IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Near-Optimal Design for Hybrid Beamforming in MmWave Massive Multi-User MIMO Systems

YANG ZHANG[1], JIANHE DU[1], (Member, IEEE), YUANZHI CHEN[1], XINGWANG LI[2], (Senior Member, IEEE), KHALED M. RABIE[3], (Member, IEEE), and RUPAK KHAREL[4], (Senior Member, IEEE)

[1]School of Information and Communication Engineering, Communication University of China, Beijing 100024, China
[2]School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China
[3]Department of Engineering, Manchester Metropolitan University, Manchester U.K., M1 5GD.
[4]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester U.K., M15 6BH.

Corresponding author: Jianhe Du (e-mail: dujianhe1@gmail.com).

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0502001, in part by the National Natural Science Foundation of China under Grant 61601414, Grant 61701448, and Grant 61702466, in part by the Fundamental Research Funds for the Central Universities under Grant CUC200A011, Grant CUC19ZD001, and Grant CUC2019D012, in part by the Doctoral Scientific Funds of Henan Polytechnic University under Grant B2016-34, and in part by the Key Scientific Research Projects of Higher Education Institutions in Henan Province Grant 20A510007.

ABSTRACT Millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) systems can obtain sufficient beamforming gains to combat severe path loss in signal propagation. The hybrid (analog/digital) beamforming with multiple data streams can be utilized to further improve mmWave spectral efficiency. In this paper, we focus on the hybrid beamforming design of a downlink mmWave massive multi-user MIMO (MU-MIMO) system based on full-connected structure, and aim to maximize the sum rate of the overall system as an objective function. In the analog beamforming stage, a piecewise successive iterative approximation (PSIA) algorithm is proposed to design the analog beamformer and combiner. This algorithm not only has a linear property, but also can obtain closed-form solutions. In the digital beamforming stage, the piecewise successive approximation method is utilized to design the digital beamforming based on the criterion to avoid the loss of information, which can help reduce the computational complexity and is also implemented simply. The results show that the proposed scheme achieves good sum-rate performance in the mmWave massive MU-MIMO system, and outperforms the state-of-the art MIMO hybrid beamforming design schemes, even when the number of base station antennas is not very large.

INDEX TERMS Millimeter wave, massive MIMO, multi-user, hybrid beamforming, sum rate.

I. INTRODUCTION

THE developments of traditional wireless communication technologies seem to encounter bottleneck constraints due to the limited bandwidths. To meet the great requirements of wireless communication, millimeter wave (mmWave) communication has been considered as a key technology for the next-generation wireless communication systems [1], [2].

MmWave communication generally corresponds to 30–300 GHz tremendous frequency bands available to support Gigabits per second (Gb/s) data rate transmission. This has gained considerable attention in wireless fronthaul/backhaul, indoor, cellular (hotspot and small cell), device-to-device (D2D) communications, etc [3], [4]. However, the inherent shortcoming of the short wavelength of mmWave is that it has very different channel propagation characteristics such as the severe propagation path loss (PL), rain attenuation, high penetration loss, high delay resolution, high directivity, and human blockage [5]–[7]. Therefore, it is challenging for mmWave to implement in practice. Fortunately, the tremendously reduced wavelength enables the equipment to put the large-scale antenna arrays in a much smaller space. Thus, mmWave systems can integrate massive multiple-input multiple-output (MIMO) transceiver elements to provide a sufficiently powerful received signal, which can enhance the signal gain and spectral efficiency [8], [9]. In addition,

the high directivity can help the design of beamforming techniques to direct the signal in a certain direction, which overcomes the high PL problem and establishes reasonable signal-to-noise ratio (SNR) links [10].

Although the rationales of beamforming are the same regardless of the carrier frequency, signal processing in mmWave systems is subject to a set of important practical constraints. For example, traditional MIMO systems often perform digital linear beamforming at baseband, which enables controlling both the signal's phase and amplitude and supporting multi-stream multi-users communications [11], [12]. However, the baseband beamforming (the digital beamforming) requires not only dedicated baseband processor, but also radio frequency (RF) hardware and analog-to-digital converter (ADC) for each antenna element. When a large number of antennas is deployed, the high hardware cost, complexity, and power consumption of digital beamforming architecture become unaffordable to be implemented in practice. Therefore, it forces mmWave systems to rely heavily on the analog or RF beamforming processing [13]. Both beamforming and combining of analog beamforming processing are implemented by a network of analog phase shifters controlling the phase of the transmitted signal at each antenna element in the RF domain, which have been applied to mmWave wireless local area network (WLAN) systems to provide a simpler architecture [14], [15]. Compared to the digital beamforming processing, since the substantially reduced number of RF chains, there are implementation benefits in terms of lower hardware complexity and lower power consumption. However, the analog beamforming is subject to additional constraints, e.g., spatial multiplexing of the beams is impossible, an RF chain only shapes one beam in a cycle and is only applied to one data stream scenario, which cannot improve the spectral efficiency. Moreover, the phase shifters are controlled digitally and obtain only quantized phase values. Therefore, these constraints limit the development of analog beamforming. For a better tradeoff between the performance and costs, a hybrid beamforming approach that combines analog and digital beamforming is proposed for the mmWave massive MIMO systems. Since the hybrid beamforming achieves similar performance to the full-digital one with much lower power consumption and hardware complexity, it has attracted a great deal of research attention [16], [17].

The achievable sum rate and the mean squared error (MSE) are two important optimization objectives for the hybrid beamforming design problems [18], [19]. Since the former is an important performance evaluation standard in mmWave systems, the design aims to maximize the sum rate for hybrid beamforming in this paper. However, many challenges have arisen from maximizing data rates under the constraints derived from the hybrid architecture. Currently, this problem is solved by two methods in the existing research work. One is to jointly design the analog and digital beamformer/combiner and the other is a two-stage method, wherein the analog beamformer/combiner is designed separately from the digital beamformer/combiner.

The joint design methods are widely used for hybrid beamforming to approach full-digital performance for single-user MIMO (SU-MIMO) and multi-user MIMO (MU-MIMO) scenarios. For SU-MIMO systems, by fully exploiting the sparsity of the channel, a least square (LS) and an approach utilizing matching pursuit (MP) can decompose full-digital beamforming into a separate analog and digital beamforming for mmWave channels [20], [21]. Based on employing an iterative algorithm and approaching the non-convex optimization with a convex problem, the method aiming at full-digital single-user solutions can be utilized to jointly design an analog and digital precoder/combiner [22]. For MU-MIMO systems, a weighted sum mean square error (WSMSE) minimization approach is proposed to jointly design analog and digital beamforming, which aims to approximate the performance of the block diagonalization (BD) solution for full-digital beamforming [23]. An over-sampling codebook (OSC)-based hybrid minimum sum-mean-square-error (min-SMSE) precoding scheme for mmWave MU-MIMO systems to optimize the BER is proposed in [24].

The two-stage method is widely utilized for designing hybrid beamforming for MU-MIMO communications to approximate the capacity. Specifically, most MU schemes prefer harvesting energy based on the channel matrix in the analog stage, and further eliminating the inter-user interference based on the baseband beamformer which takes the influences of the channel matrix and the RF beamformer into account in the following digital stage. The widely used method is the zero forcing (ZF). For example, a low-complexity hybrid BD scheme is proposed to harvest the large array gain through the RF beamforming and combining, and then digital BD processing is performed by the generalized ZF in conjunction with an equal gain transmission (EGT) scheme [25]. A hybrid beamforming design method based on the Modified Generalized Low Rank Approximation of Matrices (MGLRAM) is proposed to perform the ZF combined with an iterative procedure [26]. Further, maximizing the sum rate of the equivalent baseband channel in the analog stage, followed by excluding inter-user interference in the digital stage, is investigated in [27] and [28]. The proposed hybrid regularized channel diagonalization (HRCD) scheme is utilized by simple non-iterative processing for digital beamforming, and EGT method for analog beamforming [29]. Furthermore, an approach based on the criterion of minimizing mean square error to maximize the signal-to-leakage-plus-noise ratio of each user in the digital stage is studied in [30]. Beside the above design criterion, the authors of [31] introduced a method based on leveraging BD technology to eliminate the inter-user interference in the analog stage and minimize the mean square error of each data stream to harvest energy in the digital stage. Note that there are other schemes that utilize the two-stage method to design SU-MIMO communications. For instance, a method that maximizes the single-user spectral rate by optimizing analog processing with fixed digital processing is proposed in [32]. A low-complexity

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

iterative matrix decomposition based hybrid beamforming (IMD-HBF) scheme is proposed to obtain the optimal analog and digital solutions [33]. Among the above two-stage design schemes, hybrid beamforming can realize the optimal full-digital beamforming if and only if the number of RF chains is twice the number of data streams, e.g., the method is presented and discussed in [28], but this approach is implemented at the expense of high power consumption and costs. Then, the method of leveraging the BD technology to eliminate the inter-user interference will reduce the system performance since the overlap of the row subspace of each user channel matrix becomes significant when the number of users is large. Although the method presented in [33] outperforms the BD technology, it has a high number of iterations, which enhances the computational complexity of the system.

In this current paper, we focus on the hybrid beamforming design of mmWave massive MU-MIMO system with full-connected structure, where the single BS equipped with a large antenna array is assumed to serve several multi-antenna multi-stream users. Employing the two-stage method, perfect channel state information (CSI) is derived to design the analog and digital beamformer/combiner. The main contributions of this paper can be summarized as follows.

- In the analog beamforming stage, we minimize the MSE between the equivalent analog transceiver signals to reduce the information loss of signals in the analog channel transmission, which can achieve the purpose of maximizing the mutual information of the analog transceiver signals. Meanwhile, a piecewise successive iterative approximation (PSIA) algorithm is proposed to design the analog beamforming in the internal iteration. The proposed design scheme can obtain the optimal saturation value with fewer iterations.
- In the digital beamforming stage, the piecewise successive approximation method is utilized to design the digital beamformer and combiner, which is based on the criterion to avoid the loss of information at each stage. The complexity of the proposed design method is lower than that of the baseband BD technology which is combined to design digital beamforming by the state-of-the art schemes such as EGT-BD [25], MGLRAM [26], HyEB [27], and HySBD [28], etc. In addition, the performance of the proposed design method is superior to the baseband BD technology in terms of eliminating the inner-user and inter-user interferences.
- Under the condition that the number of RF chains is the same as the number of data streams, the performance of the proposed hybrid beamforming system outperforms the state-of-the art for hybrid beamforming systems. Further, the proposed design scheme has higher and more stable power efficiency than the existing schemes. Even when the number of BS antennas is not very large, the proposed design also shows superior performance in terms of sum rate. In addition, the solutions obtained in this paper are closed-form ones.

The remainder of this paper is organized as follows. Section II briefly introduces the channel and system models of the mmWave massive MU-MIMO system. The original problem of the system design is formulated and discussed in Section III. Section IV presents and discusses the simulation results of the proposed hybrid beamforming design scheme. Finally, conclusions are drawn in Section V.

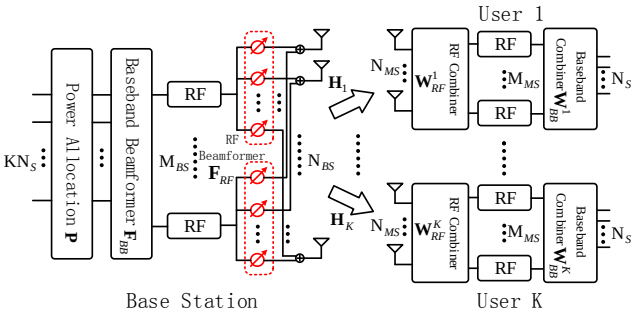

FIGURE 1. Block diagram of the mmWave massive MU-MIMO system with hybrid beamforming structure.

Notations: Bold upper-case and lower-case letters represent matrices and column vectors, respectively; $(\cdot)^{-1}$, $(\cdot)^T$, and $(\cdot)^H$ denote inversion, transpose, and conjugate transpose, respectively; The Frobenius norm of the matrix $\mathbf{A}$ and the 2-norm of the vector $\mathbf{a}$ are expressed as $\|\mathbf{A}\|_{\mathrm{F}}$ and $\|\mathbf{a}\|_2$, respectively. $\mathbf{A}(i,j)$, $\mathbf{A}(:,j)$, and $\mathbf{A}(i,:)$ respectively denote the (i,j)th complex element, jth column vector, and ith row vector of matrix $\mathbf{A}$, and $|\mathbf{A}(i,j)|$ is the amplitude; $A(i:j,:)$ and $A(:,i:j)$ represent the matrix consists of vectors from rows i to j and columns i to j of the matrix $\mathbf{A}$, respectively; $\mathbf{I}_N$ is the identity matrix of size $N \times N$; $\mathcal{CN}(0,\sigma^2)$ is the complex Gaussian distribution with mean 0 and the variance σ^2; $\angle\mathbf{A}$ denotes the operation of getting the angle of each entry in matrix $\mathbf{A}$; $\mathbb{D}^{l\times l}$ and $\mathbb{C}^{m\times n}$ describe a real diagonal matrix of dimension $l \times l$ and a complex matrix of dimension $m \times n$, respectively; $\mathrm{tr}\{\cdot\}$ and $\mathrm{Re}(\cdot)$ indicate the trace and real part taking operators, respectively; $\mathrm{vec}(\cdot)$ and $\mathrm{unvec}_{m\times n}(\cdot)$ are the vectorization and maxicization, respectively; $(\cdot)_{\ell,\ell}$ denotes the ℓth diagonal element of a matrix. Expectation operator is denoted by $\mathbb{E}[\cdot]$. The determinant and block diagonalization operation of a matrix are respectively expressed as $|\cdot|$ and $\mathrm{blk}(\cdot)$.

