

Two-timescale Beamforming Optimization for Downlink Multi-user Holographic MIMO Surfaces

Haochen Wu, Yuanbin Chen, Yang Ming and Zhaocheng Wang, *Fellow, IEEE*

Abstract—Benefiting from the rapid development of metamaterials and metasurfaces, the holographic multiple-input and multiple-output surface (HMIMOS) has been regarded as a promising solution for future wireless networks recently. By densely packing numerous radiation elements together, HMIMOS is capable of realizing accurate beamforming with low hardware complexity. However, enormous radiation elements on the HMIMOS lead to high computational complexity and signaling overhead when applying traditional beamforming schemes relying on instantaneous channel state information (CSI). To tackle this problem, we propose a two-timescale optimization scheme to minimize the required transmission power under the constraint of all users' quality-of-service (QoS). Specifically, the beam patterns at the base station (BS) and the user equipment (UE) are optimized over the slowly changing statistical CSI based on the constrained successive convex approximation (CSCA) algorithm. Then, the instantaneous CSI is utilized to design the precoding matrix in order to ensure the system performance without significant increase in computational cost, due to the small number of feeds on the HMIMOS. Simulation results demonstrate the effectiveness of our proposed method compared to other baselines.

Index Terms—Holographic MIMO surfaces, holographic beamforming, two-timescale optimization, convex optimization.

I. INTRODUCTION

Following the recent breakthroughs on the fabrication of programmable metamaterials, the holographic multiple-input and multiple-output surface (HMIMOS) paradigm has attracted increasing attention for its flexible and accurate beam steering capabilities, as well as low manufacturing and hardware costs. To be specific, inlaid with numerous radiation elements whose spacing is smaller than half wavelength of incident electromagnetic (EM) waves, HMIMOS is able to realize the continuous aperture and form sharp beams with weak sidelobes. Moreover, based on the holographic principle and advanced programmable metamaterial technology, HMIMOS can easily reconfigure its holographic pattern and control its EM response with low cost, hence generating the desired beam directions flexibly [1].

HMIMOS has been investigated in various aspects recently, such as channel modeling [2], hardware design [3], [4] and lo-

calization application [5]. Particularly, regarding the holographic beamforming techniques, in [6], a hybrid beamforming scheme is proposed to achieve accurate multi-beam steering in a HMIMOS-assisted multi-user system. In [7], the satellite communication networks are considered, and a holographic beamforming algorithm is proposed to maximize the system sum rate. Moreover, the discrete amplitude-controlled HMIMOS is investigated in [8], demonstrating a lower bound for the sum rate in quantized HMIMOS-assisted communication systems. However, obtaining instantaneous channel state information (CSI) is challenging in holographic MIMO scenarios, since numerous radiation elements on the HMIMOS lead to larger CSI size, and the signaling overhead which is typically related to the scale of CSI becomes also higher. In addition, enormous radiation elements lead to high computational complexity during beamforming optimization when applying traditional beamforming algorithms. Therefore, the conventional beamforming methods requiring instantaneous CSI like [6], [7] becomes impractical, which makes practical beamforming in the HMIMOS-assisted system a challenging task.

In this paper, a two-timescale (TTS) beamforming optimization method is proposed in a downlink multi-user HMIMOS-assisted communication system to reduce the computational complexity and the signaling overhead in comparison to the beamforming schemes relying upon the instantaneous CSI. We aim to minimize the average transmit power at the BS, subject to the quality-of-service (QoS) constraints of all users on the achievable spectral efficiency. On one hand, the long-term beam patterns at the BS and the UE side are optimized according to the slow-varying statistical CSI. On the other hand, the short-term precoding matrix at the BS side is properly designed in different time slots based on all users' reconfigured instantaneous CSI with optimized long-term beam patterns, therefore reducing the average transmit power of the BS effectively. Especially, the overhead for instantaneous CSI acquisition utilized in the precoding design is proportional to the number of feeds attached to the HMIMOS, which is much less than the number of radiation elements. As a result, the complexity can not be introduced by leveraging instantaneous CSI in the precoding step. In order to solve the formulated optimization problem spanning over different timescales, the constrained stochastic successive convex approximation (CSCA) algorithm is employed. Simulation results reveal that our proposed method is able to achieve a proper trade-off between the system transmit power and the overhead of channel acquisition.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

As shown in Fig. 1, we consider a downlink multi-user communication system between one BS and a group of U UEs denoted by $\mathcal{U} = \{1, 2, \dots, U\}$. Each user is served by one data stream.

