Movable Antenna-Aided Hybrid Beamforming for Multi-User Communications

Yichi Zhang, *Graduate Student Member, IEEE*, Yuchen Zhang, *Member, IEEE*, Lipeng Zhu, *Member, IEEE*, Sa Xiao, Wanbin Tang, Yonina C. Eldar, *Fellow, IEEE*, and Rui Zhang, *Fellow, IEEE*

***Abstract*—In this correspondence, we propose a movable antenna (MA)-aided multi-user hybrid beamforming scheme with a sub-connected structure, where multiple movable sub-arrays can independently change their positions within different local regions. To maximize the system sum rate, we jointly optimize the digital beamformer, analog beamformer, and positions of sub-arrays, under the constraints of unit modulus, finite movable regions, and power budget. Due to the non-concave/non-convex objective function/constraints, as well as the highly coupled variables, the formulated problem is challenging to solve. By employing fractional programming, we develop an alternating optimization framework to solve the problem via a combination of Lagrange multipliers, penalty method, and gradient descent. Numerical results reveal that the proposed MA-aided hybrid beamforming scheme significantly improves the sum rate compared to its fixed-position antenna (FPA) counterpart. Moreover, with sufficiently large movable regions, the proposed scheme with sub-connected MA arrays even outperforms the fully-connected FPA array.**

***Index Terms*—Movable antenna, hybrid beamforming, multi-user communication, sum-rate maximization.**

I. Introduction

Utilizing high-frequency millimeter-wave and/or terahertz bands for wireless communication has been recognized as an essential trend for advancing beyond 5 G systems. This is attributed to the enhanced data rate available by exploiting their ultra-broad bandwidths. In order to compensate for the severe path loss in these bands, large antenna arrays are usually needed to provide high beamforming gains. Unfortunately, the heavy power consumption and high cost of radio frequency (RF) chains make fully-digital beamforming impractical. Hybrid beamforming [1], [2], which leverages a small number of RF chains and an analog front end consisting of phase shifters (PSs), has been considered as a promising technique to achieve a good trade-off between hardware cost and communication performance.

Based on the fabrication of PS enabled front end, the hybrid beamformer can be categorized into fully-connected and sub-connected structures [3]. The fully-connected structure, with all RF chains connected to all antennas through PSs, allows more design degrees of freedom (DoFs). However, the excessive use of PSs entails a large cost. In comparison, by connecting each RF chain only with a portion of antennas, the sub-connected structure is more energy efficient and easier to implement [4]. To improve the communication performance with a limited number of RF chains, different algorithms have been proposed [1], [2]. In [1], the hybrid beamformer was optimized by approximating the fully-digital beamformer in an iterative manner. In [2], the authors developed a general framework for hybrid beamforming design under various hardware architectures. However, both existing fully-connected and sub-connected hybrid beamforming schemes rely on fixed-position antennas (FPAs), which limit their communication performance because the spatial variation of wireless channels is not fully exploited at the transmitter/receiver.

Recently, the movable antenna (MA), also known as fluid antenna system, has been proposed as a promising technology to enhance wireless communication performance [5], [6], [7], [8]. Specifically, an MA can flexibly tune its position and/or rotation to reconfigure the wireless channel towards a more favorable condition. Several prior works have investigated MA-aided systems for improving communication performance [9], [10], [11], [12], [13]. The authors in [9] initially introduced a field-response based channel model for MA-aided systems. Based on this model, the work [10] characterized the capacity of an MA-aided multi-input multi-output (MIMO) system. Compared to antenna selection, MA systems can yield better communication performance using fewer antennas because continuous movement of antennas in a large region can fully exploit the channel spatial variation, thereby reducing antenna costs. In [11], [12], [13], the MA was generalized to multi-user systems. The authors in [12] extended MAs to secure communication by jointly optimizing the MA array geometry and beamforming vector. Most existing works considered that each MA is connected to an RF chain and a motor (for antenna movement) [8], [9], [10], [11], [12], [13]. However, implementing a fully-digital structure of an MA array with a large number of motors brings additional power consumption and significant challenges in system implementation, such as the intertwining of antenna connecting wires during movement.

The main contributions of this correspondence are as follows:

- Instead of independently moving each antenna element, each sub-array is driven by a single motor and moves within a restricted region such that the antennas therein can move collectively, which significantly reduces the hardware costs and simplifies the implementation complexity.
- Utilizing the proposed MA-aided sub-connected structure, we consider downlink multi-user communications with the aim of maximizing the system sum rate, which is non-convex and thus difficult to be optimally solved. Facilitated by fractional programming (FP), we propose an alternating optimization (AO) framework to solve the problem, by jointly applying Lagrange multipliers, penalty method, and gradient descent.
- Numerical results demonstrate that our proposed MA-aided hybrid beamforming scheme outperforms its FPA counterpart. Moreover, under certain practical conditions, the proposed scheme with sub-connected MA arrays yields a higher sum rate compared to the fully-connected FPA array.

Received 2 April 2024; revised 15 September 2024; accepted 4 January 2025. Date of publication 23 January 2025; date of current version 20 June 2025. This work was supported by the National Natural Science Foundation of China under Grant 62301117, Grant 62001094, and Grant U19B2014. The review of this article was coordinated by Prof. Hsiao-Feng Lu. *(Corresponding authors: Lipeng Zhu; Yuchen Zhang.)*

Yichi Zhang, Yuchen Zhang, and Wanbin Tang are with the National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: yczhang@std.uestc.edu.cn; yc_zhang@std.uestc.edu.cn; wbtang@uestc.edu.cn).

