Clustered Hybrid Precoding Design for Multi-User Massive MIMO Systems

Ling Zhang, Lin Gui, Member, IEEE, Kai Ying, and Qibo Qin

Abstract—Hybrid precoding has been recognized as a promising technology to combat the path loss of millimeter-wave signals in massive multiple-input-multiple-output (MIMO) systems. However, due to the joint optimization of the digital and analog precoding matrices as well as extra constraints for the analog part, the hybrid precoding design is still a tough issue in current research. In this paper, we adopt the thought of clustering in unsupervised learning and provide design schemes for fully-connected hybrid precoding (FHP) and adaptively-connected hybrid precoding (AHP) in multi-user massive MIMO systems. For FHP, we propose the hierarchical-agglomerative-clustering-based (HAC-based) scheme to explore the relevance among radio frequency (RF) chains in the optimal hybrid precoding design. The similar RF chains are merged into an individual when insufficient RF chains are available. For AHP, we propose the modified-K-means-based (MKM-based) scheme to explore the relevance among antennas. The similar antennas are supported by the same RF chain to make full use of the flexible connection in AHP. Particularly, in the proposed MKM-based AHP design, the clustering centers are updated by alternating-optimum-based (AO-based) scheme with a specific initialization method, which is capable to independently provide feasible sub-connected hybrid precoding (SHP) design with quick convergence. Simulation results highlight the superior spectrum efficiency of the proposed HAC-based FHP scheme, and the high power efficiency of the proposed MKM-based AHP scheme. Moreover, all the proposed schemes are clarified to effectively handle the inter-user interference and outperform the existing work.

I. INTRODUCTION

With the popularity of intelligent terminals, mobile data traffic is facing exponential growth. To meet the potential capacity requirements for future wireless communications, various novel wireless techniques such as massive multiple-input-multiple-output (MIMO), advanced channel coding, and non-orthogonal multiple access have enthused much attention [1]–[3]. Nevertheless, the bandwidth shortage in the physical layer leads to the fundamental bottleneck for capacity improvement [4]. Thus, it is imminent to develop spectrum bands that have not been utilized in current cellular systems.

Millimeter-wave (mmWave) band spanning from 30 to 300 GHz has been determined as the alternative band to expand the available bandwidth in 5G systems [5]. Benefiting from the short wavelength of mmWave, it is feasible to deploy large-scale antennas in limited space at transceivers to implement massive MIMO systems. However, due to the extremely high carrier frequency, mmWave signals experience more serious propagation path loss compared with signals in 3G or LTE. It is necessary to use precoding technology to achieve highly directional beamforming [6], [7]. The full digital precoding is a mature technology for traditional MIMO systems. To harvest the channel gain, the singular value decomposition (SVD) method is optimal for point-to-point systems [8]. And matched-filter (MF), zero-forcing (ZF), and regularized zero-forcing (RZF) methods are efficient to manage the inter-user interference for multi-user conditions [9], [10]. The analog-only precoding is further developed to reduce the hardware consumption for massive MIMO systems with only analog phase shifters (APSs). However, the APSs impose constant modulus constraint on the precoding matrix, which leads to poor precoding performance [11].

As a promising precoding scheme for massive MIMO systems, the hybrid architecture has been investigated to provide a tradeoff between consumption and performance [12]. The hybrid precoding architecture combines the digital precoder in the baseband and the analog precoder in the radio frequency (RF) domain. Benefited from the low-dimension digital precoder, fewer RF chains are required for implementation.

A. Related Works

Recent research on hybrid precoding focuses on the fully-connected [8], [13]–[19] and the sub-connected structures [20]–[27], which can be distinguished by the connection state between RF chains and antennas as illustrated in Fig. 1(a) and Fig. 1(b). In the fully-connected structure, each antenna is supported by all RF chains through APSs and RF adders. Considering single-user scenarios, the precoding design is formulated as a sparse reconstruction problem to minimize the Euclidean distance between the optimal full digital precoding matrix and the hybrid precoding matrix [8]. In particular, the array response vectors are spanned to generate the codebook of the analog precoding matrix [8], [13]. Based on the similar thought, the columns of analog precoding matrix are selected from the discrete Fourier transform matrix in [14]. With the limitation of the codebook,

Digital Object Identifier 10.1109/TVT.2019.2949578
For AHP, we utilize the characteristic of the structure and for FHP, we propose to formulate the precoding problem. The authors in [26] reformulate the beam selection problem for uplink multi-user massive MIMO systems in mmWave communications. In this paper, we propose hybrid precoding schemes for multi-user massive MIMO systems in mmWave communications.

B. Contributions

In this paper, we propose hybrid precoding schemes for multi-user massive MIMO systems in mmWave communications. Innovatively, we formulate the FHP and AHP design problems as clustering problems and propose to solve them based on unsupervised learning methods. The main contributions are summarized as follows:

- For FHP, we propose to formulate the precoding problem as a clustering problem by minimizing the upper bound of the Euclidean distance between the optimal full digital precoder and the hybrid precoder. By exploring the relevance among RF chains in the optimal hybrid precoder, the hierarchical-agglomerative-clustering-based (HAC-based) FHP scheme is proposed with a novel defined distance function. Moreover, the upper bound of the proposed scheme is analyzed.
- For AHP, we utilize the characteristic of the structure and simplify the precoding problem as a semi-unitary matrix factorization problem, which is equivalent to a clustering problem. By exploring the relevance among antennas at the BS, the modified-K-means-based (MKM-based) AHP scheme is proposed to make full use of the flexible connection.

The numerical results in [25] illustrate that the fully-connected structure provides better precoding performance than the sub-connected structure. Nevertheless, the implementation of the fully-connected structure in massive MIMO systems requires high hardware consumption due to the large demand of APSs and RF adders. To provide a tradeoff between two structures, the adaptively-connected hybrid precoding (AHP) scheme is adopted in [28]–[32]. As shown in Fig. 1(c), the adaptively-connected structure is the generalization of the sub-connected structure. The adaptive connection network provides a flexible connection between RF chains and antennas, which means better precoding performance can be achieved with similar hardware consumption as the sub-connected structure. For single-user scenarios, the authors in [28] propose the connecting scheme based on maximizing the sum of the largest singular values of several subchannel matrices. For multi-user scenarios, the decoupling-based schemes are further revised in [29], [30], where the analog precoding matrix is designed to improve the users’ average achievable rate with the thought of the greedy algorithm. So far, fewer research efforts have been invested in AHP. Especially for multi-user scenarios, the decoupling-based schemes ignore the relationship between analog and digital precoders and make less use of the flexibility in the adaptive connection network, which results in poor performance.

In the sub-connected structure, each RF chain is connected to a specific subset of antennas. Since there is no overlap among antenna subsets, no RF adders are required. For single-user systems, the principle of manifold optimization and particle swarm optimization are respectively considered to develop two algorithms with different computational complexity in [20]. The authors in [21] discuss the analog precoding design for high and low SNR conditions, respectively. The original multi-stream transmission problem is decomposed into several single-stream transmission problems with per-antenna power constraint. For multi-user systems, the codebook-based and the decoupling-based schemes can still be operated [22]–[25]. Additionally, machine learning is proposed as a novel approach for the sub-connected hybrid precoding (SHP) design [26], [27]. The authors in [26] reformulate the beam selection problem for uplink precoding as a multi-class-classification problem which can be solved by the support vector machine algorithm. In [27], the analog precoder is realized by several switches and inverters. And an adaptive cross-entropy-based scheme is developed for the new architecture. Essentially, the new structure can be equivalently realized by one-bit quantized APSs. Consequently, it is common for recent study to impose extra constraints for sub-connected structure, such as codebook-based analog precoder and APSs with few quantization bits, which limits the freedom of design.

The numerical results in [25] illustrate that the fully-connected structure can provide better precoding performance than the sub-connected structure. Nevertheless, the implementation of the fully-connected structure in massive MIMO systems requires high hardware consumption due to the large demand of APSs and RF adders. To provide a tradeoff between two structures, the adaptively-connected hybrid precoding (AHP) scheme is adopted in [28]–[32]. As shown in Fig. 1(c), the adaptively-connected structure is the generalization of the sub-connected structure. The adaptive connection network provides a flexible connection between RF chains and antennas, which means better precoding performance can be achieved with similar hardware consumption as the sub-connected structure. For single-user scenarios, the authors in [28] propose the connecting scheme based on maximizing the sum of the largest singular values of several subchannel matrices. For multi-user scenarios, the decoupling-based schemes are further revised in [29], [30], where the analog precoding matrix is designed to improve the users’ average achievable rate with the thought of the greedy algorithm. So far, fewer research efforts have been invested in AHP. Especially for multi-user scenarios, the decoupling-based schemes ignore the relationship between analog and digital precoders and make less use of the flexibility in the adaptive connection network, which results in poor performance. Moreover, only a portion of the RF chains can be effectively utilized in [29], [30], the number of which is equal to that of users.

In this paper, we propose hybrid precoding schemes for multi-user massive MIMO systems in mmWave communications. Innovatively, we formulate the FHP and AHP design problems as clustering problems and propose to solve them based on unsupervised learning methods. The main contributions are summarized as follows:

- For FHP, we propose to formulate the precoding problem as a clustering problem by minimizing the upper bound of the Euclidean distance between the optimal full digital precoder and the hybrid precoder. By exploring the relevance among RF chains in the optimal hybrid precoder, the hierarchical-agglomerative-clustering-based (HAC-based) FHP scheme is proposed with a novel defined distance function. Moreover, the upper bound of the proposed scheme is analyzed.
- For AHP, we utilize the characteristic of the structure and simplify the precoding problem as a semi-unitary matrix factorization problem, which is equivalent to a clustering problem. By exploring the relevance among antennas at the BS, the modified-K-means-based (MKM-based) AHP scheme is proposed to make full use of the flexible connection.
In the MKM-based AHP scheme, we propose to update the clustering centers with the alternating-optimization-based (AO-based) algorithm, where a specific initialization scheme is developed to reduce the computational complexity. In addition, we clarify that the proposed AO-based algorithm is feasible for SHP design.

Simulation results demonstrate the superiority of the proposed clustering-based precoding schemes. Specifically, the HAC-based FHP scheme provides close spectral efficiency to the full digital precoder with insufficient RF chains. The MKM-based AHP scheme contributes to high power efficiency. Including the AO-based SHP scheme, all the proposed schemes provide satisfying performance gain compared with the existing work.

