we have $N_{\mathrm{RF}} = |\Gamma|$ [9].

However, in existing beamspace MIMO systems, one beam can only support one user at most. Therefore, the maximum number of supported users at the same time-frequency resources is equal to the number of RF chains [9], i.e., $K \leq N_{\mathrm{RF}}$, which is the fundamental limit of beamspace MIMO systems that was explicitly or implicitly considered in all published papers on beamspace MIMO [2], [4], [5], [9], [15]. The reason is that, the DoF provided by the RF chains must be larger than or equal to the DoF required by users, otherwise signal for different users cannot be separated by linear operation. To break this limit, we propose a new mmWave transmission scheme that integrates NOMA with beamspace MIMO in the next subsection.

B. Proposed Beamspace MIMO-NOMA

In order to further improve spectrum efficiency and connectivity density, we propose to leverage NOMA in beamspace mmWave massive MIMO systems. As shown in Fig. 1 (c), unlike existing beamspace MIMO systems, more than one user can be simultaneously served within each selected beam in the proposed beamspace MIMO-NOMA scheme.

Specifically, beam selection algorithms, e.g., maximum magnitude (MM) selection [4] and maximization of the signal-to-interference-plus-noise ratio (SINR) selection [5], can be used to select one beam for each user, and each RF chain corresponds to one beam. Note that different users are likely to select the same beam, which are called "conflicting users" in this paper. Particularly, for a typical mmWave massive MIMO system with $N = 256$ antennas and $K = 32$ users whose spatial directions follow the uniform distribution, the probability that there exists users selecting the same beam is 87% [9]. In contrast to existing beamspace MIMO systems, where user scheduling is performed to select only one user out of these conflicting users [9], conflicting users can be simultaneously served using the same RF chain in the proposed beamspace MIMO-NOMA system.

Although the number of selected beams is equal to N_{RF}, the number of simultaneously served users K can be larger than N_{RF}, i.e., $K \geq N_{\mathrm{RF}}$. Let S_n for $n = 1, 2, \cdots, N_{\mathrm{RF}}$ denote the set of users served by the nth beam with $S_i \cap S_j = \Phi$ for $i \neq j$ and $\sum_{n=1}^{N_{\mathrm{RF}}} |S_n| = K$. The $N_{\mathrm{RF}} \times 1$ beamspace channel vector after beam selection between the BS and the mth user in the nth beam is denoted by $\mathbf{h}_{m,n}$, and $\mathbf{w}_n$ of size $N_{\mathrm{RF}} \times 1$ denotes the uniform precoding vector for users in the nth beam. Without loss of generality, we assume that $\left\|\mathbf{h}_{1,n}^H \mathbf{w}_n\right\|_2 \geq \left\|\mathbf{h}_{2,n}^H \mathbf{w}_n\right\|_2 \geq \cdots \geq \left\|\mathbf{h}_{|S_n|,n}^H \mathbf{w}_n\right\|_2$ for $n = 1, 2, \cdots, N_{\mathrm{RF}}$. The received signal $y_{m,n}$ at the mth user in the nth beam ($n = 1, 2, \cdots, N_{\mathrm{RF}}$, and $m = 1, 2, \cdots, |S_n|$) can be expressed as

$$\begin{aligned} y_{m,n} &= \mathbf{h}_{m,n}^H \sum_{j=1}^{N_{\mathrm{RF}}} \sum_{i=1}^{|S_j|} \mathbf{w}_j \sqrt{p_{i,j}} s_{i,j} + \upsilon_{m,n} \\ &= \underbrace{\mathbf{h}_{m,n}^H \mathbf{w}_n \sqrt{p_{m,n}} s_{m,n}}_{\text{desired signal}} \\ &\quad + \underbrace{\mathbf{h}_{m,n}^H \mathbf{w}_n \sum_{i=1}^{m-1} \sqrt{p_{i,n}} s_{i,n} + \mathbf{h}_{m,n}^H \mathbf{w}_n \sum_{i=m+1}^{|S_n|} \sqrt{p_{i,n}} s_{i,n}}_{\text{intra-beam interferences}} \\ &\quad + \underbrace{\mathbf{h}_{m,n}^H \sum_{j \neq n} \sum_{i=1}^{|S_j|} \mathbf{w}_j \sqrt{p_{i,j}} s_{i,j}}_{\text{inter-beam interferences}} + \underbrace{\upsilon_{m,n}}_{\text{noise}}, \end{aligned} \tag{8}$$

where $s_{m,n}$ and $p_{m,n}$ are the transmitted signal and transmitted power for the mth user in the nth beam, and $\upsilon_{m,n}$ is the noise

following the distribution $\mathcal{CN}\left(0,\sigma^2\right)$. Note that in existing beamspace MIMO systems, only one user can be supported in each selected beam, i.e., $|S_n|=1$ $(n=1,2,\cdots N_{\mathrm{RF}})$, while $|S_n|$ $(n=1,2,\cdots N_{\mathrm{RF}})$ can be larger than one in the proposed beamspace MIMO-NOMA system.

In the second equation of (8), the first term is the desired signal, the second and third terms are intra-beam interferences, the fourth term is inter-beam interference, and the last term is the noise. Particularly, the precoding vectors $\{\mathbf{w}_n\}_{n=1}^{N_{\mathrm{RF}}}$ should be carefully designed to restrain inter-beam interferences, which will be discussed later in Section IV. Intra-beam interferences caused by superposition coding in NOMA can be suppressed by carrying out SIC according to the increasing order of equivalent channel gains[2] [17], [19], [21], i.e., the mth user in the nth beam can remove the interferences from the ith user (for all $i>m$) in the nth beam by performing SIC.

By employing NOMA in beamspace MIMO systems, more than one user can be simultaneously served within each beam, and thus the total number of supported users can be larger than the number of beams, i.e., $K \geq N_{\mathrm{RF}}$. However, in the proposed beamspace MIMO-NOMA system, in addition to intra-beam interferences caused by superposition transmission, users also suffer from interferences from other beams. Thus, a straightforward combination of NOMA and beamspace MIMO is not able to guarantee the reliable performance in practice, and precoding as well as power allocation should be designed to reduce interferences by maximizing the achievable sum rate.