Lipeng Zhu is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: zhulp@nus.edu.sg).

Sa Xiao is with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Kash Institute of Electronics and Information Industry, Kash 844000, China (e-mail: xiaosa@uestc.edu.cn).

Yonina C. Eldar is with the Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

Rui Zhang is with the The Chinese University of Hong Kong, Shenzhen, and Shenzhen Research Institute of Big Data, Shenzhen 518172, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: rzhang@cuhk.edu.cn).

Digital Object Identifier 10.1109/TVT.2025.3533078

0018-9545 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2025 at 13:05:40 UTC from IEEE Xplore. Restrictions apply.

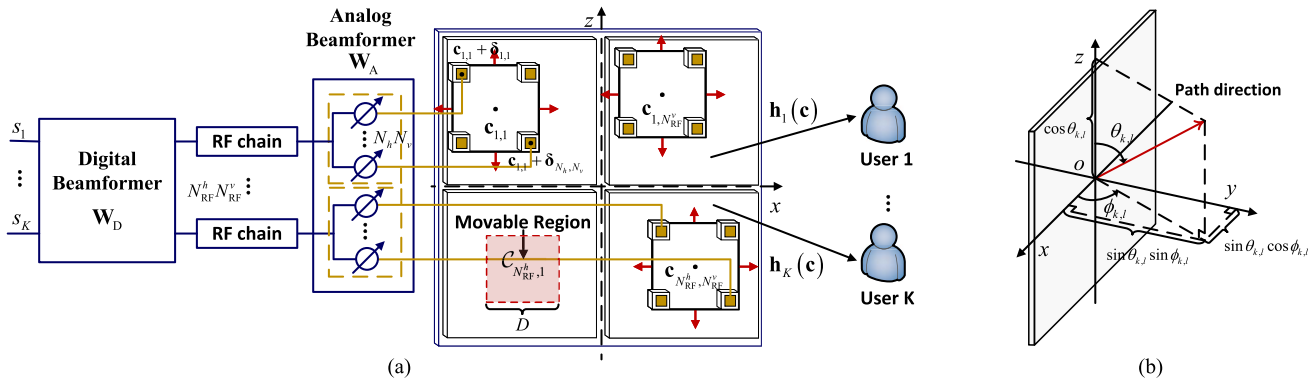

Fig. 1. (a) MA-aided sub-connected hybrid beamforming structure. (b) Illustration of spatial angles.

II. System Model

We consider an MA-aided downlink multi-user multi-input single-output (MU-MISO) system as illustrated in Fig. 1(a). The system comprises a multi-antenna base station (BS) and K single-antenna users indexed by $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$. The BS adopts hybrid beamforming with a sub-connected structure [1], [2], which consists of $N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v$ movable uniform planar arrays (UPAs), where N_{RF}^h and N_{RF}^v denote the number of UPAs along the horizontal and vertical directions, respectively.

Each movable UPA consists of $N_h N_v$ antennas, where N_h and N_v represent the number of antennas along its horizontal and vertical directions, respectively. The total number of antennas is thus given by $N = N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v N_h N_v$. The antennas within each UPA are first individually connected to a separate PS and then to a dedicated RF chain, and they (together with their connected PSs) can move collectively. To avoid collision, each movable UPA is confined in a separate rectangular frame. For convenience, the UPA located in the m-th row and n-th column is indexed by (m, n), $m = 1, 2, \ldots, N_{\mathrm{RF}}^h$ and $n = 1, 2, \ldots, N_{\mathrm{RF}}^v$. The index of the antenna in the i-th row and j-th column of a UPA is denoted as (i, j), $i = 1, 2, \ldots, N_h$ and $j = 1, 2, \ldots, N_v$. The central position of the (m, n)-th UPA in the local coordinate system is denoted as $\mathbf{c}_{m,n} = (x_{m,n}, z_{m,n})^T \in \mathcal{C}_{m,n}$, where $\mathcal{C}_{m,n}$ (enclosed by the red-dotted line in Fig. 1(a)) represents the movable region. If the central point moves continuously within this region, the UPA is guaranteed to move continuously within the given frame, ensuring that the distance between any two antennas located in different UPAs is always no less than half-wavelength. Additionally, for each UPA, we collect the relative positions of the antennas with respect to the central point in $\boldsymbol{\delta} = [\boldsymbol{\delta}_{1,1}, \boldsymbol{\delta}_{1,2}, \ldots, \boldsymbol{\delta}_{1,N_v}, \boldsymbol{\delta}_{2,1}, \ldots, \boldsymbol{\delta}_{N_h,N_v}] \in \mathbb{R}^{2 \times N_h N_v}$. The position of the (i, j)-th antenna within the (m, n)-th UPA is denoted by $\mathbf{t}_{m,n}^{i,j} = \boldsymbol{\delta}_{i,j} + \mathbf{c}_{m,n}$.

