Partially-Connected Hybrid Beamforming for Spectral Efficiency Maximization via a Weighted MMSE Equivalence

Xingyu Zhao, Student Member, IEEE, Tian Lin, Student Member, IEEE, Yu Zhu, Member, IEEE, and Jun Zhang, Senior Member, IEEE

Abstract

Hybrid beamforming (HBF) is an attractive technology for practical massive multiple-input and multiple-output (MIMO) millimeter wave (mmWave) systems. Compared with the fully-connected HBF architecture, the partially-connected one can further reduce the hardware cost and power consumption. However, the special block diagonal structure of its analog beamforming matrix brings additional design challenges. In this paper, we develop effective HBF algorithms for spectral efficiency maximization (SEM) in mmWave massive MIMO systems with the partially-connected architecture. One main contribution is that we prove the equivalence of the SEM problem and a matrix weighted sum mean square error minimization (WMMSE) problem, which leads to a convenient algorithmic approach to directly tackle the SEM problem. Specifically, we decompose the equivalent WMMSE problem into the hybrid precoding and hybrid combining subproblems, for which both the optimal digital precoder and combiner have closed-form solutions. For the more challenging analog precoder and combiner, we propose an element iteration based algorithm and a manifold optimization based algorithm. Finally, the hybrid precoder and combiner are alternatively updated. The overall HBF algorithms are proved to monotonously increase the spectral efficiency and converge. Furthermore, we also propose modified algorithms with reduced computational complexity and finite-resolution phase shifters. Simulation results demonstrate that the proposed HBF algorithms achieve significant performance gains over conventional...
I. INTRODUCTION

The strong desire for supporting ultra-high-speed data transmission has promoted the investigation and application of millimeter wave (mmWave) communications due to its advantage of providing huge spectrum resources [1]–[5]. To overcome the severe path loss and penetration loss of the mmWave propagation channel while considering the stringent constraint of the mmWave hardware cost and power consumption, the combination of massive multiple-input and multiple-output (MIMO) and hybrid beamforming (HBF) has recently become an attractive technology [6]–[9]. However, in the broadband scenario, the HBF design is quite challenging because of the joint optimization of a larger number of low-dimensional digital beamformers for all the subcarriers and a high-dimensional analog beamformer for the whole bandwidth with the specific constant modulus constraint due to the implementation of phase shifters [10]–[12].

A. Related Works and Motivations

Previous works have spent significant efforts on the fully-connected HBF architecture [9], [10], [12]–[19], where each radio frequency (RF) chain is connected to all the antennas. In [9], the authors regarded the HBF design as a matrix factorization problem by minimizing the Euclidean distance between the HBF matrix and the fully-digital beamforming matrix. The well-known orthogonal matching pursuit algorithm was applied to the HBF design with the motivation of exploiting the sparse characteristics of the mmWave propagation channel. However, the constraint that the analog beamforming matrix must be taken in a limited feasible set space led to certain performance loss. In [10], a manifold optimization (MO) based algorithm has been proposed to directly deal with the constant modulus constraint for better performance. In [13], instead of solving the matrix factorization problem, the authors directly targeted the original SEM problem and proposed some iterative algorithms. In [12], the minimum mean square error (MMSE) criterion has been taken for the HBF optimization, based on which the HBF design problem can
be shown to be decomposed into two subproblems with respect to hybrid precoding and hybrid combining optimization and solved in a unified way.

Although the fully-connected architecture has the potential of achieving the full beamforming gain for each RF chain, it requires complex circuitry and consumes relatively high power. An alternative way is to connect each RF chain only with part of the antennas, i.e., the partially-connected architecture, which can greatly reduce the hardware cost and power consumption. However, the traditional HBF design algorithms for the fully-connected architecture cannot be straightforwardly applied to the partially-connected one as the analog beamforming matrix becomes a block diagonal matrix.

Compared to its fully-connected counterpart, HBF for the partially-connected architecture has been less well studied. The first effort was in [10], where the analog and digital precoders were alternatively optimized by updating one while fixing the other. A semi-definite relaxation based algorithm was proposed to optimize the digital precoder. However, the whole design was still based on the matrix factorization approach instead of directly minimizing the spectral efficiency. In [20], the authors first designed the analog precoder for high signal to noise ratio (SNR) and low SNR regions, respectively, and then applied the water-filling algorithm to optimize the digital precoder. However, only the narrowband scenario was considered. In [11], the authors considered the design of HBF for MIMO orthogonal frequency division multiplexing (OFDM) systems. By utilizing the average of the covariance matrices of frequency domain channels, the original algorithms proposed for the narrowband scenario can be extended to the broadband one. However, to solve the problem, the original objective function had to be replaced by an upper bound, which unavoidably led to performance loss.

B. Contributions and Paper Organization

In this paper, we investigate the HBF problem for mmWave massive MIMO-OFDM systems with the partially-connected architecture, aiming at maximizing the spectral efficiency. In contrast to the previous design approaches that adopt surrogate objectives, e.g., to consider a matrix factorization problem to approximate the fully digital beamformer [9], [10] or to replace the original objective function by some bound or approximation [11], [13], we directly tackle the spectral efficiency maximization (SEM) problem based on an equivalent matrix weighted sum mean square error minimization (WMMSE) problem. The main contributions of this paper are summarized as follows:
• Inspired by previous works on the fully-digital beamforming design for narrowband systems [21]–[23], we prove that the HBF beamforming for maximizing the spectral efficiency is equivalent to the WMMSE problem, which provides a new and promising algorithmic approach for the HBF optimization. We also show that this design approach is applicable to both the partially-connected and fully-connected architectures.

• To deal with the difficulty in the highly non-convex and multivariate HBF optimization problem, we show that the WMMSE problem can be decomposed into the hybrid precoding and hybrid combining subproblems, where both the optimal digital precoder and combiner of the two subproblems have closed-form solutions. With the unit modulus constraint, the analog precoder and combiner are more challenging to optimize, for which we propose an element iteration (EI) algorithm and a MO based algorithms. Finally, an alternating optimization approach is applied which updates the hybrid precoder and combiner iteratively. The proposed WWMSE-EI and WMMSE-MO HBF optimization algorithms are proved to be able to make the spectral efficiency monotonously increase and thus converge. Simulation results show that the proposed HBF optimization algorithms can reduce the required SNR by around 2dB to achieve the same spectral efficiency when compared with the conventional algorithms.

• To reduce the computational complexity, we propose a low complexity MMSE-EI HBF optimization algorithm. We also show that the MMSE-EI algorithm can provide a good initialization point for the WMMSE-EI and WMMSE-MO algorithms to speed up their convergence and improve the spectral efficiency. Furthermore, we propose HBF optimization algorithms considering finite resolution phase shifters, which are shown via simulations to achieve higher spectral efficiently than the simple algorithm with uniform phase quantization.

The remainder of this paper is organized as follows. In Section II, we introduce the mmWave MIMO-OFDM system model with the partially-connect architecture and formulate the HBF optimization problem. In Section III, we first prove the equivalence between the SEM HBF problem and the WMMSE HBF problem, and then propose two iterative algorithms for solving the WMMSE problem. In Section IV, we present some modified HBF algorithms by considering the computational complexity and finite resolution phase shifters. In Section V, we prove the convergence of the proposed iterative algorithms and analysis their computational complexity.
Finally, we provide various simulation results in Section VII and conclude the paper in Section VII.

C. Notations

Throughout this paper, \( a \) (or \( A \)), \( \mathbf{a} \) and \( \mathbf{A} \) stand for a scaling factor, a column vector and a matrix, respectively. \( \angle (\cdot) \) denotes the angle of a complex variable. \( (\cdot)^*, (\cdot)^T \) and \( (\cdot)^H \) respectively represent the operation of the conjugate, transpose and conjugate transpose of \( a \) or \( \mathbf{A} \). \( |\cdot|, \|\cdot\|_F, \text{Tr}\{\cdot\}, (\cdot)^{-1} \) denote the determinant (or module for a complex variable), Frobenius norm, trace and inverse of matrix \( \mathbf{A} \). \( \odot \) is the Hadamard product of two matrices. \( \mathbf{A}(i,:) \), \( \mathbf{A}(:,j) \) and \( \mathbf{A}(i,j) \) denote the \( i \)-th row, the \( j \)-th column and the element in the \( i \)-th row and the \( j \)-th column of \( \mathbf{A} \), respectively. \( \mathbb{C}^{m\times n} \) is a complex space with \( m \times n \) dimensions and \( \mathbf{I}_N \) denotes the \( N \times N \) identity matrix. \( \mathbb{E}[\cdot] \) denotes the expectation operation. Finally, \( \mathbf{x} \sim \mathcal{CN}(\mathbf{\mu}, \mathbf{K}) \) means that \( \mathbf{x} \) is a circularly symmetric complex Gaussian vector whose mean is \( \mathbf{\mu} \) and covariance matrix is \( \mathbf{K} \).

II. System Model and Problem Formulation

A. System Model

Consider the downlink of an mmWave MIMO-OFDM system with the partially-connected HBF architecture, as shown in Fig. 1. At the transmitter (i.e., the base station (BS)), a digital transmit beamformer (precoder) \( \mathbf{F}_{D,k} \in \mathbb{C}^{N_{RF,t} \times N_s} \) is first employed to precode \( N_s \) data streams, denoted by vector \( \mathbf{s}_k \in \mathbb{C}^{N_s \times 1} \) with \( \mathbb{E} [\mathbf{s}_k \mathbf{s}_k^H] = \mathbf{I}_{N_s} \), at the \( k \)-th subcarrier, for \( k = 1, 2, \ldots, K \), where \( N_{RF,t} \) denotes the number of transmit RF chains. Then, each of the \( N_{RF,t} \) precoded streams are converted to the time domain by a \( K \)-point inverse fast Fourier transform (IFFT). After
adding a cyclic prefix (omitted in Fig. due to space limit), each stream is up-converted to the
carrier frequency by passing through a dedicated RF chain. Before transmitting the RF signals
at the \(N\) antennas, an analog precoder consisting of a number of phase shifters is deployed for
enhancing the beamforming gain. From an equivalent baseband point of view, the transmitted
signal vector at the \(k\)-th subcarrier of the \(N\) antennas is represented by
\[ x_k = F_{RF}F_D s_k, \]
where \(F_{RF} \in \mathbb{C}^{N \times N_{RF}}\) denotes the analog precoder. Note that \(F_{RF}\) is the same for all the \(K\) subcarriers
because of its post-IFFT processing. Considering the maximum transmit power constraint per
subcarrier, we have \(\|F_{RF}F_D\|_F^2 \leq 1\).

