Relay Hybrid Precoding Design in Millimeter-Wave Massive MIMO Systems

Xuan Xue, *Student Member, IEEE*, Yongchao Wang, *Member, IEEE*, Linglong Dai, *Senior Member, IEEE*, Christos Masouros, *Senior Member, IEEE*

***Abstract*—This paper investigates the relay hybrid precoding design in millimeter-wave (mmWave) massive MIMO systems. The optimal design of the relay hybrid precoding is highly non-convex, due to the six-order polynomial objective function, six-order polynomial constraint, and constant-modulus constraints. To efficiently solve this challenging non-convex problem, we first reformulate it into three quadratic subproblems, where one of the subproblems is convex and the other two are non-convex. Then, we propose an iterative successive approximation (ISA) algorithm to attain the high-approximate optimal solution to the original problem. Specifically, in the proposed ISA algorithm, we first convert the two non-convex subproblems to convex ones by the relaxation of the constant-modulus constraints, and then we solve the three corresponding convex subproblems iteratively. We theoretically prove that the ISA algorithm converges to a Karush-Kuhn-Tucker (KKT) point of the original problem. Simulation results demonstrate that the proposed ISA algorithm achieves good performance in terms of achievable rate in both full-connected and sub-connected relay hybrid precoding systems.**

***Index Terms*—Millimeter wave, massive MIMO, relay, hybrid precoding, non-convex optimization.**

I. Introduction

As a promising technology for the next generation of wireless communications, millimeter-wave (mmWave) communication has drawn extensive research interests in the recent years [1] [2]. By utilizing large spectrum bands between 30 GHz and 300 GHz, mmWave communication is capable of meeting the explosive growth of data rate [3]. Although the mmWave signals undergo severe path loss, the path loss can be compensated by high antenna gain using massive multiple-input multiple-output (MIMO) [4]. However, mmWave communications are mainly applied in line-of-sight (LoS) dominant scenarios, since mmWave signals are sensitive to blockage [5].

To mitigate the negative effects caused by blockage, relay can be employed in mmWave massive MIMO systems [6]. In a relay-assisted mmWave system, the channels from the source to the relay and from the relay to the destination may be LoS, and the transmission range and coverage can be extended. Similar to the conventional mmWave massive MIMO system, precoding plays an important role in the relay-assisted mmWave massive MIMO system to compensate for the high path loss by the high antenna gain [7], [8]. However, the optimal precoding design for the relay-assisted mmWave massive MIMO systems is a challenging problem due to the complicated signal processing.

It is well known that the classical full-digital precoding can achieve the optimal antenna gain [9], [10], but it is too costly for mmWave massive MIMO systems. This is caused by the fact that the traditional full-digital precoding demands the same number of radio frequency (RF) chains as that of the antennas, and each RF chain requires costly hardware and high power consumption. Therefore, when a large number of antennas are deployed in mmWave massive MIMO systems, the high hardware cost and power consumption of RF chains make the full-digital precoding unaffordable in practice [11], [12]. To this end, the recently proposed hybrid (analog/digital) precoding is a more attractive alternative, since it achieves the similar performance to the full-digital one with much fewer RF chains [13], [14]. The hybrid precoding is jointly realized in the digital and analog domains, where the digital precoding is realized by baseband signal processing, while the analog precoding is usually implemented by analog phase shifters [15], [16].

The hybrid precoding can be realized by two typical structures: full-connected structure (where each RF chain is connected to all antennas) [17], [18], and sub-connected structure (where each RF chain is connected to a subset of antennas) [19], [20]. For the relay-assisted mmWave system with the full-connected structure, downlink single-user and multi-user hybrid precoding schemes using matching pursuit (MP) algorithms have been studied in [21], [22]. Compared with the full-connected structure, the sub-connected structure is more practically attractive, since it can further reduce the hardware complexity and power consumption without an obvious performance loss. However, to the best of our knowledge, there is no existing work on the topic of sub-connected relay hybrid precoding design for mmWave massive MIMO systems.

In this paper, we try to fill this gap by proposing an efficient relay hybrid precoding algorithm for the sub-connected structure in mmWave massive MIMO systems. The proposed algorithm can be also extended to the full-connected structure. For both the sub-connected and full-connected structures, the

This work was supported in part by National Science Foundation of China under grant 61372135, 61771356, 111 project of China under grant B08038, and National Natural Science Foundation of China for Outstanding Young Scholars (Grant No. 61722109). *(Corresponding author: Yongchao Wang.)*

X. Xue and Y. Wang are with Collaborative Innovation Center of Information Sensing and Understanding, State Key Lab. of ISN, School of Telecom. Engineering, Xidian University, No.2 Taibai South Road, Xian, 710071, Shaanxi, China. (Email: xuexuan601@gmail.com, ychwang@mail.xidian.edu.cn).

L. Dai is with the Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mail: daill@tsinghua.edu.cn).

C. Masouros is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, UK. (Email: chris.masouros@ieee.org).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2018.2799201, IEEE Transactions on Signal Processing

relay hybrid precoding algorithms are designed to minimize the mean squared error (MSE) between the transmitted and received signals with the power constraint. Specially, the main contributions of this paper are summarized as follows.

- For the sub-connected structure, we propose a minimum mean squared error (MMSE)-based relay hybrid precoding design. This challenging problem is highly nonconvex due to the six-order polynomial objective function, six-order polynomial constraint, block-diagonal constraints, and constant-modulus constraints. To eliminate the block-diagonal constraints and reduce the problem dimension, we reformulate the original problem as three subproblems. Here, one of these three subproblems is a convex quadratically constrained quadratic programming (QCQP) problem, while the other two subproblems are non-convex QCQP problems with constant-modulus constraints.
- To solve these three subproblems, we propose an iterative successive approximation (ISA) algorithm with affordable complexity. In the proposed ISA algorithm, we first derive the closed-form solution to the convex QCQP subproblem. Then, for the two non-convex QCQP subproblems, we convert them to be convex by the relaxation of the constant-modulus constraints. Then, the high-approximate solution is obtained by iteratively solving these three convex problems. The theoretical analysis and simulation results demonstrate that the proposed ISA algorithm converges to a Karush-Kuhn-Tucker (KKT) point.
- For the full-connected structure, the relay hybrid precoding design problem is simpler due to the fact that the block-diagonal constraints are not needed. However, the relay hybrid precoding design problem for the full-connected structure is still challenging due to the six-order polynomial objective function and constant-modulus constraints. Fortunately, the proposed ISA algorithm can be extended to solve the hybrid precoding design problem with the full-connected structure. Since the proposed ISA algorithm does not need any predetermined candidates for the analog precoders, it can theoretically achieve better performance than the existing MP algorithm for the full-connected structure.
- Simulation results confirm that the proposed hybrid precoding for the full-connected structure is able to achieve almost the same performance as the classical full-digital precoding. The theoretical analysis and simulation results also demonstrate that the sub-connected structure is able to reduce the power consumption compared to the full-connected structure.

The rest of this paper is organized as follows. Section II briefly introduces the channel model and relay system model. In Section III, the MMSE-based hybrid precoding problem is formulated at first, and then the QCQP reformulation is presented. The ISA algorithm is proposed to solve the reformulated problem in Section IV, where the analysis of the proposed algorithm is also provided. Simulation results are shown to evaluate the performance of the proposed relay hybrid precoding design and the proposed ISA algorithm in Section V, followed by Section VI that concludes this paper.

Notations: In this paper, bold lowercase and uppercase letters denote vectors and matrices, respectively; $(\cdot)^T$ and $(\cdot)^H$ symbolize the transpose and conjugate transpose operations; The 2-norm of a vector $\mathbf{a}$ and the Frobenius norm of a matrix $\mathbf{A}$ are denoted by $||\mathbf{a}||_2$ and $||\mathbf{A}||_{\mathrm{F}}$; $\mathbb{E}[\cdot]$ represents the expectation operator; $\mathrm{Tr}(\cdot)$ and $\mathrm{blk}(\cdot)$ indicate the trace and block-diagonal operator; $\mathrm{vec}(\mathbf{A})$ is the vectorization of a matrix $\mathbf{A}$; $\otimes$ denotes the Kronecker product between two matrices; $\mathcal{CN}(0, \sigma^2)$ represents the zero-mean complex Gaussian distribution with zero mean and the variance σ^2 and the $\mathbf{I}_m$ denotes the $m \times m$ identity matrix; $\arg(\cdot)$ and $\exp(\cdot)$ represent the argument of a complex value and the exponential of a value.

II. System Description

This section briefly introduces the mmWave channel model and the relay hybrid precoding system model with both sub-connected and full-connected structures.

A. Millimeter-Wave Channel Model

As shown in Figs. 1 and 2, we consider an amplify-and-forward (AF) relay assisted mmWave massive MIMO system without direct link between the source and the destination[1]. As can be seen, the considered system includes the channels $\mathbf{H}$ from the source to the relay, and $\mathbf{G}$ from the destination to the relay. Here, we assume that the channels $\mathbf{H}$ and $\mathbf{G}$ are LoS. According to [23], the propagation loss of the LoS channels $\mathbf{H}$ and $\mathbf{G}$ obeys the Rician distribution.

In this paper, we consider the narrowband mmWave channel model widely used in the literatures [14], [21], [22]. The more challenging optimal design of the relay hybrid precoding over "delay-d" wideband mmWave channels [24] is beyond the scope of the current paper and will be studied in our future work.

Due to the fact that mmWave channels exhibit limited number of paths [25] [26], $\mathbf{H}$ and $\mathbf{G}$ often have sparse structures, which can be characterized by low-rank matrices as follows:

$$\mathbf{H} = \sum_{l=1}^{L_h} \alpha_l \mathbf{a}_l^{\mathrm{R}}(\theta_l^{\mathrm{R}})(\mathbf{a}_l^{\mathrm{S}}(\theta_l^{\mathrm{S}}))^H, \tag{1a}$$

$$\mathbf{G} = \sum_{l=1}^{L_g} \gamma_l \mathbf{a}_l^{\mathrm{R}}(\beta_l^{\mathrm{R}})(\mathbf{a}_l^{\mathrm{D}}(\beta_l^{\mathrm{D}}))^H, \tag{1b}$$

where L_h and L_g are the numbers of propagation paths in $\mathbf{H}$ and $\mathbf{G}$, α_l and γ_l are the path loss coefficients of the lth path in $\mathbf{H}$ and $\mathbf{G}$, $\mathbf{a}_l^{\mathrm{R}}$, and $\mathbf{a}_l^{\mathrm{D}}$ are the array response vectors of the source, the relay and the destination. In this paper, we

[1] In practice, the direct link from the source to the destination may not physically exist due to the blockage in the mmWave systems. However, by adopting the relays, the channels from the source to relay and from the relay to the destination may be LoS. Therefore, in this paper, we consider the mmWave relay system without direct link from the source to the destination.

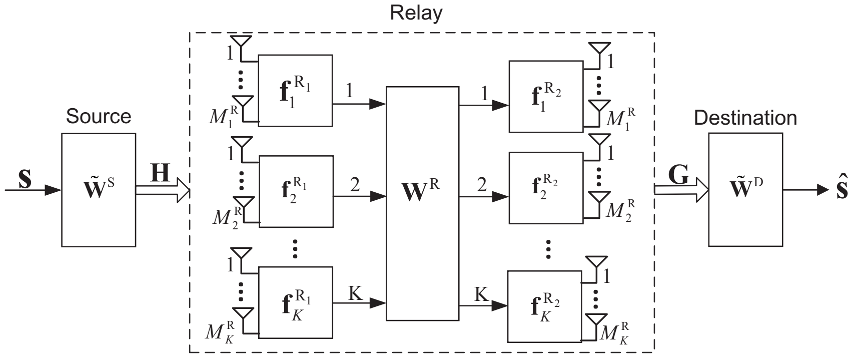
Relay
Source
Destination
s
$\tilde{\mathbf{W}}^{\mathrm{S}}$
H
$\mathbf{f}_1^{\mathrm{R}_1}$
$\mathbf{f}_2^{\mathrm{R}_1}$
$\mathbf{f}_K^{\mathrm{R}_1}$
$\mathbf{W}^{\mathrm{R}}$
$\mathbf{f}_1^{\mathrm{R}_2}$
$\mathbf{f}_2^{\mathrm{R}_2}$
$\mathbf{f}_K^{\mathrm{R}_2}$
1
2
K
M_1^{R}
M_2^{R}
M_K^{R}
G
$\tilde{\mathbf{W}}^{\mathrm{D}}$
$\hat{\mathbf{s}}$

Fig. 1. Relay hybrid precoding with the sub-connected structure.

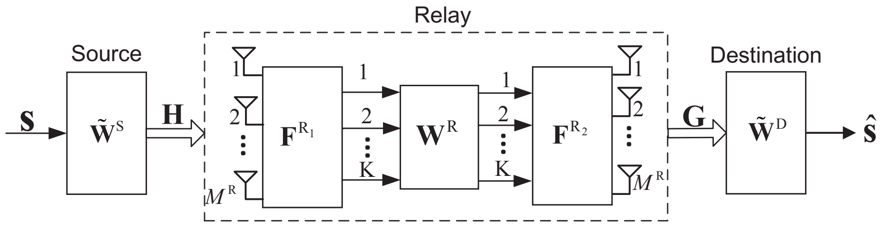
Relay
Source
Destination
s
$\tilde{\mathbf{W}}^{\mathrm{S}}$
H
1
2
M^{R}
$\mathbf{F}^{\mathrm{R}_1}$
K
$\mathbf{W}^{\mathrm{R}}$
$\mathbf{F}^{\mathrm{R}_2}$
G
$\tilde{\mathbf{W}}^{\mathrm{D}}$
$\hat{\mathbf{s}}$

Fig. 2. Relay hybrid precoding with the full-connected structure.

consider the widely used uniform linear arrays (ULA), where the array response vectors can be expressed as

$$\begin{aligned}
\mathbf{a}_l^{\mathrm{S}}(\theta) =& \frac{1}{\sqrt{M^{\mathrm{S}}}}[1, \exp^{j\frac{2\pi}{\lambda}d\sin(\theta)}, \cdots, \exp^{j(M^{\mathrm{S}}-1)\frac{2\pi}{\lambda}d\sin(\theta)}]^T,\\
\mathbf{a}_l^{\mathrm{R}}(\theta) =& \frac{1}{\sqrt{M^{\mathrm{R}}}}[1, \exp^{j\frac{2\pi}{\lambda}d\sin(\theta)}, \cdots, \exp^{j(M^{\mathrm{R}}-1)\frac{2\pi}{\lambda}d\sin(\theta)}]^T,\\
\mathbf{a}_l^{\mathrm{D}}(\theta) =& \frac{1}{\sqrt{M^{\mathrm{D}}}}[1, \exp^{j\frac{2\pi}{\lambda}d\sin(\theta)}, \cdots, \exp^{j(M^{\mathrm{D}}-1)\frac{2\pi}{\lambda}d\sin(\theta)}]^T,
\end{aligned} \tag{2}$$

where d and λ are the antenna spacing and the wave length, respectively.

To realize the precoding, we follow the assumption in [18], [27] that the channels $\mathbf{H}$ and $\mathbf{G}$ are known at the source, relay, and the destination. In practical systems, channel state information (CSI) received at the relay can be obtained via training from the source to the relay [27], and the CSI received at the destination can be obtained via training from the relay to the destination. Then the CSI is shared with the transmitter at the source via feedback from the relay to the source [28], and the CSI transmitted at the source is shared by the feedback from the relay to the destination.

B. Relay Hybrid Precoding

For both the full-connected and the sub-connected structures as shown in Figs. 1 and 2, the relay employs the hybrid precoding, where K RF chains and M^{R} antennas are used. The difference between these two structures is that, the kth $(k = 1, 2, \cdots, K)$ RF chain is connected to a subset of M_k^{R} antennas in the sub-connected structure, while the kth $(k = 1, 2, \cdots, K)$ RF chain is connected to all M^{R} antennas in the full-connected structure. For the sub-connected structure, the total number of antennas at the relay is $M^{\mathrm{R}} = \sum_{k=1}^{K} M_k^{\mathrm{R}}$. It's worth noting that the number of antennas M^{R} is much larger than that of RF chains K to achieve high antenna gain. At the source and the destination, the numbers of antennas are M^{S} and M^{D}, respectively. The transmitted symbols for L_s data streams at the source are represented as $\mathbf{s} \in \mathbb{C}^{L_s}$ with normalized power $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}_{L_s}$.