A. Signal Model

We consider far-field, quasi-static, and slow-fading channels. Given that the size of movable region for each UPA is much smaller than the signal propagation distance between the BS and the scatter points and/or receiver, the planar wave model is adopted to construct the field-response array model between the transmit region and the users [9], [10], [11]. In this model, for each channel path component, during the movement of the UPAs, the angle of departure/arrival (AoD/AoA), and amplitude of the complex path coefficient remain unchanged while only the phases vary. Define $\rho_{k,l} = [\sin\theta_{k,l}\cos\phi_{k,l}, \cos\theta_{k,l}]$, where $\theta_{k,l}$ and $\phi_{k,l}$, represent the elevation and azimuth angles of the l-th path as shown in Fig. 1(b), respectively. Let λ denote the carrier wavelength. The transmit field-response vectors (FRVs) between the (i, j)-th antenna within the (m, n)-th UPA and the k-th user are represented as [9], [10]

$$\mathbf{g}_k\left(\mathbf{t}_{m,n}^{i,j}\right) = \left[e^{j\frac{2\pi}{\lambda}\left(\mathbf{t}_{m,n}^{i,j}\right)^T \rho_{k,l}}\right]_{1 \le l \le L_k^t}^T, \tag{1}$$

where L_k^t is the number of channel paths. Define the path-response matrix (PRM) $\Sigma_k \in \mathbb{C}^{L_k^t \times L_k^r}$ as the response between all the channel paths, incorporating large-scale fading effects such as path loss, and small-scale fading effects such as the random phase of each path caused by reflection, diffraction, or scattering. The channel vector between the (m, n)-th UPA and the k-th user is then given by

$$\mathbf{h}_k\left(\mathbf{t}_{m,n}\right) = \mathbf{G}_k\left(\mathbf{t}_{m,n}\right)^H \boldsymbol{\Sigma}_k \mathbf{1}_{L_k^r}, \tag{2}$$

where $\mathbf{G}_k(\mathbf{t}_{m,n}) = [\mathbf{g}_k(\mathbf{t}_{m,n}^{1,1}), \mathbf{g}_k(\mathbf{t}_{m,n}^{1,2}), \ldots, \mathbf{g}_k(\mathbf{t}_{m,n}^{N_h,N_v})]$ is the field-response matrix (FRM) at the (m, n)-th UPA. In this correspondence, the components of field-response information $\{\theta_{k,l}, \phi_{k,l}\}$ can be estimated using classical compressed sensing algorithms, such as orthogonal matching pursuit. Based on the obtained AoA/AoD information, the components of PRM information $\{\Sigma_k\}$ are then estimated using the least square-based algorithm [14]. To characterize the performance limit, we assume that the perfect knowledge of the above field-response channel model is known at the BS for each realization of multi-user channels.

We collect the central positions of all movable UPAs in $\mathbf{c} = [\mathbf{c}_{1,1}, \mathbf{c}_{1,2}, \ldots, \mathbf{c}_{N_{\mathrm{RF}}^h, N_{\mathrm{RF}}^v}]$. The channel vector between the BS and the k-th user is given by

$$\mathbf{h}_k\left(\mathbf{c}\right) = \left[\mathbf{h}_k(\mathbf{t}_{1,1})^T, \mathbf{h}_k(\mathbf{t}_{1,2})^T, \ldots, \mathbf{h}_k(\mathbf{t}_{N_{\mathrm{RF}}^h, N_{\mathrm{RF}}^v})^T\right]^T. \tag{3}$$

Let $\mathbf{s} \in \mathbb{C}^K$ be the independent and identically distributed (i.i.d.) signal vector, with $\mathbb{E}(\mathbf{s}\mathbf{s}^H) = \mathbf{I}_K$. The received signal at the k-th user is then given by

$$\begin{aligned} y_k &= \mathbf{h}_k\left(\mathbf{c}\right)^H \mathbf{W}_{\mathrm{A}} \mathbf{W}_{\mathrm{D}} \mathbf{s} + n_k, \\ &= \underbrace{\mathbf{h}_k\left(\mathbf{c}\right)^H \mathbf{W}_{\mathrm{A}} \mathbf{w}_k s_k}_{\text{Desired signal}} + \underbrace{\mathbf{h}_k\left(\mathbf{c}\right)^H \mathbf{W}_{\mathrm{A}} \sum_{k' \ne k} \mathbf{w}_{k'} s_{k'}}_{\text{Interference}} + n_k, \end{aligned} \tag{4}$$

where $\mathbf{W}_{\mathrm{D}} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K] \in \mathbb{C}^{N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v \times K}$ denotes the digital beamformer, $\mathbf{W}_{\mathrm{A}} \in \mathbb{C}^{N \times N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v}$ denotes the analog beamformer, and $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ is zero-mean additive white Gaussian noise (AWGN) with noise power σ_k^2. Due to the sub-connected struture, we have [3]

$$\mathbf{W}_{\mathrm{A}} = \mathrm{diag}\left(\mathbf{p}_{1,1}, \mathbf{p}_{1,2}, \ldots, \mathbf{p}_{N_{\mathrm{RF}}^h, N_{\mathrm{RF}}^v}\right), \tag{5}$$

where $p_{m,n}^{i,j} = e^{j\psi_{m,n}^{i,j}}$ is the $((i-1)N_h + j)$-th element of $\mathbf{p}_{m,n}$ realized by a unit-modulus PS. Here $\psi_{m,n}^{i,j} \in [0, 2\pi)$ denotes the phase of the PS connected to the (i, j)-th antenna within the (m, n)-th UPA. Based on (4), the achievable rate of the k-th user is

$$R_k = \log_2\left(1 + \frac{\left|\mathbf{h}_k\left(\mathbf{c}\right)^H \mathbf{W}_{\mathrm{A}} \mathbf{w}_k\right|^2}{\sum_{k' \ne k}\left|\mathbf{h}_k\left(\mathbf{c}\right)^H \mathbf{W}_{\mathrm{A}} \mathbf{w}_{k'}\right|^2 + \sigma_k^2}\right). \tag{6}$$

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2025 at 13:05:40 UTC from IEEE Xplore. Restrictions apply.