At the receiver (i.e., the user equipment (UE)) with \(M\) receive antennas, the equivalent
baseband received signal at the \(k\)-th subcarrier can be expressed as
\[ \tilde{y}_k = H_k F_{RF}F_D s_k + n_k, \]
where \(n_k \sim \mathcal{CN}(0, \sigma^2 I_M)\) denotes the additive white Gaussian noise at the \(k\)-th subcarrier, and
\(H_k \in \mathbb{C}^{M \times N}\) denotes the geometric model based channel matrix at the \(k\)-th subcarrier [10]–[12],
which is given by
\[
H_k = \sqrt{\frac{M N}{N_C N_R}} \sum_{c=1}^{N_C} \sum_{l=1}^{N_R} h_{cl} a_e(\theta_{cl}^r) a_t(\theta_{cl}^t) e^{-j \frac{2\pi}{N} (k-1)},
\]
where \(N_C\) and \(N_R\) denote the number of clusters and the number of rays in each cluster, respectively. Likewise, \(h_{cl}, \theta_{cl}^r\) and \(\theta_{cl}^t\) represent the complex gain, the angles of arrival and departure
(AoA and AoD) of the \(l\)-th ray in the \(c\)-th propagation cluster, respectively. In addition, \(a_e(\cdot)\) and
\(a_t(\cdot)\) denote the array response vectors of the receiver and transmitter, respectively. For example,
as for a half-wavelength spaced uniform liner array (ULA) with \(N\) antennas at the transceiver,
the array response vector can be represented as
\[ a(\theta) = \frac{1}{\sqrt{N}} [1, e^{i \pi \sin \theta}, \ldots, e^{i (N-1) \pi \sin \theta}]^T. \]

The received signal is first processed by an analog receive beamformer (combiner) \(W_{RF} \in \mathbb{C}^{M \times N_{RF}}\) and then down-converted to the baseband, where \(N_{RF}\) denotes the number of receive RF
chains. Note that \(W_{RF}\) is also the same for all subcarriers as similar to \(F_{RF}\). After cyclic prefix
removal and fast Fourier transform (FFT), the \(N_{RF}\) baseband signal streams are then passing
through a low-dimensional digital combiner \(W_{D,k} \in \mathbb{C}^{N_{RF} \times N_s}\) at subcarrier \(k\) with the output
given by
\[
y_k = W_k^H H_k F_k s_k + W_k^H n_k,
\]
where \(F_k = F_{RF}F_{D,k}\) and \(W_k = W_{RF}W_{D,k}\). It is assumed that \(N_{RF} = N_s\) so that the digital
combiner \(W_{D,k}\) is a square matrix. The scenario with \(N_{RF} > N_s\) will be discussed in Section V-A.

In this paper, we mainly focus on the partially-connected architecture for the analog transmit
and receive beamformers and assume that each transmit RF chain is only connected with \(N/N_{RF}\)
antennas at the transmitter and each receive RF chain to \( M/N_t \) antennas at the receiver, as shown in Fig. 1. As such, both \( \mathbf{F}_{RF} \) and \( \mathbf{W}_{RF} \) become block-diagonal matrices
\[
\mathbf{F}_{RF} = \text{blkdiag} \left( \mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{N_t^{RF}} \right),
\]
\[
\mathbf{W}_{RF} = \text{blkdiag} \left( \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{N_t^{RF}} \right),
\]
where \( \mathbf{f}_q \in \mathbb{C}^{N_t^{RF} \times 1} \) for \( q = 1, 2, \ldots, N_t^{RF} \) and \( \mathbf{w}_n \in \mathbb{C}^{M \times 1} \) for \( n = 1, 2, \ldots, N_t^{RF} \). Since these two analog beamformers are implemented using phase shifters, the non-zero elements in \( \mathbf{F}_{RF} \) and \( \mathbf{W}_{RF} \) are subject to the constant modulus constraint.

### B. Problem Formulation

The achievable spectral efficiency of the aforementioned system at the \( k \)-th subcarrier is given by
\[
R_k = \log \left| I_{N_t} + \sigma^{-2} \mathbf{W}_k^H \mathbf{H}_k \mathbf{F}_k \mathbf{F}_k^H \mathbf{W}_k (\mathbf{W}_k^H \mathbf{W}_k)^{-1} \right|. \tag{4}
\]
In this work, we aim at maximizing the average spectral efficiency over the \( K \) subcarriers subject to the transmit power constraint and the constant modulus constraint of the analog beamformers. The problem can be formulated as follows
\[
\begin{aligned}
& \text{maximize} & & \frac{1}{K} \sum_{k=1}^{K} R_k \\
& \text{s.t.} & & \|\mathbf{F}_{D,k}\|_F^2 \leq \frac{N_t^{RF}}{N}, \quad \forall k \\
& & & |\mathbf{f}_q(p)| = 1, \quad \forall p, \forall q \\
& & & |\mathbf{w}_n(m)| = 1, \quad \forall m, \forall n,
\end{aligned} \tag{5}
\]
where \( \mathbf{f}_q(p) \) and \( \mathbf{w}_n(m) \) denote the \( p \)-th element in \( \mathbf{f}_q \) and the \( m \)-th element in \( \mathbf{w}_n \), respectively. The transmit power constraint in (5) comes from the requirement of \( \|\mathbf{F}_{RF}\mathbf{F}_{D,k}\|_F^2 \leq 1 \) and the fact that \( \mathbf{F}_{RF}^H \mathbf{F}_{RF} = \frac{N}{N_t^{RF}} \mathbf{I}_{N_t^{RF}} \) when \( \mathbf{F}_{RF} \) is a block diagonal matrix according to (5a). Throughout this paper, we focus on the HBF optimization and assume that perfect channel state information is available.

### III. HBF Optimization with the Partially-connected Architecture

For the highly non-convex and multivariate optimization problem in (5), it is very difficult to get the optimal solution. Our main idea is that instead of directly solving the problem, we formulate a WMMSE problem and show that it is equivalent to the original SEM problem. It is worth noting that although the WMMSE design approach has been considered in [21]–[23]
for the fully-digital beamforming optimization and in [12], [15] for the HBF optimization, the equivalence in the HBF scenario has not been proved. In this section, we first formulate the WMMSE HBF optimization problem and prove its equivalence to the SEM problem. Then, to solve the WMMSE problem, we separate it into two subproblems, namely the hybrid precoding and hybrid combining optimization subproblems, and propose several effective algorithms. Finally, an alternating minimization algorithm is applied between the two subproblems for better performance.

A. The WMMSE Problem

Similar to that in [12], we take the modified MSE as the performance metric and define the modified MSE matrix [24] as follows

$$E_k = E \left[ (s_k - \xi_k^{-1} y_k) (s_k - \xi_k^{-1} y_k)^H \right]$$

$$= I_{N_s} - \xi_k^{-1} F_k^H H_k^H W_k - \xi_k^{-1} W_k^H H_k F_k + \xi_k^{-2} W_k^H W_k + \xi_k^{-2} W_k^H H_k F_k F_k^H F_k H_k^H W_k,$$

for $k = 0, 1, \ldots, K$, where $\xi_k$ a scaling factor to be jointly optimized with the hybrid beamformers [12]. Defining a semi-positive definite matrix $\Lambda_k \succeq 0$ as the weight matrix for the $k$-th subcarrier, the WMMSE problem can be formulated as

$$\begin{align*}
\text{minimize} & \quad \frac{1}{K} \sum_{k=1}^{K} \left( \text{tr} \left( \Lambda_k E_k \right) - \log |\Lambda_k| \right) \\
\text{s.t.} & \quad \|F_{D,k}\|_F^2 \leq \frac{N_{RF}^M}{N_s}, \quad \forall k \\
& \quad |f_q(p)| = 1, \quad \forall p, \forall q \\
& \quad |w_n(m)| = 1, \quad \forall m, \forall n.
\end{align*}$$

(7)

Fixing $F_{RF}^*, F_{D,k}^*, W_{RF}^*, W_{D,k}^*, \xi_k$, and $\Lambda_k$, the solution of the optimal $W_{D,k}^*$ is given by

$$W_{D,k}^{\text{mmse}} = \left( W_{RF}^H G_k G_k^H W_{RF} + \alpha_k I_{N_{RF}} \right)^{-1} W_{RF}^H G_k,$$

(8)

where $G_k = \xi_k^{-1} H_k F_k$ and $\alpha_k = \frac{\sigma^2 \xi_k^{-2} M}{N_{RF}}$. By substituting $W_{D,k}^{\text{mmse}}$ back into (6), the corresponding MSE matrix becomes

$$E_k^{\text{mmse}} = \left( I_{N_s} + \alpha_k^{-1} G_k^H W_{RF} W_{RF}^H G_k \right)^{-1}.$$

(9)

Then, the following theorem will set up the equivalence between the WMMSE problem and the SEM problem in [5].

