IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 13, 2020, accepted May 1, 2020, date of publication May 6, 2020, date of current version May 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992713

3D Angular-Based Hybrid Precoding and User Grouping for Uniform Rectangular Arrays in Massive MU-MIMO Systems

ASIL KOC, (Student Member, IEEE), AHMED MASMOUDI, (Member, IEEE), AND THO LE-NGOC, (Life Fellow, IEEE)

Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0G4, Canada

Corresponding author: Asil Koc (asil.koc@mail.mcgill.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Development (CRD) Grant with InterDigital Canada.

ABSTRACT This paper proposes a new user grouping algorithm and three-dimensional (3D) angular-based hybrid precoding (AB-HP) scheme for massive multi-user multiple-input multiple-output (MU-MIMO) systems using uniform rectangular arrays (URA). At first, the users clustered in multiple spots are efficiently grouped according to the proposed user grouping algorithm, which only utilizes the user angle-of-departure (AoD) information and does not require prior knowledge of the number of user groups. By employing the AoD support of the user groups, the RF-beamformer of AB-HP is designed to reduce the inter-group interference, the channel state information (CSI) overhead, and the number of RF chains. Then, the digital baseband precoder of AB-HP is constructed via regularized zero-forcing (RZF) using the effective channel seen from baseband to simultaneously serve the users clustered in multiple groups, by considering three approaches: joint-group-processing (JGP), per-group-processing (PGP) and common-group-processing (CGP). For each approach, the signal-to-interference-plus-noise ratio (SINR) expressions as well as their tight deterministic approximations are derived. To further reduce the number of RF chains, we also propose a new transfer block design, which reduces the number of RF chains down to the number of independent data streams without penalizing the sum-rate performance. Illustrative results reveal that the proposed AB-HP schemes with the relaxed CSI estimation overhead and reduced hardware cost/complexity can closely approach to the sum-rate performance of the single-stage fully-digital precoding (FDP). Furthermore, AB-HP has considerably higher energy efficiency performance compared to FDP due to the reduced number of RF chains. We show through simulation that the proposed AB-HP can offer significantly better performance than existing HP techniques. The computational complexity of AB-HP is also analyzed.

INDEX TERMS Massive MIMO, 3D hybrid precoding, user grouping, angle of departure (AoD), uniform rectangular array (URA), RF chain reduction.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is one of the key technologies for the next generation wireless communication systems [3]–[6]. Massive MIMO systems adopting an excessively large number of antennas at the base station (BS) do not only provide a considerable improvement in the spectrum efficiency via the spatial multiplexing, but also dramatically enhance the energy efficiency via the capability of focusing signal energy through smaller regions (i.e., large beamforming gain) [5]. For the downlink transmission, the precoding at the BS is a crucial signal processing procedure to ensure reliable transmission quality for massive multi-user MIMO (MU-MIMO) systems. Early studies for precoding design focus on the single-stage precoders, where the single-stage fully-digital precoding (FDP) is directly designed according to the channel matrix. Hence, FDP requires full instantaneous channel state information (CSI). Furthermore, the number of radio frequency (RF) chains should be as many as the number of antennas. Among

The associate editor coordinating the review of this manuscript and approving it for publication was Abhishek Kandwal.

the linear FDP techniques, the regularized zero-forcing (RZF) precoder has been widely considered due to its near-optimal performance compared to the optimal non-linear dirty paper coding [7]–[10].

Despite the aforementioned benefits of large antenna arrays, it also brings two interesting challenges: (i) the number of RF chains increases the hardware cost/complexity as well as the power consumption and (ii) the requirement for the considerable amount of CSI overhead causes longer training sequence and increased signalling duration [11]. In order to overcome these problems, hybrid precoding (HP) has been proposed as a promising solution [11]–[13], where the precoder is partitioned into two stages: an analog RF beamformer followed by a low-dimensional digital baseband precoder. The RF-stage and baseband-stage are connected to each other via RF chains, where the selected number of RF chains is between the number of available antennas and the independent data streams or equivalently, the number of single-antenna users to be simultaneously served. In terms of input parameters for the RF beamformer design, there are two main strategies: (i) as a function of the fast time-varying instantaneous CSI [14]–[22], resulting in large CSI overhead, or (ii) using the slowly time-varying CSI (e.g., channel covariance matrix, angle-of-departure (AoD)) for the RF-stage [23]–[29], which can reduce both the hardware cost/complexity and the CSI overhead because only the effective instantaneous CSI is utilized for the baseband precoder.

A. RELATED WORKS

Among the HP schemes requiring the full CSI, a low-complexity HP technique reducing the number of RF chains while keeping the similar sum-rate performance of FDP is introduced in [14], where the RF beamformer following the constant-modulus constraint is designed via the phase of instantaneous CSI. In [15], the authors provide an algorithmic precoding solution based on the concept of orthogonal matching pursuit, which takes an optimal unconstrained FDP obtained by the full CSI and approximates it via a constrained HP. In order to find the minimum number of RF chains for HP keeping the same performance of FDP, [16] shows that when the number of RF chains is twice the total number of independent data streams, HP can realize any FDP regardless of the antenna array size. Then, [17] investigates HP design for the multi-carrier massive MIMO systems employing orthogonal frequency division multiple access transmission and proposes further reduction in hardware cost/complexity by decreasing the number of RF chains to the rank of FDP while achieving the similar sum-rate performance. In addition to decreasing the number of RF chains for massive MIMO systems, the phase-shifter reduction is also analyzed in [18], where the authors demonstrate that the sub-connected structure (each RF chain is connected to a subset of antennas) can reduce the number of phase-shifters by half and preserve the same spectral efficiency of the fully-connected structure (each RF chain is connected to all the antennas). User grouping, power allocation and HP design are investigated for non-orthogonal multiple access in millimeter-wave (mmWave) massive MIMO systems in [19], where the full CSI is not only required for designing HP but also clustering the users into different groups. Additionally, the number of user groups is also assumed to be known. In [20], an angular-based approach to develop the RF beamformer is independently adopted to establish orthogonal beam selection based HP (OBS-HP) and non-orthogonal angle space based HP (NOAS-HP) techniques, where the BS equipped with uniform linear array (ULA) utilizes RZF precoding at the baseband-stage. The users are not considered to be clustered in multiple spots in [20], hence, the orthogonalization in the angular domain among the user groups is not taken into account in that study. Moreover, the proposed significant beam selection algorithm in [20] selects the strongest beam of each user for the RF beamformer design, but it requires the complete knowledge of CSI. However, this algorithm is not capable of using all available degrees of freedom provided by the channel, especially for a rich scattering environment having wider angle spread. As stated in [20], the accuracy of the RF beamformer in OBS-HP and NOAS-HP heavily depends on the angle spread and it is shown that OBS-HP and NOAS-HP suffer from the sum-rate performance degradation when the angle spread is just increased from 1° to 3°. A HP scheme is investigated for the spatial modulation technique in the massive MU-MIMO systems in [21], where the RF beamformer is constructed via the phase of instantaneous CSI as in [14]. In [22], joint user scheduling and HP design is studied for the mmWave massive MIMO systems, where the BS equipped with uniform rectangular array (URA) builds the RF beamformer via applying the singular value decomposition to the instantaneous channel matrix.

As mentioned earlier, the above studies cannot reduce the large CSI overhead because they all utilize the full CSI to construct HP. According to the aforementioned second strategy using slowly time-varying CSI for the RF-beamformer design, joint spatial division multiplexing (JSDM) technique is proposed in [23]. The authors consider the correlation-based channel model as explained in [30]. Assuming that users within the same groups have an identical covariance matrix and the groups have non-overlapping AoD distributions, the RF beamformer is constructed according to either eigen-beamforming based HP (EBF-HP) or block-diagonalization based HP (BD-HP) techniques by using the slowly time-varying channel covariance matrix. The primary goals of the RF-stage are to reduce the CSI overhead required for the HP structure and suppress the effect of the inter-group interference. Afterwards, by using zero-forcing (ZF) and RZF precoding, two approaches having different CSI requirements are given for the baseband precoder design as joint-group processing (JGP) and per-group processing (PGP). Later on, in [24], they also propose two algorithms for user grouping (using K-means clustering and chordal distance fixed quantization) based on the channel covariance matrix, where the number of user groups are assumed to be known in both approaches. Similarly, by using a priori knowledge of the number of user groups, [25] clusters the users by minimizing the generalized Fubini-Study distance between eigenspaces to improve the sum-rate performance of JSDM. In [26], it is shown that the sum-rate performance of JSDM can be enhanced, when the baseband precoder is designed according to the weighted minimum-mean-squared error optimization method. In [27], the JSDM technique is combined orthogonal frequency division multiplexing, which provides a flexible way to transmission in order to separately decode data with a small loss in spectral efficiency. Afterwards, [28] proposes a HP scheme for uniform circular arrays (UCA), where the RF beamformer using the phase mode transformation technique does not require any channel knowledge and it reduces the CSI overhead as well as the number of RF chains by transforming a large-size UCA into a reduced-size virtual ULA. Then, [29] analyzes a non-linear HP design for multi-cell massive MIMO systems, where an approximate block-diagonalization (BD) technique is utilized at the RF-stage and a minimum mean square error vector perturbation technique is employed at the BB-stage. It is shown that applying a non-linear HP technique can improve the error performance compared to the linear HP.

Regarding the phased array antenna architecture design, there are mainly three important considerations as (i) the array geometry, (ii) the number of antenna elements and (iii) the element spacing [31]. In terms of the array geometry, ULA, UCA and URA are the most common configurations. However, unlike ULA and URA, the steering vector of UCA does not have Vandermonde structure, which means that the angular-based beamforming techniques cannot be directly applied [28]. On the other hand, URA is more widely considered than ULA because URA packs antennas on a two-dimensional (2D) grid instead of the single-dimensional (1D) line [32]. Furthermore, ULA can be only utilized to estimate the azimuth angle, while URA is also capable of resolving the elevation angle by means of its 2D structure [33]. For three-dimensional (3D) beamforming, both azimuth and elevation angles should be taken into account. This aspect also makes URA a better solution for 3D beamforming [33]–[35]. However, due to its simple structure, ULAs are often investigated in the literature (e.g., in [14], [16]–[21], [23]–[26], [29]).

B. CONTRIBUTIONS AND ORGANIZATION

In this paper, a new user grouping algorithm and 3D angular-based HP (AB-HP) technique is proposed for massive MU-MIMO systems, where the BS is equipped with URA and users are clustered in non-overlapping geographical regions. Similar to JSDM, the proposed AB-HP aims to minimize both the CSI overhead and the number of RF chains. The main contributions are summarized below:

- **User grouping:** Different from [19], [24], [25], [36], the proposed user grouping algorithm does not require the number of groups as an input parameter, which is a priori unknown at the BS in practical application. Furthermore, the previous studies utilize either the complete knowledge of instantaneous CSI or the channel covariance matrix, which may be unavailable, especially for massive MU-MIMO systems having large antenna arrays. In this work, we propose an entirely different user grouping algorithm, which employs the AoD information of the users with respect to the BS. According to the proposed algorithm, the target number of user groups is first estimated through the eigenvalue decomposition of the Laplacian matrix, whose entries are similarity measures among the users AoD information. Accordingly, the users are partitioned into groups sharing the similar characteristics based on their AoD information. We compare the proposed user grouping algorithm and the direct applications of the K-means algorithm employing various similarity measures as chordal distance [24], weighted likelihood [36] and subspace projection [36]. It is demonstrated that the proposed method can provide well-separation among the groups and satisfactory assignment of the users to their corresponding groups, which cannot be achieved via the direct application K-means algorithm with an incorrect priori knowledge.
- **3D Angular-Based Hybrid Precoding (AB-HP):** Based on the AoD support of the user groups, the RF beamformer is designed to eliminate the inter-group interference while reducing the CSI overhead as well as the number of RF chains. Considering the practical constraints, AB-HP only utilizes the phase-shifters at the RF-stage without variable-gain amplifiers. On the other hand, JSDM also requires variable-gain amplifiers to build the RF beamformer according to EBF-HP and BD-HP techniques [23]. Therefore, unlike JSDM, AB-HP can simply realize the RF-stage via phase-shifters and provide the constant-modulus constraints. According to the RZF technique, three approaches requiring different reduced CSI overheads are considered for the baseband precoder, which are JGP, PGP and common-group-processing (CGP). For each approach, the signal-to-interference-plus-noise ratio (SINR) expressions are derived to evaluate the system sum-rate. Numerical results demonstrate that the proposed AB-HP schemes with the relaxed CSI estimation overhead and reduced hardware cost/complexity can tightly approach the sum-rate performance of the single-stage FDP. Considering the imperfect channel knowledge, it is shown that AB-HP is more robust to channel estimation errors compared to FDP. Also, the computational complexity of FDP can be significantly decreased via AB-HP. Furthermore, AB-HP can provide a considerable performance improvement in comparison to existing two-stage HP techniques such as EBF-HP, BD-HP, OBS-HP and NOAS-HP.
- **Deterministic SINR Approximations:** By applying the deterministic equivalent principle for large antenna arrays [9], the deterministic SINR approximations having almost sure convergence are derived for JGP, PGP and CGP approaches. During the derivation, we utilize

the tools from the large-dimensional random matrix theory [37]. As shown in the numerical results, without running heavy Monte-Carlo simulations, the derived approximations can be efficiently utilized to attain insights into the system performance.

- **Transfer Block Design for RF Chain Reduction:** In addition to the two-stage AB-HP having the RF beamformer and baseband precoder, a new three-stage AB-HP including a transfer block matrix is also proposed for the further minimization in terms of the hardware cost/complexity. The transfer block placed at the RF-stage is realized via phase-shifters. Hence, it is located between the RF beamformer and baseband precoder. According to the transfer block design, three-stage AB-HP can decrease the number of RF chains exactly to the total number of independent data streams without sacrificing the performance.

The rest of this paper is organized as follows. In Section II, we introduce the system model. Section III expresses the user grouping algorithm. In Section IV, we propose three approaches for AB-HP and derive the instantaneous SINR expressions as well as their deterministic approximations. Section V explains the transfer block design for the RF chain reduction. The illustrative results are provided in Section VI. Finally, the paper is concluded in Section VII.

Notation: Bold upper/lower case letters denote matrices/vectors. $(\cdot)^*$, $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^\dagger$ and $\|\cdot\|_2$ represent the complex conjugate, the transpose, the conjugate transpose, the Moore-Penrose inverse and the 2-norm of a vector or matrix, respectively. $\mathbf{I}_K$, $\mathbb{E}\{\cdot\}$, $\mathrm{tr}(\cdot)$ and $\angle(\cdot)$ stand for $K \times K$ identity matrix, the expectation operator, the trace operator and the argument of a complex number, respectively. $\mathbf{X} \otimes \mathbf{Y}$ denotes the Kronecker product of two matrices $\mathbf{X}$ and $\mathbf{Y}$. We use $x \sim \mathcal{CN}(0, \sigma)$ when x is a complex Gaussian random variable with zero-mean and variance σ.

II. SYSTEM MODEL

We consider a BS equipped with a URA having $M = M_x \times M_y$ antennas to serve K single-antenna users. The configuration for URA is shown in Figure 1, where M_x and M_y are respectively the number of antennas placed along x-axis and y-axis as in [35], [38], d is the distance between two antenna elements normalized by the wavelength, θ and ψ are the elevation and azimuth angles, respectively. The main purpose for selecting URA is to squeeze the antennas in 2D grid to satisfy the area requirements in practical applications. For example, when the carrier frequency is 3 GHz, a 20 × 20 URA at half-wavelength antenna spacing only requires 1 meter in horizontal and vertical directions, whereas a ULA with the same number of antennas needs 20 meters of space in one direction. Based on the geometry-based 3D channel model [30] and URA [31], the multipath channel coefficient between the $(m,n)^{th}$ BS antenna and k^{th} user is given by:

$$h_k^{m,n} = \sum_{p=1}^{P} g_{k,p} e^{-j2\pi d\left[(m-1)\gamma_{x,k,p}+(n-1)\gamma_{y,k,p}\right]}, \tag{1}$$

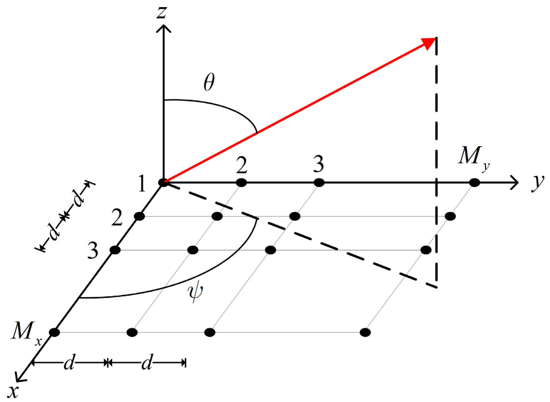


FIGURE 1. A massive MU-MIMO BS with a $M_x \times M_y$ URA configuration.

where P is the number of paths, $g_{k,p}$ is the complex coefficient reflecting the path loss and the random phase of the p^{th} path with the distribution of $\mathcal{CN}\left(0, \frac{1}{P}\right)$ so that $h_k^{m,n} \sim \mathcal{CN}(0,1)$. $\gamma_{x,k,p} = \sin\left(\theta_{k,p}\right)\cos\left(\psi_{k,p}\right)$ and $\gamma_{y,k,p} = \sin\left(\theta_{k,p}\right)\sin\left(\psi_{k,p}\right)$ carry the information of both elevation and azimuth angles of the corresponding path, where $\theta_{k,p} \in \left[\theta_k - \delta_k^\theta, \theta_k + \delta_k^\theta\right]$ with the mean elevation angle θ_k and elevation angle spread δ_k^θ, and $\psi_{k,p} \in \left[\psi_k - \delta_k^\psi, \psi_k + \delta_k^\psi\right]$ with the mean azimuth angle ψ_k and azimuth angle spread δ_k^ψ. By using (1), the received signal at the k^{th} user can be written as:

$$\begin{aligned} r_k &= \sum_{m=1}^{M_x}\sum_{n=1}^{M_y} h_k^{m,n} s_{m,n} + w_k \\ &= \sum_{m=1}^{M_x}\sum_{n=1}^{M_y}\sum_{p=1}^{P} g_{k,p} e^{-j2\pi d\left[(m-1)\gamma_{x,k,p}+(n-1)\gamma_{y,k,p}\right]} s_{m,n} + w_k \\ &= \mathbf{g}_k^T \mathbf{\Phi}_k \mathbf{s} + w_k, \end{aligned} \tag{2}$$

where $\mathbf{s} = \left[s_{1,1}, \cdots, s_{1,M_y}, \cdots, s_{M_x,1}, \cdots, s_{M_x,M_y}\right]^T \in \mathbb{C}^M$ denotes the transmitted signal vector satisfying a maximum transmit power constraint of P_T, i.e., $\mathbb{E}\left\{\|\mathbf{s}\|_2^2\right\} \leq P_T$, w_k is the complex Gaussian noise distributed as $\mathcal{CN}\left(0, \sigma^2\right)$, $\mathbf{g}_k = \left[g_{k,1}, \cdots, g_{k,P}\right]^T \in \mathbb{C}^P$ is the path gain vector and $\mathbf{\Phi}_k \in \mathbb{C}^{P \times M}$ is the phase response matrix according to the URA geometry and P paths defined as:

$$\mathbf{\Phi}_k = \begin{bmatrix} \boldsymbol{\phi}_x^T\left(\gamma_{x,k,1}\right) \otimes \boldsymbol{\phi}_y^T\left(\gamma_{y,k,1}\right) \\ \boldsymbol{\phi}_x^T\left(\gamma_{x,k,2}\right) \otimes \boldsymbol{\phi}_y^T\left(\gamma_{y,k,2}\right) \\ \vdots \\ \boldsymbol{\phi}_x^T\left(\gamma_{x,k,P}\right) \otimes \boldsymbol{\phi}_y^T\left(\gamma_{y,k,P}\right) \end{bmatrix}, \tag{3}$$

where $\boldsymbol{\phi}_x(\gamma) = \left[1, e^{-j2\pi d\gamma}, \cdots, e^{-j2\pi d(M_x-1)\gamma}\right]^T \in \mathbb{C}^{M_x}$ is the phase response vector along $x-$axis and $\boldsymbol{\phi}_y(\gamma) = \left[1, e^{-j2\pi d\gamma}, \cdots, e^{-j2\pi d(M_y-1)\gamma}\right]^T \in \mathbb{C}^{M_y}$ is the phase response vector along $y-$axis. Then, the instantaneous CSI for the k^{th} user can be defined as $\mathbf{h}_k^T = \mathbf{g}_k^T \mathbf{\Phi}_k \in \mathbb{C}^M$.

