IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 July 2022, accepted 14 August 2022, date of publication 18 August 2022, date of current version 26 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3199875

RESEARCH ARTICLE

Hybrid Precoding and Combining Strategy for MMSE-Based Rate Balancing in mmWave Multiuser MIMO Systems

WOOHYEONG PARK AND JIHOON CHOI, (Senior Member, IEEE)
School of Electronics and Information Engineering, Korea Aerospace University, Goyang-si, Gyeonggi-do 10540, South Korea

Corresponding author: Jihoon Choi (jihoon@kau.ac.kr).

This work was supported by the National Research Foundation of Korea (NRF) Grant through the Korea Government (MSIT) under Grant 2019R1A2C1006418, Grant 2021R1A4A2001316, and Grant 2022R1F1A1073999.

ABSTRACT In this paper, a new hybrid precoding and combining method is proposed for the downlink of multiuser multiple-input multiple-output (MU-MIMO) millimeter wave (mmWave) channels. The proposed method designs the precoders and combiners for radio frequency (RF) and baseband processing, respectively, based on the minimum mean square error (MMSE) criterion and the rate fairness among users. To design the RF precoder and combiners implemented by phase shifters, a new matrix factorization algorithm is devised by combining the gradient method with the orthogonal projection. Under the total transmit power constraint, the proposed factorization method increases the achievable rate by making the columns of the RF precoder near-orthogonal and growing the Frobenius norm of the baseband precoder. In addition, a new MMSE-based rate balancing algorithm is proposed to design the baseband precoder and combiners in terms of maximizing the minimum user rate. The proposed rate balancing scheme iteratively updates the baseband precoder, the transmit power constraint for the baseband precoder, the baseband combiners, and the weighting vector for rate balancing. Through theoretical analysis, it is shown that the proposed design method has a polynomial complexity order. Numerical simulations present that the proposed matrix factorization method outperforms existing schemes requiring low computational complexity and the proposed rate balancing scheme converges to a stationary point satisfying the total transmit power constraint. Moreover, simulation results in MU-MIMO channels are provided to show that the proposed design scheme performs better than existing hybrid processing schemes while achieving the minimum user rate close to the upper bound of MMSE processing.

INDEX TERMS Hybrid precoding, rate balancing, mmWave communication, MMSE, multiuser MIMO, matrix factorization.

I. INTRODUCTION

To meet the rapidly increasing demand for wireless communication services, the network capacity can be improved by employing advanced physical layer techniques such as massive multiple-input multiple-output (MIMO) [1], enhancing area spectral efficiency using small cells [2], and providing advanced cooperation through cloud radio access networks (C-RANs) [3]. On the other hand, the millimeter wave (mmWave) band from 30 to 300 GHz has been attracting a great attention as a means to fundamentally increase the capacity using more spectrum bands [4], [5], [6], [7]. The standalone mode in 5G New Radio (NR) exploits the mmWave bands in Frequency Range 2 (FR2), and the commercial NR networks adopting the standalone mode has been gradually deployed in recent years [8], [9].

MmWave cellular systems have several obstacles such as the huge path loss and rain attenuation caused by the ten-fold increase of the carrier frequency [10], [11], [12]. Fortunately, mmWave transceivers can be equipped with

The associate editor coordinating the review of this manuscript and approving it for publication was Olutayo O. Oyerinde.

large-scale antennas because the antenna form factor is reduced by virtue of small wavelength. This enables to form highly directional beams that provide significant beamforming gains for compensating for the path loss. Moreover, a spatial multiplexing gain can be achieved by concurrently transmitting multiple data streams. The mmWave system with large-scale antennas requires prohibitive cost and power consumption for fully digital precoding that controls both the magnitude and phase of digital baseband signals, because a dedicated radio frequency (RF) chain is needed for each antenna element. Considering the constraint on the number of RF chains, the two-stage hybrid precoding architecture has been widely investigated as a means for effectively interconnecting a small number of digital data streams to a large number of RF antenna elements [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. Sparse hybrid precoding and combining schemes were developed using the orthogonal matching pursuit algorithm for compressive sensing-based reconstruction in [14], and an adaptive parameter estimation method was proposed for mmWave-specific channel estimation [15]. Moreover, it was shown through a theoretical analysis that the performance of a hybrid precoder can approach that of a fully digital scheme if the number of RF chains is equal to or greater than the number of data streams [16]. The performance of hybrid precoding and/or combining has been improved by employing the alternating minimization-based design schemes [17], [18], [19] and matrix factorization techniques [20], [21]. In addition, design methods for joint hybrid precoding and combining were devised for practical mmWave transceivers with low-resolution phase shifters [22], [23], [24], and the corresponding spectral efficiency was analyzed [25]. In [26], [27], and [28], codebook-based hybrid precoding was studied to reduce the feedback information in practical mmWave systems.

The hybrid precoder and combiner design scheme has been further extended to mmWave multiuser MIMO (MU-MIMO) systems [27], [28], [29], [30], [31], [32], [33], [34]. In [29], the authors derived the lower bound on the achievable rate for single-path channels and developed a low-complexity hybrid precoding algorithm in downlink MU-MIMO systems with analog combining. Joint RF-baseband hybrid precoding was investigated based on the predetermined codebook to reduce the feedback overhead and facilitate hardware implementation in a multiuser multiple-input single-output (MU-MISO) system [27] and a MU-MIMO system with analog combining [28], respectively. The phase-shifting RF precoding can be combined with baseband precoding based on block diagonalization (BD), singular value decomposition (SVD), and regularized channel diagonalization techniques [30], [31], [32]. Also, a minimum mean squared error (MMSE) criterion is employed to design hybrid analog/digital precoders and combiners for MU-MIMO systems [33], [34]. In the downlink of fully digital MU-MIMO systems, the precoder for maximizing the achievable sum rate can be designed under total power or per-antenna power constraints using BD of multiuser channels [35], [36], [37], regularized channel diagonalization [38], [39], generalized channel inversion [40], and weighted MMSE [41]. On the other hand, the rate balancing precoding method has been studied under the MMSE criterion to ensure the rate fairness among users in the downlink MU-MIMO channels [42], [43]. The rate balancing approach was also employed to the design of hybrid precoders and combiners for the mmWave MU-MIMO system based on zero-forcing (ZF) [44].

Motivated by previous work, this paper focuses on the MMSE criterion and the rate balancing for hybrid precoding and combining in mmWave MU-MIMO systems. When the RF precoder and combiners are implemented by phase shifters, we propose a new matrix factorization technique for the design of RF precoder and combiners. By concatenating the designed RF precoder, the original MIMO channels, and the designed RF combiners, we define the effective MU-MIMO channels. From the effective channels, a new MMSE-based rate balancing algorithm is devised that computes the baseband precoder and combiners in the MMSE sense while ensuring rate fairness among users. The main contributions of this paper are summarized as follows.

- We define the optimization problem with respect to the RF and baseband precoders in terms of maximizing the minimum user rate for fairness. Considering the constant-modulus constraints, a new design method for RF precoder and combiners is proposed by combining the matrix factorization technique with orthogonal projection. In the proposed method, the fully digital precoder (or combiner) is decomposed into the RF and baseband precoders (or combiners) by iteratively updating the RF precoder (or combiner) using the gradient method and the orthogonal projecting technique in [45] and [46].
- The effective channels are defined by concatenating the RF precoder, the original MU-MIMO channels, and the RF combiners. Considering the total transmit power constraint and the MMSE-based rate balancing criterion, a new design procedure is devised for the baseband precoder and combiners. The proposed algorithm iteratively adjusts the norm constraint of the baseband precoder, updates the baseband precoder and combiners in the MMSE sense, and controls the target user rates for maximizing the minimum user rate. The proposed algorithm guarantees rate balancing among users conforming to the total transmit power constraint.
- Through theoretical analysis, the complexity order of the proposed algorithms are compared with those of existing hybrid processing methods. In addition, it is shown that the proposed entire procedure for hybrid processing has a polynomial time complexity similar to conventional MMSE-based schemes.
- Through numerical simulations, we verify the convergence of the proposed matrix factorization algorithm and the proposed rate balancing design scheme,

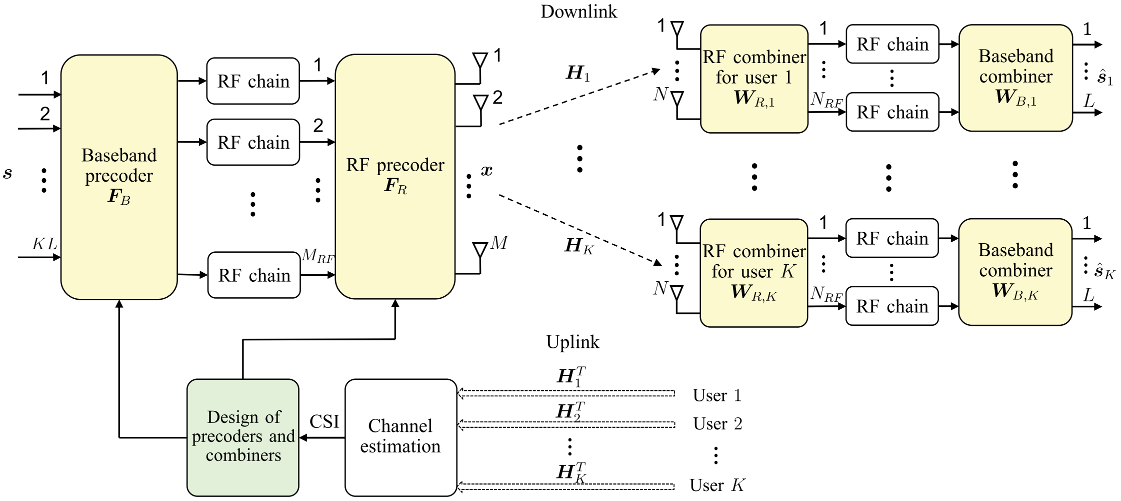

FIGURE 1. MU-MIMO system using hybrid precoding and combining composed of M_{RF} transmit RF chains, M transmit antennas, K users with N antennas each, and N_{RF} receive RF chains.

respectively. Also, simulation results show that the proposed method is advantageous than existing hybrid processing techniques for mmWave MU-MIMO systems while achieving the minimum user rate close to the upper bound. Moreover, under imperfect channel state information (CSI), it is demonstrated that the proposed method is still beneficial over conventional hybrid processing schemes.

The organization of this paper is as follows. In Section II, we introduce the MU-MIMO system with hybrid precoding and combining, and formulates the max-min optimization problem to design hybrid precoders and combiners for the downlink MU-MIMO system. In Section III, the proposed method is derived for jointly designing hybrid precoders and combiners accounting for the MMSE-based rate balancing. Section IV compares the complexity order of the proposed algorithms with those of existing methods and Section V provides numerical simulation results to present the convergence and benefits of the proposed design schemes. Finally, Section VI concludes this article.

Notations: Superscripts T, H, $*$, and -1 denote transposition, Hermitian transposition, complex conjugate, and inversion, respectively, for any scalar, vector, or matrix. $|x|$ means the absolute value of a scalar x; the notations $|\boldsymbol{X}|$, $\|\boldsymbol{X}\|$, and $\|\boldsymbol{X}\|_F$ denote the determinant, ℓ_2-norm, and Frobenius-norm of matrix $\boldsymbol{X}$, respectively; $\boldsymbol{I}_m$ represents an $m \times m$ identity matrix; $\boldsymbol{0}_{m\times n}$ and $\boldsymbol{1}_{m\times n}$ denote the $m \times n$ zero matrix and all-ones matrix, respectively; $\mathrm{tr}(\boldsymbol{A})$ is the trace operation of matrix $\boldsymbol{A}$; $\mathrm{diag}(\boldsymbol{x})$ returns a diagonal matrix whose main diagonal elements are equal to $\boldsymbol{x}$; $\mathrm{blkdiag}(\cdot)$ stands for a block-diagonal matrix with matrices on its diagonal; $\boldsymbol{A}(i, j)$ denotes the ith row and jth column of matrix $\boldsymbol{A}$; $\circ$ and $\otimes$ are Hadamard and Kronecker matrix products; $x \sim \mathcal{CN}(0, \sigma^2)$ means that a random variable x conforms to a complex normal distribution with zero mean and variance σ^2; and $\mathrm{E}[x]$ stands for the expectation value of a random variable x.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we present the system model for the downlink of a MU-MIMO communication link with hybrid precoding and combining, and define the optimization problem to design the hybrid precoders and combiners in terms of maximizing the minimum user rate for rate balancing.