C. Organization

The remainder of this paper is organized as follows. Section II introduces the system model, channel model, and problem formulation for the precoding design. In Section III and IV, two clustering-based schemes are proposed for FHP and AHP, respectively. Particularly, the AO-based algorithm in Section IV-B is proposed as a suitable scheme for SHP design. Then, simulation results are presented in Section V to demonstrate the superior performance of the proposed schemes. Finally, some conclusions are given in Section VI.

Notations: $a$, $a$ and $A$ denote a scalar, vector and matrix, respectively. For a given matrix $A$, $A^T$, $A^H$, $A^{-1}$, and $r(A)$ denote its transpose, conjugate transpose, inverse, pseudo-inverse and rank, respectively, $A(m,n)$, $A(m,:)$ and $A(:,n)$ denote the $(m,n)$-th entry, the $m$-th row and $n$-th column of $A$, respectively. The Frobenius norm and $\ell_2$ norm are noted by $\| \cdot \|_F$ and $\| \cdot \|_2$. $I_N$ denotes the $N \times N$ identity matrix, while $O_{M \times N}$ denotes the $M \times N$ all-zero matrix. $CN(\alpha, \Sigma)$ denotes the complex Gaussian distribution with mean $\alpha$ and covariance $\Sigma$. $\emptyset$ denotes the empty set. $E[\cdot]$ denotes the expectation. The magnitude and phase of a complex scalar are denoted by $|\cdot|$ and $\arg\{\cdot\}$. The SVD of $A$ is in the form of $A = U\Sigma V^H$, where $U$ and $V$ are left-singular and right-singular matrices, and $\Sigma$ is a rectangular diagonal matrix with descending ordered singular values on the diagonal.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

As shown in Fig. 1(d), we consider the downlink communication of a multi-user massive MIMO system with hybrid precoding. The base station (BS) is equipped with $N_{RF}$ RF chains and $N_t$ transmit antennas to serve $K$ users. Based on the principles of low-cost and low-power consumption for mobile terminal design [19], we further assume that the $k$-th user is equipped with one RF chain and $N_{t,k}$ receive antennas. Thus, each user shall require only a single data stream from the BS. Due to the low dimension of digital precoding in the hybrid structure, we typically have $K \leq N_{RF} \ll N_t$. Mathematically, the linear precoded transmit signal from the BS can be represented as $s = F_{RF}F_{BB}x$, where $F_{RF} \in \mathbb{C}^{N_t \times N_{RF}}$ denotes the analog precoder in the RF domain, $F_{BB} \in \mathbb{C}^{N_{RF} \times K}$ denotes the digital precoder in the baseband, and $x \in \mathbb{C}^{K \times 1}$ denotes the transmit symbol vector. Without loss of generality, the average total transmit power of the BS is set as $P$ with $x$ satisfying $E[|x|^2] = \frac{P}{K}$. Thus, the power constraint of the overall hybrid precoder can be given by

$$\|F_{RF}F_{BB}\|_F^2 = K.$$  \hspace{1cm} (2)

For simplicity, the block-fading channel model is adopted in this paper. At the $k$-th user, the received signal is further processed by own RF combiner, which can be expressed as

$$r_k = w_k^HF_kF_{RF}F_{BB}x + w_k^Hn_k,$$  \hspace{1cm} (3)

where $w_k \in \mathbb{C}^{N_{t,k} \times 1}$ denotes the RF combiner of the $k$-th user, $H_k \in \mathbb{C}^{N_{t,k} \times N_t}$ denotes the channel matrix between the BS and the $k$-th user, $n_k \in \mathbb{C}^{N_{t,k} \times 1}$ denotes a complex Gaussian noise vector with each element obeying $CN(0, \sigma_n^2)$ (assumed same for each user, i.e., $\sigma_n^2 = \sigma_r^2 \forall k$).

The Saleh-Valenzuela model is commonly accepted to characterize the limited scattering feature of a mmWave channel [16], [18], [31]. The normalized channel for the $k$-th user consists of $N_{c,k}$ scattering clusters, each of which is a sum of contributions of $N_{p,k}$ propagation paths, i.e.,

$$H_k = \sqrt{N_1 \prod_{k=1}^{K} N_{t,k} \prod_{p=1}^{N_{p,k}} \sum_{c=1}^{N_{c,k}} \sum_{p=1}^{N_{p,k}} \beta_{c,p} a_t(\theta_{c,p}^t, \phi_{c,p}^t) a_r(\theta_{c,p}^r, \phi_{c,p}^r)},$$  \hspace{1cm} (4)

where $\beta_{c,p} \sim CN(0, 1)$ denotes the complex gain of the $p$-th path in the $c$-th cluster. In addition, $a_t(\theta_{c,p}^t, \phi_{c,p}^t)$ and $a_r(\theta_{c,p}^r, \phi_{c,p}^r)$ denote the normalized receive and transmit array response vectors corresponding to the azimuth (elevation) angle of arrival $\theta_{c,p}^t$ ($\phi_{c,p}^t$) and departure $\theta_{c,p}^r$ ($\phi_{c,p}^r$), respectively. Since the proposed algorithms in this paper are applicable for arbitrary antenna arrays, the uniform planar array (UPA) will be considered for the completeness of simulations. In the case of UPA, the $W \times V$-element array’s response is variant in two angle domains, which can be expressed as

$$a_{UPA}(\theta, \phi) = \frac{1}{\sqrt{WV}} \begin{bmatrix} e^{j \frac{2\pi}{\lambda} (W \sin(\theta) \sin(\phi) + v \cos(\phi))} \\ \vdots \\ e^{j \frac{2\pi}{\lambda} ((W-1) \sin(\theta) \sin(\phi) + (V-1) \cos(\phi))} \end{bmatrix}^T,$$  \hspace{1cm} (5)

where $0 \leq w < W$ and $0 \leq v < V$. As the basis of the precoding design in this paper, $H_k$ is assumed known at the BS and the $k$-th user, i.e., the BS owns the global channel state information (CSI), while each user only holds its own part [8], [27], [29].

B. Problem Formulation for Precoder Design

Similar to the power constraint of the hybrid precoder (2), all combiners satisfy a normalized power constraint for the consistency among users [24], [25], i.e.,

$$\|w_k\|_F^2 = 1.$$  \hspace{1cm} (6)

Since $w_k$ is always a unit vector, the power of the Gaussian noise processed by RF combiner still maintains at $\sigma_r^2$ due to the unitary transformation. Recall the expression of the received signal (3), we obtain the signal-to-interference-plus-noise-ratio (SINR) of
the $k$-th user as

$$\text{SINR}_k = \frac{\mathbf{P}^H \mathbf{w}_k \mathbf{H}_k \mathbf{F}_{RF} \mathbf{F}_{BB}(i,k)}{K \sigma^2 + \sum_{l \neq k} \mathbf{P}^H \mathbf{w}_l \mathbf{H}_l \mathbf{F}_{RF} \mathbf{F}_{BB}(i,l)}.$$ (7)

Since only a part of global CSI is available for each user, the inter-user interference is not visible at the user side. Thus, the multi-user system degenerates into a point-to-point system for each user, which motivates a selfish design for each combiner to directly maximize the channel gain. According to [8], the optimal selfish combiner for the $k$-th user is given by $\mathbf{w}_k = \mathbf{U}_k (\cdot ; 1)$ with the SVD $\mathbf{U}_k \sum_k \mathbf{V}_k^H = \mathbf{H}_k$. Further, based on the constraint of APs, the RF combiner can be given by

$$\mathbf{w}_k(i) = \frac{1}{\sqrt{N_t,k}} e^{j\arg\{U_k(i,1)\}}, (8)$$

In practice, the available phases for APs are quantized in general, i.e.,

$$\mathbf{w}_{k,q}(i) \in \left\{ \frac{1}{\sqrt{N_t,k}} e^{j\frac{2\pi}{2^Q} q} : q = 0, 1, \ldots, 2^Q - 1 \right\}, \forall i, (9)$$

where $Q$ denotes the quantization bit number of APs. Accordingly, the quantized combiner can be further given by

$$\mathbf{w}_{k,q}(i) = \frac{1}{\sqrt{N_t,k}} e^{j\hat{q} \frac{2\pi q}{2^Q}}, (10)$$

with

$$\hat{q} = \arg\min_{q=1,2,\ldots,2^Q-1} \left| \mathbf{w}_k(i) - \frac{e^{j\frac{2\pi q}{2^Q}}}{\sqrt{N_t,k}} \right|. (11)$$

Since the similar operation is feasible for quantized RF precoder design, without loss of generality, the non-quantized RF combiner and precoder will be mainly discussed in the following paper.

With the global CSI, all of the combining weights can be obtained at the BS. Then, the aim of the precoding design is to manage the inter-user interference and enhance system spectral efficiency, i.e.,

$$R = \sum_{k=1}^{K} \log_2(1 + \text{SINR}_k). (12)$$

In this paper, we consider minimizing the Euclidean distance between the optimal full digital precoding matrix $\mathbf{F}_{opt}$ and the hybrid precoding matrix as follows:

$$\min_{\mathbf{F}_{RF}, \mathbf{F}_{BB}} \| \mathbf{F}_{opt} - \mathbf{F}_{RF} \mathbf{F}_{BB} \|_F^2$$

s.t. (2), constraints of $\mathbf{F}_{RF}, (13)$$

which has been proved as a sufficient precoding design scheme in [8]. In addition, the optimal full digital precoder is provided by the classical MF, ZF, and RZF method$^1$ as

$$\mathbf{F}_{opt} = \begin{cases} \mathbf{\sqrt{\mathbf{H}}^H}, \\
\mathbf{\sqrt{\mathbf{H}}} & \text{if} \text{MF} \\
\mathbf{\sqrt{\mathbf{H}}} & \text{if} \text{ZF} \\
\mathbf{\sqrt{\mathbf{H}}} (\mathbf{H}_{eq} \mathbf{H}_{eq}^H + \beta \mathbf{I}_K)^{-1} & \text{if} \text{RZF} \end{cases} (14)$$

where $\mathbf{\sqrt{\mathbf{H}}}$, $\mathbf{\sqrt{\mathbf{H}}}$, $\mathbf{\sqrt{\mathbf{H}}}$ denote normalization parameters to ensure the digital precoder satisfying the power constraint

$^1$The full digital precoder design is not the focus of this paper.

