Spatial Multiplexing With Limited RF Chains: Generalized Beamspace Modulation (GBM) for mmWave Massive MIMO

Shijian Gao, Xiang Cheng, *Senior Member, IEEE*, and Liuqing Yang, *Fellow, IEEE*

***Abstract*— Millimeter wave (mmWave) massive multiple-input multiple-output (mMIMO) has been recognized as a promising candidate for 5G communications for its capability of supporting Gb/s transmission. However, it is a common exercise to deploy a limited number of radio-frequency (RF) chains at mmWave mMIMO transceivers due to hardware complexity and cost. As a result, the potential multiplexing gain (MG), which is restricted by the smaller number of RF chains at the transmitter and receiver, is markedly compromised. In order to boost the MG and spectral efficiency (SE), we innovatively develop a novel index modulation termed as the generalized beamspace modulation (GBM). The acquisition of (sub-)beamspace is owing to a natural exploitation of the unique features of mmWave mMIMO. Based on the (sub-)beamspace, a complete GBM transceiver is designed and optimized. Unlike existing alternatives that are largely digital based, our GBM is tailored for the hybrid structure of mmWave mMIMO and can, thereby, realize efficient spatial multiplexing despite the limited RF chains. Extensive analyses and simulations have demonstrated remarkable superiority of GBM over existing counterparts in terms of the error performance and SE.**

***Index Terms*— Millimeter wave, massive multiple-input multiple-output, generalized beamspace modulation, multiplexing gain, spectral efficiency.**

I. Introduction

In mmWave mMIMO, the potential multiplexing gain (MG) is fundamentally limited by the minimum number of radio frequency (RF) chains at both ends. To cope with this fundamental limit and further boost the spectral efficiency (SE), there is an urgent need to develop index modulation (IM) techniques suitable for mmWave mMIMO.

Manuscript received December 1, 2018; revised April 8, 2019; accepted June 22, 2019. Date of publication July 25, 2019; date of current version August 19, 2019. This work was supported in part by the National Key Research and Development Project under Grant 2017YFE0121400, in part by the Major Project from Beijing Municipal Science and Technology Commission under Grant Z181100003218007, in part by the National Science and Technology Major Project under Grant 2018ZX03001031, in part by the National Natural Science Foundation of China under Grant 61622101 and Grant 61571020, and in part by NSF ECCS-1935915. This paper part has been presented in the IEEE Global Communications Conference (Globecom), Abu Dhabi, UAE, December 9-13, 2018 [1]. *(Corresponding author: Xiang Cheng.)*

S. Gao and L. Yang are with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523-1373 USA (e-mail: sjgao@colostate.edu; lqyang@engr.colostate.edu).

X. Cheng is with the State Key Laboratory of Advanced Optical Communication Systems and Networks, Department of Electronics, School of Electronics Engineering and Computing Sciences, Peking University, Beijing 100080, China (email: xiangcheng@pku.edu.cn).

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

Digital Object Identifier 10.1109/JSAC.2019.2929400

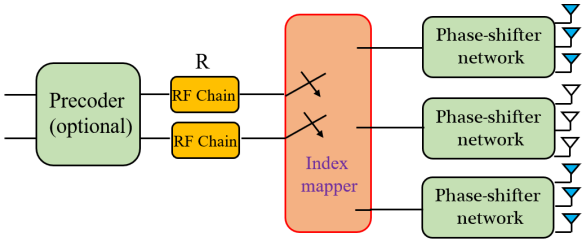

Fig. 1. The transmitter of mmWave AG-GSM.

As a matter of fact, similar problems have already been investigated in existing centimeter wave (cmWave) MIMO systems, and an effective solution is the generalized spatial modulation (GSM). The main idea of GSM is to convey the so-called index bits by utilizing the activation status of antennas [2]–[5], therefore a higher MG can be achieved by activating a subset of RF chains [6]–[8]. Although GSM has demonstrated remarkable superiority in cmWave MIMO, it does not directly apply to mmWave mMIMO. First, unlike cmWave MIMO where a digital structure is employed, mmWave mMIMO typically adopts an economic hybrid structure for power consumption and hardware cost concerns [9]–[11]. Secondly, different from the typically isotropic environment in cmWave propagation, mmWave channels are well known to exhibit limited scattering [12]–[16], thus the highly correlated channels may severely affect the error performance. As a result, a simple transplantation of existing cmWave IM techniques into mmWave mMIMO is not feasible. Instead, the ultimate solution requires a judicious design by accounting for the unique properties of mmWave mMIMO. To this end, the first step, which is also the top priority, is to seek a proper space where the index mapping can take place. We will first list existing options, and then introduce our proposed approach.

A. IM in Spatial Domain

A natural implementing space is the spatial domain; that is, the index bits directly determine which antennas are activated. In [17], an antenna-group (AG-) GSM is designed for mmWave MIMO, with transmitter structure shown in Fig. 1. Clearly, in mMIMO, directly (de-)activating each and every antenna will incur unbearable complexity, together with a huge number of RF chains. AG-GSM is adapted to hybrid mmWave structures by (de-)activating groups of (as opposed to individual) antennas. However, this approach essentially

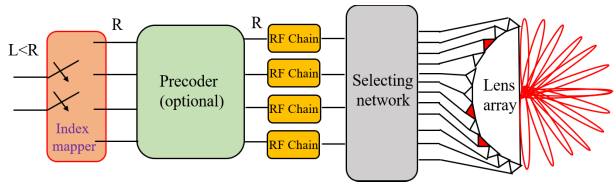

Fig. 2. The transmitter of mmWave EDC-IM.

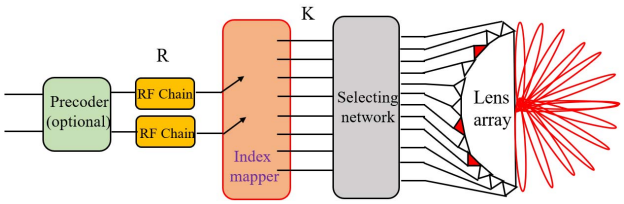

Fig. 3. The transmitter of mmWave GBM.

divides the entire array to a few groups of smaller ones, and will thus suffer from a severe loss of array gain and angle resolution. In addition, as the MG is dictated by the number of groups, there is clearly a tradeoff between the achievable MG and the array gain/angle resolution.

B. IM in Digital Domain

Recall that GSM is essentially a digital technique, thus can be directly applied to the equivalent digitized channel (EDC) that is encountered before the RF chains at the transmitter (Tx) and after the RF chains at the receiver (Rx). The Tx structure of EDC-IM is shown in Fig. 2. Although only a subset of RF chains are activated when performing IM, all antennas are employed for transmission. Hence EDC-IM is not only applicable to hybrid structures, but can also fully exploit the large array gain. However, its maximum MG is clearly limited by the minimum number of RF chains at the transceivers.

In view of the limitations of the abovementioned options, one may have realized that a proper domain has to leverage both the channel properties and the hybrid hardware structures that are unique to mmWave mMIMO. In this work, we innovatively resort to the beamspace, and the index mapping in GBM takes place neither at the antennas as the AG-GSM nor before the RF chains as the EDC-IM. Instead, the index mapping occurs after the RF chains but before the selecting network. Different from both aforementioned options, all RF chains and all antennas at the Tx are always active. As a result, not only that the array gain is fully exploited, but also the achievable MG is no longer restricted by the number of RF chains. The resultant GBM design is also perfectly compatible with prevalent mmWave mMIMO systems. Due to the unique placement of the index mapping module, the GBM design consists of two parts: i) the digital part accounting for the index mapping and demodulation functionality; and ii) the analog part involving the selecting network that optimizes the beam selection. All these, together with the precoder options, will be discussed in detail herein.

Our contributions can be briefly summarized as follows:

- By exploiting the unique propagation characteristics of mmWave channels and the hybrid transceiver structure, we propose GBM that facilitates an MG exceeding the number of Tx-end RF chains, without compromising the array gain or system compatibility.
- For the digital part of GBM, we design its modulator prototype with optimized pattern selection, and the demodulation with both the optimal maximum likelihood (ML) detector and a low-complexity zero-forcing with 2-step quantization (ZF-2Q) detector.
- For the analog part of GBM, we derive a probability bound of the maximum achievable MG and show that when the array size approaches infinity, the maximum MG approaches the rank of channel matrix with probability 1. Given an achievable MG, a low-complexity yet near-optimal beam selection scheme is proposed to optimize the error performance.
- We validate the performance of GBM in terms of the asymptotic pairwise error probability (APEP) and the SE via extensive theoretical analyses and simulations.