In the following, we will focus on sub-connected structure as shown in Fig. 1, since it is more complicated for the optimal precoder design problem due to more constraints compared to the full-connected one. The extension to the full-connected structure will be discussed in Section IV-C.

Filtered by the source precoding matrix $\tilde{\mathbf{W}}^{\mathrm{S}} \in \mathbb{C}^{M^{\mathrm{S}} \times L_s}$, the received signal $\mathbf{y}^{\mathrm{R}}$ at the relay is

$$\mathbf{y}^{\mathrm{R}} = \underbrace{\mathbf{H}\tilde{\mathbf{W}}^{\mathrm{S}}}_{\triangleq \bar{\mathbf{H}}}\mathbf{s} + \mathbf{n}^{\mathrm{R}}, \tag{3}$$

where $\mathbf{H} \in \mathbb{C}^{M^{\mathrm{R}} \times M^{\mathrm{S}}}$ is the channel matrix between the source and the relay, and $\mathbf{n}^{\mathrm{R}} \sim \mathcal{CN}(\mathbf{0}, \sigma_r^2 \mathbf{I}_{M^{\mathrm{R}}})$ is the additive noise vector at the relay.

At the relay, hybrid precoding is carried out. More specially, the received analog combiner $\mathbf{F}^{\mathrm{R}_1} \in \mathbb{C}^{M^{\mathrm{R}} \times K}$ is firstly employed for received signal $\mathbf{y}^{\mathrm{R}}$. Then, one digital precoding $\mathbf{W}^{\mathrm{R}} \in \mathbb{C}^{K \times K}$ processes the signal in the baseband. Afterwards, the analog precoder $\mathbf{F}^{\mathrm{R}_2} \in \mathbb{C}^{M^{\mathrm{R}} \times K}$ is used to forward the transmitted signal at the relay to the destination. Because each RF chain only connects to a subset of M_k^{R} antennas in the sub-connected structure, the analog combiner $\mathbf{F}^{\mathrm{R}_1}$ and the analog precoder $\mathbf{F}^{\mathrm{R}_2}$ at the relay are constrained to be block-diagonal as follows:

$$\mathbf{F}^{\mathrm{R}_1} = \mathrm{blk}(\mathbf{f}_1^{\mathrm{R}_1}, \mathbf{f}_2^{\mathrm{R}_1}, \cdots, \mathbf{f}_K^{\mathrm{R}_1}), \tag{4a}$$

$$\mathbf{F}^{\mathrm{R}_2} = \mathrm{blk}(\mathbf{f}_1^{\mathrm{R}_2}, \mathbf{f}_2^{\mathrm{R}_2}, \cdots, \mathbf{f}_K^{\mathrm{R}_2}), \tag{4b}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2018.2799201, IEEE Transactions on Signal Processing

where $\mathbf{f}_k^{\mathrm{R}_1} \in \mathbb{C}^{M_k^{\mathrm{R}} \times 1}(k = 1, 2, \cdots, K)$, and $\mathbf{f}_l^{\mathrm{R}_2} \in \mathbb{C}^{M_l^{\mathrm{R}} \times 1}(l = 1, 2, \cdots, K)$.

Since the non-zero elements in the analog precoding matrices $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are usually realized by phase shifters [14], the non-zero elements satisfy the constant-modulus constraints as follows:

$$|\mathbf{f}_k^{\mathrm{R}_1}| = \mathbf{1}, i = 1, 2, \cdots, K, \tag{5a}$$

$$|\mathbf{f}_l^{\mathrm{R}_2}| = \mathbf{1}, l = 1, 2, \cdots, K. \tag{5b}$$

Using the three matrices $\{\mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}\}$, the transmitted signal $\mathbf{x}^{\mathrm{R}}$ at the relay can be expressed as

$$\mathbf{x}^{\mathrm{R}} = \mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\bar{\mathbf{H}}\mathbf{s} + \mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{n}^{\mathrm{R}}. \tag{6}$$

In practice, the power of the transmitted signal at the relay is constrained as

$$\mathbb{E}(\|\mathbf{x}^{\mathrm{R}}\|_2^2) \leq P_{\mathrm{R}}, \tag{7}$$

where

$$\mathbb{E}(\|\mathbf{x}^{\mathrm{R}}\|_2^2) = \mathrm{Tr}\Big(\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H(\bar{\mathbf{H}}\bar{\mathbf{H}}^H + \sigma_r^2\mathbf{I}_{M^{\mathrm{R}}}) \times (\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H)^H\Big).$$

Finally, a received combining matrix $\tilde{\mathbf{W}}^{\mathrm{D}} \in \mathbb{C}^{M^{\mathrm{D}} \times L_s}$ is used at the destination, so the received signal $\mathbf{y}$ at the destination can be expressed as

$$\begin{aligned}\mathbf{y} =& (\tilde{\mathbf{W}}^{\mathrm{D}})^H\mathbf{G}^H\mathbf{x}^{\mathrm{R}} + (\tilde{\mathbf{W}}^{\mathrm{D}})^H\mathbf{n}^{\mathrm{D}} \\ =& (\tilde{\mathbf{W}}^{\mathrm{D}})^H\mathbf{G}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\bar{\mathbf{H}}\mathbf{s} \\ &+ (\tilde{\mathbf{W}}^{\mathrm{D}})^H\mathbf{G}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{n}^{\mathrm{R}} + (\tilde{\mathbf{W}}^{\mathrm{D}})^H\mathbf{n}^{\mathrm{D}},\end{aligned} \tag{8}$$

where $\mathbf{G} \in \mathbb{C}^{M^{\mathrm{R}} \times M^{\mathrm{D}}}$ is the channel matrix between the destination and the relay, and $\mathbf{n}^{\mathrm{D}} \sim \mathcal{CN}(\mathbf{0}, \sigma_d^2\mathbf{I}_{M^{\mathrm{D}}})$ is the noise at the destination.

Let $\bar{\mathbf{G}} = \mathbf{G}\tilde{\mathbf{W}}^{\mathrm{D}}$ and $\bar{\mathbf{n}}^{\mathrm{D}} = (\tilde{\mathbf{W}}^{\mathrm{D}})^H\mathbf{n}^{\mathrm{D}}$ be the equivalent channel and noise, then the received signal $\mathbf{y}$ at the destination can be rewritten as

$$\mathbf{y} = \bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\bar{\mathbf{H}}\mathbf{s} + \bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{n}^{\mathrm{R}} + \bar{\mathbf{n}}^{\mathrm{D}}. \tag{9}$$

For the full-connected structure, the analog precoder and combiner at the relay are also constrained to be constant-modulus constraints, and the received signal at the destination can also be expressed as equation (9). However, the analog precoder and combiner in the full-connected structure are not constrained to be block-diagonal. Therefore, the hybrid precoding design for the full-connected structure is simpler than that for the sub-connected one, since fewer constraints are required for the full-connected structure (i.e., only (5) is required, but (4) is not required any more).

Throughout this paper, we assume that the hybrid precoder at the source $\tilde{\mathbf{W}}^{\mathrm{S}}$ and the hybrid combiner at the destination $\tilde{\mathbf{W}}^{\mathrm{D}}$ have already been obtained (please see Appendix A for details). Therefore, we only focus on designing the optimal combiner and precoder for the hybrid precoding at the relay[2].

III. MMSE-Based Relay Hybrid Precoding Design

This section discusses the relay hybrid precoding design in mmWave massive MIMO systems for the sub-connected structure. The design goal is to minimize MSE between the transmitted symbols $\mathbf{s}$ at the source and the received signals $\mathbf{y}$ at the destination, with the constraint of the transmitted power at the relay. In this regard, we first present the original problem formulation with block-diagonal and constant-modulus constraints (i.e., (4) and (5)). Then, we eliminate the block-diagonal constraints and reformulate the original problem as three QCQP subproblems with constant-modulus constraints.

A. Original MMSE-Based Relay Hybrid Precoding Problem

As mentioned in the above section, the analog precoders $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are not only constrained to be block-diagonal, but also their non-zero elements are constrained to be constant-modulus. By utilizing (4), (5), and (7), the MMSE-based hybrid precoding problem is formulated as follows:

$$\begin{aligned}&\min_{\mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}} \ \mathbb{E}\Big(||\mathbf{s} - \mathbf{y}||_2^2\Big), \\ &\quad \text{s.t.} \quad (7),\ (4\text{a}),\ (4\text{b}),\ (5\text{a}),\ (5\text{b}),\end{aligned} \tag{10}$$

where $\mathbf{y}$ is given by (9), and P_{R} represents the maximum transmitted power at the relay.

Note that the expectation operation $\mathbb{E}(||\mathbf{s} - \mathbf{y}||_2^2)$ in the original MMSE problem (10) is difficult to be handled. For this reason, we propose the following ***Lemma 1*** to convert (10) to an equivalent problem without the expectation operation.

Lemma 1: The MMSE-based optimal hybrid precoding for (10) can be obtained by solving

$$\begin{aligned}(\mathcal{P}) \quad &\min_{\mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}} ||\tilde{\mathbf{W}}^{\mathrm{R}}\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}} - \bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}||_{\mathrm{F}}^2, \\ &\quad \text{s.t.} \quad (7),\ (4\text{a}),\ (4\text{b}),\ (5\text{a}),\ (5\text{b}),\end{aligned}$$

where

$$\begin{cases}\mathbf{R}_{\mathbf{y}^{\mathrm{R}}} \triangleq \mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H] = \bar{\mathbf{H}}\bar{\mathbf{H}}^H + \sigma_r^2\mathbf{I}_{M^{\mathrm{R}}}, \\ \tilde{\mathbf{W}}^{\mathrm{R}} \triangleq \mathbb{E}[\mathbf{s}(\mathbf{y}^{\mathrm{R}})^H]\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H]^{-1} \\ \qquad = \bar{\mathbf{H}}^H(\bar{\mathbf{H}}\bar{\mathbf{H}}^H + \sigma_r^2\mathbf{I}_{M^{\mathrm{R}}})^{-1}.\end{cases} \tag{11}$$

Proof: See the detailed proof in Appendix B[3]. ■

Unfortunately, although the converted optimization problem $(\mathcal{P})$ does not have the expectation operation any more, it is still challenging to be solved due to the following reasons. First, the objective function in $(\mathcal{P})$ and the left-hand side of constraint (7) are sixth-order polynomials, because there are three precoding matrices $\{\mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}\}$ to be jointly

[2] The source hybrid precoding and destination hybrid combining designs for a relay system are similar to the transmitter precoding and receiver combining designs in a MIMO system. The transmitter precoding and receiver combining designs in the mmWave MIMO system have been extensively studied in existing literature. Specifically, it should be pointed out that the hybrid precoding matrices $\tilde{\mathbf{W}}^{\mathrm{S}}$ and $\tilde{\mathbf{W}}^{\mathrm{D}}$ can be obtained by the algorithms in [14].

[3] Since the source precoder exploits the mmWave channel characteristics, the formulation of the original problem $(\mathcal{P})$ has adopted the sparse characteristics of the mmWave channels.

optimized. Second, due to the hardware constraints in the sub-connected structure, the analog precoding and combining matrices $\{\mathbf{F}^{\mathrm{R}_1}, \mathbf{F}^{\mathrm{R}_2}\}$ are required to be block-diagonal. Third, the non-zero elements in the analog precoders $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are constrained to be constant-modulus.

B. Reformulation of The MMSE Problem

To tackle the original high-dimensional optimization problem $(\mathcal{P})$, we decompose it into three quadratic subproblems, where one of the subproblems is convex and the other two are non-convex. Then, to facilitate the mathematical tractability, we eliminate the block-diagonal constraints (4), and reformulate the non-convex subproblems as QCQP problems with constant-modulus constraints (5).

In order to guarantee the convergence of the proposed algorithm in the following Section IV, we optimize the three QCQP subproblems as follows. First we optimize the digital precoder $\mathbf{W}^{\mathrm{R}}$ while keeping the received analog precoder $\mathbf{F}^{\mathrm{R}_1}$ and transmitted analog precoder $\mathbf{F}^{\mathrm{R}_2}$ fixed. Then for fixed $\mathbf{W}^{\mathrm{R}}$ and $\mathbf{F}^{\mathrm{R}_2}$, we optimize $\mathbf{F}^{\mathrm{R}_1}$. At last, we optimize $\mathbf{F}^{\mathrm{R}_2}$ with fixed $\mathbf{W}^{\mathrm{R}}$ and $\mathbf{F}^{\mathrm{R}_1}$. The detailed reformulation of these three subproblems are given as follows.

1) QCQP Subproblem-1: When $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are fixed, the problem $(\mathcal{P})$ can be reformulated as

$$(\mathcal{P}_1) \quad \min_{\mathbf{W}^{\mathrm{R}}} ||\tilde{\mathbf{W}}^{\mathrm{R}} \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}} - \bar{\mathbf{G}}^H \mathbf{F}^{\mathrm{R}_2} \mathbf{W}^{\mathrm{R}} (\mathbf{F}^{\mathrm{R}_1})^H \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}||_{\mathrm{F}}^2, \tag{12a}$$

$$\text{s.t.} \quad (7). \tag{12b}$$

2) QCQP Subproblem-2: Second, when $\mathbf{W}^{\mathrm{R}}$ and $\mathbf{F}^{\mathrm{R}_2}$ are fixed, the problem $(\mathcal{P})$ can be converted to

$$\begin{aligned} \min_{\mathbf{F}^{\mathrm{R}_1}} \quad & ||\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}} (\tilde{\mathbf{W}}^{\mathrm{R}})^H - \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}} \mathbf{F}^{\mathrm{R}_1} \mathbf{G}_1^H||_{\mathrm{F}}^2, \\ \text{s.t.} \quad & (7),\ (4\text{a}),\ (5\text{a}), \end{aligned} \tag{13}$$

where $\mathbf{G}_1 = \bar{\mathbf{G}}^H \mathbf{F}^{\mathrm{R}_2} \mathbf{W}^{\mathrm{R}}$.

To tackle the block-diagonal constraint (4a), we transform the matrix operator into a vector operator, based on the properties of Kronecker product: $\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A}) \mathrm{vec}(\mathbf{X})$. Then, the problem (13) can be rewritten as

$$\begin{aligned} \min_{\tilde{\mathbf{f}}^{\mathrm{R}_1}} \quad & ||\mathbf{a}^{\mathrm{R}_1} - \mathbf{A}^{\mathrm{R}_1} \tilde{\mathbf{f}}^{\mathrm{R}_1}||_2^2, \\ \text{s.t.} \quad & ||\mathbf{C}^{\mathrm{R}_1} \tilde{\mathbf{f}}^{\mathrm{R}_1}||_2^2 \leq P_{\mathrm{R}}, \\ & \tilde{\mathbf{f}}^{\mathrm{R}_1} = \mathrm{vec}(\mathrm{blk}(\mathbf{f}_1^{\mathrm{R}_1}, \mathbf{f}_2^{\mathrm{R}_1}, \cdots, \mathbf{f}_K^{\mathrm{R}_1})), \\ & |\mathbf{f}_k^{\mathrm{R}_1}| = \mathbf{1}, \quad \forall k = 1, 2, \cdots, K, \end{aligned} \tag{14}$$

where

$$\begin{cases} \mathbf{a}^{\mathrm{R}_1} \triangleq \mathrm{vec}(\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}} (\tilde{\mathbf{W}}^{\mathrm{R}})^H), \\ \mathbf{A}^{\mathrm{R}_1} \triangleq ((\mathbf{G}_1^H)^T \otimes \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}), \\ \mathbf{C}^{\mathrm{R}_1} \triangleq (((\mathbf{F}^{\mathrm{R}_2} \mathbf{W}^{\mathrm{R}})^H)^T \otimes \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}). \end{cases}$$

It is observed from (14) that the variable $\tilde{\mathbf{f}}^{\mathrm{R}_1}$ has many zero elements. There is a property of matrix multiplication, which is presented in the following ***Lemma 2***, to show that the zero elements in $\tilde{\mathbf{f}}^{\mathrm{R}_1}$ can be removed.