III. Achievable Sum Rate

As discussed in the previous section, in the nth beam using NOMA with SIC, the ith $(i>m)$ user's signal is detectable at the mth user, provided that it is detectable at itself [17], [19], [21], as the equivalent channel gain of the mth user is larger than that of the ith user, i.e., $\left\|\mathbf{h}_{1,n}^H\mathbf{w}_n\right\|_2 \geq \left\|\mathbf{h}_{2,n}^H\mathbf{w}_n\right\|_2 \geq \cdots \geq \left\|\mathbf{h}_{|S_n|,n}^H\mathbf{w}_n\right\|_2$ as assumed before. Therefore, the mth user can detect the ith user's signals for $1 \leq m < i \leq |S_n|$, and then remove the detected signals from its received signals, in a successive manner. Then, the remaining received signal at the mth user in the nth beam can be rewritten as

$$\hat{y}_{m,n} = \underbrace{\mathbf{h}_{m,n}^H\mathbf{w}_n\sqrt{p_{m,n}}s_{m,n}}_{\text{desired signal}} + \underbrace{\mathbf{h}_{m,n}^H\mathbf{w}_n\sum_{i=1}^{m-1}\sqrt{p_{i,n}}s_{i,n}}_{\text{intra-beam interferences}} + \underbrace{\mathbf{h}_{m,n}^H\sum_{j\neq n}\sum_{i=1}^{|S_j|}\mathbf{w}_j\sqrt{p_{i,j}}s_{i,j}}_{\text{inter-beam interferences}} + \underbrace{\upsilon_{m,n}}_{\text{noise}}. \quad (9)$$

Then, according to (9), the SINR at the mth user in the nth beam can be presented as

$$\gamma_{m,n} = \frac{\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n}}{\xi_{m,n}}, \quad (10)$$

[2]In this paper, we assume the beamspace channel is known by the BS. Actually, efficient tools of compressive sensing can be utilized to reliably estimate the beamspace channel with low pilot overhead thanks to the sparsity of beamspace channel in mmWave massive MIMO systems [2], [9].

where

$$\xi_{m,n} = \left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1}p_{i,n} + \sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|}p_{i,j} + \sigma^2. \quad (11)$$

As a result, the achievable rate at the mth user in the nth beam is

$$R_{m,n} = \log_2\left(1+\gamma_{m,n}\right). \quad (12)$$

Finally, the achievable sum rate of the proposed beamspace MIMO-NOMA scheme is

$$R_{\mathrm{sum}} = \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}R_{m,n}, \quad (13)$$

which can be improved by carefully designing the precoding $\{\mathbf{w}_n\}_{n=1}^{N_{\mathrm{RF}}}$ and power allocation $\left\{p_{m,n}\right\}_{m=1,n=1}^{|S_n|,N_{\mathrm{RF}}}$.

IV. Precoding in Downlink Systems

In existing beamspace MIMO systems, where only one user can be served in each beam (i.e., $K \leq N_{\mathrm{RF}}$), the classical linear ZF precoding with low complexity can be utilized to remove the inter-beam interferences [4], [5], [9], [15], which can be simply realized by the pseudo-inverse of the beamspace channel matrix for all users. However, in the proposed beamspace MIMO-NOMA system, the number of users is larger than the number of beams, i.e., $K \geq N_{\mathrm{RF}}$, which means the pseudo-inverse of the beamspace channel matrix of size $N_{\mathrm{RF}} \times K$ does not exist. As a result, the conventional ZF precoding cannot be directly used.

To address this problem, an equivalent channel can be determined for each beam to generate the precoding vector. Specifically, we introduce two methods to generate the equivalent channel for each beam, i.e., the strongest user-based equivalent channel and singular value decomposition (SVD)-based equivalent channel, which will be discussed in the following two subsections, separately.

A. The Strongest User-Based Equivalent Channel

As mentioned before, the amplitudes of NLoS components are typically 5 to 10 dB weaker than the amplitude of the LoS component [4], [33]. Therefore, the LoS component can primarily characterize the multipath channel in mmWave communications [5]. At the same time, the most important property of the beamspace channel matrix is that it has the sparse structure representing the directions of different users [4]. As a result, if the LoS component exists, the sparse beamspace channel vectors of different users in the same beam are highly correlated. That is to say, one of the beamspace channel vectors for multiplexed users in the nth beam can be regarded as the equivalent channel vector of the nth beam. Particularly, considering that the first user in each beam should perform SIC to decode all the other users' signals in this beam, we use the beamspace channel vector of the first user in each beam as the equivalent channel vector. Specifically, the equivalent channel matrix of size $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ for all

N_{RF} beams can be written as

$$\tilde{\mathbf{H}} = \left[\mathbf{h}_{1,1}, \mathbf{h}_{1,2}, \cdots, \mathbf{h}_{1,N_{\mathrm{RF}}}\right]. \tag{14}$$

Then, the precoding matrix of size $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ can be generated by

$$\tilde{\mathbf{W}} = \left[\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \cdots, \tilde{\mathbf{w}}_{N_{\mathrm{RF}}}\right] = \left(\tilde{\mathbf{H}}\right)^{\dagger} = \tilde{\mathbf{H}}\left(\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\right)^{-1}. \tag{15}$$

After normalizing the precoding vectors, the precoding vector for the nth beam ($n = 1, 2, \cdots, N_{\mathrm{RF}}$) can be written as

$$\mathbf{w}_n = \frac{\tilde{\mathbf{w}}_n}{\|\tilde{\mathbf{w}}_n\|_2}. \tag{16}$$

In this precoding scheme, the first user in each beam can completely remove the inter-beam interferences, i.e.,

$$\frac{\mathbf{h}_{1,j}^H\mathbf{w}_n}{\left\|\mathbf{h}_{1,j}^H\mathbf{w}_n\right\|_2} = \begin{cases} 0, & \text{for } j \neq n, \\ 1, & \text{for } j = n, \end{cases} \tag{17}$$

where $1 \leq j, n \leq N_{\mathrm{RF}}$. Thus, after performing SIC, the SINR at the first user in the nth beam can be rewritten as

$$\gamma_{1,n} = \frac{\left\|\mathbf{h}_{1,n}^H\mathbf{w}_n\right\|_2^2 p_{1,n}}{\sigma^2}. \tag{18}$$

B. SVD-Based Equivalent Channel

When the LoS component does not exist or the effect of NLoS components is significant, the channel correlation in the same beam may be not high enough. Therefore, we also consider another precoding scheme that exploits all of the beamspace channel vectors of users in the same beam, which is inspired by the precoding scheme used in conventional MU-MIMO systems [35]. Specifically, let $\mathbf{H}_n$ of size $N_{\mathrm{RF}} \times |S_n|$ denote the beamspace channel matrix of all $|S_n|$ users in the nth beam, i.e.,

$$\mathbf{H}_n = \left[\mathbf{h}_{1,n}, \mathbf{h}_{2,n}, \cdots, \mathbf{h}_{|S_n|,n}\right], \tag{19}$$

where $1 \leq n \leq N_{\mathrm{RF}}$. Then, by taking the SVD of $\mathbf{H}_n^T$, we have

$$\mathbf{H}_n^T = \mathbf{U}_n\Sigma_n\mathbf{V}_n^H, \tag{20}$$

where $\mathbf{U}_n$ is the left singular matrix of size $|S_n| \times |S_n|$, Σ_n is the $|S_n| \times N_{\mathrm{RF}}$ singular value matrix with its diagonal entries sorted in a non-increasing order, and $\mathbf{V}_n$ is the right singular matrix of size $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$. Then, the equivalent channel vector of the nth beam can be generated by

$$\tilde{\mathbf{h}}_n = \mathbf{H}_n\mathbf{u}_n^*, \tag{21}$$

where $\mathbf{u}_n$ is the first column of $\mathbf{U}_n$, i.e., the left singular vector corresponds to the maximum singular value. Finally, the equivalent channel matrix of size $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ can be expressed as

$$\tilde{\mathbf{H}} = \left[\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \cdots, \tilde{\mathbf{h}}_{N_{\mathrm{RF}}}\right] = \left[\mathbf{H}_1\mathbf{u}_1^*, \mathbf{H}_2\mathbf{u}_2^*, \cdots, \mathbf{H}_{N_{\mathrm{RF}}}\mathbf{u}_{N_{\mathrm{RF}}}^*\right]. \tag{22}$$

Similar to the strongest user-based equivalent channel, the ZF precoding matrix of size $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$ can be generated according to (15) and (16).

