# Dynamic Metasurface Antennas-Based Beamforming for Holographic MIMO Systems

Yuanli Ma, Bin Yan, Zheng Wang, Yongming Huang
*School of Information Science and Engineering*
*Southeast University, Nanjing, China*
Email: myl@seu.edu.cn, bin_yan@seu.edu.cn,
wznuaa@gmail.com, huangym@seu.edu.cn

***Abstract*—In this paper, to maximize sum rate for downlink multi-user holographic MIMO (HMIMO) systems, we investigate the designs of dynamic metasurface antenna (DMA) weights matrix and precoding matrix based on strategies of alternating optimization (AO) and decoupling optimization (DO), respectively. Specifically, in AO strategy, these two matrices can be designed in an alternative way. Based on it, given the fixed precoding matrix, the problem of DMA weights matrix design is approximated by a novel one, which can be efficiently solved but with better array gain. Meanwhile, for further complexity reduction, the designs of these two matrices can be decoupled in DO strategy, so that the DMA weights matrix is optimized independently from the precoding matrix. Finally, the introduced array gains and complexity reductions of the proposed algorithms based on AO and DO are validated by the simulation results.**

***Index Terms*—Dynamic metasurface antenna, holographic MIMO beamforming, alternating optimization, decoupling optimization.**

## I. Introduction

With the increasing demand for the sixth-generation (6G) communications [1], the concept of holographic MIMO (HMIMO) has been put forward for large antenna massive MIMO systems [2] with low power consumption and hardware cost [3]. HMIMO beamforming allows nearly spatially continuous aperture, such that the transceiver can accommodate huge number of elements deployed in a practically viable way [4]. As one of the typical HMIMO surfaces, dynamic metasurface antenna (DMA) is tunable with reduced cost and power consumption compared to conventional arrays [5]. Specifically, the transceiver equipped with DMA requires less radio frequency (RF) chains than the conventional one, which greatly facilitates the practical implementation [6].

Technically, DMA-based HMIMO beamforming can be realized with the proper choice of DMA weights [3]. To this end, the optimizations of DMA weights are studied in [7] and [8] to maximize the achievable sum rate in both uplink and downlink multi-user MIMO systems. Besides, based on Riemannian manifold optimization (RMO), the DMA weights and precoding matrix are alternatively optimized for near-field downlink multi-user MIMO systems [9], [10]. However, the designs for DMA weights in those works all relax the Lorentzian constraint to simplify the optimization problem, which results in inevitable considerable performance loss [11]. Recently, a

Corresponding author: Zheng Wang (email: wznuaa@gmail.com)

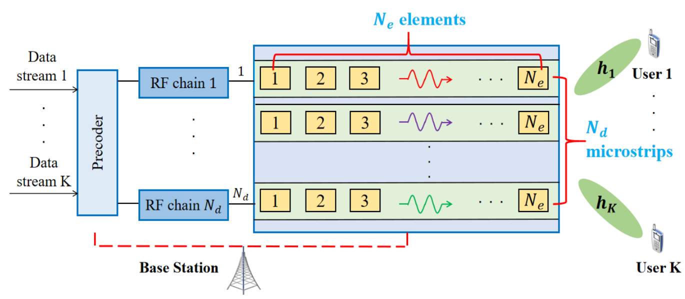


Fig. 1. System diagram of DMA-based multi-user downlink.

novel strategy that divides the Lorentzian constraint into the complex constant term and the exponential term containing the optimizable phase is proposed in [11], followed by an alternating optimization (AO) strategy to achieve the weighted sum rate (WSR) maximization based on RMO. To reduce the computational complexity arising from RMO, the work in [12] further optimizes the DMA weights by the gradient descent projection (GDP) instead of RMO. Nevertheless, the algorithms in [11], [12] still need to compute the Euclidean gradient of the sum rate, which is computationally expensive for HMIMO systems.

In this paper, the DMA weights matrix and precoding matrix designs for downlink mutli-user HMIMO systems are studied to maximize the sum rate. Based on AO strategy, these two matrices are optimized alternatively. In particular, given the fixed precoding matrix, an approximated formula of DMA weights matrix design is proposed, which improves the sum rate by obtaining array gain with less complexity cost. For further complexity reduction, the design of these two matrices based on the decoupling optimization (DO) strategy is also proposed, in which the DMA weights matrix can be optimized independently from the precoding matrix. Finally, simulation results illustrate the advantages of the proposed algorithms based on AO and DO strategy.

