IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 4, 2020, accepted November 18, 2020, date of publication November 25, 2020,
date of current version December 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040508

Partially-Connected Hybrid Beamforming for Multi-User Massive MIMO Systems

GUANGDA ZANG[1], LINGNA HU[2], FENG YANG[1], (Member, IEEE), LIANGHUI DING[3], (Member, IEEE), AND HUI LIU[1], (Fellow, IEEE)

[1]Institute of Wireless Communications, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Shanghai Institute of Satellite Engineering, Shanghai 200240, China
[3]Institute of Image Communications and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Feng Yang (yangfeng@sjtu.edu.cn)

This work was supported in part by NSFC China under Grant 61771309 and Grant 61671301, in part by the Shanghai Commission of Science and Technology Funding under Grant SCST 15DZ2270400, and in part by the Shanghai Key Laboratory Funding under Grant STCSM 18DZ1200102.

ABSTRACT Due to the high power consumption and hardware cost of radio frequency (RF) chains, the conventional fully-digital beamforming will be impractical for large-scale antenna systems (LSAS). To address this issue, hybrid beamforming has been proposed to reduce the number of RF chains. However, the fully-connected structure assumed in most hybrid beamforming schemes is still cost-intensive. Recently, the partially-connected structure employing notably fewer phase shifters has received considerable attention in both academia and industry. But the design of partially-connected hybrid beamforming has not been fully understood, especially in multi-user systems. In this article, we directly address the challenging non-convex non-smooth partially-connected hybrid beamforming design problem with individual signal-to-interference-plus-noise ratio (SINR) constraints and unit-modulus constraints in a multi-user massive multiple-input multiple-output (MIMO) system. An iterative alternating algorithm based on a penalty method is proposed to obtain a stationary point, which inevitably has relatively high computational complexity. Thus, two low-complexity algorithms are then proposed by utilizing matrix approximation. Numerical results demonstrate significant performance gains of the proposed algorithms over existing hybrid beamforming algorithms. Moreover, the proposed low-complexity algorithms can achieve near-optimal performance with dramatically reduced computational complexity.

INDEX TERMS Massive MIMO, hybrid beamforming, SINR constraints, penalty method, penalty dual decomposition.

I. INTRODUCTION

Large-scale antenna systems (LSAS), also known as massive multiple-input multiple-output (MIMO), is considered to be a key technology for 5G wireless communication [1]. The idea is to deploy an excessive number of antennas at the base station (BS) and serve multiple users at the same time and frequency band using multi-user precoding, also known as beamforming techniques. Benefits of LSAS have been widely studied to boost spectral and energy efficiencies, reduce interference among users, simplify system design, etc. Besides, LSAS can offer sufficient beamforming gain to support millimeter-wave (mmWave) communications [2], which is another key technology to address the challenge of future bandwidth shortage [3], [4]. However, due to the unaffordable power consumption and hardware cost of RF chains, the conventional fully-digital beamforming will be impractical for LSAS [5]. One promising solution to address these difficulties is hybrid (analog/digital) beamforming with a reduced number of RF chains, which has recently received considerable attention.

Hybrid beamforming schemes have been widely studied in both single-user and multi-user massive MIMO systems. The main differences in previous works are the designs of analog beamformers subject to unit-modulus constraints imposed by phase shifters. Specifically, to deal with such constraints, the analog beamformers were designed based on orthogonal matching pursuit (OMP) [5]–[7], manifold optimization (MO) [8], alternating direction of multipliers method (ADMM) [9], and majorization-minimization (MM) [10] in

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Magno.

single-user massive MIMO systems. Meanwhile, in multi-user massive MIMO systems, the analog beamformers were designed based on channel phase extraction [11], [12], channel decomposition [13], by maximizing the equivalent channel gain [14], [15], and by maximizing the analog beamforming gain [16], [17]. There were also some works dealing with the unit-modulus constraints by doubling the number of RF chains [18] or the number of phase shifters [19], yet may be impractical in the large-scale array regime.

Note that the aforementioned hybrid beamforming schemes are based on the fully-connected structure, in which each RF chain is connected to all antenna elements. The fully-connected structure is still cost-intensive in the large-scale array regime. To further reduce the hardware cost and implementation complexity, the partially-connected structure, in which the whole array is divided into sub-arrays with each sub-array being connected to one of the RF chains, has received considerable interest in both academia and industry. For example, in single-user massive MIMO systems, the works [8], [20], [21] exploited the partially-connected structure and decomposed the optimization problem into a series of subproblems. The authors of [9] used an ADMM method in the design of analog beamformers. Furthermore, the work [10] proposed an alternating minimization algorithm by using an MM-based analog beamformer and obtained a stationary point of the considered problem. In multi-user massive MIMO systems, the work [13] designed the analog beamformers by element extraction of the analog beamformer based on the fully-connected structure. The work [16] decomposed the analog beamforming design problem into a series of subproblems.

On the other hand, it is desirable to consider individual signal-to-interference-plus-noise ratio (SINR) constraints in multi-user systems to capture quality of service requirements for different users. However, there are fewer works on the partially-connected hybrid beamforming design with individual SINR constraints. The resulting problem is a challenging non-convex optimization problem even without the unit-modulus constraints. The work [22] simplified the design of analog beamformers by removing the unit-modulus constraints, which, however, is impractical in practical systems. Furthermore, the work [23] proposed a (floating-point) genetic algorithm (GA) to directly handle the unit-modulus constraints and obtain a feasible solution. But it is still a heuristic algorithm and may not be applicable in the large-scale array regime due to the high computational complexity. To the best of our knowledge, the optimal partially-connected hybrid beamforming design with individual SINR constraints and unit-modulus constraints has not been fully understood.

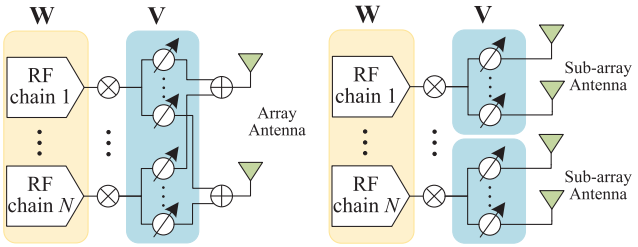
In this article, we study partially-connected hybrid beamforming design to minimize the transmission power under individual SINR constraints and unit-modulus constraints in a multi-user massive MIMO system. The resulting non-convex problem is extremely challenging due to not only the coupling between the analog beamformer and the digital beamformer but also the unit-modulus constraints. The proposed algorithm will directly solve this challenging problem (without any approximation). Our key contributions are summarized below.

- We propose an iterative alternating algorithm based on a penalty method to obtain a stationary point and show that it is so far the most promising method in terms of both transmission power and computational complexity.
- Based on matrix approximation, we then develop low-complexity algorithms in two cases that can achieve near-optimal performance with dramatically reduced computational complexity.
- The proposed algorithms can be applied to the case where the number of RF chains is smaller than the number of users (which is equal to the number of total data symbols), and hence are more general than some existing solutions.
- This is the first work that directly solves the challenging non-convex non-smooth partially-connected hybrid beamforming design problem in multi-user MIMO systems.

Notations: This article uses lower case letters for scalars and boldface lower (upper) case letters for vectors (matrices). $(\mathbf{x})_m$ is the m-th element of $\mathbf{x}$, $(\mathbf{X})_{m,:}$ is the m-th row of $\mathbf{X}$, and $(\mathbf{X})_{m,n}$ is the (m,n)-th element of $\mathbf{X}$. $(\mathbf{X})^T$, $(\mathbf{X})^H$, $(\mathbf{X})^{-1}$, and $(\mathbf{X})^\dagger$ denote the transpose, Hermitian transpose, inverse, and Moore-Penrose pseudoinverse of $\mathbf{X}$, respectively. $\angle[x]$, $|x|$, x^*, $\Re[x]$, and $\Im[x]$ denote the phase, absolute value, conjugate, real part, and imaginary part of x, respectively. $\|\mathbf{x}\|_2$ is the Euclidean norm of $\mathbf{x}$ and $\|\mathbf{X}\|_F$ is the Frobenius norm of $\mathbf{X}$. $\lambda_{\max}(\mathbf{X})$, $\mathrm{rank}(\mathbf{X})$, and $\mathrm{Tr}(\mathbf{X})$ denote the largest eigenvalue, rank, and trace of $\mathbf{X}$, respectively. $\mathrm{diag}(\cdot)$ and $\mathrm{blkdiag}(\cdot)$ denote diagonal and block diagonal operators, respectively. A complex Gaussian random variable x with zero mean and variance σ^2 is denoted by $x \sim \mathcal{CN}(0, \sigma^2)$.

II. SYSTEM MODEL

Consider a narrowband downlink unicast multi-user massive MIMO system, which consists of one BS and K single-antenna users. The BS is equipped with $M (M > K)$ antenna elements and N RF chains. Let $\mathcal{K} \triangleq \{1, \ldots, K\}$ and $\mathcal{N} \triangleq \{1, \ldots, N\}$ denote the set of user indices and the set of RF chain indices, respectively. Due to the unaffordable power consumption and hardware cost, the conventional fully-digital beamforming becomes impractical in the large-scale array regime. Therefore, we consider hybrid beamforming at the BS, which requires a reduced number of RF chains (i.e., $N < M$). Let $\mathbf{W} \triangleq [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K] \in \mathbb{C}^{N\times K}$ and $\mathbf{V} \in \mathbb{C}^{M\times N}$ denote the digital beamformer and the analog beamformer, respectively, where $\mathbf{w}_k \in \mathbb{C}^{N\times 1}$ denotes the digital beamforming vector for user k. To further reduce the implementation complexity and hardware cost, we adopt the partially-connected structure as shown in Fig. 1b, where each RF chain is connected to one dedicated sub-array that consists of a set of antenna elements. Compared with the fully-connected structure as shown in Fig. 1a,



(a) Fully-connected structure. (b) Partially-connected structure.

