Hybrid Precoding based on Tensor Decomposition for mmWave 3D-MIMO Systems

Lu Liu and Yafei Tian
School of Electronics and Information Engineering, Beihang University, Beijing, China
Email: {luliu, ytian}@buaa.edu.cn

***Abstract*—Millimeter wave (mmWave) communication is a significantly enabling technology in next generation cellular system. Combined with massive number of antennas, the throughput can be greatly improved but the computation complexity and power consumption can also be incredibly high. In this paper, we study the hybrid precoding design in mmWave three-dimensional (3D) massive multiple-input multiple-output (MIMO) systems. To exploit the characteristic of planar antenna arrays, we represent the 3D-MIMO channel response with tensor, and find the null space of interference users with tensor decomposition. The null space can be well approximated by the Kronecker product of azimuth and elevation directional array vectors, and thus the designed analog precoder can eliminate inter-user interference. Combined with the baseband digital precoding, which find the maximal projection direction of the desired channel on the null space, the conventional zero-forcing block-diagonalization (ZF-BD) precoding method is extended to tensor context with constant-modulus constraint of null space elements. Since there are massive antennas and only limited RF chains, the proposed method has larger freedom to suppress interference. Simulation results verify its superiority.**

***Index Terms*—3D-MIMO, hybrid precoding, massive-MIMO, mmWave communication, tensor decomposition**

I. Introduction

One of the primary technique in the fifth generation (5G) cellular system is millimeter wave (mmWave), which is the spectral frontier for wireless communication systems nowadays [1]. The propagation characteristic of mmWave channel is different from the microwave channel. Due to the larger penetration loss, less scattering and diffraction, mmWave channel is more sparse both in time domain and space domain. Hence, mmWave communication depends more extensively on highly directional transmission.

Since the antennas in mmWave system will be numerous, the complexity of computation and the consumption of device power such as radio frequency (RF) are especially high, which means that the full digital precoding is not practical. As an alternative, the hybrid precoding structure is generally used, where the analog precoder conducts phase-only precoding with low-complexity phase shifters, and the baseband precoder adjusts both the amplitude and phase with full-functional RF chains [1, 2].

In typical hybrid precoding designs, such as in [3–5], the analog precoder is used to improve the signal power, and the digital precoder is used to suppress inter-user interferences. More sophisticated algorithm has been proposed in [6], where the analog precoder is optimized on manifold so that the achievable rate of hybrid precoding can approach that of full digital precoding. However, the algorithm needs alternating minimization between two precoders and the computational burden is high.

Considering the massive antenna number in future cellular systems, placing antenna arrays in two dimensional grid has been proposed, where the planar array can form beams in both the azimuth and elevation dimensions, it is thus called three-dimensional (3D) multiple-input multiple-output (MIMO) system [7]. The precoding in 3D-MIMO systems might be simplified if we deal with the azimuth and elevation dimensions separately. In [8], the authors put forward that a Kronecker production of the azimuth and elevation correlations can well approximate the 3D-MIMO channel correlation matrix. In [9], a more direct approach has been taken to show the Kronecker production relationship between azimuth and elevation channels by decomposing the channel vector. A multi-layer precoding solution for 3D-massive-MIMO systems is proposed in [10] that the channel characteristics in the elevation direction is leveraged to manage inter-cell interferences. However, these works did not consider hybrid precoding architecture, and only single data stream is transmitted for each user.

The channel response of a 3D-massive-MIMO system can be represented by a large matrix, as the conventional representation of MIMO channels. But considering its unique feature, it is more natural to describe it with a tensor [11]. Then the channel in azimuth and elevation dimensions can be thought as slices of the tensor with different modes, and tensor decomposition can be used to analyze the signal subspace and null subspace in different dimensions [12]. In this paper, we first derive a null space representation of the channel tensor, and then extend the conventional matrix-based zero-forcing block-diagonalization (ZF-BD) precoding method [13] to the tensor context. The null space of interference users can be expressed by the combination of a series of Kronecker product between horizontally and vertically dimensional channel vectors, and these vectors can be well approximated by the azimuth and elevation directional antenna array vectors. In this way, we find an analog precoding method which can suppress multi-user interference and inter-cell interference, and has larger design freedom in the situation of limited RF chains. Combined with the baseband digital precoding, which finds the maximal projection direction in the null space, a new kind of hybrid precoding design methodology is proposed. The roles of outer layer analog precoding and inner layer digital precoding are conversed with conventional settings.

Notations: Scalars are denoted by lowercase letters, vectors by lowercase boldface letters, matrices by uppercase boldface letters, and higher-order tensors by calligraphic letters. $(\cdot)^T$, $(\cdot)^H$ and $(\cdot)^*$ are respectively the transpose, Hermitian and conjugate operation. The Kronecker, outer, and n-mode products are denoted by the symbols $\otimes$, $\circ$, and $\times_n$, respectively.