II. SYSTEM DESCRIPTION

In this section, we present the mmWave signal and channel model considered in this paper.

A. SYSTEM MODEL

Consider the downlink of the mmWave massive MU-MIMO system with full-connected subarray structure shown in Fig. 1 in which a BS serves K users simultaneously. The BS is equipped with a large number, N_{BS}, of antennas and M_{BS} RF chains, and each user is equipped with N_{MS} antennas and M_{MS} RF chains to support N_S data streams in a parallel mode. To enable multi-stream communication and reduce the complexity of the hardware, the number of RF chains

is constrained by $KN_S \leq M_{BS} \leq N_{BS}$ for the BS and $N_S \leq M_{MS} \leq N_{MS}$ for each user. As can be seen in Fig. 1, the transmitted symbol $\mathbf{s}$ with the total transmit power constraint P_t first passes through a diagonal power allocation matrix $\mathbf{P} \in \mathbb{D}^{KN_S \times KN_S}$ which distributes power to the transmitted symbol $\mathbf{s}_k$ of each user and satisfies $\|\mathbf{P}\|_{\mathrm{F}}^2 = P_t$. Then, the symbol $\mathbf{s}$ after power allocation is processed by digital beamforming using a baseband beamformer $\mathbf{F}_{BB} \in \mathbb{C}^{M_{BS} \times KN_S}$. After beamforming in the baseband domain, an RF beamformer $\mathbf{F}_{RF} \in \mathbb{C}^{N_{BS} \times M_{BS}}$ is applied for analog beamforming. Therefore, the discrete-time transmitted signal is finally represented as $\mathbf{x}_s = \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{P}\mathbf{s}$, where the transmit power of $\mathbf{s} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T]^T \in \mathbb{C}^{KN_S \times 1}$ is supposed to be normalized such that $\mathbb{E}\left[\mathbf{s}\mathbf{s}^H\right] = \frac{1}{KN_S}\mathbf{I}_{KN_S}$. Since $\mathbf{F}_{RF}$ is implemented by using analog phase shifters, its elements are constrained to satisfy $\left(\mathbf{F}_{RF}(:,i)\,\mathbf{F}_{RF}^H(:,i)\right)_{\ell,\ell} = N_{BS}^{-1}$, i.e., all elements of $\mathbf{F}_{RF}$ have the same amplitude. To guarantee the total transmitted power constraint, $\mathbf{F}_{BB}$ is normalized to satisfy $\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_{\mathrm{F}}^2 = KN_S$. Meanwhile, no other hardware-related constraints are placed on the baseband beamformer.

For simplicity, we consider a block fading channel model [34], e.g., the narrowband flat fading channel model, which yields a received signal of each user

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{P}\mathbf{s} + \mathbf{n}_k, \tag{1}$$

where $\mathbf{y}_k \in \mathbb{C}^{N_{MS} \times 1}$ is the kth received vector, $\mathbf{H}_k \in \mathbb{C}^{N_{MS} \times N_{BS}}$ is the channel matrix from BS to the kth user, and $\mathbf{n}_k \in \mathbb{C}^{N_{MS} \times 1}$ is the corresponding complex additive white Gaussian noise vector in which the elements follow the independent and identically distributed (i.i.d.) complex Gaussian distribution with zero mean and variance σ^2, i.e., $\mathbf{n}_k \sim \mathcal{CN}\left(0, \sigma^2\right)$. To enable beamforming, we assume that the CSI is known perfectly and instantaneously to both the BS and each user. In practical systems, CSI at the receiver can be obtained via training, then shared with the BS via limited feedback from the receiver to the BS [35].

At the receiver, each user employs its analog phase shifters and digital combiner to obtain the processed received signal

$$\hat{\mathbf{s}}_k = \mathbf{W}_{BB}^{k^H}\mathbf{W}_{RF}^{k^H}\mathbf{H}_k\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{P}\mathbf{s} + \mathbf{W}_{BB}^{k^H}\mathbf{W}_{RF}^{k^H}\mathbf{n}_k, \tag{2}$$

where $\mathbf{W}_{RF}^k \in \mathbb{C}^{N_{MS} \times M_{MS}}$ is the analog combining matrix and $\mathbf{W}_{BB}^k \in \mathbb{C}^{M_{MS} \times N_S}$ is the digital combining matrix for the kth user. Similar to analog beamforming, the analog combining is implemented by using phase shifters. Therefore, $\mathbf{W}_{RF}^k$ also satisfies the constant amplitude constraint $\left(\mathbf{W}_{RF}^k(:,i)\left(\mathbf{W}_{RF}^k(:,i)\right)^H\right)_{\ell,\ell} = N_{MS}^{-1}$. When Gaussian symbols are transmitted over the mmWave channel, the achievable sum rate is given by [36]

$$R = \log_2\left(\left|\mathbf{I}_{KN_S} + \frac{1}{KN_S}\mathbf{R}_n^{-1}\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H\right|\right), \tag{3}$$

where $\tilde{\mathbf{H}} = \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{P}$, and $\mathbf{R}_n = \sigma^2\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{W}_{RF}\mathbf{W}_{BB}$ is the noise covariance matrix after combining. $\mathbf{W}_{RF} = \mathrm{blk}\left(\mathbf{W}_{RF}^1, \ldots, \mathbf{W}_{RF}^K\right)$ and $\mathbf{W}_{BB} = \mathrm{blk}\left(\mathbf{W}_{BB}^1, \ldots, \mathbf{W}_{BB}^K\right)$ are the analog and digital combiners of K users, respectively. $\mathbf{H} = \left[\mathbf{H}_1^H, \ldots, \mathbf{H}_K^H\right]^H$ is the total channel matrix. Defining the baseband channel of the kth user as $\bar{\mathbf{H}}_k = \mathbf{W}_{RF}^{k^H}\mathbf{H}_k\mathbf{F}_{RF}$, the transmitted ith data stream of the kth user, $\hat{s}_{k_i}$, can be further expressed as

$$\begin{aligned}\hat{s}_{k_i} &= \mathbf{W}_{BB}^{k^H}(i,:)\bar{\mathbf{H}}_k\mathbf{F}_{BB}(:,k_i)\sqrt{P_{k_i}}s_{k_i} \\ &\quad + \sum_{j=1,j\neq i}^{N_S}\mathbf{W}_{BB}^{k^H}(i,:)\bar{\mathbf{H}}_k\mathbf{F}_{BB}(:,k_j)\sqrt{P_{k_j}}s_{k_j} \\ &\quad + \sum_{m=1,m\neq k}^{K}\sum_{l=1}^{N_S}\mathbf{W}_{BB}^{k^H}(i,:)\bar{\mathbf{H}}_k\mathbf{F}_{BB}(:,m_l)\sqrt{P_{m_l}}s_{m_l} \\ &\quad + \mathbf{W}_{BB}^{k^H}(i,:)\mathbf{W}_{RF}^{k^H}\mathbf{n}_k,\end{aligned} \tag{4}$$

where $k_i = (k-1)N_S + i$, s_{k_i} is the ith entry of the signal $\mathbf{s}_k$ of the kth user sent by BS, and $\sqrt{P_{k_i}}$ is the corresponding power allocation. The first term on the right side of (4) indicates the desired signal, and the other three terms represent inner-user interference, inter-user interference, and noise, respectively. Hence, the sum rate in (3) can be rewritten as

$$R = \sum_{k=1}^{K}\sum_{i=1}^{N_S}\log_2(1 + \mathrm{SINR}_{k_i}), \tag{5}$$

where SINR_{k_i} is the signal to interference to noise ratio (SINR) of the signal $\hat{s}_{k_i}$, which can be calculated by the ratio of the desired signal energy in (4) to the interference plus noise energy of the remaining terms. The SINR_{k_i} is formulated as

$$\begin{aligned}\mathrm{SINR}_{k_i} &= \frac{\tilde{S}_{k_i}}{\tilde{I}_{k_i} + \tilde{N}_{k_i}} \\ \tilde{S}_{k_i} &= \left|\mathbf{W}_{BB}^{k^H}(i,:)\bar{\mathbf{H}}_k\mathbf{F}_{BB}(:,k_i)\sqrt{P_{k_i}}\right|^2 \\ \tilde{I}_{k_i} &= \sum_{j=1,j\neq i}^{N_S}\left|\mathbf{W}_{BB}^{k^H}(i,:)\bar{\mathbf{H}}_k\mathbf{F}_{BB}(:,k_j)\sqrt{P_{k_j}}\right|^2 \\ &\quad + \sum_{m=1,m\neq k}^{K}\sum_{l=1}^{N_S}\left|\mathbf{W}_{BB}^{k^H}(i,:)\bar{\mathbf{H}}_k\mathbf{F}_{BB}(:,m_l)\sqrt{P_{m_l}}\right|^2 \\ \tilde{N}_{k_i} &= \sigma^2\left\|\mathbf{W}_{RF}^k\mathbf{W}_{BB}^k(:,i)\right\|_{\mathrm{F}}^2,\end{aligned} \tag{6}$$

where $k \in \{1,2,\ldots,K\}$, $i \in \{1,2,\ldots,N_S\}$.

B. CHANNEL MODEL

The mmWave propagation has limited spatial selectivity or scattering, which will lead to high free-space path loss. However, traditional MIMO channel models cannot accurately reflect this characteristic. Similarly, the massive tightly-packed antenna arrays adopted in mmWave transceivers lead to high levels of antenna correlation. If the statistical fading distribution in traditional MIMO analysis is used in the mmWave channel modeling, it becomes inaccurate [11]. Therefore, we adopt a narrow band channel model with uniform linear arrays (ULAs), such as the extended Saleh-Valenzuela model,

to obtain the mathematical structure of mmWave channel accurately [37].

We assume that the scattering channel has N_c scattering clusters, each of which consists of N_p propagation paths. Therefore, the discrete-time narrowband channel matrix of the kth user can be expressed as

$$\mathbf{H}_k = \gamma \sum_{i=1}^{N_c} \sum_{l=1}^{N_p} \alpha_{il}^k \Lambda_{MS}\left(\theta_{il}^k\right) \Lambda_{BS}(\phi_{il}^k) \mathbf{a}_{MS}(\theta_{il}^k) \mathbf{a}_{BS}(\phi_{il}^k)^H, \tag{7}$$

where $\gamma = \sqrt{\frac{N_{BS}N_{MS}}{N_c N_p}}$ is a normalization factor, and the channel matrix satisfies $\mathbb{E}\left[\|\mathbf{H}_k\|_{\mathrm{F}}^2\right] = N_{BS}N_{MS}$. α_{il}^k is the complex gain of the lth ray in the ith scattering cluster for the kth user, which follows the independent Gaussian distribution, i.e., $\alpha_{il}^k \sim \mathcal{CN}(0,1)$. θ_{il}^k and ϕ_{il}^k represent the azimuth angles of arrival/departure (AoAs/AoDs) of the lth ray in the ith scattering cluster for the kth user, which obey the truncated Laplacian distribution [25]. The functions $\Lambda_{MS}\left(\theta_{il}^k\right)$ and $\Lambda_{BS}(\phi_{il}^k)$ denote the transmit and receive antenna array gain at the corresponding angles of departure and arrival. Finally, $\mathbf{a}_{MS}(\theta_{il}^k)$ and $\mathbf{a}_{BS}(\phi_{il}^k)$ are the normalized antenna array response vectors at an azimuth angle of θ_{il}^k and ϕ_{il}^k, respectively. For the sake of simplicity but without loss of generality, we assume that when both the BS and each user adopt ULAs, the array response vector $\mathbf{a}_{MS}(\theta_{il}^k)$ and $\mathbf{a}_{BS}(\phi_{il}^k)$ can be presented as [37]

$$\mathbf{a}_{MS}(\theta_{il}^k) = \frac{1}{\sqrt{N_{MS}}}\left[1, e^{j\beta d\sin(\theta_{il}^k)}, \ldots, e^{j(N_{MS}-1)\beta d\sin(\theta_{il}^k)}\right]^T, \tag{8}$$

$$\mathbf{a}_{BS}(\phi_{il}^k) = \frac{1}{\sqrt{N_{BS}}}\left[1, e^{j\beta d\sin(\phi_{il}^k)}, \ldots, e^{j(N_{BS}-1)\beta d\sin(\phi_{il}^k)}\right]^T, \tag{9}$$

where $j = \sqrt{-1}$, $\beta = \frac{2\pi}{\lambda}$, λ is the carrier wavelength of the signal, and d is the inter-element spacing, e.g., $d = \frac{\lambda}{2}$.

III. MULTIUSER HYBRID BEAMFORMING DESIGN

This section discusses the hybrid beamforming design of the downlink mmWave massive MU-MIMO system with the full-connected structure. The design goal is to maximize the sum rate of the system expressed in (3), hence the optimization problem can be formulated as

$$\begin{aligned}(\mathbf{F}_{RF}, \mathbf{W}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{BB}, \mathbf{P}) = &\underset{(\mathbf{F}_{RF}, \mathbf{W}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{BB}, \mathbf{P})}{\arg\max} R \\ \text{s.t. } &\mathbf{F}_{RF}(i,j) \in \mathcal{F}_{RF}, \mathbf{W}_{RF}(i,j) \in \mathcal{W}_{RF}, \forall i,j, \\ &\mathbf{W}_{RF} = \mathrm{blk}\left[\mathbf{W}_{RF}^1, \ldots, \mathbf{W}_{RF}^K\right], \\ &\mathbf{W}_{BB} = \mathrm{blk}\left[\mathbf{W}_{BB}^1, \ldots, \mathbf{W}_{BB}^K\right], \\ &\|\mathbf{P}\|_{\mathrm{F}}^2 = P_t, \\ &\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_{\mathrm{F}}^2 = KN_S, \end{aligned} \tag{10}$$

where $\mathcal{F}_{RF}$ and $\mathcal{W}_{RF}$ are the feasible sets of constant-modulus complex numbers of $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$, respectively. Since both the objective function and the constraints are nonconvex, the original problem in (10) is nonconvex. Solving the original problem in (10) directly, the five matrix variables $(\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{RF}, \mathbf{W}_{BB}, \mathbf{P})$ need to be jointly optimized, and finding the global optima of the joint optimization problems with similar constraints is intractable for the MU-MIMO case [38]. Therefore, we utilize the two-stage design method to obtain the analog and digital beamforming solutions of the original problem, thereby reducing the difficulty of the solution process. The proposed PSIA method is utilized to obtain the closed-form optimization solutions of $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$, where each total iteration includes one external and m internal iterations. Then, the equivalent baseband channel $\tilde{\mathbf{H}}$ is exploited to design the optimal digital beamformer $\mathbf{F}_{BB}$ and combiner $\mathbf{W}_{BB}$. Finally, the power allocation matrix $\mathbf{P}$ is designed by using waterfilling.