FIGURE 1. Comparison of two hybrid beamforming structures.

the partially-connected structure is more cost-effective as it requires N times fewer phase shifters. The partially-connected structure introduces N sub-arrays that are independent of each other. Thus, the analog beamformer $\mathbf{V}$ has the following structure:

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{v}_2 & \cdots & \mathbf{0} \\ & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{v}_N \end{bmatrix}, \tag{1}$$

where $\mathbf{v}_n \in \mathbb{C}^{L\times 1}$ is the analog beamforming vector associated with n-th sub-array and $L(L = M/N)$ is the number of antenna elements constituting each sub-array. Let $\mathcal{L} \triangleq \{1, \ldots, L\}$ denote the set of antenna element indices in a sub-array. The zero off-diagonal elements of $\mathbf{V}$ indicate that each RF chain is only connected to its corresponding sub-array. Note that all the nonzero elements of $\mathbf{V}$ are subject to the unit-modulus constraints imposed by phase shifters, i.e., $|(\mathbf{v}_n)_l| = 1, n \in \mathcal{N}, l \in \mathcal{L}$.

In unicasting, the BS has an individual data symbol for each user [24]. Thus, the number of total data symbols is equal to the number of users. Let s_k denote the data symbol of user k and assume that $s_k \sim \mathcal{CN}(0, 1)$, $k \in \mathcal{K}$ are independent of each other. We now illustrate the hybrid beamforming process. 1) Through baseband digital signal processing, each data symbol s_k is first multiplied with its corresponding digital beamforming vector $\mathbf{w}_k$ and then all the weighted digital data symbols are added together. 2) The weighted digital signal $\sum_{k\in\mathcal{K}} \mathbf{w}_k s_k$ is converted to N analog signals by RF chains and sent to each sub-array. 3) The phase shifters are applied to implement analog beamforming by adjusting each analog signal's phase precisely. After the hybrid beamforming process, we obtain the transmitted signal at the BS denoted by $\sum_{k\in\mathcal{K}} \mathbf{V}\mathbf{w}_k s_k$.

We consider a narrowband block fading channel model and assume perfect channel state information at the BS.[1] Let $\mathbf{G} \triangleq [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_K]^H \in \mathbb{C}^{K\times M}$ denote the channel matrix containing the channels of K users, where $\mathbf{g}_k^H$ is the channel vector between the BS and user k. The received signal at user k is given by

$$y_k = \mathbf{g}_k^H \sum_{k\in\mathcal{K}} \mathbf{V}\mathbf{w}_k s_k + n_k = \mathbf{g}_k^H \mathbf{V}\mathbf{w}_k s_k + \sum_{j\in\mathcal{K}, j\neq k} \mathbf{g}_k^H \mathbf{V}\mathbf{w}_j s_j + n_k, \tag{2}$$

[1]It is worth mentioning that the conclusions on the proposed algorithms hold for both correlated and uncorrelated channels.

where $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ denotes the additive Gaussian noise at user k. The instantaneous SINR at user k is given by

$$\mathrm{SINR}_k = \frac{\left|\mathbf{g}_k^H \mathbf{V}\mathbf{w}_k\right|^2}{\sum_{j\in\mathcal{K}, j\neq k} \left|\mathbf{g}_k^{\mathrm{H}} \mathbf{V}\mathbf{w}_j\right|^2 + \sigma_k^2}. \tag{3}$$

We require that SINR_k is above a threshold η_k, which is the prescribed minimum SINR requirement of user k, i.e.,

$$\mathrm{SINR}_k \geq \eta_k. \tag{4}$$

This can be viewed as the quality of service (QoS) constraint for user k.

The goal in this article is to minimize the transmission power under individual SINR constraints and unit-modulus constraints, i.e., we aim to find the optimal analog beamformer $\mathbf{V}$ and digital beamformer $\mathbf{W}$ by solving

$$\begin{aligned} \mathcal{P}_{\mathrm{Ori}}: \min_{\mathbf{V},\mathbf{W}} \quad & \|\mathbf{V}\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \mathrm{SINR}_k \geq \eta_k, k \in \mathcal{K}, & (5) \\ & |(\mathbf{v}_n)_l| = 1, n \in \mathcal{N}, l \in \mathcal{L}. & (6) \end{aligned}$$

Note that even if the unit-modulus constraints in (6) are removed, the remaining hybrid beamforming design problem is still non-convex and has shown to be NP-hard for a fixed $\mathbf{W}$ [25]. Moreover, Problem $\mathcal{P}_{\mathrm{Ori}}$ is even more challenging due to the non-convex non-smooth element-wise unit-modulus constraints in (6). Due to the zero off-diagonal elements of $\mathbf{V}$ as shown in (1), the coupling between $\mathbf{V}$ and $\mathbf{W}$ for the partially-connected structure has changed compared with the one for the fully-connected structure. Thus, most of the previous works on optimal fully-connected beamforming design, such as [19], cannot be extended to address Problem $\mathcal{P}_{\mathrm{Ori}}$. In [23], the authors proposed GA to obtain a feasible point of Problem $\mathcal{P}_{\mathrm{Ori}}$ with high computational complexity. In the following, we shall directly solve Problem $\mathcal{P}_{\mathrm{Ori}}$ and seek low-complexity algorithms.

III. OPTIMAL HYBRID BEAMFORMING DESIGN

In this section, we consider the partially-connected hybrid beamforming design to minimize the transmission power under individual SINR constraints and unit-modulus constraints. We will develop an alternating algorithm based on a penalty method to obtain a stationary point of Problem $\mathcal{P}_{\mathrm{Ori}}$.

A. EQUIVALENT PROBLEM

First, the individual SINR constraints in (5) can be equivalently transformed into [19]:

$$\left\| \begin{bmatrix} (\mathbf{g}_k^H \mathbf{V}\mathbf{W})^H \\ \sigma_k \end{bmatrix} \right\|_2 \leq \sqrt{\frac{1+\eta_k}{\eta_k}} \mathbf{g}_k^H \mathbf{V}\mathbf{w}_k, \ k \in \mathcal{K}, \tag{7}$$

with additional linear constraints:

$$\Re[\mathbf{g}_k^H \mathbf{V}\mathbf{w}_k] \geq 0, \ \Im[\mathbf{g}_k^H \mathbf{V}\mathbf{w}_k] = 0, \ k \in \mathcal{K}. \tag{8}$$

This is because for any feasible point $(\mathbf{V}, \mathbf{W})$ of Problem $\mathcal{P}_{\mathrm{Ori}}$, $\mathbf{g}_k^H \mathbf{V}\mathbf{w}_k, k \in \mathcal{K}$ can be restricted to the nonnegative real domain by multiplying phase scalings $e^{j\phi_k}, k \in \mathcal{K}$ to their

right, where $\phi_k = -\angle[\mathbf{g}_k^H \mathbf{W}\mathbf{w}_k], k \in \mathcal{K}$. And it is evident that the phase scalings $e^{j\phi_k}, k \in \mathcal{K}$ construct a new feasible point, i.e., $\big(\mathbf{V}, \mathbf{W}\cdot\text{diag}(e^{j\phi_1}, e^{j\phi_2}, \ldots, e^{j\phi_K})\big)$, of Problem $\mathcal{P}_{\text{Ori}}$ without changing the feasibility and objective value. Thus, we can rearrange the SINR constraints to obtain (7) and (8).

Then, by observing the structures of $\mathbf{V}$ and $\mathbf{W}$, we construct $\mathbf{U} \in \mathbb{C}^{(L+K)\times N}$ as

$$\mathbf{U} \triangleq \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_N \\ (\mathbf{w}_1)_1^* & (\mathbf{w}_1)_2^* & \cdots & (\mathbf{w}_1)_N^* \\ (\mathbf{w}_2)_1^* & (\mathbf{w}_2)_2^* & \cdots & (\mathbf{w}_2)_N^* \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{w}_K)_1^* & (\mathbf{w}_K)_2^* & \cdots & (\mathbf{w}_K)_N^* \end{bmatrix} \triangleq [\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_N]. \tag{9}$$

The entries of $\mathbf{V}$ and $\mathbf{W}$ can be rewritten as $\mathbf{v}_n = \mathbf{S}\mathbf{u}_n,\ n \in \mathcal{N}$ and $(\mathbf{w}_k)_n = \mathbf{u}_n^H \mathbf{e}_{L+k},\ k \in \mathcal{K},\ n \in \mathcal{N}$, where $\mathbf{S} \triangleq [\mathbf{I}_L, \mathbf{0}_{L\times K}] \in \mathbb{C}^{L\times(L+K)}$ and $\mathbf{e}_{L+k} \in \mathbb{C}^{(L+K)\times 1}$ is a vector with the $(L+k)$-th element being 1 and the rest being 0. Thus, we can rewrite $\mathbf{V}\mathbf{w}_k$ as

$$\mathbf{V}\mathbf{w}_k = \begin{bmatrix} (\mathbf{w}_k)_1 \cdot \mathbf{v}_1 \\ (\mathbf{w}_k)_2 \cdot \mathbf{v}_2 \\ \vdots \\ (\mathbf{w}_k)_N \cdot \mathbf{v}_N \end{bmatrix} = \begin{bmatrix} \mathbf{S}\mathbf{u}_1\mathbf{u}_1^H \mathbf{e}_{L+k} \\ \mathbf{S}\mathbf{u}_2\mathbf{u}_2^H \mathbf{e}_{L+k} \\ \vdots \\ \mathbf{S}\mathbf{u}_N\mathbf{u}_N^H \mathbf{e}_{L+k} \end{bmatrix} = \bar{\mathbf{S}}\bar{\mathbf{\Phi}}\bar{\mathbf{e}}_{L+k}, \tag{10}$$

where $\bar{\mathbf{S}} \triangleq \text{blkdiag}(\mathbf{S}, \mathbf{S}, \ldots, \mathbf{S}) \in \mathbb{C}^{NL\times N(L+K)}$, $\bar{\mathbf{\Phi}} \triangleq \text{blkdiag}(\mathbf{u}_1\mathbf{u}_1^H, \mathbf{u}_2\mathbf{u}_2^H, \ldots, \mathbf{u}_N\mathbf{u}_N^H) \in \mathbb{C}^{N(L+K)\times N(L+K)}$, and $\bar{\mathbf{e}}_{L+k} \triangleq [\mathbf{e}_{L+k}^T, \mathbf{e}_{L+k}^T, \ldots, \mathbf{e}_{L+k}^T]^T \in \mathbb{C}^{N(L+K)\times 1}$. By introducing slack variables $\mathbf{X}_n = \mathbf{u}_n\mathbf{u}_n^H \in \mathbb{C}^{(L+K)\times(L+K)},\ n \in \mathcal{N}$, we have