II. System Model and Tensor Prerequisites

A. System model

We consider a multiple-user (MU)-MIMO downlink system, where a BS serves K users. The BS is equipped with N_t antennas, and the user k is equipped with $N_{r,k}$ antennas. The total number of antennas at all users is $N_r = \sum_{k=1}^{K} N_{r,k}$. The transmit signal of user k is denoted by $\mathbf{s}_k \in \mathbb{C}^{L_k}$, where L_k is the number of streams. $\mathbf{F}_k \in \mathbb{C}^{N_t \times L_k}$ denotes the precoding matrix of user k. The channel matrix from the BS to the user k is denoted by $\mathbf{H}_k \in \mathbb{C}^{N_{r,k} \times N_t}$. The received signal of user k can be written as

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{F}_k \mathbf{s}_k + \sum_{l=1, l\neq k}^{K} \mathbf{H}_k \mathbf{F}_l \mathbf{s}_l + \mathbf{n}_k. \tag{1}$$

In (1), the first term is the expected signal, the second term is interference, and $\mathbf{n}_k \in \mathbb{C}^{N_{r,k}}$ is the additive complex Gaussian noise. The stacked channel and precoding matrices $\mathbf{H}_S$ and $\mathbf{F}_S$ for all users can be defined as

$$\mathbf{H}_S = \begin{bmatrix} \mathbf{H}_1^T & \mathbf{H}_2^T & \cdots & \mathbf{H}_K^T \end{bmatrix}^T, \tag{2}$$

$$\mathbf{F}_S = \begin{bmatrix} \mathbf{F}_1 & \mathbf{F}_2 & \cdots & \mathbf{F}_K \end{bmatrix}. \tag{3}$$

B. Channel model

Due to the severe path loss, mmWave environment is well characterized by a clustered channel model, i.e., the Saleh-Valenzuela model [14]. In order to facilitate expression, the index k, which means the k-th user, is omitted in this subsection. Then, the channel matrix between the BS and one user is defined as

$$\mathbf{H} = \sqrt{\frac{N_t N_r}{N_{cl} N_{ray}}} \sum_{i=1}^{N_{cl}} \sum_{l=1}^{N_{ray}} \alpha_{il} \mathbf{a}_r(\phi_{il}^r, \theta_{il}^r) \mathbf{a}_t^H(\phi_{il}^t, \theta_{il}^t). \tag{4}$$

In (4), $N_t = N_{th} \times N_{tv}$ is the number of transmitter antennas on the BS, and $N_r = N_{rh} \times N_{rv}$ is the number of receiver antennas on the user end. N_{xh} and N_{xv} represent the number of antenna elements on the horizontal and vertical dimensions respectively, where $x \in \{t, r\}$ denotes the transmitter or receiver. When $N_{xh} = 1$ or $N_{xv} = 1$, the array is linear, otherwise it is a uniform planar array (UPA) shown as the structure in Fig. 1. N_{cl} and N_{ray} denote the number of clusters and the number of rays in each cluster. Generally, all of the clusters are assumed to be uniformly distributed, while the rays in one cluster follow Laplace distribution in their own angle spread. α_{il} represents the gain of the l-th ray in the i-th cluster. We suppose that it is i.i.d. and follows the distribution $\mathcal{CN}(0, \sigma_{\alpha,i}^2)$, where $\sigma_{\alpha,i}^2$ is the average power of the i-th cluster.

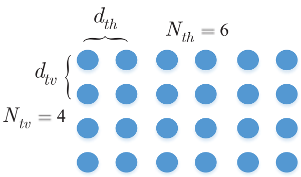

Fig. 1. An example of a $N_{tv} = 4, N_{th} = 6$ UPA at the BS.

The vector $\mathbf{a}_x(\phi_{il}^x, \theta_{il}^x)$ is the array response, in which ϕ_{il}^x is the azimuth angle and θ_{il}^x is the elevation angle. The angles with superscript t denote angles of departure (AoDs) and that with superscript r denote angles of arrival (AoAs). The array response vector can be formulated as

$$\begin{aligned} \mathbf{a}_x(\phi_{il}^x, \theta_{il}^x) = & \frac{1}{\sqrt{N_x}} [1, \mathrm{e}^{-jv_{il}^x}, \cdots, \mathrm{e}^{-j(N_{xh}-1)v_{il}^x}]^T \\ & \otimes [1, \mathrm{e}^{-ju_{il}^x}, \cdots, \mathrm{e}^{-j(N_{xv}-1)u_{il}^x}]^T \\ = & \frac{1}{\sqrt{N_x}} \mathbf{a}_{x,H}(v_{il}^x) \otimes \mathbf{a}_{x,V}(u_{il}^x), \end{aligned} \tag{5}$$

where $\mathbf{a}_{x,V}(u_{il}^x)$ and $\mathbf{a}_{x,H}(v_{il}^x)$ are the subarray vectors on the vertical and horizontal dimensions, respectively, and u_{il}^x and v_{il}^x are corresponding phase difference between adjacent elements [15].

$$u_{il}^x = \frac{2\pi d_{xv}}{\lambda} \cos \theta_{il}^x, \tag{6}$$

$$v_{il}^x = \frac{2\pi d_{xh}}{\lambda} \sin \theta_{il}^x \sin \phi_{il}^x, \tag{7}$$

where d_{xv} and d_{xh} are antenna spacing on the vertical and horizontal dimensions of the transmitter ($x = t$) or the receiver ($x = r$), and λ is the signal wavelength.

C. Tensor prerequisites

For the convenience of readers, we briefly introduce some prerequisite knowledge of tensors [11, 12]. A tensor is a multidimensional array. The order of a tensor is the number of dimensions. For instance, $\mathcal{X} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ is an N-th order tensor whose elements are denoted by $x_{i_1 i_2 \cdots i_N} = [\mathcal{X}]_{i_1 i_2 \cdots i_N}$ where $i_n \in \{1, \ldots, I_n\}, n = 1, 2, \ldots, N$. Fibers are the higher order analogue of matrix rows and columns. A fiber is defined by fixing every index but one. The mode-n unfolding of $\mathcal{X}$ is denoted by $\mathbf{X}_n$ and arranges the mode-n fibers to be the columns of the matrix.