The remainder of this paper is organized as follows. Section II introduces the generic mmWave mMIMO system and the beamspace. Sections III and IV elaborate on the digital and analog parts of GBM design, respectively. Section V provides performance analyses and discussions. Extensive simulations and comparisons are presented in Section VI, followed by conclusions in Section VII.

Notation: a, $\boldsymbol{a}$ and $\boldsymbol{A}$ represent a scalar, a vector and a matrix, respectively. $\boldsymbol{a}[i]$ represents the i-th element of $\boldsymbol{a}$. $\boldsymbol{A}[m,n]$, $\boldsymbol{A}[m,:]$ and $\|\boldsymbol{A}\|_{\mathrm{F}}$ are denoted as the (m,n) entry, the m-th row and the Frobenius norm of $\boldsymbol{A}$. $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ represent the floor and ceiling operation, respectively. $\mathbb{C}_n^k$ denotes the number of k combinations from a given set of n elements. $\mathbb{E}$ stands for expectation. $\mathcal{CN}(0,\sigma^2)$ represents the distribution of a circularly symmetric complex Gaussian random value with variance σ^2. $Q(\cdot)$, $\mathbf{I}(\cdot)$ and $\mathbb{B}(\cdot)$ represent the Gaussian Q-function, the binary indicating function and the Beta function, respectively.

II. System and the Beamspace

In this paper, we consider an uplink mmWave mMIMO system, where the mobile station (MS) and the base station (BS) are equipped with lens arrays, each having M and N antennas, respectively. Essentially, an M-dimensional lens array plays the role of an $M \times M$ spatial FFT matrix, which contains orthogonal steering beams covering the entire beam domain [16]. To alleviate the high power consumption and deployment cost, the numbers of RF chains at both ends are much smaller than that of the antennas.

At the MS, let $\mathcal{M} = \{m_1, m_2, \cdots, m_K\}$ with $K \leq \min(M,N)$ be the set containing the indices of selected beams from the M-dimensional FFT matrix $\boldsymbol{F}_M$. The function of the selecting network (SN) can be described by

$$\boldsymbol{S}_{\mathcal{M}} = [\boldsymbol{e}_M(m_1), \boldsymbol{e}_M(m_2), \cdots, \boldsymbol{e}_M(m_K)] \tag{1}$$

where $\boldsymbol{e}_M(m)$ is the m-th column of $\boldsymbol{I}_M$. Let $\boldsymbol{s} = [s_1, s_2, \cdots, s_K]^T$ be the symbol vector to be transmitted from the SN port. After propagation through the $N \times M$ channel $\boldsymbol{H}$, the received signal at the BS is

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{F}_M\boldsymbol{S}_{\mathcal{M}}\boldsymbol{s} + \boldsymbol{n} \tag{2}$$

where $\boldsymbol{n} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$ is the Gaussian noise vector.

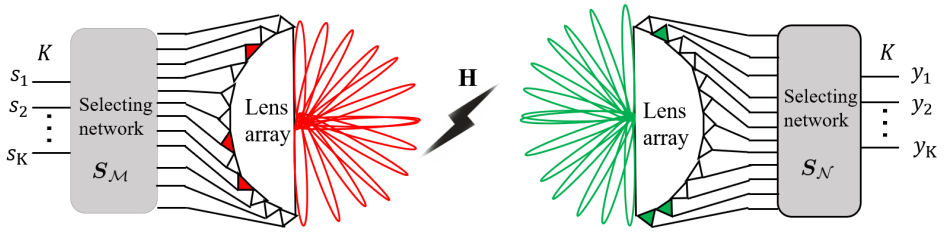

Fig. 4. The system model of the uplink mmWave beamspace mMIMO.

Similar to [18], [19], we consider a narrow-band block fading channel. The propagation environments between the MS and the BS are modeled as the widely accepted geometric channel consisting of P paths.[1] With uniform linear array (ULA) antennas configured at both ends, the channel matrix is given by

$$\boldsymbol{H} = \sqrt{\frac{MN}{P}} \sum_{p=1}^{P} \alpha_p \boldsymbol{a}_r(\theta_p) \boldsymbol{a}_t^*(\phi_p) \tag{3}$$

where $\alpha_p \sim \mathcal{CN}(0,1)$ is the complex gain of the p-th path; θ_p and ϕ_p represent the corresponding angle of arrival (AoA) and angle of departure (AoD), respectively, both modeled as uniformly distributed variables on $[0, 2\pi)$; $\boldsymbol{a}_t(\cdot)$ and $\boldsymbol{a}_r(\cdot)$ stand for the transmitting and receiving array responses, respectively. When half-wavelength spaced ULAs are employed at both ends, $\boldsymbol{a}_t(\cdot)$ and $\boldsymbol{a}_r(\cdot)$ can be written as

$$\boldsymbol{a}_t(\phi) = \frac{1}{\sqrt{M}}[1, e^{j\pi \sin\phi}, \cdots, e^{j(M-1)\pi \sin\phi}]^T \tag{4a}$$

$$\boldsymbol{a}_r(\theta) = \frac{1}{\sqrt{N}}[1, e^{j\pi \sin\theta}, \cdots, e^{j(N-1)\pi \sin\theta}]^T. \tag{4b}$$

At the BS, we define $\mathcal{N} = \{n_1, n_2, \cdots, n_K\}$, which contains the selected indices of the combing beams from the N-dimensional FFT matrix $\boldsymbol{F}_N$. Accordingly, the function of SN at the BS end can be expressed as

$$\boldsymbol{S}_{\mathcal{N}} = [\boldsymbol{e}_N(n_1), \boldsymbol{e}_N(n_2), \cdots, \boldsymbol{e}_N(n_K)]. \tag{5}$$

After analog combing, the signal to be detected in the digital baseband is given by

$$\boldsymbol{y} = \boldsymbol{S}_{\mathcal{N}}^* \boldsymbol{F}_N^* \boldsymbol{H} \boldsymbol{F}_M \boldsymbol{S}_{\mathcal{M}} \boldsymbol{s} + \boldsymbol{\xi} \tag{6}$$

where $\boldsymbol{\xi} = \boldsymbol{S}_{\mathcal{N}}^* \boldsymbol{F}_N^* \boldsymbol{n} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_K)$ remains white.

Let us now take a closer look to the effective $N \times M$ channel matrix

$$\overline{\boldsymbol{H}} = \boldsymbol{F}_N^* \boldsymbol{H} \boldsymbol{F}_M. \tag{7}$$

Note that the FFT basis is similar to the array responses shown in Eqs. (4a) and (4b). Therefore, $\overline{\boldsymbol{H}}[n,m]$ can be interpreted as the beam "path" coming from $\frac{2\pi(m-1)}{M}$ and arriving at $\frac{2\pi(n-1)}{N}$. From this perspective, $\overline{\boldsymbol{H}}$ essentially captures the channel in the "beamspace."

Proposition 1: The beam with AoA θ_p and AoD ϕ_p is mainly captured by $\overline{\boldsymbol{H}}[n,m]$, with m and n satisfying $\left|\arcsin\left(\frac{\phi_p}{2}\right) - \frac{m-1}{M}\right| \leq \frac{1}{M}$ and $\left|\arcsin(\frac{\theta_p}{2}) - \frac{n-1}{N}\right| \leq \frac{1}{N}$, respectively.

Proof: See Appendix A. ■

[1]Here, each path refers to a cluster of multipath components traveling closely in time and/or spatial domains [20].

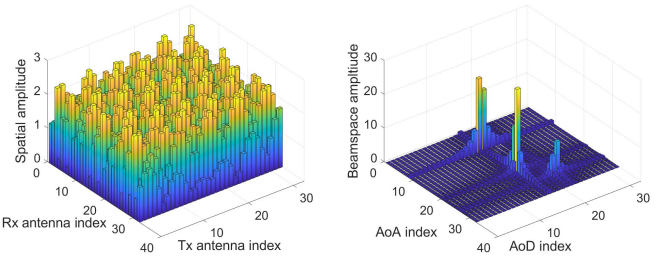

Fig. 5. The amplitude comparison between the spatial domain and beamspace.

To illustrate the capturing effect in beamspace, we randomly generate a channel with three paths, and plot the channel amplitude in the spatial domain and beamspace, respectively. It is clear that each path is localized within a small bin in the beamspace.

Given the $N \times M$ effective beamspace channel matrix $\overline{\boldsymbol{H}}$, one can potentially apply GBM directly therein. It is worth noting that though $\overline{\boldsymbol{H}}$ has the same size as $\boldsymbol{H}$, GBM on $\overline{\boldsymbol{H}}$ is fundamentally different from AG-GSM (see Section I), because GBM implements selection of beams but activate all antennas and thus exploit the full array gain. Even though this is the case, it is not wise to apply GBM directly on $\overline{\boldsymbol{H}}$. With M and N both being large, GBM directly on $\overline{\boldsymbol{H}}$ will incur high complexity and imply a huge number of RF chains. In addition, the resultant MG and error performance will both be compromised due to the sparsity in $\overline{\boldsymbol{H}}$ (see Fig. 5) caused by the limited scattering of mmWave mMIMO channels.