Theorem 1 The problem in (7) is equivalent to the problem in (5) in the sense that the global optimal solution of $F_{RF}^*, F_{D,k}^*,$ and $W_{RF}^*$ for the two problems are identical.
**Proof:** First the optimal $W_{D,k}$ can be acquired by differentiating the objective function of (7) with $W_{D,k}$ and setting the result to zero, which is exactly the same as $W_{mmse}^{D,k}$ in (8). As for the weight matrix $\Lambda_k$, it can be shown in the same way that the optimal one has a closed-form expression

$$\Lambda^{\text{opt}}_k = E_{k}^{-1}.$$  \hspace{1cm}  (10)

With the optimal $W_{D,k}$ and $\Lambda_k$, the problem in (7) can be stated as

$$\begin{align*}
\min_{F_{RF},F_{D,k}W_{RF}} & -\frac{1}{K} \sum_{k=1}^{K} \log \left| (E_k^{\text{mmse}})^{-1} \right| \\
\text{s.t.} & \|F_{D,k}\|_F^2 \leq \frac{N_{RF}}{N}, \quad \forall k \\
& |f_p(q)| = 1, \quad \forall p, \forall q \\
& |w_m(n)| = 1, \quad \forall m, \forall n,
\end{align*}$$

with

$$\log \left| (E_k^{\text{mmse}})^{-1} \right| = \log \left| I_N + \alpha_k^{-1} G^H_k W_{RF} W_{RF}^H G_k \right|$$

$$= \log \left| I_N + \frac{N_{RF}}{\sigma^2 M} W_{RF}^H H_k F_k F_k^H H_k^H W_{RF} \right|,$$  \hspace{1cm}  (12)

where (a) follows from the definition of $G_k$ and the Woodbury matrix identity that $\det (I + XY) = \det (I + YX)$. It can be seen that (12) is exactly the same as (4) for the case of $N_{RF} = N_k$. This shows that the problem of (7) is equivalent to the problem (5) in the sense that the global optimal solution of $F_{RF}, F_{D,k},$ and $W_{RF}$ for the two problems are identical. The proof is thus completed.

Theorem 1 implies that the SEM problem can be achieved by solving the problem based on the WMMSE criterion. It is also worth noting that the proof of Theorem 1 can be shown to be applicable to both the fully-connected HBF architecture and the partially-connected one as there is no constraint to the structure of the analog beamformers in the proof.

In the following subsections, we focus on solving the problem in (7). Although it is almost intractable to straightly obtain the optimal solution of (7) due to the coupled multiple variables and the non-convex constraints, we propose the following solution. First, we show that the original problem can be separated into the hybrid precoding and hybrid combining subproblems. Then, we show that both the optimal digital precoder and combiner have closed-form expressions when dealing with the two subproblems, and the analog precoder and combiner can be optimized with some iterative algorithms in a unified way. Finally, the alternative optimization between these two subproblems can be applied to iteratively update the hybrid beamformers for better performance.
B. Hybrid Precoding Design

By first fixing the hybrid combiner \( \mathbf{W}_{RF}, \mathbf{W}_{D,k} \) and the weight matrix \( \Lambda_k \) in (7), we obtain the hybrid precoding optimization subproblem. By further fixing the analog precoder \( \mathbf{F}_{RF} \) in this subproblem and according to the Karush-Kuhn-Tucker (KKT) conditions, the optimal \( \mathbf{F}_{D,k} \) and \( \xi_k \) has a closed-form solution as follows [12], [24]

\[
\mathbf{F}_{D,k}^{\text{opt}} = \xi_k \tilde{\mathbf{F}}^{-1}_k \mathbf{F}_{RF}^H \tilde{\mathbf{G}}_k^H \Lambda_k
\]

and

\[
\xi_k = \frac{1}{\sqrt{\frac{N}{N_{RF}} \| \tilde{\mathbf{F}}^{-1}_k \mathbf{F}_{RF}^H \tilde{\mathbf{G}}_k^H \Lambda_k \|^2}}
\]

where we define

\[
\tilde{\mathbf{G}}_k^H = \mathbf{H}_k^H \mathbf{W}_k, \quad \beta_k = \frac{\sigma^2 N M \text{tr} \left( \Lambda_k \mathbf{W}_{D,k}^H \mathbf{W}_{D,k} \right)}{N_{RF}^2 N_{RF}^t}, \quad \tilde{\mathbf{F}}_k = \mathbf{F}_{RF}^H \tilde{\mathbf{G}}_k^H \Lambda_k \tilde{\mathbf{G}}_k \mathbf{F}_{RF} + \beta_k \mathbf{I}_{N_{RF}}.
\]

By substituting (13) and (14) back into the objection function in (7), and assuming a fixed hybrid combiner, the original problem is now reduced to the one for optimizing the analog precoder \( \mathbf{F}_{RF} \) as follows

\[
\begin{align*}
\text{minimize} & \quad f(\mathbf{F}_{RF}) \\
\text{s.t.} & \quad |f_q(p)| = 1, \quad \forall p, \forall q,
\end{align*}
\]

where

\[
f(\mathbf{F}_{RF}) = \frac{1}{K} \sum_{k=1}^{K} \text{tr} \left( \left( \Lambda_k^{-1} + \beta_k^{-1} \tilde{\mathbf{G}}_k \mathbf{F}_{RF}^H \tilde{\mathbf{G}}_k^H \right)^{-1} \right).
\]

However, the above problem is still non-convex and it is difficult to obtain the optimal solution. Thus, we propose two iterative algorithms for obtaining a local optimal solution.

1) The EI Algorithm: We first propose an element-by-element iterative optimization algorithm. In each iteration, the contribution of an analog precoding element, say \( f_q(p) = e^{j\theta_{pq}} \) without loss of generality, to the objective function is derived and optimized when other elements are fixed. In particular, By defining \( \Omega_{q,k} \triangleq \Lambda_k^{-1} + \beta_k^{-1} \tilde{\mathbf{G}}_k \mathbf{F}_{RF}^q \mathbf{F}_{RF}^H \tilde{\mathbf{G}}_k^H \), \( \mathbf{A}_{q,k} \triangleq \beta_k^{-1} \tilde{\mathbf{G}}_k^H \Omega_{q,k}^{-2} \tilde{\mathbf{G}}_k \) and
\[ B_{q,k} \triangleq \frac{N_{RF}}{N} I_N + \beta_{k}^{-1} \widetilde{G}_k^H \Omega_{q,k}^{-1} \widetilde{G}_k, \] where \( \hat{f}_q^q \) and \( \bar{F}_q^q \) respectively denote \( F_{RF}(\cdot, q) \) and a sub-matrix of \( F_{RF} \) after removing \( F_{RF}(\cdot, q) \), the objective function in \( (16) \) can be rewritten as

\[ f(F_{RF}) = \frac{1}{K} \sum_{k=1}^{K} \text{tr}\left( \left( \Omega_{q,k} + \beta_{k}^{-1} \widetilde{G}_k^H \hat{f}_q^q (\hat{f}_R^q) (\hat{f}_R^q)^H \bar{G}_k^H \right)^{-1} \right) \]

\[ = \frac{1}{K} \sum_{k=1}^{K} \left( \text{tr}(\Omega_{q,k}^{-1}) - \frac{\text{tr}(\beta_{k}^{-1} \Omega_{q,k}^{-1} \widetilde{G}_k^H \hat{f}_q^q \hat{f}_R^q (\hat{f}_R^q)^H \bar{G}_k^H \Omega_{q,k}^{-1})}{1 + \text{tr}(\beta_{k}^{-1} \Omega_{q,k}^{-1} \widetilde{G}_k^H \hat{f}_q^q \hat{f}_R^q (\hat{f}_R^q)^H \bar{G}_k^H)} \right) \]  \( (18) \)

where (a) follows from the fact that \((X + Y)^{-1} = X^{-1} - \frac{X^{-1}YX^{-1}}{1 + \text{tr}(X^{-1}Y)}\) for a full-rank matrix \( X \) and a rank-one matrix \( Y \), and (b) follows from the property of \( \text{tr}(XY) = \text{tr}(YX) \). Note that \( e^{j\theta_{pq}} \) is contained only in the last term of \( (18) \), which can be written in the following form

\[ - \frac{1}{K} \sum_{k=1}^{K} \left( \hat{f}_R^q H A_{q,k} \hat{f}_R^q \right) = - \frac{1}{K} \sum_{k=1}^{K} \left( \hat{f}_R^q H A_{q,k} \hat{f}_R^q \right) \]

\[ = \frac{1}{K} \sum_{k=1}^{K} A_k + B_k \cos (\theta_{pq} + \theta_{1,k}) \]

\[ C_k + D_k \cos (\theta_{pq} + \theta_{2,k}) \]

where

\[ A_k = \hat{f}_q^p A_{q,k} \hat{f}_q^q + A_{q,k} (p, p), \quad B_k = 2 \left| \hat{f}_q^p A_{q,k} (\cdot, p) \right| \]

\[ C_k = \hat{f}_q^p B_{q,k} \hat{f}_q^q + B_{q,k} (p, p), \quad D_k = 2 \left| \hat{f}_q^p B_{q,k} (\cdot, p) \right| \]

where \( \hat{f}_q^p \) is defined as the resulting vector by only setting the element \( f_q^p \) in the vector \( f_R^q \) to zero without changing other elements. It can be found that \( (19) \) is a one-dimensional scalar function of \( \theta_{pq} \). However, due to the summation of the \( K \) terms, the closed-form solution of the optimal \( \theta_{pq} \) is difficult to derive. Fortunately, some simple but efficient search algorithms based on the golden section search or the simulated annealing algorithm \( [25] \) can be applied to obtain the optimal \( \theta_{pq} \) when fixing other analog beamforming elements.

The overall EI algorithm is summarized in Algorithm 1. First, the contribution of \( \theta_{pq} \) in \( F_{RF} \) to the objective function is derived in \( (19) \). Then, \( \theta_{pq} \) is optimized and updated via the one-dimensional search algorithm while guaranteeing that the objective function keeps decreasing. The element-by-element iterations will be repeated until a stop condition is satisfied finally. Therefore, the convergence of the analog precoder based on the EI algorithm can be strictly proved.