The n-mode product of $\mathcal{X}$ with a matrix $\mathbf{U} \in \mathbb{C}^{J \times I_n}$ is denoted by $\mathcal{X} \times_n \mathbf{U}$ and is of size $I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N$, whose element is

$$(\mathcal{X} \times_n \mathbf{U})_{i_1 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \cdots i_N} u_{j i_n}. \tag{8}$$

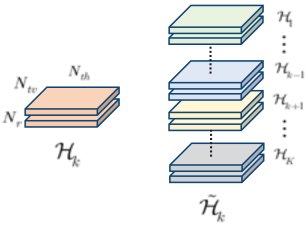

Fig. 2. The tensor description of channels.

The Tucker decomposition is a form of higher-order principal component analysis. It decomposes a tensor into a core tensor multiplied by a matrix along each mode, which is

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \cdots \times_N \mathbf{A}^{(N)} \tag{9}$$

$$= \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} g_{i_1 i_2 \cdots i_N} \mathbf{a}_{i_1}^{(1)} \circ \cdots \circ \mathbf{a}_{i_N}^{(N)} \tag{10}$$

where $\mathbf{A}^{(n)} \in \mathbb{C}^{I_n \times I_n}$, $n \in \{1, 2, \cdots, N\}$, are the factor matrices and can be thought of as the principal components in each mode. $\mathbf{a}_{i_n}^{(n)}$ is the i_nth column in the $\mathbf{A}^{(n)}$. The tensor $\mathcal{G} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ is the core tensor and the entry $g_{i_1 i_2 \cdots i_N}$ shows the level of interaction between the different components $\mathbf{a}_{i_1}^{(1)}, \cdots, \mathbf{a}_{i_N}^{(N)}$.

The matrix version of (9) is

$$\mathbf{X}_{(n)} = \mathbf{A}^{(n)} \mathbf{G}_{(n)} \Big(\mathbf{A}^{(N)} \otimes \cdots \otimes \mathbf{A}^{(n+1)} \otimes \mathbf{A}^{(n-1)} \otimes \cdots \otimes \mathbf{A}^{(1)} \Big)^T. \tag{11}$$

III. Hybrid Precoding Design based on Tensor Decomposition

In this section, we will introduce the proposed hybrid precoding scheme. For one user, the two-layer precoding structure can be written as $\mathbf{F}_k = \mathbf{F}_k^1 \mathbf{F}_k^2$, where $\mathbf{F}_k^1$ represents the outer layer analog precoding, and $\mathbf{F}_k^2$ represents the inner layer digital precoding. The focus is to design the analog precoder through tensor decomposition.

A. Tucker decomposition of channel tensor

For convenience, we assume the number of receiver antennas on each user equipment (UE) is equal to N_r, i.e., $N_r = N_{r,1} = N_{r,2} = \cdots = N_{r,K}$. We define the stacked channel matrix for all users other than user k as

$$\widetilde{\mathbf{H}}_k = \left[\mathbf{H}_1^T \quad \cdots \quad \mathbf{H}_{k-1}^T \quad \mathbf{H}_{k+1}^T \quad \cdots \quad \mathbf{H}_K^T \right]^T. \tag{12}$$

The channel matrix $\mathbf{H}_k \in \mathbb{C}^{N_r \times N_t}$ can also be viewed from a tensor perspective. Since the UPA has horizonal and vertical dimensions, the channel response of a 3D-MIMO system can be denoted as a 3rd-order tensor, i.e., $\mathcal{H}_k \in \mathbb{C}^{N_r \times N_{tv} \times N_{th}}$, as illustrated in Fig. 2. Similarly, the stacked matrix $\widetilde{\mathbf{H}}_k \in \mathbb{C}^{(K-1)N_r \times N_t}$ in (12) can be expressed as a stacked tensor

$$\widetilde{\mathcal{H}}_k \in \mathbb{C}^{(K-1)N_r \times N_{tv} \times N_{th}}. \tag{13}$$

The Tucker decompositon of $\widetilde{\mathcal{H}}_k$ can be written as

$$\widetilde{\mathcal{H}}_k = \tilde{\mathcal{G}}_k \times_1 \widetilde{\mathbf{A}}_k \times_2 \widetilde{\mathbf{B}}_k \times_3 \widetilde{\mathbf{C}}_k, \tag{14}$$

and $\widetilde{\mathcal{G}}_k \in \mathbb{C}^{(K-1)N_r \times N_{tv} \times N_{th}}$ is the core tensor, which can be computed by

$$\widetilde{\mathcal{G}}_k = \widetilde{\mathcal{H}}_k \times_1 \widetilde{\mathbf{A}}_k^H \times_2 \widetilde{\mathbf{B}}_k^H \times_3 \widetilde{\mathbf{C}}_k^H, \tag{15}$$

where $\widetilde{\mathbf{A}}_k \in \mathbb{C}^{(K-1)N_r \times (K-1)N_r}$, $\widetilde{\mathbf{B}}_k \in \mathbb{C}^{N_{tv} \times N_{tv}}$ and $\widetilde{\mathbf{C}}_k \in \mathbb{C}^{N_{th} \times N_{th}}$ are the unitary factor matrices, representing the signal space viewed from different modes of the tensor.

Tucker decomposition can be computed through high-order singular value decomposition (HOSVD) algorithm [11]. We apply $(\widetilde{\mathbf{H}}_k)_{(1)}$, $(\widetilde{\mathbf{H}}_k)_{(2)}$ and $(\widetilde{\mathbf{H}}_k)_{(3)}$ to denote the mode-1, mode-2, mode-3 unfolding of $\widetilde{\mathcal{H}}_k$ respectively. $\widetilde{\mathbf{A}}_k$, $\widetilde{\mathbf{B}}_k$ and $\widetilde{\mathbf{C}}_k$ are the left singular matrices of $(\widetilde{\mathbf{H}}_k)_{(1)}$, $(\widetilde{\mathbf{H}}_k)_{(2)}$ and $(\widetilde{\mathbf{H}}_k)_{(3)}$, respectively.