Note that in addition to ZF, other classical precoding schemes, e.g., minimum mean square error (MMSE) and Wiener filter (WF) [4], are also feasible based on the equivalent channel matrix in (14) or (22).

After obtaining the precoding vectors, power allocation for different users in different beams will be optimized in the next section to maximize the achievable sum rate (13) of the proposed beamspace MIMO-NOMA scheme.

V. Dynamic Power Allocation

The channel gain difference among users can be translated into multiplexing gains by superposition coding in NOMA systems. Therefore, power allocation has an important effect on the system performance. In fact, to suppress inter-user interferences and improve the achievable sum rate, lots of studies have been done to design power allocation in existing MIMO-NOMA systems. The integration of NOMA and MIMO was investigated in [24], where two users were considered in each beam with a random beamforming, and fixed power allocation schemes were utilized at the BS. In addition, fixed power allocation strategies have also been considered in [17]. A coordinated frequency block-dependent inter-beam power allocation was proposed in [25] to generate distinct power levels for different beams. Magnus and Neudecher [26] considered equal power allocation for different groups, and intra-group power allocation has been optimized to maximize the achievable sum rate, where each group only included two single-antenna users. The intra-group power optimization has also been investigated in [27] and [28], where a convex optimization algorithm was utilized to obtain the closed-form solution to power allocation. In [29], a non-convex power allocation problem was formulated for MIMO-NOMA systems, where only two users have been considered, and sub-optimal solutions were provided. Zhang *et al.* [30] investigated joint optimization of beamforming and power allocation, where channel uncertainties have been considered to maximize the worst-case achievable sum rate. The performance with only two users in each group was evaluated in the simulations.

Similar to existing MIMO-NOMA works, both inter-beam interferences and intra-beam interferences should be reduced to improve the achievable sum rate of the proposed beamspace MIMO-NOMA system. However, in contrast to existing MIMO-NOMA works, where fixed inter-beam power allocation and fixed number of users in each beam (e.g., two users in each beam) are usually considered, multiple users (e.g., 1, 2, or 3 users) are allowed in each beam in the proposed beamspace MIMO-NOMA scheme. Accordingly, a dynamic power allocation scheme to maximize the achievable sum rate is proposed by solving the joint power optimization problem, which not only includes the intra-beam power optimization, but also considers the inter-beam power optimization. The power allocation problem can be formulated as

$$\begin{aligned} &\max_{\{p_{m,n}\}} \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|} R_{m,n} \\ &\text{s.t. } C_1: \ p_{m,n} \geq 0, \quad \forall n, m, \\ &\qquad C_2: \ \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|} p_{m,n} \leq P, \\ &\qquad C_3: \ R_{m,n} \geq R_{\min}, \quad \forall n, m, \end{aligned} \tag{23}$$

where $R_{m,n}$ is the achievable rate of the m the user in the nth beam as defined in (12), the constraint C_1 indicates that the power allocated to each user must be positive, C_2 is the transmitted power constraint with P being the maximum total transmitted power by the BS, and C_3 is the data rate constraint for each user with $R_{\min}$ being the minimum data rate for each user. By substituting (10)-(12) into the constraint C_3 in (23), we have

$$\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n} - \eta\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1} p_{i,n} - \eta\sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|} p_{i,j} \geq \omega, \quad (24)$$

where $\eta = 2^{R_{\min}} - 1$ and $\omega = \eta\sigma^2$. In this way, the non-linear constraint C_3 has been transformed into linear constraint. Then, the optimization problem (23) can be rewritten as

$$\begin{aligned}
&\max_{\{p_{m,n}\}} \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}\log_2\left(1+\gamma_{m,n}\right)\\
&\text{s.t. } C_1: \ p_{m,n}\geq 0, \quad \forall n,m,\\
&\qquad C_2: \ \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|} p_{m,n}\leq P,\\
&\qquad C_3: \ \left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n} - \eta\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1} p_{i,n}\\
&\qquad\qquad - \eta\sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|} p_{i,j}\geq\omega, \quad \forall n,m. \quad (25)
\end{aligned}$$

We can see from (25) that all constraints, i.e., C_1, C_2, and C_3, are linear inequality constraints, while the objective function is non-convex. Therefore, this optimization problem is NP-hard, and it is very difficult to obtain the closed-form solution to the optimal power allocation problem (25).

To solve this difficult non-convex problem (25), we propose an iterative optimization algorithm to realize power allocation. Specifically, according to the extension of the Sherman-Morrison-Woodbury formula [30], [36], i.e.,

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{I}+\mathbf{CDA}^{-1}\mathbf{B}\right)^{-1}\mathbf{CDA}^{-1}, \quad (26)$$

we have

$$\begin{aligned}
&\left(1+\gamma_{m,n}\right)^{-1}\\
&\quad = 1 - \left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n}\left(\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n} + \xi_{m,n}\right)^{-1}, \quad (27)
\end{aligned}$$

where $n = 1, 2, \cdots, N_{\mathrm{RF}}$ and $m = 1, 2, \cdots, |S_n|$.

We can find that the expression (27) has the same form as the MMSE. Specifically, if MMSE detection is used to solve $s_{m,n}$ from $\hat{y}_{m,n}$ in (9), this detection problem can be formulated as

$$c_{m,n}^o = \arg\min_{c_{m,n}} e_{m,n}, \quad (28)$$

where

$$e_{m,n} = \mathrm{E}\left\{\left|s_{m,n} - c_{m,n}\hat{y}_{m,n}\right|^2\right\} \quad (29)$$

is the mean square error (MSE), $c_{m,n}$ is the channel equalization coefficient, and $c_{m,n}^o$ is the optimal value of $c_{m,n}$ to minimize the MSE. Substituting (9) into (29), we have

$$\begin{aligned}
e_{m,n} &= 1 - 2\mathrm{Re}\left(c_{m,n}\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right)\\
&\quad + \left|c_{m,n}\right|^2\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)\\
&= \left|1 - c_{m,n}\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right|^2\\
&\quad + \left|c_{m,n}\right|^2\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1} p_{i,n}\\
&\quad + \left|c_{m,n}\right|^2\sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|} p_{i,j} + \left|c_{m,n}\right|^2\sigma^2. \quad (30)
\end{aligned}$$

Then, by solving (28) based on (30), the optimal equalization coefficient $c_{m,n}^o$ can be calculated by

$$\begin{aligned}
&\left.\frac{\partial e_{m,n}}{\partial c_{m,n}}\right|_{c_{m,n}^o} = 0\\
&\Rightarrow -\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\\
&\qquad + \left(c_{m,n}^o\right)^*\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right) = 0\\
&\Rightarrow c_{m,n}^o = \left(\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right)^*\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)^{-1}. \quad (31)
\end{aligned}$$