As in [23], we assume that the phase response matrix $\mathbf{\Phi}_k$ based on AoD information changes slower than the channel coherence time. Hence, the CSI can be divided into two parts: (i) fast time-varying instantaneous CSI as $\mathbf{h}_k$ and (ii) slowly time-varying phase response matrix as $\mathbf{\Phi}_k$. Furthermore, although the users usually have different instantaneous CSIs due to the local scattering and path loss, a group of users sharing the similar AoD support results in the similar phase response matrices.

III. USER GROUPING

The first step toward the AB-HP design is to partition the users into groups sharing similar characteristics. According to the channel model given in (1), all users experience different channel coefficients. However, it is reasonable to assume that the users collocated in the same area have almost the same AoDs with respect to the BS. Our approach is based on the AoD information,[1] where the users belonging to different groups can be separated in orthogonal subspaces.

Given the AoD information associated to every user and a measure of similarity $\rho_{i,j} \geq 0$ between any pair of i^{th} and j^{th} users for $i, j = 1, \cdots, K$ and $i \neq j$, our goal is to divide the users into several groups such that users in the same group have similar AoDs and users in different groups are dissimilar to each other. In practical situations, the number of groups is a priori unknown. Therefore, we propose to jointly find the number of groups and the distribution of users among these groups. We first represent the massive MU-MIMO network as an undirected graph $\mathcal{G}$. Each vertex in this graph represents a user and the connection between two vertices i and j (called edge) is weighted by $\rho_{i,j}$ to model the similarity between two users. It is worthwhile to mention that $\rho_{i,j}$ is equal to $\rho_{j,i}$ for our undirected graph. The problem of user grouping is now reformulated using the similarity graph. Hence, our objective is to find a partition of the graph such that the edges between two users in the different groups have low weights (low similarity) by means of dissimilar AoD support and the edges within the same group have high weights (high similarity) by means of the similar AoD support.

Since the constructed graph should represent the relationships between users, the similarity measure has to model local neighborhoods. By using the Gaussian kernel function, the similarity between i^{th} and j^{th} users can be expressed as:

$$\rho_{i,j} = e^{-\frac{\|\boldsymbol{\gamma}_{x,i}-\boldsymbol{\gamma}_{x,j}\|_2^2+\|\boldsymbol{\gamma}_{y,i}-\boldsymbol{\gamma}_{y,j}\|_2^2}{\Delta^2}}, \tag{4}$$

where $\boldsymbol{\gamma}_{x,i}, \boldsymbol{\gamma}_{x,j}, \boldsymbol{\gamma}_{y,i}, \boldsymbol{\gamma}_{y,j} \in \mathbb{R}^P$ are the angle vectors constructed via the AoD information of i^{th} and j^{th} users, and Δ is the parameter controlling the width of the neighborhoods between $\{\boldsymbol{\gamma}_{x,i}, \boldsymbol{\gamma}_{x,j}\}$ and $\{\boldsymbol{\gamma}_{y,i}, \boldsymbol{\gamma}_{y,j}\}$ pairs. However, the similarity measure as defined in (4) is not suitable for the above described problem. To understand the reason, we consider the following simple example. Suppose that the angles vectors for users 1 and user 2 are $\boldsymbol{\gamma}_{x,1} = \boldsymbol{\gamma}_{y,1} = [0.21, 0.41, 0.53]$ and $\boldsymbol{\gamma}_{x,2} = \boldsymbol{\gamma}_{y,2} = [0.34, 0.38, 0.39]$, respectively. Despite of the quite similar AoD supports, the similarity measure returns a very small weight and the users will be placed in different groups. While examining the angle vectors of these users, we can see that although the difference between the second elements of the angle vectors is very small, its effect vanishes due to the high differences between the first and third element pairs. Thus, this similarity cannot be reflected via $\rho_{i,j}$ defined in (4). To overcome this problem, we first define the similarity function between i^{th} and j^{th} users as follows:

$$\rho_{i,j} = \sum_{p_x=1}^{P_i} \sum_{p_y=1}^{P_i} \sum_{q_x=1}^{P_j} \sum_{q_y=1}^{P_j} e^{-\frac{\left|\gamma_{x,i,p_x}-\gamma_{x,j,q_x}\right|^2+\left|\gamma_{y,i,p_y}-\gamma_{y,j,q_y}\right|^2}{\Delta^2}}, \tag{5}$$

where P_i and P_j denote the number of paths for the i^{th} and j^{th} users, respectively, for the case of different number of paths each user in general. The effect of close angle pairs can be now represented with above defined similarity measure in (5). The information about the similarity can be compactly represented by the weighted adjacency matrix $\mathbf{J} \in \mathbb{R}^{K \times K}$, whose $(i, j)^{th}$ element is $\rho_{i,j}$ and the diagonal elements are 0, and the diagonal matrix $\mathbf{N} \in \mathbb{R}^{K \times K}$, whose i^{th} diagonal element is $\tau_i = \sum_{j=1, j \neq i}^{P} \rho_{i,j}$. The main tool to find the number of user groups as well as the user grouping is the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{K \times K}$ defined as:

$$\mathbf{L} = \mathbf{N} - \mathbf{J} = \begin{bmatrix} \tau_1 & -\rho_{1,2} & \cdots & -\rho_{1,K} \\ -\rho_{2,1} & \tau_2 & \cdots & -\rho_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_{K,1} & -\rho_{K,2} & \cdots & \tau_K \end{bmatrix}. \tag{6}$$

The eigenvalues and eigenvectors of the Laplacian matrix describe many properties of the graph, in particular they allow us to partition the graph through eigenvalue decomposition. One can show that the Laplacian matrix $\mathbf{L}$ is a real symmetric and positive semi-definite matrix. Therefore, the eigenvalue decomposition of the Laplacian matrix can be expressed as $\mathbf{L} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q} \in \mathbb{R}^{K \times K}$ is the matrix containing the eigenvectors on its columns and $\mathbf{\Lambda} \in \mathbb{R}^{K \times K}$ is the diagonal matrix with the eigenvalues on the diagonal elements. Given the eigenvalue decomposition of $\mathbf{L}$, the user grouping is performed as follows. The multiplicity of eigenvalue 0 is equal to the number of user groups G and the eigenvector of the corresponding eigenvalue 0 are the indicator vectors[2] $\mathbf{1}_{\mathcal{G}_1}, \cdots, \mathbf{1}_{\mathcal{G}_g}, \cdots, \mathbf{1}_{\mathcal{G}_G}$ for $g = 1, \cdots, G$ [40]. To explain these properties, we consider an eigenvector $\mathbf{q}_g = \left[q_{g,1}, \cdots, q_{g,K}\right]^T \in \mathbb{R}^K$ with the corre-

[1] AoD changes over time at a much slower rate than the actual channel coefficients, Also, it can be tracked and estimated by using standard AoDs estimation algorithms [35], [38], [39].

[2] The indicator vector $\mathbf{1}_{\mathcal{G}_g}$ of group g has value 1 at position k if the k^{th} user belongs to group g, and it has value 0 at position k if the k^{th} user does not belong to group g.

Algorithm 1 Proposed User Grouping Algorithm

```
Input: $\{\theta_{k,p}, \psi_{k,p}\}$ for $k = 1 : K$ and $p = 1 : P$.
Generate angle pairs $\gamma_{x,k,p} = \sin(\theta_{k,p}) \cos(\psi_{k,p})$ and $\gamma_{y,k,p} = \sin(\theta_{k,p}) \sin(\psi_{k,p})$.
Calculate the similarity measures $\rho_{i,j}$ via (5).
Form the Laplacian matrix $\mathbf{L}$ via (6).
Find $G$, the multiplicity of the eigenvalue 0 of $\mathbf{L}$.
Find $\{\mathbf{q}_g\}_{g=1}^{G}$, the eigenvectors with the eigenvalue 0.
Form $\mathbf{Q}_G = [\mathbf{q}_1, \cdots, \mathbf{q}_G]$ via (8).
Apply $K$-means algorithm to the rows of $\mathbf{Q}_G$ to cluster $K$ users in $G$ groups.
```

sponding eigenvalue 0, hence, we have:

$$\mathbf{q}_g^T \mathbf{L} \mathbf{q}_g = \sum_{i=1}^{K} \tau_i q_{g,i}^2 - \sum_{i=1}^{K} \sum_{j=1}^{K} \rho_{i,j} q_{g,i} q_{g,j} \overset{(a)}{=} \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \rho_{i,j} \big(q_{g,i} - q_{g,j}\big)^2 = 0, \tag{7}$$

where (a) is obtained by using $\tau_i = \sum_{j=1, j\neq i}^{P} \rho_{i,j}$. As the similarity measures $\rho_{i,j}$ are non-negative, the above term can be zero, if and only if all the positive terms $\rho_{i,j}\big(q_{g,i} - q_{g,j}\big)^2$ are exactly 0, $\forall i, j$. Hence, if $\rho_{i,j}$ is strictly larger than 0 (i.e., the two users i and j belong to the same group and have a strong similarity factor), we have necessary $q_{g,i} = q_{g,j}$, where $q_{g,i}$ is equal to a constant different from 0 for the users i belonging to the same group g. Moreover, when two users i and j belong to different groups, the similarity measure $\rho_{i,j}$ is almost zero due to the disjoint AoD supports. To satisfy the orthogonality among the eigenvectors with the eigenvalue 0, the remaining entries of the eigenvector $\mathbf{q}_g$ should be equal to 0 for the corresponding users belonging to different groups rather than the group g. As a result, the corresponding eigenvector for the users belonging to the same group g becomes $\mathbf{q}_g = \mathbf{1}_{\mathcal{G}_g}$. A detailed proof can be found in [40].

Afterwards, once the number of groups G is obtained by finding the multiplicity of the eigenvalue 0 for the Laplacian matrix $\mathbf{L}$, we can compute G eigenvectors $\mathbf{q}_1, \cdots, \mathbf{q}_G \in \mathbb{R}^K$ of the Laplacian matrix $\mathbf{L}$ with the eigenvalue 0 and build:

$$\mathbf{Q}_G = [\mathbf{q}_1, \cdots, \mathbf{q}_G] \in \mathbb{R}^{K\times G}. \tag{8}$$

Finally, K-means algorithm is applied to the rows of $\mathbf{Q}_G$ in order to cluster K users in G groups. Here, we omit the explanation for the K-means algorithm and the detailed description for K-means algorithm can be found in [41]. The main complexity of the proposed grouping algorithm is the eigenvalue decomposition of the real symmetric Laplacian matrix with the size of $K \times K$. In Algorithm 1, we summarize the procedures for the proposed user grouping algorithm.

Our main objective is to collect users having similar AoD supports. One may think of applying K-means directly on the AoD information associated with each user. This approach is adopted in [36] for an ULA using the user covariance matrices with the prior knowledge of the number of user groups G. However, when we try to apply the K-means algorithm directly on the AoD information, the first obstacle is the number of groups G, which is unknown a priori. On the other hand, the Laplacian matrix given in (6) provides a systematic way to find the number of user groups by applying eigenvalue decomposition and also to cluster the users by manipulating its eigenvectors as explained in Algorithm 1.

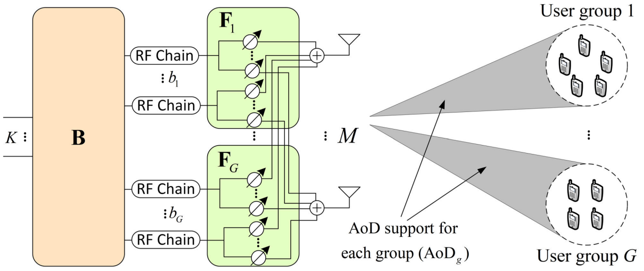


FIGURE 2. System model for the proposed two-stage AB-HP scheme.

IV. 3D ANGULAR-BASED HYBRID PRECODING

According to Figure 2, the transmitted signal is expressed as $\mathbf{s} = \mathbf{FBd}$, where $\mathbf{F} \in \mathbb{C}^{M\times b}$ is the RF beamformer designed via the slowly time-varying AoD information, $\mathbf{B} \in \mathbb{C}^{b\times K}$ is the baseband precoder designed via the effective CSI seen by the baseband and $\mathbf{d} \in \mathbb{C}^K$ is the downlink data vector encoded by i.i.d. Gaussian codebooks (i.e., i.i.d. entries of $\mathbf{d}$ follow the distribution of $\mathcal{CN}(0, 1)$, so we have $\mathbb{E}\{\mathbf{dd}^H\} = \mathbf{I}_K$). The ultimate objective for partitioning the linear precoder into two sub-blocks is to reduce the channel estimation overhead and to decrease the hardware cost/complexity. Throughout this section, we first develop the RF beamformer and then present three approaches for the baseband precoder under different CSI overhead requirement. Furthermore, deterministic equivalents of the SINR expressions are obtained for all three baseband precoder approaches.

A. RF BEAMFORMER DESIGN

Suppose that K users are clustered in G groups based on the similarity of AoD information as shown in Figure 2, where the number of users in group g is denoted as K_g such that $K = \sum_{g=1}^{G} K_g$. Since the users inside the same group have similar AoDs, the instantaneous CSI of the k^{th} user in group g is $\mathbf{h}_{g_k}^T = \mathbf{g}_{g_k}^T \mathbf{\Phi}_g \in \mathbb{C}^M$ with the index $g_k = k + \sum_{p=1}^{g-1} K_p$ and $\mathbf{\Phi}_g \in \mathbb{C}^{P\times M}$ is the phase response matrix of group g. By using (2), the received signal at group g can be given by:

$$\mathbf{r}_g = \mathbf{H}_g \mathbf{FBd} + \mathbf{w}_g, \tag{9}$$

where $\mathbf{H}_g = \mathbf{G}_g \mathbf{\Phi}_g \in \mathbb{C}^{K_g\times M}$ is the channel matrix, $\mathbf{G}_g = \big[\mathbf{g}_{g_1}, \cdots, \mathbf{g}_{g_{K_g}}\big]^T \in \mathbb{C}^{K_g\times P}$ is the path gain matrix and $\mathbf{w}_g = \big[w_{g_1}, \cdots, w_{g_{K_g}}\big]^T \in \mathbb{C}^{K_g}$ is the noise vector for group g. As mentioned earlier, the RF beamformer $\mathbf{F}$ depends on the AoD information, i.e., on the phase response matrix $\mathbf{\Phi}_g$. According to the user groups, we design G different sub-blocks for the RF stage as $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_G] \in \mathbb{C}^{M \times b}$, where $\mathbf{F}_g \in \mathbb{C}^{M \times b_g}$ denotes the RF beamformer for group g such that $b = \sum_{g=1}^{G} b_g$. Then, the received signal is written as:

$$\mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_G \end{bmatrix} = \boldsymbol{\mathcal{H}}\mathbf{B} \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_G \end{bmatrix} + \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_G \end{bmatrix}, \qquad (10)$$

where $\boldsymbol{\mathcal{H}} = \mathbf{HF} \in \mathbb{C}^{K \times b}$ is the reduced-size effective channel matrix seen from baseband. By defining $\mathbf{H} = \left[\mathbf{H}_1^T, \cdots, \mathbf{H}_G^T\right]^T \in \mathbb{C}^{K_g \times M}$, the effective channel matrix can be expanded as:

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathbf{H}_1\mathbf{F}_1 & \mathbf{H}_1\mathbf{F}_2 & \cdots & \mathbf{H}_1\mathbf{F}_G \\ \mathbf{H}_2\mathbf{F}_1 & \mathbf{H}_2\mathbf{F}_2 & \cdots & \mathbf{H}_2\mathbf{F}_G \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_G\mathbf{F}_1 & \mathbf{H}_G\mathbf{F}_2 & \cdots & \mathbf{H}_G\mathbf{F}_G \end{bmatrix}, \qquad (11)$$

where the diagonal block-matrix $\boldsymbol{\mathcal{H}}_g = \mathbf{H}_g\mathbf{F}_g = \mathbf{G}_g\mathbf{\Phi}_g\mathbf{F}_g$ is the effective channel matrix for group g and the off-diagonal block-matrix $\mathbf{H}_p\mathbf{F}_g = \mathbf{G}_p\mathbf{\Phi}_p\mathbf{F}_g$ is the effective interference channel matrix between groups g and p, $\forall p \neq g$. To eliminate the inter-group interference, the RF beamformer matrices have to be chosen as:

$$\mathbf{\Phi}_p\mathbf{F}_g \approx 0, \ \forall p \neq g \text{ and } p, g = 1, \cdots, G. \qquad (12)$$

The above approximate zero condition implies that the RF beamformer for group g should be orthogonal to the phase response matrix of the group p. This can be verified as long as the columns of $\mathbf{F}_g$ belong to the intersection of the null spaces of $\mathbf{\Phi}_p$, i.e., $\text{Span}\left(\mathbf{F}_g\right) \subset \cap_{p \neq g} \text{Null}\left(\mathbf{\Phi}_p\right)$. Moreover, in order to maximize the beamforming gain, the columns of $\mathbf{F}_g$ should belong to the subspace spanned by $\mathbf{\Phi}_g$, i.e., $\text{Span}\left(\mathbf{F}_g\right) \subset \text{Span}\left(\mathbf{\Phi}_g\right)$. Thus, the intersection of $\text{Span}\left(\mathbf{\Phi}_g\right)$ and $\text{Null}\left(\mathbf{\Phi}_p\right)$, $\forall p \neq g$, should not be empty to obtain the RF beamformer matrix satisfying the above conditions. As explained later, when the number of antennas M increases as in a massive MIMO scenario, $\text{Span}\left(\mathbf{\Phi}_g\right)$ is asymptotically included in $\text{Null}\left(\mathbf{\Phi}_p\right)$. Hence, a sufficient condition is to design the columns of $\mathbf{F}_g$ in the basis of $\text{Span}\left(\mathbf{\Phi}_g\right)$.