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This work was supported in part by National Key R&D Program of China under Grant 2018YFB1801501. (Corresponding author: Zhaocheng Wang.)

H. Wu is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wuhc23@mails.tsinghua.edu.cn).

Y. Chen is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: chen_yuanbin@163.com).

Y. Ming is with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: mingy20@mails.tsinghua.edu.cn).

Z. Wang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and also with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: zcwang@tsinghua.edu.cn).

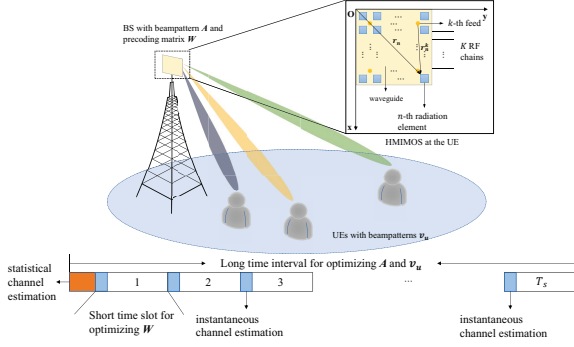


Fig. 1. Diagram of TTS method.

Moreover, we assume that both BS and UEs are equipped with HMIMOS. The HMIMOS is a kind of special leaky-wave antenna consisting of feeds, a planar array of amplitude-controlled radiation elements and a parallel-plate waveguide, which is depicted in Fig. 1. The feeds are embedded in the bottom layer of the HMIMOS in order to generate reference waves, which propagate on the waveguide and carry the radiating signals. Meanwhile, the radiation elements are located on the surface of the waveguide. Excited by the reference wave, the amplitude-controlled radiation elements can intelligently control their electromagnetic response, thus generating direct and sharp beams with low sidelobes.

As for the key features of HMIMOS, the number of antenna elements is much more than that in the conventional MIMO, which leads to better beamforming performance. In addition, these radiation elements are all amplitude-controlled since the HMIMOS is a kind of leaky-wave antenna, while the phase control method is usually adopted in the conventional MIMO. When compared with reconfigurable intelligent surfaces (RISs), HMIMOS is different in terms of its physical structure and operating mechanism. Additionally, HMIMOS is generally used as the active transmitting and receiving antenna in order to reduce the complexity of feeding networks, while the RIS is widely utilized as passive relays to extend cell coverage.

Specifically, in our considered system, the HMIMOS at the BS consists of $N = N_t^2$ radiation elements and K feeds, while those at the UEs are comprised of $M = N_r^2$ radiation elements and one feed. It is required that the number of feeds at the BS should be more than the number of active data streams, i.e. $K > U$, and the spacing of radiation elements is below half of the wavelength λ of the incident EM wave [6].

Since the HMIMOS can not perform digital signal processing, digital beamforming is performed at the baseband and at the BS. Accordingly, the received signal y_u of the u -th UE can be expressed as

$$y_u = \mathbf{v}_u^H \mathbf{B} \mathbf{H}_u^H \mathbf{A} \Theta \mathbf{w}_u x_u + \mathbf{v}_u^H \mathbf{B} \mathbf{H}_u^H \mathbf{A} \Theta \sum_{u' \neq u} \mathbf{w}_{u'} x_{u'} + \mathbf{v}_u^H \mathbf{B} \mathbf{z}_u, \quad (1)$$

where x_u is the intended signal for the u -th UE, and $\mathbf{w}_u \in \mathbb{C}^{K \times 1}$ is the u -th column of the precoding matrix $\mathbf{W}$ at the BS.