## II. System Model

Consider a downlink system in which the base station (BS) is equipped with $N$ DMA elements to serve $K$ single-antenna users. As shown in Fig. 1, the DMA array is composed of $N_d$ microstrip lines, and each microstrip holds $N_e$ elements, so as to $N = N_d \times N_e$ elements in total. The BS transmits a precoded signal $\mathbf{x} = \sum_{k=1}^{K} \mathbf{f}_k d_k$ to the DMA array, where

$\mathbf{f}_k \in \mathbb{C}^{N_d \times 1}$ is the precoding vector for information symbol $d_k$ intended for the $k$-th user. Then the output of the DMA array can be represented as:

$$\overline{\mathbf{x}} = \mathbf{G}\mathbf{Q}\mathbf{x}. \tag{1}$$

Here, $\mathbf{G} \in \mathbb{C}^{N \times N}$ is a diagonal matrix with elements $G_{(i-1)Ne+l,(i-1)Ne+l} = e^{-\rho_{i,l}(\alpha_i + j\beta_i)}$ for $i \in \{1, \cdots, N_d\}, l \in \{1, \cdots, N_e\}$. $\alpha_i$ and $\beta_i$ denote the attenuation coefficient and wavenumber of microstrip $i$, $\rho_{i,l}$ is the location of the $l$-th element in the $i$-th microstrip. Meanwhile, the block-diagonal matrix $\mathbf{Q} \in \mathbb{C}^{N \times N_d}$ in (1) contains the DMA elements with tunable weights, i.e.,

$$Q_{(i-1)N_e+l,q} \in \Phi = \begin{cases} \frac{j+e^{j\theta_{(i-1)N_e+l,q}}}{2}, & \text{if } i = q, \\ 0, & \text{if } i \neq q \end{cases} \tag{2}$$

with $\theta \in [0, 2\pi)$, which follow the Lorentzian constraint [11].

As for the downlink multi-user model, the signal received by user $k$ is given by

$$y_k = \mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{x} + z, \tag{3}$$

where $\mathbf{h}_k \in \mathbb{C}^{N \times 1}$ is the channel between the BS and the user $k$. Then, the WSR for the system is formulated as

$$R(\mathbf{f}_k, \mathbf{Q}) = \sum_{k=1}^{K} \omega_k \log_2 (1 + \mathrm{SINR}_k), \tag{4}$$

where

$$\mathrm{SINR}_k = \frac{|\mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_k|^2}{\sum_{i \neq k}^{K} |\mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_i|^2 + \sigma^2} \tag{5}$$

denotes the signal-to-interference-plus-noise ratio (SINR), $0 \leq \omega_k \leq 1$ indicates the priority of user $k$, and $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_K] \in \mathbb{C}^{N_d \times K}$ is the precoding matrix. According to the DMA weights matrix $\mathbf{Q}$ and the precoding matrix $\mathbf{F}$, the optimization problem of maximizing the sum rate $R$ in (4) becomes:

$$\max_{\mathbf{F},\mathbf{Q}} \quad R(\mathbf{F}, \mathbf{Q}) \tag{6a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \|\mathbf{G}\mathbf{Q}\mathbf{f}_k\|^2 \leq P, \tag{6b}$$

$$Q_{(i-1)N_e+l,i} \in \Phi, \quad \forall i, \forall l. \tag{6c}$$

To solve it, the algorithms in [11], [12] alternatively optimize $\mathbf{Q}$ and $\mathbf{F}$, which leads to the alternating optimization (AO) strategy. On one hand, when $\mathbf{Q}$ is fixed, $\mathbf{F}$ can be designed by classic weighted MMSE (WMMSE) algorithm. As shown in [11], the WMMSE algorithm iteratively updates the percoder matrix as follows:

$$\eta_k = \frac{\mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_k}{\sum_{i=1}^{K} |\mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_i|^2 + \sigma^2}, \tag{7a}$$

$$\psi_k = \frac{1}{1 - \eta_k^* \mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_k}, \tag{7b}$$

$$\mathbf{f}_k = \omega_k \eta_k \psi_k \left( \mu \mathbf{I}_{N_d} + \sum_{i=1}^{K} \omega_i \psi_i |\eta_i|^2 \mathbf{Q}^H \mathbf{G}^H \overline{\mathbf{h}}_i \mathbf{h}_i^T \mathbf{G}\mathbf{Q} \right)^{-1} \mathbf{Q}^H \mathbf{G}^H \overline{\mathbf{h}}_k. \tag{7c}$$