Lemma 2: Consider a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N] \in \mathbb{C}^{M \times N}$, and a vector $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T \in \mathbb{C}^{N \times 1}$, where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$, and x_i denotes the ith element of $\mathbf{x}$. If the ith $(i = 1, 2, \cdots, N)$ element in $\mathbf{x}$ is zero, we will have

$$\mathbf{Ax} = \hat{\mathbf{A}} \hat{\mathbf{x}},$$

where $\hat{\mathbf{A}} = [\mathbf{a}_1, \cdots, \mathbf{a}_{i-1}, \mathbf{a}_{i+1}, \cdots, \mathbf{a}_N] \in \mathbb{C}^{M \times (N-1)}$ and $\hat{\mathbf{x}} = [x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N]^T \in \mathbb{C}^{(N-1) \times 1}$.

Proof: The proof of this lemma can be easily obtained by the fact that the zero elements do not provide any contribution in matrix multiplying operation. ■

Now, we extract the non-zero elements in $\tilde{\mathbf{f}}^{\mathrm{R}_1}$, leading to a new variable vector given by

$$\hat{\mathbf{f}}^{\mathrm{R}_1} = [(\mathbf{f}_1^{\mathrm{R}_1})^T, (\mathbf{f}_2^{\mathrm{R}_1})^T, \cdots, (\mathbf{f}_K^{\mathrm{R}_1})^T]^T. \tag{15}$$

Replacing $\mathbf{F}^{\mathrm{R}_1}$ by $\hat{\mathbf{f}}^{\mathrm{R}_1}$, we can see that the constraint (4a) is eliminated. According to ***Lemma 2***, an equivalent problem for (14) is obtained as

$$(\mathcal{P}_2) \qquad \min_{\hat{\mathbf{f}}^{\mathrm{R}_1}} ||\mathbf{a}^{\mathrm{R}_1} - \hat{\mathbf{A}}^{\mathrm{R}_1} \hat{\mathbf{f}}^{\mathrm{R}_1}||_2^2, \tag{16a}$$

$$\text{s.t.}\ ||\hat{\mathbf{C}}^{\mathrm{R}_1} \hat{\mathbf{f}}^{\mathrm{R}_1}||_2^2 \leq P_{\mathrm{R}}, \tag{16b}$$

$$|\hat{\mathbf{f}}^{\mathrm{R}_1}| = \mathbf{1}, \tag{16c}$$

where $\hat{\mathbf{A}}^{\mathrm{R}_1}$ and $\hat{\mathbf{C}}^{\mathrm{R}_1}$ are generated by removing the corresponding columns of matrices $\mathbf{A}^{\mathrm{R}_1}$ and $\mathbf{C}^{\mathrm{R}_1}$. The vector "**1**" means an all-one vector. By mathematical manipulation, $\hat{\mathbf{A}}^{\mathrm{R}_1}$ and $\hat{\mathbf{C}}^{\mathrm{R}_1}$ can be equivalently obtained by

$$\begin{cases} \hat{\mathbf{A}}^{\mathrm{R}_1} \triangleq & [\hat{\mathbf{G}}_1^H \otimes \mathbf{R}_1, \hat{\mathbf{G}}_2^H \otimes \mathbf{R}_2, \cdots, \hat{\mathbf{G}}_K^H \otimes \mathbf{R}_K], \\ \hat{\mathbf{C}}^{\mathrm{R}_1} \triangleq & [(\mathbf{T}_1^H)^T \otimes \mathbf{R}_1, (\mathbf{T}_2^H)^T \otimes \mathbf{R}_2, \cdots, (\mathbf{T}_K^H)^T \otimes \mathbf{R}_K], \end{cases}$$

where $\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}} = [\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_K]$, $\mathbf{G}_1 = [\hat{\mathbf{G}}_1, \hat{\mathbf{G}}_2, \cdots, \hat{\mathbf{G}}_K]$, and $\mathbf{F}^{\mathrm{R}_2} \mathbf{W}^{\mathrm{R}} = [\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_K]$.

3) QCQP Subproblem-3: When $\mathbf{W}^{\mathrm{R}}$ and $\mathbf{F}^{\mathrm{R}_1}$ are fixed, the problem $(\mathcal{P})$ is expressed as

$$\begin{aligned} \min_{\mathbf{F}^{\mathrm{R}_2}} \quad & ||\tilde{\mathbf{W}}^{\mathrm{R}} \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}} - \bar{\mathbf{G}}^H \mathbf{F}^{\mathrm{R}_2} \mathbf{W}^{\mathrm{R}} (\mathbf{F}^{\mathrm{R}_1})^H \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}||_{\mathrm{F}}^2, \\ \text{s.t.} \quad & (7),\ (4\text{b}),\ (5\text{b}). \end{aligned} \tag{17}$$

Similarly, based on the property of Kronecker product, the problem (17) can be converted to the following problem

$$\begin{aligned} \min_{\tilde{\mathbf{f}}^{\mathrm{R}_2}} \quad & ||\mathbf{a}^{\mathrm{R}_2} - \mathbf{A}^{\mathrm{R}_2} \tilde{\mathbf{f}}^{\mathrm{R}_2}||_2^2, \\ \text{s.t.} \quad & ||\mathbf{C}^{\mathrm{R}_2} \tilde{\mathbf{f}}^{\mathrm{R}_2}||_2^2 \leq P_{\mathrm{R}}, \\ & \tilde{\mathbf{f}}^{\mathrm{R}_2} = \mathrm{vec}(\mathrm{blk}(\mathbf{f}_1^{\mathrm{R}_2}, \mathbf{f}_2^{\mathrm{R}_2}, \cdots, \mathbf{f}_K^{\mathrm{R}_2})), \\ & |\mathbf{f}_k^{\mathrm{R}_2}| = \mathbf{1}, \quad \forall k = 1, 2, \cdots, K, \end{aligned} \tag{18}$$

where

$$\begin{cases} \mathbf{a}^{\mathrm{R}_2} \triangleq \mathrm{vec}(\tilde{\mathbf{W}}^{\mathrm{R}} \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}), \\ \mathbf{A}^{\mathrm{R}_2} \triangleq ((\mathbf{W}^{\mathrm{R}} (\mathbf{F}^{\mathrm{R}_1})^H \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}})^T \otimes \bar{\mathbf{G}}^H), \\ \hat{\mathbf{C}}^{\mathrm{R}_2} \triangleq ((\mathbf{W}^{\mathrm{R}} (\mathbf{F}^{\mathrm{R}_1})^H \mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}})^T \otimes \mathbf{I}_{M^{\mathrm{R}}}). \end{cases}$$

Following the similar derivations from (14) to $(\mathcal{P}_2)$, the problem (18) can also be converted to the following problem by introducing a new variable $\hat{\mathbf{f}}^{\mathrm{R}_2} = [(\mathbf{f}_1^{\mathrm{R}_2})^T, (\mathbf{f}_2^{\mathrm{R}_2})^T, \cdots,$

$(\mathbf{f}_K^{\mathrm{R}_2})^T]^T$:

$$(\mathcal{P}_3)\quad \min_{\hat{\mathbf{f}}^{\mathrm{R}_2}} ||\mathbf{a}^{\mathrm{R}_2} - \hat{\mathbf{A}}^{\mathrm{R}_2}\hat{\mathbf{f}}^{\mathrm{R}_2}||_2^2, \tag{19a}$$

$$\text{s.t. } ||\hat{\mathbf{C}}^{\mathrm{R}_2}\hat{\mathbf{f}}^{\mathrm{R}_2}||_2^2 \leq P_{\mathrm{R}}, \tag{19b}$$

$$|\hat{\mathbf{f}}^{\mathrm{R}_2}| = \mathbf{1}, \tag{19c}$$

with

$$\begin{cases} \mathbf{A}^{\mathrm{R}_2} \triangleq [\mathbf{t}_1^{\mathrm{R}_2} \otimes \bar{\mathbf{G}}_1, \mathbf{t}_2^{\mathrm{R}_2} \otimes \bar{\mathbf{G}}_2, \cdots, \mathbf{t}_K^{\mathrm{R}_2} \otimes \bar{\mathbf{G}}_K], \\ \mathbf{C}^{\mathrm{R}_2} \triangleq [\mathbf{t}_1^{\mathrm{R}_2} \otimes \bar{\mathbf{I}}_1, \mathbf{t}_2^{\mathrm{R}_2} \otimes \bar{\mathbf{I}}_2, \cdots, \mathbf{t}_K^{\mathrm{R}_2} \otimes \bar{\mathbf{I}}_K], \end{cases}$$

where $\bar{\mathbf{G}} = [\bar{\mathbf{G}}_1, \bar{\mathbf{G}}_2, \cdots, \bar{\mathbf{G}}_K]^H$, $\mathbf{I}_{M^{\mathrm{R}}} = [\bar{\mathbf{I}}_1, \bar{\mathbf{I}}_2, \cdots, \bar{\mathbf{I}}_K]$, and $(\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}})^T = [\mathbf{t}_1^{\mathrm{R}_2}, \mathbf{t}_2^{\mathrm{R}_2}, \cdots, \mathbf{t}_K^{\mathrm{R}_2}]$.

Up to now, one can understand that the original hybrid precoding problem $(\mathcal{P})$ has been decomposed into three subproblems $(\mathcal{P}_1)$, $(\mathcal{P}_2)$, and $(\mathcal{P}_3)$. The subproblem $(\mathcal{P}_1)$ is convex, but both the subproblems $(\mathcal{P}_2)$ and $(\mathcal{P}_3)$ are non-convex, due to the non-convex constant-modulus constraints (16c) and (19c). In the following Section IV, we will propose the ISA algorithm to iteratively solve these three subproblems.

IV. Proposed Iterative Successive Approximation Algorithm

In this section, we will propose the ISA algorithm to attain the original problem's high-approximate optimal solution. The key idea of the ISA algorithm is as follows. First, for the nth iteration, $(\mathcal{P}_2)$ and $(\mathcal{P}_3)$ are reformulated to approximated problems denoted as $(\mathcal{Q}_2^{(n)})$ and $(\mathcal{Q}_3^{(n)})$, which are completely equivalent to the convex real-valued QCQP subproblems $(\tilde{\mathcal{Q}}_2^{(n)})$ and $(\tilde{\mathcal{Q}}_3^{(n)})$. Meanwhile, a closed-form solution to $(\mathcal{P}_1)$ can be obtained. Then, the three convex problems $(\mathcal{P}_1), (\tilde{\mathcal{Q}}_2^{(n)}), (\tilde{\mathcal{Q}}_3^{(n)})$ are iteratively solved, until the stop criterion being satisfied. Finally, we theoretically prove that the ISA algorithm converges to a KKT point of the original precoding problem $(\mathcal{P})$, and present the complexity analysis of the proposed ISA algorithm.

A. Closed-Form Solution to $(\mathcal{P}_1)$

For the subproblem $(\mathcal{P}_1)$, we can obtain a closed-form solution as follows. Let $\mathbf{A}_W = \tilde{\mathbf{W}}^{\mathrm{R}}\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}$, $\hat{\mathbf{G}}_W = \bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}$, and $\hat{\mathbf{H}}_W = (\mathbf{F}^{\mathrm{R}_1})^H\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}$. Then, the objective function of $(\mathcal{P}_1)$ can be rewritten as

$$\begin{aligned} &||\mathbf{A}_W - \hat{\mathbf{G}}_W\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W||_{\mathrm{F}}^2 \\ &= \mathrm{Tr}\Big((\mathbf{A}_W - \hat{\mathbf{G}}_W\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W)^H(\mathbf{A}_W - \hat{\mathbf{G}}_W\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W)\Big). \end{aligned} \tag{20}$$

Similarly, considering the definition of $\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}$ in (11), the left-hand side of (7) can be rewritten as

$$\begin{aligned} &\mathrm{Tr}\Big(\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}(\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H)^H\Big) \\ &= \mathrm{Tr}\Big(\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W\hat{\mathbf{H}}_W^H(\mathbf{W}^{\mathrm{R}})^H(\mathbf{F}^{\mathrm{R}_2})^H\Big). \end{aligned} \tag{21}$$

According to (20) and (21), the Lagrangian function of $(\mathcal{P}_1)$ is given by

$$\begin{aligned} L(\mathbf{W}^{\mathrm{R}}, \lambda_1) =& \mathrm{Tr}((\mathbf{A}_W - \hat{\mathbf{G}}_W\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W)^H \\ &\times (\mathbf{A}_W - \hat{\mathbf{G}}_W\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W)) \\ &+ \lambda_1(\mathrm{Tr}(\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W\hat{\mathbf{H}}_W^H \\ &\times (\mathbf{W}^{\mathrm{R}})^H(\mathbf{F}^{\mathrm{R}_2})^H) - P_{\mathrm{R}}), \end{aligned} \tag{22}$$

where $\lambda_1 \geq 0$ is the Lagrangian multiplier to satisfy the power constraint (7).

The derivative of (22) with respect to $\mathbf{W}^{\mathrm{R}}$ is

$$\begin{aligned} \frac{\partial L(\mathbf{W}^{\mathrm{R}}, \lambda_1)}{\partial \mathbf{W}^{\mathrm{R}}} =& 2(\hat{\mathbf{G}}_W^H\hat{\mathbf{G}}_W + \lambda_1(\mathbf{F}_W^{\mathrm{R}_2})^H\mathbf{F}_W^{\mathrm{R}_2})\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W\hat{\mathbf{H}}_W^H \\ &- 2\hat{\mathbf{G}}_W^H\mathbf{A}_W\hat{\mathbf{H}}_W^H. \end{aligned} \tag{23}$$

It can be seen that a closed-form solution to $(\mathcal{P}_1)$ can be obtained by letting $\frac{\partial L(\mathbf{W}^{\mathrm{R}},\lambda)}{\partial \mathbf{W}^{\mathrm{R}}} = 0$, which is expressed as follows:

$$\begin{aligned} \mathbf{W}^{\mathrm{R}} =& (\hat{\mathbf{G}}_W^H\hat{\mathbf{G}}_W + \lambda_1(\mathbf{F}^{\mathrm{R}_2})^H\mathbf{F}^{\mathrm{R}_2})^{-1} \\ &\times \hat{\mathbf{G}}_W^H\mathbf{A}_W\hat{\mathbf{H}}_W^H(\hat{\mathbf{H}}_W\hat{\mathbf{H}}_W^H)^{-1}. \end{aligned} \tag{24}$$

Here, the Lagrangian multiplier λ_1 satisfies the following equation

$$\lambda_1(\mathrm{Tr}(\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}\hat{\mathbf{H}}_W\hat{\mathbf{H}}_W^H(\mathbf{W}^{\mathrm{R}})^H(\mathbf{F}^{\mathrm{R}_2})^H) - P_{\mathrm{R}}) = 0, \tag{25}$$

where λ_1 can be obtained by the bisection method [29].

We can see that for fixed $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$, the optimal digital precoder $\mathbf{W}^{\mathrm{R}}$ at the relay for $(\mathcal{P}_1)$ can be obtained using equation (24). Now we turn to the hybrid precoding and combining design problems $(\mathcal{P}_2)$ and $(\mathcal{P}_3)$, which are non-convex due to the constant-modulus constraints (16c) and (19c).

B. Convex Approximations of $(\mathcal{P}_2)$ and $(\mathcal{P}_3)$

Up to now, two of the three challenging facts of the problem $(\mathcal{P})$, sixth-order polynomials and block-diagonal constraints, have been eliminated by problem reformulation in the Section III. In this subsection, we focus on the approximation of constant-modulus constraints (16c) and (19c) for $(\mathcal{P}_2)$ and $(\mathcal{P}_3)$. Inspired by the successive closed forms (SCF) algorithm [30], we approximate the constraints (16c) and (19c) as follows.