To this end, the SN comes as a natural help, and one can obtain the sub-beamspace as follows

$$\overline{\boldsymbol{H}}_K = \boldsymbol{S}_{\mathcal{N}}^* \overline{\boldsymbol{H}} \boldsymbol{S}_{\mathcal{M}}. \tag{8}$$

The corresponding I/O relationship accordingly becomes

$$\boldsymbol{y} = \overline{\boldsymbol{H}}_K \boldsymbol{s} + \boldsymbol{\xi}. \tag{9}$$

Evidently, the system performance heavily relies on $\overline{\boldsymbol{H}}_K$ (SNs), whose optimization will be detailed in Section IV.

At this point, it is worth emphasizing that we are simply describing the practical mmWave transceiver without any alteration, except for revealing that the lens arrays naturally project the spatial domain to the beamspace and the SNs naturally facilitate the dimensional reduction and beam selection. Next, we are going to introduce GBM that is judiciously designed for the (sub-)beamspace. As will be seen next, GBM will facilitate the multiplexing in ready-for-deployment mmWave systems under limited number of RF chains.

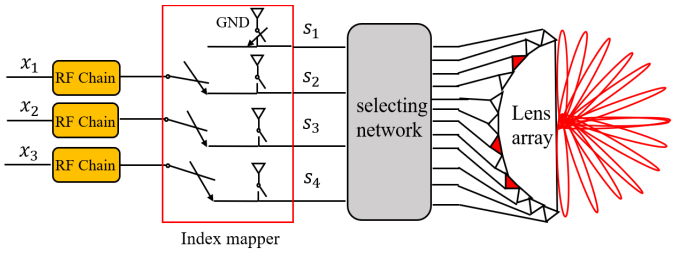

Fig. 6. An illustration of GBM modulator.

III. The Digital Part of GBM

In this section, we will focus on the digital part of GBM. The modulation along with the pattern selection is first introduced, followed by the demodulation design.

A. The GBM Modulator

The prototype of GBM modulator is shown in Fig. 6. With $R \leq K$ RF chains at Tx, there are $\eta = R\log_2 X + \lfloor \log_2(\mathbb{C}_K^R) \rfloor$ incoming bits every transmission, with $R\log_2 X$ symbol bits and $\lfloor \log_2(\mathbb{C}_K^R) \rfloor$ index bits. The symbol bits are modulated into a symbol vector $\boldsymbol{x} = [x_1, \cdots, x_R]^T$, whose elements are chosen from $\mathcal{S}$, e.g., the constellation of an X-ary phase shift keying/quadrature amplitude (PSK/QAM) modulation. In this paper, $\mathcal{S}$ is assumed to be normalized, so the signal-to-noise ratio (SNR) per bit is defined as $E_b/N_0 = R/(\eta\sigma^2)$.

Evidently, one only needs $R \leq K$ RF chains to transmit these R PSK/QAM symbols. However, it is worth emphasizing that this reduction in RF chains is not at the price of compromised SE, because the actual transmitted signal $\boldsymbol{s}$ has a higher dimension than $\boldsymbol{x}$. The conversion from $\boldsymbol{x}$ to $\boldsymbol{s}$ is realized by a $K \times R$ index mapping matrix $\boldsymbol{B}_{\mathcal{R}}$. Let $\mathcal{R}$ be a length-R lexicographical sequence, whose elements range from $[1, K]$ and are sorted in an ascending order, then $\boldsymbol{B}_{\mathcal{R}}$ is constructed as follows:

- $\forall n \notin \mathcal{R}$, $\boldsymbol{B}_{\mathcal{R}}[n,:] = \boldsymbol{0}_{1\times R}^T$
- $\boldsymbol{B}_{\mathcal{R}}[\mathcal{R},:] = \boldsymbol{I}_R$.

To make it more clear, we take $(K = 4, R = 3)$ as an example. Suppose the 1^{st}, the 2^{nd} and the 4^{th} beams are selected, i.e, $\mathcal{R} = \{1, 2, 4\}$, then $\boldsymbol{s}$ is given by

$$\boldsymbol{s} = \boldsymbol{B}_{\mathcal{R}}\boldsymbol{x} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \\ x_2 \\ x_3 \end{bmatrix}. \tag{10}$$

Each $\boldsymbol{B}_{\mathcal{R}}$ corresponds to a unique $\mathcal{R}$, which essentially represents a specific index pattern. If $\log_2 \mathbb{C}_K^R$ is an integer, all index patterns will be used. However, if $\log_2 \mathbb{C}_K^R$ is not an integer, then $\mathbb{C}_K^R - 2^{\lfloor \log_2 \mathbb{C}_K^R \rfloor}$ index patterns become redundant.

For GSM, these $2^{\lfloor \log_2 \mathbb{C}_K^R \rfloor}$ index patterns are often just arbitrarily selected or lexicographically selected via a look-up table as in [21]. For GBM, K is not entirely a design parameter but is rather dictated by the channel as we will discuss in SectionVI. In addition, the beam quality may vary significantly. Therefore, if there are redundant index patterns, pattern selection is expected to have a non-negligible influence on the overall error performance.

An algorithm is henceforth proposed herein to select the preferred index patterns. The detailed procedure is described as follows

- *Let $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ represent two index patterns,*[2] *then the pattern distance (PD) between them is defined as $d_{i,j} = |\overline{\boldsymbol{H}}_K(\boldsymbol{p}_i - \boldsymbol{p}_j)|$;*
- *Choose the index combination with the maximal minimal PD (max-MPD), and the corresponding index patterns are recognized as preferred.*

The total number of index combinations is upper-bounded by $\max\left(\mathbb{C}_K^{\lceil \frac{K}{2} \rceil}, \mathbb{C}_K^{\lfloor \frac{K}{2} \rfloor}\right)$. Since the maximum K is restricted by the low-rank mmWave channels, it is typically very small (typically less than 12 as we will discuss in Section V). Thus the additional complexity involved by the pattern selection is minimal. From simulations in Section VI, we will see that such a simple scheme can bring a noticable performance improvement.

B. The GBM Demodulator

To meet different implementing requirements, we provide two detectors, namely the ML detector and the ZF-2Q detector.

1) ML Detector: The ML criterion is expressed as

$$\widehat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}\in\mathcal{G}} \|\boldsymbol{y} - \overline{\boldsymbol{H}}_K\boldsymbol{s}\|^2 \tag{11}$$

where $\mathcal{G}$ is the ensemble containing all effective GBM vectors. Since the noise samples after RF combining remain uncorrelated, the ML detector can achieve the optimal detection performance. The overall computational complexity in terms of the number of multiplications is $O(2^\eta)$.

2) ZF-2Q: As a low-complexity alternative to the ML detector, the ZF detector is a popular option. A standard ZF detector consists of two components: i) the linear ZF filter $\overline{\boldsymbol{s}} = \overline{\boldsymbol{H}}_K^\dagger \boldsymbol{y}$; ii) the non-linear vector quantization:

$$\widehat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}\in\mathcal{G}} \|\overline{\boldsymbol{s}} - \boldsymbol{s}\|^2 . \tag{12}$$

However, in the context of IM as GBM here, the vector quantization actually induces exponential complexity $O(2^\eta)$, which is identical to ML! To this end, we propose the 2-step quantization (2Q) following the linear ZF filter.

2a: Quantization-I (Per-symbol quantization)

$$\widehat{\boldsymbol{s}}[i] = \arg\min_{s\in\mathcal{S}} |\overline{\boldsymbol{s}}[i] - s|^2. \tag{13}$$

2b: Quantization–II (Index pattern quantization)

$$\widehat{\boldsymbol{p}} = \arg\max_{\boldsymbol{p}\in\mathcal{P}} \sum_{i=1}^{K} \left(\mathbf{real}\{\overline{\boldsymbol{s}}[i]^* \times \widehat{\boldsymbol{s}}[i]\} - |\widehat{\boldsymbol{s}}[i]|^2/2\right)\boldsymbol{p}[i] \tag{14}$$

where $\mathcal{P}$ contains all preferred index patterns.

Along the lines of Page. 5 (L-C ML detector) in our earlier work [22], it can be readily proved that this 2-step quantization

[2]For a mapping matrix, its corresponding index pattern is the sum of all its columns.

is equivalent to the vector quantization given in Eq. (12), but achieves a complexity reduction from $O(2^{\eta})$ to $O(C_K^R)$. Plus the complexity of the first-step ZF equalization, the overall computational complexity of ZF-2Q is $O(C_K^R + K^3)$, which has polynomial complexity. It is worth noting that, when all beams are strictly orthogonal as the array size M and N approach infinity, no performance degradation will be incurred by ZF-2Q detector.