Here $\overline{\mathbf{h}}$ is conjugate vector of $\mathbf{h}$, and $\mu \geq 0$ is *Lagrangian* variable for the power constraint in (6a). On the other hand, when $\mathbf{F}$ is fixed, the problem to optimize $\mathbf{Q}$ becomes

$$\max_{\mathbf{Q}} \quad \sum_{k=1}^{K} \frac{1}{K} \log_2 \left( 1 + \frac{|\mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_k|^2}{\sum_{i \neq k}^{K} |\mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_i|^2 + \sigma^2} \right) \quad \text{s.t.} \quad (6c). \tag{8}$$

To solve (8), by dividing the *Lorentzian* constraint of $Q$ into complex constant term and exponential term, the algorithms in [11], [12] transform the *Lorentzian* constraint into the unit-modulus constraint, followed by RMO or GDP to get the final results of the DMA weights matrix $\mathbf{Q}$. However, computing the gradient of (8) costs $\mathcal{O}(NK^2)$ computational complexity, which results in high complexity burden.

## III. DMA Weights Matrix and Precoding Matrix Alternating Optimization Strategy

During the alternating optimization between DMA weights matrix $\mathbf{Q}$ and precoding matrix $\mathbf{F}$, to reduce the complexity cost in seeking for $\mathbf{Q}$, we attempt to get rid of the inter-user interference in DMA weights matrix design, then eliminate it with the WMMSE algorithm in precoding matrix design. In particular, assuming that each user has equal priority, the problem of DMA weights matrix design in (8) can be simplified as

$$\max_{\mathbf{Q}} \quad \sum_{k=1}^{K} \frac{1}{K} \log_2 \left( 1 + \frac{|\mathbf{h}_k^T \mathbf{G}\mathbf{Q}\mathbf{f}_k|^2}{\sigma^2} \right) \quad \text{s.t.} \quad (6c). \tag{9}$$

Then, by dividing the *Lorentzian* constraint of $Q$ as $Q_n = \frac{j+e^{j\theta_n}}{2}$, the *Lorentzian* constraint can be changed into the unit-modulus constraint. To be more specific, by introducing auxiliary vectors $\mathbf{s} \in \mathbb{C}^{1 \times N}$ and $\mathbf{b}_k \in \mathbb{C}^{N \times 1}, k \in \{1, \cdots, K\}$, the problem in (9) can be formulated in an equivalent way:

$$\max_{\mathbf{s} \in \mathcal{C}} \quad \sum_{k=1}^{K} \frac{1}{K} \log_2 \left( 1 + \frac{|a_k + \mathbf{s}\mathbf{b}_k|^2}{\sigma^2} \right) \tag{10}$$

with

$$a_k = \frac{j}{2} \sum_{i=1}^{N_d} \sum_{l=1}^{N_e} h_{(i-1)N_e+l,k} g_{i,l} f_{i,k}. \tag{11}$$

Here, the $n$-th element of $\mathbf{s}$ is $e^{j\theta_n} = 2Q_n - j$ and the $[(i-1)N_e+l]$-th element of $\mathbf{b}_k$ is $h_{(i-1)N_e+l,k} g_{i,l} f_{i,k}, \forall i, \forall l$. $g_{i,l}$ is a convenient expression of $G_{(i-1)Ne+l,(i-1)Ne+l}$, $Q_n$ denotes the only non-zero element of the $n$-th row of $\mathbf{Q}$, $h_{k,i}$ and $f_{k,i}$ denote the $i$-th element of $\mathbf{h}_k$ and $\mathbf{f}_k$ respectively. Besides, we denote the unit-modulus constraint on $\mathbf{s}$ as $\mathbf{s} \in \mathcal{C} = \{\mathbf{s} | s_n = e^{j\theta_n}, n \in \{1, \cdots, N\}\}$.