III. CLUSTERING BASED FULLY-CONNECTED HYBRID PRECODING

In this section, we first clarify the constraint of the RF precoder for FHP. Considering the optimal hybrid precoding scheme in [15], we formulate FHP with insufficient RF chains as a clustering problem by minimizing the upper bound of the Euclidean distance in the objective function. Then, with a novel defined inter-cluster distance function, we propose the HAC-based FHP scheme to solve the clustering problem, which is successively operated with the clustering and center design...
parts. The proposed scheme explores the relevance among RF chains in the optimal hybrid precoder and merges the similar RF chains to efficiently reflect the original performance.

A. Problem Derivation for FHP

Since each pair of RF chain and antenna is connected via an APS and an RF adder in FHP, the constraint of RF precoder in (13) can be specified by

$$\| F_{RF}(i, n) \| = 1. $$ \hspace{1cm} (16)

It has been pointed out in [15] that twice the number of transmit streams) RF chains are sufficient for FHP to realize the same performance as the full digital precoder, i.e., \( F_{opt} = F_{RF}^*, F_{BB}^* \). With the single-stream transmission for each user, the original problem (13) is equivalent to

$$ \begin{align*}
\min_{F_{RF}, F_{BB}} & \|F_{RF}^* F_{BB}^* - F_{RF} F_{BB}\|_F \\
\text{s.t.} & \quad (2), (16),
\end{align*} $$ \hspace{1cm} (17)

where the number of columns in \( F_{RF}^* \) and the number of rows in \( F_{BB}^* \) are both 2K. For the cases with insufficient RF chains, we have \( K \leq N_{RF} \leq 2K \). Due to the complexity of (17), we consider to minimize the upper bound of the objective function based on \( \| A + B \|_F \leq \| A \|_F + \| B \|_F \) as follows

$$ \begin{align*}
\| F_{RF} F_{BB} - F_{RF} F_{BB} \|_F & = \left\| \left( \sum_{m=1}^{2K} F_{HP}^*(m) \right) - \left( \sum_{c=1}^{N_{RF}} F_{HP}(c) \right) \right\|_F \\
\text{(a)} & = \left\| \left( \sum_{c=1}^{N_{RF}} \sum_{m \in \Gamma_c} F_{HP}^*(m) \right) - \left( \sum_{c=1}^{N_{RF}} F_{HP}(c) \right) \right\|_F \\
& \leq \sum_{c=1}^{N_{RF}} \left\| \sum_{m \in \Gamma_c} F_{HP}^*(m) - F_{HP}(c) \right\|_F, \hspace{1cm} (18)
\end{align*} $$

where \( F_{HP}^*(m) = F_{HP}(m,:), F_{HP}(c) = F_{HP}(:,c) \). In the equation (a) of (18), 2K matrices \( F_{HP}^*(m) \) are divided into \( N_{RF} \) groups, where \( \Gamma_c \) is the index set of the \( c \)-th group. To ensure the completeness of the membership and the dissimilarity among groups, \( \Gamma_c \) is constrained by

$$ \Gamma_c \cap \Gamma_d = \emptyset, \forall c \neq d, $$

$$ \bigcup_{c=1}^{N_{RF}} \Gamma_c = \{1, 2, \ldots, 2K\}. $$ \hspace{1cm} (19)

Accordingly, a clustering problem for FHP (CP-FHP) can be obtained as follows

$$ \begin{align*}
\min_{F_{RF}, F_{BB}} & \sum_{c=1}^{N_{RF}} \left\| \sum_{m \in \Gamma_c} F_{HP}^*(m) - F_{HP}(c) \right\|_F \\
\text{s.t.} & \quad (2), (16), (19).
\end{align*} $$ \hspace{1cm} (20)

CP-FHP is an atypical clustering problem because of the special form of objective function, which will be discussed in detail in Section III-B. In CP-FHP, the set of data samples is composed of 2K matrices with \( F_{HP}^*(m) \) denoting each of them. The data samples are expected to be divided into \( N_{RF} \) clusters. For the \( c \)-th cluster, \( F_{HP}(c) \) denotes the clustering center, while \( \Gamma_c \) denotes the index set of members.

In terms of data set composition, each RF chain in the optimal hybrid precoder generates a data sample, which implies the design of FHP is essentially the clustering of RF chains. And the design scheme can be developed with the successive operation of the clustering and center design parts. The clustering part is to determine \( N_{RF} \) member index sets corresponding to expected clusters. For different \( N_{RF} \) cases, the HAC method is feasible for this part, where the key point is to define the distance between clusters. Then, in the center design part, each cluster will generate its own clustering center which is essentially corresponding to a certain component of the expected hybrid precoding matrix.

B. Clustering and Center Design

With each initial cluster formed by a single data sample, the key approach of the clustering part in the HAC method is to find a pair of clusters with the smallest distance and merge them into a new cluster [33]. For the fairness among RF chains, the distance between the \( c \)-th and the \( d \)-th cluster is given by the mean distance between members of each cluster as

$$ D(c, d) = \frac{1}{|\Gamma_c|} \sum_{m \in \Gamma_c} \sum_{n \in \Gamma_d} D_F(F_{HP}^*(m), F_{HP}^*(n)), $$ \hspace{1cm} (21)

where \( |\Gamma| \) denotes the number of members in set \( \Gamma \), \( D_F(A, B) = \| A - B \|_F \) denoting the inter-sample distance between \( A \) and \( B \).

In a typical clustering problem, the inter-sample distance is commonly given by the Frobenius norm \( D_F(A, B) = \| A - B \|_F \). This is mainly owing to the typical form of the objective function, where the clustering center can be regarded as a virtual sample. Specifically, for CP-FHP, the Frobenius norm is adoptable only if the objective function is in the typical form like

$$ \sum_{c=1}^{N_{RF}} \sum_{m \in \Gamma_c} \| F_{HP}^*(m) - F_{HP}(c) \|_F, $$

where the intra-cluster distance is determined by the summation of the distance between each data sample and the clustering center. However, the atypical objective function in CP-FHP measures the intra-cluster distance by the distance between the clustering center and the summation of all data samples. Thus, we consider redefining the inter-sample distance function based on the following proposition.

**Proposition 1:** Under the condition that the column space of \( F_{RF} \) keeps unchanged, arbitrarily adjusted \( F_{RF} \) will make no difference to hybrid precoding performance if \( F_{BB} \) is appropriately adjusted.

**Proof:** Please refer to Appendix A. □

According to **Proposition 1**, the maintenance of the column space of \( F_{RF} \) is the key point for the precoding design. In CP-FHP, each sample \( F_{HP}^*(m) \) contains a part of the column space of \( F_{RF} \). With the process of clustering, the samples in the same cluster are centered by a rank-one matrix. Note that each clustering center with a low rank is a certain component of the expected hybrid precoding matrix. To reduce the loss of the column space caused by the reduction of the rank, the samples with similar column space should be merged into the same
cluster. Intuitively, it inspires us to measure the inter-sample distance based on the similarity of the column space as follows

\[ D_F (F_{\text{RF}}^\star (m), F_{\text{RF}}^\star (n)) = \left\| F_{\text{RF}}^\star (\cdot, m)^H F_{\text{RF}}^\star (\cdot, n) \right\|_F^{-1}, \]

(22)

where a large inner product means a high similarity in column space and a small distance between samples.

With the definition of the distance for clusters and samples, the clustering part is operable to determine the membership of each cluster, i.e., \( \forall \Gamma_c \in \text{CP-FHP} \) is determined. Then, CP-FHP can be decomposed into \( N_{\text{RF}} \) subproblems for center design (CD-FHP), i.e.,

\[
\min_{F_{\text{RF}}^\star (\cdot, c)} \left\| \sum_{m \in \Gamma_c} F_{\text{RF}}^\star (m) - F_{\text{RF}}^\star (\cdot, c) F_{\text{BB}} (c, : ) \right\|_F,
\]

s.t.

\[ (16), \]

(23)

where the center of the \( c \)-th cluster is obtained to provide the design for a certain RF chain. Note that \( F_{\text{BB}} (c, :) \) is not optimized in CD-FHP, which will be explained in the remark of Section III-B. Due to the decomposition of CP-FHP, the power constraint (2) is temporarily neglected in CD-FHP, which slightly brings about negative impacts [34].

**Algorithm 1:** Proposed HAC-based FHP Algorithm.

**Input:** The optimal full digital precoder \( F_{\text{opt}} \).

**Output:** \( F_{\text{RF}}, F_{\text{BB}} \).

1: Obtain the ideal fully-connected hybrid precoder design \( F_{\text{opt}} = F_{\text{RF}}^\star, F_{\text{BB}}^\star \) based on the expressions in [15].

2: Initialize the number of clusters by \( N_{\text{cl}} = 2K \).

3: for \( c = 1 \) to \( N_{\text{cl}} \) do

4: Initialize the \( c \)-th cluster with the member \( F_{\text{RF}}^\star (c) = F_{\text{RF}} (\cdot, c) F_{\text{BB}} (c, :) \) and the member index set \( \Gamma_c = \{ c \} \).

5: end for

6: For \( \forall n \neq m \), calculate the distance between \( F_{\text{RF}}^\star (n) \) and \( F_{\text{RF}}^\star (m) \) based on (22).

7: while \( N_{\text{cl}} > N_{\text{RF}} \) do

8: Calculate inter-cluster distance \( D (c, d), \forall c \neq d \) based on (21).

9: Find the pair of clusters with the smallest distance, i.e., \( (\Gamma_{c^*}, \Gamma_{d^*}) = \arg \min_{\Gamma_c, \Gamma_d} D (c, d) \).

10: Merge the nearest clusters \( \Gamma_{c^*} = \Gamma_{c^*} \cup \Gamma_{d^*}, \Gamma_{d^*} = \emptyset \).

11: Update the number of current clusters by \( N_{\text{cl}} = N_{\text{cl}} - 1 \), and renumber the non-empty set \( \Gamma \) from 1 to \( N_{\text{cl}} \).

12: end while

13: For each \( \Gamma_c \), generate \( F_{\text{RF}} (\cdot, c) \) based on (24).

14: Obtain \( F_{\text{BB}} \) by LS method, or the classical digital precoding methods (14) with the effective baseband channel.