Substituting (31) into (30), we obtain the MMSE as

$$\begin{aligned}
e_{m,n}^o &= 1 - 2\mathrm{Re}\left(c_{m,n}^o\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right)\\
&\quad + \left|c_{m,n}^o\right|^2\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)\\
&= 1 - 2\mathrm{Re}\Bigg(\left(\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right)^*\Bigg(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\\
&\qquad + \xi_{m,n}\Bigg)^{-1}\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\Bigg)\\
&\quad + \left|\left(\sqrt{p_{m,n}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right)^*\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)^{-1}\right|^2\\
&\quad \times\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)\\
&= 1 - 2p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)^{-1}\\
&\quad + p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)^{-1}\\
&= 1 - p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\left(p_{m,n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 + \xi_{m,n}\right)^{-1}. \quad (32)
\end{aligned}$$

which is equal to $\left(1+\gamma_{m,n}\right)^{-1}$ in (27), i.e., we have

$$\left(1+\gamma_{m,n}\right)^{-1}=\min_{c_{m,n}} e_{m,n}. \tag{33}$$

Then, the achievable rate of the mth user in the nth beam can be written as

$$R_{m,n}=\log_2\left(1+\gamma_{m,n}\right)=\max_{c_{m,n}}\left(-\log_2 e_{m,n}\right). \tag{34}$$

To remove the log function in (34), we introduce the following proposition [30].

Proposition 1: Let $f(a)=-\frac{ab}{\ln 2}+\log_2 a+\frac{1}{\ln 2}$ and a be a positive real number, we have

$$\max_{a>0} f(a)=-\log_2 b, \tag{35}$$

where the optimal value of a is $a^o=\frac{1}{b}$.

Proof: The function $f(a)$ is concave, and thus the maximum value of $f(a)$ can be obtained by solving

$$\left.\frac{\partial f(a)}{\partial a}\right|_{a=a^o}=0. \tag{36}$$

Then, we have $a^o=\frac{1}{b}$. By instituting a^o into $f(a)$, the maximum value of $f(a)$ is $-\log_2 b$. ∎

Using Proposition 1, (34) can be rewritten as

$$R_{m,n}=\max_{c_{m,n}}\max_{a_{m,n}>0}\left(-\frac{a_{m,n}e_{m,n}}{\ln 2}+\log_2 a_{m,n}+\frac{1}{\ln 2}\right). \tag{37}$$

As a result, the objective function for the optimization problem (25) has been transformed into quadratic programming function, and (25) can be reformulated as

$$\begin{aligned}
&\max_{\{p_{m,n}\}}\ \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}\max_{c_{m,n}}\max_{a_{m,n}>0}\left(-\frac{a_{m,n}e_{m,n}}{\ln 2}+\log_2 a_{m,n}+\frac{1}{\ln 2}\right)\\
&\text{s.t. } C_1:\ p_{m,n}\geq 0,\quad \forall n,m,\\
&\qquad C_2:\ \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}p_{m,n}\leq P,\\
&\qquad C_3:\ \left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n}-\eta\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1}p_{i,n}\\
&\qquad\qquad -\eta\sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|}p_{i,j}\geq\omega,\quad \forall n,m.
\end{aligned} \tag{38}$$

To solve the reformulated optimization problem (38), we propose to iteratively optimize $\{c_{m,n}\}$, $\{a_{m,n}\}$, and $\{p_{m,n}\}$. Specifically, given the optimal power allocation solution $\left\{p_{m,n}^{(t-1)}\right\}$ in the $(t-1)$th iteration, the optimal solution of $\left\{c_{m,n}^{(t)}\right\}$ in the tth iteration can be obtained according to (31), i.e.,

$$c_{m,n}^{(t)}=\left(\sqrt{p_{m,n}^{(t-1)}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right)^*\left(p_{m,n}^{(t-1)}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2+\xi_{m,n}^{(t-1)}\right)^{-1}, \tag{39}$$

where

$$\xi_{m,n}^{(t-1)}=\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1}p_{i,n}^{(t-1)}+\sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|}p_{i,j}^{(t-1)}+\sigma^2, \tag{40}$$

and the corresponding MMSE denoted by (32) in the tth iteration can be expressed as

$$e_{m,n}^{o(t)}=1-\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n}^{(t-1)}\left(\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n}^{(t-1)}+\xi_{m,n}^{(t-1)}\right)^{-1}. \tag{41}$$

Then, the optimal solution of $\left\{a_{m,n}^{(t)}\right\}$ in the tth iteration can be obtained by

$$a_{m,n}^{(t)}=\frac{1}{e_{m,n}^{o(t)}}. \tag{42}$$

After obtaining the optimal $\left\{c_{m,n}^{(t)}\right\}$ and $\left\{a_{m,n}^{(t)}\right\}$ in the tth iteration, the optimal $\left\{p_{m,n}^{(t)}\right\}$ in the tth iteration can be obtained by solving the following problem:

$$\begin{aligned}
&\min_{\left\{p_{m,n}^{(t)}\right\}}\ \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}a_{m,n}^{(t)}e_{m,n}^{(t)}\\
&\text{s.t. } C_1:\ p_{m,n}^{(t)}\geq 0,\quad \forall n,m,\\
&\qquad C_2:\ \sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}p_{m,n}^{(t)}\leq P,\\
&\qquad C_3:\ \left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2 p_{m,n}^{(t)}-\eta\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1}p_{i,n}^{(t)}\\
&\qquad\qquad -\eta\sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|}p_{i,j}^{(t)}\geq\omega,\quad \forall n,m,
\end{aligned} \tag{43}$$

where

$$\begin{aligned}
e_{m,n}^{(t)}=&\left|1-c_{m,n}^{(t)}\sqrt{p_{m,n}^{(t)}}\mathbf{h}_{m,n}^H\mathbf{w}_n\right|^2\\
&+\left|c_{m,n}^{(t)}\right|^2\left\|\mathbf{h}_{m,n}^H\mathbf{w}_n\right\|_2^2\sum_{i=1}^{m-1}p_{i,n}^{(t)}\\
&+\left|c_{m,n}^{(t)}\right|^2\sum_{j\neq n}\left\|\mathbf{h}_{m,n}^H\mathbf{w}_j\right\|_2^2\sum_{i=1}^{|S_j|}p_{i,j}^{(t)}+\left|c_{m,n}^{(t)}\right|^2\sigma^2.
\end{aligned} \tag{44}$$

To solve the convex optimization problem (43), we define the Lagrange function as

$$\begin{aligned}
L(p,\lambda,\mu)=&\sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}a_{m,n}^{(t)}e_{m,n}^{(t)}+\lambda\left(\sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}p_{m,n}^{(t)}-P\right)\\
&+\sum_{n=1}^{N_{\mathrm{RF}}}\sum_{m=1}^{|S_n|}\mu_{m,n}\theta_{m,n},
\end{aligned} \tag{45}$$

where

$$\theta_{m,n} = \eta \left\| \mathbf{h}_{m,n}^H \mathbf{w}_n \right\|_2^2 \sum_{i=1}^{m-1} p_{i,n}^{(t)} + \eta \sum_{j \neq n} \left\| \mathbf{h}_{m,n}^H \mathbf{w}_j \right\|_2^2 \sum_{i=1}^{|S_j|} p_{i,j}^{(t)} - \left\| \mathbf{h}_{m,n}^H \mathbf{w}_n \right\|_2^2 p_{m,n}^{(t)} + \omega, \quad (46)$$