The AoD support of the group g is expressed as the union of AoD supports for all user in the corresponding group as:

$$\begin{aligned} \text{AoD}_g &= \Big\{\left[\gamma_x, \gamma_y\right] \\ &= \sin(\theta)\left[\cos(\psi), \sin(\psi)\right] \Big| \psi \in \boldsymbol{\psi}_g, \theta \in \boldsymbol{\theta}_g\Big\}, \end{aligned} \qquad (13)$$

where $\boldsymbol{\psi}_g = \left[\psi_g^{\min}, \psi_g^{\max}\right] = \left[\min_{g_k,p}\psi_{g_k,p}, \max_{g_k,p}\psi_{g_k,p}\right]$ is the azimuth angle support for group g and $\boldsymbol{\theta}_g = \left[\theta_g^{\min}, \theta_g^{\max}\right] = \left[\min_{g_k,p}\theta_{g_k,p}, \max_{g_k,p}\theta_{g_k,p}\right]$ is the elevation angle support for group g. To achieve $\text{Span}\left(\mathbf{F}_g\right) \subset \text{Span}\left(\mathbf{\Phi}_g\right)$, the columns of the RF beamformer for the group g can be constructed as:

$$\mathbf{F}_g = \left\{\mathbf{e}\left(\gamma_x, \gamma_y\right) \mid \left(\gamma_x, \gamma_y\right) \in \text{AoD}_g\right\}, \qquad (14)$$

where $\mathbf{e}\left(\gamma_x, \gamma_y\right) = \mathbf{e}_x\left(\gamma_x\right) \otimes \mathbf{e}_y\left(\gamma_y\right) \in \mathbb{C}^M$ is the steering vector for $\left(\gamma_x, \gamma_y\right)$ angle combination. Here, $\mathbf{e}_x\left(\gamma_x\right) = \frac{1}{\sqrt{M_x}}\left[1, e^{j2\pi d\gamma_x}, \cdots, e^{j2\pi d(M_x-1)\gamma_x}\right]^T \in \mathbb{C}^{M_x}$ and $\mathbf{e}_y\left(\gamma_y\right) = \frac{1}{\sqrt{M_y}}\left[1, e^{j2\pi d\gamma_y}, \cdots, e^{j2\pi d(M_y-1)\gamma_y}\right]^T \in \mathbb{C}^{M_y}$ are the steering vector along x-axis and y-axis, respectively. By defining the quantized angle-pairs as $\lambda_i^x = -1 + \frac{2i-1}{M_x}$ and $\lambda_j^y = -1 + \frac{2j-1}{M_y}$ for $i = 1, \cdots, M_x$ and $j = 1, \cdots, M_y$, one can verify the orthogonality between the steering vectors as:

$$\mathbf{e}^H\left(\lambda_i^x, \lambda_j^y\right)\mathbf{e}(\lambda_{i'}^x, \lambda_{j'}^y) = 0, \forall \{i, j\} \neq \{i', j'\}. \qquad (15)$$

Therefore, $\left\{\left\{\mathbf{e}\left(\lambda_i^x, \lambda_j^y\right)\right\}_{i=1}^{M_x}\right\}_{j=1}^{M_y}$ provides as many orthogonal vectors as the number of antennas $M = M_xM_y$. The quantized angle-pairs provide M orthogonal steering vectors, which is the minimum number of angle-pairs to span the complete AoD space (i.e., the complete elevation and azimuth angle domain). Hence, it minimizes the RF chains utilization to cover the complete AoD support of a given user group. On the other hand, other angle-pairs can be also utilized for the RF beamformer design. However, they require a larger number of RF chains and increase the hardware cost/complexity. Furthermore, by choosing the quantized angle-pairs for the RF beamformer design, the inter-group interference can be further suppressed and well-managed compared to the non-orthogonal beamforming techniques (e.g., NOAS-HP [20] as discussed in Section VI-F). For any $\left(\lambda_i^x, \lambda_j^y\right) \in \text{AoD}_g$ and $\left(\gamma_x, \gamma_y\right) \notin \text{AoD}_g$, the inner product:

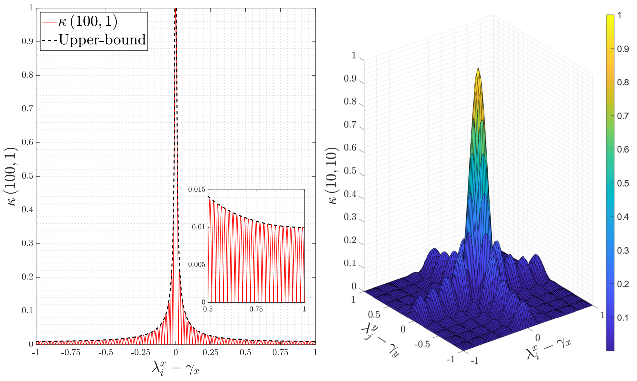
$$\mathbf{e}^H\left(\gamma_x, \gamma_y\right)\mathbf{e}\left(\lambda_i^x, \lambda_j^y\right) = \sum_{m=0}^{M_x-1}\sum_{n=0}^{M_y-1} \frac{e^{j2\pi d[m(\lambda_i^x-\gamma_x)+n(\lambda_j^y-\gamma_y)]}}{M}, \qquad (16)$$

tends to be 0 for large antenna arrays. In order to prove the convergence of the above expression for the large antenna arrays, $\kappa\left(M_x, M_y\right)$ is defined as follows:

$$\kappa\left(M_x, M_y\right) = \left|\mathbf{e}^H\left(\gamma_x, \gamma_y\right)\mathbf{e}\left(\lambda_i^x, \lambda_j^y\right)\right|. \qquad (17)$$

For the half-wavelength antenna separation (i.e., $d = 0.5$), using the finite geometric series expansion for $\lambda_i^x \neq \gamma_x$ and $\lambda_j^y \neq \gamma_y$, an upper-bound for $\kappa\left(M_x, M_y\right)$ can be obtained as:

$$\begin{aligned} \kappa\left(M_x, M_y\right) &= \frac{\left|\sum_{m=0}^{M_x-1} e^{j\pi m(\lambda_i^x-\gamma_x)}\right|\left|\sum_{n=0}^{M_y-1} e^{j\pi n\left(\lambda_j^y-\gamma_y\right)}\right|}{M} \\ &= \frac{1}{M}\frac{\left|1-e^{j\pi M_x(\lambda_i^x-\gamma_x)}\right|}{\left|1-e^{j\pi(\lambda_i^x-\gamma_x)}\right|}\frac{\left|1-e^{j\pi M_y\left(\lambda_j^y-\gamma_y\right)}\right|}{\left|1-e^{j\pi\left(\lambda_j^y-\gamma_y\right)}\right|} \\ &\overset{(a)}{=} \frac{\left|\sin\left(\frac{\pi M_x(\lambda_i^x-\gamma_x)}{2}\right)\right|\left|\sin\left(\frac{\pi M_y\left(\lambda_j^y-\gamma_y\right)}{2}\right)\right|}{M\left|\sin\left(\frac{\pi(\lambda_i^x-\gamma_x)}{2}\right)\right|\left|\sin\left(\frac{\pi\left(\lambda_j^y-\gamma_y\right)}{2}\right)\right|} \\ &\overset{(b)}{\leq} \frac{1}{M\left|\sin\left(\frac{\pi(\lambda_i^x-\gamma_x)}{2}\right)\right|\left|\sin\left(\frac{\pi\left(\lambda_j^y-\gamma_y\right)}{2}\right)\right|}, \end{aligned} \qquad (18)$$



(a) (b)

FIGURE 3. $\kappa\left(M_x, M_y\right)$ for (a) 100 × 1 ULA and (b) 10 × 10 URA.

where (a) is found via $\left|1-e^{j\alpha}\right| = \left|\sin\left(\frac{\alpha}{2}\right)\right|$ and (b) is the direct consequence of $0 \leq \left|\sin(\alpha)\right| \leq 1$. For any $\left(\lambda_i^x, \lambda_j^y\right) \in \mathrm{AoD}_g$ and $\left(\gamma_x, \gamma_y\right) \notin \mathrm{AoD}_g$ (i.e., $\frac{2}{Mx} \ll |\lambda_i^x - \gamma_x| \ll 2 - \frac{2}{Mx}$ and/or $\frac{2}{M_y} \ll |\lambda_j^y - \gamma_y| \ll 2 - \frac{2}{M_y}$, where $\frac{2}{M_x}$ and $\frac{2}{M_y}$ represent the step-size for the quantized angle-pairs along x-axis and y-axis, respectively), each sine function located in the denominator converges to a constant value higher than 0. Therefore, as antenna array size increases (i.e. M goes to infinity), the upper-bound for $\kappa\left(M_x, M_y\right)$ given in (18) converges to 0, when $\left(\lambda_i^x, \lambda_j^y\right) \in \mathrm{AoD}_g$ and $\left(\gamma_x, \gamma_y\right) \notin \mathrm{AoD}_g$.

Figure 3 illustrates the behavior of $\kappa\left(M_x, M_y\right)$ for two antenna array configurations having $M = M_x \times M_y = 100$ antennas. Firstly, in Figure 3(a), $\kappa(100, 1)$ for 100 × 1 ULA is plotted versus $\lambda_i^x - \gamma_x$, where the tightness of upper-bound expression obtained in (18) is also represented. As seen from the results, as long as the difference among the angle-pairs increases, $\kappa(100, 1)$ approaches to 0, which provides the earlier defined approximate zero condition given in (12). Moreover, when the angle-pairs are quite closed to each other (i.e., $|\lambda_i^x - \gamma_x| \approx 0$), the beamforming gain can be maximized as indicated in (14). Secondly, $\kappa(10, 10)$ for 10 × 10 URA is demonstrated versus $\lambda_i^x - \gamma_x$ and $\lambda_j^y - \gamma_y$ in Figure 3(b), where the maximum beamforming gain can be achieved by minimizing the distance between angle-pairs along both x-axis and y-axis. Otherwise, the beamforming gain approaches to 0 (i.e., it becomes nearly orthogonal).

Based on the phase response matrix given in (3), the q^{th} row of $\mathbf{\Phi}_g \in \mathbb{C}^{P\times M}$ can be represented by $\mathbf{e}^H\left(\gamma_{x,g,q}, \gamma_{y,g,q}\right) = \boldsymbol{\phi}_x^T\left(\gamma_{x,g,q}\right) \otimes \boldsymbol{\phi}_y^T\left(\gamma_{y,g,q}\right)$. Therefore, considering no intersection between the AoD supports of the user groups, $\mathbf{\Phi}_p \mathbf{e}(\lambda_i^x, \lambda_j^y)$ approaches $\mathbf{0}$ for $\left(\lambda_i^x, \lambda_j^y\right) \in \mathrm{AoD}_g$ as expressed in (18) and illustrated in Figure 3. Hence, when the RF beamformer for group g is constructed via the quantized angle-pairs covering AoD_g, it both satisfies (i) $\mathrm{Span}\left(\mathbf{F}_g\right) \subset \cap_{p\neq g} \mathrm{Null}\left(\mathbf{\Phi}_p\right)$ to suppress the inter-group interference as indicated by the approximate zero condition given in (12) and (ii) $\mathrm{Span}\left(\mathbf{F}_g\right) \subset \mathrm{Span}\left(\mathbf{\Phi}_g\right)$ to maximize the beamforming gain. Then, the quantized angle-pairs covering AoD_g are found as:

Algorithm 2 RF Beamformer Design

```
Input: $M_x, M_y$, and $\{\theta_{k,p}, \psi_{k,p}\}$ for $k = 1 : K$ and $p = 1 : P$.
Generate angle pairs $\lambda_i^x = -1 + \frac{2i-1}{M_x}$ and $\lambda_j^y = -1 + \frac{2j-1}{M_y}$.
for $g = 1 : G$, do
    Form $\mathrm{AoD}_g$ given in (13).
    Find $\left(\lambda_i^x, \lambda_j^y\right)$ pairs covering $\mathrm{AoD}_g$ via (19).
    Obtain the RF beamformer for group $g$ as $\mathbf{F}_g$ via (20).
end for
Generate the RF beamformer as $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_G]$.
```

$$\left(\lambda_i^x, \lambda_j^y\right) \mid \gamma_x \in \boldsymbol{\lambda}_i^x, \gamma_y \in \boldsymbol{\lambda}_j^x, \left(\gamma_x, \gamma_y\right) \in \mathrm{AoD}_g, \qquad (19)$$

where $\boldsymbol{\lambda}_i^x = \left[\lambda_i^x - \frac{1}{M_x}, \lambda_i^x + \frac{1}{M_x}\right]$ denotes the boundaries of quantized angle λ_i^x along x-axis, $\boldsymbol{\lambda}_j^y = \left[\lambda_j^y - \frac{1}{M_y}, \lambda_j^y + \frac{1}{M_y}\right]$ denotes the boundaries of quantized angle λ_j^y along y-axis. According to the quantized angle-pairs, the number of $\left(\lambda_i^x, \lambda_j^y\right)$ pairs satisfying (19) is denoted by b_g. Actually, b_g also represents the number of orthogonal-beams and the maximum number of data streams that can be supported for group g. Finally, by using the quantized angle-pairs, the RF beamformer for group g is obtained as:

$$\mathbf{F}_g = \left[\mathbf{e}\left(\lambda_{i_1}^x, \lambda_{j_1}^y\right), \cdots, \mathbf{e}\left(\lambda_{j_{b_g}}, \lambda_{j_{b_g}}^y\right)\right]. \qquad (20)$$

The RF beamformer design for the proposed AB-HP technique is summarized in Algorithm 2. Hence, the RF beamformer can be realized via phase-shifters by means of the constant-modulus property of the steering vectors. Also, the obtained RF beamformer matrix is unitary, i.e., $\mathbf{F}^H\mathbf{F} = \mathbf{I}_b$. Based on Algorithm 2, an illustrative example for the RF beamformer design is discussed in Section VI-A, where Figure 5 explains how to obtain b_g for a given AoD support.

It is worthwhile to mention that when the total number of users is higher than the number of available RF chains, the user scheduling policies in [42]–[44] can be applied per each group to serve a subset of users. In this paper, we suppose here that all users can be simultaneously served by the BS.

B. BASEBAND PRECODER DESIGN

After obtaining the RF beamformer, we apply three approaches: joint-group-processing (JGP), per-group-processing (PGP) and common-group-processing (CGP) to design the baseband precoding stage. Note that both JGP and PGP are also employed for the EBF-HP and BD-HP techniques in [23], where the RF-stage requiring variable-gain amplifiers as well as phase-shifters is constructed via the channel covariance matrix.

1) JOINT-GROUP-PROCESSING (JGP)

When the estimation of the entire effective channel matrix given in (11) is affordable, the baseband precoder can be jointly designed for all groups according to the JGP approach. Hence, the RZF baseband precoder is given by:

$$\mathbf{B} = \varepsilon \mathbf{W} \boldsymbol{\mathcal{H}}^H, \tag{21}$$

where $\mathbf{W} = \left(\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{H}} + K\alpha \mathbf{I}_b\right)^{-1}$, α is the regularization parameter and ε is the normalization scalar to satisfy the transmit power constraint defined as:

$$\begin{aligned}\varepsilon^2 &= \frac{P_T}{\mathrm{tr}\left(\mathbb{E}\left\{\mathbf{d}^H \boldsymbol{\mathcal{H}} \mathbf{W} \mathbf{F}^H \mathbf{F} \mathbf{W} \boldsymbol{\mathcal{H}}^H \mathbf{d}\right\}\right)} \\ &\overset{(a)}{=} \frac{P_T}{\mathrm{tr}\left(\mathbf{F} \mathbf{W} \boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{H}} \mathbf{W} \mathbf{F}^H\right)} \\ &\overset{(b)}{=} \frac{P_T}{\mathrm{tr}\left(\boldsymbol{\mathcal{H}} \mathbf{W}^2 \boldsymbol{\mathcal{H}}^H\right)},\end{aligned} \tag{22}$$

where (a) follows the linearity of trace operator and $\mathbb{E}\left\{\mathbf{d}\mathbf{d}^H\right\} = \mathbf{I}_K$, (b) is the direct consequence of the unitary property of the RF beamformer. Using (2), the received signal at the k^{th} user in group g is expanded as:

$$r_{g_k} = \underbrace{\varepsilon \mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_{g_k}^* d_{g_k}}_{\text{Desired Signal}} + \underbrace{\sum_{p \neq g_k}^{K} \varepsilon \mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_p^* d_p}_{\text{Interference}} + \underbrace{w_{g_k}}_{\text{Noise}}. \tag{23}$$

After some mathematical manipulations, the instantaneous SINR at the k^{th} user in the group g for the JGP approach is found as:

$$\mathrm{SINR}_{g_k}^J = \frac{\left|\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{h}_{g_k}^*\right|^2}{\left\|\mathbf{h}_{g_k}^T \mathbf{F} \mathbf{W} \mathbf{F}^H \mathbf{H}_{[g_k]}^H\right\|_2^2 + \frac{\sigma^2}{\varepsilon^2}}, \tag{24}$$

where $\mathbf{H}_{[g_k]} = \left[\mathbf{h}_1, \cdots, \mathbf{h}_{g_k - 1}, , \mathbf{h}_{g_k + 1} \cdots, \mathbf{h}_K\right]^T$ represents the interference channel matrix. By applying the deterministic equivalent principle for large antenna arrays [9], the SINR in (24) converges almost surely to a deterministic quantity $\overline{\mathrm{SINR}}_{g_k}^J$, which can be derived as:

$$\overline{\mathrm{SINR}}_{g_k}^J = \frac{\left(m_g^J\right)^2}{\frac{(K_g - 1)c_{g,g}^J}{\left(1 + m_g^J\right)^2} + \sum\limits_{p \neq g}^{G} \frac{K_p c_{g,p}^J}{\left(1 + m_p^J\right)^2} + \sum\limits_{p=1}^{G} \frac{K_p c_p^J \sigma^2 \left(1 + m_g^J\right)^2}{P_T \left(1 + m_p^J\right)^2}}, \tag{25}$$

where m_g^J for $g = 1, \cdots, G$ is the unique solution of:

$$\begin{aligned} m_g^J &= \frac{1}{P} \mathrm{tr}\left(\mathbf{D}_g^J \mathbf{U}^J\right), \\ \mathbf{U}^J &= \left(\sum_{g=1}^{G} \frac{K_g \mathbf{D}_g^J}{P\left(1 + m_g^J\right)} + K\alpha \mathbf{I}_b\right)^{-1}, \end{aligned} \tag{26}$$

where $\mathbf{D}_g^J = \mathbf{F}^H \boldsymbol{\Phi}_g^H \boldsymbol{\Phi}_g \mathbf{F}$. Also, c_g^J and $c_{g,p}^J$ are the elements of the following defined column vector and matrix:

$$\left[c_1^J, \cdots, c_G^J\right]^T = \left(\mathbf{I}_G - \mathbf{Z}\right)^{-1} \mathbf{v},$$

$$\begin{bmatrix} c_{1,1}^J & \cdots & c_{1,G}^J \\ \vdots & \ddots & \vdots \\ c_{G,1}^J & \cdots & c_{G,G}^J \end{bmatrix} = \left(\mathbf{I}_G - \mathbf{Z}\right)^{-1} \mathbf{V}, \tag{27}$$

where $\mathbf{Z}$, $\mathbf{V}$ and $\mathbf{v}$ are defined as:

$$\begin{aligned} \mathbf{Z}(g, p) &= \frac{K_p \mathrm{tr}\left(\mathbf{D}_g^J \mathbf{U}^J \mathbf{D}_p^J \mathbf{U}^J\right)}{P^2 \left(1 + m_p^J\right)^2}, \\ \mathbf{V}(g, p) &= \frac{\mathrm{tr}\left(\mathbf{D}_g^J \mathbf{U}^J \mathbf{D}_p^J \mathbf{U}^J\right)}{P^2}, \\ \mathbf{v}(g) &= \frac{\mathrm{tr}\left(\mathbf{D}_g^J \left(\mathbf{U}^J\right)^2\right)}{P}. \end{aligned} \tag{28}$$

Proof: The detailed proof of (25) is presented Appendix A.

While the derived deterministic equivalent given in (25) is the limit of the instantaneous SINR expression in (24) as M goes to infinity. We verify in Section VI-C that the limit is accurate for a practical number of antennas.