2) The MO Algorithm: The above EI algorithm involves alternating iterations among elements and thus may have high computational complexity. Here we propose the MO algorithm with less complexity. Note that the MO method has been applied to deal with a series of problems
Algorithm 1 The EI algorithm for the analog beamforming optimization with the partially-connected architecture

**Input:** $\beta_k, \Lambda_k, \sigma^2, \tilde{G}_k, K$

1: Initialize $F_{RF}^{(i)}$ with random phases and set $i = 0$

2: repeat

3: for $q = 1 \rightarrow N_t^{RF}$ do

4: Compute $\Omega_{q,k}, A_{q,k}$ and $B_{q,k}$ according to (18);

5: for $p = \frac{(q-1)N_t^{RF}}{N_t^{RF}} + 1 \rightarrow \frac{qN_t^{RF}}{N_t^{RF}}$ do

6: Compute $A_k, B_k, C_k, D_k, \theta_{1,k}$ and $\theta_{2,k}$ according to (20);

7: Update $F_{RF}^{(i)}(p, q)$ based on the one-dimensional search;

8: end for

9: end for

10: $F_{RF}^{(i+1)} = F_{RF}^{(i)}$ and $i \leftarrow i + 1$

11: until a stop condition is satisfied.

**Output:** $F_{RF}$

with the constant modulus constraint of the phase shifters [10], [12], [26]. The basic idea is to consider these optimization problems in a Riemannian manifold space, i.e., a complex circle manifold defined by the constant modulus constraint. Then, some gradient decent like algorithm can be adopted to iteratively update the optimization variable (the analog beamformer) in the direction of the Riemannian gradient. Finally, the updated optimization variable is retracted into a complex circle manifold to meet the constant modulus constraint (For more detail, please refer to [10]).

In the above design procedure, the crucial step is to derive the Riemannian gradient, which is the orthogonal projection of the Euclidean conjugate gradient. To the best of our knowledge, the MO method has not been applied to solve the HBF optimization problem with the partially-connected architecture. We show in this subsection that with the help of the following lemma, the Euclidean conjugate gradient in the partially-connected architecture can be derived.

**Lemma 1** For the partially-connected architecture, the Euclidean conjugate gradient $\nabla f (F_{RF})$ of the real-valued function $f (F_{RF})$ with respect to $F_{RF}$ can be given by

$$\nabla f (F_{RF}) = \nabla_{F_{RF}} f (F_{RF}) \odot P_1,$$

(21)
where $\nabla_{\mathbf{F}_{\text{RF}}}^* \mathbf{f} (\mathbf{F}_{\text{RF}}) = \frac{\partial \mathbf{f} (\mathbf{F}_{\text{RF}})}{\partial \mathbf{F}_{\text{RF}}}^*$, $\mathbf{P}_1 = \text{blkdiag}(\mathbf{p}_1, \ldots, \mathbf{p}_{N_{\text{RF}}})$ is a block-diagonal matrix and $\mathbf{p}_1 = \ldots = \mathbf{p}_{N_{\text{RF}}}$ = $\mathbf{I}_{N_{\text{RF}} \times 1}$.

**Proof:** See Appendix A. $\blacksquare$

According to Lemma 1, the first step is to derive $\nabla_{\mathbf{F}_{\text{RF}}}^* \mathbf{f} (\mathbf{F}_{\text{RF}})$ without considering the partially-connected architecture. By applying the properties of the matrix differentiation [27], we have

$$
d (\mathbf{f} (\mathbf{F}_{\text{RF}})) = \left( \frac{1}{K} \sum_{k=1}^{K} \text{tr} (\mathbf{M}_k^{-1}) \right)
$$

$$(a) \Rightarrow - \text{tr} \left( \frac{1}{K} \sum_{k=1}^{K} \mathbf{M}_k^{-2} d (\mathbf{M}_k) \right)
$$

$$(b) \Rightarrow - \text{tr} \left( \frac{1}{K} \sum_{k=1}^{K} \beta_k^{-1} \mathbf{G}_k \mathbf{M}_k^{-2} \mathbf{G}_k^H \mathbf{F}_{\text{RF}} d (\mathbf{F}^H_{\text{RF}}) \right),
$$

where (a) follows by defining $\mathbf{M}_k = \Lambda_k^{-1} + \beta_k^{-1} \mathbf{G}_k \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{RF}}^H \mathbf{G}_k^H$ and noting that $d (\text{tr} (\mathbf{X}^{-1})) = - \text{tr} (\mathbf{X}^{-2} d (\mathbf{X}))$, and (b) follows from the basic properties of the matrix differentiation and the trace, i.e., $d (\mathbf{A} \mathbf{X} \mathbf{B}) = \mathbf{A} d (\mathbf{X}) \mathbf{B}$ ($\mathbf{A}$ and $\mathbf{B}$ are constant matrices independent of $\mathbf{X}$) and $\text{tr} (\mathbf{X} \mathbf{Y}) = \text{tr} (\mathbf{Y} \mathbf{X})$. According to the relationship between the matrix differentiation of the scaling function and the Euclidean gradient [27], we have

$$
d (\mathbf{f} (\mathbf{F}_{\text{RF}})) = \text{tr} \left( \nabla_{\mathbf{F}_{\text{RF}}}^* \mathbf{f} (\mathbf{F}_{\text{RF}}) d (\mathbf{F}^H_{\text{RF}}) \right).$$

By comparing (22) and (23), we obtain $\nabla_{\mathbf{F}_{\text{RF}}}^* \mathbf{f} (\mathbf{F}_{\text{RF}})$. According to Lemma 1, we finally have

$$
\nabla \mathbf{f} (\mathbf{F}_{\text{RF}}) = - \left( \frac{1}{K} \sum_{k=1}^{K} \beta_k^{-1} \mathbf{G}_k^H \mathbf{M}_k^{-2} \mathbf{G}_k \mathbf{F}_{\text{RF}} \right) \odot \mathbf{P}_1.
$$

With the derived Euclidean conjugate gradient, the next step is to project it onto the tangent space to obtain the Riemannian gradient and update $\mathbf{F}_{\text{RF}}$ with a proper step size determined by the well-known Armijo backtracking algorithm. Finally, the retraction operation is applied to make the result satisfy the constant modulus constraint. The overall algorithm for the analog precoding with the partially-connected architecture is summarized in Algorithm 2.

It is worth noting that the MO method can guarantee the convergence to a critical point where the gradient is zero according to Theorem 4.3.1 in [26]. Moreover, the well-developed conjugate gradient descent algorithm using the Armijo backtracking line search step and the Polak-Ribiere parameter can also ensure the objective function not to increase in each iteration [28]. Thus, the entire MO analog beamforming optimization algorithm converges, which further ensures that the
Algorithm 2 The MO algorithm for the analog beamforming optimization with the partially-connected architecture

Input: $\beta_k, \Lambda_k, \sigma^2, G_k, K$

1: Initialize $F_R^{(i)}$ with random phases and set $i = 0$;
2: repeat
3: Compute $\nabla f (F_R)$ according to Lemma 1
4: Update $F_R^{(i+1)}$ based on the MO method;
5: $i \leftarrow i + 1$;
6: until a stopping condition is satisfied.

Output: $F_R$

iterations of the analog precoding optimization make the WMMSE not to increase until the stop condition is satisfied.

C. Hybrid Combining Design

In the above subsection, we have investigated the hybrid precoding optimization subproblem by assuming the hybrid combiner is fixed. Now we focus on the design of the hybrid combiner with a fixed precoder. Back to the original WMMSE problem in (7), by fixing $F_R, F_{D,k}, \xi_k$ and $\Lambda_k$, the objective function is only a function of $W_R$ and $W_{D,k}$. The closed-form solution of the optimal $W_{D,k}$ has been given by (8). By substituting it into the object function in (7) and neglecting the second unrelated term, the optimization problem for $W_R$ can be formulated as

$$
\min_{W_R} \quad g(W_R) \\
\text{s.t.} \quad |w_n(m)| = 1, \quad \forall m, \forall n,
$$

where

$$
g(W_R) = \frac{1}{K} \sum_{k=1}^{K} \text{tr} \left( (\Lambda_k^{-1} + \alpha_k^{-1} \Lambda_k^{-1} G_k^H W_R W_R^H G_k)^{-1} \right).
$$

This problem is very similar to the analog precoding optimization problem in (16) and thus can be solved in the same way as that in Section III-B. In particular, by similarly defining $\Omega_{n,k} \triangleq \Lambda_k^{-1} + \alpha_k^{-1} \Lambda_k^{-1} G_k^H \tilde{W}_R^{n} \tilde{W}_R^{nH} G_k$, $A_{n,k} \triangleq \alpha_k^{-1} G_k \Omega_{n,k}^{-2} \Lambda_k^{-1} G_k^H$ and $B_{n,k} \triangleq \frac{N_{RF}}{M} I_M + \alpha_k^{-1} G_k \Omega_{n,k}^{-1} \Lambda_k^{-1} G_k^H$, where $W_R^n$ and $\tilde{W}_R^n$ respectively denote $W_R(:, n)$ and a sub-matrix of $W_R$ after removing $W_R(:, : n)$, Algorithm 1 can be applied to optimize $W_R$ in the same way as that for $F_R$. 
Similarly, for the MO algorithm, the Euclidean gradient of the objective function of $WRF$ can be derived to be
\[
\nabla g(WRF) = -\left(\frac{1}{K} \sum_{k=1}^{K} \alpha_k^{-1} G_k N_k^{-2} G_k^H WRF\right) \odot P_2,
\]
where $N_k \triangleq \Lambda_k^{-1} + \alpha_k^{-1} \Lambda_k^{-1} G_k^H WRF WRF^H G_k$ and $P_2 = \text{blkdiag}(p'_1, \ldots, p'_{NRF})$ with $p'_1 = \ldots = p'_{NRF} = I_{M \times 1}$. Thus, the Algorithm 2 can be applied to optimize the analog combiner.