$\lambda \geq 0$, and $\mu_{m,n} \geq 0$ ($n = 1, 2, \cdots, N_{\mathrm{RF}}$, $m = 1, 2, \cdots, |S_n|$). Then, the Karush-Kuhn-Tucker (KKT) [37] conditions of (43) can be obtained by the following three equations (47)-(49):

$$\begin{aligned} \frac{\partial L}{\partial p_{m,n}} = a_{m,n}^{(t)} & \left(\left| c_{m,n}^{(t)} \right|^2 \left\| \mathbf{h}_{m,n}^H \mathbf{w}_n \right\|_2^2 - \mathrm{Re}\left(c_{m,n}^{(t)} \mathbf{h}_{m,n}^H \mathbf{w}_n \right) \left(p_{m,n}^{(t)} \right)^{-\frac{1}{2}} \right) \\ & + \sum_{u=m+1}^{|S_n|} a_{u,n}^{(t)} \left| c_{u,n}^{(t)} \right|^2 \left\| \mathbf{h}_{u,n}^H \mathbf{w}_n \right\|_2^2 \\ & + \sum_{v \neq n} \sum_{u=1}^{|S_v|} a_{u,v}^{(t)} \left| c_{u,v}^{(t)} \right|^2 \left\| \mathbf{h}_{u,v}^H \mathbf{w}_n \right\|_2^2 + \lambda \\ & - \mu_{m,n} \left\| \mathbf{h}_{m,n}^H \mathbf{w}_n \right\|_2^2 + \sum_{u=m+1}^{|S_n|} \mu_{u,n} \eta \left\| \mathbf{h}_{u,n}^H \mathbf{w}_n \right\|_2^2 \\ & + \sum_{v \neq n} \sum_{u=1}^{|S_v|} \mu_{u,v} \eta \left\| \mathbf{h}_{u,v}^H \mathbf{w}_n \right\|_2^2 \\ = 0, & \quad (47) \end{aligned}$$

$$\lambda \left(\sum_{n=1}^{N_{\mathrm{RF}}} \sum_{m=1}^{|S_n|} p_{m,n}^{(t)} - P \right) = 0, \quad (48)$$

$$\mu_{m,n} \theta_{m,n} = 0, \quad \forall n, m. \quad (49)$$

From (47), we can obtain the optimal solution of $p_{m,n}$ as follows:

$$p_{m,n}^{(t)} = \left(\frac{a_{m,n}^{(t)} \mathrm{Re}\left(c_{m,n}^{(t)} \mathbf{h}_{m,n}^H \mathbf{w}_n \right)}{\tau} \right)^2, \quad (50)$$

where

$$\begin{aligned} \tau = & \sum_{u=m}^{|S_n|} a_{u,n}^{(t)} \left| c_{u,n}^{(t)} \right|^2 \left\| \mathbf{h}_{u,n}^H \mathbf{w}_n \right\|_2^2 \\ & + \sum_{v \neq n} \sum_{u=1}^{|S_v|} a_{u,v}^{(t)} \left| c_{u,v}^{(t)} \right|^2 \left\| \mathbf{h}_{u,v}^H \mathbf{w}_n \right\|_2^2 \\ & + \lambda - \mu_{m,n} \left\| \mathbf{h}_{m,n}^H \mathbf{w}_n \right\|_2^2 + \sum_{u=m+1}^{|S_n|} \mu_{u,n} \eta \left\| \mathbf{h}_{u,n}^H \mathbf{w}_n \right\|_2^2 \\ & + \sum_{v \neq n} \sum_{u=1}^{|S_v|} \mu_{u,v} \eta \left\| \mathbf{h}_{u,v}^H \mathbf{w}_n \right\|_2^2. \quad (51) \end{aligned}$$

Since (28), $f(a)$ in (35), and (43) are convex (or concave), the obtained $\{c_{m,n}^{(t)}\}$, $\{a_{m,n}^{(t)}\}$, and $\{p_{m,n}^{(t)}\}$ are optimal solutions in the tth iteration. Therefore, iteratively updating $\{c_{m,n}\}$, $\{a_{m,n}\}$, and $\{p_{m,n}\}$ will increase or maintain the value of the objective function in (38) [30]. With the constraint of the maximum transmitted power P, we will obtain a monotonically non-decreasing sequence of the objective value in (38) with a upper bound, i.e., the global maximum. As a result, the proposed iterative optimization algorithm for power allocation will converge to a stationary solution to the problem (38). To this end, we summarize the procedure of the proposed solution in **Algorithm 1**.

Algorithm 1 Proposed Iterative Power Allocation Algorithm

Input:

Beamspace channel vectors: $\mathbf{h}_{m,n}$ for $\forall n, m$;
Precoding vectors: $\mathbf{w}_n$ for $\forall n$;
Noise variance: σ^2;
Maximum iteration times: $T_{\max}$.

Output:

Power allocation: $p_{m,n}$ for $\forall m, n$.

1: $t = 0$.
2: **while** $t < T_{\max}$ **do**
3: Obtain the optimal $\{c_{m,n}^{(t)}\}$ according to (39);
4: Obtain the optimal $\{a_{m,n}^{(t)}\}$ according to (42);
5: Obtain the optimal $\{p_{m,n}^{(t)}\}$ according to (50);
6: $t = t + 1$.
7: **end while**
8: **return** $p_{m,n} = p_{m,n}^{(t)}$ for $\forall n, m$.

In each iteration, the complexity to obtain the optimal $\{c_{m,n}\}$ in (39) and $\{a_{m,n}\}$ in (42) is linear to the number of users, i.e., $O(K)$. λ in (48) and $\mu_{m,n}$ ($n = 1, 2, \cdots, N_{\mathrm{RF}}$, $m = 1, 2, \cdots, |S_n|$) in (49) can be obtained by using Newton¡¯s or bisection method with the complexity $O(K^2 \log_2(\delta))$, where δ is the required accuracy. As a result, the complexity of the proposed power allocation algorithm is $O(T_{\max} K^2 \log_2(\delta))$, where $T_{\max}$ is the maximum iteration times. Thus, the proposed iterative power allocation algorithm can be realized with a polynomial complexity.