A. DESIGN OF INITIAL ANALOG COMBINING MATRIX

Assuming the signal transmission mode can be described as $\mathbf{s} \to \mathbf{x} \xrightarrow{\mathbf{H}} \hat{\mathbf{x}} \to \hat{\mathbf{s}}$, where $\mathbf{x}$ is the transmitted symbol in the analog stage, i.e., $\mathbf{x} = \mathbf{F}_{BB}\mathbf{P}\mathbf{s}$, and $\hat{\mathbf{x}}$ is the corresponding received signal, i.e., $\hat{\mathbf{x}} = \mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{x} + \mathbf{W}_{RF}^H\mathbf{n}$. Due to $\mathbf{F}_{BB}$ obtained by the piecewise successive approximation method in the digital stage is a unitary matrix, i.e., $\mathbf{F}_{BB}\mathbf{F}_{BB}^H = \mathbf{I}_{KN_S}$, $\mathbb{E}\|\mathbf{x}\|_2^2 = \frac{P_t}{KN_S}$ can be derived. To avoid the data loss, we firstly utilize the minimum MSE (MMSE) between $\mathbf{x}$ and $\hat{\mathbf{x}}$ in the analog stage as the objective function to obtain the analog combining matrix $\mathbf{W}_{RF\_init}$ with the optimal phase. Then $\mathbf{W}_{RF\_init}$ is used as the initial value of the analog beamforming design to maximize the baseband channel capacity, so as to reduce the number of internal iterations of the design in the analog stage.

According to the previous analysis, the analog combiner $\mathbf{W}_{RF\_init}$ is not only constrained to be constant-modulus, but also constrained to be block-diagonal. Therefore, the MMSE-based analog combining problem is formulated as follows

$$\begin{aligned}\mathbf{W}_{RF\_init} = &\min_{\mathbf{W}_{RF\_init}} \left(\mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right) \\ \text{s.t.} &\mathbf{W}_{RF\_init}^k(i,j) \in \mathcal{W}_{RF\_init}, \forall i,j, \\ &\mathbf{W}_{RF\_init} = \mathrm{blk}\left[\mathbf{W}_{RF\_init}^1, \ldots, \mathbf{W}_{RF\_init}^K\right], \end{aligned} \tag{11}$$

It is worth noting that the expectation operation in (11) is difficult to be handled. For this reason, we first assume the analog beamformer $\mathbf{F}_{RF}$ is fixed, then the objective function in (11) can be rewritten as

$$\begin{aligned}\mathbb{E}\left(\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right) &= \mathbb{E}\left(\left\|\mathbf{x} - \mathbf{W}_{RF\_init}^H\mathbf{y}_{\mathbf{x}}\right\|_2^2\right) \\ &= \mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right] - 2\mathrm{Re}\left(\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\mathbf{W}_{RF\_init}\right)\right. \\ &\quad \left. + \mathbf{W}_{RF\_init}^H\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\mathbf{W}_{RF\_init}\right), \end{aligned} \tag{12}$$

where $\mathbf{y}_{\mathbf{x}} = \mathbf{H}\mathbf{F}_{RF}\mathbf{x} + \mathbf{n}$. By introducing a constant matrix $\tilde{\mathbf{W}} = \mathbb{E}\left[\mathbf{x}\mathbf{y}_{\mathbf{x}}^H\right]\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]^{-1}$, the second term of (12) can be re-expressed as

$$\mathbb{E}\left[\mathbf{x}\mathbf{y}_{\mathbf{x}}^H\right]\mathbf{W}_{RF\_init} = \tilde{\mathbf{W}}\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\mathbf{W}_{RF\_init} \tag{13}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

Since $\tilde{\mathbf{W}}\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\tilde{\mathbf{W}}^H$ is a constant value, (12) can be reformulated by using (13) as follows

$$\begin{aligned}&\mathrm{tr}\left(\tilde{\mathbf{W}}\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\tilde{\mathbf{W}}^H-2\mathrm{Re}\left(\tilde{\mathbf{W}}\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\mathbf{W}_{RF\_init}\right)\right.\\&\quad+\mathbf{W}_{RF\_init}^H\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\mathbf{W}_{RF\_init}\\&\quad\left.+\mathbb{E}\left[\mathbf{x}\mathbf{x}^H\right]-\tilde{\mathbf{W}}\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]\tilde{\mathbf{W}}^H\right)\\&=\left\|\tilde{\mathbf{W}}\mathbf{R}_{\mathbf{x}}^{\frac{1}{2}}-\mathbf{W}_{RF\_init}^H\mathbf{R}_{\mathbf{x}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2+\text{constant}\end{aligned}\tag{14}$$

It can be seen from (14) that the solution of the objective function is independent of the constant term, by removing the constant term in (14), the minimization problem in (12) can be equivalent to solving the following problem as

$$\begin{aligned}(\mathcal{P})\ \mathbf{W}_{RF\_init}=&\arg\min_{\mathbf{W}_{RF\_init}}\left\|\tilde{\mathbf{W}}\mathbf{R}_{\mathbf{x}}^{\frac{1}{2}}-\mathbf{W}_{RF\_init}^H\mathbf{R}_{\mathbf{x}}^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\\&\mathrm{s.t.}\mathbf{W}_{RF\_init}^k(i,j)\in\mathcal{W}_{RF\_init},\forall i,j,\\&\mathbf{W}_{RF\_init}=\mathrm{blk}\left[\mathbf{W}_{RF\_init}^1,\ldots,\mathbf{W}_{RF\_init}^K\right],\end{aligned}$$

where

$$\begin{cases}\mathbf{R}_{\mathbf{x}}=\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]=\frac{P_t}{KN_S}\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{RF}^H\mathbf{H}^H+\sigma^2\mathbf{I}_{KN_{MS}},\\\tilde{\mathbf{W}}=\mathbb{E}\left[\mathbf{x}\mathbf{y}_{\mathbf{x}}^H\right]\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^H\right]^{-1}\\\quad=\mathbf{F}_{RF}^H\mathbf{H}^H\left(\frac{P_t}{KN_S}\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{RF}^H\mathbf{H}^H+\sigma^2\mathbf{I}_{KN_{MS}}\right)^{-1}.\end{cases}\tag{15}$$

Although (11) has been transformed into the problem with no-expectation operation, there are still other constraints in $(\mathcal{P})$. To tackle the block-diagonal constraint in $\mathbf{W}_{RF\_init}$, the matrix operator can be converted into a vector operator, based on the properties of Kronecker product: $\mathrm{vec}\left(\mathbf{XGY}\right)=\left(\mathbf{Y}^T\otimes\mathbf{X}\right)\mathrm{vec}\left(\mathbf{G}\right)$, where $\otimes$ represents the Kronecker product between two matrices. Meanwhile, to satisfy the constant-modulus constraints, we initialize $|\mathbf{W}_{RF\_init}(i,j)|=1$. The problem $(\mathcal{P})$ can be reformulated as

$$\begin{aligned}\mathbf{w}_{RF\_init}=&\arg\min_{\mathbf{w}_{RF\_init}}\|\mathbf{d}-\mathbf{A}\mathbf{w}_{RF\_init}\|_2^2\\&\mathrm{s.t.}\left|\mathbf{W}_{RF\_init}^k(i,j)\right|=1,\forall i,j,\\\mathbf{w}_{RF\_init}=&\mathrm{vec}\big(\mathrm{blk}\left[\mathbf{W}_{RF\_init}^1,\ldots,\mathbf{W}_{RF\_init}^K\right]\big)^H,\end{aligned}\tag{16}$$

where $\mathbf{d}=\mathrm{vec}\left(\tilde{\mathbf{W}}\mathbf{R}_{\mathbf{x}}^{\frac{1}{2}}\right)$ and $\mathbf{A}=\left(\mathbf{R}_{\mathbf{x}}^{\frac{1}{2}}\right)^H\otimes\mathbf{I}_{KM_{MS}}$. It is observed from (16) that the variable $\mathbf{w}_{RF\_init}$ has many zero elements. Since these zero elements do not contribute anything in the procedure of matrix multiplication, hence that the zero elements in $\mathbf{w}_{RF\_init}$ can be removed [18]. The new variable vector is given by

$$\hat{\mathbf{w}}_{RF\_init}=\mathrm{vec}\left(\hat{\mathbf{w}}_{RF\_init}^1,\ldots,\hat{\mathbf{w}}_{RF\_init}^K\right),\tag{17}$$

where $\hat{\mathbf{w}}_{RF\_init}^k=\mathrm{vec}\left(\mathbf{W}_{RF\_init}^{k\,H}\right),k=1,\ldots,K$. Replacing $\mathbf{w}_{RF\_init}$ by $\hat{\mathbf{w}}_{RF\_init}$, the block-diagonal constraint in (16) is eliminated. Therefore, (16) can be rewritten as

$$\begin{aligned}&\min_{\mathbf{x}}\left\|\tilde{\mathbf{d}}-\tilde{\mathbf{A}}\tilde{\mathbf{x}}\right\|_2^2\\&\mathrm{s.t.}\,|\tilde{\mathbf{x}}|=\mathbf{1},\end{aligned}\tag{18}$$

where $\tilde{\mathbf{d}}$ and $\tilde{\mathbf{A}}$ are respectively the vectors and matrices generated from $\mathbf{d}$ and $\mathbf{A}$ after removing the columns corresponding to the zero elements in $\mathbf{w}_{RF\_init}$, $\tilde{\mathbf{x}}=\hat{\mathbf{w}}_{RF\_init}$, and the vector "$\mathbf{1}$" denotes the all-one vector. Inspired by the SCF algorithm [39], the constant-modulus constraint can be eliminated by considering (18) in real domain. Hence, consider the sequence of constraint

$$\mathrm{Re}\left(\mathbf{B}^{(n)}\tilde{\mathbf{x}}\right)=\mathbf{1},\tag{19}$$

with

$$\mathbf{B}^{(n)}(i,j)=\begin{cases}e^{j\angle\tilde{x}_l^{(n-1)}},&\text{if } i=j=l,\\0,&\text{otherwise,}\end{cases}\tag{20}$$

where $\tilde{x}_l^{(n)}$, $l=1,2,\ldots,L$ is the lth element of $\tilde{\mathbf{x}}^{(n)}$, $L=KN_{MS}M_{MS}$, and (n) denotes the nth optimization procedure. Replacing the constant-modulus constraint in (18) by (19), (18) can be rewritten as

$$\left(Q^{(n)}\right)\begin{cases}\min\limits_{\tilde{\mathbf{x}}}\left\|\tilde{\mathbf{d}}-\tilde{\mathbf{A}}\tilde{\mathbf{x}}\right\|_2^2\\\mathrm{s.t.}\ \mathrm{Re}\left(\mathbf{B}^{(n)}\tilde{\mathbf{x}}\right)=\mathbf{1}.\end{cases}\tag{21}$$

To illustrate the constraint (19) is adjusted to satisfy the constant-modulus constraint, let $\tilde{\mathbf{x}}^{(n-1)}$ be the solution which satisfies the constraint $\mathrm{Re}\left(\mathbf{B}^{(n-1)}\tilde{\mathbf{x}}^{(n-1)}\right)=\mathbf{1}$, and the constant-modulus affine solution of $\tilde{\mathbf{x}}^{(n-1)}$ is given by $\tilde{\mathbf{x}}_{(n-1)}=e^{j\angle\tilde{\mathbf{x}}^{(n-1)}}$. If $\tilde{\mathbf{x}}^{(n)}=\tilde{\mathbf{x}}_{(n-1)}$, then $\mathbf{B}^{(n)}=\mathbf{B}^{(n+1)}$, and the constraints of the next problem $Q^{(n+1)}$ are the same as problem $Q^{(n)}$, i.e., $\mathrm{Re}\left(\mathbf{B}^{(n+1)}\tilde{\mathbf{x}}^{(n+1)}\right)=\mathbf{1}$ and $\mathrm{Re}\left(\mathbf{B}^{(n)}\tilde{\mathbf{x}}^{(n)}\right)=\mathbf{1}$ are equivalent. Thus, $\tilde{\mathbf{x}}^{(n+1)}=\tilde{\mathbf{x}}^{(n)}$ is derived which means the algorithm converges. Otherwise, the constraint is updated through the constant-modulus affine solution of $\tilde{\mathbf{x}}^{(n)}$ according to (20). As a conclusion, the obtained solution converges to a constant-modulus one by the adaptive constraint.