The channel matrix $\widetilde{\mathbf{H}}_k$ is equivalent to the mode-1 unfolding of $\widetilde{\mathcal{H}}_k$. According to (11), we can obtain

$$\widetilde{\mathbf{H}}_k = (\widetilde{\mathbf{H}}_k)_{(1)} = \widetilde{\mathbf{A}}_k (\widetilde{\mathbf{G}}_k)_{(1)} (\widetilde{\mathbf{C}}_k \otimes \widetilde{\mathbf{B}}_k)^T$$
$$= \sum_{t=1}^{(K-1)N_r} \sum_{i=1}^{N_{tv}} \sum_{j=1}^{N_{th}} \tilde{g}_{k_{tij}} \tilde{\mathbf{a}}_{k_t} (\tilde{\mathbf{c}}_{k_j} \otimes \tilde{\mathbf{b}}_{k_i})^T, \tag{16}$$

where $(\widetilde{\mathbf{G}}_k)_{(1)}$ is the mode-1 unfolding of $\widetilde{\mathcal{G}}_k$, and $\tilde{g}_{k_{tij}}$ is the (t, i, j)-th element of $\widetilde{\mathcal{G}}_k$. $\tilde{\mathbf{a}}_{k_t}$ is the t-th column of $\widetilde{\mathbf{A}}_k$, $\tilde{\mathbf{b}}_{k_i}$ is the i-th column of $\widetilde{\mathbf{B}}_k$, and $\tilde{\mathbf{c}}_{k_j}$ is the j-th column of $\widetilde{\mathbf{C}}_k$. Since $\widetilde{\mathbf{B}}_k$ and $\widetilde{\mathbf{C}}_k$ are all unitary matrices, the column vectors in them are orthogonal among each other, i.e.,

$$(\tilde{\mathbf{b}}_{k_v})^H \tilde{\mathbf{b}}_{k_{v'}} = (\tilde{\mathbf{b}}_{k_v})^T (\tilde{\mathbf{b}}_{k_{v'}})^* = \begin{cases} 0 & v \neq v' \\ 1 & v = v', \end{cases} \tag{17}$$

$$(\tilde{\mathbf{c}}_{k_h})^H \tilde{\mathbf{c}}_{k_{h'}} = (\tilde{\mathbf{c}}_{k_h})^T (\tilde{\mathbf{c}}_{k_{h'}})^* = \begin{cases} 0 & h \neq h' \\ 1 & h = h', \end{cases} \tag{18}$$

where $v, v' \in 1, 2, \cdots, N_{tv}$ and $h, h' \in 1, 2, \cdots, N_{th}$.

B. Null space of interference users

The optimal solution of interference management is to make all inter-user interference be zero. According to (1), it means

$$\mathbf{H}_l \mathbf{F}_k = \mathbf{H}_l \mathbf{F}_k^1 \mathbf{F}_k^2 = \mathbf{0} \quad \text{for} \quad l \neq k, \tag{19}$$

and $\mathbf{H}_S \mathbf{F}_S$ will be block diagonal (BD) [13]. Hence, in the first-layer precoding, we try to find $\mathbf{F}_k^1$ which satisfies the condition

$$\mathbf{H}_l \mathbf{F}_k^1 = \mathbf{0} \quad \text{for} \quad l \neq k. \tag{20}$$

Considering all the users, the constraint becomes to

$$[\mathbf{H}_1^T \ \cdots \ \mathbf{H}_{k-1}^T \ \mathbf{H}_{k+1}^T \ \cdots \ \mathbf{H}_K^T]^T \mathbf{F}_k^1 = \widetilde{\mathbf{H}}_k \mathbf{F}_k^1 = \mathbf{0}. \tag{21}$$

From the equation, it is obviously that $\mathbf{F}_k^1$ should lie in the null space of $\widetilde{\mathbf{H}}_k$. Hence, in the context of tensor, we should find the null space of $\widetilde{\mathcal{H}}_k$.

For a matrix, it is easy to find its rank and orthogonal basis by singular value decomposition (SVD). The null space is formed by the orthogonal basis corresponding to those zero singular values. However, in tensor field, there is no straightforward algorithm to determine the rank of a specific given tensor. For a general third-order tensor $\mathcal{X} \in \mathbb{C}^{I\times J\times K}$, only the following weak upper bound on its maximum rank is known [16]:

$$\text{rank}(\mathcal{X}) \leqslant \min\{IJ, IK, JK\}. \tag{22}$$

Hence, it is intractable to utilize the rank of a tensor to find the accurate number of linear combination of $\tilde{\mathbf{a}}_{k_t}(\tilde{\mathbf{c}}_{k_j}\otimes\tilde{\mathbf{b}}_{k_i})^T$ in (16) to represent the channel tensor, as well as the null space. For this reason, we instead resort to the element $\tilde{g}_{k_{tij}}$ of the core tensor $\tilde{\mathcal{G}}_k$. The value of $\tilde{g}_{k_{tij}}$ reflects the level of interaction among $\tilde{\mathbf{a}}_{k_t}$, $\tilde{\mathbf{b}}_{k_i}$ and $\tilde{\mathbf{c}}_{k_j}$. When $\tilde{g}_{k_{tij}} = 0$, the component $\tilde{g}_{k_{tij}}\tilde{\mathbf{a}}_{k_t}(\tilde{\mathbf{c}}_{k_j}\otimes\tilde{\mathbf{b}}_{k_i})^T$ in (16) is ineffective.