2) PER-GROUP-PROCESSING (PGP)

To reduce CSI overhead, the PGP approach builds G different baseband precoder blocks requiring the effective channel for each individual group (i.e., $\boldsymbol{\mathcal{H}}_g = \mathbf{H}_g \mathbf{F}_g$). Particularly, only the diagonal block-matrices of the reduced-size effective matrix in (11) are required for the development of the baseband stage. Hence, the baseband precoder can be denoted by a block diagonal matrix as $\mathbf{B} = \mathrm{diag}\left(\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_G\right)$. Then, the RZF baseband precoder for group g is given by:

$$\mathbf{B}_g = \varepsilon_g \mathbf{W}_g \boldsymbol{\mathcal{H}}_g^H, \tag{29}$$

where $\mathbf{W}_g = \left(\boldsymbol{\mathcal{H}}_g^H \boldsymbol{\mathcal{H}}_g + K_g \alpha \mathbf{I}_{b_g}\right)^{-1}$ and ε_g is the normalization scalar for the group g regarding the transmit power constraint. By using (22) and considering equal power allocation, the normalization scalar is defined as:

$$\varepsilon_g^2 = \frac{P_T K_g}{K} \frac{1}{\mathrm{tr}\left(\boldsymbol{\mathcal{H}}_g \mathbf{W}_g^2 \boldsymbol{\mathcal{H}}_g^H\right)}. \tag{30}$$

By again using (2), the received signal at the k^{th} user in the group g is given by:

$$\begin{aligned} r_{g_k} &= \underbrace{\varepsilon_g \mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_k}^* d_{g_k}}_{\text{Desired Signal}} + \underbrace{\sum_{q \neq k}^{K_g} \varepsilon_g \mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_q}^* d_{g_q}}_{\text{Intra - Group Interference}} \\ &+ \underbrace{\sum_{p \neq g}^{G} \sum_{q=1}^{K_p} \varepsilon_p \mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathbf{F}_p^H \mathbf{h}_{p_q}^* d_{p_q}}_{\text{Inter - Group Interference}} + \underbrace{\mathbf{w}_{g_k}}_{\text{Noise}}, \end{aligned} \tag{31}$$

as the summation of the desired information bearing signal, the intra-group interference among the users in the group g, the inter-group interference due to the users located in all other groups and the noise. The inter-group interference results from the approximate zero condition given in (12) for a large array size. Then, the instantaneous SINR for the PGP approach is found as:

$$\mathrm{SINR}_{g_k}^{P} = \frac{\left|\mathbf{h}_{g_k}^{T}\mathbf{F}_g\mathbf{W}_g\mathbf{F}_g^{H}\mathbf{h}_{g_k}^{*}\right|^2}{\left\|\mathbf{h}_{g_k}^{T}\mathbf{F}_g\mathbf{W}_g\boldsymbol{\mathcal{H}}_{g,[k]}^{H}\right\|_2^2+\sum_{p\neq g}^{G}\frac{\varepsilon_p^2}{\varepsilon_g^2}\left\|\mathbf{h}_{g_k}^{T}\mathbf{F}_p\mathbf{W}_p\boldsymbol{\mathcal{H}}_p^{H}\right\|_2^2+\frac{\sigma^2}{\varepsilon_g^2}}, \tag{32}$$

where $\boldsymbol{\mathcal{H}}_{g,[k]} = \mathbf{H}_{g,[k]}\mathbf{F}_g$ is the reduced-size effective interference channel matrix with $\mathbf{H}_{g,[k]} = [\mathbf{h}_{g_1},\cdots,\mathbf{h}_{g_{k-1}},\mathbf{h}_{g_{k+1}},\cdots,\mathbf{h}_{g_{K_g}}]^T$. According to the deterministic equivalent principle for large antenna arrays [9], similar to the JGP approach, the deterministic equivalent of the instantaneous SINR given in (32) is denoted as $\overline{\mathrm{SINR}}_{g_k}^{P}$. The deterministic equivalent of the instantaneous SINR at the k^{th} user in the group g for the PGP approach is obtained as:

$$\overline{\mathrm{SINR}}_{g_k}^{P} = \frac{\left(m_g^P\right)^2}{\frac{(K_g-1)c_{g,g}^P}{P\left(1+m_g^P\right)^2}+\sum_{p\neq g}^{G}\frac{K_p c_g^P c_{p,g}^P}{c_p^P}+\frac{K c_g^P \sigma^2}{P_T}}, \tag{33}$$

where m_g^P for $g = 1,\cdots,G$ is the unique solution of:

$$m_g^P = \frac{1}{P}\mathrm{tr}\left(\mathbf{D}_g^P\mathbf{U}_g^P\right),$$
$$\mathbf{U}_g^P = \left(\frac{K_g-1}{P\left(1+m_g^P\right)}\mathbf{D}_g^P + K_g\alpha\mathbf{I}_{b_g}\right)^{-1}, \tag{34}$$

where $\mathbf{D}_g^P = \mathbf{F}_g^H\boldsymbol{\Phi}_g^H\boldsymbol{\Phi}_g\mathbf{F}_g$. Moreover, c_g^P and $c_{g,p}^P$ are respectively defined as:

$$c_g^P = \frac{\frac{1}{P}\mathrm{tr}\left(\mathbf{D}_g^P\left(\mathbf{U}_g^P\right)^2\right)}{1-\frac{K_g}{P^2}\frac{\mathrm{tr}\left(\left(\mathbf{D}_g^P\mathbf{U}_g^P\right)^2\right)}{\left(1+m_g^P\right)^2}},$$
$$c_{g,p}^P = \frac{\frac{1}{P^2}\mathrm{tr}\left(\mathbf{D}_g^P\mathbf{U}_g^P\mathbf{D}_{g,p}^P\mathbf{U}_g^P\right)}{1-\frac{K_g}{P^2}\frac{\mathrm{tr}\left(\left(\mathbf{D}_g^P\mathbf{U}_g^P\right)^2\right)}{\left(1+m_g^P\right)^2}}, \tag{35}$$

where $\mathbf{D}_{g,p}^P = \mathbf{F}_g^H\boldsymbol{\Phi}_p^H\boldsymbol{\Phi}_p\mathbf{F}_g$.

Proof: The detailed proof of (33) is given in Appendix B.

3) COMMON-GROUP-PROCESSING (CGP)

Recall that the AoD support for each group given in (13) has two dimensions on both x-axis and y-axis. For the PGP approach, in the case of intersection between AoD_g and AoD_p on either x-axis or y-axis (not on both axes), baseband precoder for groups g and p are separately designed as in (29). We also propose a new design strategy for the baseband stage as the CGP approach, when AoD supports for group g and p have common angles either on x-axis or y-axis. A systematic way to determine the user groups having the common AoD support is expressed in Algorithm 3. Then, the common baseband precoder for group g and p is given by:

Algorithm 3 Groups Having Common AoD Support

```
Input: AoD_g given in (13) for $g = 1, \cdots, G$.
Generate $(\gamma_x, \gamma_y) \in \mathrm{AoD}_g$ angle pairs.
if $(\gamma_x, \varpi) \in \mathrm{AoD}_p$ or $(\varpi, \gamma_y) \in \mathrm{AoD}_p$ for any $(\gamma_x, \gamma_y) \in \mathrm{AoD}_g$ and $\varpi \in [-1, 1]$, then
    Consider any group $g$ and $p$ as a common group.
    Design a baseband precoder for them via (36).
else
    Consider group $g$ as a single group.
    Design a baseband precoder for group $g$ via (29).
end if
Assign the total number of common/single groups to $G'$.
```

$$\mathbf{B}_{gp} = \varepsilon_{gp}\mathbf{W}_{gp}\boldsymbol{\mathcal{H}}_{gp}^{H}, \tag{36}$$

where $\mathbf{W}_{gp} = \left(\boldsymbol{\mathcal{H}}_{gp}^H\boldsymbol{\mathcal{H}}_{gp} + K_{gp}\alpha\mathbf{I}_{b_{gp}}\right)^{-1}$, $\boldsymbol{\mathcal{H}}_{gp} = \mathbf{H}_{gp}\mathbf{F}_{gp}$ is the concatenated reduced-size effective channel matrix constructed as a function of the concatenated channel matrix $\mathbf{H}_{gp} = \left[\mathbf{H}_g^T, \mathbf{H}_p^T\right]^T$ and the concatenated RF beamformer matrix $\mathbf{F}_{gp} = \left[\mathbf{F}_g, \mathbf{F}_p\right]$, $K_{gp} = K_g + K_p$ is the total number of users served by the common baseband precoder, $b_{gp} = b_g + b_p$ is the total number of quantized angle-pairs inside the concatenated RF beamformer, ε_{gp} is the normalization scalar given by:

$$\varepsilon_{gp}^2 = \frac{P_T K_{gp}}{K}\frac{1}{\mathrm{tr}\left(\boldsymbol{\mathcal{H}}_{gp}\mathbf{W}_{gp}^2\boldsymbol{\mathcal{H}}_{gp}^H\right)}. \tag{37}$$

Let G' denote the number of baseband precoder blocks with $1 \leq G' \leq G$, the baseband precoder for the CGP approach is obtained as $\mathbf{B} = \mathrm{diag}(\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_{G'})$. Also, CGP becomes identical to JGP for $G' = 1$ and to PGP for $G' = G$. Therefore, the channel estimation overhead for CGP is between that for JGP and PGP approaches. It is also important to point out that PGP and CGP approaches become identical for ULA structure due to its single-dimensional AoD support on either x-axis or y-axis.

By using (31) and (32), after some manipulations, the instantaneous SINR is obtained as in (38), as shown at the bottom of the next page, where $\boldsymbol{\mathcal{H}}_{gp,[g_k]} = \mathbf{H}_{gp,[g_k]}\mathbf{F}_{gp}$ is the reduced-size effective interference channel matrix with $\mathbf{H}_{gp,[g_k]} = [\mathbf{H}_{g,[k]}^T, \mathbf{H}_p^T]^T$. As in JGP and PGP, the deterministic equivalent of (38) is obtained as in (39), as shown at the bottom of the next page, where c_t^P and $c_{t,g}^P$ are given in (35) for PGP, m_g^C is the unique solution of:

$$m_g^C = \frac{1}{P}\mathrm{tr}\left(\mathbf{D}_{gp,g}^C\mathbf{U}_{gp,g}^C\right),$$
$$\mathbf{U}_{gp,g}^C = \left(\frac{(K_g-1)\mathbf{D}_{gp,g}^C}{P\left(1+m_g^C\right)}+\frac{K_p\mathbf{D}_{gp,p}^C}{P\left(1+m_p^C\right)}+K_{gp}\alpha\mathbf{I}_{b_{gp}}\right)^{-1}, \tag{40}$$

where $\mathbf{D}^C_{gp,g} = \mathbf{F}^H_{gp}\mathbf{\Phi}^H_g\mathbf{\Phi}_g\mathbf{F}_{gp}$. Also, c^C_g, c^C_p and $c^C_{g,p}$ are respectively defined as:

$$\begin{aligned}\left[c^C_g, c^C_p\right]^T &= \left(\mathbf{I} - \mathbf{Z}_{gp}\right)^{-1}\mathbf{v}_{gp},\\ \begin{bmatrix} c^C_{g,g} & c^C_{g,p}\\ c^C_{p,g} & c^C_{p,p}\end{bmatrix} &= \left(\mathbf{I} - \mathbf{Z}_{gp}\right)^{-1}\mathbf{V}_{gp},\end{aligned} \tag{41}$$

where $\mathbf{Z}_{gp}$, $\mathbf{V}_{gp}$ and $\mathbf{v}_{gp}$ are defined as:

$$\mathbf{Z}^C_{gp} = \begin{bmatrix} \dfrac{\mathrm{tr}\left(\left(\mathbf{D}^C_{gp,g}\mathbf{U}^C_{gp}\right)^2\right)}{P^2\left(1+m^J_g\right)^2/K_g} & \dfrac{\mathrm{tr}\left(\mathbf{D}^C_{gp,g}\mathbf{U}^C_{gp}\mathbf{D}^C_{gp,p}\mathbf{U}^C_{gp}\right)}{P^2\left(1+m^J_p\right)^2/K_p} \\ \dfrac{\mathrm{tr}\left(\mathbf{D}^C_{gp,p}\mathbf{U}^C_{gp}\mathbf{D}^C_{gp,g}\mathbf{U}^C_{gp}\right)}{P^2\left(1+m^J_g\right)^2/K_g} & \dfrac{\mathrm{tr}\left(\left(\mathbf{D}^C_{gp,p}\mathbf{U}^C_{gp}\right)^2\right)}{P^2\left(1+m^J_p\right)^2/K_p}\end{bmatrix},$$

$$\mathbf{V}^C_{gp} = \begin{bmatrix} \dfrac{\mathrm{tr}\left(\left(\mathbf{D}^C_{gp,g}\mathbf{U}^C_{gp}\right)^2\right)}{P^2} & \dfrac{\mathrm{tr}\left(\mathbf{D}^C_{gp,g}\mathbf{U}^C_{gp}\mathbf{D}^C_{gp,p}\mathbf{U}^C_{gp}\right)}{P^2} \\ \dfrac{\mathrm{tr}\left(\mathbf{D}^C_{gp,p}\mathbf{U}^C_{gp}\mathbf{D}^C_{gp,g}\mathbf{U}^C_{gp}\right)}{P^2} & \dfrac{\mathrm{tr}\left(\left(\mathbf{D}^C_{gp,p}\mathbf{U}^C_{gp}\right)^2\right)}{P^2}\end{bmatrix},$$

$$\mathbf{v}^C_{gp} = \frac{1}{P}\left[\mathrm{tr}\left(\mathbf{D}^C_{gp,g}\left(\mathbf{U}^C_{gp}\right)^2\right), \mathrm{tr}\left(\mathbf{D}^C_{gp,p}\left(\mathbf{U}^C_{gp}\right)^2\right)\right]^T. \tag{42}$$

Proof: The detailed proof of (39) is given in Appendix C.

V. REDUCING THE NUMBER OF RF CHAINS

The proposed two-stage AB-HP scheme in Section IV-A requires $b = \sum_{g=1}^{G} b_g$ RF chains in order to exploit all degrees of freedom provided by the channel. Moreover, the total number of transmitted independent data streams should be less than or equal to the number of orthogonal-beams (i.e., $K \le b$). Hence, in comparison to the single-stage FDP, the number of RF chains can be reduced from M to b via the proposed two-stage AB-HP by placing b_g RF chains for each user group. Indeed, b_g is dictated by the AoD support of the group g given in (13) to design $\mathbf{F}_g$ given in (20).

To further minimize the hardware cost/complexity in the massive MIMO systems, the number of RF chains for the group g, $N_{RF,g}$,should be as low as the number of independent data streams in that group. However, when $N_{RF,g}$ is smaller than b_g, the RF beamformer $\mathbf{F}_g$ can only contain $N_{RF,g}$ angles and cannot span the complete range of the phase response matrix $\mathbf{\Phi}_g$. As a result, some of the paths cannot be used for transmission and the sum-rate is subsequently lower than using all paths via b_g RF chains. In order to fully exploit all paths while reducing the number of RF chains, we propose a new transfer block architecture illustrated in Figure 4, which only requires $N_{RF,g} = K_g$ RF chains for each user group without affecting the sum-rate performance. Hence, the proposed transfer block architecture is capable to reduce the number of RF chains exactly to $N_{RF} = \sum_{g=1}^{G} N_{RF,g} = \sum_{g=1}^{G} K_g = K$ without sacrificing the performance.

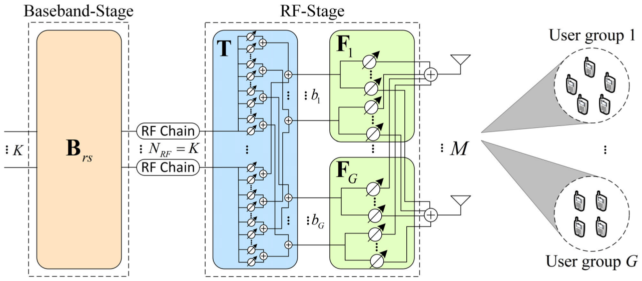


FIGURE 4. System model for the proposed three-stage AB-HP with transfer block.

A. PROBLEM FORMULATION

Given N_{RF} RF chains, the baseband precoder outputs is related to all b angles contained in RF beamformer as illustrated in Figure 2. In other words, the transmitted signal from each RF chain should be mapped to cover all b angles, even when $N_{RF} < b$. Accordingly, we propose to include a transfer block denoted by $\mathbf{T} \in \mathbb{C}^{b \times N_{RF}}$ between the RF beamformer $\mathbf{F} \in \mathbb{C}^{M \times b}$ and the reduced-size baseband precoder $\mathbf{B}_{rs} \in \mathbb{C}^{N_{RF} \times K}$ as represented in Figure 4, where the transfer block is placed at the RF-stage of AB-HP. Hence, the transfer block matrix should be realized via phase-shifters similar to the RF beamformer. According to the transfer block architecture, the transmitted signal from the BS can be rewritten as $\mathbf{s} = \mathbf{F}\mathbf{T}\mathbf{B}_{rs}\mathbf{d}$, which satisfies the maximum transmit power constraint of P_T, i.e., $\mathbb{E}\left\{\|\mathbf{s}\|^2_2\right\} = \mathrm{tr}\left(\mathbf{B}^H_{rs}\mathbf{T}^H\mathbf{T}\mathbf{B}_{rs}\right) \le P_T$.

For a given baseband precoder $\mathbf{B}$ and b RF chains, it is desired to find the transfer block $\mathbf{T}$ and the reduced-size

$$\mathrm{SINR}^C_{g_k} = \frac{\varepsilon^2_{gp}\left|\mathbf{h}^T_{g_k}\mathbf{F}_{gp}\mathbf{W}_{gp}\mathbf{F}^H_{gp}\mathbf{h}^*_{g_k}\right|^2}{\varepsilon^2_{gp}\left\|\mathbf{h}^T_{g_k}\mathbf{F}_{gp}\mathbf{W}_{gp}\boldsymbol{\mathcal{H}}^H_{gp,[g_k]}\right\|^2_2 + \sum\limits_{t\neq g,p}^{G'}\varepsilon^2_t\left\|\mathbf{h}^T_{g_k}\mathbf{F}_t\mathbf{W}_t\boldsymbol{\mathcal{H}}^H_t\right\|^2_2 + \sigma^2}. \tag{38}$$

$$\overline{\mathrm{SINR}}^C_{g_k} = \frac{\left(m^C_g\right)^2}{\frac{(K_g-1)c^C_{g,g}}{\left(1+m^C_g\right)^2} + \frac{K_p c^C_{g,p}}{\left(1+m^C_p\right)^2} + \frac{\left(1+m^C_g\right)^2}{K_{gp}}\left(\sum\limits_{t=g,p}\frac{K_t c^C_t}{\left(1+m^C_t\right)^2}\right)\left(\sum\limits_{t\neq g,p}^{G'}\frac{K_t c^P_{t,g}}{c^P_t} + \frac{K}{P_T}\sigma^2\right)}. \tag{39}$$

baseband precoder $\mathbf{B}_{rs}$ that approximate $\mathbf{B}$ while reducing the number of RF chains. The problem of interest is accordingly formulated as:

$$\min_{\mathbf{T},\mathbf{B}_{rs}} \|\mathbf{B}-\mathbf{T}\mathbf{B}_{rs}\|_2^2 ,$$
$$\text{s.t. } \mathbf{T} \in \mathcal{T}_{RF},\ \operatorname{tr}\left(\mathbf{B}_{rs}^H\mathbf{T}^H\mathbf{T}\mathbf{B}_{rs}\right) \le P_T, \tag{43}$$

where $\mathcal{T}_{RF}$ represents the set of matrices with the size of $b \times N_{RF}$ satisfying the aforementioned unit-modulus property for the transfer block constructed by the phase-shifters. However, (43) is a non-convex optimization problem since the elements of $\mathbf{T}$ have unit-modulus.

B. TRANSFER BLOCK DESIGN

Solving (43) for $\mathbf{T}$ and $\mathbf{B}_{rs}$ is a difficult problem because of the non-convexity of the constraint $\mathbf{T} \in \mathcal{T}_{RF}$. We propose to represent the transfer block as a summation of two inner matrices as follows:

$$\mathbf{T} = (\mathbf{T}_A + \mathbf{T}_B), \quad \text{s.t. } \mathbf{T}_A, \mathbf{T}_B \in \mathcal{T}_{RF}, \tag{44}$$

where $\mathbf{T}_A \in \mathbb{C}^{b\times K}$ and $\mathbf{T}_B \in \mathbb{C}^{b\times K}$ are inner transfer block matrices requiring only $N_{RF} = K$ RF chains. Therefore, each baseband precoder output is passed through two phase-shifters and their summation is connected to the corresponding RF beamformer input as shown in Figure 4. One can show that when two complex numbers having the unit-modulus are summed, then its modulus varies between 0 and 2. Therefore, by defining $\mathbf{T} = (\mathbf{T}_A + \mathbf{T}_B)$, the unit modulus constraint for $\mathbf{T}$ expressed in (43) can be converted to a modulus constraint within 0 and 2. According to (44), the optimization problem given in (43) can be reorganized as:

$$\min_{\mathbf{T}_A,\mathbf{T}_B,\mathbf{B}_{rs}} \|\mathbf{B}-(\mathbf{T}_A+\mathbf{T}_B)\,\mathbf{B}_{rs}\|_2^2 ,$$
$$\text{s.t. } \mathbf{T}_A, \mathbf{T}_B \in \mathcal{T}_{RF}, \mathbf{T} = \mathbf{T}_A + \mathbf{T}_B, \operatorname{tr}\left(\mathbf{B}_{rs}^H\mathbf{T}^H\mathbf{T}\mathbf{B}_{rs}\right) \le P_T. \tag{45}$$

Using Lemma 1 and applying the least-squares solution [45], the optimal solution for the inner transfer block and reduced-size baseband precoder matrices can be found as follows:

$$\begin{aligned} \mathbf{T}_A(i,q) &= e^{j\left(\angle\mathbf{B}(i,q)+\cos^{-1}\left(\frac{|\mathbf{B}(i,q)|}{2\mu}\right)\right)}, \\ \mathbf{T}_B(i,q) &= e^{j\left(\angle\mathbf{B}(i,q)-\cos^{-1}\left(\frac{|\mathbf{B}(i,q)|}{2\mu}\right)\right)}, \\ \mathbf{B}_{rs} &= (\mathbf{T}_A+\mathbf{T}_B)^{\dagger}\mathbf{B}. \end{aligned} \tag{46}$$

where $\mu = \frac{1}{2}\max_{i,q}|\mathbf{B}(i,q)|$ is the half of the highest modulus element at the baseband precoder $\mathbf{B}$. Afterwards, one can prove that (46) minimizes the expression given in (45) by satisfying $\mathbf{B} = (\mathbf{T}_A + \mathbf{T}_B)\,\mathbf{B}_{rs}$.

In a nutshell, to further reduce the hardware cost/complexity of the proposed AB-HP shown in Figure 2, we propose a new three-stage architecture by including the additional transfer block capable of reducing the number of RF chains N_{RF} exactly to the total number of data streams K as shown in Figure 4. Furthermore, the three-stage architecture provides the same sum-rate performance as in two-stage, which is validated later in Section VI-D. The functions of each stage can be simply explained as follows: (i) the RF beamformer designed by slowly time-varying AoD information given in (13) reduces the CSI overhead as well as the hardware cost/complexity via reducing the number of RF chains from M to b, (ii) the transfer block based on the effective channel matrix given in (11) minimizes the RF chain utilization by decreasing its number from b to K, (iii) the reduced-size baseband precoder adjusts the transmit power and mitigates the interference among the users. Furthermore, for future work (beyond the scope of this paper), the last stage can be also utilized for power allocation among the users to enhance the sum-rate capacity of the system [9].