IV. The Analog Part of GBM

In this section, we concentrate on the analog part of GBM. Via a careful beam selection design, highly reliable communications can be guaranteed by GBM.

A. Optimizing the Size of $\overline{\boldsymbol{H}}_K$

Let MG_{GBM} denote the maximum achievable MG facilitated by mmWave mMIMO beamspace channel $\boldsymbol{H}$ in Eq (7). We have the following result:

Proposition 2: In mmWave mMIMO channels modeled as in Eq. 3, the maximum achievable MG MG_{GBM} is determined by the number of exclusively resolvable beams (ERBs) sharing no common AoA or AoD.

Proof: See Appendix B. ∎

Clearly, MG_{GBM} is upper bounded by the number of paths (P) in the channel. In conventional MIMO systems, MG_{GBM} is typically regarded as the rank of the channel matrix. The unique beamspace behavior of the mmWave mMIMO channels, together with the sparsity therein lead to Proposition 2. Though MG_{GBM} is in general not precisely equal to $rank(\overline{\boldsymbol{H}})$, they remain very close to each other. The difference of these two is induced by the finite beamspace (or angle-domain resolution of the antenna array). Such limited resolution leads to beamspace leakages which may contribute to $rank(\overline{\boldsymbol{H}})$ but does not contribute meaningful MG. As the array size M and N approach infinity, the beam resolution approaches zero and the beam leakage vanishes. In such an extreme case, one will find that both MG_{GBM} and $rank(\overline{\boldsymbol{H}})$ converge to the number of paths P. We further investigate the probability distribution of MG_{GBM} and obtain the following result.

Proposition 3: In a mmWave mMIMO channel with P spatial paths, the maximum achievable MG MG_{GBM} follows a cumulative mass function (CMF) that can be upper bounded by

$$\begin{aligned}&\mathrm{CMF}_{\mathrm{MG}}(p)\\ &\triangleq \Pr(MG_{\mathrm{GBM}} < p) \le \overline{\mathrm{CMF}}_{\mathrm{MG}}(p)\\ &= \begin{cases} \left(1-\prod_{i=0}^{p-1}\left(1-f(i)\right)\right)^{\frac{P}{p}}, & \text{if } mod(P,p)=0\\ \left(1-\prod_{i=0}^{p-1}\left(1-f(i)\right)\right)^{\lfloor \frac{P}{p} \rfloor -1} & \\ \left(1-\prod_{i=0}^{p}(1-f(i)) - \left(\sum_{i=1}^{p} f(i)\right)\prod_{i=1}^{p}\left(1-f(i)\right)\right), & \text{o.w.} \end{cases}\end{aligned} \tag{15}$$

with $f(i) = \frac{i(M+N-i)}{MN}$.

Proof: See Appendix C. ∎

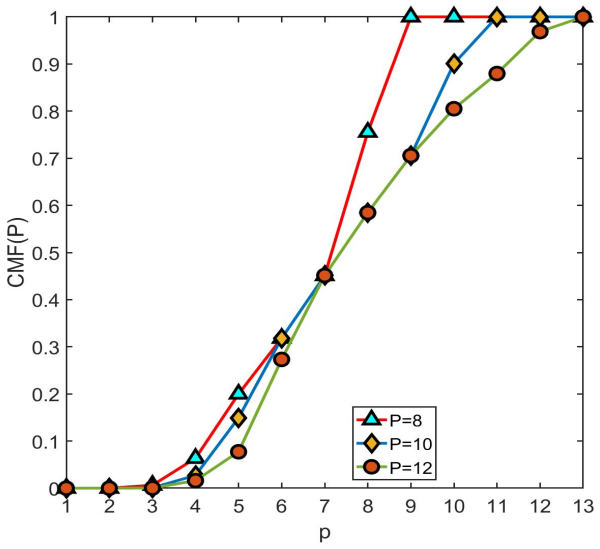

Fig. 7. The upper-bound $\mathrm{CMF}_{\mathrm{MG}}$ of the achievable MG under different number of paths.

As M and N approach infinity, $\overline{\mathrm{CMF}}_{\mathrm{MG}}(P) \to 0$, implying that $MG_{\mathrm{GBM}} \to P$. In addition, as the size of the array increases, the array resolution approaches zero and thus $rank(H) \to P$ as well. An example of $\overline{\mathrm{CMF}}_{\mathrm{MG}}(p)$ with a finite array size ($M = 32$ and $N = 64$) is presented in Fig. 7. Existing measurements show that mmWave channels typically have $8 \sim 12$ dominant paths in "rich" scattering environments [23], so we set $P = 8$, 10 and 12.

Lemma 1: The probability that the size of SNs K is no smaller than the maximum achievable MG MG_{GBM} can be lower bounded by

$$\Pr(MG_{\mathrm{GBM}} \le K) \ge 1 - \overline{\mathrm{CMF}}_{\mathrm{MG}}(K) \tag{16}$$

where $\overline{\mathrm{CMF}}_{\mathrm{MG}}(\cdot)$ has been defined in Proposition 3.

Based on Lemma 1, one can learn that choosing the size of sub-beamspace channel K too large may lead to $\overline{\boldsymbol{H}}_K$ with insufficient beams to support GBM. Whereas choosing K too small will not fully exploit MG_{GBM} facilitated by the channel. Hence given M, N and P, Eq. (15) provides a valuable guidance for choosing the size of SNs.

B. Optimizing the Entries of $\overline{\boldsymbol{H}}_K$

Even with K determined, not all $\overline{\boldsymbol{H}}_K$ candidates can support reliable communications because the beam quality may vary significantly. To optimize the entries of $\overline{\boldsymbol{H}}_K$, the immediate objective is to minimize the bit error rate (BER). Considering that the exact BER expression is mathematically intractable, we resort to the APEP. We use $\mathrm{P}_{\mathrm{GBM}}(\overline{\boldsymbol{H}}_K)$ to stand for the conditional APEP under $\overline{\boldsymbol{H}}_K$. According to [22], $\mathrm{P}_{\mathrm{GBM}}(\overline{\boldsymbol{H}}_K)$ can be approximated as

$$\mathrm{P}_{\mathrm{GBM}}(\overline{\boldsymbol{H}}_K) \approx \frac{1}{\eta 2^{\eta}} \sum_{\forall \boldsymbol{s}} \sum_{\forall \widehat{\boldsymbol{s}}} Q\left(\sqrt{\frac{d^2(\overline{\boldsymbol{H}}_K, \mathbf{s}, \widehat{\boldsymbol{s}})}{2\sigma^2}}\right) e(\boldsymbol{s}, \widehat{\boldsymbol{s}}) \tag{17}$$

where $d(\overline{\boldsymbol{H}}_K, \boldsymbol{s}, \widehat{\boldsymbol{s}}) = \| \overline{\boldsymbol{H}}_K(\boldsymbol{s} - \widehat{\boldsymbol{s}}) \|_{\mathrm{F}}$; $e(\boldsymbol{s}, \widehat{\boldsymbol{s}})$ represents the number of differing bits between $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$ after demodulation.

Existing IM-related works have shown IM manifests prominent advantages mainly at high SNR (see, e.g., [3] and [24], [25].) Therefore, we will seek a simplified form of Eq. (17) at high SNR. A proposition relating to $\mathrm{P}_{\mathrm{GBM}}(\overline{\boldsymbol{H}}_K)$ is made as follows.

Proposition 4: At high SNR, the conditional APEP $P_{\mathrm{GBM}}(\overline{\boldsymbol{H}}_K)$ can be approximated as a monotonic decreasing function of $d(\overline{\boldsymbol{H}}_K)$, where $d(\overline{\boldsymbol{H}}_K) \triangleq \min\limits_{\forall \widehat{\boldsymbol{s}} \neq \boldsymbol{s}} d(\overline{\boldsymbol{H}}_K, \boldsymbol{s}, \widehat{\boldsymbol{s}})$.

Proof: See Appendix D. ∎

Note that this monotonicity is independent of the SNR, so the original SNR-coupled APEP minimization problem can be simplified to the SNR-independent problem of $d(\overline{\boldsymbol{H}}_K)$ maximization.

Ideally, the maximal $d(\overline{\boldsymbol{H}}_K)$ can be obtained via exhaustive search among all $\mathcal{M}$'s and $\mathcal{N}$'s. However, the computational burden would be badly severe under mMIMO (over 7.8×10^{12} for $M = 64$, $N = 32$ and $K = 4$). For practical implementation, it is necessary to shrink the searching space.