VI. Simulation Results

In this section, we provide the simulation results to verify the performance of the proposed beamspace MIMO-NOMA system. Specifically, we consider a typical downlink mmWave massive MIMO system where the BS is equipped with an ULA of $N = 256$ antennas and communicates with K users. The total transmitted power is set as $P = 32$ mW (15 dBm) [9]. One LoS component and $L = 2$ NLoS components are assumed for all users' channels. We consider the channel parameters of user k as follows: 1) $\beta_k^{(0)} \sim \mathcal{CN}(0, 1)$, $\beta_k^{(l)} \sim \mathcal{CN}\left(0, 10^{-1}\right)$ for $1 \leq l \leq L$; 2) $\theta_k^{(0)}$ and $\theta_k^{(l)}$ for $1 \leq l \leq L$ follow the uniform distribution within $\left[-\frac{1}{2}, \frac{1}{2}\right]$. The signal-to-noise ratio (SNR) is defined as $\frac{E_b}{\sigma^2}$ in this paper.

In the simulations, we consider the following four typical mmWave massive MIMO schemes for comparison: (1) "Fully digital MIMO", where each antenna is connected to one RF chain, i.e., $N_{\mathrm{RF}} = N$; (2) "Beamspace MIMO" [9], where each beam only contains one user with $N_{\mathrm{RF}} = K$; (3) "MIMO-OMA" [27] with $N_{\mathrm{RF}} \leq K$, where OMA is performed for conflicting users, and users in the same beam

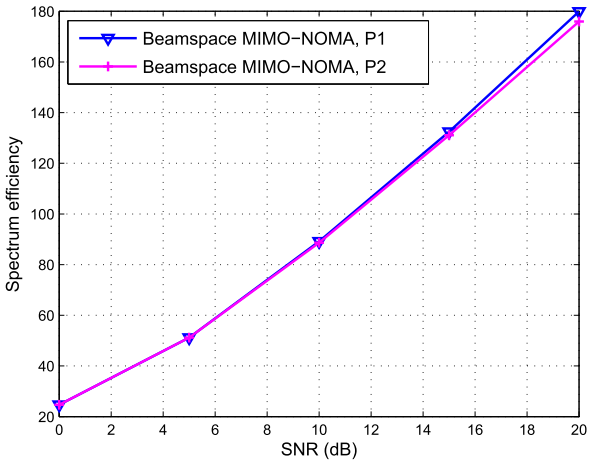

Fig. 2. Spectrum efficiency against SNR, where the number of users is $K = 32$.

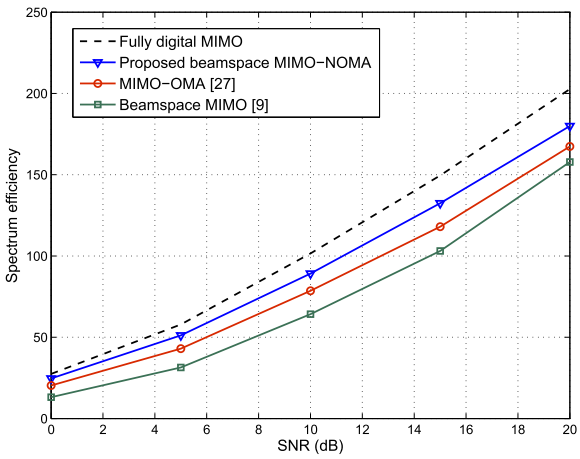

Fig. 3. Spectrum efficiency against SNR, where the number of users is $K = 32$.

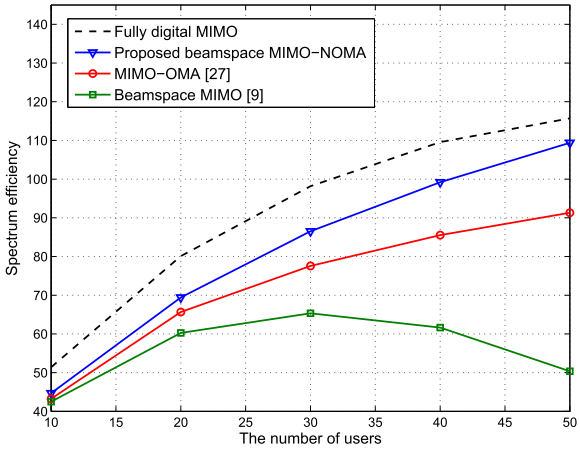

Fig. 4. Spectrum efficiency against the number of users K, where SNR $= 10$ dB.

are allocated with orthogonal frequency resources; (4) The "proposed beamspace MIMO-NOMA" with $N_{\mathrm{RF}} \leq K$, which integrates NOMA and beamspace MIMO. Both the strongest user-based equivalent channel and SVD-based equivalent channel introduced in Section IV are considered, and the power allocation algorithm proposed in Section V is performed to alleviate interferences. Particularly, ZF precoding is considered in fully digital MIMO and beamspace MIMO. In the following subsections, the performance of the proposed beamspace MIMO-NOMA scheme will be evaluated in terms of spectrum efficiency[3] and energy efficiency.

A. Spectrum Efficiency

Fig. 2 shows the spectrum efficiency against SNR of the proposed beamspace MIMO-NOMA scheme with the strongest user-based equivalent channel (denoted as "beamspace MIMO-NOMA, P1") and SVD-based equivalent channel (denoted as "beamspace MIMO-NOMA, P2"), where the number of users is $K = 32$ and the iteration times to solve the power allocation optimization problem is set as 20, which is sufficient to make the iterative power allocation algorithm converged as shown later in Fig. 7. We can see from the simulation results that "beamspace MIMO-NOMA, P1" with very low complexity and "beamspace MIMO-NOMA, P2" with much higher complexity caused by the SVD have very similar performance, which indicates that although only the strongest user's channel is considered to perform precoding in the proposed rank-deficient beamspace MIMO-NOMA scheme, it is able to achieve the similar result compared to the precoding that considers the effect of all users's channels in the same beam. This favorable result is attributed to the strong correlation of beamspace channels in the same beam. Thus, in the following simulations, we only consider the strongest user-based equivalent channel to realize low-complexity precoding since SVD is not required.

[3]When the normalized bandwidth is considered, the spectrum efficiency can defined as the achievable sum rate (13).

Fig. 3 shows the spectrum efficiency against SNR of the considered four schemes mentioned above, where the number of users is $K = 32$. We can find that the proposed beamspace MIMO-NOMA scheme can achieve higher spectrum efficiency than that of beamspace MIMO [9] as well as MIMO-OMA [27]. Particularly, the proposed beamspace MIMO-NOMA has about 3 dB SNR gain compared to the beamspace MIMO, which benefits from the use of NOMA to serve multiple users in each beam. In addition, the proposed beamspace MIMO-NOMA also outperforms MIMO-OMA, since NOMA can achieve higher spectrum efficiency than that of OMA [17], [19], [21]. It is intuitive that the fully digital MIMO can achieve the best spectrum efficiency as shown in Fig. 3, since there is not beam selection in the fully digital MIMO, and $N_{\mathrm{RF}} = N$ RF chains are used to serve all users. However, the fully digital MIMO suffers from the worst energy efficiency as shown in Fig. 5, which will be discussed later.