A. MU-MIMO SYSTEM MODEL

Fig. 1 describes the MU-MIMO system for downlink transmission using hybrid precoding at the transmitter with M antennas and M_{RF} RF chains and hybrid combining at the receiver with N antennas and N_{RF} RF chains. All data streams are concurrently transferred to K users through baseband precoding followed by RF precoding. For notational convenience, it is assumed that all users receive the same number of data streams, i.e. L data streams per user, with the same number of receive antennas and RF chains. Notice that it is straightforward to extend the proposed scheme to a MU-MIMO system with an arbitrary number of data streams and receive antennas. The number of RF chains for the transmitter and user receivers satisfy that $KL \le M_{RF} \le M$ and $L \le N_{RF} \le N$, respectively. $\boldsymbol{H}_k \in \mathbb{C}^{N\times M}$ is the downlink channel matrix for user k whose elements represent flat fading channel gains. It is assumed that the CSI for all users, $\{\boldsymbol{H}_k; 1 \le k \le K\}$, are available at the transmitter. For example, when time division duplexing is used, $\{\boldsymbol{H}_k\}$ can be estimated in the uplink from the channel reciprocity as shown in Fig. 1. It is noticeable that the effect of imperfect CSI is evaluated through numerical simulations in Section V.

When a modulated symbol vector $\boldsymbol{s} \in \mathbb{C}^{KL\times 1}$ is transmitted using the baseband precoder $\boldsymbol{F}_B \in \mathbb{C}^{M_{RF}\times KL}$ and the RF precoder $\boldsymbol{F}_R \in \mathbb{C}^{M\times M_{RF}}$, the transmit symbol vector $\boldsymbol{x} \in \mathbb{C}^{M\times 1}$ is given by

$$\boldsymbol{x} = \boldsymbol{F}_R\boldsymbol{F}_B\boldsymbol{s} \tag{1}$$

where the elements of RF precoder $\boldsymbol{F}_R$ have a constant amplitude and adjustable phases, i.e. $\boldsymbol{F}_R(i, j) = \frac{1}{\sqrt{M}}e^{j\theta_{i,j}}$,

$E[\boldsymbol{s}\boldsymbol{s}^H] = \boldsymbol{I}_{KL}$, and $E[\|\boldsymbol{x}\|^2] \leq P$ where P is the maximum transmit power. For hybrid precoding, $\boldsymbol{F}_B$ and $\boldsymbol{F}_R$ are designed to conform to the power constraint $\|\boldsymbol{F}_R\boldsymbol{F}_B\|_F^2 \leq P$. When hybrid combining is used at users, the received signal for user k is expressed as

$$\begin{aligned}\boldsymbol{y}_k &= \boldsymbol{W}_{B,k}^H \boldsymbol{W}_{R,k}^H (\boldsymbol{H}_k \boldsymbol{x} + \boldsymbol{n}_k) \\ &= \boldsymbol{W}_{B,k}^H \boldsymbol{W}_{R,k}^H \boldsymbol{H}_k \boldsymbol{F}_R \boldsymbol{F}_B \boldsymbol{s} + \boldsymbol{W}_{B,k}^H \boldsymbol{W}_{R,k}^H \boldsymbol{n}_k, \qquad (2)\end{aligned}$$

where $\boldsymbol{W}_{B,k} \in \mathbb{C}^{N_{RF} \times L}$ and $\boldsymbol{W}_{R,k} \in \mathbb{C}^{N \times N_{RF}}$ are the baseband and RF combiners for user k, respectively, and $\boldsymbol{n}_k \in \mathbb{C}^{N \times 1}$ is the noise vector composed of independent and identically distributed (i.i.d.) complex Gaussian variables with zero mean and variance σ_k^2, i.e. $\boldsymbol{n}_k \sim \mathcal{CN}(\boldsymbol{0}, \sigma_k^2 \boldsymbol{I}_N)$. Throughout the paper, it is assumed that the elements of RF combiners $\{\boldsymbol{W}_{R,k}\}$ have a constant amplitude and adjustable phases, i.e. $\boldsymbol{W}_{R,k}(i,j) = \frac{1}{\sqrt{N}} e^{j\phi_{i,j}^{(k)}}$.

B. PROBLEM FORMULATION

We design the RF precoder and combiners with a constant amplitude and adjustable phases as well as the baseband precoder and combiners with controllable amplitude and phases. When a Gaussian symbol vector $\boldsymbol{s}$ is transmitted over the downlink MU-MIMO channel with hybrid precoding and combining, the achievable rate for user k is given by [30], [34]

$$\begin{aligned}R_k = \log_2 \Big| &\boldsymbol{I}_L + \boldsymbol{C}_k^{-1} \boldsymbol{W}_{B,k}^H \boldsymbol{W}_{R,k}^H \boldsymbol{H}_k \boldsymbol{F}_R \boldsymbol{F}_{B,k} \\ &\times \boldsymbol{F}_{B,k}^H \boldsymbol{F}_R^H \boldsymbol{H}_k^H \boldsymbol{W}_{R,k} \boldsymbol{W}_{B,k} \Big| \qquad (3)\end{aligned}$$

where $\boldsymbol{C}_k \in \mathbb{C}^{L \times L}$ is defined as

$$\begin{aligned}\boldsymbol{C}_k = \boldsymbol{W}_{B,k}^H \boldsymbol{W}_{R,k}^H \Bigg(&\boldsymbol{H}_k \boldsymbol{F}_R \sum_{m=1, m\neq k}^{K} \boldsymbol{F}_{B,m} \boldsymbol{F}_{B,m}^H \\ &\times \boldsymbol{F}_R^H \boldsymbol{H}_k^H + \sigma_k^2 \boldsymbol{I}_N \Bigg) \boldsymbol{W}_{R,k} \boldsymbol{W}_{B,k}. \qquad (4)\end{aligned}$$

Here, $\boldsymbol{F}_{B,k} \in \mathbb{C}^{M_{RF} \times L}$ is the baseband precoder for transmitting the modulated symbols to user k, i.e. $\boldsymbol{F}_B = \left[\boldsymbol{F}_{B,1}\ \boldsymbol{F}_{B,2}\ \cdots\ \boldsymbol{F}_{B,K}\right]$. This paper focuses on designing the precoders and combiners that maximizes the minimum user rate for rate balancing among users, and thus the optimization problem can be formulated as

$$\max_{\boldsymbol{F}_R, \boldsymbol{F}_B, \{\boldsymbol{W}_{R,k}\}, \{\boldsymbol{W}_{B,k}\}} \quad \min\{R_1, R_2, \cdots, R_K\} \qquad (5a)$$

$$s.t. \quad |\boldsymbol{F}_R(i,j)| = \frac{1}{\sqrt{M}} \quad \text{for } \forall i,j \qquad (5b)$$

$$|\boldsymbol{W}_{R,k}(i,j)| = \frac{1}{\sqrt{N}} \quad \text{for } \forall i,j,k \qquad (5c)$$

$$\|\boldsymbol{F}_R \boldsymbol{F}_B\|_F^2 \leq P. \qquad (5d)$$

Here, the objective R_k is a nonconvex function because it includes $\boldsymbol{C}_k^{-1}$. Moreover, it is more challenging to solve the optimization problem (5) due to the nonconvex constant-modulus constraints for the RF precoder and RF combiners in (5b) and (5c), respectively. Thus, it is difficult to find a globally optimal solution of (5). Instead, in order to develop an optimization method with tractable complexity, we reformulate the rate balancing optimization problem into two separate design problems for RF processing (i.e. RF precoding and combining) and baseband processing (i.e. baseband precoding and combining).

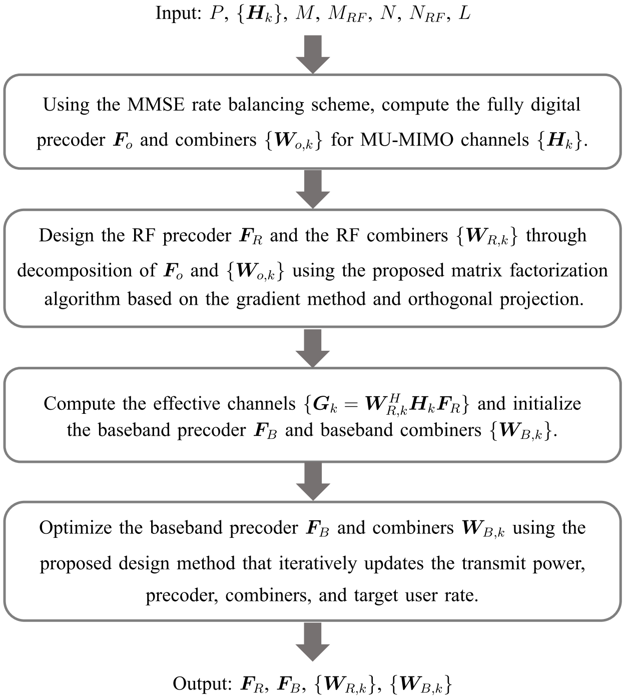
Input: P, $\{\boldsymbol{H}_k\}$, M, M_{RF}, N, N_{RF}, L
Using the MMSE rate balancing scheme, compute the fully digital precoder $\boldsymbol{F}_o$ and combiners $\{\boldsymbol{W}_{o,k}\}$ for MU-MIMO channels $\{\boldsymbol{H}_k\}$.
Design the RF precoder $\boldsymbol{F}_R$ and the RF combiners $\{\boldsymbol{W}_{R,k}\}$ through decomposition of $\boldsymbol{F}_o$ and $\{\boldsymbol{W}_{o,k}\}$ using the proposed matrix factorization algorithm based on the gradient method and orthogonal projection.
Compute the effective channels $\{\boldsymbol{G}_k = \boldsymbol{W}_{R,k}^H \boldsymbol{H}_k \boldsymbol{F}_R\}$ and initialize the baseband precoder $\boldsymbol{F}_B$ and baseband combiners $\{\boldsymbol{W}_{B,k}\}$.
Optimize the baseband precoder $\boldsymbol{F}_B$ and combiners $\boldsymbol{W}_{B,k}$ using the proposed design method that iteratively updates the transmit power, precoder, combiners, and target user rate.
Output: $\boldsymbol{F}_R$, $\boldsymbol{F}_B$, $\{\boldsymbol{W}_{R,k}\}$, $\{\boldsymbol{W}_{B,k}\}$

FIGURE 2. **Overall procedure of the proposed MMSE-based design method for hybrid precoding and combining.**

III. PROPOSED MMSE-BASED HYBRID PROCESSING

In this section, we propose a new algorithm to design the RF precoder and combiners based on the matrix factorization method, and then derive a new MMSE-based rate balancing algorithm to design the baseband precoder and combiners. The overall procedure of the proposed design method is presented in Fig. 2.

A. MATRIX FACTORIZATION FOR DESIGNING RF PRECODER AND COMBINERS

We propose a new matrix factorization method to design the RF precoder and RF combiners for hybrid processing in MU-MIMO systems. Firstly, the fully digital precoder and combiners are obtained by employing the MMSE-based rate balancing technique in [43] that iteratively updates the precoder and combiners using the MSE duality between downlink and uplink. Then, the RF precoder (or combiner) for hybrid processing is determined by factorizing the fully digital precoder (or combiner) in the least squares (LS) sense. For example, given the fully digital precoder $\boldsymbol{F}_o \in \mathbb{C}^{M \times LK}$, the LS matrix factorization problem is formulated as

$$(\boldsymbol{F}_R, \boldsymbol{F}_B) = \arg\min_{\boldsymbol{F}, \boldsymbol{F}_B} \|\boldsymbol{F}_o - \boldsymbol{F}\boldsymbol{F}_B\|_F^2 \qquad (6a)$$

$$s.t. \quad |\boldsymbol{F}(m,n)| = \frac{1}{\sqrt{M}}, \quad \forall m,n \tag{6b}$$

$$\|\boldsymbol{F}\boldsymbol{F}_B\|_F^2 = P. \tag{6c}$$

Since it is obvious that the achievable rate is maximized when the maximum transmit power is used, the constraint in (5d) is replaced with the equality constraint in (6c). Also, when the RF precoder $\boldsymbol{F}$ is fixed, the optimal baseband precoder is given by $\boldsymbol{F}_B = c(\boldsymbol{F}^H\boldsymbol{F})^{-1}\boldsymbol{F}^H\boldsymbol{F}_o$ from the LS solution [17], where c is a scaling factor to meet the transmit power constraint (6c). In [21], it was shown that the power constraint can be removed without loss of local and global optimality. Following the approach in [17] and [21], we temporarily drop the power constraint (6c) and denote the baseband precoder as $\boldsymbol{F}_B = (\boldsymbol{F}^H\boldsymbol{F})^{-1}\boldsymbol{F}^H\boldsymbol{F}_o$ (i.e. $c = 1$). Now, the matrix factorization problem (6a) can be rewritten as

$$\boldsymbol{F}_R = \arg\min_{\boldsymbol{F}} \|\boldsymbol{F}_o - \boldsymbol{F}(\boldsymbol{F}^H\boldsymbol{F})^{-1}\boldsymbol{F}^H\boldsymbol{F}_o\|_F^2 \tag{7a}$$

$$s.t. \quad |\boldsymbol{F}(m,n)| = \frac{1}{\sqrt{M}}, \quad \forall m,n. \tag{7b}$$