Thanks to the sparsity of $\overline{\boldsymbol{H}}$, a power-based criterion can be used to screen out weak beams:

$$\mathcal{P}_1 = \left\{ (i,j) \middle| \frac{|\overline{\boldsymbol{H}}[i,j]|^2}{\max\limits_{i,j} |\overline{\boldsymbol{H}}[i,j]|^2} \geq \lambda \right\}, \tag{18}$$

where λ is a small threshold (e.g., 0.05). Although the cardinality of $\mathcal{P}_1$ is much smaller than MN, further modifications are still required to avoid two extreme cases.

1) The cardinality of $\mathcal{P}_1$ may be too large such that the searching complexity is still unacceptable.
2) The cardinality of $\mathcal{P}_1$ may be too small such that it does not provide K entries for $\overline{\boldsymbol{H}}$.

Case.1 can be addressed via a trimming procedure if the cardinality of $\mathcal{P}_1$ exceeds a certain threshold. For case.2, we can choose the first K largest and exclusive indices from $\overline{\boldsymbol{H}}$, and these selected indices are collected by $\mathcal{B}$.

Proposition 5: Define $\overline{\boldsymbol{H}}(A)$ to be the A-th ($A > K$) largest entry from $\overline{\boldsymbol{H}}$, then the searching space is given by

$$\mathcal{P} = \begin{cases} \mathcal{P}_1 \bigcup \mathcal{B}; & \mathbf{Cal}(\mathcal{P}_1) \leq A \\ \left\{ (i,j) \middle| |\overline{\boldsymbol{H}}[i,j]| \geq \overline{\boldsymbol{H}}(A) \right\} \bigcup \mathcal{B}; & \mathbf{Cal}(\mathcal{P}_1) > A. \end{cases} \tag{19}$$

The union of $\mathcal{B}$ in Eq. (19) is to guarantee the effectiveness of the searching set. Denote (n_i, m_i) as the index of the i-th selected beam, then the optimal beam indices can be obtained via

$$\left\{ (\overline{n}_1, \overline{m}_1), (\overline{n}_2, \overline{m}_2), \cdots, (\overline{n}_K, \overline{m}_K) \right\} = \underset{\forall i, (n_i, m_i) \in \mathcal{P}}{\arg\max}\ d(\overline{\boldsymbol{H}}_K, \overline{\boldsymbol{s}}). \tag{20}$$

Replacing $\mathcal{N} = (\overline{n}_1, \overline{n}_2, \cdots, \overline{n}_K)$ and $\mathcal{M} = (\overline{m}_1, \overline{m}_2, \cdots, \overline{m}_K)$ in Eq. (8), we can get $\overline{\boldsymbol{H}}_K$.

In summary, the GBM design procedures can be listed as follows:

1: Determine the number of selected beams at the transceiver SNs (K) based on MG_{GBM}.
2: Choose the number of RF chains R at the MS satisfying $R < K$;
3: Select the preferred beams based on the min-APEP criterion;
4: Trim redundant index patterns based on "max-MPD" algorithm; and
5: Choose the ML or ZF-2Q detector.

V. Analyses for GBM

To gain a better understanding of GBM and evaluate its performance, extensive analyses and discussions will be provided in this section.

A. APEP Analysis

Due to the uncertainty of the sub-beamspace and the off-grid beam leakage, it is extremely difficult to derive an exact APEP. Here we attempt to derive an APEP bound to evaluate the error performance. As the array size approaches infinity, the actual APEP will also approach the derived bound.

Let $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_K]^T$ with $\beta_i = \frac{P}{MN} \overline{\boldsymbol{H}}_K^2[i,i]$, and neglect the off-grid beam leakage. The pairwise error probability $\mathrm{P}_{\mathrm{GBM}}(\boldsymbol{s}, \widehat{\boldsymbol{s}})$ can be approximated as

$$\begin{aligned} \mathrm{P}_{\mathrm{GBM}}(\boldsymbol{s}, \widehat{\boldsymbol{s}}) &= Q\left(\sqrt{\frac{d^2(\overline{\boldsymbol{H}}_K, \mathbf{s}, \widehat{\boldsymbol{s}})}{2\sigma^2}} \right) \\ &\overset{(a)}{\approx} \mathbb{E}_{\boldsymbol{\beta}} \left\{ \frac{1}{12} \exp\left(-\sum_{i=1}^{K} MN\beta_i \Delta s_i^2 / 4P\sigma^2 \right) \right. \\ &\quad \left. + \frac{1}{4} \exp\left(-\sum_{i=1}^{K} MN\beta_i \Delta s_i^2 / 3P\sigma^2 \right) \right\} \end{aligned} \tag{21}$$

where (a) comes from $Q(x) \simeq \frac{1}{12} e^{-\frac{x^2}{2}} + \frac{1}{4} e^{-\frac{2x^2}{3}}$, which is a tight approximation of $Q(x)$ [26].

Proposition 6: At high SNR, when K beams are selected from a mmWave channel with P paths, the pairwise error probability $\mathrm{P}_{\mathrm{GBM}}(\boldsymbol{s}, \widehat{\boldsymbol{s}})$ can be approximated as

$$\begin{aligned} \mathrm{P}_{\mathrm{GBM}}(\boldsymbol{s}, \widehat{\boldsymbol{s}}) \simeq\ & \frac{P^K}{12} \prod_{i=1}^{K} \mathbb{B}\left(\frac{MN\Delta s_i^2 E_b \eta}{4PRN_0} + P + 1 - \kappa, \kappa \right) \\ & + \frac{P^K}{4} \prod_{i=1}^{K} \mathbb{B}\left(\frac{MN\Delta s_i^2 E_b \eta}{3PRN_0} + P + 1 - \kappa, \kappa \right) \end{aligned} \tag{22}$$

where κ is the diversity gain ranging from 1 to P.

Proof: See Appendix E. ∎

In practice, the selected beams are generally strong, so the actual achieved diversity gain is very unlikely close to the lower-bound. However, the off-grid leakage not only incurs interference but also influences the beam selection, preventing one from achieving the diversity upper-bound. Therefore, we can infer that the actual diversity gain should be moderately high. This will be verified by simulations. In addition, Eq. (22) also reveals that the power gain achieved by GBM is MN, which is in fact the full array gain.

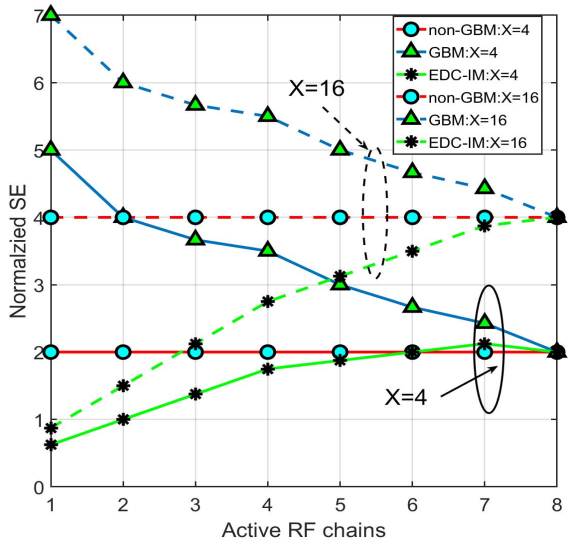

Fig. 8. The nSE comparisons among GBM, non-GBM and EDC-IM.

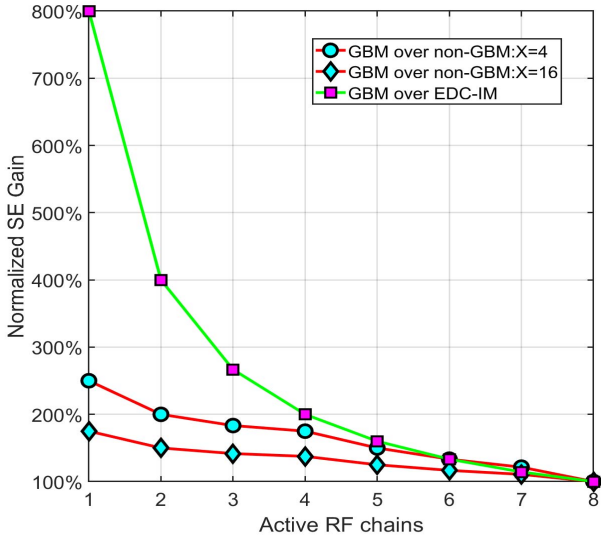

Fig. 9. The nSE gain of GBM over non-GBM and EDC-IM.