The performance comparison in terms of spectrum efficiency against the number of users is shown in Fig. 4, where SNR is set as 10 dB. We can see from the simulation results that with the increasing of the number of users K, the performance gap between the beamspace MIMO and the

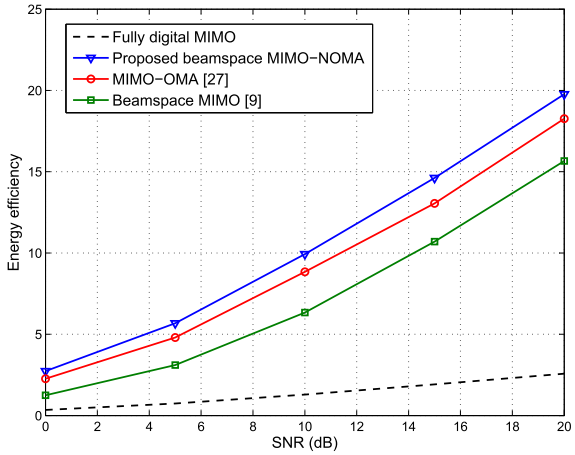

Fig. 5. Energy efficiency against SNR, where the number of users is $K = 32$.

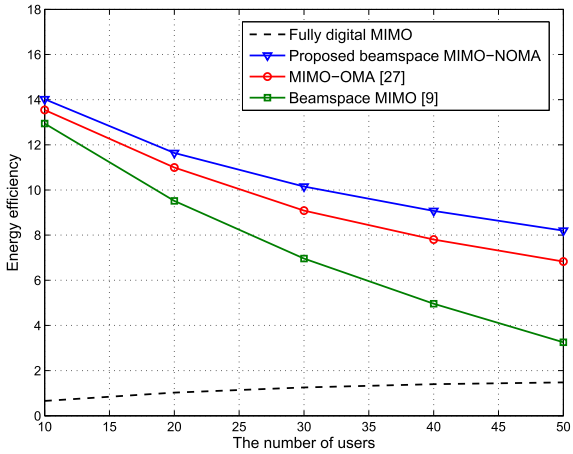

Fig. 6. Energy efficiency against the number of users K, where SNR $= 10$ dB.

proposed beamspace MIMO-NOMA becomes larger. This is because the larger the number of users, the larger the probability that the same beam is selected for different users is. As a result, existing beamspace MIMO will suffer from an obvious performance loss, while the proposed beamspace MIMO-NOMA can still perform well due to the use of NOMA.

B. Energy Efficiency

The energy efficiency ε is defined as the ratio between the achievable sum rate R_{sum} and the total power consumption [6], i.e.,

$$\varepsilon = \frac{R_{\text{sum}}}{P + N_{\text{RF}} P_{\text{RF}} + N_{\text{RF}} P_{\text{SW}} + P_{\text{BB}}} \text{ (bps/Hz/W)}, \quad (52)$$

where P is the maximum transmitted power, P_{RF} is the power consumed by each RF chain, P_{SW} is the power consumption of switch, and P_{BB} is the baseband power consumption. Particularly, we adopt the typical values $P_{\text{RF}} = 300$ mW, $P_{\text{SW}} = 5$ mW, and $P_{\text{BB}} = 200$ mW [6].

Fig. 5 shows the energy efficiency against SNR, where the number of users is also $K = 32$. We can find that the proposed beamspace MIMO-MOMA can achieve higher energy efficiency than other three schemes. Particularly, the proposed beamspace MIMO-MOMA has about 25% energy efficiency improvement compared to existing beamspace MIMO, which benefits from the use of NOMA to serve multiple users in each beam. In addition, the proposed beamspace MIMO-MOMA can achieve much higher energy efficiency than the fully digital MIMO scheme, where the number of RF chains is equal to the number of BS antennas, which leads to very high energy consumption, e.g., 300 mW for each RF chain. On the contrary, the number of RF chains is much smaller than the number of antennas in the proposed beamspace MIMO-MOMA scheme. Therefore, the energy consumption caused by the RF chains can be significantly reduced compared to the fully digital MIMO scheme.

The performance comparison in terms of energy efficiency against the number of users is shown in Fig. 6, where SNR is set as 10 dB. We can see that the energy efficiency of the proposed beamspace MIMO-MOMA scheme is higher than all of other three schemes even the number of users is very large (e.g., 50 users are simultaneously severed).

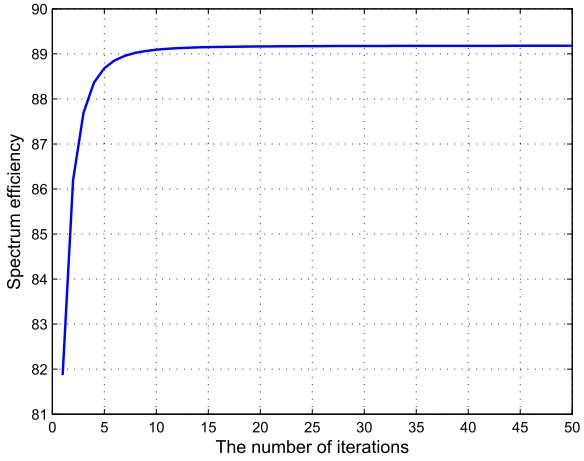

Fig. 7. Spectrum efficiency against the number of iterations for power allocation.

C. Convergence of Power Allocation

In this subsection, we evaluate the convergence of the proposed iterative power allocation algorithm in Section V, where the number of users is set as $K = 32$, and SNR $=$ 10 dB. As shown in Fig. 7, the spectrum efficiency tends to be stable after 10 times of iteration, which verifies the convergence of the proposed power allocation as discussed in Section V.

D. The User Fairness

In this subsection, we evaluate the user fairness of the proposed beamspace MIMO-NOMA scheme, where the number of users is set as $K = 32$, SNR $= 20$ dB, and $R_{\min} =$ 1 bit/s/HZ. As shown in Fig. 8, the achievable data rate for each user is larger than the minimum data rate $R_{\min}$, due to the data rate constraint for each user, i.e., C_3 in (23).

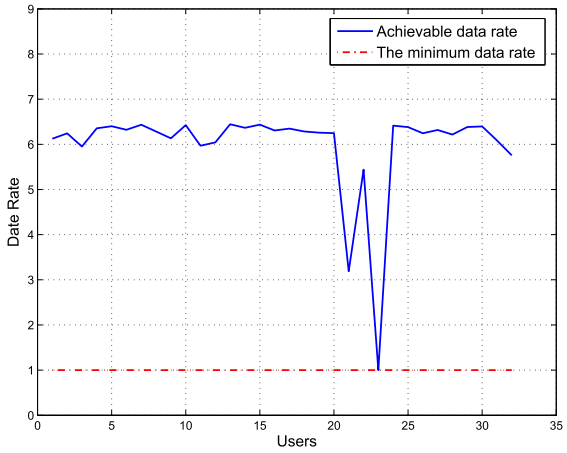

Fig. 8. The achievable data rate for each user.