Note that the transmit power constraint will be considered in the design of the baseband precoder of Section III-C. To further simplify the optimization problem, define the phase of $\boldsymbol{F}$ be $\widetilde{\boldsymbol{\Phi}} \in \mathbb{R}^{M\times M_{RF}}$, i.e. $\boldsymbol{F}(\widetilde{\boldsymbol{\Phi}}) = \frac{1}{\sqrt{M}}e^{j\widetilde{\boldsymbol{\Phi}}}$. When $\boldsymbol{F}_R$ is an optimal solution of (7a), $\boldsymbol{F}_R\boldsymbol{D}$ is also optimal for $\boldsymbol{D} = \mathrm{diag}(e^{j\theta_1}, e^{j\theta_2}, \cdots, e^{j\theta_{M_{RF}}})$ with arbitrary phases $\{\theta_1, \theta_2, \cdots, \theta_{M_{RF}}\}$. In other words, the optimal phase matrix $\widetilde{\boldsymbol{\Phi}}_R$ corresponding to the optimal RF precoder $\boldsymbol{F}_R$ is not unique. Without loss of optimality, we restrict the first row of $\widetilde{\boldsymbol{\Phi}}$ being a zero vector to obtain a unique solution. Then, we may write

$$\boldsymbol{F}(\widetilde{\boldsymbol{\Phi}}) = \boldsymbol{F}(\boldsymbol{\Phi}) = \frac{1}{\sqrt{M}}\exp\left(j\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{\Phi}\end{bmatrix}\right), \tag{8}$$

and the problem (7) can be reformulated in the following form by substituting $\boldsymbol{F}$ to $\boldsymbol{F}(\boldsymbol{\Phi})$.

$$\boldsymbol{\Phi}_R = \arg\min_{\boldsymbol{\Phi}} f(\boldsymbol{\Phi}) = \|\boldsymbol{F}_o - \boldsymbol{F}(\boldsymbol{\Phi})\boldsymbol{F}^{+}(\boldsymbol{\Phi})\boldsymbol{F}_o\|_F^2, \tag{9}$$

where $\boldsymbol{A}^{+} = (\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H$ is the pseudo-inverse of a matrix $\boldsymbol{A}$. Since the constant-modulus constraint is removed in (9) by employing $\boldsymbol{F}(\boldsymbol{\Phi})$ in (8), the optimal solution $\boldsymbol{\Phi}_R \in \mathbb{R}^{(M-1)\times M_{RF}}$ can be found by solving the unconstrained minimization problem in (9).

In an attempt to develop a low-complexity algorithm for finding the optimal phase matrix $\boldsymbol{\Phi}_R$, we employ the gradient descent method. As stated in [15], it is natural to design the RF precoder or the baseband precoder as an orthogonal matrix, in order to mitigate the transmit power increment in concatenation of $\boldsymbol{F}_R$ and $\boldsymbol{F}_B$ for hybrid precoding. By imposing this constraint to the RF precoder, we insert a matrix projection step that makes the columns of $\boldsymbol{F}(\boldsymbol{\Phi})$ be as close as orthonormal to each other, which is derived from the orthogonal projection technique in [45].

Specifically, let us denote the phase matrix at the ith iteration as $\boldsymbol{\Phi}(i)$. To apply the gradient descent method, we compute the gradient of $f(\boldsymbol{\Phi}(i))$ with respect to $\boldsymbol{\Phi}(i)$ from [21, Appendix B] as follows:

Algorithm 1 Proposed Matrix Factorization Algorithm for the Design of RF Precoder

```
**Input:** $\boldsymbol{F}_o, M, M_{RF}$
Initialize $i = -1$ and each element of $\boldsymbol{\Phi}(0)$ is set to a random phase over $[-\pi, \pi)$.
Compute the initial RF precoder $\boldsymbol{F}(\boldsymbol{\Phi}(0))$ by substituting $\boldsymbol{\Phi}(0)$ into (8).
**repeat**
    $i = i + 1$.
    Calculate $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ using (11).
    Compute the gradient $\nabla_{\boldsymbol{\Phi}} f(\boldsymbol{\Phi}(i))$ from (10).
    Update the phase matrix using the gradient descent method given by (12).
    Obtain $\boldsymbol{F}_p$ via the SVD-based orthogonal projection using (13) and (14).
    Compute the projected phase matrix for the next iteration, $\boldsymbol{\Phi}(i+1)$, using (15).
    Compute $\boldsymbol{F}(\boldsymbol{\Phi}(i+1))$ by substituting $\boldsymbol{\Phi}(i+1)$ into (8), then calculate the cost function $f(\boldsymbol{\Phi}(i+1))$ from the definition in (9).
**until** $|f(\boldsymbol{\Phi}(i+1)) - f(\boldsymbol{\Phi}(i))| < \epsilon_1$, where $\epsilon_1$ is the tolerance for termination.
**Output:** $\boldsymbol{F}_R = \boldsymbol{F}(\boldsymbol{\Phi}(i+1))$.
```

$$\nabla_{\widetilde{\boldsymbol{\Phi}}} f(\widetilde{\boldsymbol{\Phi}}(i)) = 2\,\mathrm{Im}\left((-\boldsymbol{Z}_1(i)\boldsymbol{F}_o\boldsymbol{Z}_2^H(i)) \circ \boldsymbol{F}_o^*\right) \tag{10a}$$

$$\nabla_{\boldsymbol{\Phi}} f(\boldsymbol{\Phi}(i)) = \nabla_{\widetilde{\boldsymbol{\Phi}}} f(\widetilde{\boldsymbol{\Phi}}(i))(2:M,:), \tag{10b}$$

where $\mathrm{Im}(x)$ is the imaginary part of a complex x, $\boldsymbol{A}(m:n,:)$ means the submatrix composed of rows m through n of a matrix $\boldsymbol{A}$, and $\boldsymbol{Z}_1(i)$ and $\boldsymbol{Z}_2(i)$ are given by

$$\boldsymbol{Z}_1(i) = \boldsymbol{I}_M - \boldsymbol{F}(\boldsymbol{\Phi}(i))\boldsymbol{F}^{+}(\boldsymbol{\Phi}(i)) \tag{11a}$$

$$\boldsymbol{Z}_2(i) = \boldsymbol{F}^{+}(\boldsymbol{\Phi}(i))\boldsymbol{F}_o. \tag{11b}$$

Using the gradient in (10b), the phase matrix is updated by the gradient descent method as below:

$$\boldsymbol{\Phi}_0(i+1) = \boldsymbol{\Phi}(i) - \mu_1 \nabla_{\boldsymbol{\Phi}} f(\boldsymbol{\Phi}(i)), \tag{12}$$

where μ_1 is the step-size parameter. As a next step, we conduct the matrix projection. When the RF precoder corresponding to $\boldsymbol{\Phi}_0(i+1)$ has full rank, $\boldsymbol{F}(\boldsymbol{\Phi}_0(i+1))$ is factorized by the reduced singular value decomposition (SVD) as follows:

$$\boldsymbol{F}(\boldsymbol{\Phi}_0(i+1)) = \boldsymbol{U}_p\boldsymbol{\Sigma}_p\boldsymbol{V}_p^H, \tag{13}$$

where $\boldsymbol{U}_p \in \mathbb{C}^{M\times M_{RF}}$ is a complex orthogonal matrix, $\boldsymbol{\Sigma}_p \in \mathbb{R}_+^{M_{RF}\times M_{RF}}$ is a diagonal matrix whose diagonal elements are positive, and $\boldsymbol{V}_p \in \mathbb{C}^{M_{RF}\times M_{RF}}$ is a unitary matrix. From the results in [45, Sec. III-F], the nearest tight frame to $\boldsymbol{F}(\boldsymbol{\Phi}_0(i+1))$ (i.e. a complex orthogonal matrix closest to $\boldsymbol{F}(\boldsymbol{\Phi}_0(i+1))$ in Frobenius norm) can be obtained as

$$\boldsymbol{F}_p = \boldsymbol{U}_p\boldsymbol{V}_p^H. \tag{14}$$

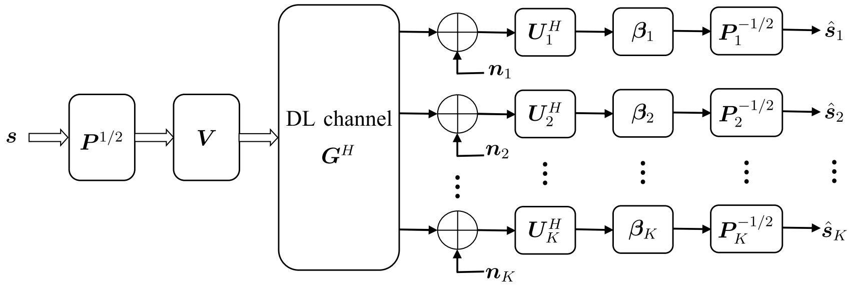

FIGURE 3. DL equivalent channel for designing the baseband precoder and combiners.

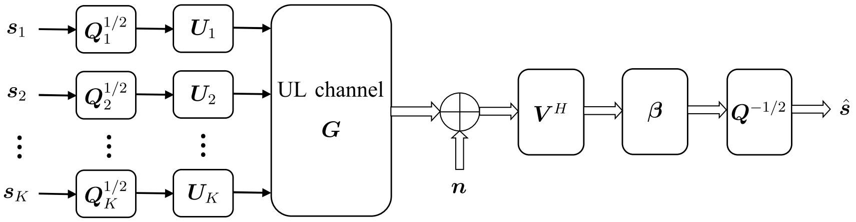

FIGURE 4. UL equivalent channel for designing the baseband precoder and combiners.

Here, the projected precoder satisfies $F_p^H F_p = I_{M_{RF}}$, yet its elements do not have a constant amplitude. To make the RF precoder have a constant amplitude, we update $\Phi(i)$ by only taking the phases of F_p as follows:

$$\Phi(i+1) = \frac{F_p(m,n)}{|F_p(m,n)|}, \quad \text{for } \forall m,n. \tag{15}$$

The proposed matrix factorization algorithm for designing F_R is summarized as Algorithm 1. The matrix factorization problem for RF combiners is identical to that for the RF precoder except that no transmit power constraint is present. As stated in (9), the transmit power constraint is not used in the design of the RF precoder but utilized in the design of the baseband precoder in Section III-C. Therefore, the matrix factorization procedure in (10)–(15) can be applied to the design of RF combiners $\{W_{R,k}\}$ as well.

B. DOWNLINK AND UPLINK EQUIVALENT CHANNELS FOR MSE DUALITY

Using the RF precoder and combiners designed in the previous subsection, we compute the downlink effective channels $\{G_k^H \in \mathbb{C}^{N_{RF} \times M_{RF}}\}$ as below:

$$G_k^H = W_{R,k}^H H_k F_R, \quad \text{for } \forall k. \tag{16}$$

From the effective channels for multiple users, we derive a new algorithm to design the baseband precoder F_B and baseband combiners $\{W_{B,k}\}$ based on the MMSE-based rate balancing criterion. The proposed algorithm exploits the user-wise MSE balancing strategy derived from the MSE duality in [42] and the rate balancing scheme derived from the weighted MSE (WMSE) optimization in [43]. When fully digital precoders and combiners are designed in the MMSE sense via an iterative algorithm, the Frobenius norm of the precoder remains constant during iterations due to the transmit power constraint [43]. In contrast, when hybrid processing is used, the Frobenius norm of the baseband precoder, i.e. $\|F_B\|_F$, is not fixed but varied at every iteration to update the precoder and combiners, because the concatenated precoder is subject to the transmit power constraint in hybrid precoding, i.e. $\|F_R F_B\|_F^2 \leq P$. To take into account this fact, the proposed algorithm iteratively adjusts the Frobenius norm of the baseband precoder so that the concatenated precoder satisfies the transmit power constraint.

Given downlink (DL) and uplink (UL) effective channels, $\{G_k^H\}$ and $\{G_k\}$, respectively, Figs. 3 and 4 present the DL and UL equivalent channels for designing the baseband precoder and combiners. In the DL channel, the modulated symbol vector s is transmitted using the precoder $F = VP^{1/2} \in \mathbb{C}^{M_{RF} \times LK}$, where $V = [V_1, V_2, \cdots, V_K]$ is the DL transmit filter composed of the kth user filtering matrix $V_k \in \mathbb{C}^{M \times L}$, and $P = \text{blkdiag}\{P_1, P_2, \cdots, P_K\}$ is the DL power allocation matrix defined by a diagonal power allocation matrix for user k, $P_k \in \mathbb{R}_+^{L \times L}$. Note that the ith column of V has a unit norm, i.e. $\|v_i\| = 1$. Similarly, the receive combiner for user k is denoted as $W_k^H = P_k^{-1/2} \beta_k U_k^H \in \mathbb{C}^{L \times N_{RF}}$, where $U_k \in \mathbb{C}^{N_{RF} \times L}$ is the receive filter for user k, the diagonal matrix $\beta_k \in \mathbb{R}_+^{L \times L}$ contains scaling factors ensuring that the columns of U have unit norm, i.e. $\|u_i\| = 1$, and $U = \text{blkdiag}\{U_1, U_2, \cdots, U_K\}$. From the total power constraint in (6c), the matrix P meets the following constraint:

$$\|F_R V P^{1/2}\|_F^2 = P. \tag{17}$$

Here, notice that the baseband equivalent transmit power, $\text{tr}(P)$, is adjusted according to F_R and V as explained in Section III-C.