TABLE 1. Number of RF Chains and Required CSI Size for AB-HP and FDP.

Precoding Scheme	# of RF chains	Required CSI Size
FDP	M	$M \times K$
AB-HP with JGP	K	$b \times K$
AB-HP with CGP	K	$b_{gp}\times K_{gp}+\sum_{t\neq g,p}^{G'} b_t\times K_t$
AB-HP with PGP	K	$\sum_g^G b_g \times K_g$

Finally, Table 1 summarizes the hardware cost/complexity and the CSI overhead for the proposed AB-HP in comparison to FDP. According to the three-stage architecture, the number of RF chains is reduced exactly to K, which is equal to M in the case of FDP. Regarding the required CSI size, each JGP, CGP and PGP approach expressed in Section IV-B has different CSI overhead according to the effective channel matrix size utilized for the baseband precoder.

VI. ILLUSTRATIVE RESULTS AND DISCUSSIONS

Monte-Carlo simulation results for the sum-rate performance of the proposed AB-HP in a single-cell environment are presented and compared to the approximations achieved by the deterministic expressions. Using the instantaneous SINR expressions given in (24), (32) and (38), the ergodic sum-rate capacity curves for JGP, PGP and CGP approaches are generated by:

$$R_{\text{sum}}^{\upsilon} = \sum_{g=1}^{G}\sum_{k=1}^{K_g}\mathbb{E}\Big[\log_2\left(1+\text{SINR}_{gk}^{\upsilon}\right)\Big] \text{ [bps/Hz]}, \tag{47}$$

where $\upsilon \in \{J, P, C\}$. Furthermore, the approximated sum-rate capacity via the deterministic expressions given in (25), (33) and (39) is obtained by:

$$R_{\text{sum,approx}}^{\upsilon} = \sum_{g=1}^{G}\sum_{k=1}^{K_g}\log_2\left(1+\overline{\text{SINR}}_{gk}^{\upsilon}\right) \text{ [bps/Hz]}. \tag{48}$$

According to the 3D urban micro-cellular scenario in [46], Table 2 summarizes the numerical values of the parameters used in the simulation setup, unless otherwise stated, where the user height and the user-BS horizontal distances are

TABLE 2. Simulation parameters.

Cell radius [45]	100m
BS height [45]	10m
User height [45]	1.5m-2.5m
User-BS horizontal distance	15m-95m
# of user groups	$G = 6$
Mean azimuth angle	$\psi_g = 20^\circ + \frac{360^\circ}{G}(g-1)$
Azimuth angle spread [45]	$\delta_g^\psi = 9^\circ$
Mean elevation angle [45]	$\theta_g = 73^\circ$
Elevation angle spread [45]	$\delta_g^\theta = 12.5^\circ$
# of paths	$P = 10$
Antenna spacing (in wavelength)	$d = 0.5$
# of Monte-Carlo simulations	2000

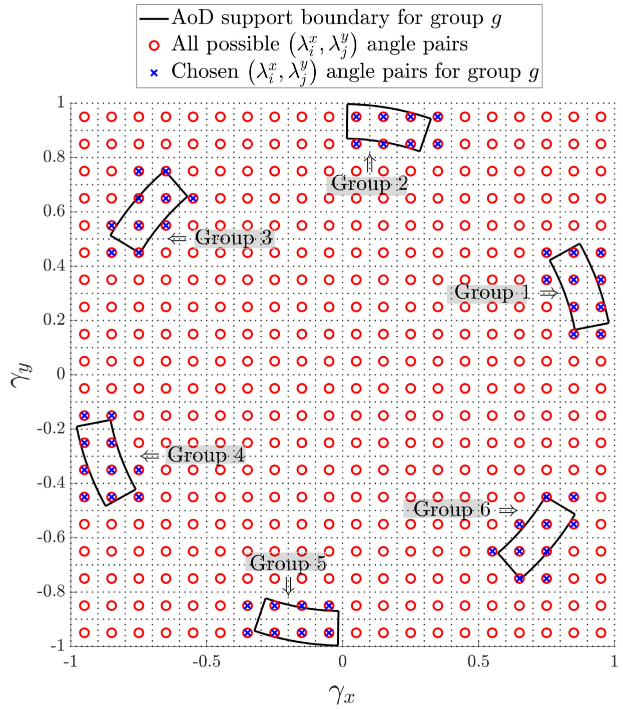


FIGURE 5. RF beamformer design and quantized angle-pairs for 20 × 20 URA.

assumed to be uniformly distributed in the specified range. Based on the geometry, the mean elevation angle and elevation angle spread can be respectively calculated as $\theta_g = 73^\circ$ and $\delta_g^\theta = 12.5^\circ$ by using [46, eq. (7.4-1)]. As in [47], the regularization parameter is chosen as $\alpha = \frac{\sigma^2}{P_T}$. We define the signal-to-noise ratio (SNR) as $\text{SNR} = \frac{P_T}{K\sigma^2}$.

A. RF BEAMFORMER DESIGN

In Figure 5, the RF beamformer design expressed in Algorithm 2 is visualized for 20 × 20 URA. According to the URA configuration, there are 20 quantized angles along both x-axis and y-axis. Totally, there are $20 \times 20 = 400$ quantized angle-pairs plotted with red circles, which is equal to the number of antennas. Firstly, based on the AoD information given in Table 2 (i.e., ψ_g, θ_g, δ_g^ψ, δ_g^θ), the AoD support for each user group is generated by using (13). Then, the quantized angle-pairs covering the corresponding AoD support are determined according to (19), which are represented by blue crosses. Furthermore, the total number of orthogonal-beams for each group (i.e., b_g for group g given in (20)) can be easily found by counting the blue-crosses. For example, there are $b_1 = 10$ quantized angle-pairs in order to cover the AoD support of group 1. Similarly, group 2 requires $b_2 = 8$ quantized angle-pairs for the RF beamformer design in order to exploit all the degrees of freedom provided by the channel matrix. As a result, the total number of quantized angle-pairs to serve all user groups can be found as $b = \sum_{g=1}^{G} b_g = 56$, where $b_1 = b_3 = b_4 = b_6 = 10$ and $b_2 = b_5 = 8$. On the other hand, the group 1 has an overlapping AoD support with the group 6 along x-axis and the group 3 along y-axis. Similarly, the AoD support of the group 4 overlaps with the group 3 along x-axis and the group 6 along y-axis. According to Algorithm 3, these four groups are considered as a common group in the CGP approach. However, neither the group 2 nor 5 has an overlapping AoD support with other groups. Therefore, the CGP approach designs $G' = 3$ baseband precoder blocks as the first two separate blocks for the group 2 and the group 5 by considering each of them as a single group and the last block for the remaining four groups by considering them as a common group.

B. USER GROUPING

In the following, we compare the performance of the proposed user grouping algorithm against the direct applications of K-means algorithm with chordal distance [24], weighted likelihood [36] and subspace projection [36] similarity measures, which are iterative methods and require a priori knowledge of the number of user groups. Figure 6 demonstrates the superiority of the proposed user grouping algorithm by correctly finding the number of user groups G. It is worthwhile to remark that user grouping algorithms in [24], [36] employ the channel covariance matrix of each user, whereas the proposed user grouping algorithm is based on the AoD information. For all figures, the user locations are plotted in 2D scale, the cross symbols in the middle represent the BS, the straight-lines demonstrate the user group boundaries on the azimuth angle domain and the users within the same group have the same mark and color. We assume that there are $K_g = 6$ users per each group and $K = 36$ users in total. In Figure 6(a), our proposed user grouping approach is applied to first find the number of groups and then cluster the users based on the eigenvectors of the Laplacian matrix defined in (6) having the eigenvalue 0. From Figure 6(a), the proposed algorithm finds the correct number of user groups $G = 6$. Then, applying K-means algorithm to $\mathbf{Q}_G$ given in (8) provides well clustered users. On the other hand, the direct application of K-means algorithm as in [24], [36] with an incorrect priori knowledge of the number of user groups results in unsatisfactory clustering. When the input number of groups to K-means is $G = 5$ (Figure 6(b)), which is lower than its actual value, the well-separated user groups having the disjoint AoD supports (e.g., yellow circles in chordal distance, green diamonds in weighted likelihood and blue squares in

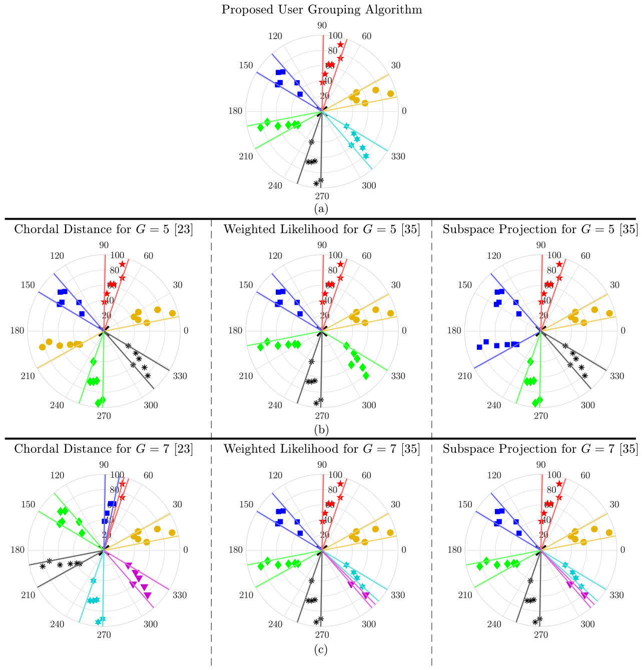


FIGURE 6. User grouping for the user deployment as in Table 2 with $K_g = 6$ and $K = 36$: (a) the proposed approach expressed in Algorithm 1 and the direct application of K-means algorithm with chordal distance [24], weighted likelihood [36] and subspace projection [36] similarity measures with an incorrect priori knowledge of (b) $G = 5$ and (c) $G = 7$.

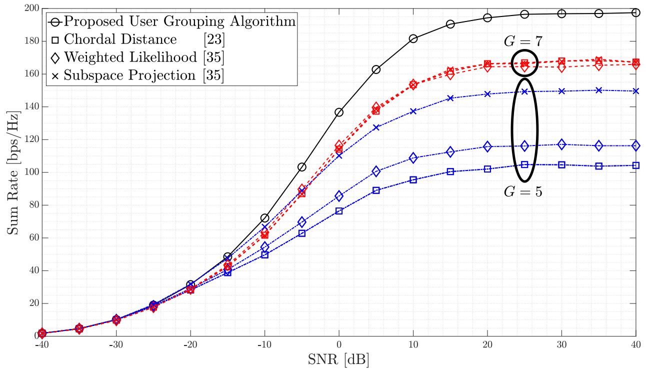


FIGURE 7. Sum-rate of AB-HP with PGP for 20 × 20 URA considering all user grouping techniques applied in Figure 6, where $K_g = 6$ and $K = 36$.

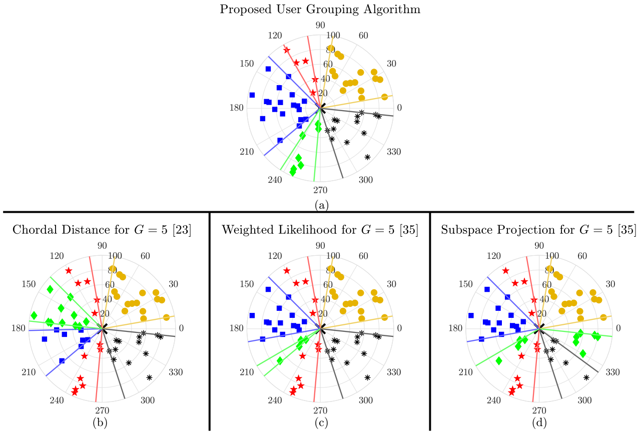
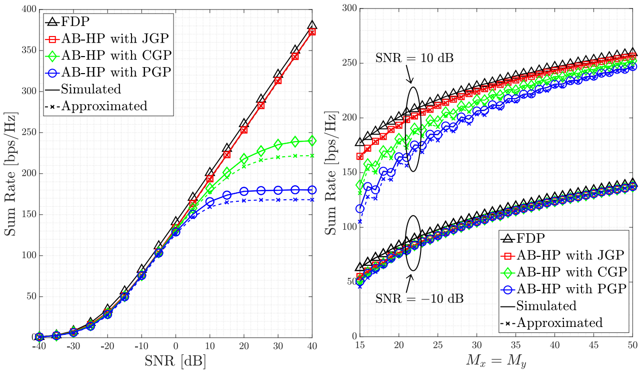


FIGURE 8. User grouping for the random user deployment with $K = 60$: (a) the proposed approach expressed in Algorithm 1 and the direct application of K-means algorithm with (b) chordal distance [24], (c) weighted likelihood [36] and (d) subspace projection [36] with a priori knowledge of $G = 5$.

subspace projection) are accidentally clustered in the same group. Considering the phase response matrix given in (3), the well-separated users symmetric about either x-axis (e.g., blue squares in subspace projection) or y-axis (e.g., green diamonds in weighted likelihood) have similar covariance matrices based on the symmetry property of cosine function about x-axis (i.e., $\cos(\alpha) = \cos(-\alpha)$) and sine function about y-axis (i.e., $\sin(\alpha) = \sin(\pi - \alpha)$). In Figure 6(c), a priori assumption for the number of groups $G = 7$ exceeds its correct value $G = 6$. The users having similar AoD supports (e.g., blue squares and red stars in chordal distance, purple triangles and turquoise hexagrams in weighted likelihood and subspace projection) can be mistakenly clustered in different groups, which may cause additional inter-group interference between those user groups.

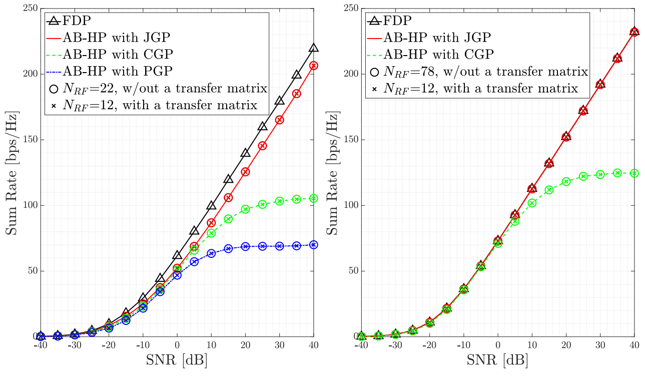
Figure 7 plots the sum-rate performance of the proposed AB-HP with PGP technique for 20 × 20 URA considering all user grouping techniques applied in Figure 6. The first and the most crucial observation is that the proposed user grouping algorithm can provide a superior performance by means of finding the correct number of user groups and well clustering them. On the other hand, when the number of user groups is presumed as $G = 7$ for K-means, shown in Figure 6(c), all K-means techniques provide almost the same sum-rate performance. Furthermore, the inter-group interference increases due to the similar AoD support of user groups and the sum-rate curves quickly saturate as the SNR increases. Besides, when the number of user groups is considered as $G = 5$ shown in Figure 6(b), subspace projection can provide better sum-rate performance compared to chordal distance and weighted likelihood similarity measures. The reason for this behaviour is that the AoD support boundaries for each user group are non-overlapping in the case of subspace projection.

The case of randomly deployed users is illustrated in Figure 8 with $K = 60$ users. The mean azimuth angle of each user ψ_k varies between 0° and 360° based on their location. In Figure 8(a), the number of user groups are estimated as $G = 5$ via the proposed user grouping algorithm, then the users are clustered. As seen from the results, the azimuth angle spread of each user group varies due to the randomness. For instance, the user group illustrated via blue squares has a wider azimuth angle spread, however, it is narrower for the user group illustrated via red stars. On the other hand, the proposed approach successfully clusters the user groups by creating well-separation on the azimuth angle domain, which is important to mitigate the inter-group interference. However, by using a priori knowledge of user groups number as $G = 5$, the direct application of K-means algorithm with chordal distance [24] in Figure 8(b), weighted likelihood [36]



(a) (b)

FIGURE 9. Sum-rate and its approximation (a) versus SNR for 20 × 20 URA and (b) versus square array sizes for SNR = ±10 dB, where $K_g = 3$ and $K = 18$.



(a) (b)

FIGURE 10. Sum-rate versus SNR: (a) 10 × 10 URA and (b) 1 × 100 ULA, where $K_g = 2$ and $K = 12$.

in Figure 8(c) and subspace projection [36] in Figure 8(d) cannot provide fine separation. To illustrate, in the K-means algorithm, the well-separated red star users, which are symmetric about x-axis, are clustered in the same group due to their similar covariance matrices as mentioned above. Also, the AoD support for red star users covers the AoD support of the green diamond and blue square users.

C. VALIDATION OF THE DETERMINISTIC ANALYSES

Figure 9 shows the sum-rate curves obtained from the derived deterministic SINR expressions and from Monte Carlo simulations, where there are $K_g = 3$ users in each group. Moreover, it also compares the sum-rate capacity of the proposed AB-HP employing all three different baseband precoder design approaches (i.e., JGP, CGP and PGP) with the single-stage FDP.

In Figure 9(a), the sum-rate curves for 20 × 20 URA are plotted versus SNR. According to Table 1 and the chosen quantized angle-pairs for 20 × 20 URA shown in Figure 5, the required CSI sizes are as 56 × 18 for JGP, $(40 \times 12) + (16 \times 3)$ for CGP and 56 × 3 for PGP. Noting that the FDP needs the entire CSI with the size of 400 × 18, the CSI overhead is decreased by 86% via JGP, 92.67% via CGP and 97.67% via PGP. It is shown that all three proposed AB-HP techniques can provide a comparable performance to FDP with only 1.2 dB performance gap in low to medium SNR regimes. However, in the high SNR regimes, the proposed AB-HP with CGP and PGP reaches a performance floor due to the inter-group interference level created by the approximate zero condition in (12). On the other hand, the CGP approach is capable of mitigating the inter-group interference better than PGP by jointly designing the baseband precoder for the groups, which have the common AoDs along either x-axis or y-axis as explained in Figure 5. Hence, CGP can provide higher capacity than PGP as expected. For instance, the performance floors observed for CGP and JGP are approximately 240 bps/Hz and 180 bps/Hz, respectively. Another important observation is that the performance gap between the proposed AB-HP and FDP remains constant for the JGP approach as the SNR increases, so it can avoid the performance floor problem by means of jointly designing the baseband precoder. Furthermore, the numerical results illustrate that the approximated sum-rate is a tight and accurate approximation of the ergodic sum-rate expression, especially in the low and medium SNR regimes. Hence, it validates the correctness of the derived deterministic equivalent SINR expressions for JGP, PGP and CGP approaches given in (25), (33) and (39), respectively.

In Figure 9(b), the sum-rate curves are produced versus various square array sizes having $M_x = M_y$ for SNR $= \pm 10$ dB. It is observed that the sum-rate gap between the proposed AB-HP and FDP vanishes as the array size increases. From Figure 9(b), the sum-rate curves for CGP are always between those for JGP and PGP for all array sizes. Moreover, we can also see that the obtained deterministic equivalent SINR expression for JGP can tightly approximate the actual SINR even for small number of antennas. Regarding the performance floor occurred in the high SNR regimes for CGP and PGP, the approximations become more accurate as the number of antennas increases. For example, when we employ 15 × 15 URA using AB-HP with CGP at SNR = 10 dB, the approximated sum-rate value as 105 bps/Hz is 10.26% below its actual value as 117 bps/Hz. But, the difference decreases to only 0.37% for 50 × 50 URA configuration.