B. Problem Formulation

We aim to maximize the sum rate of all users jointly optimizing the central positions of UPAs $\mathbf{c}$, the analog beamformer $\mathbf{W}_{\mathrm{A}}$, and the digital beamformer $\mathbf{W}_{\mathrm{D}}$. The optimization problem is formulated as

$$\max_{\mathbf{W}_{\mathrm{D}},\mathbf{W}_{\mathrm{A}},\mathbf{c}} \sum_{k=1}^{K} R_k \tag{7a}$$

$$\text{s.t.} \quad \mathbf{c}_{m,n} \in \mathcal{C}_{m,n}, \forall m,n, \tag{7b}$$

$$||\mathbf{W}_{\mathrm{A}}\mathbf{W}_{\mathrm{D}}||_F^2 \le P_{\max}, \tag{7c}$$

$$\left|p_{m,n}^{i,j}\right| = 1, \forall m,n,i,j, \tag{7d}$$

where $P_{\max}$ denotes the power budget. Due to the non-concave/non-convex objective function/constraints, as well as the coupled variables, (7) is difficult to solve. Specifically, in contrast to the existing FPA-based hybrid beamforming designs [1], [2], [15], the movable positions of the UPAs are implicitly contained in the channel vectors and introduce additional interdependencies among variables in both the numerator and denominator of the achievable rate for each user, thus posing further challenges.

III. Proposed Solution

In this section, we develop a low-complexity algorithm to obtain a suboptimal solution for (7). Specifically, we first invoke the FP framework [16], a iterative optimization framework shared similar inspiration with the well-known weighted minimum mean-square error (WMMSE) framework [17], to recast (7) into a more tractable form. Next, based on AO, we proceed to decompose the FP problem into three subproblems. Then, the penalty method is employed to address the unit modulus constraint. Finally, we propose a gradient decent method to optimize the positions of the UPAs.

A. FP-Based Reformulation

Let $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_K]$ and $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_K]$ be slack variables. To address the fractional forms in the objective function, we employ the FP framework in [16], through which (7) is equivalently reformulated as

$$\max_{\mathbf{W}_{\mathrm{D}},\mathbf{W}_{\mathrm{A}},\mathbf{c},\boldsymbol{\gamma},\boldsymbol{\omega}} \mathcal{L} = \sum_{k=1}^{K} \log(1+\gamma_k) - \gamma_k + (1+\gamma_k)\left(2\mathcal{R}\left\{\omega_k^* a_k\right\} - |\omega_k|^2 b_k\right) \tag{8a}$$

$$\text{s.t.} \gamma_k > 0, \omega_k \in \mathbb{C}, \forall k \in \mathcal{K}, \tag{8b}$$

$$(7\text{b}), (7\text{c}), (7\text{d}), \tag{8c}$$

where, for fixed $\mathbf{W}_{\mathrm{D}}$, $\mathbf{W}_{\mathrm{A}}$, and $\mathbf{c}$, the optimal values of γ_k and ω_k are given by

$$\gamma_k^* = \frac{|a_k|^2}{\sum_{k' \ne k}\left|\mathbf{h}_k(\mathbf{c})^H \mathbf{W}_{\mathrm{A}}\mathbf{w}_{k'}\right|^2 + \sigma_k^2} \tag{9}$$

and

$$\omega_k^* = \frac{a_k}{b_k}, \tag{10}$$

respectively, with $a_k = \mathbf{h}_k(\mathbf{c})^H \mathbf{W}_{\mathrm{A}}\mathbf{w}_k$ and $b_k = \sigma_k^2 + \sum_{k'=1}^{K}|\mathbf{h}_k(\mathbf{c})^H \mathbf{W}_{\mathrm{A}}\mathbf{w}_{k'}|^2$.

To decouple the variables $\{\mathbf{W}_{\mathrm{D}}, \mathbf{W}_{\mathrm{A}}, \mathbf{c}, \boldsymbol{\gamma}, \boldsymbol{\omega}\}$ in (8), we employ the AO framework to update each variable iteratively with other variables being fixed. Since the closed-form expressions of variables $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$ are provided in (9) and (10), respectively, we only need to update the other variables with given $\{\boldsymbol{\gamma}, \boldsymbol{\omega}\}$ in each iteration. Define $\alpha_k = \log(1+\gamma_k) - \gamma_k - (1+\gamma_k)|\omega_k|^2\sigma_k^2$. Since α_k is only relevant to variables $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$, we omit it in the rest of this correspondence.

B. Digital Beamformer Design

Given $\{\mathbf{W}_{\mathrm{A}}, \mathbf{c}, \boldsymbol{\gamma}, \boldsymbol{\omega}\}$, we seek the optimal value of the digital beamformer $\mathbf{W}_{\mathrm{D}}$. We recast the power constraint as [1]

$$||\mathbf{W}_{\mathrm{A}}\mathbf{W}_{\mathrm{D}}||_F^2 = N_v N_h ||\mathbf{W}_{\mathrm{D}}||_F^2 \le P_{\max}. \tag{11}$$