B. SE Analysis

With R and K RF chains at the Tx and Rx, respectively, the achievable SE of GBM is

$$\mathcal{C}=\left\{\lfloor \log_2 C_K^R \rfloor + R\log_2 X \middle| K \le MG_{\text{GBM}}; R<K\right\}. \quad (23)$$

To understand the advantages of GBM over other alternatives, we will next compare their SE peformances. Considering that comparing SE under different setups is somewhat "unfair," we adopt the normalized spectral efficiency (nSE) for comparison, and the nSE is defined as the ratio of the SE and the number of Tx-end RF chains. According to [27], the energy efficiency (EE) is defined as the ratio of the SE and the sum of the transmit power consumption and hardware power consumption. The consumed power is roughly proportional to the number of RF chains, so nSE can roughly imply the EE as well.

In Fig. 8, GBM is compared with the EDC-IM and the non-GBM cases in terms of the nSE. In Fig. 9, the nSE gain of GBM over EDC-IM and non-GBM is further presented. For all cases, we adopt two modulation orders: $X=4$ and $X=16$, both with $K=8$. Based on these two figures, we make the following remarks:

- *Using the same RF chains, GBM is always superior over non-GBM in terms of the nSE. This is because GBM can exploit a higher MG.*
- *Under the same MG, GBM is always superior over EDC-IM in terms of the nSE. This is because GBM requires less RF chains at the Tx.*
- *The nSE gain of GBM over non-GBM decreases with the modulation order, while the nSE gain of GBM over EDC-IM is irrelevant to the modulation order.*

C. Variants of GBM

Recall that in Fig. 3, the optional precoder is essentially set as $\boldsymbol{I}$ in GBM. In this case, the Tx needs minimal or no CSI. The only information needed are the AoD indices of the selected beams, leading to a lightweight feedback overhead of $K\log_2 M$ (as opposed to MN with full CSI). Nevertheless, GBM can have different variants by altering the precoder or other system parameters.

1) Spatial Scattering Modulation (SSM): If only one RF chain is employed at the Tx, GBM is similar to SSM [28]. However, SSM assumes (and works if and only if) there is no leakage among the selected beams, whereas GBM explicitly cope with beam leakage that is inevitable due to the finite array size.

2) Generalized Eigenspace Modulation (GEM): The precoder in Fig. 3 can be configured such that the combined digital precoding and analog SN selection can approximate the (sub-)eigenspace of the channel. By selecting different hybrid precoders in each transmission, GBM subsumes to GEM, which is the mmWave counterpart of the authors' earlier work in [29]. By optimizing the power allocation, GEM can potentially achieve an improved end-to-end mutual information. If the array size approaches infinity, GBM and GEM become identical. However, the GEM variant of the GBM comes at much increased complexity and compromised practicality in that i) frequent transmitter reconfiguration; ii) complicated power allocation and bit loading; and iii) extremely heavy feedback overhead.

3) Precoded Beamspace Modulation (PBM): The precoder in Fig. 3 can also implement pre-equalization, such as the ZF precoding, to lower the receiver complexity at the cost of increased feedback overhead of CSI.

VI. Simulations

In this section, extensive simulations are presented for an uncoded mmWave mMIMO system. The size of lens-array is set as $M=32$ at the MS and $N=64$ at the BS. Each BER curve is on the average of 20000 independent channel realizations, with a block length of 500 for each. For all figures, we set $P=12$, $\lambda=0.05$ and $A=2P$. Without a specific statement, the receiver adopts ML detector to perform demodulation. To understand the critical importance of $\overline{\boldsymbol{H}}_K$, in Fig. 10, we compare BER using different $\overline{\boldsymbol{H}}_K$ construction

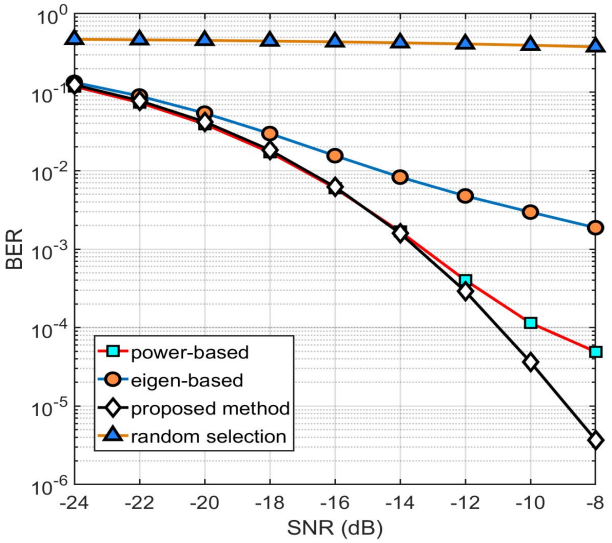

Fig. 10. BER comparisons among different sub-beamspace construction methods.

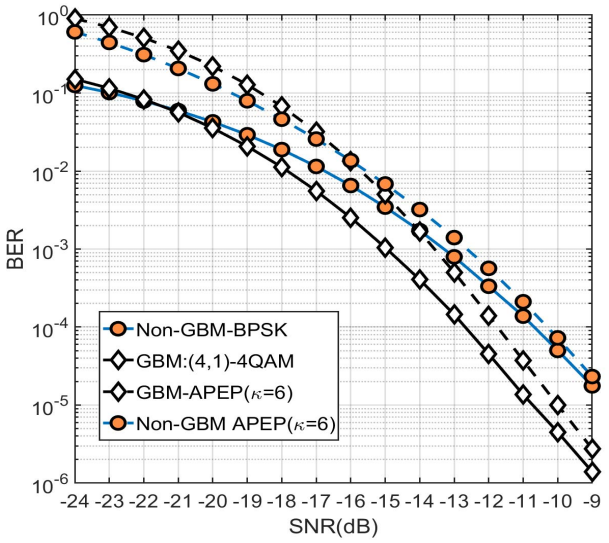

Fig. 12. BER comparisons between GBM and non-GBM under $\eta = 4$bits/Hz.

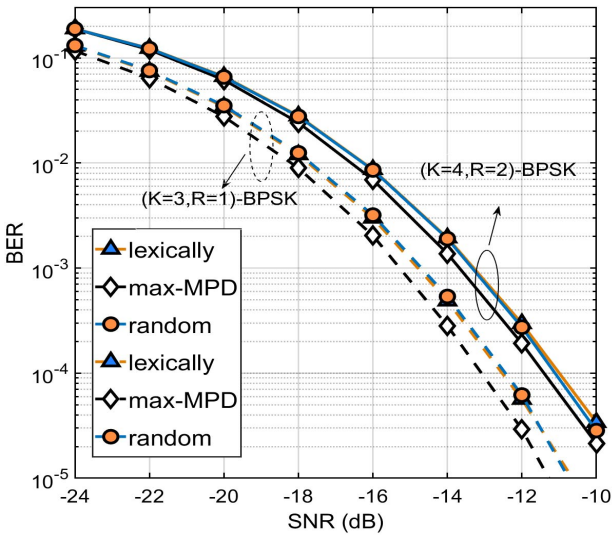

Fig. 11. BER comparisons for different index pattern selection schemes.

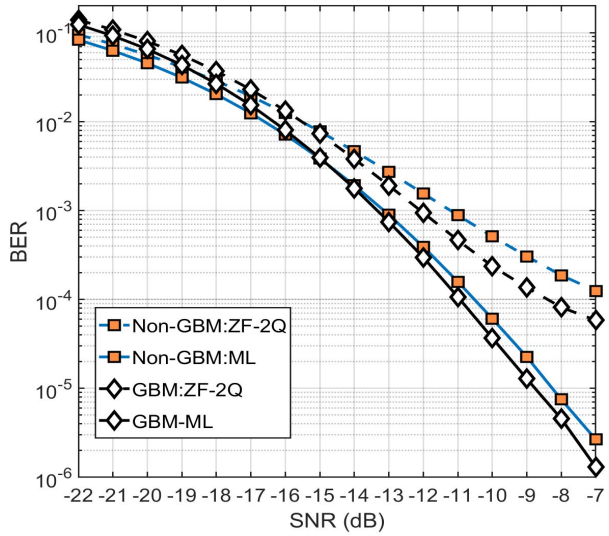

Fig. 13. BER comparisons between GBM and non-GBM under $\eta = 8$bits/Hz.

methods for non-GBM with $K = 4$ and BPSK modulation. It is clear to see that a random construction will result in an unusable system. Another two methods, namely the power-based (PB) method and eigen-based (EB) method are also provided. For PB method, those strongest beams are selected. For EB method, those transmitting/receiving beams best matching the channel left/right singular vectors (EB) will be selected. Our proposed method can achieve an optimal performance at almost whole SNR region, even though the optimality is derived at high SNR.