Consider the sequences of constraints:

$$\mathrm{Re}(\mathbf{B}^{\mathrm{R}_1(n)}\hat{\mathbf{f}}^{\mathrm{R}_1}) = \mathbf{1}, \tag{26a}$$

$$\mathrm{Re}(\mathbf{B}^{\mathrm{R}_2(n)}\hat{\mathbf{f}}^{\mathrm{R}_2}) = \mathbf{1}, \tag{26b}$$

with

$$\mathbf{B}^{\mathrm{R}_1(n)}(i,j) = \begin{cases} \exp(-j\mathrm{arg}\ (f_\ell^{\mathrm{R}_1})^{(n-1)}), & i = j = \ell, \\ 0, & \text{otherwise}, \end{cases} \tag{27a}$$

$$\mathbf{B}^{\mathrm{R}_2(n)}(i,j) = \begin{cases} \exp(-j\mathrm{arg}\ (f_\ell^{\mathrm{R}_2})^{(n-1)}), & i = j = \ell, \\ 0, & \text{otherwise}, \end{cases} \tag{27b}$$

where $f_\ell^{\mathrm{R}_m}$ is the ℓth element of $\hat{\mathbf{f}}^{\mathrm{R}_m}$ $(m = 1, 2)$.

In particular, the constraints (26) *are adjusted to satisfy the constant-modulus constraints. To illustrate this,*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2018.2799201, IEEE Transactions on Signal Processing

let $\hat{\mathbf{f}}^{\mathrm{R}_1^{(n-1)}}$ *be the solution which satisfies the constraint* $\mathrm{Re}(\mathbf{B}^{\mathrm{R}_1(n-1)}\hat{\mathbf{f}}^{\mathrm{R}_1^{(n-1)}}) = \mathbf{1}$, *then the constant-modulus mapping solution of* $\hat{\mathbf{f}}^{\mathrm{R}_1^{(n-1)}}$ *is given by* $\mathbf{x}^{\mathrm{R}_1^{(n-1)}} = \exp(j\arg(\hat{\mathbf{f}}^{\mathrm{R}_1^{(n-1)}}))$. *If* $\hat{\mathbf{f}}^{\mathrm{R}_1^{(n)}} = \mathbf{x}^{\mathrm{R}_1^{(n-1)}}$, *we have* $\mathbf{B}^{\mathrm{R}_1(n)} = \mathbf{B}^{\mathrm{R}_1(n+1)}$. *Then the constraint* $\mathrm{Re}(\mathbf{B}^{\mathrm{R}_1(n+1)}\hat{\mathbf{f}}^{\mathrm{R}_1^{(n+1)}}) = \mathbf{1}$ *is the same as the constraint* $\mathrm{Re}(\mathbf{B}^{\mathrm{R}_1(n)}\hat{\mathbf{f}}^{\mathrm{R}_1^{(n)}}) = \mathbf{1}$. *In this case, we will see that* $\hat{\mathbf{f}}^{\mathrm{R}_1^{(n+1)}} = \mathbf{x}^{\mathrm{R}^{(n)}}$. *Otherwise, the constraints are updated by the constant-modulus mapping solution of* $\hat{\mathbf{f}}^{\mathrm{R}_1^{(n)}}$ *according to* (27). *As a conclusion, the sequences produced by the adaptive constraints converge to a constant-modulus solution.*

Replacing constant-modulus constraints (16c) *and* (19c) *by* (26a) *and* (26b), *the subproblems* $(\mathcal{P}_2)$ *and* $(\mathcal{P}_3)$ *in the* n*th iteration can be reformulated as* $({\mathcal{Q}_2}^{(n)})$ *and* $({\mathcal{Q}_3}^{(n)})$, *respectively:*

$$
\begin{aligned}
({\mathcal{Q}_2}^{(n)}) \quad & \min_{\hat{\mathbf{f}}^{\mathrm{R}_1}} ||\mathbf{a}^{\mathrm{R}_1} - \mathbf{A}^{\mathrm{R}_1}\hat{\mathbf{f}}^{\mathrm{R}_1}||_2^2, \\
& \text{s.t. } ||\mathbf{C}^{\mathrm{R}_1}\hat{\mathbf{f}}^{\mathrm{R}_1}||_2^2 \le P_{\mathrm{R}}, \\
& \qquad \mathrm{Re}(\mathbf{B}^{\mathrm{R}_1(n)}\hat{\mathbf{f}}^{\mathrm{R}_1}) = \mathbf{1},
\end{aligned}
$$

$$
\begin{aligned}
({\mathcal{Q}_3}^{(n)}) \quad & \min_{\hat{\mathbf{f}}^{\mathrm{R}_2}} ||\mathbf{a}^{\mathrm{R}_2} - \mathbf{A}^{\mathrm{R}_1}\hat{\mathbf{f}}^{\mathrm{R}_2}||_2^2, \\
& \text{s.t. } ||\mathbf{C}^{\mathrm{R}_2}\hat{\mathbf{f}}^{\mathrm{R}_2}||_2^2 \le P_{\mathrm{R}}, \\
& \qquad \mathrm{Re}(\mathbf{B}^{\mathrm{R}_2(n)}\hat{\mathbf{f}}^{\mathrm{R}_2}) = \mathbf{1}.
\end{aligned}
$$

The optimal solutions of complex-valued problems $({\mathcal{Q}_2}^{(n)})$ *and* $({\mathcal{Q}_3}^{(n)})$ *can be obtained by solving their completely equivalent real-valued problems as follows:*

$$
\begin{aligned}
(\tilde{\mathcal{Q}}_2^{(n)}) \quad & \min_{\mathbf{x}_1} ||\hat{\mathbf{a}}_1 - \hat{\mathbf{A}}_1\mathbf{x}_1||_2^2 \\
& \text{s.t. } ||\hat{\mathbf{C}}_1\mathbf{x}_1||_2^2 \le P_{\mathrm{R}}, \\
& \qquad \hat{\mathbf{B}}_1^{(n)}\mathbf{x}_1 = \mathbf{1},
\end{aligned}
$$

where $\hat{\mathbf{A}}_1 = \begin{bmatrix} \mathrm{Re}(\mathbf{A}^{\mathrm{R}_1}) & -\mathrm{Im}(\mathbf{A}^{\mathrm{R}_1}) \\ \mathrm{Im}(\mathbf{A}^{\mathrm{R}_1}) & \mathrm{Re}(\mathbf{A}^{\mathrm{R}_1}) \end{bmatrix}$,
$\hat{\mathbf{b}}_1 = [\mathrm{Re}(\mathbf{a}^{\mathrm{R}_1})^T, \mathrm{Im}(\mathbf{a}^{\mathrm{R}_1}))^T]^T$,
$\hat{\mathbf{C}}_1 = \begin{bmatrix} \mathrm{Re}(\mathbf{C}^{\mathrm{R}_1}) & -\mathrm{Im}(\mathbf{C}^{\mathrm{R}_1}) \\ \mathrm{Im}(\mathbf{C}^{\mathrm{R}_1}) & \mathrm{Re}(\mathbf{C}^{\mathrm{R}_1}) \end{bmatrix}$,
$\mathbf{x}_1 = [\mathrm{Re}(\hat{\mathbf{f}}^{\mathrm{R}_1})^T, \mathrm{Im}(\hat{\mathbf{f}}^{\mathrm{R}_1})^T]^T$,

$$
\hat{\mathbf{B}}_1^{(n)}(i,j) = \begin{cases} \cos(\arg\,(\hat{f}_l^{\mathrm{R}_1})^{(n-1)}), & \text{if } \ i = j = l, \\ \sin(\arg\,(\hat{f}_l^{\mathrm{R}_1})^{(n-1)}), & \text{if } \ i = l, \text{and } j = l + M^{\mathrm{R}}, \\ 0, & \text{otherwise.} \end{cases}
$$

Similarly, we can get the problem $({\tilde{\mathcal{Q}}_3}^{(n)})$ *with variable* $\mathbf{x}_2 = [\mathrm{Re}(\hat{\mathbf{f}}^{\mathrm{R}_2})^T, \mathrm{Im}(\hat{\mathbf{f}}^{\mathrm{R}_2})^T]^T$, *which is the equivalent real-valued problem of* $({\mathcal{Q}_3}^{(n)})$ *and with same structure of* $({\tilde{\mathcal{Q}}_2}^{(n)})$.

Due to the fact that the real-valued problems $({\tilde{\mathcal{Q}}_2}^{(n)})$ *and* $({\tilde{\mathcal{Q}}_3}^{(n)})$ *are convex problems with quadratic and linear constraints, the global optimal solutions of* $\mathbf{x}_1^{(n)}$ *and* $\mathbf{x}_2^{(n)}$ *can be obtained by the existing interior point algorithms [31]. Further, since the real-valued problems* $({\tilde{\mathcal{Q}}_2}^{(n)})$ $({\tilde{\mathcal{Q}}_3}^{(n)})$ *and the complex-valued problems* $({\mathcal{Q}_2}^{(n)})$ $({\mathcal{Q}_3}^{(n)})$ *are completely equivalent, the global optimal* $\hat{\mathbf{f}}^{{\mathrm{R}_1}^{(n)}}$ *and* $\hat{\mathbf{f}}^{{\mathrm{R}_2}^{(n)}}$ *can be obtained by the equations* $\mathbf{x}_1 = [\mathrm{Re}(\hat{\mathbf{f}}^{{\mathrm{R}_1}^{(n)}})^T, \mathrm{Im}(\hat{\mathbf{f}}^{{\mathrm{R}_1}^{(n)}})^T]^T$ *and* $\mathbf{x}_2^{(n)} = [\mathrm{Re}(\hat{\mathbf{f}}^{{\mathrm{R}_2}^{(n)}})^T, \mathrm{Im}(\hat{\mathbf{f}}^{{\mathrm{R}_2}^{(n)}})^T]^T$. *Then, the variables* $\hat{\mathbf{f}}^{{\mathrm{R}_1}^{(n+1)}}$ *and* $\hat{\mathbf{f}}^{{\mathrm{R}_2}^{(n+1)}}$ *in the* $(n+1)$*th iteration are updated by solving the updating subproblems* $({\tilde{\mathcal{Q}}_2}^{(n+1)})$ *and* $({\tilde{\mathcal{Q}}_3}^{(n+1)})$.

Algorithm 1 The proposed ISA Algorithm

```
**Initialize**: $\hat{\mathbf{f}}_k^{\mathrm{R}_1^{(0)}} = \mathbf{1}$ and $\hat{\mathbf{f}}_k^{\mathrm{R}_2^{(0)}} = \mathbf{1}$;
**while** $|f^{(n+1)} - f^{(n)}| > \epsilon$ **do**
    set $j = 0$;
    **while** $|f_1^{(n)} - f_2^{(n)}| + |f_2^{(n)} - f_3^{(n)}| + |f_3^{(n)} - f_1^{(n)}| > \epsilon_{\mathrm{in}}$ **do**
        For fixed $\mathbf{F}^{\mathrm{R}_1^{(n-1)}}$ and $\mathbf{F}^{\mathrm{R}_2^{(n-1)}}$, update $\mathbf{W}^{\mathrm{R}^{(n)}}$ using (24), and calculate $f_1^{(n)^j}$;
        For fixed $\mathbf{W}^{\mathrm{R}^{(n)}}$ and $\mathbf{F}^{{\mathrm{R}_2}^{(n-1)}}$ update $\mathbf{F}^{{\mathrm{R}_1}^{(n)}}$ by solving $({\tilde{\mathcal{Q}}_2}^{(n)})$, and the equation $\mathbf{x}_1 = [\mathrm{Re}(\hat{\mathbf{f}}^{\mathrm{R}_1})^T, \mathrm{Im}(\hat{\mathbf{f}}^{\mathrm{R}_1})^T]^T$; and then calculate $f_2^{(n)^j}$;
        For fixed $\mathbf{W}^{\mathrm{R}^{(n)}}$ and $\mathbf{F}^{{\mathrm{R}_1}^{(n)}}$ update $\mathbf{F}^{{\mathrm{R}_2}^{(n)}}$ by solving $(\tilde{\mathcal{Q}_3^{(n)}})$, and the equation $\mathbf{x}_2 = [\mathrm{Re}(\hat{\mathbf{f}}^{\mathrm{R}_2})^T, \mathrm{Im}(\hat{\mathbf{f}}^{\mathrm{R}_2})^T]^T$; and then calculate $f_3^{(n)^j}$;
        Let $j = j + 1$;
    **end while**
    Let $n = n+1$, and update $\mathbf{B}^{\mathrm{R}_2(n)}$ and $\mathbf{B}^{\mathrm{R}_2(n)}$ via (26).
**end while**
**return** $\mathbf{F}^{\mathrm{R}_1}$, $\mathbf{W}^{\mathrm{R}_1}$, and $\mathbf{F}^{\mathrm{R}_2}$.
```

Note that the subproblems $(\mathcal{P}_2)$ and $(\mathcal{P}_3)$ are difficult to be solved by the existing algorithms due to the non-convex constant-modulus constraints. Fortunately, by replacing the constant-modulus constraints with (26), the global optimal solutions of the reformulated subproblems $({\mathcal{Q}_2}^{(n)})$ and $({\mathcal{Q}_3}^{(n)})$ can be obtained. Moreover, since the problems $({\mathcal{Q}_2}^{(n)})/({\mathcal{Q}_3}^{(n)})$ is completely equivalent to $({\tilde{\mathcal{Q}}_2}^{(n)})/({\tilde{\mathcal{Q}}_3}^{(n)})$, the following clarification and analysis are expressed based on the complex-valued problems $({\mathcal{Q}_2}^{(n)})$ and $({\mathcal{Q}_3}^{(n)})$ for better expression.

Although the reformulated problems $({\mathcal{Q}_2}^{(n)})$ and $({\mathcal{Q}_3}^{(n)})$ do not result in constant-modulus solutions, they can produce the non-increasing sequences that converge to the constant-modulus solutions. The convergence is guaranteed by ***Proposition 1*** in Section IV-D later. Further, since the non-zero elements in the analog precoders $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are required to be constant-modulus in the original problem, we initialize the non-zero elements in $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ to be 1. If the original initialization of non-zero elements in $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are not constant-modulus, the optimal solution of $(\mathcal{Q}_1^{(n)})$ in the first iteration will not be seek in the feasible spaces of $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$. In this case, the optimization may tend to some bad local optima.

For better expression, let $(\mathcal{Q}_1^{(n)})$ be equal to $(\mathcal{P}_1)$ in the nth iteration. It is worthy noting that the quadratically constrained quadratic programming (QCQP) problems $\{(\mathcal{Q}_i^{(n)})\}_{i=1,2,3}$ can be regarded as being decomposed from the problem $(\tilde{\mathcal{P}}^{(n)})$, which is given by

$$
\begin{aligned}
(\tilde{\mathcal{P}}^{(n)}) \min_{\mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}} \quad & \mathbb{E}\Big(||\mathbf{s} - \mathbf{y}||_2^2\Big), \\
\text{s.t.} \quad & (7),\ (4\text{a}),\ (4\text{b}),\ (26\text{a}),\ (26\text{b}).
\end{aligned}
$$

Therefore, the procedure of the proposed ISA algorithm is to iteratively solve the problem $(\tilde{\mathcal{P}}^{(n)})$, which is the external iteration. To solve the problem $(\tilde{\mathcal{P}}^{(n)})$, we need to solve the three subproblems $\{(\mathcal{Q}_i^{(n)})\}_{i=1,2,3}$ iteratively, which is the internal iteration. Since the global optimal solutions to the subproblems $\{(\mathcal{Q}_i^{(n)})\}_{i=1,2,3}$ can be obtained in each external and internal iteration, an optimal solution to $(\tilde{\mathcal{P}}^{(n)})$ can be guaranteed.