In order to maximize the sum rate shown in (10), one can design $\mathbf{s}$ under unit-modulus constraint based on the equal gain transmission (EGT) method [13], which can improve sum rate by harvesting considerable array gain [14]. In particular, from (10), we present that for a specific index $k$, the value $|a_k + \mathbf{s}\mathbf{b}_k|^2$ in (10) can be maximized if and only if each term of the following summation keeps the same phase with that in $a_k$:

$$\mathbf{s}\mathbf{b}_k = \sum_{i=1}^{N_d}\sum_{l=1}^{N_e} h_{(i-1)N_e+l,k}\, g_{i,l} f_{i,k} \frac{e^{j\theta_{(i-1)N_e+l}}}{2}. \tag{12}$$

Typically, for $\forall i, \forall l$,

$$\angle\left(h_{(i-1)N_e+l,k}\, g_{i,l} f_{i,k} \frac{e^{j\theta_{(i-1)N_e+l}}}{2}\right) = \phi_k, \tag{13}$$

where $\angle$ indicates the phase and $\phi_k = \angle(a_k)$. To be more specific, $\mathbf{s}\mathbf{b}_k$ aims to be

$$t_k = \sum_{i=1}^{N_d}\sum_{l=1}^{N_e} \frac{1}{2}\left|h_{(i-1)N_e+l,k}\, g_{i,l} f_{i,k}\right| e^{\phi_k}. \tag{14}$$

However, due to the fact that $\theta_n$ can not be arranged to satisfy (13) for all $k \in \{1, \cdots, K\}$, only few items of (13) can be fulfilled at the same time. Motivated by this, to minimize the difference between $\mathbf{s}\mathbf{b}_k$ and $t_k$ for $\forall k$, we approximate the problem (10) by the following formula:

$$\min_{\mathbf{s}\in\mathcal{C}} \quad \sum_{k=1}^{K} \frac{1}{K}\log_2\left(1 + \frac{|t_k - \mathbf{s}\mathbf{b}_k|^2}{\sigma^2}\right). \tag{15}$$

After that, the computationally efficient GDP can be applied to solve (15) as follows:

$$\begin{aligned} &\nabla f(\mathbf{s}) = -\frac{2}{\sigma^2 K \ln 2}\sum_{k=1}^{K}\frac{t_k - \mathbf{s}\mathbf{b}_k}{1+\frac{1}{\sigma^2}|t_k - \mathbf{s}\mathbf{b}_k|^2}\mathbf{b}_k^H, \\ &\mathbf{x} = \mathbf{s} - \mu\nabla f(\mathbf{s}), \\ &\mathbf{s} = e^{j\angle \mathbf{x}}. \end{aligned} \tag{16}$$

Here, $\mu$ is the step size given by *Armijo rule* [12]. Because the objective function in (15) has *Lipschitz* continuous gradient, the convergence of GDP method can be guaranteed [15]. By solving (15), considerable array gain of DMA can be harvested to maximize the sum rate in (10).

When $\mathbf{Q}$ is fixed, $\mathbf{F}$ can be updated by WMMSE as shown in (7a), (7b) and (7c) until convergence. Then the DMA weights matrix $\mathbf{Q}$ and the precoding matrix $\mathbf{F}$ are alternatively optimized until the objective function in (6a) converges. To summarize, the proposed algorithm is shown in *Algorithm 1*. Note that, to speed up the convergence, a better initial value $s_{(i-1)Ne+l} = g_{i,l}, \forall i, \forall l$ is presented, instead of random initialization in [11], [12].

Now, we analyze computational complexity of the proposed algorithm based on AO. Respectively, we denote $R_W$, $R_G$ and $R_A$ as the WMMSE, GDP and outer AO iteration number. On one hand, when optimizing $\mathbf{F}$, the proposed algorithm based on AO and schemes in [11], [12] all adopt the WMMSE algorithm, in which the complexity is given by $\mathcal{O}(R_W N_d^3 K)$.

On the other hand, when optimizing $\mathbf{Q}$, the complexity of GDP method is given by $\mathcal{O}(R_G N K)$. Compared to the algorithms in [11], [12], the complexity on the DMA weights matrix design is reduced from $\mathcal{O}(NK^2)$ to $\mathcal{O}(NK)$, which reveals that the proposed algorithm has reduced complexity. Due to DMA weights matrix and precoding matrix are alternatively optimized, the total complexity of proposed algorithm based on AO is given by $\mathcal{O}\left(R_A\left(R_W N_d^3 K + R_G N K\right)\right)$.