$$\mathbf{V}\mathbf{w}_k = \bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{e}}_{L+k}, \tag{11}$$

where $\bar{\mathbf{X}} \triangleq \text{blkdiag}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$. Note that $\mathbf{X}_n$ can be rewritten as $\mathbf{X}_n = \mathbf{u}_n\mathbf{u}_n^H$ if and only if $\mathbf{X}_n$ is a rank-one positive semidefinite matrix, i.e.,

$$\mathbf{X}_n \succeq \mathbf{0},\ n \in \mathcal{N}. \tag{12}$$

$$\text{rank}(\mathbf{X}_n) = 1,\ n \in \mathcal{N}. \tag{13}$$

Thus, the objective function of Problem $\mathcal{P}_{\text{Ori}}$ can be recast as

$$\|\mathbf{V}\mathbf{W}\|_F^2 = \|\bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{E}}\|_F^2, \tag{14}$$

where $\bar{\mathbf{E}} \triangleq [\bar{\mathbf{e}}_{L+1}, \bar{\mathbf{e}}_{L+2}, \ldots, \bar{\mathbf{e}}_{L+K}]$. And (7) can be recast as

$$\left\| \begin{bmatrix} (\mathbf{g}_k^H \bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{E}})^H \\ \sigma_k \end{bmatrix} \right\|_2 \le \sqrt{\frac{1+\eta_k}{\eta_k}}\, \mathbf{g}_k^H \bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{e}}_{L+k},\ k \in \mathcal{K}, \tag{15}$$

with additional linear constraints:

$$\Re[\mathbf{g}_k^H \bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{e}}_{L+k}] \ge 0,\ \Im[\mathbf{g}_k^H \bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{e}}_{L+k}] = 0,\ k \in \mathcal{K}. \tag{16}$$

Thanks to the special properties of the rank-one positive semidefinite matrix, we can rewrite the unit-modulus constraints in (6) in an equivalent form as shown in Lemma 1.

Lemma 1: If $\mathbf{X}_n$ satisfies (12), (13), and

$$(\mathbf{X}_n)_{l,l} = 1,\ n \in \mathcal{N},\ l \in \mathcal{L}, \tag{17}$$

then it can be decomposed as $\mathbf{X}_n = \mathbf{u}_n\mathbf{u}_n^H$, where $\mathbf{u}_n = [\mathbf{v}_n^T, (\mathbf{w}_1)_n^*, (\mathbf{w}_2)_n^*, \ldots, (\mathbf{w}_K)_n^*]^T$ and $|(\mathbf{v}_n)_l| = 1,\ l \in \mathcal{L}$.

Proof: It is clear that (12) and (13) define a rank-one positive semidefinite matrix $\mathbf{X}_n$, which can be decomposed as $\mathbf{X}_n = \mathbf{u}_n\mathbf{u}_n^H$. Besides, the l-th diagonal element of $\mathbf{X}_n$ is $|(\mathbf{u}_n)_l|^2$. Thus, $(\mathbf{X}_n)_{l,l} = 1$ means $|(\mathbf{u}_n)_l|^2 = 1$. Since $|(\mathbf{u}_n)_l| \ge 0$, we have $|(\mathbf{u}_n)_l| = 1$, i.e., $|(\mathbf{v}_n)_l| = 1$. ■

Thus, based on Lemma 1, we can directly handle the unit-modulus constraints in (6).

Finally, we obtain the following equivalent formulation of Problem $\mathcal{P}_{\text{Ori}}$:

$$\mathcal{P}_{\text{Eq}}: \min_{\{\mathbf{X}_n\}_{n\in\mathcal{N}}} \ \|\bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{E}}\|_F^2 \quad \text{s.t.}\ (12),\ (13),\ (15),\ (16),\ (17).$$

Problem $\mathcal{P}_{\text{Ori}}$ and Problem $\mathcal{P}_{\text{Eq}}$ are equivalent in the sense that they share the same optimal value. Moreover, by following [25, Proposition 3], it can be verified that if $\{\mathbf{X}_n\}_{n\in\mathcal{N}}$ is a stationary point of Problem $\mathcal{P}_{\text{Eq}}$, $(\mathbf{V}, \mathbf{W})$ is a stationary point of Problem $\mathcal{P}_{\text{Ori}}$. In the following, we focus on solving Problem $\mathcal{P}_{\text{Eq}}$ instead of Problem $\mathcal{P}_{\text{Ori}}$. Note that Problem $\mathcal{P}_{\text{Eq}}$ is convex except for the rank-one constraints in (13), which are non-smooth and hard to tackle. Inspired by [19], [26], we resort to a similar approach for efficiently solving Problem $\mathcal{P}_{\text{Eq}}$.

B. PENALTY METHOD

To address the challenge of the rank-one constraints in (13), based on the non-smooth optimization approach in [26], we rewrite Problem $\mathcal{P}_{\text{Eq}}$ in an equivalent form, which is shown in Theorem 1 with a penalty weight μ.

Theorem 1: Problem $\mathcal{P}_{\text{Eq}}$ is equivalent to

$$\mathcal{P}_{\text{Pen}}: \min_{\{\mathbf{X}_n\}_{n\in\mathcal{N}}} \ \|\bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{E}}\|_F^2 + \frac{1}{\mu}\sum_{n\in\mathcal{N}}\big(\text{Tr}(\mathbf{X}_n) - \lambda_{\max}(\mathbf{X}_n)\big) \quad \text{s.t.}\ (12),\ (15),\ (16),\ (17).$$

Moreover, there exists $\mu_0 \in (0, +\infty)$ such that whenever $\mu < \mu_0$, $\sum_{n\in\mathcal{N}}\big(\text{Tr}(\mathbf{X}_n) - \lambda_{\max}(\mathbf{X}_n)\big) = 0$ and $(\mathbf{V}, \mathbf{W})$ is a stationary point of Problem $\mathcal{P}_{\text{Ori}}$, where $\{\mathbf{X}_n\}_{n\in\mathcal{N}}$ is a stationary point of Problem $\mathcal{P}_{\text{Pen}}$.

Proof: The theorem can be proven by following [26, Theorem 1]. We here omit the proof for brevity. ■

Theorem 1 shows that if the rank-one constraints are satisfied, the penalty term does not contribute to the objective function of Problem $\mathcal{P}_{\text{Pen}}$. However, the objective function of Problem $\mathcal{P}_{\text{Pen}}$ is still non-smooth because $\lambda_{\max}(\cdot)$ is not differentiable. In fact, $\text{Tr}(\mathbf{X}_n) - \lambda_{\max}(\mathbf{X}_n)$ computes the sum of the $(L+K-1)$ smallest eigenvalues of $\mathbf{X}_n$. Thus, Problem $\mathcal{P}_{\text{Pen}}$ can be recast as [27]

$$\mathcal{P}_{\text{Alt}}: \min_{\{\mathbf{X}_n\}_{n\in\mathcal{N}}} \min_{\{\mathbf{P}_n\}_{n\in\mathcal{N}}} \ \|\bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{E}}\|_F^2 + \frac{1}{\mu}\sum_{n\in\mathcal{N}}\text{Tr}(\mathbf{P}_n^T\mathbf{X}_n) \quad \text{s.t.}\ \mathbf{P}_n \in \mathbf{\Psi}_{L+K,1},\ n \in \mathcal{N},\ (12),\ (15),\ (16),\ (17),$$

where $\mathbf{\Psi}_{L+K,1} \triangleq \{\mathbf{P} \in \mathbb{S}^{L+K},\ \mathbf{0} \preceq \mathbf{P} \preceq \mathbf{I},\ \mathrm{Tr}(\mathbf{P}) = L+K-1\}$ denotes a convex hull of the rank-one projection matrices. Problem $\mathcal{P}_{\mathrm{Alt}}$ involves two disjoint blocks of variables, i.e., $\{\mathbf{X}_n\}_{n\in\mathcal{N}}$ and $\{\mathbf{P}_n\}_{n\in\mathcal{N}}$. By leveraging the two-block GS method [28], we can solve Problem $\mathcal{P}_{\mathrm{Alt}}$ through an iterative alternating procedure, which leads to the following two subproblems.

1) THE SUBPROBLEM w.r.t. $\{\mathbf{X}_n\}_{n\in\mathcal{N}}$

For fixed $\{\mathbf{P}_n\}_{n\in\mathcal{N}}$, the subproblem w.r.t. $\{\mathbf{X}_n\}_{n\in\mathcal{N}}$ is accordingly given by

$$\begin{aligned}\min_{\{\mathbf{X}_n\}_{n\in\mathcal{N}}}\ & \|\bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{E}}\|_F^2 + \frac{1}{\mu}\sum_{n\in\mathcal{N}}\mathrm{Tr}(\mathbf{P}_n^T\mathbf{X}_n)\\ \text{s.t. }& (12),\ (15),\ (16),\ (17).\end{aligned} \tag{18}$$

The subproblem in (18) is convex and can be efficiently solved using standard interior-point toolboxes [29].

2) THE SUBPROBLEM w.r.t. $\{\mathbf{P}_n\}_{n\in\mathcal{N}}$

$\{\mathbf{P}_n\}_{n\in\mathcal{N}}$ only appear in the second term of the objective function of Problem $\mathcal{P}_{\mathrm{Alt}}$. Thus, the subproblem w.r.t. $\{\mathbf{P}_n\}_{n\in\mathcal{N}}$ is accordingly given by

$$\begin{aligned}\min_{\{\mathbf{P}_n\}_{n\in\mathcal{N}}}\ & \sum_{n\in\mathcal{N}}\mathrm{Tr}(\mathbf{P}_n^T\mathbf{X}_n)\\ \text{s.t. }& \mathbf{P}_n \in \mathbf{\Psi}_{L+K,1},\ n\in\mathcal{N}.\end{aligned} \tag{19}$$

The subproblem in (19) can be divided into N independent problems, and each has a closed-form solution expressed as

$$\mathbf{P}_n^* = \mathbf{\Delta}_n\mathbf{\Delta}_n^H, \tag{20}$$

where $\mathbf{\Delta}_n \in \mathbb{C}^{(L+K)\times(L+K-1)}$ is composed of the $(L+K-1)$ eigenvectors corresponding to the $(L+K-1)$ smallest eigenvalues of $\mathbf{X}_n$.