To guarantee the convergence of the proposed ISA algorithm, we first solve the subproblem $(\mathcal{Q}_1^{(n)})$, and then we optimize the other two subproblems $(\mathcal{Q}_2^{(n)})$ and $(\mathcal{Q}_3^{(n)})$. This is because the problem $(\mathcal{Q}_1^{(n+1)})$ is updated by the results of $(\mathcal{Q}_2^{(n)})$ and $(\mathcal{Q}_3^{(n)})$, and $(\mathcal{Q}_1^{(n+1)})$ is uncorrelated to the adaptively changing constraints (26). Therefore, the non-increasing of the sequences produced by the subproblems $\{(\mathcal{Q}_i^{(n)}), i=1,2,3\}$ can be guaranteed by optimizing the subproblem $(\mathcal{Q}_1^{(n)})$ firstly.

Moreover, according to convergence analysis in Section IV-D, the stop criterion of the external iteration for the ISA algorithm is set as $|f^{(n+1)}-f^{(n)}|\leq\epsilon$, where $f^{(n)}$ is the objective value of $(\mathcal{P})$ in the nth external iteration, and ϵ is a small factor. Denote $\{f_1^{(n)^j}\}_{i=1,2,3}$ as the objective values of subproblems $\{(\mathcal{Q}_i^{(n)}), i=1,2,3\}$ in the jth internal iteration. The stop criterion of the internal iteration of the ISA algorithm is set as $|f_1^{(n)}-f_2^{(n)}|+|f_2^{(n)}-f_3^{(n)}|+|f_3^{(n)}-f_1^{(n)}|\leq\epsilon_{\text{in}}$. The overall procedure of the proposed ISA algorithm is summarized as **Algorithm 1**.

The parameters ϵ and ϵ_{in} determine the accuracy of the proposed algorithm. There is a fact that the smaller the parameters are chosen, the more accurate optimal solution will be obtained. However, if the parameters ϵ and ϵ_{in} are setting to be too large, the accuracy of the proposed algorithm can not be guaranteed. On the other hand, if the parameters ϵ and ϵ_{in} are setting to be too small, the iteration number will be very large, leading to high implementation complexity.

C. ISA Application in Full-Connected Structure

As mentioned in Section II, the hybrid precoding design for the full-connected structure is simpler than that for the sub-connected one, since fewer constraints are required for the full-connected structure (i.e., only constant-modulus constraints are required, but block-diagonal constraints are not required any more). However, it still faces the challenges including six-order polynomials objective function and power constraint and the constant-modulus constraints. Fortunately, the ISA algorithm can also be applied to overcome these challenges, which is presented as follows.

When the full-connected structure is adopted at the relay as shown in Fig. 2, the MMSE-based relay hybrid precoding problem can be expressed as

$$\begin{aligned}\min_{\mathbf{F}^{\mathrm{R}_1},\mathbf{W}^{\mathrm{R}},\mathbf{F}^{\mathrm{R}_2}}\quad & ||\tilde{\mathbf{W}}^{\mathrm{R}}\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}-\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}||_{\mathrm{F}}^2,\\ \text{s.t.}\quad & \mathbb{E}(\|\mathbf{x}^{\mathrm{R}}\|_2^2)\leq P_{\mathrm{R}},\\ & |\mathbf{F}^{\mathrm{R}_1}|=\mathbf{1},\\ & |\mathbf{F}^{\mathrm{R}_2}|=\mathbf{1}.\end{aligned}\tag{29}$$

It is worth noting that no block-diagonal constraints exist in (29), due to each RF chain is connected to all antennas in the full-connected structure. Similar to reformulation of problem $(\mathcal{P})$ given in Section III-B, the problem (29) can also be reformulated as three QCQP subproblems with constant-modulus constraints. The three subproblems can be similarly obtained by solving one of $\{\mathbf{F}^{\mathrm{R}_1},\mathbf{W}^{\mathrm{R}},\mathbf{F}^{\mathrm{R}_2}\}$, with the other two fixed.

First, when $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are fixed, the subproblem is as same as $(\mathcal{P}_1)$.

Then, for fixed $\mathbf{F}^{\mathrm{R}_2}$ and $\mathbf{W}^{\mathrm{R}}$, using the properties of Kronecker product, the second subproblem can be formulated as

$$\begin{aligned}\min_{\tilde{\mathbf{f}}^{\mathrm{R}_1}}\quad & ||\mathbf{a}^{\mathrm{R}_1}-\mathbf{A}^{\mathrm{R}_1}\tilde{\mathbf{f}}^{\mathrm{R}_1}||_2^2,\\ \text{s.t.}\quad & ||\mathbf{C}^{\mathrm{R}_1}\tilde{\mathbf{f}}^{\mathrm{R}_1}||_2^2\leq P^{\mathrm{R}},\\ & |\tilde{\mathbf{f}}^{\mathrm{R}_1}|=\mathbf{1},\end{aligned}\tag{30}$$

where $\tilde{\mathbf{f}}^{\mathrm{R}_1}$, $\mathbf{a}^{\mathrm{R}_1}$, $\mathbf{A}^{\mathrm{R}_1}$, and $\mathbf{C}^{\mathrm{R}_1}$ are defined in (14).

Finally, when $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{W}^{\mathrm{R}}$ are fixed, the third subproblem is given by

$$\begin{aligned}\min_{\tilde{\mathbf{f}}^{\mathrm{R}_2}}\quad & ||\mathbf{a}^{\mathrm{R}_2}-\mathbf{A}^{\mathrm{R}_2}\tilde{\mathbf{f}}^{\mathrm{R}_2}||_2^2,\\ \text{s.t.}\quad & ||\mathbf{C}^{\mathrm{R}_2}\tilde{\mathbf{f}}^{\mathrm{R}_2}||_2^2\leq P^{\mathrm{R}},\\ & |\tilde{\mathbf{f}}^{\mathrm{R}_2}|=\mathbf{1},\end{aligned}\tag{31}$$

where $\tilde{\mathbf{f}}^{\mathrm{R}_2}$, $\mathbf{a}^{\mathrm{R}_2}$, $\mathbf{A}^{\mathrm{R}_2}$, and $\mathbf{C}^{\mathrm{R}_2}$ are defined in (18).

Now the full-connected hybrid precoding problem (29) can be reformulated as three QCQP subproblems $(\mathcal{P}_1)$, (30), and (31), where (30) and (31) only have constant-modulus constraints. Recall that the hybrid precoding design problem for the sub-connected structure can be reformulated as three subproblems $(\mathcal{P}_i,\ i=1,2,3)$. We confirm that the three subproblems for the full-connected structure have the same structures as the corresponding three subproblems for the sub-connected structure $(\mathcal{P}_i,\ i=1,2,3)$. Consequently, the three QCQP subproblems for the full-connected structure can also be efficiently solved by the proposed ISA algorithm.

D. Convergence Analysis

The ISA algorithm can be applied to design the hybrid precoders in both the sub-connected and the full-connected structures by iteratively solving the three subproblems. The convergence of the proposed ISA algorithm is guaranteed by the following proposition.

Proposition 1: The sequence of objective values generated by the proposed ISA algorithm monotonically non-increases and eventually converges.

Proof: As definition in Section IV-II, for the subproblems $\{(\mathcal{Q}_i^{(n)})\}_{i=1,2,3}$ in the nth external iteration, $\{f_i^{(n)^j}\}_{i=1,2,3}$ are denoted as their objective values in the jth internal iteration.

In the nth external iteration, as the subproblems $\{(\mathcal{Q}_i^{(n)})\}_{i=1,2,3}$ are obtained from the problem $(\tilde{\mathcal{P}}^{(n)})$ and their global optimal solutions can be obtained, their objective values monotonically non-increase and eventually converge, i.e., $f_1^{{(n)}^1} \geq f_2^{{(n)}^1} \geq f_3^{{(n)}^1} \geq f_1^{{(n)}^2} \geq \cdots \geq f_i^{{(n)}^j} \geq \cdots \geq f_1^{{(n)}^*} = f_2^{{(n)}^*} = f_3^{{(n)}^*} \triangleq f^{{(n)}^*}$.

Then, in the $(n+1)$th external iteration, the subproblem $(\mathcal{Q}_1^{(n+1)})$ is updated by the results of $(\mathcal{Q}_2^{(n)})$ and $(\mathcal{Q}_3^{(n)})$, which are uncorrelated to the adaptively changing constraints (26). Therefore, the objective value of $(\mathcal{Q}_1^{(n+1)})$ satisfies $f_1^{{(n+1)}^1} \leq f^{{(n)}^*}$. As a conclusion, we can see that the objective values of the sequences satisfy $f^{{(1)}^*} \geq f^{{(2)}^*} \geq \cdots \geq f^{{(n)}^*} \geq \cdots \geq f^*$, which means that the objective values are non-increasing. On the other hand, it is known that the objective value of $(\mathcal{P})$ is not smaller than zero. Therefore, it can be concluded that the sequence of objective values converges to f^* when n is sufficiently large. ■

Moreover, it is worthy noting that when the sequence of objective values converges, $\mathbf{f}^{\mathrm{R}_1^{(n)}}$ and $\mathbf{f}^{\mathrm{R}_2^{(n)}}$ converge to constant-modulus solutions. As mentioned in Section IV-B, the sequence produced by the adaptively changing constraints converges to a constant-modulus constraint. Introducing this result into problems $(\mathcal{P})$, we can see that the sequence $\mathbf{f}^{\mathrm{R}_1^{(n)}}$ and $\mathbf{f}^{\mathrm{R}_2^{(n)}}$ satisfy the constant-modulus constraints.

E. Optimality Analysis

Based on the convergence analysis in the previous subsection, we now present the optimality analysis of the proposed ISA algorithm, which is given by the following proposition.

Proposition 2: As the increasing of the iteration number n, the proposed ISA algorithm converges to a KKT point of $(\mathcal{P})$.

Proof: We here assume that the sequence of variables $\{\mathbf{F}^{\mathrm{R}_1^{(n)}}, \mathbf{W}^{\mathrm{R}^{(n)}}, \mathbf{F}^{\mathrm{R}_2^{(n)}}\}$ obtained from the proposed the ISA algorithm converges to $\{\mathbf{F}^{\mathrm{R}_1^*}, \mathbf{W}^{\mathrm{R}^*}, \mathbf{F}^{\mathrm{R}_2^*}\}$.

As the block-diagonal constraints (4a) and (5b) are naturally satisfied by the reformulation as the three QCQP subproblems, we only consider the convex power constraint (7) and constant-modulus constraints (5a) and (5b). For generic representation to problem $(\mathcal{P})$, we denote $f(\cdot)$ as the objective functions, $g(\cdot)$ as the power constraint.

In the nth iteration, since the subsequences generated by the internal iterations converge, the optimal solution to $(\tilde{\mathcal{P}}^{(n)})$ can be obtained. Due to the equivalent convexity of the subproblems $({\mathcal{Q}_i}^{(n)})$ and the fact that the objective functions and power constraint of $(\tilde{\mathcal{P}}^{(n)})$ and $(\mathcal{P})$ are the same, the KKT conditions of $(\tilde{\mathcal{P}}^{(n)})$ can be expressed as follows:

$$\begin{aligned}
&\nabla_{\mathbf{W}^{\mathrm{R}}} f(\mathbf{W}^{\mathrm{R}^{(n)}}) + \lambda^{(n)} \nabla_{\mathbf{W}^{\mathrm{R}}} g(\mathbf{W}^{\mathrm{R}^{(n)}}) = 0,\\
&\nabla_{\mathbf{F}^{\mathrm{R}_1}} f(\mathbf{F}^{{\mathrm{R}_1}^{(n)}}) + \lambda^{(n)} \nabla_{\mathbf{F}^{\mathrm{R}_1}} g(\mathbf{F}^{{\mathrm{R}_1}^{(n)}}) + \frac{1}{2}\sum_{k=1}^{M^R} \mu_{1k}^{(n)} \mathbf{b}_k^{\mathrm{R}_1(n)} = 0,\\
&\nabla_{\mathbf{F}^{\mathrm{R}_2}} f(\mathbf{F}^{{\mathrm{R}_2}^{(n)}}) + \lambda^{(n)} \nabla_{\mathbf{F}^{\mathrm{R}_2}} g(\mathbf{F}^{\mathrm{R}_2^{(n)}}) + \frac{1}{2}\sum_{k=1}^{M^R} \mu_{2k}^{(n)} \mathbf{b}_k^{\mathrm{R}_2(n)} = 0,\\
&\lambda^{(n)} \geq 0,\ \lambda^{(n)} g(\mathbf{W}^{\mathrm{R}^{(n)}}, \mathbf{F}^{\mathrm{R}_1^{(n)}}, \mathbf{F}^{\mathrm{R}_2^{(n)}}) = 0,\\
&\mu_{1k}^{(n)}((\mathrm{Re}(\mathbf{b}_k^{\mathrm{R}_1(n)})^T \mathbf{f}_k^{\mathrm{R}_1^{(n)}}) - 1) = 0,\\
&\mu_{2k}^{(n)}((\mathrm{Re}(\mathbf{b}_k^{\mathrm{R}_2(n)})^T \mathbf{f}_k^{\mathrm{R}_2^{(n)}}) - 1) = 0,
\end{aligned} \tag{32}$$

where $\mathbf{B}^{\mathrm{R}_1(n)} = [\mathbf{b}_1^{\mathrm{R}_1(n)}, \mathbf{b}_2^{\mathrm{R}_1(n)}, \cdots, \mathbf{b}_{M^{\mathrm{R}}}^{\mathrm{R}_1(n)}]^T$.

When n is sufficiently large, the satisfaction of the stop criterion indicates that $\lim_{n\to\infty} ||\mathbf{f}_k^{{\mathrm{R}_1}^{(n-1)}} - \mathbf{f}_k^{{\mathrm{R}_1}^{(n)}}||_2^2 = 0$. Introducing this equation into (26) and (27a), we will have $\lim_{n\to\infty} \bar{f}_{\ell,k}^{{\mathrm{R}_1}^{(n-1)}} f_{\ell,k}^{{\mathrm{R}_1}^{(n)}} = 1$, where $\bar{f}$ is the conjugation of f. In other words, when $\mathbf{f}_k^{\mathrm{R}_1^{(n)}}$ converges to $\mathbf{f}_k^{\mathrm{R}_1^*}$, the constant-modulus constraint $|\mathbf{f}_k^{\mathrm{R}_1^*}| = \mathbf{1}$ holds. On the other hand, when $\mathbf{f}_k^{\mathrm{R}_1^{(n)}} \rightarrow \mathbf{f}_k^{{\mathrm{R}_1}^*}$, $\mathbf{b}_k^{\mathrm{R}_1^{(*)}}$ is obtained. Therefore, the equation (26) is equivalent to the constant-modulus constraint (5a) once the stop criterion is satisfied. Similarly, when $n \rightarrow \infty$, the approximation of linear constraint is equivalent to the constant-modulus constraint for variable $\mathbf{F}^{\mathrm{R}_2}$. Combining the satisfaction of constant-modulus constraints with (32), it is concluded that the KKT conditions of problem $(\mathcal{P})$ hold. ■

The ***Proposition 2*** demonstrates that the solution solved by the ISA algorithm satisfies the KKT conditions. Although the global optimal solution cannot be guaranteed due to inherent characteristics of the non-convex problem, the KKT conditions guarantee the first-order necessary conditions for the solution solved by ISA algorithm to be optimal.

On the other hand, the following ***Lemma 3*** demonstrates the relation between the optimal solutions of the original problem $(\mathcal{P})$ and the three subproblems $(\mathcal{Q}_i^{(n)},\ i = 1, 2, 3)$.

Lemma 3: Let $\{\mathbf{F}^{\mathrm{R}_1^*}, \mathbf{W}^{\mathrm{R}^*}, \mathbf{F}^{\mathrm{R}_2^*}\}$ is a local minimizer of the original problem $(\mathcal{P})$. If $\mathrm{Re}(\mathbf{B}_1^{(n)} \hat{\mathbf{f}}^{\mathrm{R}_1}) = \mathbf{1}$ and $\mathrm{Re}(\mathbf{B}_2^{(n)} \hat{\mathbf{f}}^{\mathrm{R}_2}) = \mathbf{1}$, then $\mathbf{W}^{\mathrm{R}^*}$, $\mathbf{f}^{\mathrm{R}_1^*}$, $\mathbf{f}^{\mathrm{R}_2^*}$ are the global minimizers of $(\mathcal{Q}_1^{(n)})$, $(\mathcal{Q}_2^{(n)})$, and $(\mathcal{Q}_3^{(n)})$, respectively.