**Algorithm 1** Proposed algorithm based on AO

**Input:** $\mathbf{G}, \mathbf{h}_k, \forall k \in [1, \cdots, K]$
1: Initialize $\mathbf{s}$, thus $Q_n = \frac{j+s_n}{2}, \forall n$.
2: **repeat**
*Optimize* $\mathbf{F}$ *when* $\mathbf{Q}$ *is fixed.*
3: **while** Convergence is not met **do**
4: Compute $\mathbf{f}_k$ using the WMMSE algorithm shown in (7a), (7b), (7c), $\forall k$.
5: **end while**
*Optimize* $\mathbf{Q}$ *when* $\mathbf{F}$ *is fixed.*
6: **while** Convergence is not met **do**
7: Compute $\mathbf{s}$ using GDP shown in (16).
8: **end while**
9: $Q_n = \frac{j+s_n}{2}, \forall n$.
10: **until** Objective function $R(\mathbf{F}, \mathbf{Q})$ converges
**Output:** $\mathbf{Q}$ and $\mathbf{F}$

## IV. DMA Weights Matrix and Precoding Matrix Decoupling Optimization Strategy

To solve the problem in (6a), by decoupling the designs of $\mathbf{Q}$ and $\mathbf{F}$, the optimizations of them can be significantly simplified. Motivated by it, DO strategy can be adopted to decouple the designs of DMA weights matrix and precoding matrix. Specifically, the problem of DMA weights matrix design based on DO is given by

$$\begin{aligned} \max_{\mathbf{Q}} \quad & \sum_{k=1}^{K}\frac{1}{K}\log_2\left(1 + \frac{||\mathbf{h}_k^T\mathbf{G}\mathbf{Q}||_2^2}{\sigma^2}\right), \\ \text{s.t.} \quad & (6c). \end{aligned} \tag{17}$$

From (17), it is noted that designing the DMA weights matrix $\mathbf{Q}$ based on DO is independent from the precoding matrix $\mathbf{F}$, so that $\mathbf{F}$ can be determined after $\mathbf{Q}$ is finally designed by solving (17). In contrast, based on the AO strategy, $\mathbf{F}$ and $\mathbf{Q}$ are alternatively optimized, which inevitably results in high computational complexity on multiple times of GDP and WMMSE convergence as shown in the following *Table 1*.

By changing the *Lorentzian* constraint on $Q$ into the unit-modulus constraint, the problem in (17) can be equivalently expressed as

$$\max_{\mathbf{s}_{(i)}\in\mathcal{M},\forall i} \quad \sum_{k=1}^{K}\frac{1}{K}\log_2\left(1 + \frac{\sum_{i=1}^{N_d}\left|c_{k,i} + \mathbf{s}_{(i)}\mathbf{d}_{k(i)}\right|^2}{\sigma^2}\right) \tag{18}$$

with

$$c_{k,i} = \frac{j}{2}\sum_{l=1}^{N_e} h_{(i-1)N_e+l,k}\, g_{i,l}. \tag{19}$$

Here, the $l$-th elements of introduced $\mathbf{s}_{(i)} \in \mathbb{C}^{1\times N_e}$ and $\mathbf{d}_{k(i)} \in \mathbb{C}^{N_e\times 1}$ are $e^{j\theta_{(i-1)N_e+l}}$ and $h_{(i-1)N_e+l,k}\, g_{i,l}, \forall i, \forall k$, respectively. Furthermore, we denote $\mathcal{M} = \{\mathbf{s} | s_l = e^{j\theta_l}, l \in \{1, \cdots, N_e\}\}$. Inspired by the proposed algorithm based on AO, the problem in (18) can be approximated by

$$\min_{\mathbf{s}_{(i)}\in\mathcal{M},\forall i} \quad \sum_{k=1}^{K}\frac{1}{K}\log_2\left(1 + \frac{|r_{k,i} - \mathbf{s}_{(i)}\mathbf{d}_{k(i)}|^2}{\sigma^2}\right) \tag{20}$$

**Algorithm 2** Proposed algorithm based on DO

**Input:** $\mathbf{G}, \mathbf{h}_k, \forall k \in [1, \cdots, K]$

*Optimize* $\mathbf{Q}$ *while* $\mathbf{F}$ *is ignored.*

Initialize $\mathbf{s}$ and compute $\sum_{m=1}^{N_d} |r_{k,m} - \mathbf{s}_{(m)} \mathbf{d}_{k(m)}|^2$.

**for** $i = 1, \cdots, N_d$ **do**

  **while** Convergence is not met **do**

    Compute $\mathbf{s}_{(i)}$ using the GDP method in (22).