In the UL channel, we switch the role of the transmit and receive filters. Thus, the transmit filter for user k is denoted as $W_k = U_k Q_k^{1/2}$ and the multiuser receive filter is given by

$\mathbf{F}^H = \mathbf{Q}^{-1/2}\boldsymbol{\beta}\mathbf{V}^H$, where $\mathbf{Q} = \text{blkdiag}\{\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_K\}$ is the UL power allocation matrix composed of diagonal power allocation matrices for user k, $\mathbf{Q}_k \in \mathbb{R}_+^{L\times L}$, and $\boldsymbol{\beta} = \text{blkdiag}\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K\}$. In addition, we denote the overall UL channel as $\mathbf{G} = [\mathbf{G}_1\ \mathbf{G}_2\ \cdots\ \mathbf{G}_K]$.

C. DESIGN OF BASEBAND PRECODER AND COMBINERS FOR MMSE-BASED RATE BALANCING

This subsection describes the proposed algorithm to design the baseband precoder and combiners based on the MMSE and rate balancing criteria. When describing the proposed iterative algorithm, we omit the index for iteration to avoid clutter. For notational convenience, it is assumed that the noise variance of a DL receiver is identical for all users and also assumed that the noise variance of a DL receiver is the same as that of the UL receiver,[1] i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2$ and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_{M_{RF}})$.

Given $\mathbf{U}$ and $\mathbf{Q}$, we compute the UL receiver filters $\{\mathbf{V}_k\}$ and scaling matrices $\{\boldsymbol{\beta}_k\}$ for MMSE combining as follows:

$$\mathbf{A}_k = (\mathbf{G}\mathbf{U}\mathbf{Q}\mathbf{U}^H\mathbf{G}^H + \sigma^2\mathbf{I}_{M_{RF}})^{-1}\mathbf{G}^H\mathbf{U}_k\mathbf{Q}_k \tag{18a}$$
$$\boldsymbol{\beta}_k = \text{diag}(b_{k,1}^{UL}, b_{k,2}^{UL}, \cdots, b_{k,L}^{UL}) \tag{18b}$$
$$\mathbf{V}_k = \mathbf{A}_k\boldsymbol{\beta}_k^{-1} \tag{18c}$$

where $b_{k,\ell}^{UL} = \sum_{m=1}^{M_{RF}} |\mathbf{A}_k(m,\ell)|^2$ and $k = 1, 2, \cdots, K$. Using the DL transmit filter $\mathbf{V}$ and receive filter $\mathbf{U}$, the equivalent channels for DL and UL are given by

$$\widetilde{\mathbf{G}}_{DL} = \mathbf{U}^H\mathbf{G}^H\mathbf{V} \tag{19a}$$
$$\widetilde{\mathbf{G}}_{UL} = \widetilde{\mathbf{G}}_{DL}^H = \mathbf{V}^H\mathbf{G}\mathbf{U}, \tag{19b}$$

respectively. In case that an optimal MMSE combiner is used, the UL MSE is computed as

$$\boldsymbol{\epsilon}_{UL} = \text{Re}\{\text{diag}(\boldsymbol{\beta}^2\mathbf{Q}^{-1}\widetilde{\mathbf{G}}_{UL}^H\mathbf{Q}\widetilde{\mathbf{G}}_{UL} - 2\boldsymbol{\beta}\widetilde{\mathbf{G}}_{UL}^H)\} + \sigma^2\,\text{diag}(\boldsymbol{\beta}^2\mathbf{Q}^{-1}) + \mathbf{1}_{LK\times 1}, \tag{20}$$

where $\text{Re}(x)$ is the real part of a complex x and $\mathbf{1}_{LK\times 1}$ means the $LK \times 1$ all-ones vector.

On the other hand, using the MSE duality between the UL and DL, the DL power allocation matrix $\mathbf{P}_0$ is obtained as

$$\boldsymbol{\Psi}_{DL} = \boldsymbol{\beta}^2\boldsymbol{\Psi}^T - 2\boldsymbol{\beta}\,\text{diag}(\text{diag}(\widetilde{\mathbf{G}}_{UL})) + \mathbf{I}_{LK} \tag{21a}$$
$$\mathbf{p} = \sigma^2(\text{diag}(\boldsymbol{\epsilon}_{UL}) - \boldsymbol{\Psi}_{DL})^{-1}\boldsymbol{\beta}^2\mathbf{1}_{LK\times 1} \tag{21b}$$
$$\mathbf{P}_0 = \text{diag}(\mathbf{p}), \tag{21c}$$

where $\boldsymbol{\Psi}(m,n) = |\widetilde{\mathbf{G}}_{UL}(m,n)|^2$, $\text{diag}(\text{diag}(\mathbf{A}))$ means the diagonal matrix composed of the diagonal elements of $\mathbf{A}$, and $\mathbf{I}_{LK}$ is the $LK \times LK$ identity matrix. When fully digital processing is used at the transmitter and receiver based on the MMSE criterion as in [42] and [43], the DL precoding matrix defined as $\mathbf{V}\mathbf{P}_0^{1/2}$ satisfies the transmit power constraint, i.e. $\|\mathbf{V}\mathbf{P}_0^{1/2}\|_F^2 \le P$. In case that hybrid precoding is used, the transmit power constraint is changed to (17), as explained in Section III-B. Thus, the power allocation matrix $\mathbf{P}$ needs to be designed so that the baseband precoder $\mathbf{F}_B = \mathbf{V}\mathbf{P}^{1/2}$ satisfies the transmit power constraint (17). Unfortunately, the power allocation matrix $\mathbf{P}_0$ obtained by (21) does not satisfy the power constraint for hybrid precoding. To tackle this issue, we propose a new procedure to adjust the transmit power of the hybrid precoder by scaling the diagonal matrix $\mathbf{P}$. Specifically, given $\mathbf{V}$ and $\mathbf{P}_0$, the power allocation matrix conforming to (17) is obtained as

$$\mathbf{P}_s = \frac{P}{\|\mathbf{F}_R\mathbf{V}\mathbf{P}_0^{1/2}\|_F^2}\mathbf{P}_0 \tag{22a}$$
$$\mathbf{P} = \mathbf{P}_0 + \mu_2(\mathbf{P}_s - \mathbf{P}_0) \tag{22b}$$

where μ_2 is a step-size parameter to control the speed of power adjustment and $\mathbf{P}_s \in \mathbb{R}^{LK\times LK}$ is a scaled power allocation matrix.

In a similar manner to (18), we can compute the DL receiver filters $\{\mathbf{U}_k\}$ and scaling matrices $\{\boldsymbol{\beta}_k\}$ as follows:

$$\mathbf{B}_k = (\mathbf{G}_k^H\mathbf{V}\mathbf{P}\mathbf{V}^H\mathbf{G}_k + \sigma^2\mathbf{I}_{N_{RF}})^{-1}\mathbf{G}_k^H\mathbf{V}_k\mathbf{P}_k \tag{23a}$$
$$\boldsymbol{\beta}_k = \text{diag}(b_{k,1}^{DL}, b_{k,2}^{DL}, \cdots, b_{k,L}^{DL}) \tag{23b}$$
$$\mathbf{U}_k = \mathbf{B}_k\boldsymbol{\beta}_k^{-1}, \tag{23c}$$

where $b_{k,\ell}^{DL} = \sum_{m=1}^{N_{RF}} |\mathbf{B}_k(m,\ell)|^2$ and $k = 1, 2, \cdots, K$. Again, when an optimal MMSE combiner is used, we get the DL MSE vector for all data streams as

$$\boldsymbol{\epsilon}_{DL} = \text{Re}\{\text{diag}(\boldsymbol{\beta}^2\mathbf{P}^{-1}\widetilde{\mathbf{G}}_{DL}\mathbf{P}\widetilde{\mathbf{G}}_{DL}^H - 2\boldsymbol{\beta}\widetilde{\mathbf{G}}_{DL})\} + \sigma^2\,\text{diag}(\boldsymbol{\beta}^2\mathbf{P}^{-1}) + \mathbf{1}_{LK\times 1}. \tag{24}$$

From the dual expression of (21), the UL power allocation matrix $\mathbf{Q}$ is given by

$$\boldsymbol{\Psi}_{UL} = \boldsymbol{\beta}^2\boldsymbol{\Psi} - 2\boldsymbol{\beta}\,\text{diag}(\text{diag}(\widetilde{\mathbf{G}}_{UL})) + \mathbf{I}_{LK} \tag{25a}$$
$$\mathbf{q} = \sigma^2(\text{diag}(\boldsymbol{\epsilon}_{DL}) - \boldsymbol{\Psi}_{UL})^{-1}\boldsymbol{\beta}^2\mathbf{1}_{LK\times 1} \tag{25b}$$
$$\mathbf{Q} = \text{diag}(\mathbf{q}). \tag{25c}$$

Now, by modifying the user-MSE optimization technique in [42] and [43], we update the UL power allocation $\mathbf{Q}$. Let us define $\xi_k \in \mathbb{R}^{K\times 1}$ be the MSE weight for user k. Considering MSE balancing among users, the weighted UL MSE optimization problem can be formulated as

$$\min_{\mathbf{V},\mathbf{U},\{\boldsymbol{\Omega}_k\}} \quad \max_{1\le k\le K} \frac{\epsilon_{w,k}^{UL}}{\xi_k} \tag{26a}$$
$$s.t. \quad \text{tr}(\mathbf{Q}) \le P_B \tag{26b}$$

where $\epsilon_{w,k}^{UL} = \text{tr}(\boldsymbol{\Omega}_k\mathbf{E}_k^{UL})$, $\boldsymbol{\Omega}_k \in \mathbb{C}^{L\times L}$ is a MSE weight matrix for user k, $P_B = \text{tr}(\mathbf{P})$ is the total transmit power allowed for the baseband precoding in the UL, and $\mathbf{E}_k^{UL}$ is given by

$$\mathbf{E}_k^{UL} = (\mathbf{I}_L - \mathbf{F}_k^H\mathbf{G}_k\mathbf{W}_k)(\mathbf{I} - \mathbf{F}_k^H\mathbf{G}_k\mathbf{W}_k)^H \times \sum_{j=1,j\neq k}^{K} \mathbf{F}_k^H\mathbf{G}_j\mathbf{W}_j\mathbf{W}_j^H\mathbf{G}_j^H\mathbf{F}_k + \sigma^2\mathbf{F}_k^H\mathbf{F}_k \tag{27}$$

[1] It is straightforward to extend the proposed MMSE-based rate balancing algorithm derived in the following subsection for multiuser systems with different noise variance.

where $\boldsymbol{F} = [\boldsymbol{F}_1 \ \boldsymbol{F}_2 \ \cdots \ \boldsymbol{F}_K]$. We decompose $\boldsymbol{Q}_k = \tilde{q}_k \widetilde{\boldsymbol{Q}}_k$, where $\tilde{q}_k = \mathrm{tr}(\boldsymbol{Q}_k)$ is the individual power allocation for user k and $\widetilde{\boldsymbol{Q}}_k$ is the normalized power allocation matrix for user k satisfying $\mathrm{tr}(\widetilde{\boldsymbol{Q}}_k) = 1$. For fixed $\{\widetilde{\boldsymbol{Q}}_k\}$, we adjust $\{\tilde{q}_k\}$ to update the UL power allocation for MSE balancing. By substituting $\boldsymbol{E}_k^{UL}$ in (27) to $\epsilon_{w,k}^{UL} = \mathrm{tr}(\boldsymbol{\Omega}_k \boldsymbol{E}_k^{UL})$, we have

$$\epsilon_{w,k}^{UL} = a_k + \tilde{q}_k^{-1} \sum_{j=1, j\neq k}^{K} \tilde{q}_j b_{kj} + \tilde{q}_k^{-1} c_k \sigma^2 \tag{28}$$

where a_k, b_{kj}, and c_k are given by

$$\begin{aligned} a_k &= \mathrm{tr}(\boldsymbol{\Omega}_k) + \mathrm{tr}(\boldsymbol{\Omega}_k \widetilde{\boldsymbol{F}}_k^H \boldsymbol{G}_k \widetilde{\boldsymbol{W}}_k \widetilde{\boldsymbol{W}}_k^H \boldsymbol{G}_k^H \widetilde{\boldsymbol{F}}_k) \\ &\quad -2\,\mathrm{Re}\left\{\mathrm{tr}(\boldsymbol{\Omega}_k \widetilde{\boldsymbol{F}}_k^H \boldsymbol{G}_k \widetilde{\boldsymbol{W}}_k)\right\} && \text{(29a)} \\ b_{kj} &= \mathrm{tr}(\boldsymbol{\Omega}_k \widetilde{\boldsymbol{F}}_k^H \boldsymbol{G}_j \widetilde{\boldsymbol{W}}_j \widetilde{\boldsymbol{W}}_j^H \boldsymbol{G}_j^H \widetilde{\boldsymbol{F}}_k) && \text{(29b)} \\ c_k &= \mathrm{tr}(\boldsymbol{\Omega}_k \widetilde{\boldsymbol{F}}_k^H \widetilde{\boldsymbol{F}}_k). && \text{(29c)} \end{aligned}$$