Proof: See the detailed proof in Appendix C. ■

The ***Lemma 3*** implies that the global optimal solutions of $(\mathcal{Q}_i^{(n)},\ i = 1, 2, 3)$ are the necessary conditions for the optimal solution of the original problem $(\mathcal{P})$.

According to ***Proposition 2*** and ***Lemma 3***, we can see that the proposed ISA algorithm guarantees the theoretically necessary conditions for the optimal solution to the original problem $(\mathcal{P})$.

F. Complexity Analysis of the Proposed ISA Algorithm

Since the dimension of the analog precoder and combiner M^{R} is much larger than that of the digital precoder K, the complexity of the proposed algorithm is dominated by the analog precoder and combiner [4] .

For the sub-connected structure, the quadratic subproblems $(\tilde{\mathcal{Q}}_2)$ and $(\tilde{\mathcal{Q}}_3)$ have $2M^{\mathrm{R}}$ real variables, and the number of iterations is upper bounded by $\mathcal{O}(\sqrt{2M^{\mathrm{R}}}\mathrm{log}(1/\epsilon_1))$, where ϵ_1 is the accuracy parameter [29]. Therefore, the overall complexity of the ISA algorithm for each external iteration is $\mathcal{O}(2(2M^{\mathrm{R}})^{2.5}\mathrm{log}(1/\epsilon_1)))$.

It can be observed that by solving $(\tilde{\mathcal{Q}_2^{(n)}})$ and $(\tilde{\mathcal{Q}_3^{(n)}})$ instead of solving (13) and (17), the dimension of variables can be reduced from $\mathbb{C}^{M_{\mathrm{R}} \times K}$ to $\mathbb{R}^{2M_{\mathrm{R}} \times 1}$.

[4] To further reduce the computational complexity of the hybrid precoding algorithm, the mmWave channel characteristics can be exploited, e.g., the analog precoding can be designed by exploiting the mmWave channels characteristics, and then the digital precoding design is as same as (24). The detailed mathematical analysis and explanation is omitted due to space constraints.

For the full-connected structure, the quadratic subproblems of analog precoder and combiner have $2KM^{\mathrm{R}}$ real variables, and the number of iterations is upper bounded by $\mathcal{O}(\sqrt{2KM^{\mathrm{R}}}\log(1/\epsilon_2))$, where ϵ_2 is the accuracy parameter. Therefore, the overall complexity of the ISA algorithm for each external iteration is $\mathcal{O}(2(2KM^{\mathrm{R}})^{2.5}\log(1/\epsilon_2)))$.

It can be seen that the complexity of the ISA algorithm for sub-connected structure is much less than that for the full-connected one, which indicates that the ISA algorithm is more applicable for the sub-connected structure.

V. Numerical Simulation

In this section, we present numerical results of the proposed relay hybrid precoding design in mmWave massive MIMO systems. For simplicity, we assume that the number of antennas connected to each RF chain is the same, i.e., $M_k^{\mathrm{R}} = M^{\mathrm{R}}/K, (k = 1, 2, \cdots, K)$. The number of antennas at the source and destination are set as $M^{\mathrm{S}} = 64$ and $M^{\mathrm{D}} = 32$. The noise variance is $\sigma_r^2 = 1$, and the SNR is defined as $\mathrm{SNR} \triangleq \frac{\mathrm{P_R}}{\sigma_\mathrm{r}^2}$. The factor ϵ and ϵ_{in} in Algorithm 1 is setting as $\epsilon = 10^{-4}$ and $\epsilon_{\mathrm{in}} = 10^{-4}$ [31]. We adopt the mmWave channel model in (1a). The propagation loss α_ℓ and γ_ℓ in the channels $\mathbf{H}$ and $\mathbf{G}$ obeys the Rician distribution, where the factor κ is set to be 13.2 according to the practical measurements [23].

It is worth noting that we focus on the relay hybrid precoding design in the paper. Since the full-digital precoding and combining can achieve the best performance of the mmWave system, the full-digital precoders and combiners are the upper bound of the hybrid ones. Therefore, to illustrate the performance of the proposed relay hybrid precoding design, one of the important criteria is the gap between our proposed hybrid precoding scheme and the joint source-relay-destination full-digital precoding scheme. The reduced performance gap demonstrates the superior performance of our proposed hybrid precoding scheme. However, if the source, the relay, and the destination are adopted by the hybrid precoding architectures, the gap between the full-digital and the hybrid precoders will be brought by the source, the relay, or the destination. In this case, to demonstrate the performance of the proposed hybrid precoding schemes, we assume that the source and the destination employ the full-digital precoding, which are realized by the algorithm in [32].

In our simulations, the sub-connected hybrid precoding using the proposed ISA algorithm is compared with five existing precoding schemes: 1) full-connected hybrid precoding using MP algorithm [22]; 2) full-digital precoding [32], which provides the performance upper bound; 3) full-analog precoding [33], which provides the performance low bound; 4) full-connected hybrid precoding using the algorithm in [12]; 5) sub-connected hybrid precoding using the algorithm in [34]. Further, recalling that the proposed ISA algorithm can also be used for the full-connected structure, which has been discussed in Section IV-C, the performance of full-connected precoding using the ISA algorithm is also provided for comparison. In addition, to demonstrate the impact of the hybrid precoding at the source, the relay and the destination, we compare the above precoding schemes to the scheme with the hybrid precoding at the source, the relay and the destination.

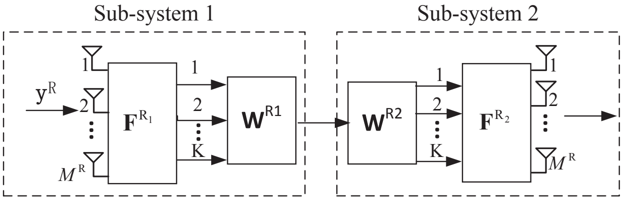
Sub-system 1
Sub-system 2
$\mathbf{y}^{\mathrm{R}}$
1
2
M^{R}
$\mathbf{F}^{\mathrm{R}_1}$
1
2
K
$\mathbf{W}^{\mathrm{R1}}$
$\mathbf{W}^{\mathrm{R2}}$
1
2
K
$\mathbf{F}^{\mathrm{R}_2}$
1
2
M^{R}

Fig. 3. Sub-systems for the relay hybrid precoding system.

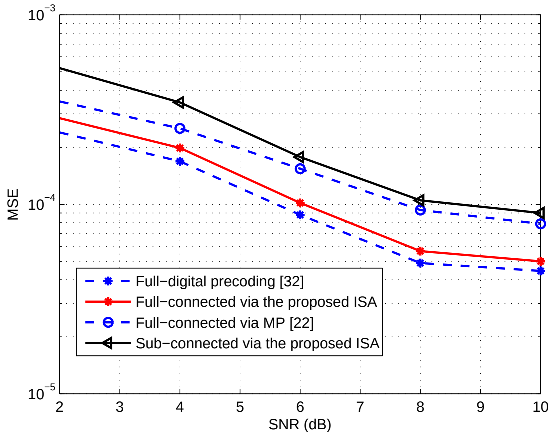

Fig. 4. MSE comparison for different precoding schemes, where $M^{\mathrm{R}} = 48$, $K = 4$, and $L_s = 2$.

It shall be noted that the algorithms in [12] and [34] are proposed for the hybrid precoding design in the single-hop MIMO systems without relays. So it is not easy to directly compare our work with them. To this end, we decompose our relay hybrid precoding system as two MIMO cascade sub-systems as shown in Fig. 3. Accordingly, the digital precoder $\mathbf{W}^{\mathrm{R}}$ for the hybrid precoding system is decomposed into two parts: $\mathbf{W}^{\mathrm{R}_1}$ and $\mathbf{W}^{\mathrm{R}_2}$, with $\mathbf{W}^{\mathrm{R}} = \mathbf{W}^{\mathrm{R}_2}(\mathbf{W}^{\mathrm{R}_1})^H$. Then the relay hybrid precoder can be expressed as $\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}_2}(\mathbf{W}^{\mathrm{R}_1})^H(\mathbf{F}^{\mathrm{R}_1})^H$. Without loss of generality, the relay hybrid precoder can also be regarded as cascade of a receive hybrid precoder $\mathbf{F}^{\mathrm{R}_1}\mathbf{W}^{\mathrm{R}_1}$ and a transmit hybrid precoder $\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}_2}$, both of which can be solved by the hybrid precoding algorithms in [12] and [34].

A. MMSE Performance

Since our objective function is the MSE between the transmitted signal and the received signal at the destination, we compare the MSE performance of different precoding schemes.

Fig. 4 compares the MSE performance of different precoding schemes against SNR. It can be seen clearly that the MSE performance of the full-connected hybrid precoding using the ISA algorithm is close to that of the full-digital precoding. For the full-connected structure, the proposed ISA algorithm outperforms the MP algorithm [22], which demonstrates the effectiveness of our proposed ISA algorithm. For the sub-connected structure, the proposed ISA algorithms achieves the MSE performance close to the full-digital precoding using the MP algorithm, indicating that the proposed ISA algorith-

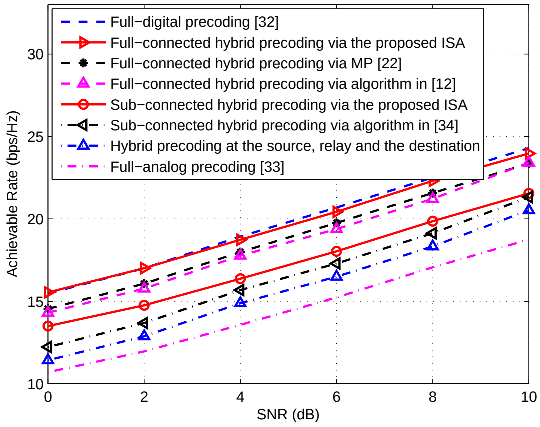

Fig. 5. Achievable rate comparison for different precoding schemes, where $M^{\mathrm{R}} = 48$, $K = 4$, and $L_s = 2$.

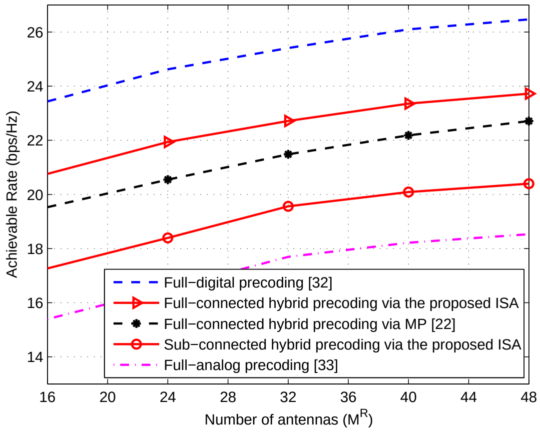

Fig. 6. Achievable rate comparison for different precoding schemes, where $L_s = K = 2$ and SNR = 10 dB.

m achieves a reasonable performance for the sub-connected structure.

B. Achievable Rate Performance

The achievable rate and the MSE are two optimization objectives for the precoding design problems. Both of them have been widely investigated, such as [12], [34] for the maximization of achievable rate problems and [35], [36] for the minimization of MSE problems. To facilitate the solving of the optimization problem, in this paper we choose to minimize the MSE, since it is difficult to solve the maximization of the achievable rate problem directly. In our paper, we first compare the MSE performance. Then, due to the fact that the achievable rate is a crucial criterion in the mmWave systems, we provide simulations in terms of the achievable rate. The achievable rate from the source to the destination is defined as:

$$R = \frac{1}{2}\log_2(1 + \mathrm{SNR}),$$

where

$$\mathrm{SNR} = \frac{||(\tilde{\mathbf{W}}^{\mathrm{D}})^H \mathbf{G}^H \mathbf{F}^{\mathrm{R}_2} \mathbf{W}^{\mathrm{R}} (\mathbf{F}^{\mathrm{R}_1})^H \mathbf{H} \tilde{\mathbf{W}}^{\mathrm{S}}||_{\mathrm{F}}}{\sigma_r^2 ||(\tilde{\mathbf{W}}^{\mathrm{D}})^H \mathbf{G}^H \mathbf{F}^{\mathrm{R}_2} \mathbf{W}^{\mathrm{R}} (\mathbf{F}^{\mathrm{R}_1})^H||_{\mathrm{F}} + \sigma_d^2 ||(\tilde{\mathbf{W}}^{\mathrm{D}})^H||_{\mathrm{F}}}.$$

Fig. 5 compares the achievable rate performance of different precoding schemes against SNR. For the full-connected structure, the proposed ISA algorithm outperforms the MP algorithm, since MP algorithm does not guarantee the optimality [21] [22]. We can find from Fig. 5 that the proposed ISA algorithm also outperforms the algorithm in [12], since the relay hybrid precoding system is decomposed into two MIMO cascade sub-systems for the application of algorithm in [12]. Furthermore, for the sub-connected structures, our proposed ISA algorithm outperforms the precoding algorithm in [34]. It can be seen that our proposed ISA algorithm achieves good performance in both the full-connected and sub-connected structures. Moreover, the achievable rate of the full-connected hybrid precoding using the ISA algorithm is close to that of the full-digital precoding, which demonstrates the effectiveness of the proposed ISA algorithm. For the sub-connected structure, the proposed ISA algorithms also outperforms the full-analog precoding, which indicates that the proposed ISA algorithm achieves a reasonable performance for the sub-connected structure. Furthermore, it can be seen clearly that the precoding scheme where the source and the destination adopt the full-digital precoding and the relay employs the hybrid one, outperforms the joint source-relay-destination hybrid precoding scheme. This indicates that the gap between the full-digital precoding and the hybrid precoding may be affected by the source, the relay and the destination.

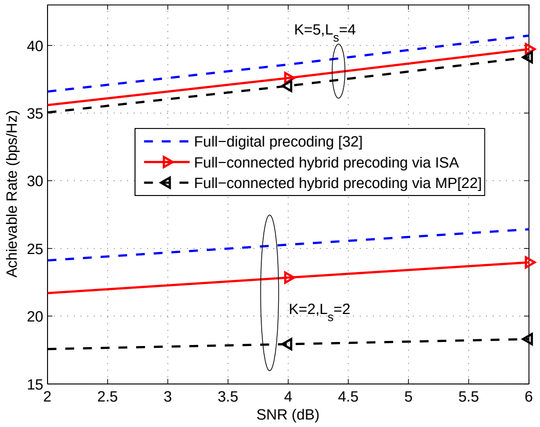

Fig. 7. Achievable rate comparison for full-connected hybrid precoding schemes and the full-digital precoding, where $M^{\mathrm{R}} = 96$.

Fig. 6 compares the achievable rate performance of different precoding algorithms for different number of antennas M^{R} when the number of RF chains K is fixed. As can be seen from this figure, when the number of antennas at the relay increases, the performance of different algorithms improves. This is expected, as the antenna gain increases when the number of antennas increases.

Fig. 7 plots the achievable rates achieved by different precoding schemes when the number of data streams L_s is different. It compares the full-connected precoding schemes

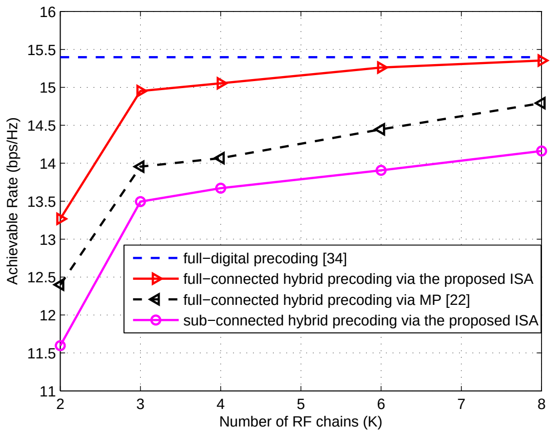

Fig. 8. Achievable rate comparison for different precoding schemes, where $L_s = 2$, $M^{\mathrm{R}} = 48$, and $\mathrm{SNR} = 0\mathrm{dB}$.