The difficulty of CD-FHP lies in the modulus constraint (16). To solve CD-FHP, we consider a heuristic scheme to provide a near-optimal solution. Note that the objective function of CD-FHP is in the same form of a low-rank matrix approximation problem. If the modulus constraint (16) is neglected, the optimal solution can be given by the Eckart-Young-Mirsky theorem with \( F_{\text{RF}} (\cdot, c) = U_c (\cdot, 1) \), where \( U, \Sigma, \Sigma \dagger = \sum_{m \in \Gamma_c} F_{\text{RF}}^\star (m) \) [35]. Therefore, a near-optimal solution can be given by

\[
F_{\text{RF}} (i, c) = e^{j \arg (U_c (i, 1))},
\]

(24)

Owing to the inter-sample distance definition, the samples in the same cluster tend to have similar column space, which means more power of \( \sum_{m \in \Gamma_c} F_{\text{RF}}^\star (m) \) is converged in the maximum singular value, and contributes to a smaller approximation error in CD-FHP.

**Remark:** Actually, the digital and analog precoders can be jointly optimized in CD-FHP, since \( F_{\text{BB}} (c, :) \) is also involved in the objective function. However, it makes no sense to optimize \( F_{\text{BB}} (c, :) \) in CD-FHP, since the power constraint is not considered. More importantly, \( F_{\text{BB}} (c, :) \) is only optimized for the subproblem. But for the original problem (17), it is obviously a better scheme to revise the overall \( F_{\text{BB}} \) with the near-optimal \( F_{\text{RF}} \) in (24). Thus, we only focus on the optimization of \( F_{\text{RF}} (\cdot, c) \) in CD-FHP.

**C. Proposed HAC-Based FHP Scheme**

The HAC-based FHP design is summarized in Algorithm 1 to solve CP-FHP. At the beginning of the algorithm, each sample is regarded as a single cluster. From step 7 to step 12, the pair of clusters with minimum distance are merged in each loop, until the number of clusters is reduced to \( N_{\text{RF}} \). With the determined clusters, the RF precoder is obtained based on (24). Finally, the baseband precoder can be obtained by two alternative schemes. One scheme is based on the least-square (LS) method, i.e., \( F_{\text{BB}} = \sqrt{\lambda_{\text{eq}}} F_{\text{RF}} \), where \( \sqrt{\lambda_{\text{eq}}} \) denotes the normalization parameter to satisfy the precoding power constraint. Obviously, the performance of this scheme mainly depends on the similarity between the column spaces of \( F_{\text{RF}} \) and \( F_{\text{opt}} \). The other scheme is based on classical digital precoding methods. Similar to (14), the second scheme is operated with the effective baseband channel \( H_{\text{eq}} = \sqrt{\lambda_{\text{eq}}} F_{\text{RF}} \), which abates the influence of column space similarity to a certain extent. In terms of the upper bound, we provide the comparison of two alternative schemes in the following lemma. Detail discusses for the selection of these two schemes will be presented in Section V. Algorithm 1 is convergent with the overall complexity \( o(N_{\text{RF}} K N_{\text{cl}}^2) \) for massive MIMO systems, which is analyzed in Appendix C.

**Lemma 1:** With \( F_{\text{opt}} = \sqrt{\lambda_{eq}} H_{eq}^\dagger \) as the input, the upper bounds of the system spectral efficiency for the HAC-based FHP design with the LS and ZF refinement schemes are respectively given by

\[
R_{\text{LS}} = K \log_2 \left( 1 + \frac{P}{\sigma^2 \left\| H_{eq} \right\|_F^2} \right),
\]

\[
R_{\text{ZF}} = K \log_2 \left( 1 + \frac{P}{\sigma^2 \left\| F_{\text{RF}} H_{eq} \right\|_F^2} \right).
\]

(25)

And \( R_{\text{ZF}} \leq R_{\text{LS}} \) is always satisfied.

**Proof:** Please refer to Appendix B.
In the proposed HAC-based FHP scheme, the optimal hybrid precoding is decomposed into $2K$ components, with each RF chain corresponding to each of them. In order to reduce the performance loss caused by insufficient RF chains, similar components in the optimal hybrid precoder are clustered based on the relevance among RF chains. And the clustered components are represented by a clustering center corresponding to a new RF chain design, which can efficiently reflect the performance of the overall cluster.

**Remark:** Actually, the K-means algorithm is also operable to solve CP-FHP [36]. However, due to the non-convexity of the problem and the fact that the number of expected clusters is even more than half the number of samples, the performance of the K-means algorithm is greatly affected by the initial value. Thus, it is not advisable to solve CP-FHP with the K-means algorithm.

### IV. Clustering Based Adaptively-Connected Hybrid Precoding

In this section, we first analyze the constraint of the RF precoder for AHP, and formulate AHP as a clustering problem based on the SVD of the precoding matrix. Then, we propose the MKM-based AHP scheme with the iterative refinement of the clustering and the center updating parts. Particularly, the clustering centers are updated by the AO-based algorithm with a specific initialization scheme to reduce the computational complexity, which is also capable to provide SHP design independently. The MKM-based AHP scheme explores the relevance among antennas and connects similar antennas with a single RF chain to make full use of the flexible connection.

#### A. Problem Derivation for AHP

The major characteristic of AHP is that each RF chain is connected with a flexible subset of the antennas. Accordingly, the hardware constraint in (13) can be expressed as

$$\sum_{n=1}^{N_{RF}} |\mathbf{F}_{RF}(i, n)| = 1, \forall i,$$  \hspace{1cm} (26a)

$$\sum_{i=1}^{N_{t}} |\mathbf{F}_{RF}(i, n)| = M, \forall n.$$  \hspace{1cm} (26b)

The constraint (26a) results from that each antenna is supported by an individual RF chain, while the constraint (26b) results from that each RF chain supports $M = N_t / N_{RF}$ (commonly assumed as an integer [29]-[31]) antennas. $\mathbf{F}_{RF}(i, n) \neq 0$ represents that the $i$-th antenna is supported by the $n$-th RF chain via an APS. Owing to the adaptive connection, the position of nonzero entries in $\mathbf{F}_{RF}$ is flexible.

Utilizing the SVD method, the AHP design problem can be given as

$$\min_{\mathbf{F}_{RF}, \mathbf{U}_{BB}} \left\| \mathbf{U}_{BB} \Sigma_{BB} \mathbf{V}_{BB}^H - \mathbf{F}_{RF} \mathbf{U}_{BB} \Sigma_{BB} \mathbf{V}_{BB}^H \right\|_F^2,$$

s.t. \hspace{1cm} (2), (26a), (26b),  \hspace{1cm} (27)

where $\mathbf{U}_{opt} \in \mathbb{C}^{N_t \times K}$, $\mathbf{U}_{BB} \in \mathbb{C}^{N_{RF} \times K}$, and $\Sigma_{opt}, \Sigma_{BB} \in \mathbb{C}^{K \times K}$. Since the special characteristic of AHP keeps the equation $\mathbf{F}_{RF}^H \mathbf{F}_{RF} = M \mathbf{I}_{N_{RF}}$ always established, the singular values of the hybrid precoding matrix are mainly determined by the baseband precoder. In other words, $\mathbf{F}_{RF} \mathbf{U}_{BB} \Sigma_{BB} \mathbf{V}_{BB}^H$ is still the SVD of the hybrid precoder, whereas $\mathbf{F}_{RF} \mathbf{U}_{BB}$ is the left-singular matrix. To tackle the complexity of (27), we consider the approximation problem for each part of the SVD of the hybrid precoder. Specifically, we design $\Sigma_{BB}$ and $\mathbf{V}_{BB}$ based on corresponding SVD parts of $\mathbf{F}_{opt}$ as follows

$$\Sigma_{BB} = \Sigma_{opt} / \sqrt{M},$$

$$\mathbf{V}_{BB} = \mathbf{V}_{opt},$$  \hspace{1cm} (28)

where the coefficient $\sqrt{M}$ is to satisfy the power constraint (2). Further, the approximation for the left-singular matrix can be formulated as a semi-unitary matrix factorization problem, i.e.,

$$\min_{\mathbf{F}_{RF}, \mathbf{U}_{BB}} \left\| \sqrt{M} \mathbf{U}_{opt} - \mathbf{F}_{RF} \mathbf{U}_{BB} \right\|_F^2,$$

s.t. \hspace{1cm} (26a), (26b), $\mathbf{U}_{BB}^H \mathbf{U}_{BB} = \mathbf{I}_K.$  \hspace{1cm} (29)

By minimizing the distance between corresponding SVD parts of $\mathbf{F}_{RF} \mathbf{U}_{BB}$ and $\mathbf{F}_{opt}$, the objective function value in the original problem can be reduced. Actually, because of the unequal singular values in $\Sigma_{BB}$, the derivation from (27) to (29) is a kind of approximation which will lead to performance loss. However, it provides an extra unitary constraint, which is helpful to solve the problem. Despite the performance loss caused by the approximation, the proposed scheme can still provide better performance than the existing work as shown in Section V.

With the constraint (26a), the rows in $\mathbf{F}_{RF} \mathbf{U}_{BB}$ are rotated by the rows in $\mathbf{U}_{BB}$ with nonzero $\mathbf{F}_{RF}(m, n)$. Further according to $\| \mathbf{A} \|_F^2 = \sum_m \| \mathbf{A}(m, :) \|_2^2$, a clustering problem for AHP (CP-AHP) can be formulated as

$$\min_{\mathbf{F}_{RF}, \mathbf{U}_{BB}} \sum_{n=1}^{N_{RF}} \sum_{m \in \Gamma_n} \left\| \sqrt{M} \mathbf{U}_{opt}(m, :) - \mathbf{F}_{AHP, m}(n, :) \right\|_2^2,$$

s.t. \hspace{1cm} (26a), (26b), $\mathbf{U}_{BB}^H \mathbf{U}_{BB} = \mathbf{I}_K,$  \hspace{1cm} (30)

where $\mathbf{F}_{AHP, m}(n, :) \triangleq \mathbf{F}_{RF}(m, n) \mathbf{U}_{BB}(n, :), \Gamma_n$ is the index set of nonzero entries $\mathbf{F}_{RF}(m, n)$ in $n$-th column of $\mathbf{F}_{RF}$. In CP-AHP, the set of data samples is consisted of the rows of $\sqrt{M} \mathbf{U}_{opt}$, which are expected to be clustered into $N_{RF}$ clusters. For the $n$-th cluster, $\mathbf{U}_{BB}(n, :)$ denotes the clustering center, while $\Gamma_n$ denotes the index set of members.