**D. Alternating Optimization for HBF**

With the alternative optimization, the hybrid precoder and hybrid combiner can be jointly optimized by alternatively and iteratively using the proposed EI or MO algorithm in the above two subsections. Besides, another crucial step is to optimize the weight matrix $\Lambda_k$. According to Theorem 1, there exists a closed-form solution, i.e., $\Lambda_k^{opt} = E_k^{-1}$. Hence, we come up to the whole optimization process, which consists of three steps.

Without loss of generality, we assume that each iteration starts with the optimization of the precoder. Thus, in the first step, by fixing the hybrid combiner and the weight matrix, $F_{RF}$ and $F_{D,k}$ along with $\xi_k$ are optimized according to Algorithm 1 (or Algorithm 2), (13) and (14), respectively. Then, in the second step, by fixing the hybrid precoder and the weight matrix, $W_{RF}$ and $W_{D,k}$ are optimized based Algorithm 1 (or Algorithm 2) and (8), respectively. In the last step, $\Lambda_k$, which is related to the SEM, is obtained according to (10). The three steps are repeated until the stop condition is satisfied. The overall HBF optimization for mmWave MIMO-OFDM systems with the partially-connected architecture is summarized in Algorithm 3. We refer to the whole algorithm as the WMMSE-EI or WMMSE-MO algorithm according to whether the EI algorithm (i.e., Algorithm 1) or MO algorithm (i.e., Algorithm 2) is used when optimizing the analog beamformer. It is worth noting that if each iteration starts with the optimization of the combiner, the weight matrix should be updated before the optimization of the precoder to ensure the convergence of the spectral efficiency. The detailed proof of the convergence of the proposed WMMSE-EI and WMMSE-MO algorithms will be provided in Section V-A.

**IV. MODIFIED HBF DESIGN ALGORITHMS**

In this section, we first consider the HBF optimization by modifying the WMMSE objective function to the MMSE and obtain a low complexity algorithm. We also modify the proposed WMMSE-EI algorithm with the practical consideration of finite resolution phase shifters.
Algorithm 3 HBF optimization with alternating minimization for mmWave MIMO-OFDM Systems with the partially-connected architecture

**Input:** $H_k$ and $\sigma^2$

1. Initialize $W_{RF}^{(i)}, F_{RF}^{(i)}, W_{D,k}^{(i)}, \Lambda_k^{(i)}$ and $i = 0$.
2. repeat
   3. Compute $F_{RF}^{(i)}$ based on Algorithm 1 (or Algorithm 2);
   4. Compute $F_{D,k}^{(i)}$ according to (13);
   5. Compute $W_{RF}^{(i)}$ based on Algorithm 1 (or Algorithm 2);
   6. Compute $W_{D,k}^{(i)}$ according to (12);
   7. Set $\Lambda_k^{(i)} = E_k^{-1}$ based on (9);
   8. $i \leftarrow i + 1$;
3. until a stopping condition is satisfied;

**Output:** $F_{RF}, F_{D,k}, W_{RF}, W_{D,k}$

A. HBF Design Based on the MMSE Criterion

In the previous section, we have investigated the HBF design for the partially-connected architecture aiming at maximizing the spectral efficiency by solving an equivalent WMMSE problem. We now consider its special case when the weight matrix is reduced to an identity matrix, i.e., without any weight. We refer to it as the MMSE HBF design. The motivation for the MMSE HBF design comes from three aspects. First, it can be regarded as a low-complexity version of the WMMSE design as the weight matrix does not need to be optimized. Second, it can be regarded as an initialization step for the WMMSE design by providing some good initial HBF matrices instead of random initialization for the WMMSE based algorithm. Third, in some cases, when a practical system is constrained to some particular modulation and coding scheme instead of the Gaussian code, the MSE metric becomes a direct performance measure to characterize the transmission reliability.

By setting $\Lambda_k = I_{N_s}$ in (7), the MMSE HBF optimization problem is formulated as

\[
\begin{align*}
\text{minimize}_{F_{RF}, F_{D,k}, W_{RF}, W_{D,k}, \xi_k} & \quad \frac{1}{K} \sum_{k=1}^{K} (\text{tr}(E_k)) \\
\text{s.t.} & \quad \|F_{D,k}\|_F^2 \leq \frac{N_{RF}}{N_s}, \quad \forall k \\
& \quad |f_q(p)| = 1, \quad \forall p, \forall q \\
& \quad |w_n(m)| = 1, \quad \forall m, \forall n.
\end{align*}
\]
Similar to the design approach for the WMMSE HBF problem, it can be separated into the hybrid precoding and combining subproblems. It can be shown that the digital precoder $F_{D,k}$, the scaling factor $\xi_k$, and the digital combiner $W_{D,k}$ can be expressed in the same form as those in (13), (14), and (8) by just replacing $\Lambda_k$ by $I_{N_t}$ in them. Furthermore, the optimization problems for the analog precoder and combiner are similar and can be solved in the same way. We take the analog precoder optimization problem with the MMSE criterion for example, which is expressed as follows by replacing $\Lambda_k$ by $I_{N_t}$ in (16):

$$\min_{F_{RF}} \quad J(F_{RF})$$

s.t. $|f_q(p)| = 1, \forall p, \forall q,$

with

$$J(F_{RF}) = \frac{1}{K} \sum_{k=1}^{K} \text{tr} \left( \left( I_{N_t} + \phi_k^{-1} \tilde{G}_k F_{RF} F_{RF}^H \tilde{G}_k^H \right)^{-1} \right),$$

where $\phi_k \triangleq \frac{\sigma^2 N_{RF} N_t}{N_{RF} N_{RF}^t} \text{tr} (W_{D,k}^H W_{D,k}).$ Although it can be solved using the EI or MO algorithm in Section III some new low complexity algorithm can be found. One way is to replace the objective function by one of its upper bounds. Using the Courant-Fisher min-max theorem [12], [29], an upper bound of the objective function $J(F_{RF})$ in (29) can be derived as follows

$$J(F_{RF}) \leq \frac{1}{K} \text{tr} \left( F_{RF}^H N_{RF} I_{N_t} F_{RF} F_{RF}^H \tilde{G}_k \tilde{G}_k^H F_{RF} \right),$$

which is defined as $J_{UB}(F_{RF}).$ Using the matrix inversion equality and after some manipulation, we have

$$J_{UB}(F_{RF}) = \frac{1}{K} \text{tr} \left( \sum_{k=1}^{K} N_{RF} F_{RF} F_{RF}^H F_{RF} - \frac{N}{N_{RF} N_{RF}^t} F_{RF}^H A F_{RF} \right) = \frac{N^2}{N_{RF}^t} - \frac{N}{K N_{RF}^t} \text{tr} (F_{RF}^H A F_{RF}),$$

where $A = \sum_{k=1}^{K} \left( \phi_k^{-1} \tilde{G}_k^H \left( \frac{N_{RF}^t}{N} I_{N_t} + \phi_k^{-1} \tilde{G}_k \tilde{G}_k^H \right)^{-1} \tilde{G}_k \right).$ The $J_{UB}$ minimization problem is then equivalent to the following problem

$$\max_{F_{RF}} \quad \text{tr} (F_{RF}^H A F_{RF})$$

s.t. $|f_q(p)| = 1, \forall p, \forall q.$

Either the EI algorithm or the MO algorithm can be used to solve (33). From the low computational complexity point of view, we focus on the EI algorithm which can be shown to have a closed-form solution of $F_{RF}.$ Following the element-by-element optimization approach of the
EI algorithm and taking the element of $f_q(p) = e^{j\theta_{pq}}$ for example, the objective function in (33) can be expressed as the summation of two terms with or without $\theta_{pq}$. That is,

$$\text{tr} \left( F_{RF}^H A F_{RF} \right) = 2B \cos (\theta_1 + \theta_{pq}) + C,$$

(34)

where $B = 2 \left| \hat{f}_{q,p}^H A (\cdot, p) \right|$, $\theta_1 = \angle \left( \hat{f}_{q,p}^H A (\cdot, p) \right)$, and $C$ is a term unrelated to $\theta_{pq}$. It is not hard to find that $\theta_{pq}^{\text{opt}} = -\theta_1$.

Similarly, it can be found that the above low complexity EI algorithm can be applied to the optimization of the analog combiner. Finally, by using the alternating minimization for the joint hybrid precoding and combining optimization until the stop conditional is satisfied, we obtain the optimized hybrid precoder and combiner. We refer to this HBF optimization algorithm as the MMSE-EI algorithm. Our computational complexity analysis in Section VII along with the simulation results in Section VIII will show that the MMSE-EI algorithm can reduce the computational complexity by more than 90% at the cost of less than 0.5 bits/s/Hz in the spectral efficiency when compared to the WMMSE-EI and WMMSE-MO algorithms.

B. HBF Design with Finite Resolution Phase Shifters

In practical systems, as the phase shifters may have finite resolution, it is necessary to consider the HBF design in this scenario. According to [30], a simple but efficiency way is to design the analog beamforming matrix with the assumption of infinite resolution, for example using the proposed WMMSE-EI, WMMSE-MO, and MMSE-EI algorithms, and then project the resulting phases into the quantized phase shifts set. We refer to this design approach as the projection based approach.

However, for the EI related HBF algorithms, some specific optimization with finite resolution phase shifters can be conducted. As the analog precoding and combining optimization problem can be formulated and processed in a unified way, we take the analog precoding optimization as example. The original precoding optimization problem (16) with infinite resolution can now be formulated as the following one with quantized phase shifts

$$\min_{F_{RF}} f \left( F_{RF} \right)$$

s.t.  

$$f_q(p) \in \mathcal{F}, \quad \forall p, \forall q,$$

(35)

where $\mathcal{F} = \{ f^0, f^1, \ldots, f^{2^B-1} \}$ with $f^n = e^{j2\pi n}$ is the set of all the possible quantized phase shifts and $B$ is the number of quantization bits. Then, based on the previous derivation (19) in the
WMMSE-EI algorithm, the contribution of $f_q(p)$ in the objective function can also be obtained. Thus, with the constraint of the quantized phase shifts, the optimization of $f_q(p) = e^{j\theta_{pq}}$ becomes

$$\theta_{pq}^{opt} = \arg \max_{e^{j\theta_{pq}} \in F} \sum_{k=1}^{K} \frac{A_k + B_k \cos (\theta_{pq} + \theta_{1k})}{C_k + D_k \cos (\theta_{pq} + \theta_{2k})}.$$  \hspace{1cm} (36)

When the number of quantization bits is small, for example, $B = 1$ or $B = 2$, the optimal phase can be quickly obtained through a few comparisons. We refer to this modification of WMMSE-EI as the WMMSE-EI-Q algorithm.