D. REDUCING THE NUMBER OF RF CHAINS

Figure 10 illustrates the sum-rate comparison between the proposed AB-HP and FDP for the utilization of 10 × 10 URA and 1 × 100 ULA configurations, both having $M = 100$ antennas to serve $K_g = 2$ users per group. We mention to remind that PGP is equivalent to CGP in the case of ULA due to its 1D structure, so the sum-rate curves are only generated for JGP and CGP in Figure 10(b). The "circle" signs represent the two-stage AB-HP requiring $N_{RF} = b$ RF chains demonstrated in Figure 2, whereas the "cross" signs represent the three-stage AB-HP requiring $N_{RF} = K = 12$ RF chains demonstrated in Figure 4. We numerically show that the hardware cost/complexity of AB-HP can be minimized without sacrificing the sum-rate by employing

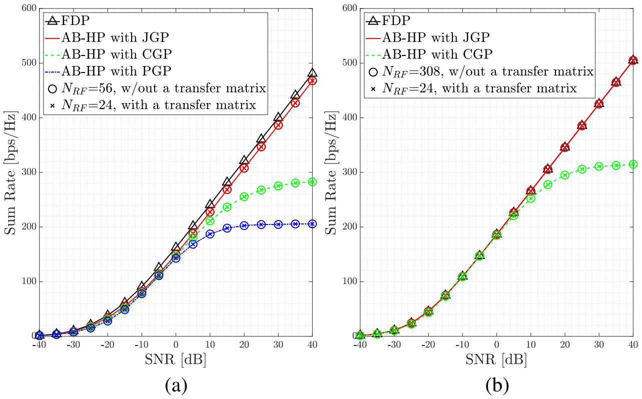


FIGURE 11. **Sum-rate versus SNR: (a) 20 × 20 URA and (b) 1 × 400 ULA, where $K_g = 4$ and $K = 24$.**

a transfer matrix, which reduces the number RF chains to $N_{RF} = 12$. Compared to FDP with $N_{RF} = M = 100$, the proposed AB-HP with JGP suffers 3 dB in SNR in the case of URA (Figure 10(a)), but offers a comparable the performance in the case of ULA (Figure 10(b)). On the other hand, although the number of RF chains can be decreased from $b = 22$ to $K = 12$ for URA and $b = 78$ to $K = 12$ for ULA, the required CSI size given in Table 1 is still function of b. Therefore, considering the antenna array configuration, it brings an interesting trade-off between the CSI overhead and sum-rate performance.

Figure 11 shows the sum-rate performance versus SNR for a larger array having $M = 400$ antennas, i.e., 20 × 20 URA and 1×400 ULA, for $K_g = 4$. In comparison to Figure 10(a), although doubling the number of users in each group from $K_g = 2$ to $K_g = 4$ increases the interference, the performance difference between FDP and AB-HP with JGP in Figure 11(a) decreases to 1.4 dB in SNR due to larger array size. The main factor for this outcome is that the RF beamformer expressed in (20) has a better angle resolution, when the number of antennas increases. Hence, the RF beamformer enables the higher beamforming gain via focusing the signal energy on the region, where the user groups are located. Furthermore, the performance gap becomes indistinguishable for ULA illustrated in Figure 11(b).

E. IMPERFECT CSI

In the previous studies, we considered perfect CSI (P-CSI) for designing of AB-HP and FDP. In practical applications, CSI is contaminated by estimation error and the precoder matrices are constructed via the imperfect CSI (I-CSI). In presence of channel estimation error, similar to [23], [48], we assume that the BS has a perfect estimate of AoD information because it varies slower than the fast time-varying instantaneous CSI. Therefore, the RF-stage of AB-HP is not affected by I-CSI. On the other hand, the single-stage FDP requires the full I-CSI. In this study, two different models are considered regarding the channel estimation error.

In the first model, we assume that the variance of channel estimation error is fixed for all SNR values as in [48]–[50].

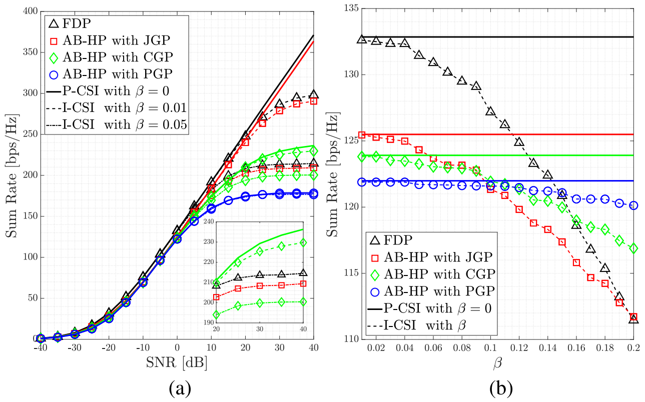


FIGURE 12. **Sum-rate comparison of P-CSI and I-CSI based on the model given in (49) for 20 × 20 URA: (a) versus SNR and (b) versus the quality of CSI estimation (β) for SNR = 0 dB, where $K_g = 3$ and $K = 18$.**

According to the effective P-CSI seen from the baseband given in (11), the effective I-CSI can be given by:

$$\tilde{\mathcal{H}} = \sqrt{1-\beta^2}\,\mathcal{H} + \beta \boldsymbol{E}_1, \tag{49}$$

where $\boldsymbol{E}_1 \sim \mathcal{CN}(0, \boldsymbol{I}_b)$ and the parameter β characterizes the quality of CSI estimation. Hence, $\beta = 0$ represents the P-CSI case and $0 < \beta \leq 1$ represents the I-CSI case at the BS. In Figure 12, the effect of channel estimation error is investigated for both the proposed AB-HP and FDP, where the BS is equipped with 20 × 20 URA to serve $K_g = 3$ users per group. The sum-rate curves are plotted versus SNR in Figure 12(a), where $\beta = 0$ and $\beta = \{0.01, 0.05\}$ represent P-CSI and I-CSI, respectively. In addition to AB-HP with CGP and PGP, when the BS utilizes I-CSI for both FDP and AB-HP with JGP, these precoding techniques also reach a performance floor in the high SNR regimes due to the fixed channel estimation error defined in (49). However, in the low and medium SNR regimes, the sum-rate performance under I-CSI with both $\beta = 0.01$ and $\beta = 0.05$ remains approximately identical as in P-CSI. In Figure 12(b), the sum-rate performance is plotted versus the quality of CSI estimation (i.e., $0.01 \leq \beta \leq 0.2$) for SNR $= 0$ dB, where the straight lines demonstrate the case of P-CSI. It is shown that AB-HP with PGP is more robust to channel estimation errors compared to other precoding techniques due to the smaller CSI overhead as shown in Table 1. On the other hand, the performance of FDP requiring full CSI deteriorates in the case of high estimation error. To illustrate, in comparison to P-CSI, when $\beta = 0.2$, the sum-rate of AB-HP with PGP only decreases from 122 bps/Hz to 120 bps/Hz, whereas the sum-rate of FDP declines from 133 bps/Hz to 111 bps/Hz.

In the second model, we assume that the power of the estimation error is inversely proportional to SNR and the number of pilot symbols used during the training phase as in [51]–[53]. According to the effective P-CSI seen from the baseband given in (11), the effective I-CSI can be given by:

$$\tilde{\mathcal{H}} = \mathcal{H} + \boldsymbol{E}_2, \tag{50}$$

where $\boldsymbol{E}_2 \sim \mathcal{CN}\big(0, \frac{\sigma^2}{\zeta P_T}\boldsymbol{I}_b\big)$. Here, ζ depends on the number of pilot symbols used during the training phase and the applied channel estimation method [51]–[53]. So, ζ approaches to infinity for the case of P-CSI. Based on the second model given in (50), the sum-rate comparison of P-CSI and I-CSI is presented for 20 × 20 URA and $K_g = 3$ in Figure 13. Due to the smaller required CSI size, AB-HP with PGP can outperform FDP in the low and medium SNR regimes. For example, when $\zeta = 1$ and SNR $\leq$ 5.8 dB, the sum-rate of AB-HP with PGP is higher than its of FDP as shown in Figure 13(a). Moreover, in contrast to the first model given in (49), as long as SNR increases, no performance floor behavior is observed for neither AB-HP with JGP nor FDP because the power of the estimation error is inversely proportional to SNR in the second model given in (50). In Figure 13(b), the sum-rate curves are produced versus the quality of CSI estimation (i.e., $1 \leq \zeta \leq 20$) for SNR = 0 dB, where the straight lines display the case of P-CSI. Similar to the first model represented in Figure 12(b), AB-HP with PGP is again more robust to the channel estimation errors.

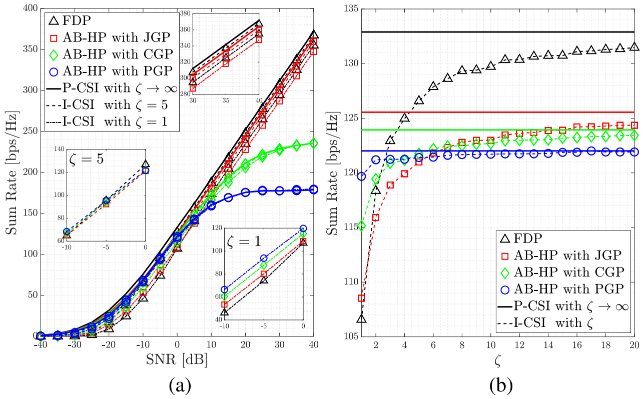


FIGURE 13. Sum-rate comparison of P-CSI and I-CSI based on the model given in (50) for 20 × 20 URA: (a) versus SNR and (b) versus the quality of CSI estimation (ζ) for SNR = 0 dB, where $K_g = 3$ and $K = 18$.

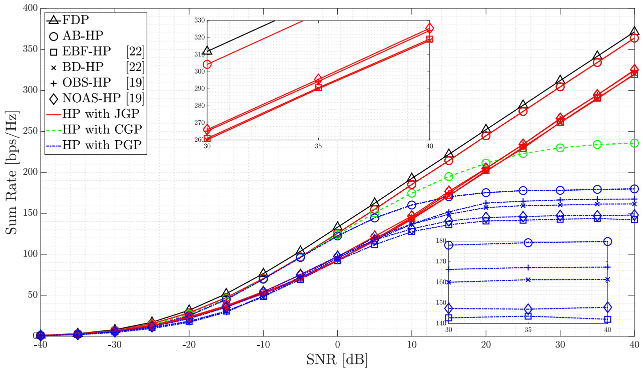


FIGURE 14. Sum-rate comparison of FDP and HP techniques versus SNR for 20 × 20 URA, where $K_g = 3$ and $K = 18$.

F. BENCHMARK

In addition to the single-stage FDP, the proposed AB-HP technique is also compared with the exiting HP solutions as eigen-beamforming based HP (EBF-HP) [23], block-diagonalization based HP (BD-HP) [23], orthogonal beam selection based HP (OBS-HP) [20] and non-orthogonal angle space based HP (NOAS-HP) [20]. It is important to remark that although EBF-HP and BD-HP use the channel covariance matrix as the slowly time-varying CSI for the RF beamformer design, OBS-HP and NOAS-HP construct the RF beamformer via instantaneous CSI and require complete channel knowledge. By defining $\mathbf{R}_g \in \mathbb{C}^{M\times M}$ as the covariance matrix of $\mathbf{H}_g = \mathbf{G}_g\boldsymbol{\Phi}_g \in \mathbb{C}^{K_g\times M}$, the eigenvectors of $\mathbf{R}_g$ with dominant eigenvalues are utilized to design the RF beamformer for group g in EBF-HP and BD-HP. Considering that the channel covariance matrix may change over time under the same AoD support, then RF beamformer is not required to be updated in the proposed AB-HP scheme, however, EBF-HP and BD-HP need to reconstruct the RF beamformer for every channel covariance matrix. Moreover, due to the non-constant modulus characteristic of the eigenvectors, EBF cannot satisfy the constant modulus requirement and the RF beamformer cannot be realized via phase-shifters. Even though non-constant modulus elements are utilized at the RF beamformer of EBF-HP and BD-HP and the full CSI knowledge is required for OBS-HP and NOAS-HP, they nevertheless serve as benchmark for the sum-rate performance. In terms of the baseband precoder design, we consider that either JGP or PGP is employed for all benchmark techniques.

In Figure 14, the sum-rate curves are plotted versus SNR, where all HP schemes employ $N_{RF} = K = 18$ RF chains while $N_{RF} = 400$ for FDP. The numerical results imply that although NOAS-HP provides the best sum-rate performance among all benchmark techniques in JGP, the proposed AB-HP with JGP outperforms NOAS-HP with JGP by 6.5 dB. The primary factor for this performance enhancement is the proposed transfer block design capable of exploiting all degrees of freedom provided by the channel, while NOAS-HP selects only a set of significant beams as many as number of RF chains. Furthermore, in the high SNR regimes, when the PGP approach is utilized at the baseband precoder design, the performance floor for OBS-HP at 167 bps/Hz, BD-HP at 161 bps/Hz, NOAS-HP at 148 bps/Hz and EBF-HP at 142 bps/Hz can be increased to 180 bps/Hz via AB-HP. In other words, the proposed AB-HP with PGP is capable of providing 7.2%, 10.6%, 17.8% and 21.1% sum-rate enhancement with respect to OBS-HP, BD-HP, NOAS-HP and EBF-HP, respectively. There are two main reasons for the beneficial aspect of the proposed AB-HP with PGP. Firstly, the RF beamformer better deal with the effect of the inter-group interferences by providing the approximate zero condition given in (12) satisfied by (16), which cannot be provided by EBF-HP and NOAS-HP. Secondly, the direct application of the AoD support boundary for the RF beamformer given in (20) provides higher beamforming gain instead of only selecting the significant beams within the AoD support in OBS-HP and employing the eigenvectors of the covariance matrix as an indirect solution in BD-HP.

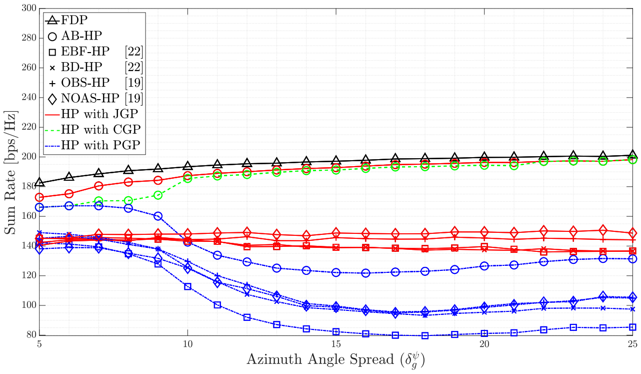


FIGURE 15. Sum-rate comparison of FDP and HP techniques versus azimuth angle spread (δ_g^{ψ}) for 20 × 20 URA, where SNR = 10 dB and $K_g = 3$ and $K = 18$.

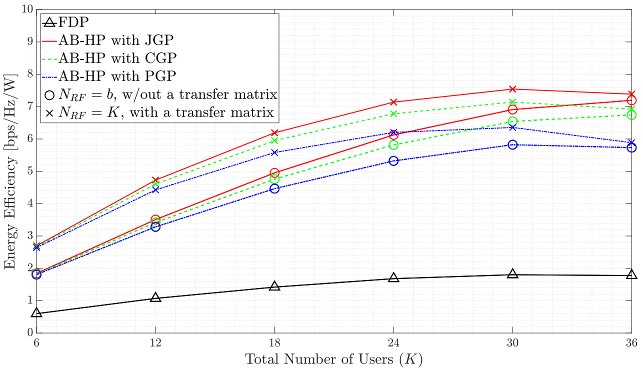


FIGURE 16. Energy efficiency versus total number of users (K) for 20 × 20 URA and SNR = 10 dB.

In Figure 15, the sum-rate curves are demonstrated versus the azimuth angle spread δ_g^{ψ} for SNR = 10 dB. As seen from the curves, the sum-rate of the proposed AB-HP with JGP and CGP increases for the wider azimuth angle spread similar to their single-stage counterpart FDP. However, the sum-rate of all benchmark techniques with JGP remains approximately constant all across the azimuth angle spread values. Additionally, we observe that the proposed CGP approach can provide an intermediate solution between JGP and PGP as expected. For instance, when $\delta_g^{\psi} = 5°$ and $\delta_g^{\psi} = 6°$, the number of baseband precoder block for CGP is found as $G' = 6$, which means that CGP is identical with PGP for the corresponding values of δ_g^{ψ}. Then, for $\delta_g^{\psi} = 7°$ and $\delta_g^{\psi} = 8°$, CGP with $G' = 4$ can provide slightly better sum-rate in comparison to PGP. Similarly, when the azimuth angle spread is increased to $\delta_g^{\psi} = 9°$, we have $G' = 3$ for CGP as illustrated in Figure 5. However, the sum-rate for PGP starts to decrease due the enhanced inter-group interference via the common AoD support along either x-axis or y-axis among the user groups. Afterwards, when $10° \leq \delta_g^{\psi} \leq 21°$, the performance of CGP with $G' = 2$ closely reaches to JGP. Ultimately, when $22° \leq \delta_g^{\psi} \leq 25°$, CGP with $G' = 1$ now becomes identical with JGP.

G. ENERGY EFFICIENCY

Finally, we represent the energy efficiency of the proposed AB-HP technique in comparison to FDP. According to the power consumption model in [54]–[56], the energy efficiency η is defined as:

$$\eta = \frac{R_{\text{sum}}^{\upsilon}}{P_{total}} = \frac{R_{\text{sum}}^{\upsilon}}{P_T + N_{RF}P_{RF} + N_{PS}P_{PS}} [\text{bps/Hz/W}] \qquad (51)$$

where $P_{total} = P_T + N_{RF}P_{RF} + N_{PS}P_{PS}$ represents the total power consumption, P_T is the total transmission power, P_{RF} and P_{PS} are the power consumption by each RF chain and phase-shifter, respectively, N_{RF} and N_{PS} are the number of RF chains and phase-shifters, respectively. As in [54], [55], we assume that $P_T = 1$ W, $P_{RF} = 250$ mW and $P_{PS} = 1$ mW.

$N_{RF} = b$ RF chains and $N_{PS} = b \times M$ phase-shifters are required for two-stage AB-HP without a transfer matrix as illustrated in Figure 2. Similarly, three-stage AB-HP with a transfer matrix needs $N_{RF} = K$ RF chains and $N_{PS} = b \times M + 2 \times b \times K$ phase-shifters as demonstrated in Figure 4. On the other hand, for the single-stage FDP technique, we need to place $N_{RF} = M$ RF chains in total without any phase-shifters.

Figure 16 plots the energy efficiency curves versus total number of users K for 20 × 20 URA and **SNR** = 10 dB. Considering $G = 6$ groups in Table 2, the number of users in each group varies between $K_g = 1$ and $K_g = 6$ in Figure 16. The proposed three-stage AB-HP technique with a transfer matrix, which reduces the number of RF chain to $N_{RF} = K$, provides higher energy efficiency compared to the proposed two-stage AB-HP technique with $N_{RF} = b = 56$ RF chains (please see Figure 5). As expected, the gap between their energy efficiency curves becomes smaller as K increases since it converges to b. Moreover, we observe that AB-HP remarkably improves the energy efficiency in comparison to the single-stage FDP technique.

H. COMPLEXITY ANALYSIS

The computational complexity of the proposed AB-HP techniques with and without a transfer matrix are compared to the single-stage FDP in Table 3. Firstly, the design of FDP requires $\mathcal{O}\left(M^3 + KM^2\right)$ operations, where $\mathcal{O}\left(KM^2\right)$ is related to the computation of $\mathbf{H}^H\mathbf{H}$ and $\mathcal{O}\left(M^3\right)$ is originated from the matrix inversion as in (21) [57]. Regarding the proposed AB-HP technique, the RF beamformer design in Algorithm 2 only needs $\mathcal{O}\left(GM\right)$ operations to find the corresponding quantized angle-pairs covering the AoD support of each user group, where there are M possible quantized angle-pairs and G groups. Afterwards, building the baseband precoder in the JGP approach requires a computational complexity of $\mathcal{O}\left(b^3 + Kb^2\right)$, where $\mathcal{O}\left(Kb^2\right)$ represents the computation of $\mathcal{H}^H\mathcal{H}$ and $\mathcal{O}\left(b^3\right)$ arises from the matrix inverse operation expressed in (21). In overall, the proposed AB-HP with JGP without a transfer matrix has a computational complexity of $\mathcal{O}\left(GM + b^3 + Kb^2\right)$. Furthermore, when we include a transfer block matrix as shown in Figure 4, according to (46), the inner transfer block matrices (i.e., $\mathbf{T}_A$ and $\mathbf{T}_B$) and the reduced-size baseband

TABLE 3. Computational complexity comparison for AB-HP and FDP.