CP-AHP can be solved based on the K-means algorithm. The key approach of the K-means algorithm is the iterative refinement of the clustering and the center updating parts with the thought of the greedy algorithm [36]. However, due to the extra constraints in CP-AHP, the classical K-means algorithm is no longer applicable, which means modification is imperative.

#### B. Clustering and AO-Based Center Updating

The clustering part of the iterative refinement is to assign each sample to the nearest cluster with fixed clustering centers $\mathbf{U}_{BB}$.
The key to this part is to define the distance between samples and clustering centers.

Since clustering centers can be arbitrarily rotated by the nonzero entries in $F_{RF}$, the Euclidean distance cannot be directly used for distance definition. Further considering that if $\sqrt{M}U_{opt}(m,:)$ is assigned to the cluster centered by $U_{BB}(t,:)$, $F_{RF}(m,n)$ should be nonzero (i.e., $m \in \Gamma_n$) and determined by the following problem (DD-AHP) which also helps for distance definition

$$\min_{F_{RF}(m,n)} \left\| \sqrt{M}U_{opt}(m,:) - F_{RF}(m,n)U_{BB}(n,:) \right\|_F^2$$

s.t. $|F_{RF}(m,n)| = 1.

There exists a closed-form solution for DD-AHP given by [34]

$$F_{RF}(m,n) = e^{j \arg(U_{opt}(m,:))U_{BB}(n,:)^H}. \quad (32)$$

Thus, the distance between the sample vector $s$ and the clustering center $c$ can be defined by the value of the objective function in DD-AHP as follows

$$D_A(s,c) = \| s - e^{j \arg(\sigma c^H)c} \|_2^2. \quad (33)$$

The other part of the iterative refinement is to update clustering centers with fixed $\Gamma_n$. The fixed nonzero positions of $F_{RF}$ mean the connection between antennas and RF chains is determined, which degenerates AHP into SHP. Mathematically, the center updating problem for AHP (CU-AHP) can be given as

$$\min_{F_{RF}, U_{BB}} \left\| \sqrt{M}RU_{opt} - RF_{RF}U_{BB} \right\|_F^2$$

s.t. (26a), (26b), $U_{BB}^H U_{BB} = I_K$, \quad (34)

where $R \in \mathbb{R}^{N_x \times N_t}$ is obtained by rearranging the rows of $I_{N_t}$ to make the equivalent analog precoding $\tilde{F}_{RF} = RF_{RF}$ a block diagonal matrix based on fixed $\Gamma_n$, i.e.,

$$\tilde{F}_{RF} = \begin{bmatrix} \tilde{f}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tilde{f}_{N_{RF}} \end{bmatrix} \in \mathbb{C}^{MN_{RF} \times N_{RF}}, \quad (35)$$

where $\tilde{f}_n \in \mathbb{C}^{M \times 1}$, Actually, the block diagonal analog precoding matrix is the key characteristic of SHP.

Since CU-AHP is a joint optimization problem with non-convex constraints, we propose to utilize the AO method to find a near-optimal solution. One part of the AO-based method (P1-AO) is to optimize $F_{RF}$ with fixed $U_{BB}$ as follows

$$\min_{F_{RF}} \left\| \sqrt{M}U_{opt} - RF_{RF}U_{BB} \right\|_F^2$$

s.t. $\left\| f_n(m) \right\|_1 = 1. \quad (36)$

where $U_{opt} = RU_{opt}$. Observing that $\tilde{F}_{RF}U_{BB}$ is formed from the rotated rows of $U_{BB}$, we equivalently decompose P1-AO into $N_t$ subproblems as

$$\min_{f_n(m)} \left\| \sqrt{M}U_{opt}(t,:) - \tilde{f}_n(m)U_{BB}(n,:) \right\|_2^2$$

s.t. $\left\| f_n(m) \right\|_1 = 1. \quad (37)$

Algorithm 2: Proposed AO-based Center Updating Algorithm.

**Input:** The left-singular matrix of the full digital precoder $U_{opt}$, the index set of nonzero entries $\Gamma_n$.

**Output:** $F_{RF}$, $U_{BB}$.

1: Calculate $R$ based on $\Gamma_n$ to ensure a block diagonal $F_{RF}$.

2: Initialize $\tilde{F}_{RF}$ based on (41).

3: Initialize the value of objective function as infinite $\nu_0 = +\infty$, and set $k = 1$.

4: repeat

5: Calculate the SVD $\tilde{F}_{RF}^H \tilde{U}_{opt} = U_B \Sigma_B V_B^H$.

6: Obtain $U_{BB}$ based on (39).

7: Update the equivalent analog precoding matrix $\tilde{F}_{RF}$ based on $\tilde{f}_n(m) = e^{j \arg(U_{opt}(t,:))U_{BB}(n,:)^H}$.

8: $k = k + 1$.

9: $\nu_k = \left\| \sqrt{M}U_{opt} - RF_{RF}U_{BB} \right\|_F^2.$

10: until $\nu_{k-1} - \nu_k < \varepsilon$

11: $F_{RF} = R^{-1} \tilde{F}_{RF}$.

where $t = Mn - M + m$. The problem (37) is in the same form as (31), which can be solved by $f_n(m) = e^{j \arg(U_{opt}(t,:))U_{BB}(n,:)^H}$.

The other part of the AO method (P2-AO) is to optimize $U_{BB}$ with fixed $F_{RF}$ as follows

$$\min_{U_{BB}} \left\| \sqrt{M}U_{opt} - RF_{RF}U_{BB} \right\|_F^2$$

s.t. $U_{BB}^H U_{BB} = I_K$. \quad (38)

P2-AO is a typical semi-orthogonal procrustes problem [37], which can be solved by

$$U_{BB} = U_BK_B^H \quad (39)$$

where $U_B \Sigma_B V_B^H = \tilde{F}_{RF}^H \tilde{U}_{opt}$, $U_B \in \mathbb{C}^{N_{RF} \times K}$.

Generally, the AO method is initialized with a random optimization variable. In CU-AHP, it could be $F_{RF}$ with random phases or a random unitary $U_{BB}$. Since each optimization variable gradually approaches the local optimum in each iteration, the initialization scheme is closely related to the convergence time. Here, we propose a simple initialization scheme for $F_{RF}$.

Regardless of the unitary constraint in CU-AHP, $f_n$ could be initialized by the following problem (INT-AO)

$$\min_{f_n} \left\| \sqrt{M}U_{opt} - f_nU_{BB}(n,:) \right\|_F^2$$

s.t. $\left\| f_n(m) \right\|_1 = 1. \quad (40)$

where $U_{opt,n} \in \mathbb{C}^{M \times K}$ denotes the submatrix of $U_{opt} = [(U_{opt,1})^H, \ldots, (U_{opt,N_{RF}})]^H$. INT-AO is in the same form as (23), which suggests to initial $F_{RF}$ with

$$\tilde{f}_{n,\text{int}}(m) = e^{j \arg(U_n(m,:))}. \quad (41)$$

where $U_n \Sigma_n V_n^H = U_{opt,n}$.

As a summary, the AO-based scheme for updating clustering centers is shown in Algorithm 2. At the beginning of the algorithm, $F_{RF}$ is initialized by (41) to reduce the computational
Algorithm 3: Proposed MKM-based AHP Algorithm.

**Input:** Full digital precoder $F_{RF}^*$, the number of RF chains $N_{RF}$.

**Output:** $F_{RF}^*, F_{BB}^*$.

1. Calculate the SVD $F_{opt}^* = U_{opt} \Sigma_{opt} V_{opt}^H$.
2. Initialize $U_{BB}$ as a random semi-unitary matrix.
3. Initialize the value of objective function as infinite $v_0 = +\infty$, and set $k = 1$.

4. **repeat**
   5. For each pair of $\sqrt{M}U_{opt}(m,:)$ and $U_{BB}(n,:)$, calculate the distance based on (33).
   6. $F_{RF} = 0_{N_{RF} \times N_{RF}}$, $\Gamma = \{1,2,\ldots,N_i\}$, $\Gamma_n = \emptyset, \forall n$.
   7. **for** $p = 1$ to $M$ **do**
   8. **for** $q = 1$ to $N_{RF}$ **do**
   9. $i^* = \arg\min_{i \in F} D_A(\sqrt{M}U_{opt}(i,:),U_{BB}(q,:))$.
   10. $\Gamma^q = \Gamma_q \cup \{i^*\}$, $\Gamma = \Gamma \setminus \{i^*\}$.
   **end for**
   **end for**
   11. **end for**
   12. Fixed $\Gamma_n$, refine $U_{BB}$ and $F_{RF}$ by Algorithm 2.
   13. $k = k + 1$.
   14. $v_k = \|\sqrt{M}U_{opt} - F_{RF}U_{BB}\|_F^2$.
   15. **until** $v_{k-1} - v_k < \varepsilon$
   16. **Obtain** $F_{BB}^*$ by the classical digital precoding methods (14) with the effective baseband channel.

With DD-AHP and CU-AHP solved in Section IV-B, we propose the MKM-based AHP scheme in Algorithm 3 to solve CP-AHP. With a randomly initialized $U_{BB}$, the clustering part is performed from step 7 to step 12. Due to the constraints in CP-AHP, it is worth mentioning that greedy clustering each sample into the nearest cluster is not advisable, since it will lead to inconsistent numbers of members in different clusters. For the fairness among RF chains, which is similarly considered in [29], [31], we assign members to different clusters in turn from step 8 to step 11 as an inner loop. And the inner loop will be repeated $M$ times for the complete design. Then, the clustering centers are updated based on Algorithm 2 for further iterations, which will be terminated when the difference between the objective function values in adjacent loops reaches a certain threshold. At the end of the scheme, the baseband precoder is further refined by the classical digital precoding methods. Actually, similar to the proposed FHP scheme, the LS method is also operable for baseband precoder design in AHP. However, since the degree of freedom for RF precoder design significantly decreases in AHP, it is inadvisable to approximate the column space of $F_{RF}$ to that of $F_{opt}$ with the LS method. Thus, the classical precoding method is proposed. Algorithm 3 is convergent with the overall complexity $o(K_{out} K_{iter} (K N_{RF}^2 + K^3))$, where $K_{out}$ denotes the number of the outer iterations from step 4 to step 16$^3$. Please refer to Appendix C for the analysis of the convergence and complexity.

In the proposed MKM-based AHP scheme, the optimal digital precoding is decomposed into $N_t$ components, with each antenna corresponding to each of them. Based on the relevance among antennas, the similar components are merged into one cluster, which implies the similar antennas will be supported by the same RF chain.