Thus, $\mathbf{W}_{\mathrm{A}}$ and $\mathbf{W}_{\mathrm{D}}$ are decoupled in (7c). The subproblem of digital beamformer design is then formulated as

$$\max_{\mathbf{W}_{\mathrm{D}}} -\sum_{k=1}^{K}\mu_k \sum_{k'=1}^{K}|\boldsymbol{\xi}_k^H \mathbf{w}_{k'}|^2 - 2\mathcal{R}\left\{\boldsymbol{\beta}_k^H \mathbf{w}_k\right\} \tag{12a}$$

$$\text{s.t.} \quad ||\mathbf{W}_{\mathrm{D}}||_F^2 \le \frac{P_{\max}}{N_v N_h}, \tag{12b}$$

where $\boldsymbol{\beta}_k^H = (1+\gamma_k)\omega_k^* \boldsymbol{\xi}_k^H$, $\boldsymbol{\xi}_k^H = \mathbf{h}_k(\mathbf{c})^H \mathbf{W}_{\mathrm{A}}$, and $\mu_k = (1+\gamma_k)|\omega_k^2|$. Let $\boldsymbol{\Xi} = \sum_{k=1}^{K}\mu_k \boldsymbol{\xi}_k \boldsymbol{\xi}_k^H$. According to the Lagrangian multipliers, the convex quadratic optimization problem admits the closed-form solution given by

$$\mathbf{w}_k^*(\lambda) = \left(\boldsymbol{\Xi} + \lambda \mathbf{I}_{N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v}\right)^{-1}\boldsymbol{\beta}_k, \tag{13}$$

where λ is a Lagrange multiplier such that $\lambda(||\mathbf{W}_{\mathrm{D}}||_F^2 - \frac{P_{max}}{N_v N_h}) = 0$. That is, if $||\mathbf{W}_{\mathrm{D}}^*||_F^2 = \sum_{k=1}^{K}||\mathbf{w}_k^*(0)||^2 \le \frac{P_{max}}{N_v N_h}$, $\lambda = 0$. Otherwise, $\lambda > 0$ can be found by a bisection search.

C. Analog Beamformer Design

With given $\{\mathbf{W}_{\mathrm{D}}, \mathbf{c}, \boldsymbol{\gamma}, \boldsymbol{\omega}\}$, we set out to optimize $\mathbf{W}_{\mathrm{A}}$. Define $\tilde{\mathbf{p}} = [\mathbf{p}_{1,1}^T, \mathbf{p}_{1,2}^T \cdots, \mathbf{p}_{N_{\mathrm{RF}}^h, N_{\mathrm{RF}}^v}^T]^T$, $\tilde{\mathbf{h}}_{k,k'}^H = \mathbf{h}_k(\mathbf{c})^H(\mathrm{diag}(\mathbf{w}_{k'}) \otimes \mathbf{I}_{N_h N_v})$, and $\tilde{\boldsymbol{\beta}}_k^H = (1+\gamma_k)\omega_k^* \tilde{\mathbf{h}}_{k,k}^H$. Based on the penalty method, the subproblem of analog beamformer design is formulated as

$$\max_{\tilde{\mathbf{p}},\boldsymbol{\phi}} \sum_{k=1}^{K}\left(-\mu_k \sum_{k'=1}^{K}\left|\tilde{\mathbf{h}}_{k,k'}^H \boldsymbol{\phi}\right|^2 + 2\mathcal{R}\left\{\tilde{\boldsymbol{\beta}}_k^H \boldsymbol{\phi}\right\}\right) - \eta||\boldsymbol{\phi} - \tilde{\mathbf{p}}||_2^2 \tag{14a}$$

$$\text{s.t.} \quad |[\tilde{\mathbf{p}}]_s| = 1, \forall s = 1, 2, \ldots, N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v N_h N_v, \tag{14b}$$

where $\boldsymbol{\phi}$ is the introduced continuous variable and $\eta > 0$ is the penalty parameter. Problem (14) can be solved via updating $\tilde{\mathbf{p}}$ and $\boldsymbol{\phi}$ iteratively. With given $\tilde{\mathbf{p}}$, the optimal solution of $\boldsymbol{\phi}$ is

$$\boldsymbol{\phi}^* = \left[\tilde{\boldsymbol{\Xi}}\right]^{-1}\tilde{\boldsymbol{\beta}}, \tag{15}$$

where $\tilde{\boldsymbol{\Xi}} = \sum_{k=1}^{K}\sum_{k'=1}^{K}(\mu_k \tilde{\mathbf{h}}_{k,k'}\tilde{\mathbf{h}}_{k,k'}^H) + \eta \mathbf{I}_N$ and $\tilde{\boldsymbol{\beta}} = \sum_{k=1}^{K}\tilde{\boldsymbol{\beta}}_k + \eta\tilde{\mathbf{p}}$. Given $\boldsymbol{\phi}$, the optimization of $\tilde{\mathbf{p}}$ can be extracted from (14) as

$$\min_{\tilde{\mathbf{p}}} ||\boldsymbol{\phi} - \tilde{\mathbf{p}}||_2^2 \tag{16a}$$

$$\text{s.t.} \quad |[\tilde{\mathbf{p}}]_s| = 1, \forall s = 1, 2, \ldots, N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v N_h N_v. \tag{16b}$$

The optimal phase of $\tilde{\mathbf{p}}$ is given by

$$\arg\{\tilde{\mathbf{p}}\} = \arg\{\boldsymbol{\phi}\}. \tag{17}$$

D. MA Position Design

Given $\{\mathbf{W}_{\mathrm{D}}, \mathbf{W}_{\mathrm{A}}, \boldsymbol{\gamma}, \boldsymbol{\omega}\}$, the subproblem of MA position design can be decomposed into iteratively optimizing $\mathbf{c}_{m,n}$ with the other

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2025 at 13:05:40 UTC from IEEE Xplore. Restrictions apply.