Similarly, in the MMSE-EI algorithm with finite phase resolution, the precoding optimization problem becomes the one to select the best phase shift in the $F$ that maximizes (34), i.e., the closest quantized phase shift to $\theta_1$. We refer to this modification of MMSE-EI as the MMSE-EI-Q algorithm.

V. SYSTEM EVALUATION

In this section, we first show the convergence of the proposed WMMSE-EI and WMMSE-MO HBF optimization algorithms. We also analyze and compare the computational complexity of different HBF algorithms.

A. Convergence Analysis

We have shown in Section III that the proposed WMMSE-EI and WMMSE-MO HBF optimization algorithms have two levels of iterations. One is the iteration between the hybrid precoding and the hybrid combining optimization with the alternating minimization, which we refer to as the outer iteration, the other is the iteration within the EI or MO algorithm when performing the analog precoding or combining optimization, which we refer to as the inner iteration. Furthermore, we have mentioned in Section III-D that each outer iteration of both the WMMSE-EI and WMMSE-MO algorithms consists of three steps, i.e., hybrid precoding optimization, hybrid combining optimization, weight matrix optimization in series.

In this subsection, we first show that both the WMMSE-EI and WMMSE-MO algorithms lead to a non-increasing sequence of the matrix weight sum-MSE until the stop criterion is satisfied. Then, from the relationship between the WMMSE problem and the SEM problem, we show that these two algorithms lead to a non-decreasing sequence of the spectral efficiency.

**Proposition 1** Define the resulting matrix weighted sum-MSE (i.e., the objective function of (7)) of the three steps in the $n$-th outer iteration of the WMMSE-EI or the WMMSE-MO algorithms as...
as \( J_{s1}^{(n)}, J_{s2}^{(n)}, \) and \( J_{s3}^{(n)} \), respectively. Then, \( \{ J_{s1}^{(n)}, J_{s2}^{(n)}, J_{s3}^{(n)} \} \) for \( n = 0, 1, \ldots \), is a non-increasing sequence until the stop condition is satisfied. That is,

\[
J_{s1}^{(n)} \geq J_{s2}^{(n)} \geq J_{s3}^{(n)} \geq J_{s1}^{(n+1)} \geq J_{s2}^{(n+1)} \geq J_{s3}^{(n+1)}.
\] (37)

**Proof:** See Appendix B.

---

**Proposition 2** Define the resulting spectral efficiency in the \( n \)-th outer iteration of the WMMSE-EI or the WMMSE-MO algorithms as \( R^{(n)} \). Then, \( \{ R^{(n)} \} \) for \( n = 0, 1, \ldots \), is a non-decreasing sequence until the stop condition is satisfied. That is, \( R^{(n)} \leq R^{(n+1)} \).

**Proof:** See Appendix C.

---

**Remark 1:** The above convergence proof is based on the optimization order of the hybrid precoder, the hybrid combiner, and the weight matrix within each outer iteration. As we mentioned in Section III-D, the outer iteration can also start with the optimization of hybrid combiner. However, once the hybrid combiner is optimized, the weight matrix should be updated immediately before the optimization of the precoder to ensure the convergence. This can be explained as follows. According to Theorem 1, the equivalence between the WMMSE problem and the SEM problem is established using the optimal weight matrix, which is obtained by substituting the optimized \( W_{\text{mmse}}^{D,k} \) into the MSE matrix. In particular, considering the \( n \)-th outer iteration and assuming that the iteration starts with the hybrid combiner optimization, the optimized hybrid combiner is denoted by \( W_k^{(n)} = W_{\text{RF}} W_{D,k}^{(n)} \), which is a function of the hybrid precoder in the \( (n-1) \)-th iteration, i.e., \( F_{k}^{(n-1)} \). By then substituting the optimized hybrid combiner into the MSE matrix and the weight matrix, the current spectral efficiency can be expressed as

\[
R^{(n)} = \frac{1}{K} \sum_{k=1}^{K} \log \left| \left( E_k^{(n)} \right)^{-1} \right| = \frac{1}{K} \sum_{k=1}^{K} \log \left| I_{N_s} + \frac{N_f^{\text{RF}}}{\sigma^2 M} \left( W_{\text{RF}}^{(n)} \right)^H H_k F_k^{(n-1)} H_k^H W_k^{(n)} \right|. \] (38)

The convergence of \( R^{(n)} \) can be proved in the same procedure as that in Proposition 2. Note that \( R^{(n)} \) in (38) is a function of \( F_{k}^{(n-1)} \). The optimized precoder \( F_{k}^{(n)} \) in the \( n \)-th iteration will be used in optimization of combiner and weight matrix in the next iteration.

**Remark 2:** So far, we have mainly focused on the HBF design in the case when \( N_f^{\text{RF}} = N_s \), and shown that the SEM problem is equivalent to the WMMSE problem from Theorem 1. We have
TABLE I
Computational complexity of different HBF algorithms

| Proposed Algorithms | Computational Complexity | $N_{\text{in}}$ | $N_{\text{out}}$ | $N_g$ |
|---------------------|--------------------------|----------------|----------------|------|
| The WMMSE-EI algorithm | $N_{\text{out}}N_{\text{in}}N_{\text{ant}}K(2N_{\text{ant}}^2N_{\text{RF}} + 3N_{\text{ant}}N_{\text{RF}}^2 + 4N_{\text{ant}}^2 + 2N_{\text{ant}} + 3N_{\text{RF}}^3 - N_{\text{RF}}^2 - N_{\text{ant}}N_{\text{RF}} + N_g + 2O(N_{\text{RF}}^3))$ | 3 | 10 | 8.1 |
| The WMMSE-MO algorithm | $N_{\text{out}}N_{\text{in}}(K(5N_{\text{ant}}^2N_{\text{RF}} + 6N_{\text{ant}}N_{\text{RF}}^2 + 4N_{\text{RF}}^2 + 4O(N_{\text{RF}}^2)) + 3N_{\text{ant}}N_{\text{RF}} + N_{\text{ant}})$ | 21.2 | 10 | -- |
| The MMSE-EI algorithm | $N_{\text{out}}N_{\text{in}}(2N_{\text{ant}}^2N_{\text{RF}} + 3N_{\text{ant}}N_{\text{RF}}^2 + N_{\text{RF}}^2 + O(N_{\text{RF}}^2)) + N_{\text{ant}}^2$ | 4 | 5.2 | -- |

also proposed the WMMSE-EI and WMMSE-MO logarithms to solve the WMMSE problem with guaranteed convergence. In the case when $N_{\text{RF}}^r > N_s$, from the proof of Theorem 1 and (12), we can see that the WMMSE problem is equivalent to the SEM problem maximization if we define the spectral efficiency as the one at the output of the analog combiner $W_{\text{RF}}$. In this case, if a linear digital combiner is employed at each subcarrier after the analog combining, there will be always rate reduction after the digital combining. According to [31], an MMSE estimator along with successive interference cancellation can compensate for such rate reduction. Nevertheless, the linear digital combiner obtained in the WMMSE-EI and WMMSE-MO algorithms, which is given by (8), can still be used. Simulation results in Section VI-B will show that the proposed WMMSE based HBF design approach can significantly outperform the conventional one in both the case of $N_{\text{RF}}^r > N_s$ and the case of $N_{\text{RF}}^r = N_s$.

B. Complexity Analysis

In this subsection we analyze and compare the computational complexity of different HBF optimization algorithms in terms of the number of complex multiplications. As the optimized digital beamformers and the weight matrix have closed-form expressions and have much lower dimension than those of the analog beamformers, we ignore their computational complexity. Besides, as both the analog precoder and combiner can be solved in the same procedure, we focus on the complexity analysis of the analog precoder. To simplify the notation, denote $N_{\text{ant}} = \max\{M, N\}$, $N_{\text{out}}$ as the number of the outer iterations and $N_{\text{in}}$ as the number of the inner iterations and assume $N_{\text{RF}}^r = N_{\text{RF}} = N_s = N_{\text{RF}}$.

1) The WMMSE-EI Algorithm: The complexity of the WMMSE-EI algorithm mainly includes the following two parts:
Computation of some related parameters: According to (18), the complexity for computing \( \Omega_{q,k}, A_{q,k} \) and \( B_{q,k} \) is
\[
K(2N_{\text{ant}}^2N_{\text{RF}} + 3N_{\text{ant}}N_{\text{RF}}^2 + 3N_{\text{RF}}^3 - N_{\text{RF}}^2 - N_{\text{ant}}N_{\text{RF}} + 2O(N_{\text{RF}}^3)) ,
\]
where \( 2O(N_{\text{RF}}^3) \) results from the inversion of two \( N_{\text{RF}} \times N_{\text{RF}} \) matrices. Furthermore, the complexity for computing \( A_k, B_k, C_k, D_k, \theta_{1,k} \) and \( \theta_{2,k} \) in (20) is
\[
K(4N_{\text{ant}}^2 + 2N_{\text{ant}} + 3N_{\text{RF}}^3 - N_{\text{RF}}^2 - N_{\text{ant}}N_{\text{RF}} + N_{\text{ant}}N_{\text{RF}} + N_{\text{RF}}^3 + 2O(N_{\text{RF}}^3)).
\]

One-dimensional line search: We take the golden section search as an example. The complexity is \( KN_g \) if \( N_g \) iterations are needed.