  **end while**

  $Q_{(i-1)N_e+l} = \frac{j+e^{j\theta_{(i-1)N_e+l}}}{2}, l \in \{1, \cdots, N_e\}$.

**end for**

*Optimize* $\mathbf{F}$ *after* $\mathbf{Q}$ *is obtained.*

**while** Convergence is not met **do**

  Compute $\mathbf{f}_k$ using the WMMSE algorithm shown in (7a), (7b), (7c), $\forall k \in [1, \cdots, K]$.

**end while**

**Output:** $\mathbf{Q}$ and $\mathbf{F}$

with

$$r_{k,i} = \sum_{l=1}^{N_e} \frac{1}{2} \left| h_{(i-1)N_e+l,k} g_{i,l} \right| e^{\alpha_{k,i}} \tag{21}$$

and $\alpha_{k,i} = \angle (c_{k,i})$, followed by the GDP operations:

$$\begin{aligned} \nabla g(\mathbf{s}_{(i)}) &= -\frac{2}{\sigma^2 K \ln 2} \sum_{k=1}^{K} \frac{\left( r_{k,i} - \mathbf{s}_{(i)} \mathbf{d}_{k(i)} \right) \mathbf{d}_{k(i)}^H}{1 + \frac{1}{\sigma^2} \sum_{m=1}^{N_d} |r_{k,m} - \mathbf{s}_{(m)} \mathbf{d}_{k(m)}|^2}, \\ \mathbf{y} &= \mathbf{s}_{(i)} - \mu \nabla g(\mathbf{s}_{(i)}), \\ \mathbf{s}_{(i)} &= e^{j \angle \mathbf{y}}. \end{aligned} \tag{22}$$

By adopting GDP to get solution of each $\mathbf{s}_{(i)}, \forall i$, the DMA weights matrix $\mathbf{Q}$ can be determined.

After $\mathbf{Q}$ is finally determined by solving (17), the precoding matrix $\mathbf{F}$ can be optimized by WMMSE as shown in (7a), (7b) and (7c) until convergence. To conclude, the proposed algorithm based on DO is shown in *Algorithm 2*.

Now, we analyze its computational complexity and compare it with those algorithms based on AO strategy. As shown in *Algorithm 2*, first of all, the complexity of calculating $\sum_{m=1}^{N_d} |r_{k,m} - \mathbf{s}_{(m)} \mathbf{d}_{k(m)}|^2$ in (22) is $KN_d(N_e + 1)$. Then, the GDP method in (22) is adopted for $N_d$ times to get solution of each $\mathbf{s}_{(i)}, \forall i$. The complexity of GDP method in (22) is $\mathcal{O}(R_G N_e K)$. Finally, $\mathbf{F}$ is designed by the WMMSE algorithm with $\mathcal{O}\left(R_W N_d^3 K\right)$ complexity. Overall, the complexity of the proposed algorithm based on DO is given by $\mathcal{O}\left(R_W N_d^3 K + R_G N K\right)$.

Obviously, compared with those algorithms based on AO, the computational complexity of the proposed algorithm based on DO can be greatly reduced by avoiding multiple times of GDP and WMMSE convergence. Note that though the ignorance of $\mathbf{F}$ in DMA weights matrix design in DO results in inevitable performance loss, the following simulation results show that it is acceptable.

## V. Simulation Results

In this section, we compare the sum rate achieved by the proposed algorithms based on AO and DO with other algorithms in [11], [12]. To enable HMIMO beamforming, the spacing between DMA elements and micrstrips is set to be $\frac{\lambda}{5}$ [3] and $\alpha = 0.6\mathrm{m}^{-1}$, $\beta = 827.67\mathrm{m}^{-1}$ [11]. The transmitter power and receiver noise is set to be $23\mathrm{dBm}$ and -$80\mathrm{dBm}$. We consider the practical Saleh-Valenzuela millimeter wave channel model [16] with 28GHz carrier frequency. The large-scale fading factor is described by the square root of the distance-dependent path loss [11], given as

$$\Lambda(\tau, \xi)[\mathrm{dB}] = -45 - 25 \log_{10}(\tau), \tag{23}$$

where $\tau$ is the BS-user distance in meters, randomly distributed from 35 to 300. The user priority is set to be equal. We further consider fully digital (FD) antenna architecture obtained by WMMSE algorithm [17].