FDP	$\mathcal{O}\left(M^3+KM^2\right)$
AB-HP with JGP	w/out a transfer matrix: $\mathcal{O}\left(GM+b^3+Kb^2\right)$
	with a transfer matrix: $\mathcal{O}\left(GM+b^3+Kb^2+K^2b+K^3\right)$
AB-HP with CGP	w/out a transfer matrix: $\mathcal{O}\big(GM+b_{gp}^3+K_{gp}b_{gp}^2+\sum_{t\neq g,p}^{G'} b_t^3+K_t b_t^2\big)$
	with a transfer matrix: $\mathcal{O}\big(GM+b_{gp}^3+K_{gp}b_{gp}^2+K_{gp}^2 b_{gp}+K_{gp}^3$ $+\sum_{t\neq g,p}^{G'} b_t^3+K_t b_t^2+K_t^2 b_t+K_t^3\big)$
AB-HP with PGP	w/out a transfer matrix: $\mathcal{O}\big(GM+\sum_{g=1}^{G} b_g^3+K_g b_g^2\big)$
	with a transfer matrix: $\mathcal{O}\big(GM+\sum_{g=1}^{G} b_g^3+K_g b_g^2+K_g^2 b_g+K_g^3\big)$

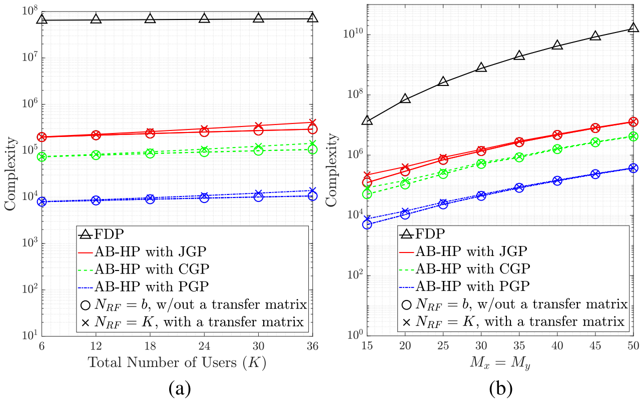


FIGURE 17. Computational complexity (a) versus total number of users (K) for 20 × 20 URA and (b) versus square array sizes for $K_g = 3$ and $K = 18$.

precoder (i.e., $\mathbf{B}_{rs}$) increase the computational complexity by $\mathcal{O}\left(K^2b+K^3\right)$. Hence, the computational complexity of the proposed AB-HP with JGP employing a transfer matrix is $\mathcal{O}\left(GM+b^3+Kb^2+K^2b+K^3\right)$. Following similar process, the complexity of AB-HP with CGP and PGP are obtained as in Table 3.

Figure 17 compares the complexity of FDP and AB-HP. As shown in Figure 17(a), the complexity of FDP can be significantly decreased by the proposed AB-HP technique. For example, when we employed a transfer matrix for AB-HP to serve $K = 36$ users, the complexity of JGP, CGP and PGP can be respectively reduced to 0.6%, 0.2% and 0.02% of the complexity of FDP. Furthermore, the efficiency of AB-HP improves as the antenna array size increases as demonstrated in Figure 17(b).

VII. CONCLUSIONS

A new user grouping algorithm and 3D angular-based hybrid precoding (AB-HP) technique have been proposed for massive MU-MIMO systems equipped with URA. The proposed user grouping algorithm can accurately find the target number of groups, then cluster users in the corresponding groups. Considering the practical implementations of the massive MU-MIMO systems with the reduced CSI overhead and hardware cost/complexity, the RF beamformer has been individually designed for each user group to eliminate the inter-group interference. The RF beamformer design only requires the slowly time-varying AoD information. Three different approaches as JGP, PGP and CGP have been examined for the baseband precoder design. By applying the deterministic equivalent principle onto the instantaneous SINR expressions, their tight deterministic approximations have been obtained for all three JGP, PGP and CGP approaches. The accuracy of the derived approximations is verified by Monte-Carlo simulation results. In addition, a transfer block design has been suggested to further reduce the number of RF chains exactly to the number of users without affecting the sum-rate performance. In comparison to FDP, the proposed AB-HP technique (i) requires low hardware and computational complexity, (ii) significantly enhances the energy efficiency performance and (iii) provides a comparable sum-rate performance, where the performance gap between them decreases as the size of URA increases. Furthermore, it has been shown that AB-HP is more robust to the channel estimation errors with respect to FDP. AB-HP provides a superior performance enhancement compared to its two-stage HP counterparts as EBF-HP [23], BD-HP [23], OBS-HP [20] and NOAS-HP [20].

APPENDIX A
DETERMINISTIC EQUIVALENT OF THE SINR FOR JGP

The instantaneous SINR for JGP obtained in (24) consists of two terms: (i) the desired signal power at the k^{th} user in group g as $\mathrm{S}_{gk}^J = \varepsilon^2\left|\mathbf{h}_{gk}^T\mathbf{F}\mathbf{W}\mathbf{F}^H\mathbf{h}_{gk}^*\right|^2$, (ii) the interference power at the corresponding user as $\mathrm{IN}_{gk}^J = \varepsilon^2\left\|\mathbf{h}_{gk}^T\mathbf{F}\mathbf{W}\mathbf{F}^H\mathbf{H}_{[gk]}^H\right\|_2^2$, hence, we have:

$$\mathrm{SINR}_{gk}^J = \frac{\mathrm{S}_{gk}^J}{\mathrm{IN}_{gk}^J+\sigma^2}. \tag{52}$$

A. DESIRED SIGNAL POWER

According to Lemma 2, the desired user power is written as:

$$\mathrm{S}_{gk}^J = \varepsilon^2\left|\mathbf{h}_{gk}^T\mathbf{F}\mathbf{W}\mathbf{F}^H\mathbf{h}_{gk}^*\right|^2 = \varepsilon^2\left|\frac{\mathbf{h}_{gk}^T\mathbf{F}\mathbf{W}_{[gk]}\mathbf{F}^H\mathbf{h}_{gk}^*}{1+\mathbf{h}_{gk}^T\mathbf{F}\mathbf{W}_{[gk]}\mathbf{F}^H\mathbf{h}_{gk}^*}\right|^2, \tag{53}$$

where $\mathbf{W}_{[gk]} = \left(\mathbf{F}^H\mathbf{H}_{[gk]}^H\mathbf{H}_{[gk]}\mathbf{F}+K\alpha\mathbf{I}_b\right)^{-1}$. The term in the numerator of (48) converges to:

$$\mathbf{h}_{gk}^T\mathbf{F}\mathbf{W}_{[gk]}\mathbf{F}^H\mathbf{h}_{gk}^* \xrightarrow{(a)} \frac{1}{P}\mathrm{tr}\left(\mathbf{F}^H\mathbf{\Phi}_g^H\mathbf{\Phi}_g\mathbf{F}\mathbf{W}_{[gk]}\right) \xrightarrow{(b)} m_g^J, \tag{54}$$

where m_g^J is given in (26), (a) can be obtained by using Lemma 3 and (b) is the direct consequence of [9, Theorem 1] for the large random matrices (i.e., $M \to \infty$). The deterministic equivalent of the normalization scalar ε given in (22) can be obtained by approximating $\mathrm{tr}\left(\boldsymbol{\mathcal{H}}\mathbf{W}^2\boldsymbol{\mathcal{H}}^H\right)$ as:

$$\mathrm{tr}\left(\boldsymbol{\mathcal{H}}\mathbf{W}^2\boldsymbol{\mathcal{H}}^H\right) \overset{(a)}{=} \sum_{p=1}^{G}\sum_{q=1}^{K_p}\mathbf{g}_{p_q}^T\mathbf{\Phi}_p\mathbf{F}\mathbf{W}^2\mathbf{F}^H\mathbf{\Phi}_p^H\mathbf{g}_{p_q}^*$$

$$\overset{(b)}{=} \sum_{p=1}^{G}\sum_{q=1}^{K_p} \frac{\mathbf{g}_{p_q}^T \mathbf{\Phi}_p \mathbf{F}\mathbf{W}_{[p_q]}^2 \mathbf{F}^H \mathbf{\Phi}_p^H \mathbf{g}_{p_q}^*}{\left(1+\mathbf{g}_{p_q}^T \mathbf{\Phi}_p \mathbf{F}\mathbf{W}_{[p_q]} \mathbf{F}^H \mathbf{\Phi}_p^H \mathbf{g}_{p_q}^*\right)^2}$$
$$\xrightarrow{(c)} \sum_{p=1}^{G}\sum_{q=1}^{K_p} \frac{\frac{1}{P}\mathrm{tr}\left(\mathbf{D}_p^J \mathbf{W}_{[p_q]}^2\right)}{\left(1+\frac{1}{P}\mathrm{tr}\left(\mathbf{D}_p^J \mathbf{W}_{[p_q]}\right)\right)^2}$$
$$\xrightarrow{(d)} \sum_{p=1}^{G} \frac{K_p c_p^J}{\left(1+m_p^J\right)^2}, \qquad (55)$$

where (a) follows from Lemma 3, (b) is obtained via utilizing Lemma 2 twice, (c) is derived via Lemma 3 and (d) is found via applying both (54) and [9, Theorem 2] with c_p^J expressed in (27). Then, by substituting (22), (54) and (55) into (53), the deterministic equivalent of the desired user power for JGP approach can be obtained as:

$$\mathrm{S}_{g_k}^J \xrightarrow[M\to\infty]{\text{(a.s.)}} \frac{P_T}{\sum\limits_{p=1}^{G} \frac{K_p c_p}{\left(1+m_p^J\right)^2}} \left(\frac{m_g^J}{1+m_g^J}\right)^2. \qquad (56)$$

B. INTERFERENCE POWER

The interference power for JGP approach is given by:

$$\mathrm{IN}_{g_k}^J = \varepsilon^2 \left\|\mathbf{h}_{g_k}^T \mathbf{F}\mathbf{W}\mathbf{F}^H \mathbf{H}_{[g_k]}^H \cdot\right\|_2^2. \qquad (57)$$

By applying Lemma 2 twice, Lemma 3 and (54), the following expression can be approximated as:

$$\left\|\mathbf{h}_{g_k}^T \mathbf{F}\mathbf{W}\mathbf{F}^H \mathbf{H}_{[g_k]}^H\right\|_2^2 \rightarrow \frac{\frac{1}{P}\mathrm{tr}\left(\mathbf{H}_{[g_k]}\mathbf{F}\mathbf{W}_{[g_k]}\mathbf{D}_g^J \mathbf{W}_{[g_k]}\mathbf{F}^H \mathbf{H}_{[g_k]}^H\right)}{\left(1+m_g^J\right)^2}. \qquad (58)$$

To simplify the above expression, the following equality can be written via Lemma 4 as follows:

$$\mathrm{tr}\left(\mathbf{H}_{[g_k]}\mathbf{F}\mathbf{W}_{[g_k]}\mathbf{D}_g^J \mathbf{W}_{[g_k]}\mathbf{F}^H \mathbf{H}_{[g_k]}^H\right) = \sum_{p=1}^{G} \sum_{\substack{q=1 \\ p_q\neq g_k}}^{K_p} \mathbf{h}_{p_q}^T \mathbf{F}\mathbf{W}_{[g_k]}\mathbf{D}_g^J \mathbf{W}_{[g_k]}\mathbf{F}^H \mathbf{h}_{p_q}^*. \qquad (59)$$

By using (59), (58) can be approximated as:

$$\left\|\mathbf{h}_{g_k}^T \mathbf{F}\mathbf{W}\mathbf{F}^H \mathbf{H}_{[g_k]}^H\right\|_2^2$$
$$\rightarrow \sum_{p=1}^{G} \sum_{\substack{q=1 \\ p_q\neq g_k}}^{K_p} \frac{\mathbf{h}_{p_q}^T \mathbf{F}\mathbf{W}_{[g_k]}\mathbf{D}_g^J \mathbf{W}_{[g_k]}\mathbf{F}^H \mathbf{h}_{p_q}^*}{P\left(1+m_g^J\right)^2}$$
$$\overset{(a)}{=} \sum_{p=1}^{G} \sum_{\substack{q=1 \\ p_q\neq g_k}}^{K_p} \frac{\frac{1}{P}\mathbf{h}_{p_q}^T \mathbf{F}\mathbf{W}_{[g_k,p_q]}\mathbf{D}_g^J \mathbf{W}_{[g_k,p_q]}\mathbf{F}^H \mathbf{h}_{p_q}^*}{\left(1+\mathbf{h}_{p_q}^T \mathbf{F}\mathbf{W}_{[g_k,p_q]}\mathbf{F}^H \mathbf{h}_{p_q}^*\right)^2\left(1+m_g^J\right)^2}$$
$$\xrightarrow{(b)} \sum_{p=1}^{G} \sum_{\substack{q=1 \\ p_q\neq g_k}}^{K_p} \frac{\frac{1}{P^2}\mathrm{tr}\left(\mathbf{D}_p^J \mathbf{W}_{[g_k,p_q]}\mathbf{D}_g^J \mathbf{W}_{[g_k,p_q]}\right)}{\left(1+\frac{1}{P}\mathrm{tr}\left(\mathbf{D}_p^J \mathbf{W}_{[g_k,p_q]}\right)\right)^2\left(1+m_g^J\right)^2}$$
$$\xrightarrow{(c)} \frac{\left(K_g-1\right)c_{g,g}^J}{\left(1+m_g^J\right)^4} + \sum_{p\neq g}^{G} \frac{K_p c_{g,p}^J}{\left(1+m_p^J\right)^2\left(1+m_g^J\right)^2}, \qquad (60)$$

where (a) follows from Lemma 2, (b) is obtained via Lemma 3 and (c) is approximated via [9, Theorem 2] with $c_{g,p}^J$ expressed in (27). Therefore, by combining (22), (55), (57) and (60), the deterministic equivalent of the interference power is derived as:

$$\mathrm{IN}_{g_k}^J \xrightarrow[M\to\infty]{\text{(a.s.)}} \frac{P_T\left(\frac{(K_g-1)c_{g,g}^J}{\left(1+m_g^J\right)^4} + \sum\limits_{p\neq g}^{G} \frac{K_p c_{g,p}^J}{\left(1+m_p^J\right)^2\left(1+m_g^J\right)^2}\right)}{\sum\limits_{p=1}^{G} \frac{K_p c_p}{\left(1+m_p^J\right)^2}}. \qquad (61)$$

Finally, by injecting (56) and (61) into (52), the deterministic equivalent of the SINR expression for JGP approach can be obtained as given in (25).

APPENDIX B
DETERMINISTIC EQUIVALENT OF THE SINR FOR PGP

The instantaneous SINR for PGP given in (32) consists of three terms: (i) the desired signal power at the k^{th} user in group g as $\mathrm{S}_{g_k}^P = \varepsilon_g^2\left|\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_k}^*\right|^2$, (ii) the intra-group interference power as $\mathrm{IaGI}_{g_k}^P = \varepsilon_g^2\left\|\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \boldsymbol{\mathcal{H}}_{g,[k]}^H\right\|_2^2$, (iii) the inter-group interference power as $\mathrm{IGI}_{g_k}^P = \sum_{p\neq g}^{G} \varepsilon_p^2\left\|\mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \boldsymbol{\mathcal{H}}_p^H\right\|_2^2$. Then, $\mathrm{SINR}_{g_k}^P$ given in (32) can be rewritten as:

$$\mathrm{SINR}_{g_k}^P = \frac{\mathrm{S}_{g_k}^P}{\mathrm{IaGI}_{g_k}^P + \mathrm{IGI}_{g_k}^P + \sigma^2}. \qquad (62)$$

A. DESIRED SIGNAL POWER

By using Lemma 2, the desired user power can be given by:

$$\mathrm{S}_{g_k}^P = \varepsilon_g^2\left|\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_k}^*\right|^2$$
$$= \varepsilon_g^2\left|\frac{\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_{g,[k]} \mathbf{F}_g^H \mathbf{h}_{g_k}^*}{1+\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_{g,[k]} \mathbf{F}_g^H \mathbf{h}_{g_k}^*}\right|^2, \qquad (63)$$

where $\mathbf{W}_{g,[k]} = \left(\mathbf{F}_g^H \mathbf{H}_{g,[k]}^H \mathbf{H}_{g,[k]} \mathbf{F}_g + K_g \alpha \mathbf{I}_{b_g}\right)^{-1}$. To derive the deterministic equivalent of (63), by using Lemma 3 and [9, Theorem 1], the following term can be approximated as:

$$\mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_{g,[k]} \mathbf{F}_g^H \mathbf{h}_{g_k}^* \rightarrow \frac{\mathrm{tr}\left(\mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}_g \mathbf{W}_{g,[k]}\right)}{P} \rightarrow m_g^P. \qquad (64)$$

where m_g^P is expressed in (34). Then, the deterministic equivalent of the normalization scalar for group g given in (30) can be obtained by approximating $\mathrm{tr}\left(\boldsymbol{\mathcal{H}}_g \mathbf{W}_g^2 \boldsymbol{\mathcal{H}}_g^H\right)$ as:

$$\mathrm{tr}\left(\boldsymbol{\mathcal{H}}_g \mathbf{W}_g^2 \boldsymbol{\mathcal{H}}_g^H\right) \overset{(a)}{=} \sum_{q=1}^{K_g} \mathbf{g}_{g_q}^T \mathbf{\Phi}_g \mathbf{F}_g \mathbf{W}_g^2 \mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{g}_{g_q}^*$$

$$\xrightarrow{(b)} \sum_{q=1}^{K_g} \frac{\frac{1}{P}\mathrm{tr}\left(\mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}_g \mathbf{W}_{g,[q]}^2\right)}{\left(1+\frac{1}{P}\mathrm{tr}\left(\mathbf{F}_g^H \mathbf{\Phi}_g^H \mathbf{\Phi}_g \mathbf{F}_g \mathbf{W}_{g,[q]}\right)\right)^2}$$
$$\xrightarrow{(c)} \frac{K_g c_g^P}{\left(1+m_g^P\right)^2}, \qquad (65)$$

where (a) is the direct consequence of Lemma 4, (b) is obtained via Lemma 2 and Lemma 3, (c) is derived by using (64) and [9, Theorem 2] with c_g^P given in (35). By substituting (30), (64) and (65) into (63), the deterministic equivalent of the desired signal power can be given by:

$$\mathrm{S}_{g_k}^P \xrightarrow[M\to\infty]{\text{(a.s.)}} \frac{P_T m_g^2}{K c_g}. \qquad (66)$$

B. INTRA-GROUP INTERFERENCE POWER

The intra-group interference power at k^{th} user in the group g is given by:

$$\mathrm{IaGI}_{g_k}^P = \varepsilon_g^2 \left\| \mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \boldsymbol{\mathcal{H}}_{g,[k]}^H \right\|_2^2$$
$$= \varepsilon_g^2 \sum_{q\neq k}^{K_g} \left| \mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_q}^* \right|^2. \qquad (67)$$

For the deterministic equivalent of (67), the following term can be approximated as:

$$\left| \mathbf{h}_{g_k}^T \mathbf{F}_g \mathbf{W}_g \mathbf{F}_g^H \mathbf{h}_{g_q}^* \right|^2 \xrightarrow{(a)} \frac{\mathrm{tr}\left(\mathbf{h}_{g_q}^T \mathbf{F}_g \mathbf{W}_{g,[k,q]} \mathbf{D}_g^P \mathbf{W}_{g,[k,q]} \mathbf{F}_g^H \mathbf{h}_{g_q}^*\right)}{P\left(1+m_g^P\right)^4}$$
$$\xrightarrow{(b)} \frac{\frac{1}{P^2}\mathrm{tr}\left(\mathbf{D}_g^P \mathbf{W}_{g,[k,q]} \mathbf{D}_g^P \mathbf{W}_{g,[k,q]}\right)}{\left(1+m_g^P\right)^4}$$
$$\xrightarrow{(c)} \frac{c_{g,g}^P}{\left(1+m_g^P\right)^4}, \qquad (68)$$

where $\mathbf{W}_{g,[k,q]} = \left(\mathbf{F}_g^H \mathbf{H}_{g,[k,q]}^H \mathbf{H}_{g,[k,q]} \mathbf{F}_g + K_g \alpha \mathbf{I}_{b_g}\right)^{-1}$, (a) is obtained by using Lemma 2 twice, Lemma 3 and (64), respectively, (b) is derived by applying Lemma 3, (c) is found by using [9, Theorem 2] with $c_{g,g}^P$ given in (35). By combining (30), (64), (67) and (68), the deterministic equivalent of the intra-group interference power is found as:

$$\mathrm{IaGI}_{g_k}^P \xrightarrow[M\to\infty]{\text{(a.s.)}} \frac{\left(K_g-1\right)}{K} \frac{P_T c_{g,g}}{c_g\left(1+m_g\right)^2}. \qquad (69)$$

C. INTER-GROUP INTERFERENCE POWER

The inter-group interference power for PGP approach is given by:

$$\mathrm{IGI}_{g_k}^P = \sum_{p\neq g}^{G} \varepsilon_p^2 \left\| \mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \boldsymbol{\mathcal{H}}_p^H \right\|_2^2$$
$$= \sum_{p\neq g}^{G} \sum_{q=1}^{K_p} \varepsilon_p^2 \left| \mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathbf{F}_p^H \mathbf{h}_{p_q}^* \right|^2. \qquad (70)$$

Similar to (68), the following term can be approximated as:

$$\left| \mathbf{h}_{g_k}^T \mathbf{F}_p \mathbf{W}_p \mathbf{F}_p^H \mathbf{h}_{p_q}^* \right|^2 \to \frac{c_{p,g}}{\left(1+m_p\right)^2}, \qquad (71)$$

with $c_{p,g}^P$ given in (35). By substituting (30), (64) and (71) into (70), the deterministic equivalent of the inter-group interference power is obtained as:

$$\mathrm{IGI}_{g_k}^P \xrightarrow[M\to\infty]{\text{(a.s.)}} \sum_{p\neq g}^{G} \frac{P_T K_p c_{p,g}}{K c_p}. \qquad (72)$$

By applying (66), (69) and (72) into (62), the deterministic equivalent of the SINR expression for PGP approach can be obtained as given in (33).