The details are summarized in Algorithm 1. Since the objective value of Problem $\mathcal{P}_{\mathrm{Alt}}$ for a given μ is nonnegative and is monotonically non-increasing with i, the iterative alternating procedure (Step 3-Step 7 of Algorithm 1) converges to a limit point. As the constraint sets of the subproblem in (18) and the subproblem in (19) are disjoint, the iterative alternating procedure is actually a 2Block GS method and the obtained limit point is a stationary point of Problem $\mathcal{P}_{\mathrm{Alt}}$ [28, Corollary 2].[2] By the equivalence between Problem $\mathcal{P}_{\mathrm{Alt}}$ and Problem $\mathcal{P}_{\mathrm{Pen}}$ and by Theorem 1, we can conclude that Algorithm 1 converges to a stationary point of Problem $\mathcal{P}_{\mathrm{Ori}}$.

C. IMPLEMENT ISSUES

1) INITIALIZATION

Algorithm 1 involves an outer loop for finding a sufficiently small μ and an inner loop for iterative alternating. Generally, the inner loop requires an initial feasible point as initialization. Here, we provide an efficient method for finding a good initial feasible point. First, we drop the rank-one constraints in (13) to obtain a relaxed version of Problem $\mathcal{P}_{\mathrm{Eq}}$, i.e., a semidefinite relaxation (SDR) problem as follows.

$$\begin{aligned}\min_{\{\mathbf{X}_n\}_{n\in\mathcal{N}}}\ & \|\bar{\mathbf{S}}\bar{\mathbf{X}}\bar{\mathbf{E}}\|_F^2\\ \text{s.t. }& (12),\ (15),\ (16),\ (17).\end{aligned} \tag{21}$$

Consequently, the problem in (21) reduces into a convex problem and can be solved using standard interior-point toolboxes [29]. Let $\{\mathbf{X}_n^*\}_{n\in\mathcal{N}}$ denote the optimal solution of the problem in (21). Then, a practical initial point $\{\mathbf{P}_n^{(0)}\}_{n\in\mathcal{N}}$ can be obtained by solving

$$\begin{aligned}\{\mathbf{P}_n^{(0)}\}_{n\in\mathcal{N}} = \arg\min_{\{\mathbf{P}_n\}_{n\in\mathcal{N}}}\ & \sum_{n\in\mathcal{N}}\mathrm{Tr}(\mathbf{P}_n^T\mathbf{X}_n^*)\\ \text{s.t. }\ & \mathbf{P}_n \in \mathbf{\Psi}_{L+K,1},\ n\in\mathcal{N},\end{aligned}$$

which has a closed-form solution as shown in (20).

Algorithm 1 Optimal Hybrid Beamforming Design

Initialization: Choose proper μ and $\{\mathbf{P}_n^{(0)}\}_{n\in\mathcal{N}}$;
1: **while** $\sum_{n\in\mathcal{N}}\left(\mathrm{Tr}(\mathbf{X}_n) - \lambda_{\max}(\mathbf{X}_n)\right) > 0$ **do**
2: $i := 0$
3: **repeat**
4: Given $\{\mathbf{P}_n^{(i)}\}_{n\in\mathcal{N}}$, update $\{\mathbf{X}_n^{(i+1)}\}_{n\in\mathcal{N}}$ by solving the subproblem in (18);
5: Given $\{\mathbf{X}_n^{(i+1)}\}_{n\in\mathcal{N}}$, update $\{\mathbf{P}_n^{(i+1)}\}_{n\in\mathcal{N}}$ according to (20);
6: $i \leftarrow i+1$;
7: **until** convergence criterion is met;
8: $\mu := c\mu$;
9: **end while**
10: Construct $\mathbf{V}$ and $\mathbf{W}$ using $\{\mathbf{X}_n^{(i)}\}_{n\in\mathcal{N}}$.

2) COMPLEXITY

In each iteration of Algorithm 1, the subproblem in (19) is solved by computing a closed-form solution, and the subproblem in (18) is solved using an interior-point method that requires $\mathcal{O}\left(\sqrt{N(L+K)}\right)$ iterations with each iteration requiring at most $\mathcal{O}\left(N^3(L+K)^6 + KN(L+K)^2\right)$ arithmetic operations. With large system parameters (e.g., K, N, and L), the computational complexity of Algorithm 1 may be relatively high. Nevertheless, it should be noted that the computational complexity of Algorithm 1 is still much lower than that of existing GA. Besides, Algorithm 1 can provide a theoretically guaranteed solution for Problem $\mathcal{P}_{\mathrm{Ori}}$. Numerical results also show remarkable performance improvement over existing GA in terms of both computational complexity and transmission power. Hence, Algorithm 1 can serve as a reference providing benchmark performance.

IV. LOW-COMPLEXITY HYBRID BEAMFORMING DESIGN

Although Algorithm 1 can jointly optimize the analog beamformer and the digital beamformer and provide a stationary

[2] As shown in Step 8 of Algorithm 1, a sufficiently small μ (i.e., $\mu < \mu_0$) can be obtained from an outer loop by setting $\mu := c\mu$ until $\sum_{n\in\mathcal{N}}\left(\mathrm{Tr}(\mathbf{X}_n) - \lambda_{\max}(\mathbf{X}_n)\right) = 0$, where $c \in (0,1)$ is a weight control parameter.

point of the original problem, its computational complexity may be relatively high. Thus, it is necessary to develop algorithms with lower computational complexity and acceptable performance loss. In this section, we will propose low-complexity hybrid beamforming designs based on matrix approximation. The corresponding matrix approximation problem can be formulated as follows:[3]

$$\mathcal{P}_{\text{MAP}}: \min_{\mathbf{V},\mathbf{W}} \ \|\mathbf{VW}-\mathbf{W}_{\text{FD}}\|_F^2 \quad \text{s.t.} \ \ (5),\ (6),$$

where $\mathbf{W}_{\text{FD}} \in \mathbb{C}^{M\times K}$ is the optimal fully-digital beamformer and can be obtained using an efficient method based on a semi-closed form solution [25]. Note that the transmission power achieved by hybrid beamforming is lower bounded by the one achieved by the fully-digital beamforming. Hence, we can obtain good performance if the hybrid beamformers are carefully designed to be sufficiently close to the optimal fully-digital beamformer.

However, dealing with the non-convex individual SINR constraints in (5) will inevitably incur relatively high computational complexity. To derive low-complexity solutions, we approximate the individual SINR constraints based on a ZF method [30]. Specifically, ZF processing is performed to cancel inter-user interference, i.e., $|\mathbf{g}_k^H\mathbf{V}\mathbf{w}_j| = 0$, for $k \neq j$. Therefore, we can approximately solve Problem $\mathcal{P}_{\text{MAP}}$ by addressing the following problem.

$$\mathcal{P}_{\text{MAP-ZF}}: \min_{\mathbf{V},\mathbf{W}} \ \|\mathbf{VW}-\mathbf{W}_{\text{FD}}\|_F^2 \quad \text{s.t.} \ \ \mathbf{GVW}=\mathbf{\Xi},\ (6), \tag{22}$$

where $\mathbf{\Xi} \triangleq \text{diag}(\sqrt{\eta_1\sigma_1^2}, \sqrt{\eta_2\sigma_2^2}, \ldots, \sqrt{\eta_K\sigma_K^2}) \in \mathbb{C}^{K\times K}$ collects the prescribed SINR requirements of K users.

In the following, we focus on solving Problem $\mathcal{P}_{\text{MAP-ZF}}$, which has a simpler form than Problem $\mathcal{P}_{\text{MAP}}$. According to the concept of ZF processing, it requires a restriction on the number of users, i.e., $K \leq N$, to completely cancel inter-user interference. Hence, in the case of $K \leq N$, any $(\mathbf{V}, \mathbf{W})$ satisfies the constraint in (22) also satisfies the individual SINR constraints in (5). However, in the case of $K > N$, we cannot cancel all the inter-user interference, i.e., the constraint in (22) cannot be satisfied by any $\mathbf{W}$. This motivates us to first violate the constraint in (22) by performing some projection and transform Problem $\mathcal{P}_{\text{MAP-ZF}}$ into an analog beamformer design problem. The digital beamformer is then designed to directly satisfy the individual SINR constraints in (5). Thus, we shall consider two cases, i.e., $K \leq N$ and $K > N$.

[3]Hybrid beamforming designs based on matrix approximation (without the individual SINR constraints in (5)) have been widely investigated in single-user massive MIMO systems [5], [6], [8], [10] and shown good performance. In this article, the objection function $\|\mathbf{VW}-\mathbf{W}_{\text{FD}}\|_F^2$ cannot be driven to zero due to the special structure of $\mathbf{V}$. A feasible solution satisfying (5) cannot be obtained using existing methods and we impose (5) in Problem $\mathcal{P}_{\text{MAP}}$ to sustain feasibility in the design of low-complexity algorithms. Consequently, due to the non-convex individual SINR constraints in (5), Problem $\mathcal{P}_{\text{MAP}}$ is even more challenging and cannot be addressed by existing methods.