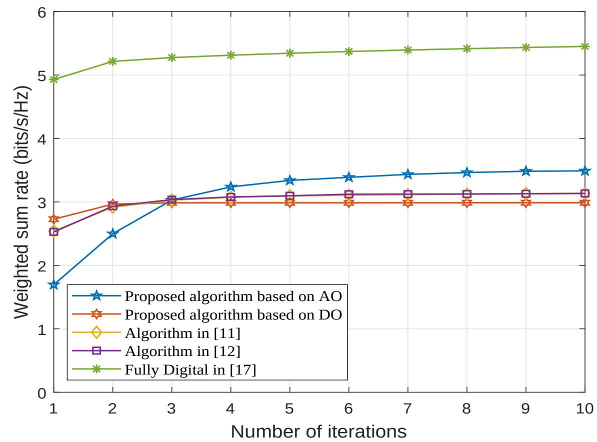


Fig. 2. Convergence behavior of different algorithms under the setting of $N_d$ = 15, $N_e$ = 10, $K$ = 5 and $P$ = 23dBm.

Fig. 2 shows the convergence behaviors of different algorithms and *Table 1* lists the corresponding computational complexities. Compared to the algorithms in [11], [12], the proposed algorithm based on AO achieves higher sum rate with lower complexity. Besides, the proposed algorithm based on DO has much lower complexity with acceptable performance loss, which illustrates that it can achieve trade-off between performance and complexity. Though the FD architecture can attain the highest sum rate, it suffers expensive RF chain hardware cost.

To evaluate the performance of various algorithms under different DMA array size, the mircostrip number is varied in Fig. 3. It illustrates that the proposed algorithm based on AO always attain better performance compared to other algorithms regardless of DMA array size. Besides, we note that with larger DMA array size, the performance advantage of the proposed algorithm based on AO over other algorithms is more obvious. It can be attributed to the fact that the proposed algorithm based on AO can make better use of DMA array gain.

In Fig. 4, it can be observed that as the user number increases, each algorithm suffers performance degradation due to enhanced inter-user interference. We observe that the proposed algorithm based on AO still remains more excellent than other algorithms even under high inter-user interference situation, because it can still obtain array gain to improve sum rate in DMA weights matrix design.

TABLE I
THE COMPUTATIONAL COMPLEXITIES OF PROPOSED ALGORITHMS AND OTHER ALGORITHMS

| | Overall complexity | $R_W$ | $R_G$ | $R_A$ | Complexity value |
|---|---|---|---|---|---|
| Proposed algorithm based on AO | $\mathcal{O}\left(R_A\left(R_W N_d^3 K + R_G N K\right)\right)$ | 3 | 12 | 5 | 298125 |
| Proposed algorithm based on DO | $\mathcal{O}\left(R_W N_d^3 K + R_G N K\right)$ | 3 | 2 | – | 52125 |
| The algorithms in [11], [12] | $\mathcal{O}\left(R_A\left(R_W N_d^3 K + R_G N K^2\right)\right)$ | 3 | 16 | 3 | 331875 |

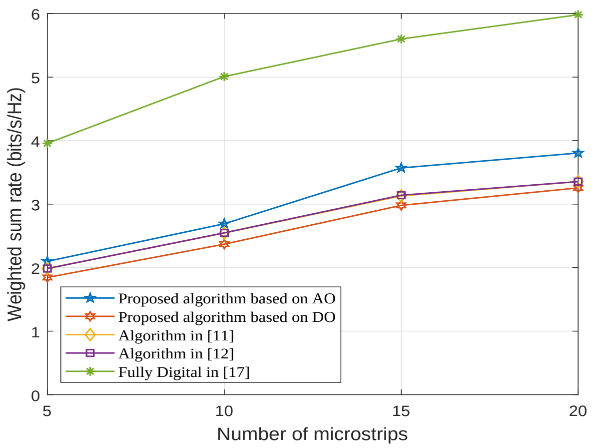


Fig. 3. WSR achieved by different algorithms with the increasing number of microstrips for multi-user downlink HMIMO systems with $N_e$ = 10, $K$ = 5 and $P$ = 23dBm.