In Fig. 11, to validate the advantages of the proposed index pattern selection algorithm, we consider two GBM cases: ($K = 3$, $R = 1$, BPSK) and ($K = 4$, $R = 2$, BPSK), both of which have a proportion of 1/3 redundant patterns. Simulations show that with max-MPD scheme, more than 0.5dB BER advantage can be achieved over the lexicographically sequential selection (LSS) and random selection. Unlike LSS, max-MPD method requires an additional feedback of $\lceil \mathbb{C}_K^R \rceil$ bits to indicate the selected index combination, but this overhead is obviously negligible. Therefore, our method is an appealing option when index patterns have redundancy.

In Fig. 12, we compare the BER performance of GBM and non-GBM with a spectral efficiency of 4 bits/Hz. For the GBM, we set $(K = 4, R = 1)$ with 4-QAM modulation. To achieve the same spectral efficiency, the non-GBM uses BPSK modulation. With an increase of SNR, the BER advantage of GBM over non-GBM gradually becomes noticeable. At high SNR, the BER advantage is over 2dB. This advantages owe to a large ratio (50%) of index bits, which are more robust compared to the index bits in high SNR region. Furthermore, we observe that the BER curves of GBM and non-GBM can be well described by Eq. (22), and the diversity gain of GBM is slightly larger than that of non-GBM. Besides, the actual

diversity gain of both systems is about 6 ($1 < 6 < 12$), which is also consistent with our previous analysis.

In Fig. 13, we compare the BER performance with a higher system spectral efficiency. ($K = 4, R = 3$, 4-QAM) is set for GBM. To get the same spectral efficiency (8 bits/Hz), the non-GBM adopts 4-QAM modulation. At low SNR, GBM is inferior to non-GBM, as the index bits are vulnerable to be wrongly detected. However, GBM soon outperforms non-GBM with a small increase of SNR. Even though the ratio of index bits is only 25%, the BER advantage is still over 0.5dB combined with a 33% nSE enhancement. In addition to the ML detector, the BER performance using ZF-2Q detector is also presented here. Although the detection complexity has been largely reduced, the performance gap compared to the ML detector is also notable. However, when both adopt ZF-2Q detector, we find the advantage of GBM over non-GBM is about 1.5dB. This signifies the GBM can partially compensate the performance loss arising from the suboptimal equalizer. In the future, it is of great significance to design a near-optimal detector for GBM, which can constitute a better tradeoff between the complexity and BER performance.

VII. Conclusions

By exploiting the unique features of mmWave channels and the hybrid transceiver structures, a novel IM design termed as GBM has been judiciously devised for mmWave mMIMO. Under limited RF chains, GBM can achieve improved MG and SE, without compromising the array gain or the compatibility with prevalent mmWave mMIMO systems. Based on the (sub-)beamspace, a complete GBM transceiver is designed and optimized from the digital and analog parts. Extensive theoretical analyses and numerical simulations have demonstrated the remarkable advantages of GBM over non-GBM alternatives in terms of both the BER and SE.

Appendix A
Proof of Proposition 1

Without loss of generality, we take the p-th path as an example. The 2-D beamforming gain at grid point $[n, m]$ is $\boldsymbol{F}_N^*[:, n]\boldsymbol{a}_r(\theta_p) \times \boldsymbol{a}_t^*(\phi_p)\boldsymbol{F}_M[:, m]$. Due to the symmetry, we consider the left part only, which is calculated as

$$\boldsymbol{a}_t^*(\phi_p)\boldsymbol{F}_M[:,m]=\frac{1}{M}\left|\frac{\sin\pi M\big(\arcsin\big(\frac{\phi_p}{2}\big)-\frac{m-1}{M}\big)}{\sin\pi\big(\arcsin\big(\frac{\phi_p}{2}\big)-\frac{m-1}{M}\big)}\right|. \quad (24)$$

The main-lobe of Eq. (24) is limited to $\big|\arcsin\big(\frac{\phi_p}{2}\big)-\frac{m-1}{M}\big| \le \frac{1}{M}$, and the beamforming gain decreases rapidly in large M. Similarly, the main-lobe of the BS-end beamforming gain is limited to $\big|\arcsin\big(\frac{\theta_p}{2}\big)-\frac{n-1}{N}\big| \le \frac{1}{N}$. Both M and N are quite large in mmWave mMIMO, thus the 2-D beamforming gain at $[m, n]$ tends to be negligible for either $\big|\arcsin\big(\frac{\phi_p}{2}\big)-\frac{m-1}{M}\big| > \frac{1}{M}$ or $\big|\arcsin(\frac{\theta_p}{2})-\frac{n-1}{N}\big| > \frac{1}{N}$.

Appendix B
Proof of Proposition 2

The sufficiency is easy to be verified thus being omitted, so we focus on the necessity only. When less than K exclusive elements exist in beamspace, at least one common column (row) will be shared by $\boldsymbol{S}_{\mathcal{N}}$ (or $\boldsymbol{S}_{\mathcal{M}}$), leading to two columns (rows) in $\overline{\boldsymbol{H}}_K$ linearly dependent. Therefore, the rank of $\overline{\boldsymbol{H}}_K$ will be smaller than K.

Appendix C
Proof of Proposition 3

Let $p(a, b)$ represent the probability that at least b out of $a(b \le a)$ entries are exclusive, then one can readily get

$$p(m,m)=\prod_{i=0}^{m-1}\left(1-\frac{i(M+N-i)}{MN}\right).$$

$p(m+1, m)$ includes two cases: all $m+1$ entries are exclusive, and only m entries are exclusive, thus $p(m+1, m)$ is lower-bounded by

$$\begin{aligned}p(m+1,m) \ge\ & p(m+1,m+1)\\ &+\left(\sum_{i=1}^{m}\frac{i(M+N-i)}{MN}\right)\\ &\times\prod_{i=1}^{m}\left(1-\frac{i(M+N-i)}{MN}\right).\end{aligned}$$

If $\lfloor\frac{a}{b}\rfloor = 1$, it is clear that $P(a, b) \ge p(\min(b+1, a), b)$. Thus when $\lfloor\frac{a}{b}\rfloor > 1$, if $\mod(a, b) = 0$, one can get $p(a, b) \ge 1-(1-p(b, b))^{\lfloor\frac{a}{b}\rfloor}$; if $\mod(a, b) \ne 0$, one can get $p(a, b) \ge (1 - p(b+1, b))(1 - p(b, b))^{\lfloor\frac{a}{b}\rfloor-1}$.

Appendix D
Proof of Proposition 4

Let $d(\overline{\boldsymbol{H}}_K) = \min_{\forall\widehat{\boldsymbol{s}}\ne\boldsymbol{s}} d(\overline{\boldsymbol{H}}_K, \boldsymbol{s}, \widehat{\boldsymbol{s}})$. By using the following result in [30] inductively

$$a_1Q(x_1) + a_2Q(x_2) \simeq aQ(\min(x_1, x_2))$$

with

$$a=\begin{cases}a_\mu, & \mu=\arg\min\limits_{i=1,2}x_i;\ \textit{if } x_1\ne x_2\\ 2; & \textit{if } x_1=x_2\end{cases}$$

at high SNR, $\mathrm{P}_{\mathrm{GBM}}$ in Eq. (17) can be approximated as

$$\mathrm{P}_{\mathrm{GBM}}(\overline{\boldsymbol{H}}_K) \simeq \tfrac{1}{\eta 2^\eta}Q\big(\sqrt{d(\overline{\boldsymbol{H}}_K)/2\sigma^2}\big)\mathcal{C}$$

where $\mathcal{C} = \sum_{\forall\boldsymbol{s}}\sum_{\forall\widehat{\boldsymbol{s}}} e(\boldsymbol{s}, \widehat{\boldsymbol{s}})\mathbf{I}(d(\overline{\boldsymbol{H}}_K, \boldsymbol{s}, \widehat{\boldsymbol{s}}) == d(\overline{\boldsymbol{H}}_K))$. Hence P_{GBM} can be approximated as a monotonically decreasing function of $d(\overline{\boldsymbol{H}}_K)$.