For transforming the cost function into the completely equivalent real-valued version, let $\tilde{\mathbf{s}}^{(n)}=\left[\mathrm{Re}\{\tilde{\mathbf{x}}^{(n)}\}^T\ \mathrm{Im}\{\tilde{\mathbf{x}}^{(n)}\}^T\right]^T$ be the optimal solution of $Q^{(n)}$ and $\tilde{\mathbf{x}}^{(n)}$ be the complex version defined as

$$\tilde{x}_l^{(n)}=\tilde{s}_l^{(n)}+j\tilde{s}_{l+L}^{(n)},\ l=1,2,\ldots,L,\tag{22}$$

where $\tilde{s}_l^{(n)}$ is the lth element of $\tilde{\mathbf{s}}^{(n)}$. In this case, the matrix $\mathbf{B}^{(n)}$ is defined as

$$\mathbf{B}^{(n)}(i,j)=\begin{cases}\cos\left(\angle\tilde{x}_l^{(n-1)}\right),\text{if } i=j=l,\\\sin\left(\angle\tilde{x}_l^{(n-1)}\right),\text{if } i=l,j=l+L,\\0,\qquad\qquad\text{otherwise.}\end{cases}\tag{23}$$

Therefore, the optimal solution of complex-valued problem $Q^{(n)}$ can be obtained by solving the completely equivalent real-valued problem as follows

$$\left(\tilde{Q}^{(n)}\right)\begin{cases}\min\limits_{\mathbf{s}}\left\|\hat{\mathbf{d}}-\hat{\mathbf{A}}\tilde{\mathbf{s}}\right\|_2^2\\\mathrm{s.t.}\ \mathbf{B}^{(n)}\tilde{\mathbf{s}}=\mathbf{1},\end{cases}\tag{24}$$

where

$$\hat{\mathbf{d}}=\left[\mathrm{Re}\left(\tilde{\mathbf{d}}\right)^T\ \mathrm{Im}\left(\tilde{\mathbf{d}}\right)^T\right]^T$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathrm{Re}\left(\tilde{\mathbf{A}}\right) & -\mathrm{Im}\left(\tilde{\mathbf{A}}\right) \\ \mathrm{Im}\left(\tilde{\mathbf{A}}\right) & \mathrm{Re}\left(\tilde{\mathbf{A}}\right) \end{bmatrix}.$$

Note that, $\tilde{Q}^{(n)}$ is a convex optimization problem with linear equality constraints. Therefore, according to (24), the Lagrange function of $\tilde{Q}^{(n)}$ is given by

$$\begin{aligned} L\left(\tilde{\mathbf{s}}, \boldsymbol{\lambda}\right) &= \left\| \hat{\mathbf{d}} - \hat{\mathbf{A}}\tilde{\mathbf{s}} \right\|_2^2 - \boldsymbol{\lambda}^H \left(\mathbf{B}^{(n)}\tilde{\mathbf{s}} - \mathbf{1} \right) \\ &= \mathrm{tr}\left\{ \left(\hat{\mathbf{d}} - \hat{\mathbf{A}}\tilde{\mathbf{s}} \right)^H \left(\hat{\mathbf{d}} - \hat{\mathbf{A}}\tilde{\mathbf{s}} \right) - \boldsymbol{\lambda}^H \left(\mathbf{B}^{(n)}\tilde{\mathbf{s}} - \mathbf{1} \right) \right\} \\ &= 2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\tilde{\mathbf{s}} - 2\hat{\mathbf{A}}^H\hat{\mathbf{d}} + \hat{\mathbf{d}}^H\hat{\mathbf{d}} - \boldsymbol{\lambda}^H \left(\mathbf{B}^{(n)}\tilde{\mathbf{s}} - \mathbf{1} \right), \end{aligned} \quad (25)$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier. Let the partial derivative of (25) with respect to $\mathbf{s}$ and $\boldsymbol{\lambda}$ are $\frac{\partial L(\tilde{\mathbf{s}}, \boldsymbol{\lambda})}{\partial \tilde{\mathbf{s}}} = 0$ and $\frac{\partial L(\tilde{\mathbf{s}}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 0$, respectively. Thus, we have the equation set as follows

$$\begin{bmatrix} 2\hat{\mathbf{A}}^H\hat{\mathbf{A}} & -\left(\mathbf{B}^{(n)}\right)^H \\ -\mathbf{B}^{(n)} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{s}} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} 2\hat{\mathbf{A}}^H\hat{\mathbf{d}} \\ -\mathbf{1} \end{bmatrix}. \quad (26)$$

Since the coefficient matrix (Lagrange matrix) of the equation set is non-singular, Lagrange matrix is invertible. Then the inverse matrix can be expressed as

$$\begin{bmatrix} 2\hat{\mathbf{A}}^H\hat{\mathbf{A}} & -\left(\mathbf{B}^{(n)}\right)^H \\ -\mathbf{B}^{(n)} & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C} & -\mathbf{D}^H \\ -\mathbf{D} & \mathbf{Z} \end{bmatrix}. \quad (27)$$

Based on the properties of inverse matrix: $\mathbf{X}^{-1}\mathbf{X} = \mathbf{I}$, we can derive

$$\begin{cases} 2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\mathbf{C} + \left(\mathbf{B}^{(n)}\right)^H\mathbf{D} = \mathbf{I}_{2KN_{MS}M_{MS}} \\ -2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\mathbf{D}^H - \left(\mathbf{B}^{(n)}\right)^H\mathbf{Z} = \mathbf{0}_{2KN_{MS}M_{MS} \times KN_{MS}M_{MS}} \\ -\mathbf{B}^{(n)}\mathbf{C} = \mathbf{0}_{KN_{MS}M_{MS} \times 2KN_{MS}M_{MS}} \\ \mathbf{B}^{(n)}\mathbf{D}^H = \mathbf{I}_{KN_{MS}M_{MS}} \end{cases} \quad (28)$$

Since $\hat{\mathbf{A}}^H\hat{\mathbf{A}}$ is the Hermitian matrix, its inverse matrix exists. Then $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{Z}$ can be expressed as

$$\begin{cases} \mathbf{C} = \left(2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\right)^{-1} \\ \quad -\left(2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\right)^{-1}\left(\mathbf{B}^{(n)}\right)^H\left(\mathbf{B}^{(n)}\left(2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\right)^{-1}\left(\mathbf{B}^{(n)}\right)^H\right)^{-1}\mathbf{B}^{(n)}\left(2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\right)^{-1} \\ \mathbf{D} = \left(\mathbf{B}^{(n)}\left(2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\right)^{-1}\left(\mathbf{B}^{(n)}\right)^H\right)^{-1}\mathbf{B}^{(n)}\left(2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\right)^{-1} \\ \mathbf{Z} = -\left(\mathbf{B}^{(n)}\left(2\hat{\mathbf{A}}^H\hat{\mathbf{A}}\right)^{-1}\left(\mathbf{B}^{(n)}\right)^H\right)^{-1} \end{cases} \quad (29)$$

Both sides of (26) are multiplied by the inverse of the Lagrange matrix, the solution of $\tilde{Q}^{(n)}$ can be obtained as

$$\tilde{\mathbf{s}}^{(n)} = 2\mathbf{C}\hat{\mathbf{A}}^H\hat{\mathbf{d}} + \mathbf{D}^H\mathbf{1}, \quad (30)$$

where $\boldsymbol{\lambda} = -2\mathbf{D}\hat{\mathbf{A}}^H\hat{\mathbf{D}} - \mathbf{Z}\mathbf{1}$. Although the problem in (30) cannot obtain an optimal amplitude solution, an optimal phase solution can be obtained.

B. ANALOG BEAMFORMING AND COMBINING MATRICES DESIGN

In the previous subsection, the analog combining matrix $\mathbf{W}_{RF\_init}$ with optimal phase has been obtained. Thus, the PSIA method is utilized to solve the optimal analog beamforming $\mathbf{F}_{RF}^{opt}$ and combining $\mathbf{W}_{RF}^{opt}$ in this subsection.

According to the rationales of information theory, when the inner-user and inter-user interference are eliminated and the downlink broadcast channel capacity of the overall baseband channel $\bar{\mathbf{H}}$ can be reached, the capacity of the system is equal to the mutual information between $\mathbf{x}$ and $\hat{\mathbf{x}}$, i.e., $R = I\left(\mathbf{x}, \hat{\mathbf{x}}\right)$ [28]. Since the proposed digital beamforming is designed based on SVD, the resulting solutions are unitary matrices and can make the baseband channels of different users mutually orthogonal. In addition, the digital beamforming matrix satisfies $\|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_{\mathrm{F}}^2 = KN_S$. Therefore, maximizing the capacity of the overall system can be approximately equal to maximizing $I\left(\mathbf{x}, \hat{\mathbf{x}}\right)$, i.e., $R \approx I\left(\mathbf{x}, \hat{\mathbf{x}}\right)$, and the expression of $I\left(\mathbf{x}, \hat{\mathbf{x}}\right)$ is described as

$$\begin{aligned} I\left(\mathbf{x}, \hat{\mathbf{x}}\right) = \log_2 \big(& \left| \mathbf{I}_{KM_{MS}} \right. \\ & \left. + \tfrac{P_t}{\sigma^2 KN_S}\left(\mathbf{W}_{RF}^H\mathbf{W}_{RF}\right)^{-1}\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{RF}^H\mathbf{H}^H\mathbf{W}_{RF} \right| \big) \end{aligned} \quad (31)$$

Assuming the designed analog combining matrix is a unitary matrix, i.e., $\mathbf{W}_{RF}^H\mathbf{W}_{RF} = \frac{1}{N_{MS}}\mathbf{I}_{KM_{MS}}$, (31) can be rewritten as

$$\begin{aligned} & I\left(\mathbf{x}, \hat{\mathbf{x}}\right) \\ & = \log_2 \left(\left| \mathbf{I}_{KM_{MS}} + \tfrac{P_t N_{MS}}{\sigma^2 KN_S}\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{RF}^H\mathbf{H}^H\mathbf{W}_{RF} \right| \right) \end{aligned} \quad (32)$$

It can be seen from (32) that maximizing $I\left(\mathbf{x}, \hat{\mathbf{x}}\right)$ is equivalent to maximizing $\left\|\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\right\|_{\mathrm{F}}^2$, which means that the maximum equivalent baseband channel is only related to the solution of analog beamforming $\mathbf{F}_{RF}^{opt}$ with $\mathbf{W}_{RF}$ is fixed. Thus, the problem of analog beamforming can be formulated as

$$\begin{aligned} (\mathcal{P}_1)\ \ & \mathbf{F}_{RF} = \arg\max_{\mathbf{F}_{RF}} \left\|\mathbf{H}_{comp}\mathbf{F}_{RF}\right\|_{\mathrm{F}}^2 \\ & \text{s.t. } \mathbf{F}_{RF}\left(i,j\right) \in \mathcal{F}_{RF}, \forall i, j, \end{aligned} \quad (33)$$

where $\mathbf{H}_{comp} = \mathbf{W}_{RF\_init}^H\mathbf{H}$. Inspired by the method which tries to avoid the loss of information [28], the SVD of composite channel is defined as $\frac{1}{\sqrt{N_{BS}}}\mathbf{H}_{comp} = \mathbf{U}_{comp}\boldsymbol{\Sigma}_{comp}\mathbf{V}_{comp}^H$, where $\mathbf{U}_{comp}$ and $\mathbf{V}_{comp}$ are the left and right singular value vector matrices, respectively, and $\boldsymbol{\Sigma}_{comp}$ is the singular value vector matrix sorted by descending order. Therefore, the objective function in (33) can be reformulated as

$$\begin{aligned} \left\|\mathbf{H}_{comp}\mathbf{F}_{RF}\right\|_{\mathrm{F}}^2 &= N_{BS}\left\|\frac{1}{\sqrt{N_{BS}}}\mathbf{H}_{comp}\mathbf{F}_{RF}\right\|_{\mathrm{F}}^2 \\ &= N_{BS}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{comp}^2\mathbf{V}_{comp}^H\mathbf{F}_{RF}\mathbf{F}_{RF}^H\mathbf{V}_{comp}\right\}. \end{aligned} \quad (34)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

Further, we define the following two partitions of the matrices $\mathbf{\Sigma}_{comp}$ and $\mathbf{V}_{comp}$ as

$$\mathbf{\Sigma}_{comp}=\begin{bmatrix}\mathbf{\Sigma}_{comp}^{1} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_{comp}^{2}\end{bmatrix},\mathbf{V}_{comp}=\begin{bmatrix}\mathbf{V}_{comp}^{1} & \mathbf{V}_{comp}^{2}\end{bmatrix}, \tag{35}$$

where $\mathbf{\Sigma}_{comp}^{1}\in\mathbb{C}^{KN_{MS}\times M_{BS}}$ and $\mathbf{V}_{comp}^{1}\in\mathbb{C}^{N_{BS}\times M_{BS}}$. Substituting (35) into (34) gives rise to the following approximate expression of (34)

$$\begin{aligned}&\mathrm{tr}\left\{\mathbf{\Sigma}_{comp}^{2}\mathbf{V}_{comp}^{H}\mathbf{F}_{RF}\mathbf{F}_{RF}^{H}\mathbf{V}_{comp}\right\}\\&\quad\simeq\mathrm{tr}\left\{\mathbf{\Sigma}_{comp}^{1^2}\mathbf{V}_{comp}^{1^H}\mathbf{F}_{RF}\mathbf{F}_{RF}^{H}\mathbf{V}_{comp}^{1}\right\},\end{aligned} \tag{36}$$