A. CASE OF $K \leq N$

The main obstacles in Problem $\mathcal{P}_{\text{MAP-ZF}}$ are still the coupling between the matrix variables $\mathbf{V}$ and $\mathbf{W}$, and the non-convex non-smooth unit-modulus constraints in (6). To address these issues and derive a low-complexity algorithm, we adopt a penalty dual decomposition (PDD) method which is a double-loop general framework for solving non-convex non-smooth optimization problems involving non-convex coupling constraints [31]. The main idea of the PDD method is to penalize and dualize the coupling constraints into the objective function in the outer loop and use the block successive upper-bound minimization (BSUM) method to iteratively solve the resulting augmented Lagrangian problem in the inner loop. Besides, due to the special structure of the analog beamformer $\mathbf{V}$, the unit-modulus constraints in (6) can be well exploited by the BSUM method, which facilitates the design of low-complexity algorithms.

Specifically, we first introduce an auxiliary variable $\mathbf{Y}$ such that $\mathbf{Y} = \mathbf{VW}$ and a corresponding dual variable $\mathbf{Z}$. With appropriate penalty, we have the following augmented Lagrangian problem.

$$\mathcal{P}_{\text{Au}}: \min_{\mathbf{V},\mathbf{W},\mathbf{Y}} \ \|\mathbf{VW}-\mathbf{W}_{\text{FD}}\|_F^2 + \frac{1}{2\rho}\|\mathbf{Y}-\mathbf{VW}+\rho\mathbf{Z}\|_F^2 \quad \text{s.t.} \ \ \mathbf{GY}=\mathbf{\Xi},\ (6), \tag{23}$$

where ρ is a penalty weight. In the outer loop, the dual variable $\mathbf{Z}$ and the penalty weight ρ are updated according to certain constraint violation condition, which will be described later. Then, the inner loop is devoted to solving the non-convex Problem $\mathcal{P}_{\text{Au}}$ for fixed $\mathbf{Z}$ and ρ. As we have disjoint constraints on three different blocks of variables (i.e., $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{Y}$), it is desirable to adopt the BSUM method to solve Problem $\mathcal{P}_{\text{Au}}$.

Note that the main difficulty of the PDD method lies in the inner loop for efficiently solving augmented Lagrangian problems. Therefore, our main efforts are on developing a low-complexity algorithm based on BSUM/block coordinate descent (BCD) for Problem $\mathcal{P}_{\text{Au}}$ with unit-modulus constraints.[4] The principle of BCD is to optimize the objective function with respect to one block of variables while fixing the other blocks in each iteration. This leads to the following three subproblems.

1) THE SUBPROBLEM w.r.t. $\mathbf{V}$

For fixed $\mathbf{W}$ and $\mathbf{Y}$. The subproblem with respect to $\mathbf{V}$ is accordingly given by

$$\mathcal{P}_{\mathbf{V}}: \min_{\mathbf{V}} \ \|\mathbf{VW}-\mathbf{W}_{\text{FD}}\|_F^2 + \frac{1}{2\rho}\|\mathbf{Y}-\mathbf{VW}+\rho\mathbf{Z}\|_F^2 \quad \text{s.t.} \ \ (6).$$

From the special structure of the analog beamformer $\mathbf{V}$, as shown in (1), we observe that the m-th ($m = (n-1)L + l$, $n \in \mathcal{N}$, $l \in \mathcal{L}$) row of $\mathbf{VW}$ is the product of $(\mathbf{v}_n)_l$ and

[4]In this article, the objective function itself is chosen as the upper bound of the BSUM method. Thus, the BSUM reduces to the special case of BCD [32].

the n-th row of $\mathbf{W}$. Thus, Problem $\mathcal{P}_{\mathbf{V}}$ can be divided into M independent subproblems. Specifically, the m-th subproblem is given by

$$\min_{(\mathbf{v}_n)_l} \|(\mathbf{v}_n)_l(\mathbf{W})_{n,:} - (\mathbf{W}_{\text{FD}})_{m,:}\|_2^2 + \frac{1}{2\rho}\|(\mathbf{v}_n)_l(\mathbf{W})_{n,:} - (\mathbf{A})_{m,:}\|_2^2, \quad \text{s.t. } |(\mathbf{v}_n)_l| = 1,$$

where $\mathbf{A} \triangleq \mathbf{Y} + \rho\mathbf{Z}$. With some appropriate rearrange, the above problem can be reformulated as

$$\min_{(\mathbf{v}_n)_l} \ -2\Re\Big[\big((\mathbf{W}_{\text{FD}})_{m,:} + \frac{1}{2\rho}(\mathbf{A})_{m,:}\big)(\mathbf{W})_{n,:}^H(\mathbf{v}_n)_l^*\Big] \quad \text{s.t.} \ \ |(\mathbf{v}_n)_l| = 1.$$

Note that these M subproblems can be solved in parallel. For the m-th subproblem, there exists a closed-form solution expressed as

$$\angle\,[(\mathbf{v}_n)_l] = \angle\Big[\big((\mathbf{W}_{\text{FD}})_{m,:} + \frac{1}{2\rho}(\mathbf{A})_{m,:}\big)(\mathbf{W})_{n,:}^H\Big]. \tag{24}$$

2) THE SUBPROBLEM w.r.t. W

The variable $\mathbf{W}$ only appears in the objective function of Problem $\mathcal{P}_{\text{Au}}$. Thus, the subproblem with respect to $\mathbf{W}$ is accordingly given by

$$\mathcal{P}_{\mathbf{W}}: \min_{\mathbf{W}} \ \|\mathbf{V}\mathbf{W} - \mathbf{W}_{\text{FD}}\|_F^2 + \tfrac{1}{2\rho}\|\mathbf{Y} - \mathbf{V}\mathbf{W} + \rho\mathbf{Z}\|_F^2$$

It can be verified that Problem $\mathcal{P}_{\mathbf{W}}$ is an unconstrained convex problem for fixed $\mathbf{V}$ and $\mathbf{Y}$, which can be equivalently reformulated as

$$\min_{\mathbf{W}} \ \alpha\cdot\text{Tr}(\mathbf{W}^H\mathbf{V}^H\mathbf{V}\mathbf{W}) - 2\Re\big[\,\text{Tr}\big(\mathbf{W}^H\mathbf{V}^H(\mathbf{W}_{\text{FD}} + \frac{1}{2\rho}\mathbf{A})\big)\big]$$

where $\alpha \triangleq (1 + \frac{1}{2\rho})$. Since $\mathbf{V}^H\mathbf{V} = L\cdot\mathbf{I}$ is positive definite, we can express the optimal $\mathbf{W}$ as

$$\mathbf{W} = \frac{1}{\alpha L}\mathbf{V}^H(\mathbf{W}_{\text{FD}} + \frac{1}{2\rho}\mathbf{A}). \tag{25}$$

3) THE SUBPROBLEM w.r.t. Y

The auxiliary variable $\mathbf{Y}$ only appears in the second term of the objective function of Problem $\mathcal{P}_{\text{Au}}$. Thus, for fixed $\mathbf{V}$ and $\mathbf{W}$, the subproblem with respect to $\mathbf{Y}$ is accordingly given by

$$\mathcal{P}_{\mathbf{Y}}: \min_{\mathbf{Y}} \ \|\mathbf{Y} - \mathbf{V}\mathbf{W} + \rho\mathbf{Z}\|_F^2 \quad \text{s.t. (23).}$$

Since $\mathbf{\Xi}$ is a diagonal matrix with full rank, the constraint in (23) can be rewritten as $\tilde{\mathbf{G}}\mathbf{Y} = \mathbf{I}$, where $\tilde{\mathbf{G}} \triangleq \mathbf{\Xi}^{-1}\mathbf{G}$. This constraint is related to the concept of the generalized inverse in linear algebra. In fact, $\mathbf{Y}$ denotes the generalized inverse of $\tilde{\mathbf{G}}$. According to [30], the generalized inverse is not unique and any generalized inverse can be expressed as

$$\mathbf{Y} = \tilde{\mathbf{G}}^\dagger + \mathbf{D}_\perp\mathbf{F}, \tag{26}$$

where $\mathbf{D}_\perp \triangleq \mathbf{I} - \tilde{\mathbf{G}}^\dagger\tilde{\mathbf{G}}$ is the orthogonal projection onto the null space of $\tilde{\mathbf{G}}$ and $\mathbf{F}$ is an arbitrary matrix. By substituting (26) into the objective function of Problem $\mathcal{P}_{\mathbf{Y}}$ and removing constant terms, Problem $\mathcal{P}_{\mathbf{Y}}$ can be reformulated as

$$\min_{\mathbf{F}} \text{Tr}(\mathbf{F}^H\mathbf{D}_\perp^H\mathbf{D}_\perp\mathbf{F}) - 2\Re\big[\,\text{Tr}(\mathbf{F}^H(\mathbf{D}_\perp^H\mathbf{B} - \mathbf{D}_\perp^H\tilde{\mathbf{G}}^\dagger))\big]$$

where $\mathbf{B} \triangleq (\mathbf{V}\mathbf{W} - \rho\mathbf{Z})$. The above unconstrained problem has a closed-form solution expressed as

$$\mathbf{F} = (\mathbf{D}_\perp^H\mathbf{D}_\perp)^{-1}(\mathbf{D}_\perp^H\mathbf{B} - \mathbf{D}_\perp^H\tilde{\mathbf{G}}^\dagger). \tag{27}$$

The details are summarized in Algorithm 2, which consists of an inner loop and an outer loop. In the inner loop, we solve Problem $\mathcal{P}_{\text{Au}}$ based on BCD. For given ρ and $\mathbf{Z}$, since the objective value of Problem $\mathcal{P}_{\text{Au}}$ is nonnegative and is monotonically non-increasing with i, the BCD iteration converges to a limit point. As the three constraints sets of Problem $\mathcal{P}_{\mathbf{V}}$, Problem $\mathcal{P}_{\mathbf{W}}$, and Problem $\mathcal{P}_{\mathbf{Y}}$ are disjoint, the limit point obtained by the BCD iteration in the inner loop is a stationary point of Problem $\mathcal{P}_{\text{Au}}$ [32]. In the outer loop, the dual variable $\mathbf{Z}$ and the penalty weight ρ for the augmented Lagrangian Problem $\mathcal{P}_{\text{Au}}$ are updated according to the constraint violation condition related to τ (see Step 12-Step 16 of Algorithm 2, where $c \in (0, 1)$ is a weight control parameter). For a sufficiently small ρ, the penalty term $\|\mathbf{Y} - \mathbf{V}\mathbf{W} + \rho\mathbf{Z}\|_F^2$ will be driven to zero, and it no longer contributes to the objective function. The outer loop is terminated when the penalized constraint $\mathbf{Y} = \mathbf{V}\mathbf{W}$ is achieved [31] and a practical choice of the termination condition is $\|\mathbf{Y} - \mathbf{V}\mathbf{W}\|_F^2 \le \epsilon$, where ϵ is a predefined small constant, e.g., $\epsilon = 10^{-6}$. Note that the inner loop and outer loop of Algorithm 2 set up a standard PDD framework [31]. Thus, we can conclude that Algorithm 2 converges to a stationary point of Problem $\mathcal{P}_{\text{MAP-ZF}}$.