Appendix E
Proof of Proposition 6

Since the amplitude of each path obeys a complex norm distribution, its square obeys a unit exponential distribution. The distribution of β_i corresponding to the strongest path is $f(\beta_i) = P(1 - e^{-\beta_i})^{P-1}e^{-\beta_i}$. Denote $\mathcal{C} = \frac{E_b MN\eta}{N_0 P}$, and the first item in Eq. (21) can be bounded by

$$\frac{1}{12}\prod_{i=1}^{K}\int_0^\infty f(\beta_i)e^{-\mathcal{C}\beta_i\frac{\Delta s_i^2}{4}}d\beta_i \overset{(b)}{=} \frac{P^K}{12}\prod_{i=1}^{K}\mathbb{B}\big(\frac{\mathcal{C}\Delta s_i^2}{4R}+1, P\big)$$

where (b) is according to Eq.(3.251) in [31]. Thus lower-bound of Eq. (21) is derived as

$$\frac{P^K}{12}\prod_{i=1}^{K}\mathbb{B}\left(\frac{\mathcal{C}\Delta s_i^2}{4R}+1,P\right)+\frac{P^K}{4}\prod_{i=1}^{K}\mathbb{B}\left(\frac{\mathcal{C}\Delta s_i^2}{3R}+1,P\right).$$

For the weakest path, $f(\beta_i)=Pe^{-P\beta_i}$. Following the same procedure, the upper-bound of Eq. (21) can be derived as

$$\frac{P^K}{12}\prod_{i=1}^{K}\mathbb{B}\left(\frac{\mathcal{C}\Delta s_i^2}{4R}+P,1\right)+\frac{P^K}{4}\prod_{i=1}^{K}\mathbb{B}\left(\frac{\mathcal{C}\Delta s_i^2}{3R}+P,1\right).$$

At high SNR, it can be verified that

$$\begin{aligned}&\prod_{i=1}^{K}\mathbb{B}\left(\frac{\mathcal{C}\Delta s_i^2}{4R}+P+1-\kappa,\kappa\right)\\&\quad\simeq\prod_{i=1}^{K}\frac{(\mathcal{C}\Delta s_i^2/4R+P-\kappa)!(\kappa-1)!}{(P+\mathcal{C}\Delta s_i^2/4R)!}\\&\quad\simeq\mathcal{M}_0\left(\frac{E_b}{N_0}\right)^{-\kappa}+o\left\{\left(\frac{E_b}{N_0}\right)^{-\kappa}\right\}\end{aligned}$$

where $\mathcal{M}_0$ is a constant irrelevant to E_b/N_0. Therefore κ represents the diversity gain, whose lower-bound and upper-bound is 1 and P, respectively. Thus, the APEP can be described as the form of Eq. (22).

REFERENCES

[1] S. Gao, X. Cheng, and L. Yang, "Generalized beamspace modulation for mmWave MIMO," in *Proc. Global Telecommun. Conf.*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[2] Y. Bian, X. Cheng, M. Wen, L. Yang, H. V. Poor, and B. Jiao, "Differential spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 64, no. 7, pp. 3262–3268, Jul. 2015.

[3] M. Di Renzo, H. Haas, A. Ghrayeb, S. Sugiura, and L. Hanzo, "Spatial modulation for generalized MIMO: Challenges, opportunities and implementation," *Proc. IEEE*, vol. 102, no. 1, pp. 56–103, Jan. 2014.

[4] P. Yang, M. Di Renzo, Y. Xiao, S. Li, and L. Hanzo, "Design guidelines for spatial modulation," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 6–26, 1st Quart., 2015.

[5] M. Zhang, M. Wen, X. Cheng, and L. Yang, "Pre-coding aided differential spatial modulation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[6] W. Qu, M. Zhang, X. Cheng, and P. Ju, "Generalized spatial modulation with transmit antenna grouping for massive MIMO," *IEEE Access*, vol. 5, pp. 26798–26807, 2017.

[7] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2228–2241, Jul. 2008.

[8] A. Younis, N. Serafimovski, R. Mesleh, and H. Haas, "Generalised spatial modulation," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2010, pp. 1498–1502.

[9] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[10] C. Chen, Y. Dong, X. Cheng, and L. Yang, "Low-resolution PSs based hybrid precoding for multiuser communication systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6037–6047, Jul. 2018.

[11] X. Gao, L. Dai, S. Han, C.-L. I, and X. Wang, "Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6010–6021, Sep. 2017.

[12] Y. Dong, C. Chen, and Y. Jin, "Joint beamforming with low-resolution PSs for millimetre-wave communications," *Electron. Lett.*, vol. 52, no. 18, pp. 1541–1543, Sep. 2016.

[13] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.

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

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

[16] 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.

[17] L. He, J. Wang, and J. Song, "Spatial modulation for more spatial multiplexing: RF-chain-limited generalized spatial modulation aided mm-wave MIMO with hybrid precoding," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 986–998, Mar. 2018.

[18] S. Gao, Y. Dong, C. Chen, and Y. Jin, "Hierarchical beam selection in mm wave multiuser MIMO systems with one-bit analog phase shifters," in *Proc. IEEE 8th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Yangzhou, China, Oct. 2016, pp. 1–5.

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

[20] S. Sun and T. S. Rappaport, "Millimeter wave MIMO channel estimation based on adaptive compressed sensing," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2017, pp. 47–53.

[21] B. Zheng, X. Wang, M. Wen, and F. Chen, "Soft demodulation algorithms for generalized spatial modulation using deterministic sequential Monte Carlo," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3953–3967, Jun. 2017.

[22] S. Gao, M. Zhang, and X. Cheng, "Precoded index modulation for multi-input multi-output OFDM," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 17–28, Jan. 2018.

[23] C. Gustafson, K. Haneda, S. Wyne, and F. Tufvesson, "On mm-wave multipath clustering and channel modeling," *IEEE Trans. Antennas Propag.*, vol. 62, no. 3, pp. 1445–1455, Mar. 2014.

[24] X. Cheng, M. Zhang, M. Wen, and L. Yang, "Index modulation for 5G: Striving to do more with less," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 126–132, Apr. 2018.

[25] M. Wen, X. Cheng, and L. Yang, *Index Modulation for 5G Wireless Communications*. Cham, Switzerland: Springer, 2017.

[26] E. Basar, U. Aygölü, E. Panayirci, and H. V. Poor, "Orthogonal frequency division multiplexing with index modulation," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5536–5549, Nov. 2013.

[27] 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.

[28] Y. Ding, K. J. Kim, T. Koike-Akino, M. Pajovic, P. Wang, and P. Orlik, "Spatial scattering modulation for uplink millimeter-wave systems," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1493–1496, Jul. 2017.

[29] J. Li, M. Wen, M. Zhang, and X. Cheng, "Virtual spatial modulation," *IEEE Access*, vol. 4, pp. 6929–6938, 2016.

[30] T. P. Do, J. S. Wang, I. Song, and Y. H. Kim, "Joint relay selection and power allocation for two-way relaying with physical layer network coding," *IEEE Commun. Lett.*, vol. 17, no. 2, pp. 301–304, Feb. 2013.

[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Elsevier, 2007.

Shijian Gao received the B.Sc. and M.Sc. degrees in electrical engineering from Nankai University and Peking University in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA. His research interests are in the areas of wireless communications and related fields.

Xiang Cheng (S'05–M'10–SM'13) received the Ph.D. degree from Heriot-Watt University and The University of Edinburgh, Edinburgh, U.K., in 2009, where he received the Postgraduate Research Thesis Prize. He is currently a Professor with Peking University. His general research interests are in areas of channel modeling, wireless communications, and data analytics, subject on which he has published more than 200 journal and conference papers, five books, and holds six patents. He was a recipient of the IEEE Asia Pacific (AP) Outstanding Young Researcher Award in 2015, the co-recipient for the 2016 IEEE JSAC Best Paper Award: Leonard G. Abraham Prize, the NSFC Outstanding Young Investigator Award, the both First-Rank and Second-Rank Award in Natural Science, Ministry of Education in China. He has also received the Best Paper Awards at IEEE ITST'12, ICCC'13, ITSC'14, ICC'16, ICNC'17, GLOBECOM'18, ICCS'18, and ICC'19. He has served as the Symposium Leading-Chair, the Co-Chair, and a member of the Technical Program Committee for several international conferences. He is currently an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and *Journal of Communications and Information Networks*, and an IEEE Distinguished Lecturer.

Liuqing Yang (S'02–M'04–SM'06–F'15) received the Ph.D. degree from the University of Minnesota, Minneapolis, MN, USA, in 2004. Her main research interests include communications and signal processing. She received the Office of Naval Research Young Investigator Program Award in 2007, the National Science Foundation Career Award in 2009, the IEEE GLOBECOM Outstanding Service Award in 2010, the George T. Abell Outstanding Mid-Career Faculty Award, and the Art Corey Outstanding International Contributions Award from CSU in 2012 and 2016, respectively, and Best Paper Awards at IEEE ICUWB'06, ICCC'13, ITSC'14, GLOBECOM'14, ICC'16, WCSP'16, GLOBECOM'18, ICCS'18, and ICC'19. She has been actively serving in the technical community, including the organization of many IEEE international conferences, and on the editorial boards of a number of journals, including the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING. She is currently serving as the Editor-in-Chief for *IET Communications.*