Thus, the total complexity of the analog precoder optimization using Algorithm 1 is given by
\[
C_{\text{WMMSE-EI}} = N_{\text{out}}N_{\text{in}}N_{\text{ant}}K(2N_{\text{ant}}^2N_{\text{RF}} + 3N_{\text{ant}}N_{\text{RF}}^2 + 4N_{\text{ant}}^2
+ 2N_{\text{ant}} + 3N_{\text{RF}}^3 - N_{\text{RF}}^2 - N_{\text{ant}}N_{\text{RF}} + N_{\text{ant}}N_{\text{RF}} + N_{\text{RF}}^3 + 2O(N_{\text{RF}}^3)). \tag{39}
\]

2) The WMMSE-MO Algorithm: The complexity of the WMMSE-MO algorithm mainly includes the following three parts:

- Computation of the conjugate gradient: From the expression of the conjugate gradient in (24), the computational complexity is
\[
K(4N_{\text{ant}}^2N_{\text{RF}} + 4N_{\text{ant}}N_{\text{RF}}^2 + 3N_{\text{RF}}^3 + 2O(N_{\text{RF}}^3)) + N_{\text{ant}}N_{\text{RF}}, \text{ where } 2O(N_{\text{RF}}^3) \text{ results from the inversion of } \Lambda_k \text{ and } \Sigma_k, \text{ and the last term } N_{\text{ant}}N_{\text{RF}} \text{ comes from the operation of the Hadamard production in (21).}
\]
- Orthogonal projection and retraction operations: In the MO method, the orthogonal projection and retraction operations are the key steps which map the Euclidean gradient into the Riemannian manifold and guarantee the satisfaction of the constant modulus constraint. According to [10], the complexity of the orthogonal projection and retraction operations is \( 2N_{\text{ant}}N_{\text{RF}} \) and \( N_{\text{ant}} \), respectively.
- Armijo backtracking line search: In order to determine a suitable step size, the main complexity of the well-known Armijo backtracking line search is
\[
K(N_{\text{ant}}^2N_{\text{RF}} + 2N_{\text{ant}}N_{\text{RF}}^2 + N_{\text{RF}}^3 + 2O(N_{\text{RF}}^3)) \text{ matrices.}
\]

Thus, the total complexity of the analog precoder design using Algorithm 2 is given by
\[
C_{\text{WMMSE-MO}} = N_{\text{out}}N_{\text{in}}K(5N_{\text{ant}}^2N_{\text{RF}} + 6N_{\text{ant}}N_{\text{RF}}^2 + 4N_{\text{RF}}^3 + 4O(N_{\text{RF}}^3)) + 3N_{\text{ant}}N_{\text{RF}} + N_{\text{ant}}). \tag{40}
\]

3) The MMSE-EI Algorithm: The complexity of the MMSE-EI algorithm proposed in Section IV-A mainly includes the following two parts:

- Computation of some related parameters: In each inner iteration, the complexity for computing \( A_k \) in (32) is
\[
K(2N_{\text{ant}}^2N_{\text{RF}} + 3N_{\text{ant}}N_{\text{RF}}^2 + N_{\text{RF}}^3 + O(N_{\text{RF}}^3)), \text{ where } O(N_{\text{RF}}^3) \text{ results from the inversion of an } N_{\text{RF}} \times N_{\text{RF}} \text{ matrix.}
• Update of the optimal phase: The complexity of computing $\theta_1$ is $N_{\text{ant}}$. Hence, the complexity of optimizing $F_{\text{RF}}$ is $N_{\text{ant}}^2$.

Thus, the total complexity of the analog precoder optimization is given by

$$C_{\text{MMSE-EI}} = N_{\text{out}} N_{\text{in}} (K(2N_{\text{ant}}^2 N_{\text{RF}} + 3N_{\text{ant}} N_{\text{RF}}^2 + N_{\text{RF}}^3) + \mathcal{O}(N_{\text{RF}}^3)) + N_{\text{ant}}^2).$$ (41)

In summary, the complexity of all the three proposed HBF algorithms is listed in Table I. It can be seen from this table that the dominant term for the three algorithms is $2N_{\text{out}} N_{\text{in}} N_{\text{ant}}^3 K N_{\text{RF}}$, $5N_{\text{out}} N_{\text{in}} N_{\text{ant}}^2 K N_{\text{RF}}$, and $2N_{\text{out}} N_{\text{in}} N_{\text{ant}}^2 K N_{\text{RF}}$, respectively, which imply that $C_{\text{WMMSE-EI}} > C_{\text{WMMSE-MO}} > C_{\text{MMSE-EI}}$. For more precise comparison and for a more intuitive expression, the average numbers of iterations in different levels are provided over 100 independent channel realizations in simulations, where $N_{\text{ant}} = 32$, $N_{\text{RF}} = 4$ and $K = 64$. Using these parameters, the number of complex multiplications required by the three algorithms are about $9.0 \times 10^8$, $3.3 \times 10^8$ and $1.3 \times 10^7$, respectively. Thus, the MMSE-EI algorithm has the lowest complexity, with at least one order of magnitude lower than that of the other two algorithms.

VI. SIMULATION RESULTS

In this section, we first present some simulation results to evaluate the convergence of the proposed HBF optimization algorithms. We then compare the spectral efficiency performance of different HBF algorithms for various system configurations. Finally, we present some results with the consideration of finite resolution phase shifters.

Consider an mmWave MIMO-OFDM systems with the partially-connected HBF architecture as that in Fig. 1. Unless otherwise specified, we assume that the transmitter takes a half-wavelength spaced ULA with $N = 64$ antennas and $N_{\text{RF}}^t = 4$ RF chains for the transmission of $N_s = 2$ streams, and the receiver takes a ULA with $M = 32$ antennas and $N_{\text{RF}}^r = 2$ RF chains. The total number of subcarriers is set to $K = 64$. The MIMO channel is generated according to the model in (1) in Section II where the number of clusters and the number of rays in each cluster are set to $N_C = 5$ and $N_R = 10$, respectively, as similar to that in [11], [12]. The complex gain of each ray is assumed to satisfy the circularly symmetric complex Gaussian distribution with $h_{cl} \sim \mathcal{CN}(0, 1)$. The AoA, $\theta_{cr}$, and AoD, $\theta_{ct}$, are generated according to the Laplacian distribution with random mean cluster angles $\bar{\theta}_{cr} \in [0, 2\pi)$ and $\bar{\theta}_{ct} \in [0, 2\pi)$. 
A. Convergence Behavior

First, we evaluate the convergence properties of the proposed HBF WMMSE-EI and WMMSE-MO algorithms. It is worth noting that the performance of these two iterative algorithms is highly related to the initialization of the beamforming optimization variables. We selected two different initialization methods for testing: one was random initialization (labeled with ‘Random-ini’), the other was to use the low complexity MMSE-EI algorithm for initialization (labeled with ‘MMSE-ini’). Fig. 2 illustrates the performance of average spectral efficiency as a function of the number of outer iterations (the iteration between the hybrid precoding and combining using the alternative minimization) for the proposed WMMSE-EI and WMMSE-MO algorithms with the two initialization methods when SNR is fixed at $-6$ dB. According to Fig. 2 with either of these two initialization methods, the WMMSE-EI and WMMSE-MO algorithms both converge, which verifies the convergence proof in Section V-A. Furthermore, Fig. 2 also shows that the ‘MMSE-ini’ method can significantly speed up the convergence and further improve the spectral efficiency compared with the random initialization. Thus, the ‘MMSE-ini’ method is used for initialization in the following simulations.
B. Spectral efficiency v.s. SNR

Fig. 3 shows the performance of spectral efficiency as a function of SNR for the proposed WMMSE-EI, WMMSE-MO and MMSE-EI algorithms. For comparison, the performance of two conventional HBF algorithms for the partially-connected architecture (labeled with ‘HBF in [10]’ and ‘HBF in [11]’) and that of the optimal fully-digital beamforming (labeled with ‘FD-BF’) are also provided in Fig. 3. It is shown that the proposed WMMSE-EI and WMMSE-MO HBF algorithms perform almost the same, and significantly outperform the conventional HBF algorithm in [10] and that in [11] by about more than 4.5 dB and 2.0 dB in SNR, respectively, for a target spectral efficiency of 8 bits/s/Hz. Besides, the proposed low complexity MMSE-EI algorithm has a gap of about 0.5 dB in SNR when compared to the proposed WMMSE-EI and WMMSE-MO algorithms. Furthermore, it can be seen from this figure that the performance gap between the optimal fully-digital beamforming and the HBF with the partially-connected architecture is relatively large mainly due to the great reduction of the number RF chains and the number of phase shifters in the partially-connected architecture. Nevertheless, the proposed WMMSE based HBF algorithms provide a more promising design approach to balance the performance loss and hardware cost and consumption for the HBF design with the partially-connected architecture.

To verify the generality of the proposed HBF algorithms, we consider two other mmWave MIMO system configurations, where more receive RF chains ($N_{RF}^r = 4$) are employed in Fig. 4 and more transmit antennas ($N = 144$) are further employed in Fig. 5. It can be seen from these two figures that the proposed HBF algorithms can also achieve similar performance improvement over the conventional counterparts.

C. Performance with Finite Resolution Phase Shifters

Considering the fact that practical phase shifters may have limited resolution, we compare the performance of different HBF algorithms with different numbers of quantization bits, denoted by $q$, in Fig. 6 when the SNR is fixed at $-6$ dB. For the WMMSE-EI and MMSE-EI algorithms, the modified algorithms with the consideration of finite resolution proposed in Section IV-B, i.e., the WMMSE-EI-Q and MMSE-EI-Q algorithms, were applied in the simulation. For the WMMSE-MO algorithm, we first obtained the optimized analog beamforming matrices under the condition of infinite resolution and then simply uniformly quantized the phase of each entry with $q$ bits, which is labeled with ‘WMMSE-MO-U’ in the figure. For comparison, we also
provide the performance of the HBF design with finite resolution phase shifters in [11], which is labeled as ‘HBF-Q in [11]’ in the figure. It can be seen from this figure that the proposed HBF algorithms still outperform the conventional counterpart with finite phase shift resolution. Meanwhile, the WMMSE-EI-Q algorithm achieves higher spectral efficiency for small $q$ such as $q = 1$ or $q = 2$ than the uniform quantization method. The MMSE-EI-Q algorithm even outperforms the WMMSE-MO-U algorithm when $q = 1$. Fig. 6 also shows that the performance loss caused by finite resolution is almost negligible when $q \geq 4$.