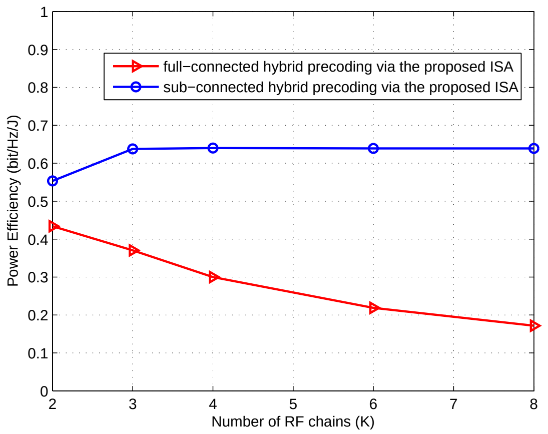

Fig. 9. Energy Efficiency comparison for the sub-connected and the full-connected hybrid precoding schemes using ISA algorithm, where $L_s = 2$, $M^{\mathrm{R}} = 48$, and $\mathrm{SNR} = 0\mathrm{dB}$.

using ISA algorithm and MP algorithm with the full-digital precoding. It turns out that the ISA algorithm always outperforms the MP algorithm under different parameters. We find that when $K = L_s = 2$, the MP algorithm works poorly, while the proposed ISA algorithm always obtains close performance to the full-digital one.

Fig. 8 compares the achievable rate performance of different precoding schemes against the number of RF chains K. It can be clearly found that the performance gap between the full-digital precoding and other hybrid precoding schemes decreases as the increasing number of RF chains. Specially, when the number of RF chains is large enough, the full-connected hybrid precoding using the proposed ISA algorithm almost performs the same as the full-digital one. However, the MP algorithm can not achieve such good performance even with a sufficiently large number of RF chains. Furthermore, the comparison between the two hybrid precoding structures using the same ISA algorithm shows that the sub-connected structure has to pay some unavoidable performance loss compared to the full-connected one. In theory, to realize the full-digital precoding, it is sufficient that the number of RF chains should be greater than or equal to twice of the number of data streams, i.e., $K \geq 2L_s$. When the number of RF chains increases from 1 to 2 (or 2 to 3), the achievable rate performance improves a lot because larger dimensional digital precoding is deployed. However, when the number of RF chains is larger enough, the performance improvement achieved by the digital precoding will saturate. Furthermore, when the number of RF chains is larger enough (but still much less than the number of antennas), the the hybrid precoding achieves very close performance to the full-digital one. In this case, adding further RF chains would increase the power consumption and hardware complexity, with diminishing performance returns.

C. Power Efficiency Performance

As mentioned in Section I, the power consumption is a key issue which should be considered for both the sub-connected and full-connected hybrid precoding structures. In this subsection, we aim at comparing the full-connected and sub-connected structures using the proposed ISA algorithm in terms of power efficiency [37].

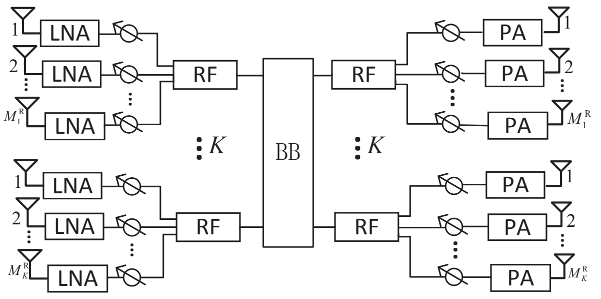

Fig. 10. Power consumption model for the hybrid precoding architecture with the sub-connected structure.

Considering the hybrid precoding architecture at the relay, the approximate power consumption models for the sub-connected and the full-connected structures are shown in figs. 10 and 11, respectively. From these figures, we can see that in the hybrid precoding architecture, the power consumed by five blocks: a) the low noise amplifiers (LNA) on the receiver side; b) the phase shifter on both receiver and transmitter sides; c) the RF chains on both receiver and transmitter sides; d) the base-band (BB) processor; e) the power amplifiers (PA) on the transmitter side.

Based on the the above five blocks, the total power consumption P_{c} in the relay hybrid precoding architecture can be written as

$$P_{\mathrm{c}} = P_{\mathrm{BB}} + 2KP_{\mathrm{RF}} + M^{\mathrm{R}}P_{\mathrm{LNA}} + M^{\mathrm{R}}P_{\mathrm{PA}} + 2N_{\mathrm{PS}}P_{\mathrm{PS}}, \quad (33)$$

where P_{BB} is the power consumed by baseband processor, and N_{PS} is the number of phase shifters. P_{RF}, P_{PA}, P_{LNA}, and P_{PS} are the power of each RF chain, the power of each low noise amplifier, and the power of each power amplifier, the power of each phase shifter, respectively.

Here we shall notice that for the full-connected and the sub-connected structures, the number of phase shifters N_{PS} can be

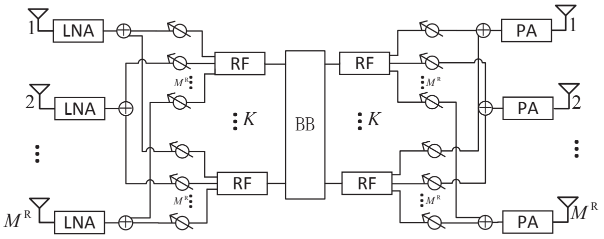

Fig. 11. Power consumption model for the hybrid precoding architecture with the full-connected structure.

expressed as:

$$N_{\text{PS}} = \begin{cases} KM^{\text{R}}, & \text{full-connected}, \\ M^{\text{R}}, & \text{sub-connected}. \end{cases} \tag{34}$$

Introducing (34) into (33), it can be seen that the full-connected structure requires more phase shifters than the sub-connected structure, which indicates that the power consumed by the full-connected structure is larger than that of the sub-connected structure.

To better compare the performance of the two hybrid precoding structures, the power efficiency η is defined as the ratio between the achievable rate R and the total power consumption P_c, which is expressed as follows:

$$\eta = \frac{R}{P_c}, \tag{35}$$

where the unite of η is bps/Hz/J.

In the following, we will simulate the power efficiency performance of the hybrid precoding using the proposed ISA algorithm in the sub-connected and the full-connected structures.

Fig. 9 compares the power efficiency performance of the sub-connected and the full-connected hybrid precoding schemes using the proposed ISA algorithm. The simulation parameters in (33) and (35) are set as follows: $P_{\text{BB}} = 10$W, $P_{\text{RF}} = 100$mW, $P_{\text{PS}} = 10$mW, and $P_{\text{LNA}} = P_{\text{PA}} = 100$mW [38]. It can be see clearly that for the full-connected structure, the power efficiency will increase tremendously as the increasing of the number of RF chains K. While for the sub-connected structure, the power efficiency remains almost stable over different number of RF chains K, due to the fact that the number of phase shifters is independent of the number of RF chains K. On the other hand, as shown in Fig. 8, the achievable rates achieved by the proposed ISA algorithm in both the full-connected and sub-connected structures are improved as the increasing of the number of RF chains K. Based on the above facts, the achievable rate increases with more power consumption for the full-connected structure, while it increases with the unchangeable power consumption for the sub-connected structure.

D. Convergence Performance

Fig. 12 shows the convergence of the proposed ISA algorithm in terms of MSE. We can find that the proposed ISA algorithm always converges within a reasonable number of iterations, regardless the number of antennas and the value of SNR, which verifies the convergence of our proposed scheme. In addition, the MSE decreases monotonically as the number of iteration increases for different the number of antennas and the value of SNR. Moreover, the MSE decreases as the increasing of either the number of antennas or the value of SNR, which is expected, as the diversity increases as the increasing of the number of antennas.

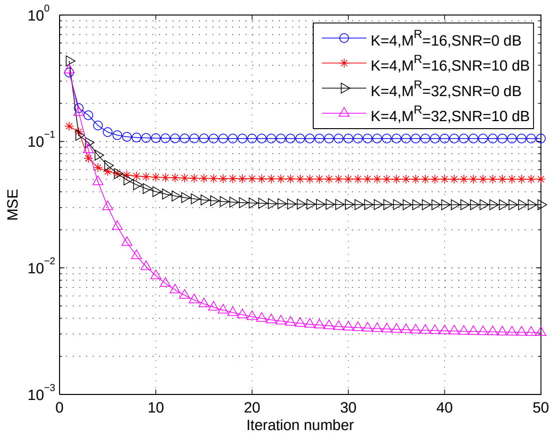

Fig. 12. Convergence of the proposed ISA algorithm.

TABLE I
Comparison of Different Parameters ϵ

External Accuracy Parameter	Objective Value	Iteration Number
$\epsilon = 10^{-1}$	0.0097	6
$\epsilon = 10^{-2}$	0.0056	17
$\epsilon = 10^{-4}$	0.0052	38
$\epsilon = 10^{-6}$	0.0051	61

As discussed above, the parameters ϵ and ϵ_{in} determine the external and internal iteration numbers of the proposed algorithm. To demonstrate the impact of the parameters, we compare the different ϵ with the same channels $\mathbf{H}$ and $\mathbf{G}$. In this case, the original problems remain unchanged when different ϵ is setting. Here, we focus on the comparison among different parameter ϵ in terms of objective value and the external iteration number when the stop criterion is satisfied.

Table I shows the objective value and the iteration number for different ϵ when the stop criterion is satisfied. From this table, we can find that when the external accuracy parameter ϵ is too large (such as 10^{-1}), the optimal solution cannot be guaranteed. On the other hand, when the ϵ is decreasing from 10^{-4} to 10^{-6}, the objective value decreases a little bit, while the iteration number increases almost 2 times (from 38 to 61). As a conclusion, the external accuracy parameter should be setting moderated to guarantee the convergence of the proposed algorithm and the appropriate implementation complexity.

E. Quantized Precoding Schemes

According to the proposed hybrid precoding design and algorithm, each non-zero element of the analog precoders

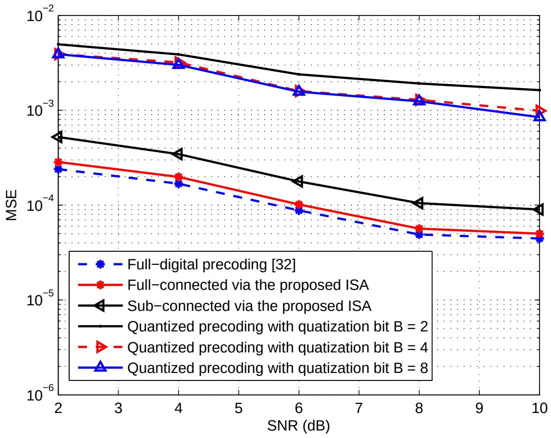

Fig. 13. MSE comparison for full-digital precoding, sub-connected precoding via ISA algorithm, and different quantized precoding schemes, where $M^{\mathrm{R}} = 48$, $K = 4$, and $L_s = 2$.

$\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ has continuous values in phases. However, in practical implementation, the phase of each non-zero element is chosen to be quantized due to practical adoption of phase shifters. Therefore, we need to investigate the performance of our proposed precoding scheme and ISA algorithm in this realistic situations, i.e., phases of the non-zero elements of $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are quantized up to B bits. The phase of each non-zero element of $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ can thus be written as $\hat{\phi} = (2\pi\hat{n})/(2^B)$, where $\hat{n}$ is chosen according to

$$\hat{n} = \arg \min_{n\in\{0,1,\cdots,2^B-1\}} |\phi - \frac{2\pi n}{2^B}|, \tag{36}$$

where ϕ is the unquantized phase of $\mathbf{F}^{\mathrm{R}_1}$ or $\mathbf{F}^{\mathrm{R}_2}$ obtained from the proposed ISA algorithm. Then the digital precoder is computed with the quantized analog precoders.

Fig. 13 compares the MSE performance of the full-digital precoding, the sub-connected precoding via ISA algorithm, and the different quantized precoding schemes. In the quantized precoding scheme, the phases of the non-zero elements of $\mathbf{F}^{\mathrm{R}_1}$ and $\mathbf{F}^{\mathrm{R}_2}$ are quantized up to B bits. It can be see clearly that the MSE decreases as the increasing of the quantization bit B. It shall be noted that the power consumption and cost of the phase shifter devise increase with increasing the quantization bit B. Moreover, it can be seen that the benefits with the quantization bit B saturate at higher values, indicating that moderate-precision phase shifters are appropriate.

VI. Conclusions

In this paper, we have proposed the relay hybrid precoding design for mmWave massive MIMO systems. The hybrid precoding design is to minimize the MSE between the transmitted symbols at the source and the received signals at the destination. To solve this challenging hybrid precoding design problem, an efficient ISA algorithm was proposed for both the sub-connected structure and the full-connected structure. This algorithm attains the high-approximate optimal solution to the original hybrid precoding design problem. It is theoretically proved that the ISA algorithm converges to a KKT point of the original precoding problem. Simulation results demonstrate that the proposed ISA algorithm achieves superior performance in terms of achievable rate in both full-connected and sub-connected structures. In the future, the joint hybrid precoding design at the source, the relay, and the destination can be studies based on the work in this paper.

Appendix A

Hybrid Precoding Designs at the Source and the Destination

In the mmWave systems, we consider that the source, the relay, and the destination deploy the hybrid precoding architectures. Specifically, the source is filtered by the digital precoder $\mathbf{W}^{\mathrm{S}}$ and the analog precoder $\mathbf{F}^{\mathrm{S}}$. Similarly, the destination combining matrices are $\mathbf{F}^{\mathrm{D}}$ and $\mathbf{W}^{\mathrm{D}}$, where $\mathbf{W}^{\mathrm{D}}$ is the digital combiner and the $\mathbf{F}^{\mathrm{D}}$ is the analog combiner. In addition, the relay hybrid precoders are denoted by $\mathbf{F}^{\mathrm{R}_2}$, $\mathbf{W}^{\mathrm{R}}$, and $\mathbf{F}^{\mathrm{R}_1}$ as mentioned before.

Based on the hybrid precoders and combiners, the joint design of the source-relay-destination hybrid beamforming matrices can be formulated as maximizing the achievable rate of the whole system, which can be written as:

$$R = \frac{1}{2}\log_2\Bigg(1+ \frac{||(\mathbf{F}^{\mathrm{D}}\mathbf{W}^{\mathrm{D}})^H\mathbf{G}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{H}\mathbf{F}^{\mathrm{S}}\mathbf{W}^{\mathrm{S}}||_{\mathrm{F}}}{\sigma_r^2||(\mathbf{F}^{\mathrm{D}}\mathbf{W}^{\mathrm{D}})^H\mathbf{G}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H||_{\mathrm{F}} + \sigma_d^2||(\mathbf{F}^{\mathrm{D}}\mathbf{W}^{\mathrm{D}})^H||_{\mathrm{F}}}\Bigg). \tag{37}$$

Our goal is to design the precoders $(\mathbf{F}^{\mathrm{S}}, \mathbf{W}^{\mathrm{S}}, \mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}, \mathbf{F}^{\mathrm{D}}, \mathbf{W}^{\mathrm{D}})$ by maximizing the achievable rate R in (37). However, directly maximizing (37) requires a joint optimization over the matrices $(\mathbf{F}^{\mathrm{S}}, \mathbf{W}^{\mathrm{S}}, \mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}, \mathbf{F}^{\mathrm{D}}, \mathbf{W}^{\mathrm{D}})$, which leads to be intractable global optimal solutions. To this end, we design the source precoding at first, and then seek the relay and destination precoders and combiners utilizing the optimized source precoding results.

Source precoding design: To simplify transceiver design, we temporarily seperate the joint source-relay-destination optimization problem and focus on the design of the source precoder. Therefore, instead of maximizing the achievable rate, the source precoder design problem can be formulated as maximizing the mutual information over the mmWave channel from the source to the relay:

$$\mathcal{I} = \log_2\Big(|\mathbf{I} + \mathbf{H}\mathbf{F}^{\mathrm{S}}\mathbf{W}^{\mathrm{S}}\mathbf{H}^H(\mathbf{F}^{\mathrm{S}}\mathbf{W}^{\mathrm{S}})^H|\Big). \tag{38}$$

Here, we maximize the mutual information in (38) instead of the achievable rate in (37), and we assume that the relay and destination can perform near optimal decoding based on the received signal $\mathbf{y}$.