Remark: The HAC method is not advisable to solve CP-AHP. Since the key approach of HAC is merging clusters until the desired number of clusters is obtained, it is of high possibility to merge two large clusters in the last few loops, which makes the clustering result improbable to satisfy the constraints in CP-AHP.

V. SIMULATION

In this section, we present the numerical results based on Monte Carlo simulations to evaluate the performance of proposed HAC-based FHP and MKM-based AHP schemes. In addition, the proposed SHP design is provided by the AO-based algorithm in Section IV-B for more details of performance comparison. In the simulation system, the BS is equipped with $8 \times 8$ antenna array to serve 8 users, while each user is equipped with $2 \times 2$ antenna array. The antenna elements are separated by $d = \lambda/2$ in the UPA structure. The mmWave channel between the BS and each user consists of $N_c = 5$ scattering clusters, each of which contains $N_{p,k} = 10$ propagation paths. The azimuth and elevation angles of arrival and departure are uniformly distributed in $[0, 2\pi)$ with a 10-degree angular spread. All simulation results are calculated over 1000 channel realizations.

A. System Spectral Efficiency

In this subsection, we investigate the precoding performance of different hybrid structures in terms of the system spectral efficiency.

$^3$The average number of outer iterations for convergence in numerical simulation of Section V is approximately $K_{out} = 6$. 
To illustrate the superiority of the proposed schemes, in Fig. 3, the FHP scheme in [16], the compressive-sensing-based (CS-based) scheme and the decoupling scheme for AHP in [32] and [29], the alternating-minimization-based (AM-based) scheme and the decoupling scheme for SHP in [34] and [25] are presented as benchmarks. The ZF method is utilized to provide the optimal full digital precoder, and the baseband precoder revision scheme for the last step of each proposed method. In addition, the non-quantized APSs are adopted at both combiners and precoders. To enable the existing schemes, we consider the case that the number of RF chains is equal to that of the users, i.e., $N_{RF} = K$. As shown in Fig. 3, FHP always provides the closest performance to the full digital precoding. The proposed HAC-based scheme shows a slight advantage compared with the scheme in [16]. For AHP, the proposed MKM-based scheme achieves higher spectral efficiency than the scheme in [32]. This is mainly because the proposed scheme determines the connection state between RF chains and antennas based on the relevance among antennas, which is not considered in the CS-based scheme. For SHP, the proposed AO-based scheme outperforms the scheme in [34], which is owing to the less constrained digital precoder. Since the decoupling scheme only harvests the channel gain in the analog precoder design, the schemes in [29] and [25] cause large performance loss for AHP and SHP, respectively.

Fig. 4 shows the impact of the quantized APSs on the system spectral efficiency. All three hybrid structures designed by the proposed schemes are considered for the comparison when $\text{SNR} = -5 \text{dB}$. By increasing the quantization bits of APSs, the performance of quantized schemes rapidly approaches that of non-quantized schemes owing to the improvement of phase resolution in the RF precoder. When $Q = 4$, the curves of quantized schemes almost coincide with that of corresponding non-quantized schemes with the gap at about 0.1 bps/Hz, which provides guidance for practical APS design. Moreover, since AHP and SHP share the same number of nonzero entries in the RF precoding matrix ($N_i$ in total), the impacts of few quantization bits are at almost the same level. However, the fully-connected structure with few quantization bits suffers more serious performance loss, since all entries in the RF precoding matrix are non-zero ($N_i N_{RF}$ in total). Fig. 5 shows the impact of the array scale at the BS with the comparison $N_i = \{8 \times 8, 8 \times 32\}$. The proposed MKM-based AHP scheme is adopted to provide the precoder design. Obviously, growing spectral efficiency can be observed with the enlargement of the antenna array. In addition, Fig. 5 shows the impact of the target full digital precoder provided by different digital precoding methods including MF, ZF, and RZF. For consistency, at the end of the algorithm, the baseband precoder is revised by the same method as getting the input. Compared with the ZF scheme, the RZF scheme shows a slight advantage in low SNR conditions, owing to the consideration of the noise parameter. The increasing SNR weakens the influence of the noise parameter in (14) and makes the performance gap disappear. The MF scheme provides the poorest performance at high SNR conditions, since it only harvests the channel gain but is incapable to eliminate the inter-user interference.
To clarify the performance of the proposed HAC-based FHP scheme with insufficient RF chains, in Fig. 6, we present the spectral efficiency versus the number of RF chains when \( \text{SNR} = -10 \text{ dB} \). As proposed in Section III, we adopt the LS and ZF schemes to refine the baseband precoder and analyze the performance difference, respectively. The FHP scheme in [16] and the upper bounds in (25) are also presented as the benchmarks. It is observed that the spectral efficiencies of the proposed scheme with the LS and ZF refinement schemes gradually get close to the corresponding upper bounds, which further verifies the tightness especially in the cases of large \( N_{\text{RF}} \). However, the spectral efficiency of the scheme in [16] maintains at a constant level. The defect results from the limitation that only \( K \) RF chains are effectively used, while other \( N_{\text{RF}} - K \) RF chains keep silent and make no difference to the system performance.

A more important conclusion from Fig. 6 is that the spectral efficiency of the LS scheme suffers a much sharper decline with the decreasing \( N_{\text{RF}} \) than the ZF scheme. This is because the insufficient RF chains lead to the low similarity of column spaces between the proposed RF precoder and the full digital precoder, which reduces the approximation accuracy of the LS method and exacerbates the inter-user interference. As for the ZF scheme, the inter-user interference can always be eliminated, which abates the influence of column space similarity to a certain extent. Further, a threshold of \( N_{\text{RF}} \) for the selection of two alternative refinement schemes can be given based on the intersection point of two performance curves. In the given simulation, the ZF method shall be adopted when \( N_{\text{RF}} \leq 10 \). Otherwise, it shall be the LS method.

The cases of different multi-user systems and multi-path channels are considered in Fig. 7(a) and Fig. 7(b), respectively. Simulation parameters are announced in legends. As a typical joint optimization scheme of FHP, the orthogonal matching pursuit (OMP) algorithm in [8] is presented as the benchmark. In Fig. 7(a), the proposed scheme only shows a slight advantage in the 2-user systems. However, the performance gap dramatically increases when more users are involved. Fig. 7(b) shows that the richness of the multiple paths in the mmWave channel affects the precoding performance to some extent for both HAC-based and OMP-based schemes. The proposed scheme still provides better performance than the OMP-based scheme in different multi-path cases. The superiority of the proposed scheme is mainly owing to the less constrained analog precoder which is designed free from the codebook.

### B. System Power Efficiency

In the previous simulation, FHP provides the highest system spectral efficiency. However, the fully-connected structure also leads to the highest power consumption. In this subsection, the power consumption is further taken into account for the comparison of different precoding schemes.

The power consumption of FHP can be given by [34]

\[
P_{\text{full}} = P_{\text{com}} + N_{\text{RF}}P_{\text{RF}} + N_{1}P_{\text{PA}} + N_{\text{RF}}N_{1}P_{\text{APS}},
\]

where \( P_{\text{com}} \) is the common power of the transmitter, \( P_{\text{RF}}, P_{\text{PA}} \), and \( P_{\text{APS}} \) are the power of a single RF chain, power amplifier, and APS, respectively. For SHP, which only requires \( N_{i} \) APSs, the power consumption can be given by

\[
P_{\text{sub}} = P_{\text{com}} + N_{\text{RF}}P_{\text{RF}} + N_{1}P_{\text{PA}} + N_{1}P_{\text{APS}}.
\]

As for AHP, it further requires \( N_{i} \) switches to implement the adaptive connection network, and incurs power consumption given by

\[
P_{\text{adv}} = P_{\text{com}} + N_{\text{RF}}P_{\text{RF}} + N_{1}P_{\text{PA}} + N_{1}(P_{\text{APS}} + P_{\text{SW}}),
\]

where \( P_{\text{SW}} \) is the power of a single switch. The power efficiency of a certain precoding design can be defined as the ratio \( \eta = R/P \).

In Fig. 8, we present the power efficiency versus the number of RF chains for three hybrid precoding structures with the power parameters set as \( P_{\text{com}} = 10 \text{ W}, P_{\text{RF}} = 100 \text{ mW}, P_{\text{PA}} = 100 \text{ mW}, P_{\text{APS}} = 20 \text{ mW}, P_{\text{SW}} = 10 \text{ mW} \) [34], [38]. For the proposed FHP scheme with the optimal baseband precoder refinement, although the spectral efficiency sufficiently approaches that of the optimal full digital precoder with \( N_{\text{RF}} \) increasing from 8 to 16 in Fig. 6, the performance growth
rate does not exceed 20%. However, the power consumption increases by 40%, which results in a negative impact on the power efficiency. Thus, a downward trend can be observed in Fig. 8. If the RF chains are over-equipped, i.e., $N_{RF} > 2K$, the spectral efficiency will be limited by the performance of the full digital precoder since the extra RF chains cannot be efficiently utilized. Consequently, the FHP scheme suffers a dramatic decrease in the power efficiency with the increasing RF chains.

For both the proposed AHP and SHP schemes, in Fig. 8, the curves of power efficiency monotonically increase with $N_{RF}$. Increasing spectral efficiency is one of the reasons for the improvement of power efficiency. More importantly, the almost unchanged power consumption gives rise to the positive impact, since the increasing RF chains will not result in any further requirements of APSs. In the cases of $N_{RF} < N_t$, the AHP scheme achieves much higher power efficiency than the SHP scheme. This is mainly because the adaptive connection network significantly increases the degree of freedom for the RF precoder design with several switchers which require low power consumption. At the special point $N_{RF} = N_t$ (extremely impractical in massive MIMO systems), the SHP scheme slightly outperforms the AHP scheme. This results from that the RF precoders in AHP and SHP are in the same form when $N_{RF} = N_t$, which degenerates the adaptively-connected structure into the sub-connected structure and makes the extra switches unable to provide any flexibility.

C. Initialization Scheme Analysis for AO-Based Algorithm

In Section IV-B, we propose a specific initialization scheme for the AO-based algorithm. In this subsection, we will show the impact of the initialization scheme in terms of the spectral efficiency and computational complexity.