We assume that the number of

$$\left|\tilde{g}_{k_{tij}}\right| > 0 \tag{23}$$

is R. At first, compute each $|\tilde{g}_{k_{tij}}|$ and sort them in decreasing order. The tensor element indexes of $\tilde{g}_{k_{tij}}$ corresponding to the largest R values are denoted by (t_s, i_s, j_s) and $s \in \{1, 2, \cdots, R\}$.

Then, we can simplify (16) as the equation (24) at the top of next page. We define the third matrix in (24) as

$$\widetilde{\mathbf{V}}_R = [\tilde{\mathbf{c}}_{k_{j_1}}\otimes\tilde{\mathbf{b}}_{k_{i_1}} \quad \cdots \quad \tilde{\mathbf{c}}_{k_{j_R}}\otimes\tilde{\mathbf{b}}_{k_{i_R}}]^T. \tag{25}$$

Define the pair $\{j_s, i_s\}$ as the index of $\tilde{\mathbf{c}}_{k_{j_s}}\otimes\tilde{\mathbf{b}}_{k_{i_s}}$. There are some repetitive indexes in $\widetilde{\mathbf{V}}_R$. For instance, if $|\tilde{g}_{k_{123}}| > 0$ and $|\tilde{g}_{k_{223}}| > 0$, we have the column $\tilde{\mathbf{c}}_{k_3}\otimes\tilde{\mathbf{b}}_{k_2}$ and the index $\{3, 2\}$ in $\widetilde{\mathbf{V}}_R$ twice. Delete the column with the same $\{j_s, i_s\}$ and make every column unique. Then, we obtain

$$\widetilde{\mathbf{V}}_{\text{signal}} = [\tilde{\mathbf{c}}_{k_{x_1}}\otimes\tilde{\mathbf{b}}_{k_{y_1}} \;\cdots\; \tilde{\mathbf{c}}_{k_{x_d}}\otimes\tilde{\mathbf{b}}_{k_{y_d}} \;\cdots\; \tilde{\mathbf{c}}_{k_{x_D}}\otimes\tilde{\mathbf{b}}_{k_{y_D}}]^T, \tag{26}$$

where $d \in \{1, 2, \cdots, D\}$ is the index of column of $\widetilde{\mathbf{V}}_{\text{signal}}$, and $D \leq R$ is the total number of columns. The pair index $\{x_d, y_d\}$ is picked from $\{j_s, i_s\}$, which refers to the x_d-th column of $\widetilde{\mathbf{C}}_k$ and the y_d-th column of $\widetilde{\mathbf{B}}_k$.

Theorem 1. *A part of the basis for the null space of $\widetilde{\mathbf{H}}_k$ is $\widetilde{\mathbf{V}}_{null}$, where each column is*

$$(\tilde{\mathbf{c}}_{k_j}\otimes\tilde{\mathbf{b}}_{k_i})^*, \quad \textit{for } j \neq x_d \textit{ or } i \neq y_d. \tag{27}$$

That means $\widetilde{\mathbf{V}}_{null}$ is composed by $(\widetilde{\mathbf{C}}_k\otimes\widetilde{\mathbf{B}}_k)^*$ excluding the columns of $\widetilde{\mathbf{V}}^H_{signal}$.

Proof. For the sake of clarity, we define

$$\widetilde{\mathbf{V}}_{\text{null}} = \left[\tilde{\mathbf{c}}_{k_{m_1}}\otimes\tilde{\mathbf{b}}_{k_{n_1}} \;\cdots\; \tilde{\mathbf{c}}_{k_{m_e}}\otimes\tilde{\mathbf{b}}_{k_{n_e}} \;\cdots\; \tilde{\mathbf{c}}_{k_{m_E}}\otimes\tilde{\mathbf{b}}_{k_{n_E}}\right]^*, \tag{28}$$

where $e \in \{1, \cdots, E\}$ is the index of column of $\widetilde{\mathbf{V}}_{\text{null}}$, and $E = N_t - D$ is the total number of columns. From the theorem above, we know that $m_e \neq x_d$ or $n_e \neq y_d$. Further, since the pair index $\{x_d, y_d\}$ is picked from $\{j_s, i_s\}$, we can get

$$m_e \neq j_s \quad \text{or} \quad n_e \neq i_s. \tag{29}$$

Combining (24) and (28), we get the equation (30) on the top of next page. Define the product of last two matrixes in (30) as $\mathbf{\Gamma}$. Each block of $\mathbf{\Gamma}$ is

$$\left(\tilde{\mathbf{c}}_{k_{j_s}}\otimes\tilde{\mathbf{b}}_{k_{i_s}}\right)^T\left(\tilde{\mathbf{c}}_{k_{m_e}}\otimes\tilde{\mathbf{b}}_{k_{n_e}}\right)^*. \tag{31}$$

Because

$$(\mathbf{X}\otimes\mathbf{Y})^T = \mathbf{X}^T\otimes\mathbf{Y}^T, \quad (\mathbf{X}\otimes\mathbf{Y})^* = \mathbf{X}^*\otimes\mathbf{Y}^*, \tag{32}$$

$$(\mathbf{X}\otimes\mathbf{Y})(\mathbf{M}\otimes\mathbf{N}) = (\mathbf{X}\mathbf{M})\otimes(\mathbf{Y}\mathbf{N}), \tag{33}$$

and combined with (17)(18)(29), the block (31) can be converted to

$$\left(\tilde{\mathbf{c}}^T_{k_{j_s}}\tilde{\mathbf{c}}^*_{k_{m_e}}\right)\otimes\left(\tilde{\mathbf{b}}^T_{k_{i_s}}\tilde{\mathbf{b}}^*_{k_{n_e}}\right) = 0. \tag{34}$$

It follows that $\mathbf{\Gamma} = \mathbf{0}$ and

$$\widetilde{\mathbf{H}}_k\widetilde{\mathbf{V}}_{\text{null}} = \mathbf{0}. \tag{35}$$

Hence, $\widetilde{\mathbf{V}}_{\text{null}} \in N(\widetilde{\mathbf{H}}_k)$.