As can be seen obviously from (36) that the objective function $\mathrm{tr}\left\{\mathbf{\Sigma}_{comp}^{1^2}\mathbf{V}_{comp}^{1^H}\mathbf{F}_{RF}\mathbf{F}_{RF}^{H}\mathbf{V}_{comp}^{1}\right\}$ is maximized when the unconstrained analog beamforming $\mathbf{F}_{RF}=\mathbf{V}_{comp}^{1}$. Unfortunately, $\mathbf{V}_{comp}^{1}$ does not satisfy constant-modulus constraint. However, the Frobenius norm can be employed to compute the distance between the unconstrained and the constrained solutions. Therefore, $\mathbf{F}_{RF}$ can be obtained by solving

$$\begin{aligned}&\min_{\mathbf{F}_{RF}}\left\|\mathbf{V}_{comp}^{1}-\mathbf{F}_{RF}\right\|_{\mathrm{F}}^{2}\\&\text{s.t. }\mathbf{F}_{RF}(i,j)\in\mathcal{F}_{RF},\forall i,j.\end{aligned} \tag{37}$$

Then, the objective function in (37) is expanded as follows

$$\begin{aligned}&\left\|\mathbf{V}_{comp}^{1}-\mathbf{F}_{RF}\right\|_{F}^{2}=\mathrm{tr}\left\{\left(\mathbf{V}_{comp}^{1}-\mathbf{F}_{RF}\right)^{H}\left(\mathbf{V}_{comp}^{1}-\mathbf{F}_{RF}\right)\right\}\\&=2M_{BS}-\mathrm{tr}\left\{2\mathrm{Re}\left(\mathbf{F}_{RF}^{H}\mathbf{V}_{comp}^{1}\right)\right\}\\&=2M_{BS}-2\sum_{j=1}^{M_{BS}}\sum_{i=1}^{N_{BS}}\mathrm{Re}\left\{\left|\mathbf{F}_{RF}(i,j)\right|\left|\mathbf{V}_{comp}^{1}(i,j)\right|e^{j\varphi(i,j)}\right\}\end{aligned} \tag{38}$$

where $\varphi(i,j)=\angle\mathbf{F}_{RF}(i,j)-\angle\mathbf{V}_{comp}^{1}(i,j)$. From (38), we observe that when $\varphi(i,j)=0$, i.e., $\angle\mathbf{F}_{RF}(i,j)=\angle\mathbf{V}_{comp}^{1}(i,j)$, the objective function $\left\|\mathbf{V}_{comp}^{1}-\mathbf{F}_{RF}\right\|_{\mathrm{F}}^{2}$ is minimized. Therefore, the optimal beamforming matrix can be expressed as

$$\mathbf{F}_{RF}=\frac{1}{\sqrt{N_{BS}}}e^{j\angle\mathbf{V}_{comp}^{1}} \tag{39}$$

Next, the obtained $\mathbf{F}_{RF}$ is substituted into the objective function of maximizing the equivalent baseband channel to solve the optimal analog combining $\mathbf{W}_{RF}$. Since $\mathbf{W}_{RF}$ is block-diagonal, the objective function is formulated as

$$\begin{aligned}(\mathcal{P}_2)\ \mathbf{W}_{RF}^{k}&=\arg\max_{\mathbf{W}_{RF}^{k}}\left\|\mathbf{W}_{RF}^{k}\hat{\mathbf{H}}_{comp}^{k}\right\|_{\mathrm{F}}^{2}\\&\text{s.t. }\mathbf{W}_{RF}^{k}(i,j)\in\mathcal{W}_{RF},\forall i,j,\\&\quad\mathbf{W}_{RF}=\mathrm{blk}\left[\mathbf{W}_{RF}^{1},\ldots,\mathbf{W}_{RF}^{K}\right],\end{aligned} \tag{40}$$

where $\hat{\mathbf{H}}_{comp}^{k}=\mathbf{H}_{k}\mathbf{F}_{RF}$, The SVD of $\hat{\mathbf{H}}_{comp}^{k}$ is defined as $\frac{1}{\sqrt{N_{MS}}}\hat{\mathbf{H}}_{comp}^{k}=\hat{\mathbf{U}}_{comp}^{k}\hat{\mathbf{\Sigma}}_{comp}^{k}\hat{\mathbf{V}}_{comp}^{k^H}$. Moreover, $\tilde{\mathbf{U}}_{comp}^{k}$ is expressed as the first M_{MS} columns of matrix $\hat{\mathbf{U}}_{comp}^{k}$. Substituting this decomposition into (40) gives rise to the following approximate expression

$$\begin{aligned}&\mathrm{tr}\left\{\hat{\mathbf{\Sigma}}_{comp}^{k^2}\hat{\mathbf{U}}_{comp}^{k^H}\mathbf{W}_{RF}\mathbf{W}_{RF}^{H}\hat{\mathbf{U}}_{comp}^{k}\right\}\\&\quad\simeq\mathrm{tr}\left\{\tilde{\mathbf{\Sigma}}_{comp}^{k^2}\tilde{\mathbf{U}}_{comp}^{k^H}\mathbf{W}_{RF}\mathbf{W}_{RF}^{H}\tilde{\mathbf{U}}_{comp}^{k}\right\},\end{aligned} \tag{41}$$

where $\tilde{\mathbf{\Sigma}}_{comp}^{k}$ is an $M_{MS}\times M_{MS}$ diagonal matrix. Compared (41) with (36), they are similar in nature. Hence, the optimal solution of matrix $\mathbf{W}_{RF}^{k}$ can be expressed as

$$\mathbf{W}_{RF}^{k}=\frac{1}{\sqrt{N_{MS}}}e^{j\angle\tilde{\mathbf{U}}_{comp}^{k}}. \tag{42}$$

Then, the total analog combining matrix $\mathbf{W}_{RF}$ is obtained by block diagonalization. It is worth noting that the initial value of subproblems $\left\{\mathcal{P}_{i}^{(n)}\right\}_{i=1,2}$ (internal iteration) derives from the problem $\mathcal{P}$ (external iteration). In the mth internal iteration of the nth external iteration, the subproblem $\mathcal{P}_{1}^{(n)^{(m+1)}}$ is updated by the results of $\mathcal{P}_{2}^{(n)^{(m)}}$, and the subproblem $\mathcal{P}_{2}^{(n)^{(m+1)}}$ is updated by the results of $\mathcal{P}_{1}^{(n)^{(m+1)}}$. Therefore, the monotonically non-increasing of the sequences produced by the subproblems $\left\{\mathcal{P}_{i}^{(n)}\right\}_{i=1,2}$ satisfy $f_{*}^{(n)^{(m+1)}}\leq f_{*}^{(n)^{(m)}}$ and eventually the objective values converge [18]. Then, the convergence of $\mathcal{P}^{(n)}$ can be guaranteed by $\mathbf{B}^{(n)}$. Therefore, the proposed analog beamforming design method is monotonically non-increase and eventually converges. In addition, the stop criterion of the external iteration for the proposed method is set as $\left|f^{(n+1)}-f^{(n)}\right|\leq\varepsilon$, where ε is a small factor, and $f^{(n)}$ is the objective value of (11) in the nth external iteration, namely,

$$\begin{aligned}f^{(n)}=&\mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^{H}\right]-2\mathrm{Re}\left(\mathbb{E}\left[\mathbf{x}\mathbf{y}_{\mathbf{x}}^{H}\right]\mathbf{W}_{RF\_init}\right)\right.\\&\left.+\mathbf{W}_{RF\_init}^{H}\mathbb{E}\left[\mathbf{y}_{\mathbf{x}}\mathbf{y}_{\mathbf{x}}^{H}\right]\mathbf{W}_{RF\_init}\right)\end{aligned} \tag{43}$$

In conclusion, the advantages of the analog beamformer and combiner designed by the above methods are that the channel information can be utilized more adequately to improve the sum rate of system. The overall procedure of the proposed analog beamforming scheme is summarized as **Algorithm 1**.

C. DIGITAL BEAMFORMING AND COMBINING MATRICES DESIGN

This subsection discusses the design of digital beamforming and combining matrices based on the analog beamforming and combining matrices obtained in the previous subsection. Considering the baseband BD scheme, which employs ZF to eliminate inter-user interference by the null space orthogonal bases of baseband channels of different users and ensure zero inner-user interference by digital combiner in the digital beamforming stage. Thus, a MU-MIMO downlink channel can be decomposed into multiple parallel independent SU-MIMO channels [40]. However, when the number of users is large, the overlap of the row subspace of channel matrix per user becomes significant, which results in a quite poor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

Algorithm 1 : The proposed analog beamforming design scheme

Input: $\mathbf{H}$, N_{iter};
Randomly generate the initial matrix $\mathbf{F}_{RF}$ (satisfied with the constant-modulus constraint);
while $\left|f^{(n+1)}-f^{(n)}\right|>\varepsilon$ **do**;
  Set $n=1$;
  Compute $\mathbf{B}^{(n)}$ according to (23);
  Compute $\mathbf{s}^{(n)}$ according to (30);
  Set $\tilde{x}_l^{(n)}=\tilde{s}_l^{(n)}+j\tilde{s}_{l+L}^{(n)}, l=1,2,\ldots,L$;
  Get $\mathbf{W}_{RF\_init}^{(n)}=\left\{\mathrm{blk}\left[\mathrm{unvec}_{M_{MS}\times KN_{MS}}\left(e^{\left(\angle\tilde{\mathbf{x}}^{(n)}\right)}\right)\right]\right\}^H$ and assign $\mathbf{W}_{RF}^{(n)}=\mathbf{W}_{RF\_init}^{(n)}$;
  for $m=1$ to N_{iter} **do**
    Compute $\mathbf{F}_{RF}^{(n)^{(m)}}=\frac{1}{\sqrt{N_{BS}}}e^{j\angle\mathbf{V}_{comp}^{(m)}(:,1:M_{BS})}$, where $\mathbf{V}_{comp}^{(m)}$ is derived by $\frac{1}{\sqrt{N_{BS}}}\mathbf{W}_{RF}^{(n)^{(m)}}\mathbf{H}=\mathbf{U}_{comp}^{(m)}\mathbf{\Sigma}_{comp}^{(m)}\mathbf{V}_{comp}^{(m)^{(H)}}$;
    for $k=1$ to K **do**
      Compute $\left(\mathbf{W}_{RF}^{k}\right)^{(n)^{(m)}}=\frac{1}{\sqrt{N_{MS}}}e^{j\angle\hat{\mathbf{U}}_{comp}^{k^{(m)}}(:,1:M_{MS})}$, where $\hat{\mathbf{U}}_{comp}^{k^{(m)}}$ is derived by $\frac{1}{\sqrt{N_{MS}}}\mathbf{H}_k\mathbf{F}_{RF}=\hat{\mathbf{U}}_{comp}^{k}\hat{\mathbf{\Sigma}}_{comp}^{k}\hat{\mathbf{V}}_{comp}^{k^H}$;
    end for
    Obtain the total matrix $\mathbf{W}_{RF}^{(n)^{(m)}}=\mathrm{blk}\left[\left(\mathbf{W}_{RF}^{1}\right)^{(n)^{(m)}},\ldots,\left(\mathbf{W}_{RF}^{K}\right)^{(n)^{(m)}}\right]$;
  end for
  Set $n=n+1$;
end while
Output: $\mathbf{F}_{RF}$, $\mathbf{W}_{RF}$

performance [41]. Moreover, the operational dimension of baseband BD scheme also becomes large, thereby increasing the computational complexity. However, in [28], the criterion for trying to avoid the loss of information is proposed to design the analog beamforming, which can reduce the computational complexity. Therefore, inspired by the scheme proposed in [25] and [28], we design the digital beamforming and combining matrices by empolying the criterion which tries to avoid the loss of information. Meanwhile, the complexity analysis shows the advantages over the baseband BD scheme.

The equivalent baseband channel is defined as $\bar{\mathbf{H}}=\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}$. According to (3) and (10), the problem of digital beamforming design can be formulated as

$$\begin{aligned}(\mathbf{F}_{BB},\mathbf{W}_{BB})=&\underset{(\mathbf{F}_{BB},\mathbf{W}_{BB})}{\arg\max}\\ &\log_2\left(\left|\mathbf{I}_{KN_S}+\frac{1}{KN_S}\mathbf{R}_n^{-1}\mathbf{W}_{BB}^H\bar{\mathbf{H}}\mathbf{F}_{BB}\mathbf{P}\mathbf{P}^H\mathbf{F}_{BB}^H\bar{\mathbf{H}}^H\mathbf{W}_{BB}\right|\right)\\ &\text{s.t. }\mathbf{W}_{BB}=\mathrm{blk}\left[\mathbf{W}_{BB}^1,\ldots,\mathbf{W}_{BB}^K\right],\\ &\quad\left\|\mathbf{F}_{RF}\mathbf{F}_{BB}\right\|_{\mathrm{F}}^2=KN_S.\end{aligned}\tag{44}$$

Based on the criterion proposed in [28], the SVD of equivalent baseband channel of each user is expressed as

$$\frac{1}{\sqrt{N_{BS}}}\bar{\mathbf{H}}_k=\bar{\mathbf{U}}_k\bar{\mathbf{\Sigma}}_k\bar{\mathbf{V}}_k^H,\tag{45}$$

Design the digital combining matrix $\mathbf{W}_{BB}^k$ as the first N_S column vector of $\bar{\mathbf{U}}_k$, i.e., $\mathbf{W}_{BB}^k=\bar{\mathbf{U}}_k\left(:,1:N_S\right), k\in\{1,\ldots,K\}$, and bring it into (44), then the objective function can be rewritten as

$$\begin{aligned}\mathbf{F}_{BB}=&\underset{\mathbf{F}_{BB}}{\arg\max}\\ &\log_2\left(\left|\mathbf{I}_{KN_S}+\frac{N_{BS}}{KN_S}\mathbf{R}_n^{-1}\bar{\mathbf{H}}_{comp}\mathbf{F}_{BB}\mathbf{P}\mathbf{P}^H\mathbf{F}_{BB}^H\bar{\mathbf{H}}_{comp}^H\right|\right),\end{aligned}\tag{46}$$

where

$$\begin{aligned}\bar{\mathbf{H}}_{comp}&=\frac{1}{\sqrt{N_{BS}}}\mathbf{W}_{BB}^H\bar{\mathbf{H}}\\ &=\frac{1}{\sqrt{N_{BS}}}\begin{bmatrix}\mathbf{W}_{BB}^{1^H}&\cdots&\mathbf{0}\\ \vdots&\ddots&\vdots\\ \mathbf{0}&\cdots&\mathbf{W}_{BB}^{1^H}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{H}}_1\\ \vdots\\ \bar{\mathbf{H}}_K\end{bmatrix}.\end{aligned}\tag{47}$$