Here, $\widetilde{\boldsymbol{F}}_k = \sqrt{\tilde{q}_k} \boldsymbol{F}_k$ and $\widetilde{\boldsymbol{W}}_k = \frac{1}{\sqrt{\tilde{q}_k}} \boldsymbol{W}_k$. Define matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ as

$$\begin{aligned} \boldsymbol{A}(k,j) &= \begin{cases} a_k, & k = j \\ b_{kj}, & k \neq j \end{cases} && \text{(30a)} \\ \boldsymbol{C} &= \mathrm{diag}(c_1, c_2, \cdots, c_K), && \text{(30b)} \end{aligned}$$

and denote $\tilde{\boldsymbol{q}} = [\tilde{q}_1, \tilde{q}_2, \cdots, \tilde{q}_K]^T$. Then, we can rewrite (28) in a vector-matrix form as

$$\boldsymbol{\epsilon}_w^{UL} \tilde{\boldsymbol{q}} = \boldsymbol{A}\tilde{\boldsymbol{q}} + \sigma^2 \boldsymbol{C} \boldsymbol{1}_{K\times 1} \tag{31}$$

where $\boldsymbol{\epsilon}_w^{UL} = \mathrm{diag}(\epsilon_{w,1}^{UL}, \epsilon_{w,2}^{UL}, \cdots, \epsilon_{w,K}^{UL})$. Denote that $\boldsymbol{\xi} = \mathrm{diag}(\xi_1, \xi_2, \cdots, \xi_K)$. By multiplying $\boldsymbol{\xi}^{-1}$ to both sides of (31), we have

$$\Delta^{UL} \tilde{\boldsymbol{q}} = \boldsymbol{\xi}^{-1} \boldsymbol{\epsilon}_w^{UL} \tilde{\boldsymbol{q}} = \boldsymbol{\xi}^{-1} \boldsymbol{A} \tilde{\boldsymbol{q}} + \sigma^2 \boldsymbol{\xi}^{-1} \boldsymbol{C} \boldsymbol{1}_{K\times 1} \tag{32}$$

where Δ^{UL} is a constant at the optimal point of (26). To combine the first and second terms of the right-hand side of (32), we define

$$\tilde{\boldsymbol{q}} = \frac{P_B}{\boldsymbol{1}_{K\times 1}^T \boldsymbol{q}'} \boldsymbol{q}', \tag{33}$$

where $\boldsymbol{q}' \in \mathbb{R}_+^{K\times 1}$ is an unconstrained power allocation vector. By replacing $\tilde{\boldsymbol{q}}$ with $\boldsymbol{q}'$ in (32), we can rewrite

$$\Delta^{UL} \boldsymbol{q}' = \boldsymbol{\xi}^{-1} (\boldsymbol{A} + \frac{\sigma^2}{P_B} \boldsymbol{C} \boldsymbol{1}_{K\times K}) \boldsymbol{q}'. \tag{34}$$

Therefore, the optimal $\boldsymbol{q}'$ is given by the principal eigenvector corresponding to the maximum eigenvalue of $\boldsymbol{\xi}^{-1}(\boldsymbol{A} + \frac{\sigma^2}{P_B} \boldsymbol{C} \boldsymbol{1}_{K\times K})$, and the optimal vector $\tilde{\boldsymbol{q}}^o$ can be obtained by scaling the optimal $\boldsymbol{q}'$ using (33). Also, we update the UL power allocation matrix as

$$\boldsymbol{Q}_k = \tilde{q}_k^o \widetilde{\boldsymbol{Q}}_k, \quad \text{for } \forall k, \tag{35}$$

where $\tilde{q}_k^o$ is the kth element of $\tilde{\boldsymbol{q}}^o$. Note that the normalized power allocation matrix $\widetilde{\boldsymbol{Q}}_k$ is not changed but the individual power allocation $\tilde{q}_k$ is updated for MSE balancing according to (26).

Algorithm 2 Proposed MMSE-Based Rate Balancing Algorithm for Designing Baseband Precoder and Combiners

```
Input: $\{\boldsymbol{G}_k\}, \boldsymbol{F}_R, M_{RF}, N_{RF}, L$
Initialize $\boldsymbol{U}_k^{(0)} = \begin{bmatrix} \boldsymbol{I}_L \\ \boldsymbol{0}_{N_{RF}-L} \end{bmatrix}$, $\boldsymbol{Q}_k^{(0)} = \frac{1}{LK}\boldsymbol{I}_L$, $\boldsymbol{\Omega}_k^{(0)} = \boldsymbol{I}_L$, $\xi_k^{(0)} = L$, $\rho_k^{(0)} = 1$, for all $k$, and $i = 0$.
Compute the initial $\boldsymbol{V}_k^{(0)}$ and $\boldsymbol{P}_0^{(0)}$ using (18)–(21), and the initial transmit power $P_B^{(0)} = \mathrm{tr}(\boldsymbol{P}^{(0)})$ from (22).
repeat
    for $j = 1 : J$ do
        $i = i + 1$.
        Compute the UL receiver filters $\{\boldsymbol{V}_k^{(i)}\}$ and scaling matrices $\{\boldsymbol{\beta}_k^{(i)}\}$ with (18).
        Obtain $\boldsymbol{\epsilon}_{UL}$ from (19) and (20), and calculate the DL power allocation matrix $\boldsymbol{P}_0^{(i)}$ using (21).
        Update the DL power allocation matrix using (22) and adjust the baseband transmit power as $P_B^{(i)} = \mathrm{tr}(\boldsymbol{P}^{(i)})$.
        Update the DL receiver filters $\{\boldsymbol{U}_k^{(i)}\}$ and scaling matrices $\{\boldsymbol{\beta}_k^{(i)}\}$ with (23).
        Obtain $\boldsymbol{\epsilon}_{DL}$ from (24) and calculate the UL power allocation matrix $\boldsymbol{Q}^{(i)}$ using (25).
        Find the optimal UL power allocation for MSE balancing $\tilde{\boldsymbol{q}}^{o,(i)}$ using (29)–(30), the eigendecomposition of $(\boldsymbol{\xi}^{(i-1)})^{-1}(\boldsymbol{A} + \frac{\sigma^2}{P_B^{(i)}}\boldsymbol{C}\boldsymbol{1}_{K\times K})$, and (33).
        Then, update $\boldsymbol{Q}_k^{(i)}$ utilizing (35).
    end for
    Calculate $\boldsymbol{E}_k^{DL,(i)}$ and $R_k^{(i)}$ using (39) and (42).
    Using (43), update the weight for user rate $\rho_k^{(i)}$ and the weight for MSE $\xi_k^{(i)}$, respectively.
until $|g(\{R_k^{(i)}\}) - g(\{R_k^{(i-1)}\})| < \epsilon_2$, where $g(\{x_k\}) = \min(x_1, x_2, \cdots, x_K)$ and $\epsilon_2$ is the tolerance for termination.
Obtain the baseband precoder $\boldsymbol{F}_B = \boldsymbol{V}^{(i)}(\boldsymbol{P}^{(i)})^{1/2}$ and baseband combiners $\boldsymbol{W}_{B,k} = (\boldsymbol{U}_k^{(i)})^H \boldsymbol{\beta}_k^{(i)} (\boldsymbol{P}_k^{(i)})^{-1/2}$ for $k = 1, 2, \cdots, K$.
Output: $\boldsymbol{F}_B, \{\boldsymbol{W}_{B,k}\}$.
```

Finally, we formulate the max-min user rate optimization problem for ensuring rate balancing as follows:

$$\begin{aligned} &\max_{\boldsymbol{V},\boldsymbol{P},\boldsymbol{U},\boldsymbol{\beta}} \ \min_{1\leq k\leq K} \ \frac{R_k}{\rho_k} && \text{(36a)} \\ &s.t. \quad \mathrm{tr}(\boldsymbol{P}) \leq P_B && \text{(36b)} \end{aligned}$$

where ρ_k is a weight for adjusting the achievable rate of user k. By defining the minimum ratio as a scaling factor t, we can write

$$\frac{R_k}{\rho_k} \geq t = \min\left(\frac{R_1}{\rho_1}, \frac{R_2}{\rho_2}, \cdots, \frac{R_K}{\rho_K}\right). \tag{37}$$

From [43, Lemma 1], the DL achievable rate for user k can be expressed as

$$R_k = \log_2 |\boldsymbol{\Omega}_k| - \mathrm{tr}(\boldsymbol{\Omega}_k \boldsymbol{E}_k^{DL}) + L, \tag{38}$$

where $\boldsymbol{E}_k^{DL} = E[(\hat{\boldsymbol{s}}_k - \boldsymbol{s}_k)(\hat{\boldsymbol{s}}_k - \boldsymbol{s}_k)^H]$ is the DL MSE matrix given by

$$\boldsymbol{E}_k^{DL} = (\boldsymbol{I}_L - \boldsymbol{W}_k^H \boldsymbol{G}_k^H \boldsymbol{F}_k)(\boldsymbol{I}_L - \boldsymbol{W}_k^H \boldsymbol{G}_k^H \boldsymbol{F}_k)^H + \boldsymbol{W}_k^H \boldsymbol{G}_k^H \sum_{j=1, j\neq k}^{K} \boldsymbol{F}_j \boldsymbol{F}_j^H \boldsymbol{G}_k \boldsymbol{W}_k + \sigma^2 \boldsymbol{W}_k^H \boldsymbol{W}_k. \tag{39}$$

By substituting (38) into (37), we can obtain

$$\log_2 |\boldsymbol{\Omega}_k| - \mathrm{tr}(\boldsymbol{\Omega}_k \boldsymbol{E}_k^{DL}) + L \geq t\rho_k, \tag{40}$$

and by manipulating both sides of (40) and using $\epsilon_{w,k}^{DL} = \mathrm{tr}(\boldsymbol{\Omega}_k \boldsymbol{E}_k^{DL})$, we have

$$\frac{\mathrm{tr}(\boldsymbol{\Omega}_k \boldsymbol{E}_k^{DL})}{\log_2 |\boldsymbol{\Omega}_k| + L - t\rho_k} = \frac{\epsilon_{w,k}^{DL}}{\xi_k} \leq 1, \tag{41}$$

where $\xi_k = \log_2 |\boldsymbol{\Omega}_k| + L - \tilde{R}_k$ is the MSE weight and $\tilde{R}_k = t\rho_k$ is the target rate for user k (i.e. $R_k \geq \tilde{R}_k$). Because the optimal MSE weight matrix is given by $\boldsymbol{\Omega}_k = (\boldsymbol{E}_k^{DL})^{-1}$, the maximum DL user rate is computed from (38) as follows:

$$R_k = -\log_2 |\boldsymbol{E}_k^{DL}|, \tag{42}$$

and the variables for rate balancing can be updated as below:

$$t = \min(\frac{R_1}{\rho_1}, \frac{R_2}{\rho_2}, \cdots, \frac{R_K}{\rho_K}) \tag{43a}$$

$$\rho_k = t\rho_k \tag{43b}$$

$$\xi_k = R_k - \rho_k + L. \tag{43c}$$

The overall design procedure for baseband processing is summarized as Algorithm 2. As mentioned before, the concatenated hybrid precoder needs to meet the transmit power constraint in (17), so the initial transmit power $P_B^{(0)}$ is computed using (22). Notice that the proposed algorithm iteratively adjusts the baseband transmit power at ith iteration $P_B^{(i)}$ with (22) whenever $\boldsymbol{V}$ and $\boldsymbol{P}_0$ are changed, whereas the MSE balancing method in [42] and rate balancing scheme in [43] update the precoder and combiners under a fixed transmit power constraint. The convergence of Algorithm 2 will be shown through numerical simulations in Section V-A.