VII. Conclusions

In this paper, we have proposed a new transmission scheme, i.e., beamspace MIMO-NOMA, to integrate NOMA and beamspace MIMO to break the fundamental limit of existing beamspace MIMO that only one user can be served in each beam at the same time-frequency resources. Particularly, the number of users can be larger than the number of RF chains in the proposed beamspace MIMO-NOMA scheme, which is essentially different from existing beamspace MIMO systems. In addition to realizing massive connectivity, the use of NOMA can also improve the capacity bound of beamspace MIMO. To restrain the inter-beam interferences, the equivalent channel vector is determined for each beam to realize precoding based on the principle of ZF, which considers the high correlation of users' beamspace channels in the same beam at mmWave frequencies. Furthermore, to suppress both inter-beam and intra-beam interferences, we proposed to jointly optimize the power allocation of all users by maximizing the achievable sum rate, and an iterative optimization algorithm has been developed to realize power allocation. Simulation results have shown that the proposed beamspace MIMO-NOMA can achieve better performance in terms of spectrum and energy efficiency compared to existing beamspace MIMO, e.g., 25% energy efficiency gain can be achieved. In the future, we will consider sophisticated user pairing/clustering for the proposed beamspace MIMO-NOMA scheme in ultra-dense network (UDN).

References

[1] S. Mumtaz, J. Rodriquez, and L. Dai, *MmWave Massive MIMO: A Paradigm for 5G*. Orlando, FL, USA; Academic, 2016.

[2] J. Brady, N. Behdad, and A. Sayeed, "Beamspace MIMO for millimeterwave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.

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

[4] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 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. Han, C.-L. I, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

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

[8] X. Cheng *et al.*, "Communicating in the real world: 3D MIMO," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 136–144, Aug. 2014.

[9] X. Gao, L. Dai, Z. Chen, Z. Wang, and Z. Zhang, "Near-optimal beam selection for beamspace mmWave massive MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1054–1057, May 2016.

[10] 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.

[11] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1089–1098, Aug. 2004.

[12] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[13] A. F. Molisch and M. Z. Win, "MIMO systems with antenna selection," *IEEE Commun. Mag.*, vol. 5, no. 1, pp. 46–56, Mar. 2004.

[14] S. Sanayei and A. Nosratinia, "Capacity of MIMO channels with antenna selection," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 4356–4362, Nov. 2007.

[15] 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.

[16] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

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

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

[19] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[20] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[21] 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.

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

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

[24] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. 77th IEEE VTC-Spring*, Dresden, Germany, Jun. 2013, pp. 1–5.

[25] Y. Hayashi, Y. Kishiyama, and K. Higuchi, "Investigations on power allocation among beams in non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2013, pp. 1–5.

[26] B. Kim *et al.*, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Military Commun. Conf. (MILCOM)*, Nov. 2013, pp. 1278–1283.

[27] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation in non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.

[28] M. S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, Dec. 2016.

[29] Q. Sun, S. Han, C. L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Aug. 2015.
[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] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.
[32] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.
[33] T. S. Rappaport, E. Ben-Dor, J. N. Murdock, and Y. Qiao, "38 GHz and 60 GHz angle-dependent propagation for cellular & peer-to-peer wireless communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 4568–4573.
[34] 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.
[35] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.
[36] J. R. Magnus and H. Neudecher, *Matrix Differential Calculus with Application in Statistics and Econometrics*. New York, NY, USA: Wiley, 1988.
[37] S. Boyd and L. Vandenberghe, *Convex Optimication*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Bichai Wang (S'15) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015, where she is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. Her research interests are in wireless communications, with the emphasis on new multiple access techniques. She has received the Freshman Scholarship of Tsinghua University in 2011, the Academic Merit Scholarships of Tsinghua University in 2012, 2013, and 2014, respectively, the Excellent Thesis Award of Tsinghua University in 2015, and the National Scholarship in 2016.

Linglong Dai (M'11–SM'14) received the B.S. degree from Zhejiang University in 2003, the M.S. degree (Hons.) from the China Academy of Telecommunications Technology in 2006, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, China, in 2011. From 2011 to 2013, he was a Post-doctoral Research Fellow with the Department of Electronic Engineering, Tsinghua University, where he has been an Assistant Professor since 2013 and then an Associate Professor since 2016. He has authored over 50 IEEE journal papers and over 30 IEEE conference papers. He also holds 13 granted patents. His current research interests include massive MIMO, millimeter-wave communications, multiple access, and sparse signal processing. He has received four conference Best Paper Awards from the IEEE ICC 2013, the IEEE ICC 2014, WCSP 2016, and the IEEE ICC 2017. He has also received the IEEE TRANSACTIONS ON BROADCASTING Best Paper Award in 2015. He currently serves as the Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE COMMUNICATIONS LETTERS.

Zhaocheng Wang (M'09–SM'11) received the B.S., M.S., and Ph.D. degrees from Tsinghua University, Beijing, China, in 1991, 1993, and 1996, respectively. From 1996 to 1997, he was a Post-Doctoral Fellow with Nanyang Technological University, Singapore. From 1997 to 1999, he was with the OKI Techno Center (Singapore) Pte. Ltd., Singapore, where he was a Research Engineer and then became a Senior Engineer. From 1999 to 2009, he was with Sony Deutschland GmbH, where he was a Senior Engineer and then became a Principal Engineer. He is currently a Professor with the Department of Electronic Engineering, Tsinghua University, and also serves as the Director of the Broadband Communication Key Laboratory and the Tsinghua National Laboratory for Information Science and Technology. He has authored or co-authored over 130 journal papers and holds 34 granted U.S./EU patents. He has co-authored two books, one of which, *Millimeter Wave Communication Systems*, was selected by the IEEE Series on Digital and Mobile Communication (Wiley–IEEE Press). His research interests include wireless communications, visible light communications, millimeter wave communications, and digital broadcasting. He received the 2013 Beijing Science and Technology Award (First Prize), the IEEE ICC 2013 Best Paper Award, the OECC 2015 Best Student Award, the 2016 IEEE Scott Helt Memorial Award (Best Paper Award of the IEEE TRANSACTIONS ON BROADCASTING), the 2016 National Award for Science and Technology Progress (First Prize), and the IEEE ICC 2017 Best Paper Award. He is a fellow of the Institution of Engineering and Technology. He served as an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2011 to 2015 and an Associate Editor of the IEEE COMMUNICATIONS LETTERS from 2013 to 2016, and has also served as the technical program committee co-chairs of various international conferences.

Ning Ge (M'97) received the B.S. and Ph.D. degrees from Tsinghua University, China, in 1993 and 1997, respectively. From 1998 to 2000, he was involved in the development of ATM switch fabric ASIC with ADC Telecommunications, Dallas. Since 2000, he has been with the Department of Electronics Engineering, Tsinghua University, where he is currently a Professor and also serves as the Director of Communication Institute. His research interests include ASIC design, short range wireless communication, and wireless communications. He is a senior member of CIC and CIE.

Shidong Zhou (M'98) received B.S. and M.S. degrees from Southeast University, Nanjing, China, in 1991 and 1994, respectively, and the Ph.D. degree from Tsinghua University, Beijing, China, in 1998. He is currently a Professor with the Department of Electronic Engineering, Tsinghua University. He was involved in several major national projects on 3G and 4G mobile communication technique research and development. His research interests include mobile communication system architectures, advanced transmission technique, wireless channel sounding and modeling, radio resource management, and high energy efficient wireless networks.