The power allocation obtained by utilizing waterfilling will be approximately performed by equal power allocation in the digital stage, i.e., $\mathbf{P}\approx\sqrt{\frac{P_t}{KN_S}}\mathbf{I}_{KN_S}$ for N_{BS} approaches infinity. Due to $\mathbf{W}_{RF}$ derived in the above subsection is a para-unitary matrix (i.e., $\mathbf{W}_{RF}^H\mathbf{W}_{RF}=\mathbf{I}_{M_{BS}}$), we can obtain

$$\mathbf{R}_n=\sigma^2\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{W}_{RF}\mathbf{W}_{BB}=\sigma^2\mathbf{I}_{KN_S}.\tag{48}$$

Thus, (46) can be further written as

$$\mathbf{F}_{BB}=\underset{\mathbf{F}_{BB}}{\arg\max}\log_2\left(\left|\mathbf{I}_{KN_S}+\frac{N_{BS}P_t}{\sigma^2KN_S}\mathbf{F}_{BB}^H\bar{\mathbf{H}}_{comp}^H\bar{\mathbf{H}}_{comp}\mathbf{F}_{BB}\right|\right).\tag{49}$$

It can be seen from (49) that the optimal digital beamforming $\mathbf{F}_{BB}$ of the maximized objective function can be obtained by the first KN_S columns of the right singular vector of $\bar{\mathbf{H}}_{comp}$. Defining the SVD of $\bar{\mathbf{H}}_{comp}$ as $\bar{\mathbf{H}}_{comp}=\bar{\mathbf{U}}_{comp}\bar{\mathbf{\Sigma}}_{comp}\bar{\mathbf{V}}_{comp}^H$, then we set $\mathbf{F}_{BB}=\bar{\mathbf{V}}_{comp}\left(:,1:KN_S\right)$.

Up to now, the optimal digital beamforming $\mathbf{F}_{BB}$ and combining $\mathbf{W}_{BB}$ obtained to eliminate inter and inner-interference can be illustrated in the following.

For a massive MIMO system with multi-antenna users, the asymptotic orthogonality of different user channels has been proven as the number of BS antennas is large [28]. Here, we extend the conclusion to the baseband channel model, then the correlation matrix of different user baseband channels follows

$$\lim_{N_{BS}\to\infty}\frac{1}{N_{BS}}\bar{\mathbf{H}}_p\bar{\mathbf{H}}_q^H=\mathbf{0}_{M_{MS}\times M_{MS}},\ p,q\in\{1,\ldots,K\},\forall p\neq q.\tag{50}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

Proof: For a massive MIMO system with multi-antenna users, the correlation matrices of different user analog channels satisfy the following asymptotic orthogonality [28]

$$\lim_{N_{BS}\to\infty}\frac{1}{N_{BS}}\mathbf{H}_p\mathbf{H}_q^H=\mathbf{0}_{N_{MS}\times N_{MS}},\ p,q\in\{1,\ldots,K\},\forall p\neq q. \tag{51}$$

Then the correlation matrix of different user baseband channels can be expressed as

$$\lim_{N_{BS}\to\infty}\frac{1}{N_{BS}}\bar{\mathbf{H}}_p\bar{\mathbf{H}}_q^H=\lim_{N_{BS}\to\infty}\frac{1}{N_{BS}}\mathbf{W}_{RF}^{p\,H}\mathbf{H}_p\mathbf{F}_{RF}\mathbf{F}_{RF}^H\mathbf{H}_q^H\mathbf{W}_{RF}^q,\quad p,q\in\{1,\ldots,K\},\forall p\neq q. \tag{52}$$

Since the obtained analog beamforming $\mathbf{F}_{RF}$ is a unitary matrix, i.e., $\mathbf{F}_{RF}\mathbf{F}_{RF}^H=\mathbf{I}_{N_{BS}\times N_{BS}}$, according to the properties of (51), (50) holds up.

Substituting (45) into (50), (50) can be rewritten as

$$\begin{aligned}\lim_{N_{BS}\to\infty}\frac{1}{N_{BS}}\bar{\mathbf{H}}_p\bar{\mathbf{H}}_q^H&=\lim_{N_{BS}\to\infty}\bar{\mathbf{U}}_p\bar{\mathbf{\Sigma}}_p\bar{\mathbf{V}}_p^H\bar{\mathbf{V}}_q\bar{\mathbf{\Sigma}}_q^H\bar{\mathbf{U}}_q^H\\&=\mathbf{0}_{M_{MS}\times M_{MS}}.\end{aligned} \tag{53}$$

Since $\bar{\mathbf{U}}_p$ is a unitary matrix, (53) can be further written as

$$\lim_{N_{BS}\to\infty}\bar{\mathbf{\Sigma}}_p\bar{\mathbf{V}}_p^H\bar{\mathbf{V}}_q\bar{\mathbf{\Sigma}}_q^H=\mathbf{0}_{M_{MS}\times M_{MS}}. \tag{54}$$

Defining the following two partitions of matrices $\bar{\mathbf{\Sigma}}_k$ and $\bar{\mathbf{V}}_k$ as:

$$\bar{\mathbf{\Sigma}}_k=\begin{bmatrix}\bar{\mathbf{\Sigma}}_k^1 & \mathbf{0}\\ \mathbf{0} & \bar{\mathbf{\Sigma}}_k^2\end{bmatrix},\bar{\mathbf{V}}_k=\begin{bmatrix}\bar{\mathbf{V}}_k^1 & \bar{\mathbf{V}}_k^2\end{bmatrix},k=1,\ldots,K, \tag{55}$$

where $\bar{\mathbf{\Sigma}}_k^1=\bar{\mathbf{\Sigma}}_k\left(1:N_S,1:N_S\right)$ and $\bar{\mathbf{V}}_k^1=\bar{\mathbf{V}}_k\left(:,1:N_S\right)$. Substituting (55) into (54), we have

$$\begin{aligned}\lim_{N_{BS}\to\infty}\bar{\mathbf{\Sigma}}_p\bar{\mathbf{V}}_p^H\bar{\mathbf{V}}_q\bar{\mathbf{\Sigma}}_q^H&\approx\lim_{N_{BS}\to\infty}\bar{\mathbf{\Sigma}}_p^1\bar{\mathbf{V}}_p^{1\,H}\bar{\mathbf{V}}_q^1\bar{\mathbf{\Sigma}}_q^{1\,H}\\&=\mathbf{0}_{M_{MS}\times M_{MS}}\end{aligned} \tag{56}$$

Since $N_S\leq\operatorname{rank}\left(\bar{\mathbf{H}}_k\right)$, then $\bar{\mathbf{\Sigma}}_k^1\neq\mathbf{0}_{N_S\times N_S}$. Hence, we can obtain

$$\lim_{N_{BS}\to\infty}\bar{\mathbf{V}}_p^H\left(1:N_S,:\right)\bar{\mathbf{V}}_q\left(:,1:N_S\right)=\mathbf{0}_{N_S\times N_S} \tag{57}$$

i.e., every of the first N_S right singular vectors belonged to two different user equivalent baseband channels are asymptotically mutual orthogonal in massive MIMO regimen.

Using $\mathbf{W}_{BB}^k=\bar{\mathbf{U}}_k\left(:,1:N_S\right),k=1,\ldots,K$, (45), and (47), we can obtain

$$\begin{aligned}\bar{\mathbf{H}}_{comp}&=\begin{bmatrix}\bar{\mathbf{\Sigma}}_1\left(1:N_S,:\right)\bar{\mathbf{V}}_1^H\\ \vdots\\ \bar{\mathbf{\Sigma}}_K\left(1:N_S,:\right)\bar{\mathbf{V}}_K^H\end{bmatrix}\\&=\begin{bmatrix}\tilde{\mathbf{\Sigma}}_1 & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \tilde{\mathbf{\Sigma}}_K\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{V}}_1^H\\ \vdots\\ \tilde{\mathbf{V}}_K^H\end{bmatrix}\\&=\mathbf{I}_{KN_S}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^H,\end{aligned} \tag{58}$$

where $\tilde{\mathbf{\Sigma}}_k=\bar{\mathbf{\Sigma}}_k\left(1:N_S,1:N_S\right)$, $\tilde{\mathbf{V}}_k=\bar{\mathbf{V}}_k\left(:,1:N_S\right)$, $k=1,\ldots,K$. Due to $\tilde{\mathbf{V}}$ is KN_S-order square matrix, and leveraging the conclusion of (57), it can be deduced that $\tilde{\mathbf{V}}$ is the right singular vectors of $\bar{\mathbf{H}}_{comp}$ in massive MIMO regimen, i.e., $\bar{\mathbf{V}}_{comp}\left(:,1:KN_S\right)=\left[\bar{\mathbf{V}}_1\left(:,1:N_S\right),\ldots,\bar{\mathbf{V}}_K\left(:,1:N_S\right)\right]$ as the number of antennas at BS approaches infinity.

To sum up, when $\mathbf{F}_{BB}=\bar{\mathbf{V}}_{comp}\left(:,1:KN_S\right)$, we have

$$\bar{\mathbf{H}}_k\mathbf{F}_{BB}=\bar{\mathbf{U}}_k\tilde{\mathbf{\Sigma}}_k,k=1,\ldots,K. \tag{59}$$

In other words, the data streams of different users can be independently transmitted on subchannel $\bar{\mathbf{H}}_k\tilde{\mathbf{V}}_k$, so as to eliminate inter-interference in baseband channel. In addition, let $\mathbf{W}_{BB}^k=\bar{\mathbf{U}}_k\left(:,1:N_S\right),k=1,\ldots,K$, the inner-interference can also be eliminated. For satisfying the constraint $\left\|\mathbf{F}_{RF}\mathbf{F}_{BB}\right\|_{\mathrm{F}}^2=KN_S$, we adjust each column of $\mathbf{F}_{BB}$ to be $\mathbf{F}_{BB}\left(:,i\right)=\frac{\mathbf{F}_{BB}(:,i)}{\left\|\mathbf{F}_{RF}\mathbf{F}_{BB}(:,i)\right\|_{\mathrm{F}}},\ i\in\{1,\ldots,KN_S\}$. Therefore, all the constraints in (44) are satisfied and the specific procedure is summarized as **Algorithm 2**.

Algorithm 2 : The proposed digital beamforming design scheme

1: **Input**: $\mathbf{H}$, $\mathbf{F}_{RF}$, $\mathbf{W}_{RF}$;
2: **for** $k=1$ to K **do**
3: Compute the baseband channel for each user $\bar{\mathbf{H}}_k=\mathbf{W}_{RF}^{k\,H}\mathbf{H}_k\mathbf{F}_{RF}$;
4: Compute $\mathbf{W}_{BB}^k=\bar{\mathbf{U}}_k\left(:,1:N_S\right)$, where $\bar{\mathbf{U}}_k$ is derived by (39);
5: **end for**
6: Compute $\mathbf{W}_{BB}=\mathrm{blk}\left[\mathbf{W}_{BB}^1,\ldots,\mathbf{W}_{BB}^K\right]$;
7: Obtain the composite channel $\bar{\mathbf{H}}_{comp}=\frac{1}{\sqrt{N_{BS}}}\mathbf{W}_{BB}^H\bar{\mathbf{H}}$;
8: Compute $\mathbf{F}_{BB}=\bar{\mathbf{V}}_{comp}\left(:,1:KN_S\right)$, where $\bar{\mathbf{V}}_{comp}$ is derived by $\bar{\mathbf{H}}_{comp}=\bar{\mathbf{U}}_{comp}\bar{\mathbf{\Sigma}}_{comp}\bar{\mathbf{V}}_{comp}^H$;
9: Adjust each column of $\mathbf{F}_{BB}$ to be $\mathbf{F}_{BB}\left(:,i\right)=\frac{\mathbf{F}_{BB}(:,i)}{\left\|\mathbf{F}_{RF}\mathbf{F}_{BB}(:,i)\right\|_{\mathrm{F}}},i\in\{1,\ldots,KN_S\}$;
10: Obtain the total equivalent baseband channel $\mathbf{H}_{total}=\mathbf{W}_{BB}^H\bar{\mathbf{H}}\mathbf{F}_{BB}$;
11: Compute $\mathbf{P}$ by using waterfilling power allocation of the total equivalent channel $\mathbf{H}_{total}$;
12: **Output**: $\mathbf{F}_{BB}$, $\mathbf{W}_{BB}$, $\mathbf{P}$

Then, we compare the computational complexity of the proposed digital beamforming design method and the baseband BD scheme in [25]. Since both of them need to compute the equivalent baseband channel of each user, we compare computational complexity from the solution procedure of digital beamforming and combining.

It can be found from **Algorithm 2** that the complexity of the proposed scheme except computing the baseband channels mainly comes from steps 4, 7, and 8. In step 4, the equivalent baseband channel $\bar{\mathbf{H}}_k\in\mathbb{C}^{M_{MS}\times M_{BS}}$ of each user performs SVD to obtain the digital combining matrix, and the corresponding complexity is $\mathcal{O}\left(KM_{MS}^2M_{BS}\right)$ [42]. In step 7, the composite matrix $\bar{\mathbf{H}}_{comp}$ is obtained by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

one multiplication of two matrices, hence the complexity is $\mathcal{O}\left(K^2 M_{MS} N_S M_{BS}\right)$. The last one originates from computing the digital beamforming matrix $\mathbf{F}_{BB}$ in step 8. Since this part requires one SVD of the composite matrix $\bar{\mathbf{H}}_{comp}$, the complexity is $\mathcal{O}\left(K^2 N_S^2 M_{BS}\right)$.