IV. COMPLEXITY ANALYSIS

This section compares the time complexity of the proposed algorithms with existing schemes. Firstly, Table 1 presents the complexity order for various matrix factorization methods including the proposed Algorithm 1. Here, J_1 is the number of iterations for each factorization method to satisfy the termination condition. The complexity order of the gradient projection (GP) method is identical to the alternating optimization (AO) scheme with $O(M^2 M_{RF})$. However, as shown in [18], the GP method requires less computational complexity than the AO scheme in numerical runtime simulations, because the AO approach necessitates more complicated procedures for updating the RF precoder. The proposed matrix factorization technique has the same complexity order as the GP method, yet requires slightly more computational load for orthogonal projection via SVD. When M_{RF} is proportional to M, the BFGS scheme has the highest complexity order. It is noticeable that the complexity order is identical to all algorithms when M_{RF} is fixed. In Section V-B, it is demonstrated that the runtime of the BFGS scheme is comparable to that of the AO through numerical simulations with fixed M_{RF}.

TABLE 1. Complexity of various matrix factorization algorithms.

AO [17]	GP [18]	BFGS [21]	Proposed (Alg. 1)
Stepsize: $O(MM_{RF}^2)$	Descent direction: $O(MM_{RF}^2)$	Descent direction: $O(M^2M_{RF}^2)$	$\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$: $O(M^2M_{RF})$
Riemann gradient: $O(MM_{RF}^2)$	Gradient update: $O(MM_{RF})$	Stepsize: $O(M^2M_{RF})$	Gradient update: $O(M^2M_{RF})$
Conj. direction: $O(MM_{RF}^2)$	Update $\boldsymbol{F}_B$: $O(M^2M_{RF})$	Update $\boldsymbol{\Phi}_n$: $O(MM_{RF})$	Projection: $O(M^2M_{RF})$
Update $\boldsymbol{F}_B$: $O(M^2M_{RF})$	–	Update $\boldsymbol{B}_n$: $O(M^2M_{RF}^2)$	Update $F(\boldsymbol{\Phi})$: $O(M^2M_{RF})$
Overall: $O(J_1M^2M_{RF})$	Overall: $O(J_1M^2M_{RF})$	Overall: $O(J_1M^2M_{RF}^2)$	Overall: $O(J_1M^2M_{RF})$

Table 2 compares the complexity order of the proposed Algorithm 2 with those of existing design methods for the baseband precoder and combiners. Here, J_2 is the number of iterations for optimizing the power allocation in the ZF-based sum rate maximization (ZF-SRM) [35] and the ZF-based rate balancing (ZF-RB) [44], while it is the number of iterations for adjusting MMSE-based filters and RB-based power allocation in the MMSE-based design scheme [43] and the proposed Algorithm 2. $N_{ns} = M_{RF} - (K-1)L$ denotes the dimension of the null space obtained by BD of effective channels in ZF-based techniques. Whereas ZF-based schemes calculate the BD procedure of multiuser channels only once, the MMSE hybrid method and the proposed algorithm iteratively computes the DL and UL filters in combination with power adjustment. In general, it holds that $J_2 M_{RF} \gg N_{ns}K$ and $M_{RF} > LK$, and thus the MMSE hybrid method and the proposed algorithm require more computational complexity than the ZF-SRM and ZF-RB schemes. We compare the runtime of overall hybrid processing algorithms in Section V-C.

TABLE 2. Complexity of design methods for baseband precoder and combiners.

ZF-SRM [35]	ZF-RB [44]	MMSE hybrid [43]	Proposed (Alg. 2)
BD of channels: $O(M_{RF}^2N_{ns}K)$	BD of channels: $O(M_{RF}^2N_{ns}K)$	UL filters: $O(M_{RF}^3)$	UL filters: $O(M_{RF}^3)$
Power allocation: $O(LK)$	Numerical gradient: $O(LK^2)$	DL power alloc.: $O(M_{RF}L^2K^2)$	Power adjustment: $O(M_{RF}^2LK)$
Rate computation: $O(LK)$	Power adjustment: $O(LK)$	DL filters: $O(M_{RF}^3)$	DL filters: $O(M_{RF}^2LK)$
–	–	UL power alloc.: $O(M_{RF}L^2K^2)$	UL power alloc.: $O(L^3K^3)$
		Power optimization: $O(M_{RF}^2LK)$	Rate balancing: $O(M_{RF}^2LK)$
Overall: $O(M_{RF}^2N_{ns}K)$ $+O(J_2LK)$	Overall: $O(M_{RF}^2N_{ns}K)$ $+O(J_2LK^2)$	Overall: $O(J_2M_{RF}^3)$	Overall: $O(J_2M_{RF}^3)$

V. SIMULATION RESULTS

Through numerical simulations, we present the convergence of the proposed algorithms in Section V-A, and the proposed Algorithm 1 is compared to conventional matrix factorization methods in Section V-B. In addition, the performance of the proposed hybrid processing with Algorithms 1 and 2 is compared with those of existing hybrid processing schemes under the perfect CSI and imperfect CSI, respectively. The baseline schemes considered in the simulations are as follows:

- *Fully digital MMSE processing [43]*: the rate balancing technique in [43] is used to design the fully digital MMSE precoder and combiners for MU-MIMO systems. This method denotes the performance upper bound of MMSE-based precoding and combining in terms of maximizing the minimum user rate.
- *Proposed MMSE-based hybrid method*: the hybrid precoders and combiners are designed according to the proposed algorithms in Section III. The RF precoder $\boldsymbol{F}_R$ and combiners $\{\boldsymbol{W}_{R,k}\}$ are determined by Algorithm 1, and the baseband precoder $\boldsymbol{F}_B$ and combiners $\{\boldsymbol{W}_{B,k}\}$ are obtained by Algorithm 2.
- *ZF-RB hybrid method [18], [44]*: the RF precoder and combiners are designed by the matrix factorization method based on the GP method [18]. The baseband precoder and combiners are obtained by combining the BD technique with the power allocation method for rate balancing in [44].
- *MMSE hybrid method [17], [43]*: the RF precoder and combiners are designed by the matrix factorization method based on the AO method [17]. The baseband precoder and combiners are designed by the MMSE-based iterative algorithm in [43] and then scaled by multiplying a constant to meet the transmit power constraint.
- *Corr.-based MMSE hybrid method [34]*: following the approach in [34], the RF precoder and combiners are jointly constructed by sequentially selecting the beamformer and combiner with the maximum correlation from predetermined codebooks. Algorithm 2 is used to design the baseband precoder and combiners.
- *ZF-SRM hybrid method [18], [35]*: the RF precoder and combiners are designed by the matrix factorization method based on the GP method [18]. The baseband precoder and combiners are obtained by the BD technique with the water-filling algorithm for sum-rate maximization in [35].
- *Random RF processing*: the RF precoder and combiners are defined as random matrices whose elements have a constant magnitude and random phases, respectively. Algorithm 2 is used to design the baseband precoder and combiners. This scheme presents the performance lower bound of RF precoding and combining.

The following parameters are commonly used in numerical simulations unless otherwise stated: $M_{RF} = KN_{RF}$ for the transmitter; $N = 8$, $N_{RF} = 3$ and $L = 2$ for receivers; $\mu_1 = \frac{3\sqrt{M}}{2}$ and $\epsilon_1 = 0.01$ for Algorithm 1; $J = 50$, $\mu_2 = 0.99$, and $\epsilon_2 = 0.005$ for Algorithm 2; and the mmWave MU-MIMO channels $\{\boldsymbol{H}_k\}$ are generated by the Saleh-Valenzuela channel model as in [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], and [34]. We used the parameters identical to those in [44] for the ZF-RB and ZF-SRM hybrid methods. Using the design strategy in [14], we constructed two codebooks with 128 quantized phase shifting vectors, respectively, to determine the RF precoder and combiners for the corr.-based MMSE hybrid method as in [34].

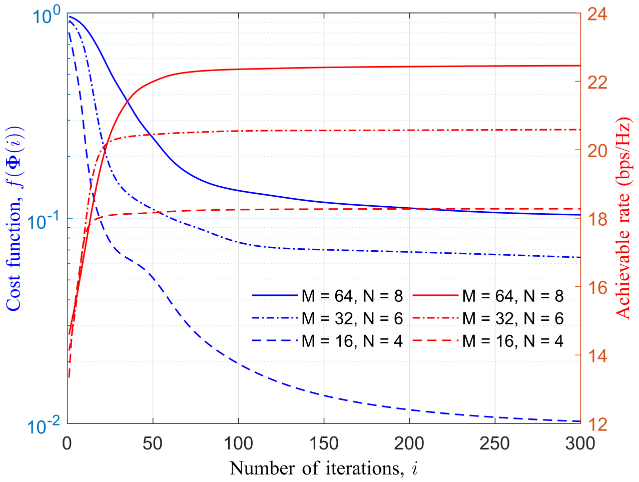

FIGURE 5. Convergence characteristics of the proposed matrix factorization method (Algorithm 1) when $K = 1$ and SNR = 20 dB.

To generate the mmWave MU-MIMO channels, we set the regarding parameters as follows: the carrier frequency is 28 GHz; the number of clusters is 3; the number of subpaths per cluster is 8; the angle-of-departure (AoD) and angle-of-arrival (AoA) for each cluster are uniformly distributed from $-\pi$ to π in the azimuth direction and from -0.5π to 0.5π in the elevation direction, respectively; the subpath angular spread is set to $\pi/64$ and $\pi/16$ for the transmitter and receiver, respectively, by assuming it is identical for azimuth and elevation directions; and the inter-element spacing is equal to half wavelength for both the transmitter and receiver. Considering the distance variation from the transmitter to the receiver in MU-MIMO systems, the average channel gains are set asymmetrically with 10 dB deviation. Specifically, we set $E[\|\boldsymbol{H}_K\|_F^2] = 0.1\ E[\|\boldsymbol{H}_1\|_F^2]$, and $E[\|\boldsymbol{H}_k\|_F^2] = \zeta_k E[\|\boldsymbol{H}_1\|_F^2]$ for $2 \leq k \leq K-1$ where ζ_k is a random variable with uniform distribution in the range of (0.1, 1.0). The nominal signal-to-noise ratio (SNR) is defined as the total transmit power over the noise variance, i.e. P/σ^2. The convergence behaviors in Figs. 5–7 present numerical results obtained from an instantaneous channel realization, while the results in Figs. 8–14 are obtained by averaging the minimum user rate or runtime over more than 100 independent channel realizations.

A. CONVERGENCE OF PROPOSED ALGORITHMS

This subsection verifies the convergence of Algorithms 1 and 2 through numerical simulations, when the CSI is perfectly

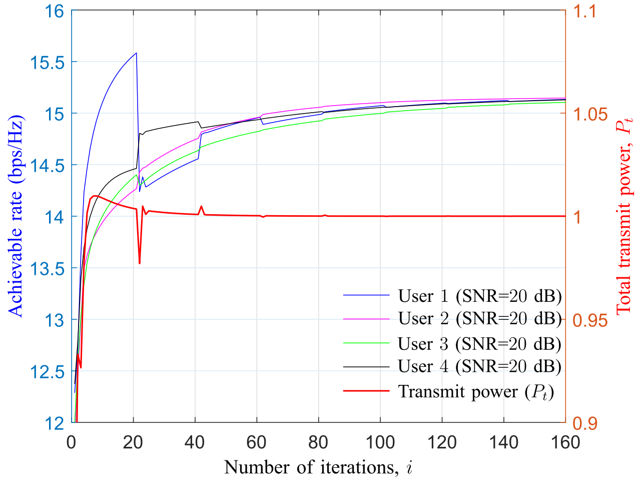

FIGURE 6. Convergence behavior of the proposed rate balancing algorithm (Algorithm 2) when $P = 1$, $K = 4$, $M = 32$, and SNR = 20 dB.

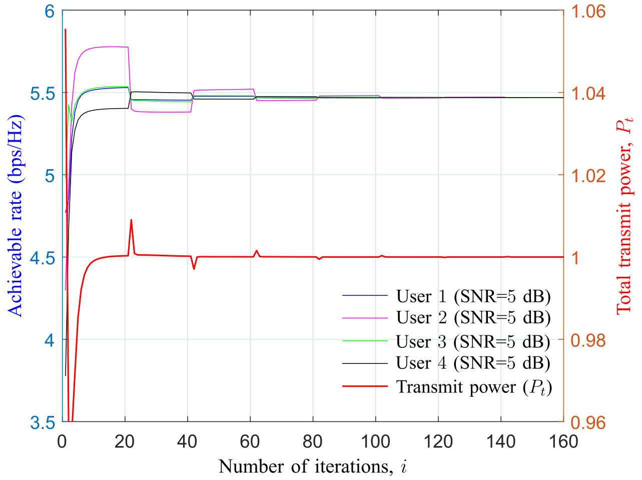

FIGURE 7. Convergence behavior of the proposed rate balancing algorithm (Algorithm 2) when $P = 1$, $K = 4$, $M = 32$, and SNR = 5 dB.

known to the transmitter. Fig. 5 shows the convergence behavior of the proposed matrix factorization scheme described as Algorithm 1, when $K = 1$, $M_{RF} = 3$, and SNR = 20 dB. Blue curves mean the cost function $f(\mathbf{\Phi}(i))$ defined in (9) representing the squared Frobenius norm of the error matrix, and red curves denote the achievable rate obtained by the factorized hybrid precoding matrices when the receiver uses the optimal fully digital combining matrix. For all cases, as the number of iterations increases, the cost function gradually decreases while the achievable rate rapidly grows. Specifically, the cost function $f(\mathbf{\Phi}(i))$ converges to a steady state after about 300 iterations for all antenna configurations, and the steady-state value increases with the increment of the number of transmit and receive antennas because the number of elements in $\boldsymbol{F}_o$ is proportional to the number of transmit antennas M. The achievable rate converges faster than the cost function so that the achievable rate reaches a near-peak value after about 100 iterations regardless of the number of antennas.