Firstly, we compare the proposed AHP and SHP schemes with random and specific initialization schemes in terms of the spectral efficiency. The random initialization scheme means the phases in APSs are randomly initialized, which is generally adopted in the classical AO algorithm. The specific initialization scheme refers to the proposed initialization scheme in (41). Fig. 9(a) clarifies a common conclusion for AHP and SHP that the different initialization schemes contribute to almost the same average spectral efficiency. For more details, Fig. 9(b) plots the probability density function (PDF) curves of the spectral efficiency for the specific point $SNR = -5\, dB$ in Fig. 9(a). The random and specific initialization schemes provide almost the same regularity of distribution.

To compare the computational complexity of the algorithms with two alternative initialization schemes, in Fig. 10(a), we count the average number of iterations $K_{iter}$ in the AO-based algorithm. For SHP, the proposed specific initialization scheme requires 33% fewer iterations for convergence than the random initialization scheme. For AHP, the rate of reduction even reaches 50%. Further, in Fig. 10(b), we present the spectral efficiency of the AO-based SHP scheme versus the number of iterations. It demonstrates that the specific initialization scheme contributes to less convergence time than the random initialization scheme.

Remark: Although FHP achieves the closest performance to full digital precoding, it requires a complex implementation scheme in practical systems. From the perspective of power efficiency of FHP, it is not wise to equip many RF chains at the BSs, which means few users can be served by a single BS. In
addition, since the complexity of the HAC-based algorithm is related to \( N_t \), it raises a high demand for the computing capability in massive MIMO systems. Thus, the FHP scheme is suitable for the cases where spectral efficiency is extremely desirable. The sub-connected structure has the simplest connection between RF chains and antennas, which leads to performance loss, but also reduces power consumption especially for large \( N_{RF} \) cases. The adaptively-connected structure utilizes low-cost switches to improve the flexibility for precoding design, and achieves the best power efficiency performance. The complexity of the proposed AHP and SHP schemes is independent of \( N_t \), where the number of iterations is further reduced with the specific initialization scheme. Thus, it is more possible to adopt the AHP or SHP scheme at the BSs which cannot afford the complex structure and heavy computing tasks.

VI. CONCLUSION

In this paper, we have proposed hybrid precoding schemes for fully-connected and adaptively-connected structures in multi-user massive MIMO systems. Innovatively, the thought of clustering in unsupervised learning has been adopted. The hierarchical-agglomerative-clustering-based (HAC-based) scheme has been proposed for the fully-connected structure to explore the relevance among RF chains. The modified-K-means-based (MKM-based) scheme has been proposed to make full use of the flexibility in the adaptively-connected structure based on the relevance among antennas. Particularly, the alternating-optimum-based (AO-based) center updating algorithm in the MKM-based scheme can provide a feasible sub-connected structure design.

Simulation results have illustrated that the HAC-based scheme achieves close spectral efficiency to the full digital precoder with insufficient RF chains. The MKM-based scheme provides high power efficiency with the full use of the low-cost adaptive connection network. For the sub-connected structure, the AO-based scheme effectively enhances both spectral and power efficiency with the lowest power consumption. All proposed schemes outperform the existing work for corresponding hybrid structure designs. Moreover, the specific initialization scheme for the AO-based algorithm dramatically reduces the convergence time of proposed schemes.

Finally, our work has proved the feasibility of the clustering method in hybrid precoding design. However, only two representative clustering algorithms (HAC and K-means) have been considered in this paper. Actually, there is a great possibility for other clustering algorithms to achieve better performance, which will require further investigation. In addition, it is also worth to consider the practical cases that the perfect CSI is no longer available at the BS.

APPENDIX A

PROOF OF THE PROPOSITION 1

With the adjusted RF precoding matrix denoted by \( \bar{F}_{RF} \in \mathbb{C}^{N_t \times N_{RF}} \), the unchanged column space implies that the columns in \( \bar{F}_{RF} \) can be linear represented by the columns in \( F_{RF} \), i.e., \( \bar{F}_{RF} = F_{RF}A \), where \( A \in \mathbb{C}^{N_{RF} \times N_{RF}} \) is an arbitrary matrix.

Further, the basis vectors of the column space of \( F_{RF} \) and \( \bar{F}_{RF} \) can be respectively given by the columns of \( U_{RF} \in \mathbb{C}^{N_r \times r} \) and \( U_{RF} \in \mathbb{C}^{N_r \times r} \), where \( U_{RF} = F_{RF} = F_{RF} \), \( \bar{U}_{RF} = \bar{F}_{RF} \), and \( r = r(F_{RF}) = r(\bar{F}_{RF}) \). And the standard bases can be represented by each other with unitary transformation \( \bar{U}_{RF} = U_{RF}\Psi \), where \( \Psi \in \mathbb{C}^{r \times r} \) is a unitary matrix.

Accordingly, we have the following derivation
\[
\begin{align*}
F_{RF}A &= \bar{F}_{RF} \\
U_{RF}F_{RF}V_{RF}H &= \bar{U}_{RF}F_{RF}V_{RF}H \\
U_{RF}F_{RF}V_{RF}H &= U_{RF}\Psi F_{RF}V_{RF}H \\
\bar{U}_{RF}F_{RF}V_{RF}H &= \bar{F}_{RF} \Psi F_{RF}V_{RF}H.
\end{align*}
\]
Note that \( P = \Sigma F_{RF}H_{RF} \) has full row rank and satisfies \( PP^\dagger = I \), we adjust the baseband precoding matrix as \( F_{BB} = P^\dagger \Sigma F_{RF}V_{RF}F_{BB} \). Thus, the adjusted hybrid precoding matrix satisfies
\[
\begin{align*}
F_{RF}F_{BB} &= F_{RF}AP_{opt}^\dagger \Sigma F_{RF}V_{RF}F_{BB} \\
&= U_{RF}F_{RF}V_{RF}AP_{opt}^\dagger \Sigma F_{RF}V_{RF}F_{BB} \\
&= U_{RF}P^\dagger \Sigma F_{RF}V_{RF}F_{BB} \\
&= F_{RF}F_{BB},
\end{align*}
\]
which means no performance difference.

APPENDIX B

PROOF OF THE LEMMA 1

According to Proposition 1, the ideal case for the HAC-based FHP scheme is that \( F_{RF} \) holds the same column space as \( F_{opt} \), i.e., \( U_{RF} = U_{opt} \Theta \), where \( U_{RF} \in \mathbb{C}^{N_r \times r} \) and \( U_{opt} \in \mathbb{C}^{N_r \times r} \) denote the left-singular matrices of \( F_{RF} \) and \( F_{opt} \), \( r = r(F_{RF}) = r(F_{opt}) \). \( \Theta \in \mathbb{C}^{r \times r} \) is a unitary matrix.

With \( F_{BB} = F_{RF}F_{opt} \) obtained by the LS refinement scheme, we have the following derivation
\[
\begin{align*}
F_{RF}F_{BB} &= F_{RF}F_{RF}F_{opt} \\
&= U_{RF}(U_{RF})^H U_{opt} \Sigma_{opt} V_{opt}^H \\
&= U_{opt} \Theta(U_{opt} \Theta)^H U_{opt} \Sigma_{opt} V_{opt}^H \\
&= F_{opt} = \sqrt{K/\|H_{eq}\|_F^2} H_{eq} x + n,
\end{align*}
\]
Hence, the overall received signal in (3) can be given by
\[
r = \sqrt{K/\|H_{eq}\|_F^2} H_{eq} x + n,
\]
where \( r = [r_1, \ldots, r_K]^T \), \( n = [w_1^T n_1, \ldots, w_K^T n_K]^T \). Since the inter-user interference is completely eliminated in (48), the SINR of the \( k \)-th user is \( SINR_k = P/\sigma^2 \|H_{eq}\|_F^2 \). Thus, the system spectral efficiency is bounded by
\[
R_{LS} \leq K \log_2 \left( 1 + \frac{P}{\sigma^2 \|H_{eq}\|_F^2} \right).
\]

With \( F_{BB} = H_{eq}^\dagger = (H_{eq} F_{RF})^\dagger \) obtained by the ZF refinement scheme, the received signal can be given by
\[
r = \sqrt{K/\|F_{RF}H_{BB}\|_F^2} H_{eq} F_{RF} H_{BB} x + n,
\]
which also eliminates the inter-user interference and further bounds the system spectral efficiency as follows

\[ R_{ZF} \leq K \log_2 \left( 1 + \frac{P}{\sigma^2 \|F_{RF}H_{BB}^*\|_F^2} \right). \quad (51) \]

Since the row space of \( H_{eq} \) is consistent with the column space of \( F_{opt} \), the standard basis vectors of the row space of \( H_{eq} \) can be linear represented by the columns of \( F_{RF} \) with \( T \in \mathbb{C}^{N_{RF} \times r} \), i.e.,

\[
F_{RF} T = V_{eq},
\]

\[
V_{eq}^H F_{RF} T = I_r,
\]

where \( U_{eq} \Sigma_{eq} V_{eq}^H = H_{eq} \) and \( \Sigma_{eq} \in \mathbb{C}^{r \times r} \) is a diagonal matrix with \( r \) nonzero singular values of \( H_{eq} \) on the diagonal. According to (52), since \( N_{RF} \geq r \), the matrix \( V_{eq}^H F_{RF} \) has full row rank and satisfies \( V_{eq}^H F_{RF} = 1 \).

Further, we derive as follows

\[
\|H_{eq}\|_F^2 = \|\Sigma_{eq}^1 U_{eq}\|_F^2
\]

\[
\leq \|V_{eq}^H F_{RF}(V_{eq}^H F_{RF})^\dagger \Sigma_{eq}^{-1} U_{eq}\|_F^2
\]

\[
= \|F_{RF}(V_{eq}^H F_{RF})^\dagger \Sigma_{eq}^{-1} U_{eq}\|_F^2
\]

\[
\leq \|F_{RF}(H_{eq} V_{eq}^H F_{RF})\|_F^2 = \|F_{RF} H_{BB}^*\|_F^2,
\]

(53)

where (a) comes to equality if and only if \( V_{eq} \) fits the left-singular matrix of \( F_{RF} H_{BB}^* \), and (b) results from the product property of pseudo inverse. According to (53), \( R_{ZF} \leq R_{LS} \) is always satisfied.