APPENDIX C
DETERMINISTIC EQUIVALENT OF THE SINR FOR CGP

The instantaneous SINR for CGP given in (38) consists of (i) the desired signal power at the k^{th} user in group g as $\mathrm{S}_{g_k}^C = \varepsilon_{gp}^2 \left| \mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^* \right|^2$, (ii) the common-group interference power as $\mathrm{CGI}_{g_k}^C = \varepsilon_{gp}^2 \left\| \mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \boldsymbol{\mathcal{H}}_{gp,[g_k]}^H \right\|_2^2$, (iii) the inter-group interference power as $\mathrm{IGI}_{g_k}^C = \sum_{t\neq g,p}^{G'} \varepsilon_t^2 \left\| \mathbf{h}_{g_k}^T \mathbf{F}_t \mathbf{W}_t \boldsymbol{\mathcal{H}}_t^H \right\|_2^2$. Hence, we have:

$$\mathrm{SINR}_{g_k}^C = \frac{\mathrm{S}_{g_k}^C}{\mathrm{CGI}_{g_k}^C + \mathrm{IGI}_{g_k}^C + \sigma^2}. \qquad (73)$$

A. DESIRED SIGNAL POWER

By applying Lemma 2, The desired signal power can be rewritten as:

$$\mathrm{S}_{g_k}^C = \varepsilon_{gp}^2 \left| \mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^* \right|^2$$
$$= \varepsilon_{gp}^2 \left| \frac{\mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp,[g_k]} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^*}{1+\mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp,[g_k]} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^*} \right|^2, \qquad (74)$$

where $\mathbf{W}_{gp,[g_k]} = \left(\mathbf{F}_{gp}^H \mathbf{H}_{gp,[g_k]}^H \mathbf{H}_{gp,[g_k]} \mathbf{F}_{gp} + K_{gp} \alpha \mathbf{I}_{b_{gp}}\right)^{-1}$. Similar to (64), the following term can be approximated as:

$$\mathbf{h}_{g_k}^T \mathbf{F}_{gp} \mathbf{W}_{gp,[g_k]} \mathbf{F}_{gp}^H \mathbf{h}_{g_k}^* \to m_g^C, \qquad (75)$$

where m_g^C is given in (40). Then, the normalization scalar belonging to the common baseband precoder for group g and p given in (37) can be easily approximated by considering the case of JGP processing. Hence, by using (55), we can derive the following approximation as:

$$\mathrm{tr}\left(\boldsymbol{\mathcal{H}}_{gp} \mathbf{W}_{gp}^2 \boldsymbol{\mathcal{H}}_{gp}^H\right) \to \sum_{t=g,p} \frac{K_t c_t^C}{\left(1+m_t^C\right)^2}, \qquad (76)$$

with c_g^C and c_p^C defined in (41). By substituting (37), (75) and (76) into (74), the deterministic equivalent of the desired signal power can be given by:

$$\mathrm{S}_{g_k}^C \xrightarrow[M\to\infty]{\text{(a.s.)}} \frac{P_T K_{gp}}{K \sum_{t=g,p} \frac{K_t c_t^C}{\left(1+m_t^C\right)^2}} \left(\frac{m_g^C}{1+m_g^C}\right)^2. \qquad (77)$$

B. COMMON-GROUP INTERFERENCE POWER

By using (37), (57), (60) and (76), the deterministic equivalent of the common-group interference power is given by:

$$\begin{aligned}\mathrm{CGI}_{gk}^{C} &= \varepsilon_{gp}^{2}\left\|\mathbf{h}_{gk}^{T}\mathbf{F}_{gp}\mathbf{W}_{gp}\boldsymbol{\mathcal{H}}_{gp,[gk]}^{H}\right\|_{2}^{2}\\ &\xrightarrow[M\to\infty]{\text{(a.s.)}} \frac{\left(\frac{(K_g-1)c_{g,g}^{C}}{\left(1+m_g^J\right)^4}+\frac{K_p c_{g,p}^{C}}{\left(1+m_p^J\right)^2\left(1+m_g^J\right)^2}\right)}{\frac{K}{P_T K_{gp}}\sum\limits_{t=g,p}\frac{K_t c_t^C}{\left(1+m_t^C\right)^2}}, \qquad (78)\end{aligned}$$

with $c_{g,g}^{C}$ and $c_{g,p}^{C}$ defined in (41).

C. INTER-GROUP INTERFERENCE POWER

As in the PGP approach, by using both (70) and (72), the deterministic equivalent of the inter-group interference power for CGP can be found as:

$$\begin{aligned}\mathrm{IGI}_{gk}^{C} &= \sum_{t\neq g,p}^{G'}\varepsilon_t^2\left\|\mathbf{h}_{gk}^{T}\mathbf{F}_t\mathbf{W}_t\boldsymbol{\mathcal{H}}_t^H\right\|_2^2\\ &\xrightarrow[M\to\infty]{\text{(a.s.)}}\sum_{t\neq g,p}^{G'}\frac{P_T K_t c_{t,g}^P}{K c_t^P}, \qquad (79)\end{aligned}$$

with c_t^P and $c_{t,g}^P$ defined in (35), which are obtained for the PGP approach. Then, by substituting (77), (78) and (79) into (73), the deterministic equivalent of the SINR expression for the CGP approach can be found as given in (39).

APPENDIX D
IMPORTANT LEMMAS

Lemma 1 [17]: Let $\mathbf{a} = [a_1, \cdots, a_N]^T \in \mathbb{C}^N$ be a complex vector with $a_n = |a_n|\, e^{j\angle a_n} \in \mathbb{C}$. Then, we have:

$$a_n = |a_n|\, e^{j\angle a_n} = \mu\left(e^{j\vartheta_{n,1}} + e^{j\vartheta_{n,2}}\right) \qquad (80)$$

where $\mu = \frac{1}{2}\max\limits_n |a_n|$, $\vartheta_{n,1} = \cos^{-1}\left(\frac{|a_n|}{2\mu}\right) + \angle a_n$ and $\vartheta_{n,2} = -\cos^{-1}\left(\frac{|a_n|}{2\mu}\right) + \angle a_n$.

Proof: By applying the trigonometric identities for the summation of sine and cosine functions, we can write:

$$\begin{aligned}&\mu\left(e^{j\vartheta_{n,1}} + e^{j\vartheta_{n,2}}\right)\\ &= \mu\left[\cos\left(\vartheta_{n,1}\right) + j\sin\left(\vartheta_{n,1}\right) + \cos\left(\vartheta_{n,2}\right) + j\sin\left(\vartheta_{n,2}\right)\right]\\ &= 2\mu\cos\left(\frac{\vartheta_{n,1}-\vartheta_{n,2}}{2}\right)\left[\cos\left(\frac{\vartheta_{n,1}+\vartheta_{n,2}}{2}\right)\right.\\ &\qquad\left. + j\sin\left(\frac{\vartheta_{n,1}+\vartheta_{n,2}}{2}\right)\right]\\ &= 2\mu\cos\left(\frac{\vartheta_{n,1}-\vartheta_{n,2}}{2}\right)e^{j\left(\frac{\vartheta_{n,1}+\vartheta_{n,2}}{2}\right)}. \qquad (81)\end{aligned}$$

By using $\mu = \frac{1}{2}\max\limits_n |a_n|$, we have the following inequality $0 \leq \frac{|a_n|}{2\mu} = \frac{|a_n|}{\max\limits_n |a_n|} \leq 1$. Therefore, by substituting $\frac{\vartheta_{n,1}+\vartheta_{n,2}}{2} = \angle a_n$, $\frac{\vartheta_{n,1}-\vartheta_{n,2}}{2} = \frac{|a_n|}{2\mu}$ and $\cos\left(\cos^{-1}\left(\frac{|a_n|}{2\mu}\right)\right) = \frac{|a_n|}{2\mu}$ into (81), the proof of (80) is completed as follows:

$$\mu\left(e^{j\vartheta_{n,1}} + e^{j\vartheta_{n,2}}\right) = 2\mu\frac{|a_n|}{2\mu}e^{j\angle a_n} = |a_n|\, e^{j\angle a_n} = a_n. \qquad (82)$$

Lemma 2 (Matrix Inversion Lemma) [37]: Let $\mathbf{X} \in \mathbb{C}^{N\times N}$ and $\mathbf{X} + \zeta\mathbf{a}^H\mathbf{a} \in \mathbb{C}^{N\times N}$ be invertible matrices with $\mathbf{a} \in \mathbb{C}^N$ and $\zeta \in \mathbb{C}$. Then, we have;

$$\mathbf{a}^H\left(\mathbf{X} + \zeta\mathbf{a}^H\mathbf{a}\right)^{-1}\mathbf{a} = \frac{\mathbf{a}^H\mathbf{X}^{-1}\mathbf{a}}{1 + \mathbf{a}^H\mathbf{X}^{-1}\mathbf{a}}. \qquad (83)$$

Lemma 3 [37]: Let $\mathbf{X} \in \mathbb{C}^{N\times N}$ be a random matrix generated by the probability space $(\Omega, \mathcal{F}, P)$ such that $w \in X \subset \Omega$ with $P(X) = 1$ and $\|\mathbf{X}(w)\| < \infty$. Let $\mathbf{a} \in \mathbb{C}^N$ be a random vector with i.i.d. entries having have zero mean, variance $\frac{1}{N}$, eight order moment of $O\left(\frac{1}{N^4}\right)$ and independent of $\mathbf{X}$. Then, $\mathbf{a}^H\mathbf{X}\mathbf{a}$ almost surely converges to $\mathrm{tr}(\mathbf{X})$ as $N \to \infty$, thus, we have:

$$\mathbf{a}^H\mathbf{X}\mathbf{a} - \frac{1}{N}\mathrm{tr}(\mathbf{X}) \xrightarrow[N\to\infty]{\text{(a.s.)}} 0. \qquad (84)$$

Lemma 4 (Trace of a Matrix Product) [45]: Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{C}^{N\times A}$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N] \in \mathbb{C}^{N\times A}$ be two complex matrices having the same size. Then, we have:

$$\mathbf{X}^H\mathbf{Y} = \begin{bmatrix}\mathbf{x}_1^H\mathbf{y}_1 & \cdots & \mathbf{x}_1^H\mathbf{y}_N\\ \vdots & \ddots & \vdots\\ \mathbf{x}_N^H\mathbf{y}_1 & \cdots & \mathbf{x}_N^H\mathbf{y}_N\end{bmatrix}. \qquad (85)$$

Hence, the trace of the above matrix product can be expanded as follows:

$$\mathrm{tr}\left(\mathbf{X}^H\mathbf{Y}\right) = \sum_{n=1}^{N}\mathbf{x}_n^H\mathbf{y}_n. \qquad (86)$$

ACKNOWLEDGMENT

This article was presented in part at the 5G and Beyond Technologies for Ultra-Dense Environments Workshop in 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, September 2019 [1].

REFERENCES

[1] H. V. Vu and T. Le-Ngoc, "Performance analysis of underlaid full-duplex D2D cellular networks," *IEEE Access*, vol. 7, pp. 176233–176247, 2019, doi: 10.1109/ACCESS.2019.2958300.

[2] A. Koc, A. Masmoudi, and T. Le-Ngoc, "Angular-based 3D hybrid precoding for URA in multi-user massive MIMO systems," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–5.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[6] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[7] N. Fatema, G. Hua, Y. Xiang, D. Peng, and I. Natgunanathan, "Massive MIMO linear precoding: A survey," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3920–3931, Dec. 2018.

[8] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[9] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, Jul. 2012.

[10] C. Zhang, Y. Jing, Y. Huang, and L. Yang, "Performance analysis for massive MIMO downlink with low complexity approximate zero-forcing precoding," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3848–3864, Sep. 2018.

[11] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.

[12] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.

[13] P. Sudarshan, N. Mehta, A. Molisch, and J. Zhang, "Channel statistics-based RF pre-processing with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3501–3511, Dec. 2006.

[14] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.

[15] O. E. 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.

[16] 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.

[17] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog–digital beamforming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3311–3326, May 2016.

[18] S. Payami, M. Ghoraishi, M. Dianati, and M. Sellathurai, "Hybrid beamforming with a reduced number of phase shifters for massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 4843–4851, Jun. 2018.

[19] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.

[20] H. Lin, F. Gao, S. Jin, and G. Y. Li, "A new view of multi-user hybrid massive MIMO: Non-orthogonal angle division multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2268–2280, Oct. 2017.

[21] M. Maleki and K. Mohamed-Pour, "Hybrid preprocessing aided spatial modulation in multi-user massive MIMO systems," *IET Commun.*, vol. 12, no. 15, pp. 1812–1821, Sep. 2018.

[22] P. Zhao and Z. Wang, "Joint user scheduling and hybrid precoding for multi-user mmWave systems with two-layer PS network," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[23] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and Multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[24] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876–890, Oct. 2014.

[25] J. Nam, Y.-J. Ko, and J. Ha, "User grouping of two-stage MU-MIMO precoding for clustered user geometry," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1458–1461, Aug. 2015.

[26] Y. Jeon, C. Song, S.-R. Lee, S. Maeng, J. Jung, and I. Lee, "New beamforming designs for joint spatial division and multiplexing in large-scale MISO multi-user systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3029–3041, May 2017.

[27] T. Ketseoglou and E. Ayanoglu, "Downlink precoding for massive MIMO systems exploiting virtual channel model sparsity," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 1925–1939, May 2018.

[28] A. Koc, A. Masmoudi, and T. Le-Ngoc, "Hybrid beamforming for uniform circular arrays in multi-user massive MIMO systems," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2019, pp. 1–4.

[29] R. Mai and T. Le-Ngoc, "Nonlinear hybrid precoding for coordinated multi-cell massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2459–2471, Mar. 2019.

[30] X. Cheng, B. Yu, L. Yang, J. Zhang, G. Liu, Y. Wu, and L. Wan, "Communicating in the real world: 3D MIMO," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 136–144, Aug. 2014.

[31] C. Balanis, *Antenna Theory: Analysis and Design*. Hoboken, NJ, USA: Wiley, 2015.

[32] J. Song, J. Choi, and D. J. Love, "Common codebook millimeter wave beam design: Designing beams for both sounding and communication with uniform planar arrays," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1859–1872, Apr. 2017.

[33] M. D. Zoltowski, M. Haardt, and C. P. Mathews, "Closed-form 2-D angle estimation with rectangular arrays in element space or beamspace via unitary ESPRIT," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 316–328, Feb. 1996.

[34] Y.-H. Nam, B. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, Jun. 2013.

[35] T. Wang, B. Ai, R. He, and Z. Zhong, "Two-dimension direction-of-arrival estimation for massive MIMO systems," *IEEE Access*, vol. 3, pp. 2122–2128, 2015.

[36] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access*, vol. 2, pp. 947–959, 2014.

[37] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[38] Z. Zheng, W.-Q. Wang, H. Meng, H. C. So, and H. Zhang, "Efficient beamspace-based algorithm for two-dimensional DOA estimation of incoherently distributed sources in massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11776–11789, Dec. 2018.

[39] H. Kim and M. Viberg, "Two decades of array signal processing research," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[40] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[41] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[42] G. Lee, Y. Sung, and M. Kountouris, "On the performance of random beamforming in sparse millimeter wave channels," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 560–575, Apr. 2016.

[43] E. Bjornson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[44] G. Lee and Y. Sung, "A new approach to user scheduling in massive multi-user MIMO broadcast channels," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1481–1495, Apr. 2018.

[45] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, vol. 71. Philadelphia, PA, USA: SIAM, 2000.

[46] *5G; Study on Channel Model for Frequencies From 0.5 to 100 GHz, V14.0.0 (2017-05)*, document 3GPP TR 38.901, Jul. 2018.

[47] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser Communication—Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[48] B. Nosrat-Makouei, J. G. Andrews, and R. W. Heath, Jr., "MIMO interference alignment over correlated channels with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2783–2794, Jun. 2011.

[49] D. Mi, M. Dianati, L. Zhang, S. Muhaidat, and R. Tafazolli, "Massive MIMO performance with imperfect channel reciprocity and channel estimation error," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3734–3749, Sep. 2017.

[50] L. Chu, F. Wen, and R. C. Qiu, "Eigen-inference precoding for coarsely quantized massive MU-MIMO system with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8729–8743, Sep. 2019.

[51] A. Koc, I. Altunbas, and E. Basar, "Full-duplex spatial modulation systems under imperfect channel state information," in *Proc. 24th Int. Conf. Telecommun. (ICT)*, Limassol, Cyprus, May 2017, pp. 1–5.

[52] E. Basar, U. Aygolu, E. Panayirci, and H. V. Poor, "Performance of spatial modulation in the presence of channel estimation errors," *IEEE Commun. Lett.*, vol. 16, no. 2, pp. 176–179, Feb. 2012.

[53] W. M. Gifford, M. Z. Win, and M. Chiani, "Diversity with practical channel estimation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1935–1947, Jul. 2005.

[54] 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.

[55] Y. Wang, W. Zou, and Y. Tao, "Analog precoding designs for millimeter wave communication systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11733–11745, Dec. 2018.

[56] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, Sep. 2005.

[57] J.-C. Chen, "A low complexity data detection algorithm for uplink multiuser massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1701–1714, Aug. 2017.

ASIL KOC (Student Member, IEEE) received the B.Sc. degree (Hons.) in electronics and communication engineering and the M.Sc. degree (Hons.) in telecommunication engineering from Istanbul Technical University, Istanbul, Turkey, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with McGill University, Montreal, QC, Canada.

From 2015 to 2017, he was a Research and a Teaching Assistant with the Electronics and Communication Engineering Department, Istanbul Technical University. Since 2017, he has been a Teaching Assistant with the Electrical and Computer Engineering Department, McGill University. His research interests include, but not limited to wireless communications, massive MIMO, full-duplex, spatial modulation, energy harvesting, and cooperative networks. He was a recipient of the Erasmus Scholarship Award funded by the European Union, the McGill Engineering Doctoral Award, the STARaCom Collaborative Grant funded by the FRQNT, and the Graduate Research Enhancement and Travel Award funded by McGill University.

AHMED MASMOUDI (Member, IEEE) was born in Ariana, Tunisia. He received the Diplôme d'Ingénieur in communications from the Ecole Supérieure des Communications de Tunis-Sup'Com, Ariana, in 2010, the M.Sc. degree from the Institut National de la Recherche Scientifique (INRS), Montreal, QC, Canada, in 2012, and the Ph.D. degree from McGill University, Montreal. He is currently a Postdoctoral Fellow with McGill University. His research interests include statistical signal processing and wireless communications.

THO LE-NGOC (Life Fellow, IEEE) received the B.Eng. degree (Hons.) in electrical engineering, in 1976, the M.Eng. degree in microprocessor applications from McGill University, Montreal, in 1978, and the Ph.D. degree in digital communications from the University of Ottawa, Canada, in 1983.

From 1977 to 1982, he was with Spar Aerospace Ltd., Sainte-Anne-de-Bellevue, QC, Canada, involved in the development and design of satellite communications systems. From 1982 to 1985, he was with SRTelecom Inc., Saint-Laurent, QC, Canada, where he developed the new point-to-multipoint DA-TDMA/TDM Subscriber Radio System SR500. From 1985 to 2000, he was a Professor with the Department of Electrical and Computer Engineering, Concordia University, Montreal. Since 2000, he has been with the Department of Electrical and Computer Engineering, McGill University. His research interest includes broadband digital communications. He is a Fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada. He was a recipient of the 2004 Canadian Award in Telecommunications Research and the IEEE Canada Fessenden Award, in 2005. He holds the Canada Research Chair (Tier I) on Broadband Access Communications.