To sum up, the overall computational complexity of the proposed digital beamforming design scheme is $\mathcal{O}\left(K^2 N_S^2 M_{BS}\right)$. In contrast, since the baseband BD scheme proposed in [25] performs SVD for the matrix with dimension $(K-1) M_{MS} \times M_{BS}$ to obtain the null space orthogonal basis of each user channel under the same parameter configuration, the corresponding computational complexity is $\mathcal{O}\left(K(K-1)^2 M_{MS}^2 M_{BS}\right)$. In addition, due to the same method is utilized to obtain the digital beamformer and combiner in [26]–[28], the corresponding computational complexity required by each literature for designing digital beamforming is also $\mathcal{O}\left(K(K-1)^2 M_{MS}^2 M_{BS}\right)$. Therefore, the proposed design method enjoys much lower computational complexity compared with baseband BD scheme.

IV. NUMERICAL SIMULATION

To evaluate the performance of the proposed hybrid beamforming design scheme, the corresponding simulation results are presented in this section. All simulation results are obtained by averaging over 1,000 random channel realizations based on MATLAB platform. For simplicity, the propagation environment is modeled as a $N_c = 8$ cluster with $N_p = 10$ rays per cluster. The AoAs/AoDs of all channels are assumed to follow the uniform distribution within $[0, 2\pi]$. In the simulation, we consider that the BS with $N_{BS} = 256$ antennas and $M_{BS} = 16$ RF chains serves $K = 8$ users, where each user is equipped with $N_{MS} = 16$ antennas and $M_{MS} = 2$ RF chains to support $N_S = 2$ data streams simultaneously. The noise variance at each user is $\sigma^2 = 1$, and the SNR is defined as $\frac{P_t}{\sigma^2}$. Furthermore, we set the maximum number of iterations as $N_{\text{iter}} = 7$, and the factor in **Algorithm 1** is set as $\varepsilon = 10^{-6}$ [18].

It is worth noting that we focus on the hybrid beamforming design of massive MIMO system with full-connected structure in the paper. We compare the performance of the proposed scheme and the state-of-the art hybrid beamforming design schemes, which include the least number of RF chains (the least number of RF chains is equal to the number of the transmitted streams) based HySBD scheme [28], the full-digital dirty paper coding (DPC) method [43], the EGT-BD based scheme [25], the iterative based MGLRAM method [26] and HyEB approach [27]. Since the DPC implemented with the iterative waterfilling algorithm has been certified to be capacity-reaching in the broadcast channel, it is used as the performance upper bound of the hybrid ones.

A. PERFORMANCE FOR SUM RATE

Fig. 2 compares the sum rate performance of different beamforming schemes versus SNR when the number of BS antennas is large ($N_{BS} = 256$). In the simulation, the number of iterations is set to 20 in the MGLRAM and HyEB schemes. It can be seen from this figure, since the baseband BD technology leads that the overlap of the row subspace of each user channel matrix becomes significant, the performance of the system decreases as the number of users increases. However, the analog hybrid beamforming design scheme in Algorithm 1 improves the capacity of the equivalent baseband channel, and the digital hybrid beamforming derived in massive MIMO regimen can eliminate both inner-user and inter-user interferences. Therefore, we can observe from Fig. 2 that the proposed hybrid beamforming design scheme is superior to the state-of-the art for hybrid beamforming ones. Meanwhile, the result also verifies the effectiveness of the proposed design scheme in BS with a large number of antennas. In addition, the HySBD is slightly better than MGLRAM with the least number of RF chains.

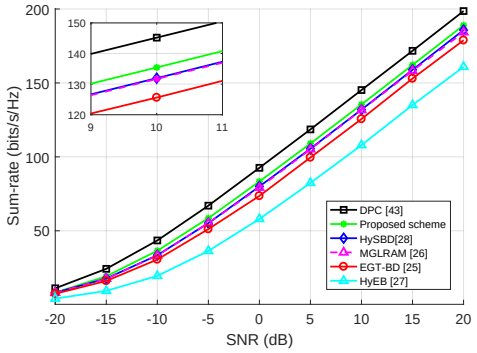

FIGURE 2. Sum rate comparison for different beamforming schemes versus SNR, where $N_{BS} = 256$.

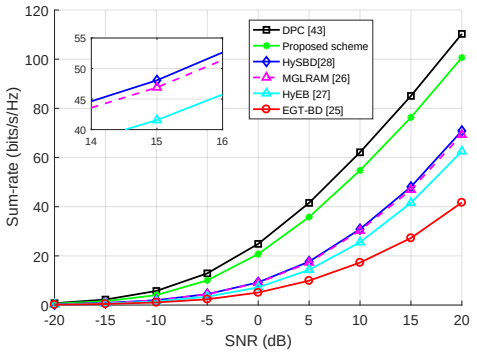

FIGURE 3. Sum rate comparison for different beamforming schemes versus SNR, where $N_{BS} = 32$.

To further investigate the performance of the proposed design scheme in small antenna arrays, Fig. 3 demonstrates the sum rate comparison for different beamforming schemes versus SNR when the number of BS antennas is small ($N_{BS} = 32$). We assume the BS with $M_{BS} = 16$ RF chains, each user with $N_{MS} = 4$ antennas, and the least number of RF chains $M_{MS} = 2$. As can be seen from the figure that although the proposed design scheme is derived from the theoretical analysis for a massive MIMO regimen, it is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

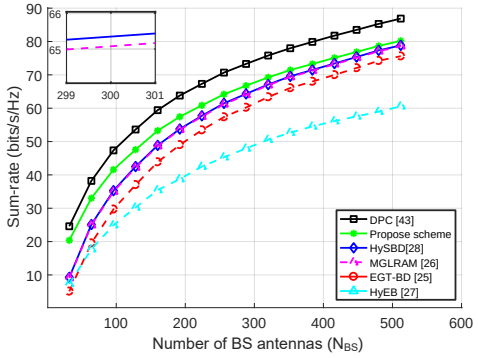

FIGURE 4. Sum rate comparison for different beamforming schemes versus the number of BS antennas, where $\mathrm{SNR} = 0\mathrm{dB}$.

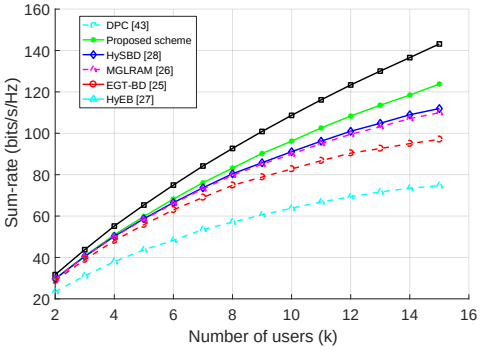

FIGURE 5. Sum rate comparison for different beamforming schemes versus the number of users, where $\mathrm{SNR} = 0\mathrm{dB}$.

outstanding compared with the state-of-the art schemes even if the number of BS antennas is not very large.

B. PERFORMANCE FOR NUMBER OF BS ANTENNAS

Fig. 4 compares the sum rate performance of different beamforming schemes versus the BS antennas, where $\mathrm{SNR} = 0\mathrm{dB}$. As can be seen from this figure, the sum rate performance of different design schemes improves correspondingly as the number of BS antennas increase, where the proposed design scheme is better than others. When the number of BS antennas is large, the performance gap between the proposed beamforming and other hybrid beamforming schemes (except for HyEB) is small. However, compared with the small number of BS antennas, the performance gap between the proposed beamforming scheme and the DPC scheme is larger. Furthermore, although the proposed scheme is derived from the massive MIMO system, it works relatively well even with not a very large number of BS antennas. For illustration, the sum rate of the proposed beamforming scheme with $N_{BS} = 64$ and $N_{BS} = 96$ respectively approaches 86% and 88% of that reached by the DPC scheme, while the sum rate of HySBD with the same settings only approaches about 66% and 74%.

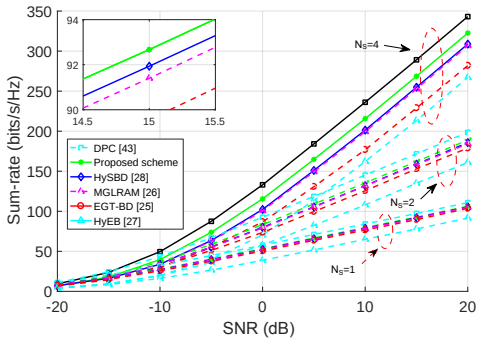

FIGURE 6. Sum rate comparison for different beamforming schemes versus SNR, where the data streams per user is set as $N_S = 1, 2$, and 4.

C. PERFORMANCE FOR NUMBER OF USERS

Fig. 5 compares the sum rate performance of different beamforming schemes versus the number of users, where the number of users changes from 2 to 15, and the transmission SNR is $\mathrm{SNR} = 0\mathrm{dB}$. We can observe from the figure that the gap between different design schemes is very small when the number of users is small (except for HyEB), e.g., $k \leq 3$, and the performance of all schemes is closer to that of the DPC scheme. As the number of users increases, the sum rate performance of different design schemes becomes large, where the proposed design scheme is better than others. Furthermore, it can also be explained that with the increase of the scale of the system, the proposed design scheme effectively eliminates the inner and inter-user interference, so as to improve the performance of system.

D. PERFORMANCE FOR DATA STREAMS PER USER

Fig. 6 plots the sum rates achieved by different beamforming schemes when the number of data streams per user is different, where $N_S = 1, 2$, and 4. Considering the costs and power consumption, the simulation environment of the BS and each user is set as the number of RF chains is equal to the number of data streams, i.e., $M_{MS} = N_S$, and $M_{BS} = KM_{MS}$. We find that the performance of all different beamforming schemes is similar and very close to that of the DPC scheme as the number of data streams per user is small, i.e., $N_S = 1$. When the number of data streams supported by the system increases, the gaps between the sum rates of different schemes become larger correspondingly. However, the proposed hybrid beamforming scheme outperforms other schemes when the number of data streams is different.

E. PERFORMANCE FOR POWER EFFICIENCY

As mentioned in Sec. I, the power consumption is a key issue which deserves our consideration for different hybrid beamforming schemes. In this subsection, we aim to compare the power efficiency performance of different hybrid beamforming design schemes [44]. Considering the hybrid beamforming design based on full-connected structure, the power consumption mainly includes the following parts: a) the power amplifier (PA) connected to each antenna at the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3009238, IEEE Access

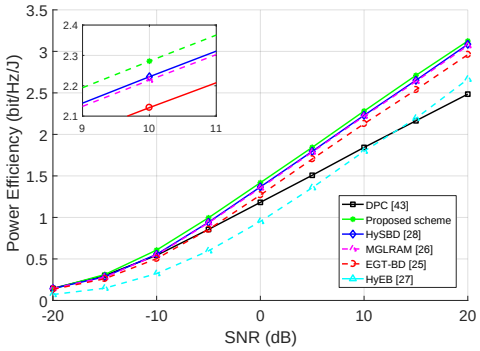

FIGURE 7. Power efficient comparison for different beamforming schemes versus SNR.

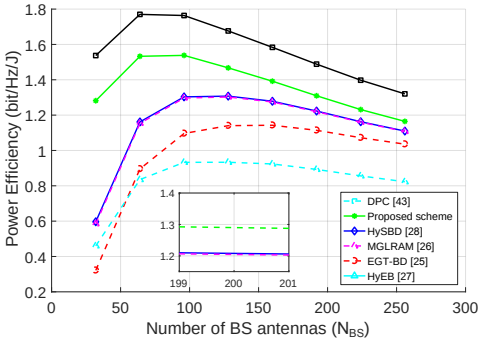

FIGURE 8. Power efficient comparison for different beamforming schemes versus the number of BS antennas, where $\mathrm{SNR} = 0\mathrm{dB}$.

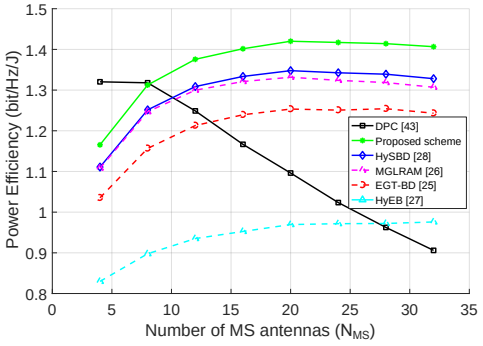

FIGURE 9. Power efficient comparison for different beamforming schemes versus the number of each user antennas, where $\mathrm{SNR} = 0\mathrm{dB}$.

BS; b) the low noise amplifier (LNA) at the receiver; c) the phase shifter (PS) and the RF chain on both receiver and transmitter sides; d) the digital baseband (BB) processor; e) the digital-to-analog converter (DAC) on the receiver side and the ADC on the transmitter side.