Similarly, we randomly choose the p-th and q-th column of $\widetilde{\mathbf{V}}_{\text{null}}$, and compute their correlation coefficient

$$\begin{aligned}&\left((\tilde{\mathbf{c}}_{k_{m_p}}\otimes\tilde{\mathbf{b}}_{k_{n_p}})^*\right)^H\left(\tilde{\mathbf{c}}_{k_{m_q}}\otimes\tilde{\mathbf{b}}_{k_{n_q}}\right)^*\\ &= \left(\tilde{\mathbf{c}}^T_{k_{m_p}}\tilde{\mathbf{c}}^*_{k_{m_q}}\right)\otimes\left(\tilde{\mathbf{b}}^T_{k_{n_p}}\tilde{\mathbf{b}}^*_{k_{n_q}}\right)\\ &= \begin{cases}0 & p\neq q\\ 1 & p = q\end{cases},\end{aligned} \tag{36}$$

where $p, q \in \{1, 2, \cdots, E\}$. It means that the columns of $\widetilde{\mathbf{V}}_{\text{null}}$ are orthogonal, and they are linearly independent of each other.

Apply the same method in (36) to (26). We will find that the columns of $\widetilde{\mathbf{V}}_{\text{signal}}$ are also orthonormal, and $\widetilde{\mathbf{V}}_{\text{signal}}$ is the unique form of $\widetilde{\mathbf{V}}_R$, so

$$\text{rank}(\widetilde{\mathbf{V}}_R) = \text{rank}(\widetilde{\mathbf{V}}_{\text{signal}}) = D. \tag{37}$$

Because the rank of matrix follows

$$\text{rank}(\mathbf{XYZ}) \leq \min\left\{\text{rank}(\mathbf{X}), \text{rank}(\mathbf{Y}), \text{rank}(\mathbf{Z})\right\}, \tag{38}$$

the rank of $\widetilde{\mathbf{H}}_k$ is

$$\text{rank}(\widetilde{\mathbf{H}}_k) \leq \text{rank}(\widetilde{\mathbf{V}}_R) = D. \tag{39}$$

Combined with $\widetilde{\mathbf{H}}_k \in \mathbb{C}^{(K-1)N_r\times N_t}$, the dimension of the null space of $\widetilde{\mathbf{H}}_k$ is

$$\dim N(\widetilde{\mathbf{H}}_k) = N_t - \text{rank}(\widetilde{\mathbf{H}}_k) \geq N_t - D, \tag{40}$$

then we can get

$$\dim N(\widetilde{\mathbf{H}}_k) \geq E. \tag{41}$$

Hence, the E columns in $\widetilde{\mathbf{V}}_{\text{null}}$ is not the full basis forming the null space of $\widetilde{\mathbf{H}}_k$. □

$$\widetilde{\mathbf{H}}_k = \sum_{s=1}^{R} \tilde{g}_{k_{t_s i_s j_s}} \tilde{\mathbf{a}}_{k_{t_s}} (\tilde{\mathbf{c}}_{k_{j_s}} \otimes \tilde{\mathbf{b}}_{k_{i_s}})^T = \Big[\tilde{\mathbf{a}}_{k_{t_1}} \ \cdots \ \tilde{\mathbf{a}}_{k_{t_R}}\Big] \mathrm{diag}\Big(\tilde{g}_{k_{t_1 i_1 j_1}} \ \cdots \ \tilde{g}_{k_{t_R i_R j_R}}\Big) \Big[\tilde{\mathbf{c}}_{k_{j_1}} \otimes \tilde{\mathbf{b}}_{k_{i_1}} \ \cdots \ \tilde{\mathbf{c}}_{k_{j_R}} \otimes \tilde{\mathbf{b}}_{k_{i_R}}\Big]^T . \tag{24}$$

$$\widetilde{\mathbf{H}}_k \widetilde{\mathbf{V}}_{\text{null}} = \Big[\tilde{\mathbf{a}}_{k_{t_1}} \ \cdots \ \tilde{\mathbf{a}}_{k_{t_R}}\Big] \mathrm{diag}\Big(\tilde{g}_{k_{t_1 i_1 j_1}} \ \cdots \ \tilde{g}_{k_{t_R i_R j_R}}\Big) \begin{bmatrix} \Big(\tilde{\mathbf{c}}_{k_{j_1}} \otimes \tilde{\mathbf{b}}_{k_{i_1}}\Big)^T \\ \vdots \\ \Big(\tilde{\mathbf{c}}_{k_{j_R}} \otimes \tilde{\mathbf{b}}_{k_{i_R}}\Big)^T \end{bmatrix} \Big[\Big(\tilde{\mathbf{c}}_{k_{m_1}} \otimes \tilde{\mathbf{b}}_{k_{n_1}}\Big)^* \ \cdots \ \Big(\tilde{\mathbf{c}}_{k_{m_E}} \otimes \tilde{\mathbf{b}}_{k_{n_E}}\Big)^*\Big] \tag{30}$$

C. Hybrid precoding design

From the expression of $\widetilde{\mathbf{V}}_{\text{null}}$, we know that it is a combination of a series of Kronecker product of vertical and horizontal factor vectors. If we use $\widetilde{\mathbf{V}}_{\text{null}}$ as the first layer precoder, it can definitely inhibit inter-user interference, but it does not satisfy the constraints for analog precoding.