Finally, Fig. 7 shows the spectral efficiency v.s. SNR for these HBF algorithms in the special case of 1-bit phase shifters. It can be seen from this figure that for all the algorithms with $q = 1$, about $3 \sim 4$ dB more power needs to be paid to achieve the same target spectral efficiency with infinite resolution. Furthermore, the performance gain achieved by the proposed algorithms over the conventional algorithms with infinite resolution is maintained and even increased for system with 1-bit phase shifters by using the WMMSE-EI-Q and MMSE-EI-Q algorithms in Section IV-B.

VII. CONCLUSIONS

We have proposed a WMMSE based design approach to equivalently solve the HBF optimization problem with the objective of maximizing the spectral efficiency. To deal with the highly
non-convex and multivariate WMMSE problem, we separated it into the hybrid precoding and combining subproblems and applied the alternating minimization algorithm to iteratively optimize the hybrid precoder and combiner. The proposed the WMMSE-EI and WMMSE-MO algorithms have been shown to achieve 2dB SNR gain over the conventional algorithms with guaranteed convergence for various MIMO configurations. Furthermore, we have modified the WMMSE based algorithms to the MMSE based ones to reduce the computational complexity and provide much better initial beamformers to the WMMSE based algorithms than the random initialization. Considering the practical finite resolution phase shifters, we have also proposed some modified HBF algorithms, which perform better than the one using the uniform quantization and the conventional algorithm, especially in the case of 1-bit phase shifters.

APPENDIX A

PROOF OF LEMMA 1

Proof: According to [27], the direction of the conjugate gradient indicates the direction of curvature for an unconstrained optimization problem. Therefore, the general conjugate gradient of $f(F_{RF})$ respect to $F_{RF}$ in the Euclidean space without considering the partially-connected
architecture is given by
\[
\frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}^*} = \begin{bmatrix}
\frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(1,1)} & \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(1,2)} & \cdots & \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(1,N_{RF}^t)} \\
\frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(2,1)} & \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(2,2)} & \cdots & \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(2,N_{RF}^t)} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(N_{RF}^t,1)} & \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(N_{RF}^t,2)} & \cdots & \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(N_{RF}^t,N_{RF}^t)}
\end{bmatrix}
\tag{42}
\]

Now considering the partially-connected architecture, as a RF chain is only connected to part of the antennas, the objective function \( f (\mathbf{F}_{RF}) \) is only related to the block-diagonal terms in \( \mathbf{F}_{RF} \) given by \((3a)\). Thus, at those positions where the entries of the matrix \( \mathbf{F}_{RF} \) are equal to zero (i.e., the non-block-diagonal entries), the corresponding partial derivative entries in \((42)\) should be zero. Therefore, the Euclidean conjugate gradient \( \nabla f (\mathbf{F}_{RF}) \) for the partially-connected architecture must be a block-diagonal matrix as follow
\[
\nabla f (\mathbf{F}_{RF}) = \text{blkdiag} \left( \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(1,1)}, \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(1,2)}, \cdots, \frac{\partial f (\mathbf{F}_{RF})}{\partial \mathbf{F}_{RF}(N_{RF}^t,1)} \right) = \nabla_{\mathbf{F}_{RF}}^* f (\mathbf{F}_{RF}) \odot \mathbf{P}_1.
\tag{43}
\]

The proof is completed.

**APPENDIX B**

**PROOF OF PROPOSITION 1**

*Proof:* First, as we have shown in \((13)\) and \((8)\) that the optimal digital precoder (along with the optimal \( \xi_k \) in \((14)\)) and combiner have a closed-form solution obtained via the KKT conditions, for a given analog precoder or combiner, the corresponding digital one always ensures the decrease of the weighted sum-MSE \([32]\). Therefore, it is the optimization of analog beamformers that decides the convergence of the HBF algorithms. For the EI algorithm, as it is an element-by-element optimization algorithm, for each inner iteration, the updated phase shifter element always guarantees that the objective function will not increase. For the MO algorithm, as we have shown in Section III-B2, the inner iteration converges to a local optimal point. Thus, the processing within the hybrid precoding or hybrid combining optimization always ensures the convergence of the objective function.

Next, when the optimization switches from the hybrid precoding to the hybrid combining, as the updated hybrid precoder is taken as a fixed one during the whole inner iteration of the hybrid combining optimization, the objective function of the matrix weighted sum-MSE does not increase. This is also similar to the optimization of the weight matrix and that of the hybrid precoding, where all the other optimization variables are fixed except for the one that needs to be optimized in the current step. The proof is thus completed.
APPENDIX C
PROOF OF PROPOSITION 2

Proof: For the $n$-th outer iteration, after the first and second step, both the hybrid precoder and the combiner have been updated. Then, from (11), we obtain the updated MSE matrix for each subcarrier, which is defined as $E_k^{(n)}$. According to Theorem 1, the optimal weight matrix in the third step of the WMMSE-EI or WMMSE-MO algorithm should be $\Lambda_k^{(n)} = (E_k^{(n)})^{-1}$, the resulting objective function of (28) in the $n$-th iteration is then given by

$$J_{s3}^{(n)} = N_s - \frac{1}{K} \log \left| \left( E_k^{(n)} \right)^{-1} \right| = N_s - R^{(n)},$$

(44)

where the second equality follows from (12) in Theorem 1. According to Proposition 1, where we have shown $J_{s3}^{(n)} \geq J_{s3}^{(n+1)}$, we now have $R^{(n)} \leq R^{(n+1)}$. Therefore, the proposed WMMSE-EI and WMMSE-MO algorithms, which are designed based on the WMMSE criterion, can indeed make the spectral efficiency monotonously increase until convergence. The proof is thus completed.

REFERENCES

[1] 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.
[2] W. Roh, J. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, “Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results,” IEEE Commun. Mag., vol. 52, no. 2, pp. 106–113, Feb. 2014.
[3] 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.
[4] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[5] Y. Xue, X. Zheng, and V. K. N. Lau, “Line-of-sight MIMO for high capacity millimeter wave backhaul in FDD systems,” J. Commun. Inf. Netw., vol. 5, no. 2, pp. 177–193, Jun. 2020.
[6] J. Zhang, X. Yu, and K. B. Letaief, “Hybrid beamforming for 5G and beyond millimeter-wave systems: A holistic view,” IEEE Open J. Commun. Soc., vol. 1, no. 1, pp. 77–91, Dec. 2019.
[7] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” IEEE Commun. Mag., vol. 52, no. 2, pp. 186–195, Feb. 2014.
[8] 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.
[9] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, “Spatially sparse precoding in millimeter wave MIMO systems,” IEEE Trans. Wireless Commun., vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
[10] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, “Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems,” IEEE J. Sel. Topics Signal Process., vol. 10, no. 3, pp. 485–500, Apr. 2016.
[11] F. Sohrabi and W. Yu, “Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1432–1443, Jul. 2017.

[12] T. Lin, J. Cong, Y. Zhu, J. Zhang, and K. B. Letaief, “Hybrid beamforming for millimeter wave systems using the MMSE criterion,” *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3693–3708, May. 2019.

[13] 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.

[14] J. Cong, T. Lin, and Y. Zhu, “Hybrid MMSE beamforming for multiuser millimeter-wave communication systems,” *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2390–2393, Nov. 2018.

[15] D. H. N. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, “Hybrid MMSE precoding and combining designs for mmWave multiuser systems,” *IEEE Access*, vol. 5, pp. 19 167–19 181, 2017.

[16] C. Han, J. Wang, J. Wang, and L. Bai, “Hybrid beamforming design for uplink mmWave systems with a predefined low-resolution codebook,” *J. Commun. Inf. Netw.*, vol. 4, no. 3, pp. 1–8, Sep. 2019.

[17] C. Rusu, R. Méndez-Rial, N. González-Prelcic, and R. W. Heath, “Low complexity hybrid precoding strategies for millimeter wave communication systems,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8380–8393, Dec. 2016.

[18] J. Mirza, B. Ali, S. Saud Naqvi, and S. Saleem, “Hybrid precoding via successive refinement for millimeter wave MIMO communication systems,” *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 991–994, May. 2017.

[19] X. Yu, J. Zhang, and K. B. Letaief, “Doubling phase shifters for efficient hybrid precoder design in millimeter-wave communication systems,” *J. Commun. Inf. Netw.*, vol. 4, no. 2, pp. 51–67, Jun. 2019.

[20] 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. 15 142–15 151, 2017.

[21] H. Sampath, P. Stoica, and A. Paulraj, “Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion,” *IEEE Trans. Commun.*, vol. 49, no. 12, pp. 2198–2206, Dec. 2001.

[22] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, “An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[23] S. S. Christensen, R. Agarwal, E. De 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.

[24] M. Joham, W. Utschick, and J. A. Nossek, “Linear transmit processing in MIMO communications systems,” *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005.

[25] B. Chen, *Optimization Theory and Algorithms*. China: Tsinghua Univ. Press, 2005.

[26] P. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2008.

[27] X. Zhang, *Matrix Analysis and Applications*. China: Tsinghua Univ. Press, 2013.

[28] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[29] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[30] X. Li, Y. Zhu, and P. Xia, “Enhanced analog beamforming for single carrier millimeter wave MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4261–4274, Jul. 2017.

[31] D. Tes and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.