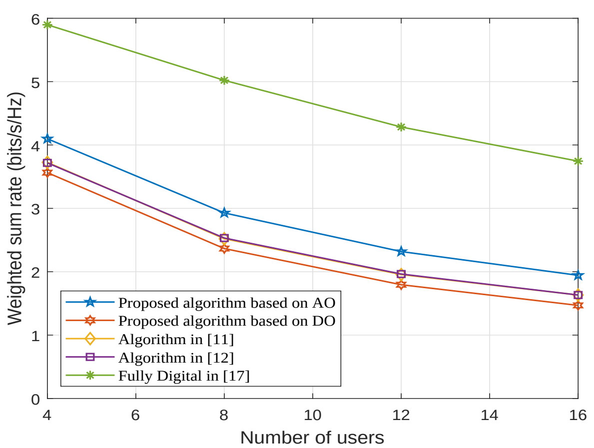


Fig. 4. WSR achieved by different algorithms with the increasing number of users for multi-user downlink HMIMO systems with $N_d$ = 16, $N_e$ = 10 and $P$ = 23dBm.

## VI. CONCLUSION

In this paper we focus on DMA weights matrix design and precoding matrix design in AO and DO strategy to maximize sum rate in the downlink multi-user HMIMO system. Specifically, these two matrices can be alternatively designed in AO. Based on it, we propose novel problem formulation of the DMA weights matrix design to harvest array gain, which can be solved with reduced complexity. To achieve performance and complexity trade-off, an algorithm based on DO is also proposed, in which the DMA weights matrix is optimized independently from the precoding matrix. Finally, simulation results confirm the advantages of the proposed algorithms on performance and complexity compared to other algorithms.

## ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under Grants No.2023YFC2205501, and in part by National Natural Science Foundation of China under Grants No.62371124.

## REFERENCES


[1] W. Chaowei et al., "Collaborative caching in vehicular edge network assisted by cell-free massive MIMO," *Chin. J. Electron.*, vol. 32, no. 6, pp. 1218-1229, Nov. 2023.

[2] N. Li and P. Fan, "Distributed cell-free massive MIMO versus cellular massive MIMO under UE hardware impairments," *Chin. J. Electron.*, vol. 33, no. 5, pp. 1274-1285, Sept. 2024.

[3] T. Gong et al., "Holographic MIMO Communications: Theoretical Foundations, Enabling Technologies, and Future Directions," *IEEE Commun. Surv. Tutorials*, vol. 26, no. 1, pp. 196-257, 2024.

[4] E. Björnson et al., "Massive MIMO is a reality What is next?: Five promising research directions for antenna arrays," *Dig. Signal Process.*, vol. 94, pp. 3-20, 2019.

[5] N. Shlezinger et al., "Dynamic Metasurface Antennas for 6G Extreme Massive MIMO Communications," *IEEE Wirel. Commun.*, vol. 28, no. 2, pp. 106-113, Apr. 2021.

[6] L. You et al., "Energy Efficiency Maximization of Massive MIMO Communications With Dynamic Metasurface Antennas," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 393-407, Jan. 2023.

[7] N. Shlezinger et al., "Dynamic Metasurface Antennas for Uplink Massive MIMO Systems," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6829-6843, Oct. 2019.

[8] H. Wang et al., "Dynamic Metasurface Antennas Based Downlink Massive MIMO Systems," in *Proc. 2019 IEEE 20th Int. Workshop Signal Process. Adv. Wirel. Commun. (SPAWC)*, Cannes, France, 2019, pp. 1-5.

[9] H. Zhang et al., "Beam focusing for multi-user MIMO communications with dynamic metasurface antennas," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4780 C4784.

[10] H. Zhang et al., "Beam Focusing for Near-Field Multiuser MIMO Communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7476-7490, Sept. 2022.

[11] S. F. Kimaryo and K. Lee, "Downlink Beamforming for Dynamic Metasurface Antennas," *IEEE Trans. Wireless Commun.*, vol. 22, no. 7, pp. 4745-4755, Jul. 2023.

[12] J. -C. Chen and C. -H. Hsu, "Beamforming Design for Dynamic Metasurface Antennas-Based Massive Multiuser MISO Downlink Systems," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 1387-1398, 2024.

[13] 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.

[14] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201-211, Jan. 2015.

[15] J. Tranter, et al., "Fast Unit-Modulus Least Squares With Applications in Beamforming," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2875-2887, Jun. 2017.

[16] O. E. Ayach et al., "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499-1513, Mar. 2014.

[17] Q. Shi et al., "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, Sept. 2011.