**APPENDIX C**

**CONVERGENCE AND COMPLEXITY OF THE PROPOSED ALGORITHMS**

Algorithm 1 is obviously convergent, since two clusters are merged in each iteration until \( N_{RF} \) clusters are obtained. The computational complexity of Algorithm 1 is mainly caused by step 6 and step 13. In step 6, we need to calculate \( K(2K - 1) \) inner products with the complexity of \( o(N_i) \) for each, which totals \( o(K^2N_i) \). In step 13, we need to calculate SVDs for \( N_{RF} \) matrices in the size of \( N_i \times K \), which totals \( o(N_{RF}(K^2N_i^2 + K^3)) \). Since \( N_i \gg K \) in massive MIMO systems, the overall complexity is \( o(N_{RF} K^2N_i^2) \).

To proof the convergence of Algorithm 2, we first clarify that the solutions of (36) and (38) are both optimized to minimize the value of the objective function in (34). Since \( \|A\|_F^2 = \sum_i \|A(m, ;)\|_F^2 \), (36) can be equivalently written as

\[
\min_{\tilde{f}_n(m)} \quad \frac{N_{RF}}{M} \sum_{n=1}^{N_{RF}} \sum_{m=1}^{M} \|\sqrt{M} \tilde{U}_{opt}(t,:) - \tilde{f}_n(m) U_{BB}(n,:)\|_F^2
\]

s.t. \( \tilde{f}_n(m) = 1 \),

(54)

where the nonzero entries \( \tilde{f}_n(m) \) can be separately optimized. Thus, (36) can be equivalently decomposed into \( N_i \) subproblems (37). The optimal solution of (37) is in step 7 of Algorithm 2.

As for (38), since \( F_{RF}^H F_{RF} = M I_{N_{RF}} \) is still satisfied, (38) can be equivalently written as

\[
\min_\{U_{BB}\} \|\sqrt{M} F_{RF}^H U_{opt} - U_{BB}\|_F^2
\]

s.t. \( U_{BB}^H U_{BB} = I_K \),

(55)

which is a typical semi-orthogonal procrustes problem with the optimal solution given by (39). Thus, the objective function in (34) successively decreases in step 6 and step 7 of Algorithm 2. Note that the value of objective function is further lower bounded by zero. Algorithm 2 shall converge to a feasible solution. The computational complexity of Algorithm 2 is mainly caused by the calculation of the SVD in step 5, where the dimension of the matrix is \( N_{RF} \times K \). Assuming that the number of iterations from step 4 to step 10 is \( K_{iter} \), the overall complexity can be given by \( o(K_{iter}(K^2N_{RF}^2 + K^3)) \).

The convergence of Algorithm 3 can be similarly clarified as Algorithm 2, since it refines the objective function value in the iterative refinement of the clustering and the center updating parts [36]. In the clustering part, each sample is assigned to its nearest cluster in step 9. In the center updating part, Algorithm 2 optimizes \( U_{BB} \) and \( F_{RF} \) for a lower value of the objective function (30) in step 13. The computational complexity of Algorithm 3 is mainly caused by the repetition of the Algorithm 2. With \( K_{out} \) outer iterations from step 4 to step 16, the overall complexity can be similarly given by \( o(K_{out}K_{iter}(K^2N_{RF}^2 + K^3)) \).

**REFERENCES**

[1] 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.

[2] D. Hui, S. Sandberg, Y. Blankenship, M. Andersson, and L. Hanzo, “Channel coding in 5G new radio: A tutorial overview and performance comparison with 4G LTE,” IEEE Veh. Technol. Mag., vol. 13, no. 4, pp. 60–69, Dec. 2018.

[3] J. Dai, K. Niu, Z. Si, C. Dong, and J. Lin, “Polar-coded non-orthogonal multiple access,” IEEE Trans. Signal Process., vol. 66, no. 5, pp. 1374–1389, May 2018.

[4] T. S. Rappaport et al., “Millimeter wave mobile communications for 5G cellular: It will work!” IEEE Access, vol. 1, pp. 335–349, 2013.

[5] "User equipment radio transmission and reception; Part 1: Range 1 standalone (release 15)," 3GPP, Sophia Antipolis, France, 3GPP TS 38.101-1, Dec. 2017.

[6] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link budget," IEEE Comm. Surv. Tuts., vol. 20, no. 2, pp. 870–913, Apr.–Jun. 2018.

[7] Y. Kim et al., “Feasibility of mobile cellular communications at millimeter wave frequency,” IEEE J. Sel. Topics Signal Process., vol. 10, no. 3, pp. 589–599, Apr. 2016.

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

[9] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 742–758, Oct. 2014.

[10] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "Channel inversion and regularization," IEEE Trans. Wireless Commun., vol. 53, no. 1, pp. 195–202, Jan. 2005.

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

[12] A. F. Molisch et al., "Hybrid beamforming for massive MIMO: A survey," IEEE Commun. Mag., vol. 55, no. 9, pp. 134–141, Sep. 2017.
[13] Y. Lee, C. Wang, and Y. Huang, “A hybrid RF/baseband precoding processor based on parallel-index-selection matrix-inversion-bypass simultaneous orthogonal matching pursuit for millimeter wave MIMO systems,” *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 305–317, Jan. 2015.

[14] J. Kim and Y. H. Lee, “MSE-based hybrid RF/baseband processing for millimeter-wave communication systems in MIMO interference channels,” *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2714–2720, Jun. 2015.

[15] 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.

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

[17] D. H. N. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, “Hybrid MMSE precoding and combining designs for mmWave multiuser systems,” *IEEE Access*, vol. 5, pp. 19167–19181, 2017.

[18] Y. Ahn, T. Kim, and C. Lee, “A beam steering based hybrid precoding for MU-MIMO mmWave systems,” *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2726–2729, Dec. 2017.

[19] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, “Hybrid MMSE precoding for mmWave multiuser MIMO systems,” in *Proc. IEEE Int. Commun. Conf.*, May 2016, pp. 1–6.

[20] O. Aluluabi, Q. Z. Ahmed, J. Wang, and H. Zhu, “Hybrid digital-to-analog precoding design for mmWave systems,” in *Proc. IEEE Int. Commun. Conf.*, May 2017, pp. 1–6.

[21] N. Li, Z. Wei, H. Yang, X. Zhang, and D. Yang, “Hybrid precoding for mmWave massive MIMO systems with partially connected structure,” *IEEE Access*, vol. 5, pp. 15142–15151, 2017.

[22] C. Kim, T. Kim, and J. Seol, “Multi-beam transmission diversity with hybrid beamforming for MIMO-OFDM systems,” in *Proc. IEEE Globecom Workshops*, Dec. 2013, pp. 61–65.

[23] J. Singh and S. Ramakrishna, “On the feasibility of codebook-based beamforming in millimeter wave systems with multiple antenna arrays,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2670–2683, May 2015.

[24] A. Li and C. Masouros, “Hybrid analog-digital millimeter-wave MU-MIMO transmission with virtual path selection,” *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 438–441, Feb. 2017.

[25] A. Li and C. Masouros, “Hybrid precoding and combining design for millimeter-wave multi-user MIMO based on SVD,” in *Proc. IEEE Int. Commun. Conf.*, May 2017, pp. 1–6.

[26] Y. Long, Z. Chen, J. Fang, and C. Tellambura, “Data-driven-based analog beam selection for hybrid beamforming under mm-Wave channels,” *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 340–352, May 2018.

[27] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, “Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems,” in *Proc. IEEE Int. Commun. Conf.*, May 2017, pp. 1–6.

[28] S. Park, A. Alkhateeb, and R. W. Heath, “Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2907–2920, May 2017.

[29] X. Zhu, Z. Wang, L. Dai, and Q. Wang, “Adaptive hybrid precoding for multi-user massive MIMO,” *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 776–779, Apr. 2016.

[30] Q. Yu, X. Zhai, and M. Zhao, “An energy-efficient hybrid precoding algorithm for multiuser mmWave massive MIMO systems,” in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.

[31] L. Zhang, L. Gui, Q. Qin, and Y. Tu, “Adaptively-connected structure for hybrid precoding in multi-user massive MIMO systems,” in *Proc. IEEE 29th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun.*, Sep. 2018, pp. 1–7.

[32] Q. Qin, L. Gui, L. Zhang, and Y. Tu, “Compressive sensing based hybrid beamforming for adaptively-connected structure,” in *Proc. IEEE/CIC Int. Conf. Commun. China*, Aug. 2018, pp. 309–314.

[33] S. Zhou, Z. Xu, and F. Liu, “Method for determining the optimal number of clusters based on agglomerative hierarchical clustering,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 3007–3017, Dec. 2017.

[34] X. Yu, J. Shen, J. Zhang, and K. B. Letaief, “Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[35] S. J. Lee, “Matrix approximation based design of a combined analog and digital beamformer in frequency-selective fading channels,” *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 325–328, Feb. 2014.

[36] A. K. Jain, “Data clustering: 50 years beyond k-means,” in *Machine Learning and Knowledge Discovery in Databases*, current neumans, B. Goethals, and K. Morik, Eds. Berlin, Germany: Springer, 2008, pp. 3–4.

[37] J. C. Gower et al., *Procrustes Problems*, vol. 30. London, U.K.: Oxford Univ. Press, 2004.

[38] R. Mindez-Rial, C. Rusu, A. Alkhateeb, N. Gonzalez-Prelcic, and R. W. Heath, “Channel estimation and hybrid combining for mmWave: Phase shifters or switches?” in *Proc. Inf. Theory Appl. Workshop*, Feb. 2015, pp. 90–97.

[39] M. Shehata, M. Crussire, M. Hlard, and P. Pajusco, “Leakage based users selection for hybrid beamforming in millimeterwave MIMO,” in *Proc. IEEE 29th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun.*, Sep. 2018, pp. 1144–1150.

Ling Zhang received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. He is currently working toward the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include millimeter wave communications, hybrid precoding, and interference management.

Lin Gui (M’08) received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2002. Since 2002, she has been with the Institute of Wireless Communication Technology, Shanghai Jiao Tong University, Shanghai, China, where she is currently a Professor. Her current research interests include HDTV and wireless communications.

Kai Ying (S’14–M’17) received the B.S. and M.E. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2010 and 2013, respectively, and the M.S. and Ph.D. degrees from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2013 and 2016, respectively. He is currently a Senior Researcher with the Sharp Laboratories of America, Camas, WA, USA. His research interests include nonlinear signal processing and ultra reliable low latency communications.

Qibo Qin received the B.S. degree in electronic and information engineering from Xian Jiao Tong University, Xian, China, in 2014. He is currently working toward the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include doubly selective channel estimation, massive multiple-input multiple-output systems and millimeter wave communications.