B. CASE OF K > N

In this case, due to the limited matrix dimension of the digital beamformer $\mathbf{W}$, the constraint in (22) cannot be satisfied by any $\mathbf{W}$. Thus, the feasibility of the penalized constraints $\mathbf{Y} = \mathbf{V}\mathbf{W}$ cannot be achieved by Algorithm 2. To derive a low-complexity algorithm, we first transform Problem $\mathcal{P}_{\text{MAP-ZF}}$ into an analog beamformer design problem by performing some projection. Specifically, we perform a projection onto $\mathbf{\Xi}$ to obtain $\mathbf{W}$, as[5]

$$\mathbf{W} = (\mathbf{G}\mathbf{V})^\dagger\mathbf{\Xi}. \tag{28}$$

Substituting (28) into the objective function of Problem $\mathcal{P}_{\text{MAP-ZF}}$, we have the following analog beamformer design problem.

$$\min_{\mathbf{V}} \ \|\mathbf{V}(\mathbf{G}\mathbf{V})^\dagger\mathbf{\Xi} - \mathbf{W}_{\text{FD}}\|_F^2 \quad \text{s.t. (6).} \tag{29}$$

Due to the unit-modulus constraints in (6), the problem in (29) is still difficult to deal with. In the following, we develop an

[5]The projection in (28) leads to a temporary violation of the individual SINR constraints in (5), which will be dealt with later in the design of the digital beamformer.

Algorithm 2 Low-Complexity Hybrid Beamforming Design for the Case of $K \leq N$

Initialization: Choose proper ρ, τ, c, $\mathbf{Y}$, and $\mathbf{Z}$;
1: Obtain $\mathbf{W}_{\mathrm{FD}}$ using the method in [25];
2: **while** $\|\mathbf{Y}-\mathbf{V}\mathbf{W}\|_F^2 \leq \epsilon$ **do**
3: Initialize proper $\mathbf{W}^{(0)}$. $\mathbf{Y}^{(0)}=\mathbf{Y}$;
4: $i:=0$;
5: **repeat**
6: Given $\mathbf{W}^{(i)}$, $\mathbf{Y}^{(i)}$, ρ, and $\mathbf{Z}$, update $\mathbf{V}^{(i+1)}$ according to (24);
7: Given $\mathbf{V}^{(i+1)}$, $\mathbf{Y}^{(i)}$, ρ, and $\mathbf{Z}$, update $\mathbf{W}^{(i+1)}$ according to (25);
8: Given $\mathbf{V}^{(i+1)}$, $\mathbf{W}^{(i+1)}$, ρ, and $\mathbf{Z}$, update $\mathbf{Y}^{(i+1)}$ according to (26) and (27);
9: $i \leftarrow i+1$;
10: **until** convergence criterion is met;
11: $\mathbf{Y}=\mathbf{Y}^{(i)}$;
12: **if** $\|\mathbf{Y}^{(i)}-\mathbf{V}^{(i)}\mathbf{W}^{(i)}+\rho\mathbf{Z}\|_\infty \leq \tau$ **then**
13: $\mathbf{Z}=\mathbf{Z}+\frac{1}{\rho}(\mathbf{Y}^{(i)}-\mathbf{V}^{(i)}\mathbf{W}^{(i)})$;
14: **else**
15: $\rho:=c\rho$;
16: **end if**
17: $\tau:=0.95\|\mathbf{Y}^{(i)}-\mathbf{V}^{(i)}\mathbf{W}^{(i)}\|_\infty$;
18: **end while**

iterative procedure to address this problem. The main idea is to successively solve a sequence of approximations of the problem in (29). Specifically, the approximate problem at iteration i is given by

$$\mathbf{V}^{(i+1)} = \arg\min_{\mathbf{V}} \ \|\mathbf{V}(\mathbf{G}\mathbf{V}^{(i)})^\dagger\boldsymbol{\Xi}-\mathbf{W}_{\mathrm{FD}}\|_F^2 \quad \text{s.t.} \quad (6).$$

Thanks to the special structure of the analog beamformer $\mathbf{V}$ as described in Section IV-A1, this problem can be divided into M independent subproblems, and the m-th ($m=(n-1)L+l,\ n\in\mathcal{N},\ l\in\mathcal{L}$) subproblem can be reformulated as

$$(\mathbf{v}_n^{(i+1)})_l = \arg\min_{(\mathbf{v}_n)_l} \ -2\Re\left[(\mathbf{W}_{\mathrm{FD}})_{m,:}(\mathbf{C}^{(i)})_{n,:}^H(\mathbf{v}_n)_l^*\right] \quad \text{s.t.} \quad |(\mathbf{v}_n)_l|=1,$$

where $\mathbf{C}^{(i)} \triangleq (\mathbf{G}\mathbf{V}^{(i)})^\dagger\boldsymbol{\Xi}$. These M independent subproblems can be solved in parallel and each has a closed-form solution expressed as

$$\angle\left[(\mathbf{v}_n^{(i+1)})_l\right] = \angle\left[(\mathbf{W}_{\mathrm{FD}})_{m,:}(\mathbf{C}^{(i)})_{n,:}^H\right]. \tag{30}$$

After finding an analog beamformer of the problem in (29), the digital beamformer is designed to satisfy the individual SINR constraints in (5). By treating the product of $\mathbf{G}$ and the obtained $\mathbf{V}^{(i)}$ as an equivalent channel, we obtain the following digital beamformer design problem.

$$\min_{\mathbf{W}} \ \|\mathbf{V}^{(i)}\mathbf{W}\|_F^2 \quad \text{s.t.} \quad \frac{\left|\mathbf{g}_k^H\mathbf{V}^{(i)}\mathbf{w}_k\right|^2}{\sum_{j\in\mathcal{K},j\neq k}\left|\mathbf{g}_k^{\mathrm{H}}\mathbf{V}^{(i)}\mathbf{w}_j\right|^2+\sigma_k^2} \geq \eta_k,\ k\in\mathcal{K}. \tag{31}$$

Algorithm 3 Low-Complexity Hybrid Beamforming Design for the Case of $K > N$

Initialization: Choose proper $\mathbf{V}^{(0)}$;
1: Obtain $\mathbf{W}_{\mathrm{FD}}$ using the method in [25];
2: $i:=0$
3: **repeat**
4: Given $\mathbf{V}^{(i)}$, update $\mathbf{V}^{(i+1)}$ according to (30);
5: $i \leftarrow i+1$;
6: **until** convergence criterion is met;
7: Given $\mathbf{G}\mathbf{V}^{(i)}$, obtain the optimal $\mathbf{W}$ of problem in (31) using the method in [25].

The problem in (31) is essentially a low-dimensional fully-digital beamforming design problem (with N RF chains). Its optimal solution can be obtained using an efficient method based on a semi-closed form solution [25]. Therefore, we summarize our hybrid beamforming design for the case of $K > N$ as Algorithm 3.[6]

Finally, we remark on the computational complexity of the proposed algorithms. Algorithm 2 and Algorithm 3 are both iterative algorithms. The optimization problem in each iteration can be solved by computing a closed-form solution. Therefore, compared with Algorithm 1, the computational complexity of Algorithm 2 and Algorithm 3 is significantly reduced. This motivates us to generate multiple initial points (i.e., initial $\mathbf{Y}$ and $\mathbf{Z}$ for Algorithm 2 and initial $\mathbf{V}^{(0)}$ for Algorithm 3) to run Algorithm 2 and Algorithm 3 multiple times and choose the smallest transmission power for improving performance. Meanwhile, as Algorithm 2 and Algorithm 3 can update the hybrid beamformer efficiently, the considered objective functions can serve as loss functions in a Neural Network framework. This will offer valuable insights for the implementation of hybrid beamforming based on Neural Network.

V. NUMERICAL RESULTS

In this section, we present the numerical results of the proposed algorithms in the multi-user MIMO systems. We consider two baseline schemes, GA in [23] and the optimal fully-digital beamformer $\mathbf{W}_{\mathrm{FD}}$ (optimal FD) in [25]. For GA, the population size and the evolution generation are set to 100 and 20, respectively. In each generation, the thresholds for arithmetic crossover and non-uniform mutation are set to 0.5 and 0.1, respectively. The maximum number of attempts in the crossover is 5. For the optimal FD, the number of RF chains N equals to M. In the simulations, we generate i.i.d. Gaussian channels with zero mean and unit variance and show the performance averaged over 50 channel realizations for each scheme. We choose $\sigma_k^2=1$ and $\eta_k=\eta,\ k\in\mathcal{K}$. Based on [31], we choose a practical weight control parameter by setting $c=0.9$. For Algorithm 1, the initial

[6]It is worth noting that although the convergence issue remains open, numerical results still show good convergence performance for Algorithm 3. As its computational complexity is greatly reduced, Algorithm 3 can serve as an excellent candidate for the implementation of hybrid beamforming.

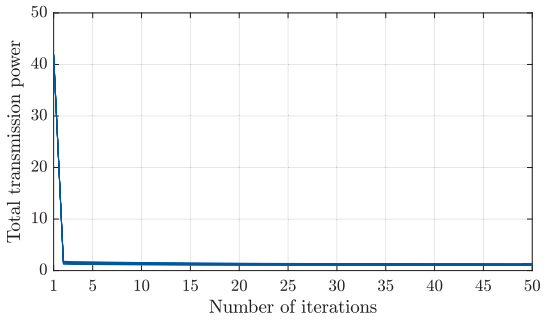


FIGURE 2. Ten examples of the convergence behavior of Algorithm 1 when $N = 8$, $L = 6$, $K = 8$, and $\eta = 2$ (dB).