To obtain precoding matrix with constant-modulus elements, we can approximate each $\tilde{\mathbf{b}}_{k_{n_e}}$ and $\tilde{\mathbf{c}}_{k_{m_e}}$ with the subarray vectors $\mathbf{a}_{t,V}(u_e^t)$ and $\mathbf{a}_{t,H}(v_e^t)$, respectively, i.e.,

$$u_e^t = \arg\max_u |\tilde{\mathbf{b}}_{k_{n_e}}^H \mathbf{a}_{t,V}(u)|, \tag{42}$$

$$v_e^t = \arg\max_v |\tilde{\mathbf{c}}_{k_{m_e}}^H \mathbf{a}_{t,H}(v)|. \tag{43}$$

Furthermore, due to the limited number of RF chains, we can not use all E columns of $\widetilde{\mathbf{V}}_{\text{null}}$ to approximate the first layer precoder, and only N_k^{RF} columns are allocated to user k. The column number of $\mathbf{F}_k^1$ should be less than N_k^{RF}. Thus we can pick N_k^{RF} columns from $\widetilde{\mathbf{V}}_{\text{null}}$ on which the channel matrix $\mathbf{H}_k$ of the desired user can have maximal projection. The finally obtained analog precoder is expressed as

$$\mathbf{F}_k^1 = \big[\mathbf{a}_{t,V}(u_1^t) \otimes \mathbf{a}_{t,H}(v_1^t), \cdots, \mathbf{a}_{t,V}(u_{N_k^{\text{RF}}}^t) \otimes \mathbf{a}_{t,H}(v_{N_k^{\text{RF}}}^t)\big]. \tag{44}$$

To obtain the second layer digital precoding matrix $\mathbf{F}_k^2$, we need first construct the equivalent channel $\mathbf{H}_k\mathbf{F}_k^1$, which is an interference-free channel. Thus we can use the technique of SU-MIMO to design $\mathbf{F}_k^2$. Actually, it is well known that the best precoding matrix for $\mathbf{F}_k^2$ is the right singular matrix of $\mathbf{H}_k\mathbf{F}_k^1$.

In this way, the new hybrid precoding method is proposed, where the outer layer analog precoder is to suppress interference and the inner layer digital precoder is to maximize transmit power. We can see that the design procedure is like the ZF-BD algorithm. But with the special constraints of analog precoding, we have resorted to tensor decomposition to obtain the approximation of null space using subarray vectors. Due to the large number of antennas, we have large freedom to inhibit inter-user interference as well as inter-cell interference, by introducing more interference channel matrices into (12).

IV. Simulation Results

In this section, we evaluate the performance of our proposed scheme in different situations and compare it with other hybrid precoding schemes. In our scenario, the BS serves K users, and each user transmit L_k data streams. The UPA is amounted with 64 antennas, where $N_{tv} = N_{th} = 8$, $d_v = d_h = \frac{1}{2}\lambda$. The number of RF chains is equal to the total number of data streams of K users. In the channel model, $N_{cl} = 8$, $N_{ray} = 10$, and each cluster's average power is $\sigma_{\alpha,i}^2 = 1$. The azimuth and elevation dimensional AOAs and AODs follow the Laplacian distribution with uniformly distributed mean angles and angular spread of 5 and 2.5 degrees.

Through the simulations, we find that sometimes there is no $\tilde{g}_{k_{tij}} = 0$ in the core tensor $\widetilde{\mathcal{G}}_k$. The reason is that when we compute the Tucker decomposition with rank-$(R_1, R_2, \cdots, R_N)$, if $R_n < \text{rank}_n(\mathcal{X})$ for one or more mode n, the decomposition will be truncated and the result does not exactly reproduce $\mathcal{X}$. To implement our proposed algorithm, we define a threshold $\eta > 0$ to replace the 0 in (23). The appropriate η defines the size of the null space of $\widetilde{\mathcal{H}}_k$.

In Fig. 3, we compare the performance of the proposed precoding scheme with the full-size digital ZF-BD, an extended orthogonal matching pursuit (OMP) scheme, and a simple beam steering scheme. The extended OMP scheme is based on the spatially sparse precoding proposed in [17], and we extend it to hybrid precoding in MU-MIMO scenarios through setting $\mathbf{F}_{opt}$ as that of ZF-BD scheme, where the analog layer maximizes signal power and the digital layer deals with interference. The simple beam steering solution just steers its data streams to the channel's best propagation paths. In this figure, we set $K = 4$ and $N_k^{\text{RF}} = L_k = 2$. As can be seen from the figure, the proposed hybrid precoding schemes outperforms the extended OMP scheme and the simple beam steering scheme. Along with the increasing of SNR, the performance gap keeps growing.