central points of UPAs fixed. The optimization problem is formulated as

$$\max_{\mathbf{c}_{m,n}\in\mathcal{C}_{m,n}} \mathcal{L}(\mathbf{c}_{m,n}) = \sum_{k=1}^{K} 2\mathcal{R}\left\{\mathbf{h}_k\left(\mathbf{t}_{m,n}\right)^H \bar{\boldsymbol{\beta}}_{k,m,n}\right\} - \mu_k \sum_{k'=1}^{K} \Bigg| \underbrace{\sum_{m=1}^{N_{\mathrm{RF}}^h}\sum_{n=1}^{N_{\mathrm{RF}}^v} \mathbf{h}_k\left(\mathbf{t}_{m,n}\right)^H \bar{\boldsymbol{\xi}}_{k',m,n}}_{\bar{h}_{k,k'}\left(\mathbf{t}_{m,n}\right)} \Bigg|^2, \tag{18}$$

where $\bar{\boldsymbol{\beta}}_{k,m,n} = (1+\gamma_k)\omega_k^* \bar{\boldsymbol{\xi}}_{k,m,n}$, $\bar{\boldsymbol{\xi}}_{k,m,n} = w_{k,m,n}\mathbf{p}_{m,n}$, with $w_{k,m,n}$ denoting the $((m-1)N_{\mathrm{RF}}^h + n)$-th element of $\mathbf{w}_k$. Due to the complicated expression of $\mathcal{L}(\mathbf{c}_{m,n})$, we propose a gradient descent method to optimize $\mathbf{c}_{m,n}$. In each iteration, the gradient can be calculated as

$$\nabla_{\mathbf{c}_{m,n}} \mathcal{L}\left(\mathbf{c}_{m,n}\right) = 2\sum_{k=1}^{K} \mathcal{R}\left\{ \frac{\partial \mathbf{h}_k\left(\mathbf{t}_{m,n}\right)}{\partial \mathbf{c}_{m,n}} \bar{\boldsymbol{\beta}}_{k,m,n}^* - \mu_k \sum_{k'=1}^{K} \frac{\partial \mathbf{h}_k\left(\mathbf{t}_{m,n}\right)}{\partial \mathbf{c}_{m,n}} \bar{\boldsymbol{\xi}}_{k',m,n}^* \bar{h}_{k,k'}\left(\mathbf{t}_{m,n}\right) \right\}, \tag{19}$$

where the expression of partial derivation is given by

$$\left[\frac{\partial \mathbf{h}_k\left(\mathbf{t}_{m,n}\right)}{\partial \mathbf{c}_{m,n}}\right]_{:,(i-1)N_v+j} = \sum_{l=1}^{L_k} -j\frac{2\pi}{\lambda}\sigma_{k,l} e^{-j\frac{2\pi}{\lambda}\left(\mathbf{t}_{m,n}^{i,j}\right)^T \boldsymbol{\rho}_{k,l}} \boldsymbol{\rho}_{k,l}, \forall i,j. \tag{20}$$

with $\sigma_{k,l} = [\boldsymbol{\Sigma}_k \mathbf{1}_{L_k^r}]_l$. Then, the (m,n)-th central point in the t-th iteration is updated by moving along the gradient direction and checking whether the new point is still located within the feasible region, i.e.,

$$\mathbf{c}_{m,n}^{(t+1)} = \begin{cases} \tilde{\mathbf{c}}_{m,n}, & \text{if } \tilde{\mathbf{c}}_{m,n} \in \mathcal{C}_{m,n} \text{ and } \mathcal{L}(\tilde{\mathbf{c}}_{m,n}) \geq \mathcal{L}(\mathbf{c}_{m,n}^{(t)}), \\ \mathbf{c}_{m,n}^{(t)}, & \text{otherwise}, \end{cases} \tag{21}$$

where $\tilde{\mathbf{c}}_{m,n} = \mathbf{c}_{m,n}^{(t)} + \kappa \nabla_{\mathbf{c}_{m,n}} \mathcal{L}(\mathbf{c}_{m,n}^{(t)})$ with κ being the step size. Note that the performance of gradient descent heavily depends on the step size. Thus, for each iteration, we initialize κ with a large positive number and update $\kappa \leftarrow \frac{\kappa}{2}$ until finding a feasible point such that $\mathcal{L}(\tilde{\mathbf{c}}_{m,n}) \geq \mathcal{L}(\mathbf{c}_{m,n}^{(t)})$ or achieve the stopping criterion $\kappa < \kappa_{min}$. It can be observed that the sequence $\{\mathcal{L}(\mathbf{c}^{(t)})\}$ keeps non-decreasing based on update guidelines (21), leading to the convergence of the sequence. Therefore, the gradient descent method is applicable to any objective function with a specific expression.

Algorithm 1 summarizes the workflow for solving problem (7). We next provide a brief proof for the convergence of the proposed algorithm. For analysis convenience, let $L(\mathbf{W}_{\mathrm{D}}^t, \mathbf{W}_{\mathrm{A}}^t, \mathbf{c}^t)$ denote the value of the penalty objective function (14a) over t-th iteration. Since the penalty term $\eta||\boldsymbol{\phi} - \tilde{\mathbf{p}}||_2^2$ is irreverent to digital beamformer and central positions of UPAs, Algorithm 1 generates a non-decreasing sequence $\{L(\mathbf{W}_{\mathrm{D}}^t, \mathbf{W}_{\mathrm{A}}^t, \mathbf{c}^t)\}$, which has a finite upper bound, thus ensuring the convergence of the penalty objective value sequence [18]. Moreover, the penalty factor can be set such that the penalty term tends to zero, causing the penalized objective function (14a) to reduce to the original problem (8a), thus guaranteeing the convergence of Algorithm 1. The computational complexities of solving the variables $\{\mathbf{W}_{\mathrm{D}}, \mathbf{W}_{\mathrm{A}}, \mathbf{c}, \boldsymbol{\gamma}, \boldsymbol{\omega}\}$ are scaled with $\mathcal{O}((N_{\mathrm{RF}}^h N_{\mathrm{RF}}^v)^3)$, $\mathcal{O}(N^3)$, $\mathcal{O}(KNN_{\mathrm{RF}}^h N_{\mathrm{RF}}^v \sum_{k=1}^K L_k)$, $\mathcal{O}(KN)$, and $\mathcal{O}(KN)$, respectively. Furthermore, we denote the number of iterations needed for the convergence of the AO algorithm as I_{AO}. Considering the dominant computational steps, the maximum computational complexity of Algorithm 1 can be expressed as $\mathcal{O}(I_{AO}(N^3 + KNN_{\mathrm{RF}}^h N_{\mathrm{RF}}^v \sum_{k=1}^K L_k))$.