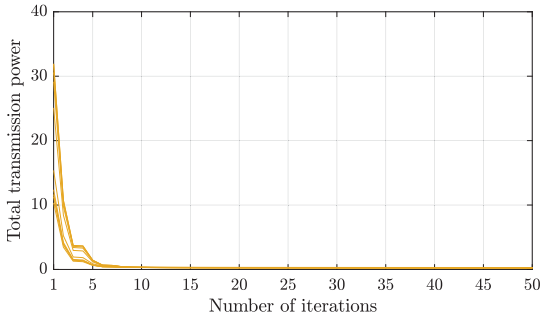


FIGURE 3. Ten examples of the convergence behavior of Algorithm 2 when $N = 8$, $L = 6$, $K = 8$, and $\eta = 2$ (dB).

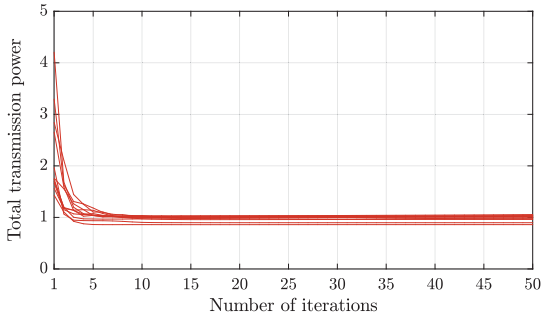


FIGURE 4. Ten examples of the convergence behavior of Algorithm 3 when $N = 8$, $L = 6$, $K = 10$, and $\eta = 2$ (dB).

parameter μ is set to 1. For Algorithm 2, the initial parameters τ and ρ are set to 1000 and 1, respectively. In evaluating the proposed algorithms, we use the same convergence criterion: $|\nu^{(i)} - \nu^{(i-1)}|/\nu^{(i-1)} \leq 10^{-3}$ or the number of iterations exceeds 50, where $\nu^{(i)}$ is the objective value at iteration i. Besides, we show the performance in terms of normalized transmission power which is unit-less, while all the individual SINR constraints in (5) are satisfied.

A. CONVERGENCE BEHAVIOR OF THE PROPOSED ALGORITHMS

The convergence behavior of the proposed algorithms is examined in Fig. 2, Fig. 3, and Fig. 4. In each figure, 10 examples of the convergence behavior of the corresponding algorithm are presented. Each curve corresponds to an example of a randomly generated multi-user channel realization. Note that Algorithm 1 and Algorithm 2 both involve an outer loop for finding a penalty weight or a dual variable and an inner loop for iteratively solving a corresponding optimization problem. Thus, we here focus on the convergence behavior of the inner loop in the first outer loop iteration with initial parameters. Fig. 2 shows that Algorithm 1 converges very fast (within almost 3 iterations) with an initial μ. This demonstrates the efficiency and practicability of the initialization method as described in Section III-C1. Fig. 3 and Fig. 4 show that Algorithm 2 and Algorithm 3 typically converge within 10 iterations for all observed channel realizations. The curves in Fig. 3 converge to very close values, showing that the convergent value of the inner loop of Algorithm 2 is not sensitive to random channel realizations. Moreover, it is observed from Fig. 4 that even if there are small fluctuations on some curves, these curves will always converge, which again verifies the excellent convergence performance of Algorithm 3.

B. PERFORMANCE EVALUATION

In this subsection, we show the performance evaluation of the proposed algorithms and compare them with the existing GA in [23] and the optimal FD in [25]. Note that Algorithm 1 directly solves Problem $\mathcal{P}_{\text{Ori}}$ under individual SINR constraints and unit-modulus constraints, which will improve the power consumption compared with existing methods. Therefore, Algorithm 1 can serve as a benchmark of the performance in terms of transmission power.

Fig. 5 illustrates the average transmission power for different K when $N = 8$ and $\eta = 2$ (dB). We can observe that when $L = 6$, Algorithm 1 outperforms the other algorithms in both the case of $K \leq N$ and the case of $K > N$, indicating that Algorithm 1 can provide benchmark performance. Furthermore, in the case of $K \leq N$, Algorithm 2 can achieve transmission power that is sufficiently close to that of Algorithm 1. In the case of $K > N$, there is a small performance gap between Algorithm 3 and Algorithm 1 increasing with K. This is because in this case, the inter-user interference cannot be completely canceled, and thus the projection in (28) entails some non-negligible performance loss. Nevertheless, the transmission power achieved by Algorithm 3 is still much lower than the one achieved by GA, and the performance gap between Algorithm 3 and GA increases rapidly with K. When $L = 16$, Algorithm 2 (when $K \leq N$) and Algorithm 3 (when $K = 9$) can achieve lower transmission power than Algorithm 1. Meanwhile, the performance gap between Algorithm 3 and Algorithm 1 is reduced, especially when K is small. Thus, Fig. 5 demonstrates the huge performance improvement of Algorithm 1 over GA and shows the effectiveness of Algorithm 2 and Algorithm 3, which have much lower computational complexity.

Fig. 6 shows the average simulation time (reflecting computational complexity) of the proposed algorithms and GA versus K. It is observed that compared with Algorithm 1 and GA, the computational complexity of Algorithm 2 and Algorithm 3 is dramatically reduced. Besides, the

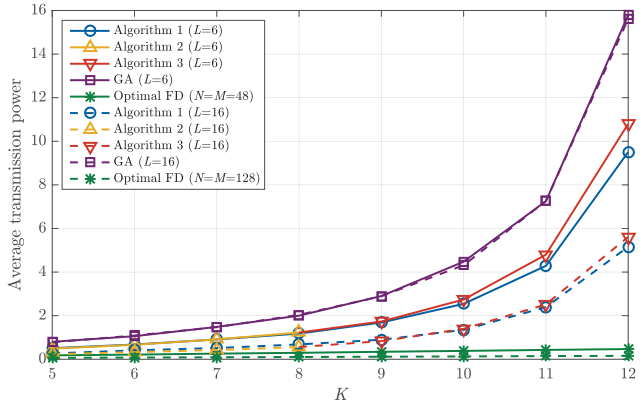


FIGURE 5. Average transmission power versus K when $N = 8$ and $\eta = 2$ (dB).

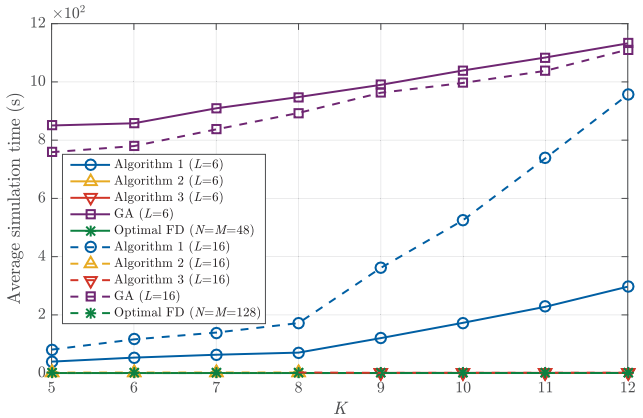
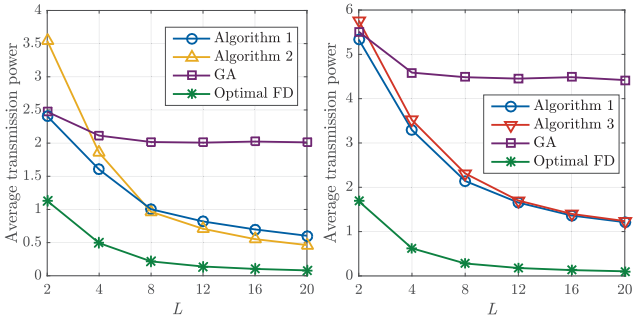


FIGURE 6. Average simulation time versus K when $N = 8$ and $\eta = 2$ (dB).

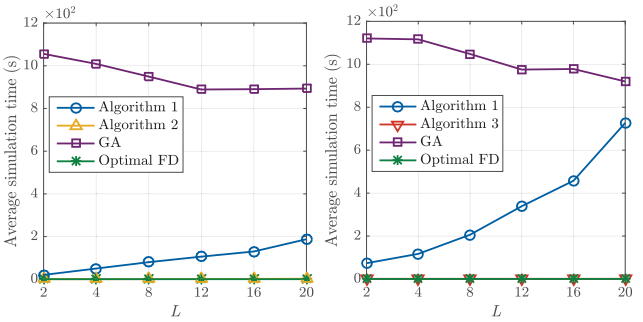
computational complexity of Algorithm 2 and Algorithm 3 remains almost unchanged over the investigated range of K, while the computational complexity of Algorithm 1 and GA increases nearly linearly with K. It is also found in Fig. 6 that although the computational complexity of Algorithm 1 is relatively high and grows faster when $L = 16$ in the case of $K > N$, it is still much lower than that of GA. Fig. 5 and Fig. 6 indicate that Algorithm 1 outperforms existing GA in terms of both transmission power and computational complexity. As Algorithm 2 and Algorithm 3 can achieve most of the performance of Algorithm 1 with dramatically reduced computational complexity, they can serve as excellent candidates for the implementation of hybrid beamforming in practical systems.

Next, we investigate the impact of the other system parameters (i.e., L and η) on the performance of the proposed algorithms in both the case of $K \leq N$ and the case of $K > N$. Fig. 7 compares the average transmission power versus L. It is seen that while the performance gap between GA and Algorithm 1 increases with L, the performance gap between Algorithm 2 and Algorithm 1 and the performance gap between Algorithm 3 and Algorithm 1 both decrease with L. The performance gap between the proposed algorithms and the optimal FD decreases with L, while the



(a) $N = 8$ and $K = 8$. (b) $N = 8$ and $K = 10$.

FIGURE 7. Average transmission power versus L when $\eta = 2$ (dB).