In Fig. 4, we evaluate the performance of the proposed scheme with different number of data streams, where the user number is fixed as $K = 4$ and $N_k^{\text{RF}} = L_k$. When $\text{SNR} = -10, -5$ dB, with more data streams our method can even outperforms the conventional ZF-BD scheme. Since the ZF-BD scheme is a sub-optimal solution for the precoding

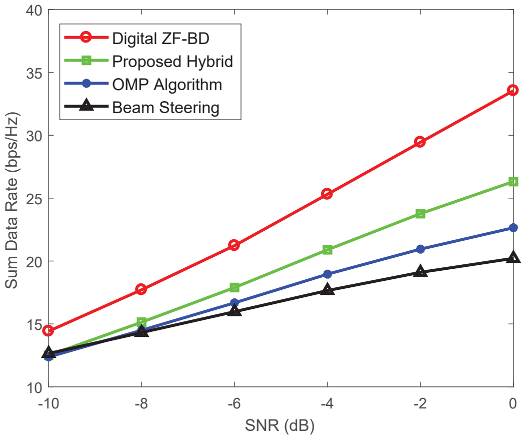

Fig. 3. Sum date rate versus SNR when $L_k = 2, K = 4$.

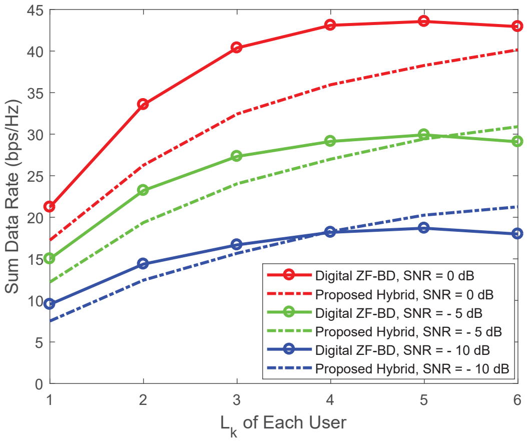

Fig. 4. Sum date rate versus streams of each user when $\text{SNR} = -10, -5, 0$ dB, where $K = 4$.

of MU-MIMO systems, the performance will decline when the number of streams increases [18]. By incorporating approximation of the null space by the subarray vectors, our method achieves a tradeoff between suppressing interference and collecting power, thus behaves better in low SNR region. When the SNR is larger, the performance of our scheme is worse than that of ZF-BD, but the gap shrinks when the system serves more data streams.

V. Conclusions

We have designed a two-layer hybrid precoding scheme for mmWave 3D-massive-MIMO systems. Exploiting the nature of planar antenna arrays, the 3D-MIMO channel response is represented by a tensor, and the null space of interference users is found through Tucker decomposition. The idea of ZF-BD precoding method is extended to accommodate the constraints of analog precoding. The null space is approximated by the Kronecker product of azimuth and elevation directional array vectors, and thus the designed analog precoder can eliminate inter-user and inter-cell interference. With limited number of RF chains, the proposed method has larger freedom for interference suppression than conventional hybrid precoding designs. Simulation results demonstrated that the performance of the proposed scheme is close to ZF-BD scheme when the number data streams grows, and exceeds other hybrid precoding method which only collect power using the analog precoder.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61371077.

References

[1] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, 2016.

[2] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, “Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[3] L. Liang, W. Xu, and X. Dong, “Low-complexity hybrid precoding in massive multiuser MIMO systems,” *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 653–656, December 2014.

[4] A. Alkhateeb, R. W. Heath, and G. Leus, “Achievable rates of multi-user millimeter wave systems with hybrid precoding,” in *Proc. of IEEE ICCW*, June 2015, pp. 1232–1237.

[5] R. A. Stirling-Gallacher and M. S. Rahman, “Multi-user MIMO strategies for a millimeter wave communication system using hybrid beam-forming,” in *Proc. of IEEE ICC*, June 2015, pp. 2437–2443.

[6] X. Yu, J. C. Shen, J. Zhang, and K. B. Letaief, “Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 485–500, April 2016.

[7] G. Liu, X. Hou, F. Wang, J. Jin, H. Tong, and Y. Huang, “Achieving 3D-MIMO with massive antennas from theory to practice with evaluation and field trial results,” *IEEE Systems Journal*, vol. 11, no. 1, pp. 62–71, March 2017.

[8] D. Ying, F. W. Vook, T. A. Thomas, D. J. Love, and A. Ghosh, “Kronecker product correlation model and limited feedback codebook design in a 3D channel model,” in *Proc. of IEEE ICC*, June 2014, pp. 5865–5870.

[9] J. Choi, K. Lee, D. J. Love, T. Kim, and R. W. Heath, “Advanced limited feedback designs for FD-MIMO using uniform planar arrays,” in *Proc. of IEEE GLOBECOM*, Dec 2015, pp. 1–6.

[10] A. Alkhateeb, G. Leus, and R. W. Heath, “Multi-layer precoding for full-dimensional massive MIMO systems,” in *Proc. of Asil. Conf. on Sig. Sys. and Comp.*, Nov 2014, pp. 815–819.

[11] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[12] N. D. Sidiropoulos, L. D. Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, “Tensor decomposition for signal processing and machine learning,” *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, July 2017.

[13] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, “Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels,” *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[14] T. S. Rappaport, R. W. Heath, R. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Pearson Education, 2014.

[15] W. Tan, S. Jin, J. Wang, and Y. Huang, “Achievable sum-rate analysis for massive mimo systems with different array configurations,” in *Proc. of IEEE WCNC*, March 2015, pp. 316–321.

[16] J. B. Kruskal, “Rank, decomposition, and uniqueness for 3-way and N-way arrays,” in *Multiway Data Analysis*. North-Holland Publishing Co., 1989, pp. 7–18.

[17] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, “Spatially sparse precoding in millimeter wave MIMO systems,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, March 2014.

[18] W. Ni and X. Dong, “Hybrid block diagonalization for massive multiuser MIMO systems,” *IEEE Transactions on Communications*, vol. 64, no. 1, pp. 201–211, Jan 2016.