Relay and destination precoding designs: By using a linear MMSE receiver, the joint relay and destination precoding can be realized. The joint relay and destination precoding problem can be solved by alternating optimizing relay/destination precoding with the other one being fixed.

When the destination precoding is fixed, the relay precoding design can be formulated as problem (10).

When the relay precoding is fixed, the destination precoding design can be formulated as

$$\min_{\tilde{\mathbf{W}}^{\mathrm{D}}} \mathbb{E}\Big(||\mathbf{s}-\mathbf{y}||_2^2\Big). \tag{39}$$

It shall be noted that the destination combining design (39) is the same as the receiver precoding design in conventional MIMO system. To determine the precoders $(\mathbf{W}^{\mathrm{S}}, \mathbf{F}^{\mathrm{S}}, \mathbf{F}^{\mathrm{D}}, \mathbf{W}^{\mathrm{D}})$, the conventional MP algorithm can be applied to obtain the sub-optimal solutions [14].

Appendix B
Proof of Lemma 1

The objective function of (10) can be rewritten as

$$\begin{aligned}
&\mathbb{E}\Big(||\mathbf{s}-(\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{y}^{\mathrm{R}}+\bar{\mathbf{n}}^{\mathrm{D}})||_2^2\Big)\\
&=\mathrm{Tr}(\mathbb{E}[\mathbf{s}\mathbf{s}^H]-2\mathrm{Re}(\mathbb{E}[\mathbf{s}(\mathbf{y}^{\mathrm{R}})^H](\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H)^H\\
&+\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H]\mathbf{F}^{\mathrm{R}_1}(\mathbf{F}^{\mathrm{R}_2})^H(\mathbf{W}^{\mathrm{R}})^H\bar{\mathbf{G}}\\
&+\hat{\sigma}_d^2\mathbf{I}),
\end{aligned} \tag{40}$$

where $\hat{\sigma}_d^2=\mathrm{Tr}(\tilde{\mathbf{W}}^{\mathrm{D}}(\tilde{\mathbf{W}}^{\mathrm{D}})^H)\sigma_d^2$.

By introducing a constant matrix $\tilde{\mathbf{W}}^{\mathrm{R}} \triangleq \mathbb{E}[\mathbf{s}(\mathbf{y}^{\mathrm{R}})^H]$ $\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H]^{-1}$ in (11), the second term of (40) can be reexpressed as

$$\begin{aligned}
&\mathbb{E}[\mathbf{s}(\mathbf{y}^{\mathrm{R}})^H](\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H)^H\\
&=\tilde{\mathbf{W}}^{\mathrm{R}}\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H]\mathbf{F}^{\mathrm{R}_1}(\mathbf{W}^{\mathrm{R}})^H(\mathbf{F}^{\mathrm{R}_2})^H\bar{\mathbf{G}}.
\end{aligned} \tag{41}$$

Since $\mathrm{Tr}\big(\tilde{\mathbf{W}}^{\mathrm{R}}\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H](\tilde{\mathbf{W}}^{\mathrm{R}_1})^H$ is a constant value, we can formulate the problem (40) using the equation (41) as follows

$$\begin{aligned}
&\mathrm{Tr}\big(\tilde{\mathbf{W}}^{\mathrm{R}}\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H](\tilde{\mathbf{W}}^{\mathrm{R}_1})^H\\
&-2\mathrm{Re}(\tilde{\mathbf{W}}^{\mathrm{R}}\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H]\mathbf{F}^{\mathrm{R}_1}(\mathbf{F}^{\mathrm{R}_2})^H(\mathbf{W}^{\mathrm{R}})^H\bar{\mathbf{G}})\\
&+\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H]\mathbf{F}^{\mathrm{R}_1}(\mathbf{F}^{\mathrm{R}_2})^H(\mathbf{W}^{\mathrm{R}})^H\bar{\mathbf{G}}\\
&+\mathbb{E}[\mathbf{s}\mathbf{s}^H]+\sigma_d^2\mathbf{I}-\tilde{\mathbf{W}}^{\mathrm{R}}\mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H](\tilde{\mathbf{W}}^{\mathrm{R}})^H\big)\\
&=||\tilde{\mathbf{W}}^{\mathrm{R}}\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}-\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}})^H\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}||_{\mathrm{F}}^2+\mathrm{constant},
\end{aligned} \tag{42}$$

where $\mathbf{R}_{\mathbf{y}^{\mathrm{R}}} \triangleq \mathbb{E}[\mathbf{y}^{\mathrm{R}}(\mathbf{y}^{\mathrm{R}})^H]$ is given in (11).

By subtracting the constant term, minimizing the problem (42) is equivalent to solving the following minimization problem

$$\min_{\mathbf{F}^{\mathrm{R}_1},\mathbf{W}^{\mathrm{R}},\mathbf{F}^{\mathrm{R}_2}} \quad ||\tilde{\mathbf{W}}^{\mathrm{R}}\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}-\bar{\mathbf{G}}^H\mathbf{F}^{\mathrm{R}_2}\mathbf{W}^{\mathrm{R}}(\mathbf{F}^{\mathrm{R}_1})^H\mathbf{R}_{\mathbf{y}^{\mathrm{R}}}^{\frac{1}{2}}||_{\mathrm{F}}^2. \tag{43}$$

Appendix C
Proof of Lemma 3

Since $\{\mathbf{F}^{\mathrm{R}_1^*}, \mathbf{W}^{\mathrm{R}^*}, \mathbf{F}^{\mathrm{R}_2^*}\}$ is a local minimizer of the original problem $(\mathcal{P})$, any matrices $\{\mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}\}$ who is in the neighborhood of $\{\mathbf{F}^{\mathrm{R}_1^*}, \mathbf{W}^{\mathrm{R}^*}, \mathbf{F}^{\mathrm{R}_2^*}\}$ satisfy $f(\mathbf{F}^{\mathrm{R}_1}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2}) \geq f(\mathbf{F}^{\mathrm{R}_1^*}, \mathbf{W}^{\mathrm{R}^*}, \mathbf{F}^{\mathrm{R}_2^*})$.

Then for fixed $\mathbf{F}^{\mathrm{R}_1^*}$ and $\mathbf{F}^{\mathrm{R}_2^*}$, we have $f(\mathbf{F}^{\mathrm{R}_1^*}, \mathbf{W}^{\mathrm{R}}, \mathbf{F}^{\mathrm{R}_2^*}) \geq f(\mathbf{F}^{\mathrm{R}_1^*}, \mathbf{W}^{\mathrm{R}^*}, \mathbf{F}^{\mathrm{R}_2^*}))$, which means $\mathbf{W}^{\mathrm{R}^*}$ is a local minimum for $(\mathcal{Q}_1^{(n)})$. Since $(\mathcal{Q}_1^{(n)})$ is a convex problem, then $\mathbf{W}^{\mathrm{R}^*}$ is the global minimizer. Similarly, $\mathbf{f}^{\mathrm{R}_1^*}$ and $\mathbf{f}^{\mathrm{R}_2^*}$ are the global minimizers of $(\mathcal{Q}_2^{(n)})$ and $(\mathcal{Q}_3^{(n)})$, respectively.

References

[1] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Bjornson, K. Yang, C. L. I, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

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

[3] R. C. Daniels and R. W. Health, "60 GHz wireless communications: Emerging requirements and design recommendations," *IEEE Veh. Technol. Mag.*, vol. 2, no. 3, pp. 41–50, Sep. 2007.

[4] S. Mumtaz, J. Rodriquez, and L. Dai, "MmWave Massive MIMO: A Paradigm for 5G", *Academic Press, Elsevier,* 2016.

[5] C. X. Wang, F. Haider, X. Gao, X. H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Comm. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[6] J. Garcia-Rois, F. Gomez-Cuba, M. Akdeniz, F. J. Gonzalez-Castano, J. C. Burguillo-Rial, S. Rangan, and B. Lorenzo, "On the analysis of acheduling in dynamic duplex multihop mmWave cellular systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6028–6042, Nov. 2015.

[7] H. Q. Ngo, H. A. Suraweera, M. Matthaiou, and E. G. Larsson, "Multipair full-duplex relaying with massive arrays and linear processing," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1721–1737, Sept. 2014.

[8] H. A. Suraweera, H. Q. Ngo, T. Q. Duong, C. Yuen, and E. G. Larsson, "Multi-pair amplify-and-forward relaying with very large antenna arrays," in *Proc. IEEE International Conference on Communications (ICC)*, Budapest, Hungary, Jun. 2013, pp. 4635–4640.

[9] A. H. Phan, H. D. Tuan, H. H. Kha, and H. H. Nguyen, "Beamforming optimization in multi-user amplify-and-forward wireless relay networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1510–1520, Apr. 2012.

[10] J. Joung and H. S. Ali, "Multiuser two-way amplify-and-forward relay processing and power control methods for beamforming systems," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1833–1846, Mar. 2010.

[11] W. Roh, J. Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Choi, and K. Cheun, "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.

[12] 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.

[13] X. Gao, L. Dai, S. Han, C.-L. I, and X. Wang, "Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 16, no. 19, pp. 6010–6021, Sep. 2017.

[14] 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.

[15] B. Wang, L. Dai, Z. Waqng, N. Ge, and S. Zhou, "Spectrum and energy efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[16] 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.

[17] Y. Y. Lee, C. H. Wang, and Y. H. Huang, "A hybrid RF/Baseband precoding processor based on parallel-index-selection matrix-inversionbypass simultaneous orthogonal matching pursuit for millimeter wave MIMO systems," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 305–317, Jan. 2015.

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

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

[20] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-complexity technologies for 5G millimeter-wave MIMO systems with large antenna arrays," to appear in IEEE Communications Magazine..

[21] J. Lee and Y. H. Lee, "AF relaying for millimeter wave communication systems with hybrid RF/baseband MIMO processing," in *Proc. IEEE International Conference on Communications (ICC)*, Sydney, Australia, Jun. 2014, pp. 5838–5842.

[22] X. Xue, T. E. Bogale, X. Wang, Y. Wang, and L. B. Le, "Hybrid analog-digital beamforming for multiuser MIMO millimeter wave relay systems," in *Proc. IEEE/CIC International Conference on Communications in China (ICCC)*, Shenzhen, China, Nov. 2015, pp. 1-7.

[23] 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.

[24] A. Alkhateeb, and R. W. Heath,, "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1801–1818, May 2016.

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

[26] J. He, T. Kim, H. Ghauch, K. Liu, and G. Wang, "Millimeter wave MIMO channel tracking systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, pp. 416–421, 2014.

[27] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[28] D. Love and R. W. Heath, Jr., "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 2967–2976, Aug. 2005.

[29] J. Nocedal, S. Wright, "Numerical Optimization," Berlin: Springer-Verlag Press, pp. 224–227.

[30] A. Omar, V. Monga ,and M. Rangaswamy. "Tractable MIMO beampattern design under constant modulus waveform constraint," in *Proc.2016 IEEE Radar Conference (RadarConf)*, Philadelphia, USA, May 2016, pp. 1–6.

[31] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optim. Meth. Softw.*, vol. 11–12, pp. 625–653, 1999.

[32] K. Muhammad and Y. Rong, "Joint transceiver optimization for multiuser MIMO relay communication systems," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5977–5986, Nov. 2012.

[33] D. J. Love and R. W. Heath, "Equal gain transmission in multiple-input multiple-output wireless systems," *IEEE Trans. Commun.*, vol. 51, no. 7, pp. 1102–1110, Jul. 2003.

[34] 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.

[35] M. Fadel, A. El-Keyi, and A. Sultan, "QOS-constrained multiuser peer-to-peer amplify-and-forward relay beamforming", *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1397–1408, Mar. 2012.

[36] C. Li, X. Wang, L. Yang, and W. Zhu, "A joint source and relay power allocation scheme for a class of MIMO relay systems", *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4852–4860, Dec. 2009.

[37] R. Mendez-Rial, C. Rusu, N. Gonzalez-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, Jan. 2016.

[38] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock, "Millimeter wave wireless communications," Prentice Hall, 2015.

Xuan Xue received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China in 2010 and 2017, respectively. From 2013 to 2015, she was a visiting Ph.D. student at Western University, Canada. Her research interests include massive MIMO, millimeter-wave systems, and cooperative communications.

Yongchao Wang received the B.E. degree in communication engineering, M.E. and Ph.D. degrees in information and communication engineering from Xidian University, Xian, China, in 1998, 2004, and 2006, respectively. From Sept. 2008 to Jan. 2010, he was a Visiting Scholar in ECE department of University of Minnesota, USA. He is currently a full processor of ISN key state Lab. in Xidian University. His current research interests lie in the areas of signal processing for communications, Massive MIMO, mathematical programming methods and their applications.

Linglong Dai (M11-SM14) received the B.S. degree from Zhejiang University in 2003, the M.S. degree (with the highest honor) from the China Academy of Telecommunications Technology in 2006, and the Ph.D. degree (with the highest honor) from Tsinghua University, Beijing, China, in 2011. From 2011 to 2013, he was a Post-Doctoral Research Fellow with the Department of Electronic Engineering, Tsinghua University, where he was an Assistant Professor from 2013 to 2016 and has been an Associate Professor since 2016. He co-authored the book mmWave Massive MIMO: A Paradigm for 5G (Academic Press, Elsevier, 2016). He has published over 50 IEEE journal papers and over 40 IEEE conference papers. He also holds 13 granted patents. His current research interests include massive MIMO, millimeter-wave communications, NOMA, sparse signal processing, and machine learning. He has received four conference Best Paper Awards at the IEEE ICC 2013, the IEEE ICC 2014, the IEEE ICC 2017, and the IEEE VTC 2017-Fall. He has also received the Tsinghua University Outstanding Ph.D. Graduate Award in 2011, the Beijing Excellent Doctoral Dissertation Award in 2012, the China National Excellent Doctoral Dissertation Nomination Award in 2013, the URSI Young Scientist Award in 2014, the IEEE Transactions on Broadcasting Best Paper Award in 2015, the Second Prize of Science and Technology Award of China Institute of Communications in 2016, the IEEE Communications Letters Exemplary Editor Award in 2017, the National Natural Science Foundation of China for Outstanding Young Scholars in 2017, and the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2017. He currently serves as an Editor of the IEEE Transactions on Communications, the IEEE Transactions on Vehicular Technology, and the IEEE Communications Letters.

Christos Masouros (SMIEEE, MIET) received the Diploma degree in Electrical and Computer Engineering from the University of Patras, Greece, in 2004, and MSc by research and PhD in Electrical and Electronic Engineering from the University of Manchester, UK in 2006 and 2009 respectively. In 2008 he was a research intern at Philips Research Labs, UK. Between 2009-2010 he was a Research Associate in the University of Manchester and between 2010-2012 a Research Fellow in Queen's University Belfast. He has held a Royal Academy of Engineering Research Fellowship between 2011-2016.

He is currently an Associate Professor in the Communications and Information Systems research group, Dept. Electrical and Electronic Engineering, University College London. His research interests lie in the field of wireless communications and signal processing with particular focus on Green Communications, Large Scale Antenna Systems, Cognitive Radio, interference mitigation techniques for MIMO and multicarrier communications. He was the recipient of the Best Paper Award in the IEEE GlobeCom 2015 conference, and has been recognised as an Exemplary Editor for the IEEE Communications Letters, and as an Exemplary Reviewer for the IEEE Transactions on Communications. He is an Editor for IEEE Transactions on Communications, an Associate Editor for IEEE Communications Letters, and was a Guest Editor for IEEE Journal on Selected Topics in Signal Processing issues "Exploiting Interference towards Energy Efficient and Secure Wireless Communications" and "Hybrid Analog / Digital Signal Processing for Hardware-Efficient Large Scale Antenna Arrays".