Algorithm 1: Proposed Algorithm for Solving Problem (7).

Input: $\mathbf{W}_{\mathrm{D}}^{(0)}, \mathbf{W}_{\mathrm{A}}^{(0)}, \mathbf{c}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\omega}^{(0)}, \{R_k^{(0)}\}, \tilde{\kappa}, \kappa_{min}, \varepsilon, I$.
Output: $\mathbf{W}_{\mathrm{D}}^*, \mathbf{W}_{\mathrm{A}}^*, \mathbf{c}^*$.

```
Set iteration index t = 1.
repeat
  Update γ^(t) and ω^(t) via (9) and (10), respectively.
  Obtain W_D^(t) via (13).
  Obtain φ^(t) via (15).
  Obtain W_A^(t) via (5) and (17).
  for all m = 1 : N_RF^h do
    for all n = 1 : N_RF^v do
      Calculate the gradient ∇_{c_{m,n}} L(c_{m,n}^(t)) via (19).
      Initialize the step size κ = κ̃.
      repeat
        Compute c̃_{m,n} = c_{m,n}^(t) + κ ∇_{c_{m,n}} L(c_{m,n}^(t)).
        Shrink the step size κ ← κ/2.
        Update c_{m,n}^(t+1) according to (21).
      until c̃_{m,n} ∈ C_{m,n} and L(c̃_{m,n}) ≥ L(c_{m,n}^(t)) or
      κ < κ_min.
    end for
  end for
  Calculate the sum rate Σ_{k=1}^K R_k^(t+1) according to (6).
  Update t = t + 1.
until |Σ_{k=1}^K R_k^(t+1) − R_k^(t)| < ε or the maximum iteration
number I is reached.
```

IV. Numerical Results

In this section, numerical results are provided to evaluate the performance of our proposed scheme. Unless stated otherwise, we consider a scenario where a 16-antenna BS operating at $f_c = 30$ GHz serves $K = 16$ users. Besides, $N_{\mathrm{RF}}^h = N_{\mathrm{RF}}^v = 4$, $N_h = N_v = 2$. The central point of each UPA remains in the x-z plane, where the inter-element spacing is $\frac{\lambda}{2}$ to avoid antenna coupling and $\lambda = 0.01$ m represents the wavelength. The distance between the k -th user and the BS is uniformly distributed from 20 to 100m, i.e., $d_k \sim \mathcal{U}[20, 100]$. For each user, we assume an equal number of transmit and receive channel paths $L_k^t = L_k^r = L = 6$. The PRM is defined as a diagonal matrix $\Sigma_k = \mathrm{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_L\}$, where $\sigma_{k,l} \sim \mathcal{CN}(0, (\rho_0 d_k^{-\alpha})^2/L)$, $\rho_0 = -40$ dB is the channel power gain at the reference distance 1m, and $\alpha = 2.8$ denotes the corresponding path loss exponent. The elevation and azimuth angles are assumed to follow the joint probability density function $f(\theta_{k,l}^{\chi}, \phi_{k,l}^{\chi}) = \frac{\cos\theta_{k,l}^{\chi}}{2\pi}$ [9], $\chi \in \{t, r\}$. For each UPA, the size of movable frame is $D \times D$. In addition, $P_{max} = 10$ dBm, $\sigma_k^2 = -80$ dBm, $D = 2\lambda$, $\tilde{\kappa} = 10, \varepsilon = 10^{-3}, \eta = 10, \kappa_{min} = 10^{-3}$, and $I = 200$.

To verify the effectiveness of the proposed scheme with Algorithm 1, we consider two baselines for comparison: 1) Fixed array, sub-connected [3]. 2) Fixed array, fully-connected [3]. Besides, by finding the optimal positions of all UPAs under the sub-connected structure within the whole region based on exhaustive search, the upper bound on the sum rate is also obtained for comparison.

Fig. 2 depicts the sum-rate versus the power budget for different schemes. It is observed that with the same power budget, our proposed algorithm achieves a higher sum-rate compared to its FPA counterpart. This suggests that the performance gain from MAs can compensate the performance loss of applying sub-connected structure. Moreover, when transmit power is sufficiently low, the MA-aided sub-connected structure can even outperform the fully-connected FPAs. This indicates that with lower inter-user interference, the MAs' ability to enhance received signal power gain significantly contributes to performance improvement.

Fig. 3 demonstrates the sum-rate versus D/λ. Note that the sum-rate increases as the normalized region size increases until reaching a

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2025 at 13:05:40 UTC from IEEE Xplore. Restrictions apply.