To show the convergence of the rate balancing algorithm summarized as Algorithm 2, we present the change of user rates according to the number of iterations in Figs. 6 and 7, when $P = 1$, $K = 4$, $M = 32$, and SNR = 20 dB or SNR = 5 dB. Here, the RF precoder and combiners were designed by decomposing the fully digital precoder $\boldsymbol{F}_o$ and the fully digital combiners $\{\boldsymbol{W}_{o,k}\}$ using the proposed matrix factorization algorithm, respectively. At every iteration, the red curve represents the instantaneous total transmit power defined as $P_t = \|\boldsymbol{F}_R\boldsymbol{F}_B\|_F^2 = \|\boldsymbol{F}_R\boldsymbol{V}^{(i)}(\boldsymbol{P}^{(i)})^{1/2}\|_F^2$ and the other curves denote the achievable rates of four users, respectively. The instantaneous transmit power P_t converges to the maximum transmit power $P = 1$ as the number of iterations increases. Whereas huge rate variations appear among users during the initial transient period, user rates gradually converge to a common steady-state value after 150 iterations in Fig. 6 and 100 iterations in Fig. 7, respectively. In general, the achievable rates of users tend to converge faster in the low SNR region than in the high SNR regime, because the steady-state user rate is lower in the low SNR region.

B. PERFORMANCE COMPARISON OF MATRIX FACTORIZATION METHODS

Various matrix factorization techniques are evaluated to design the RF precoder and combiners in terms of the minimum user rate and runtime. Specifically, the proposed Algorithm 1 is compared to existing matrix factorization methods such as the AO algorithm [17], the gradient method [18], and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [21]. Algorithm 2 was commonly utilized to design the baseband precoder and combiners. Fig. 8 compares the minimum user rate according to SNR, when $K = 4$ and $M = 32$. The fully MMSE processing denotes the performance upper bound achieved by the fully digital precoding and combining in [43]. In combination with a proper matrix factorization algorithm for RF processing, the baseband precoder and combiners are designed using the Proposed Algorithm 2 in the proposed hybrid method and the MMSE-based iterative scheme in the MMSE hybrid method, respectively. In the proposed hybrid method, the proposed Algorithm 1 outperforms the conventional factorization schemes such as the AO algorithm, the GP method, and the BFGS algorithm, while achieving the minimum user rate comparable to the fully digital MMSE processing denoting the upper bound. On the other hand, the AO algorithm obtains the highest minimum user rate in the MMSE hybrid method. When a ZF-based method is used to design the baseband precoder and combiners, the inter-user interference is completely removed and the performance is not so sensitive to the matrix factorization method but the power allocation scheme for ensuring rate balancing among users. For this reason, we use the GP method with the lowest complexity for the ZF-based baseline schemes such as the ZF-RB and ZF-SRM hybrid methods.

To compare the computational complexity of various matrix factorization methods, we present the average runtime across the number of transmit antennas when $K = 4$ and

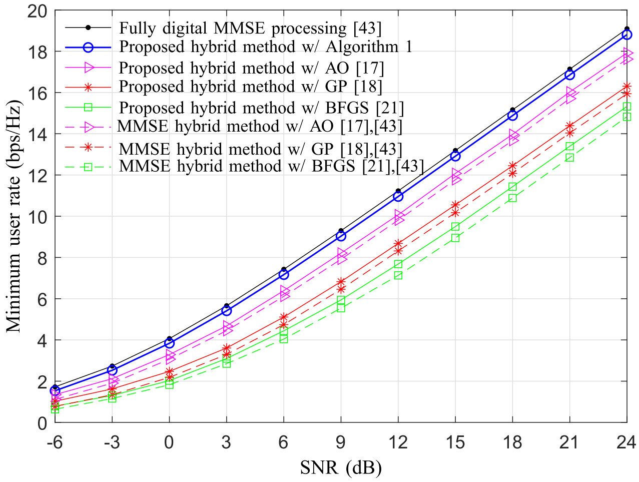

FIGURE 8. SNR versus minimum user rate for various matrix factorization methods when $K = 4$, $M = 32$, and MMSE-based schemes are used to design the baseband precoder and combiners.

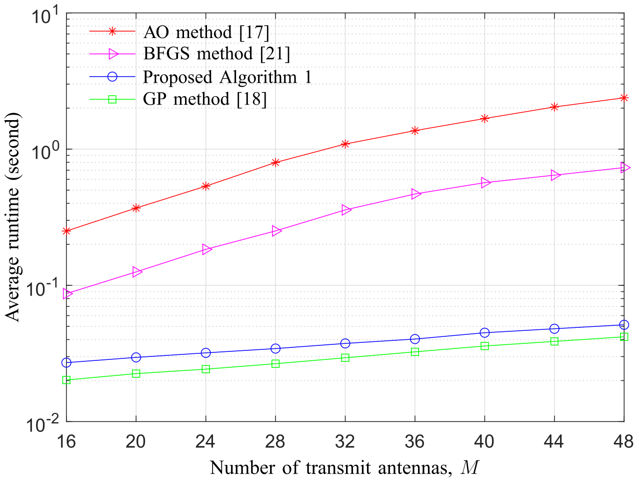

FIGURE 9. Number of transmit antennas versus average runtime for various matrix factorization methods when $K = 4$ and SNR = 20 dB.

SNR = 20 dB. The average runtime was measured using a software implemented by MATLAB R2022a and a server with i7-12700 4.9 GHz CPU, 16 GB RAM, and 64-bit operating system, and every point was obtained by averaging the execution time over more than 100 independent channel realizations. Since $M_{RF} = KN_{RF} = 12$ irrespective of the number of transmit antennas, all matrix factorization algorithms have the complexity order $O(M^2)$ from Table 1. In Fig. 9, the GP method requires the lowest runtime regardless of the number of transmit antennas. As mentioned in Section IV, the proposed matrix factorization algorithm necessitates slightly more computations for orthogonal projection compared to the GP method, and thus Algorithm 1 has slightly larger runtime than the GP method. It is noticeable that the proposed method achieves at least 4 dB SNR gain compared to the GP method in Fig. 8. Moreover, because the AO and BFGS methods require more complicated procedures for updating the RF precoder than Algorithm 1, the runtime of the proposed method is just 2.2 ~ 10.8% and 7.0 ~ 31.1% compared to those of the AO and BFGS algorithms, respectively.

C. PERFORMANCE EVALUATION UNDER PERFECT CSI

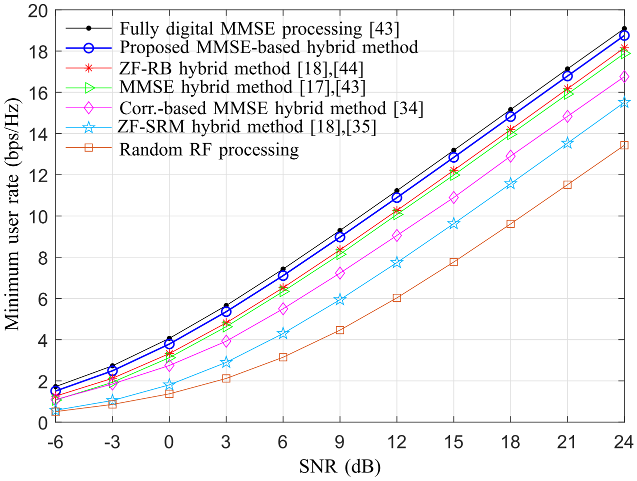

FIGURE 10. Minimum user rate of various hybrid processing schemes according to SNR when $K = 4$ and $M = 32$.

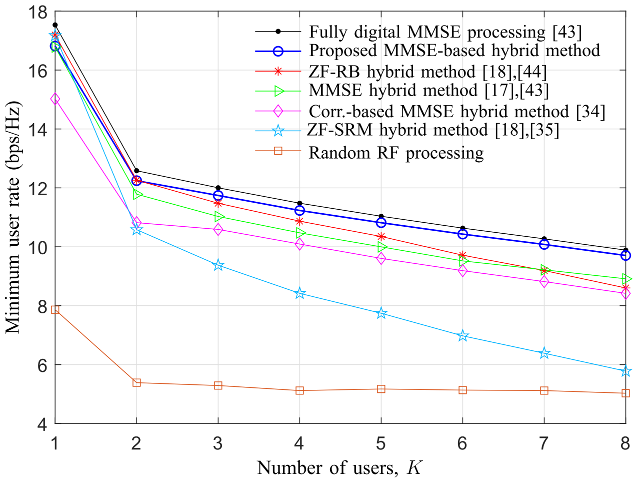

FIGURE 11. Minimum user rate for various hybrid processing methods across the number of users when $M = 64$ and SNR = 10 dB.

This subsection compares the minimum user rate of the proposed method with existing hybrid processing schemes when the perfect CSI is available at the transmitter. The minimum user rate is presented for various hybrid processing schemes according to the SNR in Fig. 10, the number of users in Fig. 11, and the number of transmit antennas in Fig. 12. The proposed MMSE-based hybrid method performs very close to the fully digital MMSE processing attaining the performance upper bound, irrespective of the SNR regions, the number of users, and the number of transmit antennas. The proposed method mitigates the inter-user interference by designing the baseband precoder and combiners in the MMSE sense, whereas the ZF-based techniques enforce the baseband precoder to completely remove the inter-user interference through additional constraints. Thus, the proposed method obtains better minimum user rate than the

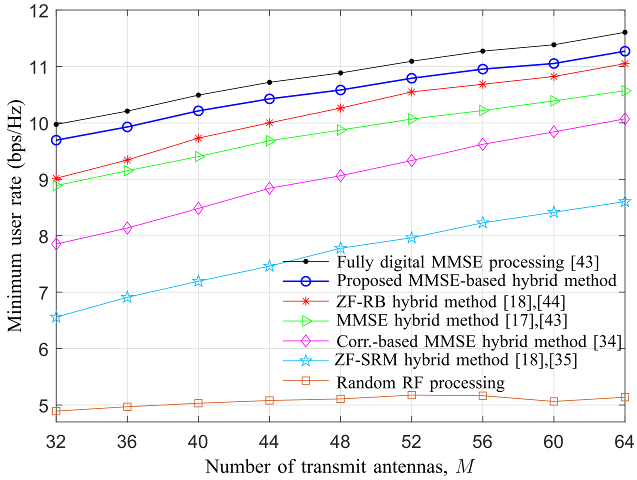

FIGURE 12. Minimum user rate across the number of transmit antennas when $K = 4$ and SNR=10 dB.

ZF-RB and ZF-SRM hybrid methods except the case with no inter-user interference like $K = 1$ of Fig. 11. As the number of users increases in Fig. 11, the minimum user rate decreases faster in the ZF-based methods than other MMSE-based schemes, because the MMSE-based schemes mitigate the inter-user interference more effectively than the ZF-based methods. In Fig. 10, the MMSE hybrid method shows slightly worse minimum user rate than the ZF-RB method due to the performance loss of the scaling procedure for complying with the transmit power constraint. The ZF-RB hybrid method exhibits much better performance than the ZF-SRM hybrid method, because the power allocation is conducted for rate balancing in the ZF-RB method and for sum-rate maximization in the ZF-SRM method. Moreover, the proposed MMSE-based hybrid method outperforms the corr.-based MMSE hybrid method and the random RF processing for all cases. In Fig. 12, the performance difference between the proposed scheme and the ZF-based hybrid method decreases with the increment of the number of transmit antennas due to the reduction of inter-user interference. Also, notice that the minimum user rate for the random RF processing is almost the same regardless of the number of transmit antennas, because the RF precoder does not achieve beamforming gains.