Considering the full-digital beamforming for MIMO, there is a DAC, an RF chain, a PS and a PA for each antenna at the BS. Then, a BB which adapts all the data streams to the transmit antennas is required. At the receiver, each antenna is equipped with an ADC, an RF chain, a PS, and an LNA, plus a baseband digital combiner that combines all the outputs of ADC to obtain the soft estimate of the transmitted symbols. Therefore, the amounts of power consumed by BS and users in full-digital MIMO architecture are expressed respectively as

$$P_{DPC\_BS} = N_{BS}\left(P_{RF} + P_{DAC} + P_{PA}\right) + P_{BB}, \quad (60)$$

$$P_{DPC\_MS} = K\left(N_{MS}\left(P_{RF} + P_{ADC} + P_{LNA}\right) + P_{BB}\right), \quad (61)$$

where P_{BB}, P_{RF}, P_{LNA}, P_{PA}, P_{PS}, P_{DAC}, and P_{ADC}, are the power of BB, the power of each RF chain, the power of each LNA, the power of each PA, the power of each PS, the power of each DAC, and the power of each ADC, respectively. Different from the full-digital beamforming for MIMO, each RF chain requires the same number of phase shifters as that of all antennas to control the phase of all antennas in the full-connected structure. Therefore, the amounts of power consumed by BS and users in full-connected MIMO architecture are expressed respectively as

$$P_{TOTAL\_BS} = M_{BS}\left(P_{RF} + P_{DAC} + N_{BS}P_{PS}\right) + N_{BS}P_{PA} + P_{BB}, \quad (62)$$

$$P_{TOTAL\_MS} = K\left(M_{MS}\left(P_{RF} + P_{ADC} + N_{MS}P_{PS}\right) + N_{MS}P_{LNA} + P_{BB}\right). \quad (63)$$

To better compare the performance of different hybrid beamforming design schemes, the power efficiency η is used as the standard of measurement, which is expressed as follows:

$$\eta = \frac{R}{P}(\mathrm{bps/Hz/J}), \quad (64)$$

where P is the total power consumption of the system.

Fig. 7 presents the power efficiency performance of different beamforming design schemes versus SNR. The simulation parameters according to [44] are set as follows: $P_{BB} = 243\mathrm{mW}$, $P_{RF} = 40\mathrm{mW}$, $P_{LNA} = 30\mathrm{mW}$, $P_{PA} = 16\mathrm{mW}$, $P_{DAC} = 110\mathrm{mW}$, and $P_{ADC} = 200\mathrm{mW}$. Furthermore, the power consumed by each PS is assumed to be $10\mathrm{mW}$ as in [45]. It can be seen clearly that the proposed scheme can transmit the signal more efficiently with the same SNR and power consumption, which means it has higher power efficiency. Further, since the full-digital MIMO architecture requires more hardware and produces higher power consumption, its power efficiency performance is relatively low compared with the full-connected architecture. Therefore, the full-digital MIMO architecture is rarely used for signal propagation in practical applications. In addition, as shown in Fig. 2, based on the fact that the sum rates of the system achieved by the HySBD and MGLRAM algorithms are similar, hence the power efficiencies of the two algorithms are approximate with the same power consumption.

Figs. 8 and 9 compare the power efficiency performance of different beamforming schemes versus the number of BS

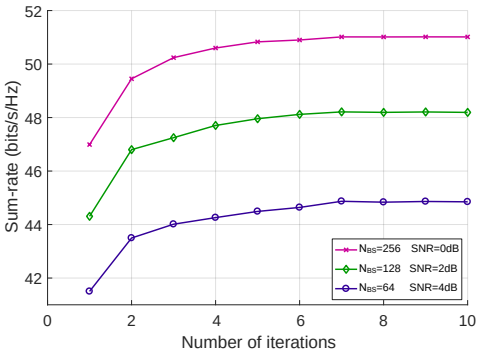

FIGURE 10. Convergence of the proposed design scheme in the analog beamforming stage, where $K = 4$.

and each user antennas, respectively, where $\mathrm{SNR} = 0\mathrm{dB}$. The number of each user antennas is set as $N_{MS} = 4$ for different the number of BS antennas. It can be seen from (60) and (62) that the power consumption of the PS is dominant when the number of BS antennas increases. Thus, under the number of large-scale BS antennas, the power consumption of the full-connected structure is higher than that of the full-digital structure. According to the simulation results shown in Fig.4, the power efficiency of different schemes is lower than that of the full-digital DPC scheme. The number of BS antennas is set as $N_{BS} = 256$ for different the number of each user antennas. It can be seen from (61) and (63) that the power consumption of the DAC is dominant when the number of each user antennas increases. Since the sum rate increases with more power consumption for the full-digital DPC scheme, its power efficiency declines considerably as the number of each user antennas increases. In summary, the proposed hybrid beamforming design scheme has stable and high power efficiency regardless of changing the number of the BS and each user antennas.

F. PERFORMANCE FOR ITERATION

Fig. 10 shows the convergence of the proposed PSIA algorithm in terms of the sum rate versus the increasing of the number of iteration N_{iter}, where $K = 4$. As can be seen from the figure that the proposed PSIA algorithm converges tremendously regardless of the number of BS antennas and the value of SNR. Especially, the sum rate performance of the system reaches 90% of the convergence value after one iteration. In addition, after 7 iterations, the sum rate performance of the system tends to saturate and no longer grows, reaching the maximum value. Therefore, the maximum number of iterations is set as $N_{\mathrm{iter}} = 7$ in all simulations.

V. CONCLUSION

In this paper, we have investigated the hybrid beamforming design of a downlink mmWave massive MU-MIMO system with full-connected structure. A two-stage linear hybrid beamforming design scheme has been proposed to obtain optimal close-form solutions. The criterion of trying to avoid the loss of information has been adopted for the sub-procedure of each communication in both analog and digital stages to approach the performance of full-digital as far as possible. Further, we have solved the initial problem by approximating different optimal targets in the analog and digital stages, respectively, which not only ensured the maximum channel capacity in each stage, but also maximized the overall capacity of system. Finally, the simulation results show that the performance of the proposed hybrid beamforming design scheme outperforms the existing methods regardless of whether BS is equipped with large or small antenna arrays. Since the proposed scheme for the hybrid beamforming system is based on perfect CSI, perspectives of this work include an extension to consider the case of imperfect CSI as in [46], which is more practical in future applications. In addition, considering the huge available bandwidth of mmWave communications, the hybrid beamforming design for massive MU-MIMO orthogonal frequency-division multiplexing (OFDM) systems can also be considered for our future work.

REFERENCES

[1] J. Huang, Y. Liu, C. Wang, J. Sun, and H. Xiao, "5G Millimeter Wave Channel Sounders, Measurements, and Models: Recent Developments and Future Challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 138–145, Jan. 2019.

[2] Y. Chen, D. Chen, T. Jiang, and L. Hanzo, "Millimeter-Wave Massive MIMO Systems Relying on Generalized Sub-Array-Connected Hybrid Precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8940–8950, Sep. 2019.

[3] J. Huang, C.-X. Wang, R. Feng, J. Sun, W. Zhang, and Y. Yang, "Multi-Frequency mmWave Massive MIMO Channel Measurements and Characterization for 5G Wireless Communication Systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1591–1605, July 2017.

[4] Z. Gao, L. Dai, D. Mi, Z. Wang, M. Imran, and M. Shakir, "MmWave massive MIMO based wireless backhaul for 5G ultra-dense network," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 13–21, Oct. 2015.

[5] J. G. Andrews, T. Bai, M. Kulkarni, and A. Alkhateeb, "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[6] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[7] F. Al-Ogaili, R.M. Shubair, "Millimeter-Wave Mobile Communications for 5G: Challenges and Opportunities," in *Proc. 2016 IEEE International Symposium on Antennas and Propagation* (*APSURSI*), Fajardo, Puerto Rico, July 2016, pp. 1003–1004.

[8] X. Liu, X. Li, S. Cao, Q. Deng, R. Ran, N. Kien, and T. Pei, "Hybrid Precoding for Massive mmWave MIMO Systems," *IEEE Access*, vol. 7, pp. 33577–33586, Mar. 2019.

[9] M. Alouzi and F. Chan, "Millimeter Wave Massive MIMO with Alamouti Code and Imperfect Channel State Information," in *Proc. 2018 IEEE 5G World Forum* (*5GWF*), Silicon Valley, CA, USA, Nov. 2018, pp. 507–511.

[10] S. Payami, M. Shariat, M. Ghoraishi, and M. Dianati, "Effective RF codebook design and channel estimation for millimeter wave communication systems," in *Proc. 2015 IEEE Int. Conf. Commun. Workshops* (*ICCW*), London, UK, Jun. 2015, pp. 1226–1231.

[11] O. El.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.

[12] E. Torkildson, C. Sheldon, U. Madhow, and M. Rodwell, "Millimeter-wave spatial multiplexing in an indoor environment," in *Proc. 2009 IEEE Globecom Workshops*, Honolulu, HI, USA, Dec. 2009, pp. 1–6.

[13] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[14] Y. Kabalci and H. Arslan, "Hybrid precoding for mmWave massive MIMO systems with generalized triangular decomposition," in *Proc. 2018 IEEE 19th Wireless and Microwave Technol. Conf.* (*WAMI-CON*), Sand Key, FL, USA, Apr. 2018, pp. 1–6.

[15] L. Zhou and Y. Ohashi, "Efficient codebook-based MIMO beamforming for millimeter-wave WLANs," in *Proc. 2012 IEEE 23rd Int. Symp. Pers., Indoor Mobile Radio Commun.* (*PIMRC*), Sydney, NSW, Australia, Sep. 2012, pp. 1885–1889.

[16] 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, Jun. 2017.

[17] 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.

[18] X. Xue , Y. C. Wang, L. Dai, and C. Masouros, "Relay hybrid precoding design in millimeter-wave massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 66, no. 8, pp. 2011–2016, Apr. 2018.

[19] Y. Zhang, J. Du, Y. Chen, M. Han, and X. Li, "Optimal Hybrid Beamforming Design for Millimeter-Wave Massive Multi-User MIMO Relay Systems," *IEEE Access*, vol. 7, pp. 157212–157225, Oct. 2019.

[20] G. Kwon, Y. Shim, H. Park, and H. M. Kwon, "Design of millimeter wave hybrid beamforming systems," in *Proc. 2014 IEEE 80th Veh. Technol. Conf.* (*VTC2014-Fall*), Vancouver, BC, Canada, Dec. 2014, pp. 1–5.

[21] A. Alkhateeb, O. El.Ayach, G. Leus, and R.W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[22] W. Ni, X. Dong, and W.-S. Lu. (2015), "Near-optimal hybrid processing for massive MIMO systems via matrix decomposition," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 3922–3933, Apr. 2017.

[23] T. E. Bogale and L. B. Le, "Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital," in *Proc. 2014 IEEE Global Commun. Conf.* (*GLOBECOM*), Dec. 2014, pp. 4066–4071.

[24] J. Mao, Z. Gao, Y. Wu and M. Alouini, "Over-Sampling Codebook-Based Hybrid Minimum Sum-Mean-Square-Error Precoding for Millimeter-Wave 3D-MIMO," in *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 938–941, Dec. 2018.

[25] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, Jan. 2016.

[26] N. Song, H. Sun, and T. Yang, "Coordinated hybrid beamforming for millimeter wave multi-user massive MIMO systems," in *Proc. 2016 IEEE Global Commun. Conf.* (*GLOBECOM*), Dec. 2016, pp. 1–6.

[27] C. Hu, J. Liu, X. Liao, Y. Liu, and J. Wang, "A novel equivalent baseband channel of hybrid beamforming in massive multiuser MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 764–767, Apr. 2018.

[28] X. Wu, D. Liu, and F. Yin, "Hybrid Beamforming for Multi-User Massive MIMO Systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3879–3891, Apr. 2018.

[29] F. Khalid, "Hybrid Beamforming for Millimeter Wave Massive Multiuser MIMO Systems Using Regularized Channel Diagonalization," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 705–708, Jun. 2019.

[30] D. Zhang, Y. Wang, X. Li, and W. Xiang, "Hybrid beamforming for downlink multiuser millimetre wave MIMO-OFDM systems," *IET Commun.*, vol. 13, no. 11, pp. 1557–1564, Jul. 2019.

[31] Z. Liu, Y. Chen, M. Yang, and R. Jian, "Hybrid Precoding Based on MMSE-EDS for Multi-user Multi-stream Massive MIMO Systems," in *Proc. 2019 28th Wireless and Opt. Commun. Conf.* (*WOCC*), Beijing, China, May 2019, pp. 1–5.

[32] 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.

[33] R. Rajashekar and L. Hanzo, "Iterative Matrix Decomposition Aided Block Diagonalization for mm-Wave Multiuser MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1372–1384, Mar. 2013.

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

[35] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *Proc. Inf. Theory Appl. Workshop* (*ITA*), San Diego, CA, USA, Feb. 2013, pp. 1–5.

[36] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, 2003.

[37] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, "Energy-efficient hybrid analogue and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Mar. 2016.

[38] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: a unified framework for convex optimization," *IEEE Trans. Signal Process.*, vol. 51, no. 9, pp. 2381–2401, Sep. 2003.

[39] A. Omar, V. Monga, and M. Rangaswamy, "Tractable MIMO beampattern design under constant modulus waveform constraint," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2588–2599, May 2017.

[40] L. U. Choi and R. D. Murch, "A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 20–24, Jan. 2004.

[41] V. Stankovic, and M. Haardt, "Generalized Design of Multi-User MIMO Precoding Matrices," *IEEE Trans. Wireless Commun.*, vol. 7, no. 3, pp. 953–961, Mar. 2008.

[42] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, USA: JHU Press, 2012.

[43] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

[44] S. Buzzi and C. D'Andrea, "Energy Efficiency and Asymptotic Performance Evaluation of Beamforming Structures in Doubly Massive MIMO mmWave Systems," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 2, pp. 385–396, Jun. 2018.

[45] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Englewood Cliffs, NJ, USA: Prentice Hall, 2015.

[46] X. Li, J. Li, Y. Liu, Z. Ding, and A. Nallanathan, "Residual Transceiver Hardware Impairments on Cooperative NOMA Networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 680–695, Oct. 2019.