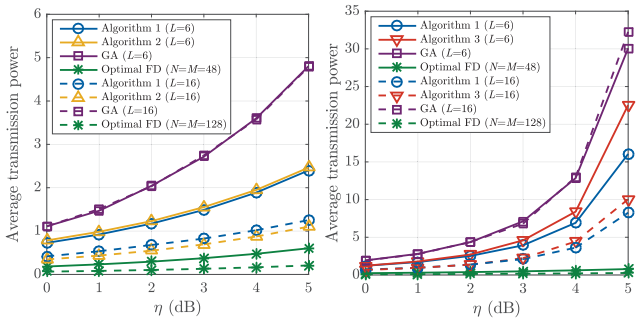


(a) $N = 8$ and $K = 8$. (b) $N = 8$ and $K = 10$.

FIGURE 8. Average simulation time versus L when $\eta = 2$ (dB).

performance gap between GA and the optimal FD increases with L. Besides, in Fig. 7a, there is an intersection point of the curves of Algorithm 1 and Algorithm 2, showing that Algorithm 2 can provide better performance with lower computational complexity when L is relatively large. This can be explained by the fact that increasing the number of antenna elements in each sub-array can provide higher channel gain and thus improve the performance of ZF processing adopted in deriving Algorithm 2 and Algorithm 3. Fig. 8 shows the average simulation time versus L of the proposed algorithms, GA, and the optimal FD. In both the case of $K \leq N$ and the case of $K > N$, Algorithm 2 and Algorithm 3 have much lower computational complexity than GA. Fig. 7 and Fig. 8 again imply the excellent effectiveness of Algorithm 2 and Algorithm 3 while enjoying the benefit of much lower computational complexity especially when the number of antenna elements in each sub-array is comparable with the number of RF chains.

Fig. 9 illustrates the average transmission power achieved by the proposed algorithms and GA when different prescribed SINR requirements are set. When $L = 6$, we can observe that Algorithm 1 outperforms GA, and their performance gap increases rapidly with η. Meanwhile, Fig. 9 indicates that increasing L will improve the performance of the low-complexity Algorithm 2 and Algorithm 3. Moreover, in Fig. 9a, the curve of Algorithm 2 nearly coincides with that of Algorithm 1, which indicates that in the case of $K \leq N$, increasing the SINR requirement η will not incur more performance loss for Algorithm 2. When



(a) $N = 8$ and $K = 8$. (b) $N = 8$ and $K = 10$.

FIGURE 9. Average transmission power versus η.

$L = 16$, Algorithm 2 outperforms Algorithm 1 and the performance gap between Algorithm 3 and Algorithm 1 is reduced. In Fig. 9b, Algorithm 3 achieves similar transmission power compared with Algorithm 1 when the prescribed SINR requirements are relatively small.

As can be observed in Fig. 5, Fig. 7, and Fig. 9, when L is relatively small, there is a tradeoff between transmission power and computational complexity. Algorithm 1 achieves lower transmission power while Algorithm 2 and Algorithm 3 achieve lower computational complexity. Moreover, by increasing L (at the sacrifice of more hardware cost and implementation complexity), the performance loss of Algorithm 2 and Algorithm 3 can be reduced and the performance of Algorithm 2 and Algorithm 3 can even outperform Algorithm 1 with relatively large L. This can provide guidance for implementing hybrid beamforming in practical systems.

VI. CONCLUSION

In this article, we considered the partially-connected hybrid beamforming design in a downlink unicast multi-user massive MIMO system to minimize the transmission power under individual SINR constraints and unit-modulus constraints. By applying the penalty method, we proposed an iterative alternating algorithm to solve the challenging non-convex problem. By utilizing matrix approximation, we also proposed low-complexity algorithms in two cases depending on the number of users and the number of RF chains. Numerical results demonstrated significant performance gains of the proposed algorithms over the existing GA. The computational complexity of the proposed low-complexity algorithms is significantly reduced while incurring some almost negligible performance loss. In the future, we will extend the proposed algorithms to MIMO-orthogonal frequency division multiplexing (OFDM) systems, as well as to systems with multiple-antenna users. Furthermore, it will be interesting to combine the low-complexity algorithms with Neural Network frameworks to develop more effective algorithms. Specifically, low-complexity algorithms can be employed to generate a large-scale training set. Also, the theoretical convergence analysis of the low-complexity algorithm for the case where the number of users is larger than that of RF chains will need further investigation.

REFERENCES

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

[2] 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.

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

[4] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 949–973, 2nd Quart., 2016.

[5] 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.

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

[7] N. Garcia, H. Wymeersch, C. Fager, and E. G. Larsson, "mmWave hybrid array with more users than RF chains," 2017, *arXiv:1709.05200*. [Online]. Available: http://arxiv.org/abs/1709.05200

[8] 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.

[9] C. G. Tsinos, S. Maleki, S. Chatzinotas, and B. Ottersten, "On the energy-efficiency of hybrid analog–digital transceivers for single- and multi-carrier large antenna array systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1980–1995, Sep. 2017.

[10] A. Arora, C. G. Tsinos, B. S. M. R. Rao, S. Chatzinotas, and B. Ottersten, "Hybrid transceivers design for large-scale antenna arrays using majorization-minimization algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 701–714, Jan. 2020.

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

[12] 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.

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

[14] Z. Wang, M. Li, X. Tian, and Q. Liu, "Iterative hybrid precoder and combiner design for mmWave multiuser MIMO systems," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1581–1584, Jul. 2017.

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

[16] C. Hu, Y. Liu, L. Liao, and R. Zhang, "Hybrid beamforming for multi-user MIMO with partially-connected RF architecture," *IET Commun.*, vol. 13, no. 10, pp. 1356–1363, Jun. 2019.

[17] J. Noh, T. Kim, J.-Y. Seol, and C. Lee, "Zero-forcing based hybrid beamforming for multi-user millimeter wave systems," *IET Commun.*, vol. 10, no. 18, pp. 2670–2677, Dec. 2016.

[18] X. Wu, D. Liu, and F. Yin, "Hybrid beamforming for multi-user massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3879–3891, Sep. 2018.

[19] G. Zang, Y. Cui, H. V. Cheng, F. Yang, L. Ding, and H. Liu, "Optimal hybrid beamforming for multiuser massive MIMO systems with individual SINR constraints," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 532–535, Apr. 2019.

[20] 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.

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

[22] J. Geng, W. Xiang, D. Yang, Z. Wei, and N. Li, "Multi-user hybrid analogue/digital beamforming for relatively large-scale antenna systems," *IET Commun.*, vol. 8, no. 17, pp. 3038–3049, Nov. 2014.

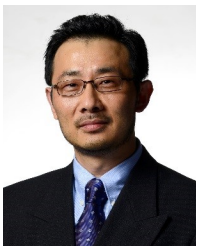
[23] C. Hong, Z. Wei, J. Geng, and D. Yang, "Multiuser hybrid phase-only analog/digital beamforming with genetic algorithm," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2014, pp. 517–521.
[24] M. Alodeh, D. Spano, A. Kalantari, C. G. Tsinos, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Symbol-level and multicast precoding for multiuser multiantenna downlink: A state-of-the-art, classification, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1733–1757, 3rd Quart., 2018.
[25] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.
[26] A. H. Phan, H. D. Tuan, H. H. Kha, and D. T. Ngo, "Nonsmooth optimization for efficient beamforming in cognitive radio multicast transmission," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2941–2951, Jun. 2012.
[27] R. A. Delgado, J. C. Agüero, and G. C. Goodwin, "A rank-constrained optimization approach: Application to factor analysis," *IFAC Proc. Vols.*, vol. 47, no. 3, pp. 10373–10378, 2014.
[28] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, no. 3, pp. 127–136, Apr. 2000.
[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[30] A. Wiesel, Y. C. Eldar, and S. Shamai (Shitz), "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.
[31] Q. Shi and M. Hong, "Spectral efficiency optimization for millimeter wave multiuser MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 455–468, Jun. 2018.
[32] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, CA, USA: Athena scientific, 1999.

GUANGDA ZANG received the B.S. degree from the Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu, China, in 2014. He is currently pursuing the Ph.D. degree in electronic engineering with Shanghai Jiao Tong University, Shanghai, China.

LINGNA HU was born in 1978. She received the Ph.D. degree in information and communication from Shanghai Jiao Tong University. Her research interest includes satellite communication systems.

FENG YANG (Member, IEEE) received the Ph.D. degree in information and communication from Shanghai Jiao Tong University. Since 2008, he has been on the Faculty of Shanghai Jiao Tong University, where he is currently an Associate Professor with the Department of Electronic Engineering. He takes part in the programme of Beyond 3G Wireless Communication Testing System and is in charge of system design. He is also the PI of some national projects, including the National High Technology Research and Development Program of China (863 Program) and the National Natural Science Foundation of China. His research interests include wireless video communication and multihop communication.

LIANGHUI DING (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University (SJTU), China, in 2009. From 2009 to 2010, he was a Researcher of signals and systems with Uppsala University, Sweden. He is currently an Assistant Professor with the Institute of Image Communication and Network Engineering, SJTU. His research interests include network optimization, resource allocation, scheduling, and network coding.



HUI LIU (Fellow, IEEE) received the B.S. degree in electrical engineering from Fudan University, Shanghai, China, in 1988, and the Ph.D. degree in electrical engineering from The University of Texas at Austin, Austin, TX, USA, in 1995. He was a Full Professor and an Associate Chairman with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, and also a Chair Professor and an Associate Dean of the School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University. He was one of the principal designers of the 3G TD-SCDMA mobile technologies. He was the Founder of Adaptix, which pioneered the development of OFDMA-based mobile broadband networks (mobile WiMAX and 4G LTE). He is currently the President and the CTO with Silkwave Holdings and also an Affiliate Professor of the University of Washington. He contributed to the global standards for broadband cellular and mobile broadcasting. His research interests include broadband wireless networks, satellite communications, digital broadcasting, and multimedia signal processing. He was a recipient of the 1997 NSF CAREER Award, the Gold Prize Patent Award in China, three IEEE best conference paper awards, and the 2000 ONR Young Investigator Award.