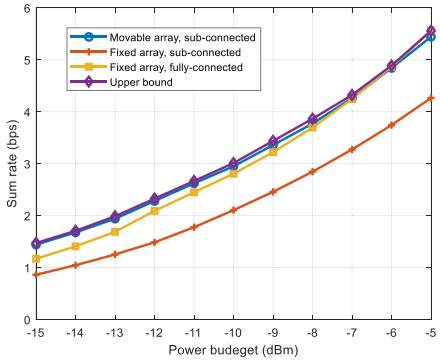

Fig. 2. Sum-rate versus P_{max} with $D = 2\lambda$.

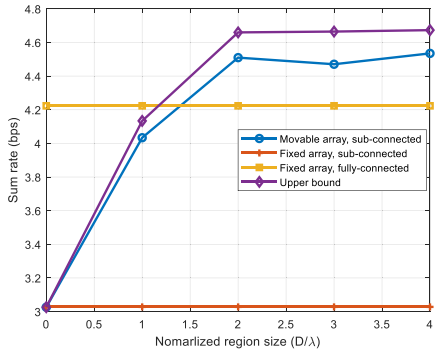

Fig. 3. Sum-rate versus the normalized region size with $P_{max} = -10$ dBm.

constant value. The reason for it is that increasing region size provides more DoFs for the MAs to exploit the channel spatial variation. However, for a sufficiently large region size, the channel gain may exhibit periodic character in the spatial domain, which indicates a strong spatial correlation [9]. For this case, the effective channel gain for each user cannot be further increased by antenna position optimization, while the multiuser interference has already approached the lower bound. Therefore, further increasing the region size offers negligible improvement, which indicates that a moderate size of the spatial region is sufficient for deploying MAs Moreover, Fig. 3 also shows that with sufficiently large region size, our proposed scheme is superior to all baselines. However, the performance gap from the upper bound is still substantial, due to that the gradient method is prone to converging to local optimum.

V. Conclusion

This correspondence investigated MA-aided multi-user hybrid beamforming under the sub-connected structure. We studied the sum rate maximization problem by jointly designing the digital beamformer, analog beamformer, and movable UPAs' positions. To solve the non-convex optimization problem, we employed FP to transform it into a more tractable form and then developed an AO-based algorithm by applying the techniques of Lagrange multiplier, penalty method and gradient descent. Numerical results demonstrated the superiority of the proposed MA-aided sub-connected structure compared to the FPA-based system. Moreover, under certain conditions, the proposed scheme with sub-connected MA arrays even outperforms the fully-connected FPA array.

References

[1] 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.

[2] S. S. Ioushua and Y. C. Eldar, "A family of hybrid analog–digital beamforming methods for massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3243–3257, Jun. 2019.

[3] J. Du, W. Xu, H. Shen, X. Dong, and C. Zhao, "Hybrid precoding architecture for massive multiuser MIMO with dissipation: Sub-connected or fully connected structures?," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5465–5479, Aug. 2018.

[4] O. E. Ayach, R. W. Heath, S. Rajagopal, and Z. Pi, "Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays," in *Proc. IEEE Glob. Commun. Conf.*, Dec. 2013, pp. 3476–3480.

[5] L. Zhu and K. K. Wong, "Historical review of fluid antenna and movable antenna," 2024, *arXiv:2401.02362*.

[6] K. -K. Wong, W. K. New, X. Hao, K. -F. Tong, and C. -B. Chae, "Fluid antenna system—Part I: Preliminaries," *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 1919–1923, Aug. 2023.

[7] K. -K. Wong, K. -F. Tong, and C. -B. Chae, "Fluid antenna system—Part II: Research opportunities," *IEEE Commun. Lett.*, vol. 27, no. 8, pp. 1924–1928, Aug. 2023.

[8] L. Zhu, W. Ma, and R. Zhang, "Movable antennas for wireless communication: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 62, no. 6, pp. 114–120, Jun. 2024.

[9] L. Zhu, W. Ma, and R. Zhang, "Modeling and performance analysis for movable antenna enabled wireless communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 6234–6250, Jun. 2024.

[10] W. Ma, L. Zhu, and R. Zhang, "MIMO capacity characterization for movable antenna systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3392–3407, Apr. 2024.

[11] L. Zhu, W. Ma, B. Ning, and R. Zhang, "Movable-antenna enhanced multiuser communication via antenna position optimization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7214–7229, Jul. 2024.

[12] Y. Wu, D. Xu, D. W. K. Ng, W. Gerstacker, and R. Schober, "Movable antenna-enhanced multiuser communication: Optimal discrete antenna positioning and beamforming," in *Proc. IEEE Glob. Commun. Conf.*, Dec. 2023, pp. 7508–7513.

[13] G. Hu, Q. Wu, K. Xu, J. Si, and N. Al-Dhahir, "Secure wireless communication via movable-antenna array," *IEEE Signal Process. Lett.*, vol. 31, pp. 516–520, 2024.

[14] W. Ma, L. Zhu, and R. Zhang, "Compressed sensing based channel estimation for movable antenna communications," *IEEE Commun. Lett.*, vol. 27, no. 10, pp. 2747–2751, Oct. 2023.

[15] W. Zhu, H. D. Tuan, E. Dutkiewicz, H. V. Poor, and L. Hanzo, "Max-min rate optimization of low-complexity hybrid multi-user beamforming maintaining rate-fairness," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 5648–5662, Jun. 2024.

[16] K. Shen and W. Yu, "Fractional programming for communication systems-Part II: Uplink scheduling via matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, May 2018.

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

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

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2025 at 13:05:40 UTC from IEEE Xplore. Restrictions apply.