Fig. 13 presents the average runtime of the overall hybrid processing methods according to the number of transmit antennas when $K = 4$ and SNR $= 10$ dB. We used the same server as in Fig. 9 for measuring the average runtime. The overall hybrid processing method is composed of the matrix factorization and the design procedure for the baseband precoder and combiners in Tables 1 and 2, respectively. The time complexity for the baseband design procedure is dominant in the proposed method, ZF-RB and ZF-SRM hybrid methods, and the corr.-based MMSE hybrid method, i.e. the complexity order is given by $O(J_2M_{RF}^3)$ for the proposed method, $O(J_2LK^2)$ for the ZF-RB method, $O(J_2LK)$ for the ZF-SRM method, and $O(J_2M_{RF}^3)$ for the corr.-based MMSE hybrid method. In contrast, the complexities for the matrix factorization and the baseband design procedure are comparable in the MMSE hybrid method, and thus its complexity order is given by $O(J_1M^2M_{RF}) + O(J_2M_{RF}^3)$. When the parameters are set as $L = 2$, $K = 4$, and $M_{RF} = KN_{RF} = 12$, the complexity order is given by $O(J_1M^2)+O(J_2)$ for the MMSE hybrid method and $O(J_2)$ for the other methods. In Fig. 13, the average runtime increases with the increment of M in the MMSE hybrid method, whereas the runtime is almost the same irrespective of the number of transmit antennas in the other hybrid processing schemes including the proposed method. The proposed method has higher runtime than the ZF-based methods, because the MMSE-based baseband design requires more computational load than the ZF-based design as shown in Table 2.

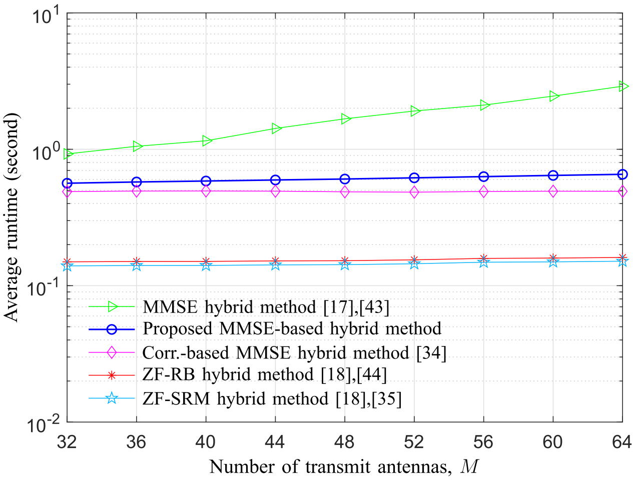

FIGURE 13. Average runtime of various hybrid processing methods across the number of transmit antennas when $K = 4$ and SNR $= 10$ dB.

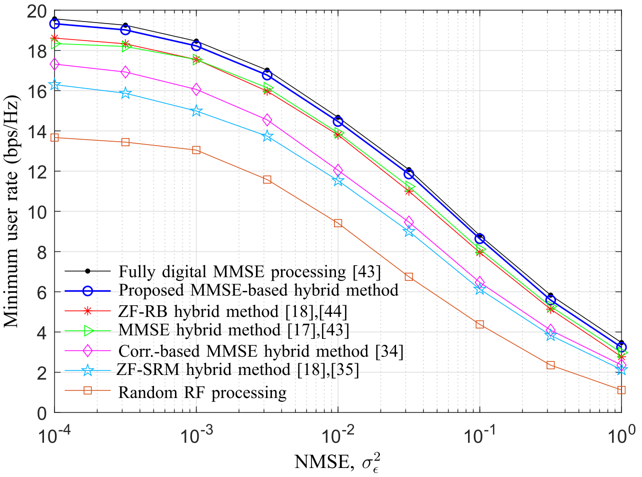

FIGURE 14. Minimum user rate according to NMSE of the channel when $K = 4$, $M = 32$, and SNR $= 25$ dB.

D. PERFORMANCE EVALUATION UNDER CSI UNCERTAINTY

Considering CSI errors in practical systems, we compare the performance of various hybrid processing methods.

CSI uncertainty is caused by the channel estimation error and/or the outdate of CSI in time-varying channels. The channel with CSI errors can be represented as

$$\hat{\boldsymbol{H}}_k = \boldsymbol{H}_k + \Xi_k \tag{44}$$

where $\Xi_k \in \mathbb{C}^{N \times M}$ is a CSI error matrix for user k whose elements are i.i.d. complex Gaussian random variables with zero mean, and $k = 1, 2, \cdots, K$. To describe the power of CSI errors relative to the channel power gains, we define the normalized MSE (NMSE) as follows:

$$\sigma_\epsilon^2 = \frac{E[\| \Xi_k \|_F^2]}{E[\|\boldsymbol{H}_k\|_F^2]}. \tag{45}$$

Applying $\{\hat{\boldsymbol{H}}_k\}$ in (44) instead of $\{\boldsymbol{H}_k\}$ as the input of Fig. 2, we design the precoders and combiners for hybrid processing, $\hat{\boldsymbol{F}}_R$, $\hat{\boldsymbol{F}}_B$, $\{\hat{\boldsymbol{W}}_{R,k}\}$, and $\{\hat{\boldsymbol{W}}_{B,k}\}$, from erroneous CSI. Then, by substituting $\boldsymbol{F}_R, \boldsymbol{F}_B, \{\boldsymbol{W}_{R,k}\}$, and $\{\boldsymbol{W}_{B,k}\}$ of (3) and (4) into $\hat{\boldsymbol{F}}_R$, $\hat{\boldsymbol{F}}_B$, $\{\hat{\boldsymbol{W}}_{R,k}\}$, and $\{\hat{\boldsymbol{W}}_{B,k}\}$, respectively, we can compute $\hat{R}_k$, i.e. the achievable rate for user k under CSI uncertainty.

Fig. 14 shows the minimum user rate of various hybrid processing schemes according to the NMSE, when $K = 4$, $M = 32$, and SNR = 25 dB. For simplicity, we assume that the NMSE is identical to all users. As expected, the minimum user rate gradually decreases with the increment of the NMSE for all processing methods. As in the perfect CSI scenarios, the proposed MMSE-based hybrid method performs better than the ZF-RB and ZF-SRM hybrid methods irrespective of NMSE, while achieving the minimum user rate very close to that of the fully digital MMSE processing. Moreover, the proposed scheme obtains huge gains in the minimum user rate compared to the corr.-based MMSE hybrid method and the random RF processing.

VI. CONCLUSION

A new MMSE-based design method was proposed for hybrid processing in the downlink of mmWave MU-MIMO systems that computes the RF precoder and combiners using the proposed matrix factorization algorithm and obtains the baseband precoder and combiners via the proposed rate balancing algorithm. Considering the matrix concatenation for hybrid precoding, the proposed matrix factorization scheme makes the columns of the RF precoder near-orthogonal and the proposed rate balancing algorithm adjusts the internal transmit power for baseband precoding. Various numerical simulations demonstrate that the proposed method performs better than existing hybrid processing techniques in terms of maximizing the minimum user rate with reasonable computational complexity.

The proposed method can be utilized to design the hybrid precoders and combiners for future 5G-Advanced and 6G mobile systems with large-scale antenna elements deployed in mmWave and Terahertz bands. In addition, the proposed matrix factorization scheme for constant-modulus RF processing can be exploited to a wireless communication link with an intelligent reflecting surface (IRS) which enhances the link performance by controlling phase shifts of IRS elements.

REFERENCES

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

[2] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[3] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 84–91, Jun. 2015.

[4] O. El Ayach, R. W. Heath, Jr., S. Abu-surra, S. Rajagopal, and Z. Pi, "The capacity optimality of beam steering in large millimeter wave MIMO systems," in *Proc. IEEE 13rd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2012, pp. 100–104.

[5] 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, 2013.

[6] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, Sep. 2014.

[7] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2nd Quart., 2018.

[8] *3GPP; Technical Specification Group Radio Access Network; NR; User Equipment (UE) Radio Transmission and Reception—Part 2: Range 2 Standalone (Release 15)*, Standard TS 38.101-2 V15.14.0, 3GPP Technical Specification, Jun. 2021, pp. 1–139.

[9] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62367–62414, 2020.

[10] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[11] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[12] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2nd Quart., 2018.

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

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

[15] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "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.

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

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

[18] J.-C. Chen, "Gradient projection-based alternating minimization algorithm for designing hybrid beamforming in millimeter-wave MIMO systems," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 112–115, Jan. 2019.

[19] X. Qiao, Y. Zhang, M. Zhou, and L. Yang, "Alternating optimization based hybrid precoding strategies for millimeter wave MIMO systems," *IEEE Access*, vol. 8, pp. 113078–113089, 2020.

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

[21] J. Jin, Y. R. Zheng, W. Chen, and C. Xiao, "Hybrid precoding for millimeter wave MIMO systems: A matrix factorization approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3327–3339, May 2018.

[22] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, Jr., "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.

[23] Z. Wang, M. Li, Q. Liu, and A. L. Swindlehurst, "Hybrid precoder and combiner design with low-resolution phase shifters in mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 256–269, May 2018.

[24] F. Dong, W. Wang, and Z. Wei, "Low-complexity hybrid precoding for multi-user mmWave systems with low-resolution phase shifters," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9774–9784, Oct. 2019.

[25] A. N. Uwaechia, N. M. Mahyuddin, M. F. Ain, N. M. A. Latiff, and N. F. Za'bah, "On the spectral-efficiency of low-complexity and resolution hybrid precoding and combining transceivers for mmWave MIMO systems," *IEEE Access*, vol. 7, pp. 109259–109277, 2019.

[26] A. W. Shaban, O. Damen, Y. Xin, and E. Au, "Statistically-aided codebook-based hybrid precoding for millimeter wave channels," *IEEE Access*, vol. 8, pp. 101500–101513, 2020.

[27] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, "Codebook-based hybrid precoding for millimeter wave multiuser systems," *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5289–5304, Oct. 2017.

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

[29] A. Alkhateeb, R. W. Heath, Jr., and G. Leus, "Achievable rates of multiuser millimeter wave systems with hybrid precoding," in *Proc. IEEE Int. Conf. Commun.* London, U.K., Jun. 2015, pp. 1232–1237.

[30] 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.

[31] K. Duan, H. Du, and Z. Wu, "Hybrid alternating precoding and combining design for mmWave multi-user MIMO systems," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2018, pp. 217–221.

[32] 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.

[33] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, "Hybrid MMSE precoding for mmWave multiuser MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

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

[35] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[36] R. Zhang, "Cooperative multi-cell block diagonalization with per-base-station power constraints," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1435–1445, Dec. 2010.

[37] J. Choi, S. Han, and J. Joung, "Low-complexity multiuser MIMO precoder design under per-antenna power constraints," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 9011–9015, Sep. 2018.

[38] J. Joung and Y. H. Lee, "Regularized channel diagonalization for multiuser MIMO downlink using a modified MMSE criterion," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1573–1579, Apr. 2007.

[39] 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.

[40] H. Sung, S. R. Lee, and I. Lee, "Generalized channel inversion methods for multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3489–3499, Nov. 2009.

[41] S. S. Christensen, R. Agarwal, E. D. Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

[42] S. Shi, M. Schubert, and H. Boche, "Downlink MMSE transceiver optimization for multiuser MIMO systems: MMSE balancing," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3702–3712, Aug. 2008.

[43] I. Ghamnia, D. Slock, and Y. Yuan-Wu, "Rate balancing for multiuser MIMO systems," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[44] W.-H. Lim, S. Jang, W. Park, and J. Choi, "ZF-based downlink hybrid precoding and combining for rate balancing in mmWave multiuser MIMO systems," *IEEE Access*, vol. 9, pp. 162731–162742, 2021.

[45] J. A. Tropp, I. S. Dhillon, R. W. Heath, and T. Strohmer, "Designing structured tight frames via an alternating projection method," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 188–209, Jan. 2005.

[46] J. Choi, B. Mondal, and R. W. Heath, "Interpolation based unitary precoding for spatial multiplexing MIMO-OFDM with limited feedback," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4730–4740, Dec. 2006.

WOOHYEONG PARK received the B.S. degree from Korea Aerospace University (KAU), Goyang-si, South Korea, in February 2022. He is preparing for his graduate studies.

In 2020, he joined the Intelligent Signal Processing Laboratory (ISPL), School of Electronics and Information Engineering, KAU, as an Undergraduate Research Assistant, where he performed research on signal processing for MIMO and IRS aided communications, compressive sensing algorithms, and transceiver design for mmWave communications. His research interests include mobile communication techniques, satellite communications, signal processing for IRS aided communication networks, and transceiver design for next generation cellular networks.

JIHOON CHOI (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1997, 1999, and 2003, respectively.

From 2003 to 2004, he was with the Department of Electrical and Computer Engineering, The University of Texas at Austin, where he performed research on MIMO-OFDM systems as a Postdoctoral Fellow. From 2004 to 2008, he was with the Samsung Electronics, South Korea, where he worked on developments of radio access stations for M-WiMAX and base stations for CDMA 1xEV-DO Rev.A/B. In 2008, he joined Korea Aerospace University (KAU), Goyang-si, South Korea, as a Faculty Member. He is currently a Professor with the School of Electronics and Information Engineering, KAU, where he is also the Chief Investigator of the ISPL. His research interests include MIMO communications and signal processing algorithms, IRS aided communication networks, secure transmission in the physical layer, radar signal processing, UAV trajectory optimization, mobile edge computing, modem design for future cellular networks, wireless LANs, the IoT devices, and digital broadcasting systems.