Moreover, we denote $\boldsymbol{\alpha} \in \mathbb{R}^{N \times 1}$ as the beampattern of the HMIMOS at the BS, and each element in $\boldsymbol{\alpha}$ represents the gain of one corresponding radiation element on the HMIMOS at the transmitter side, which is obtained based on the holographic

principle. For the n -th radiation element on the HMIMOS, we denote its position as $\mathbf{r}_n$, and its distance vector from the k -th feed as $\mathbf{r}_n^k$. In addition, the propagation vector of the object beam Ψ_{obj} and the reference waves generated by all feeds Ψ_{ref} are regarded as $\mathbf{k}_s$ and $\mathbf{k}_f$, respectively. Therefore, the interference Ψ_{intf} between the object beam and the reference wave, which is named as the beampattern, can be expressed as

$$\Psi_{intf} = \Psi_{obj} \Psi_{ref}^* = e^{-j(\mathbf{k}_f \mathbf{r}_n - \mathbf{k}_s \mathbf{r}_n^k)}. \quad (2)$$

If the acquired beampattern is excited by the reference wave Ψ_{ref} , the generated beam satisfies $\Psi_{ref} \Psi_{intf} \propto \Psi_{obj} |\Psi_{ref}|^2$, whose direction is exactly the same as the object beam. Note that the radiation elements are all amplitude-controlled, $\Re(\Psi_{intf})$ is usually chosen to parameterize the radiation amplitude to generate the object beam. In order to avoid the negative values, $\Re(\Psi_{intf})$ is normalized to $[0, 1]$. Therefore, the amplitude for $\alpha(\mathbf{r}_n^k)$ can be expressed by

$$\alpha(\mathbf{r}_n^k) = \frac{\Re(\Psi_{intf}(\mathbf{r}_n^k)) + 1}{2}. \quad (3)$$

We denote $\mathbf{A} = \text{diag}(\boldsymbol{\alpha}) = \text{diag}([\alpha_1, \alpha_2, \dots, \alpha_N])$, where $\alpha_n = \frac{\sum_{k=1}^K \alpha(\mathbf{r}_n^k)}{K}$. Similarly, for the u -th UE, the beampattern of the HMIMOS is $\mathbf{v}_u \in \mathbb{R}^{M \times 1}$, which represents the gain weight of each radiation elements of different HMIMOS at the receiver side, and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_U]$. According to the basic principles of the HMIMOS, the elements in $\mathbf{v}_u$ are also real numbers in $[0, 1]$ [1].

Furthermore, Θ is a $N \times K$ complex matrix composed of element $\{e^{-j\mathbf{k}_s \mathbf{r}_n^k}\}$ which represents the phase difference between the n -th radiation element and the k -th feed at the transmitter side. Both $\mathbf{A}$ and Θ are dependent on the distance vector $\mathbf{r}_n^k$, which allows the holographic beamforming to be different from the conventional hybrid precoding in massive MIMO systems. Since the phase matrix is determined once the positions of feeds and radiation elements are fixed, we assume the phase matrices for all UEs to be the same, which is similarly denoted by $\beta \in \mathbb{C}^{M \times 1}$, and $\mathbf{B} = \text{diag}(\beta)$. The additive white Gaussian noise is denoted by $\mathbf{z}_u \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_U)$.

As for the channel model, the widely used Saleh-Valenzuela model for millimeter-wave communications is adopted [9]. The channel matrix $\mathbf{H}_u$ for the u -th UE consisting of one line-of-sight (LoS) path and L non-LoS (NLoS) paths can be expressed as

$$\mathbf{H}_u = \beta_u^{(0)} \mathbf{a}_r(\omega_u^{(0)}, \phi_u^{(0)}) \mathbf{a}_t(\theta_u^{(0)}, \psi_u^{(0)})^H + \sum_{l=1}^L \beta_u^{(l)} \mathbf{a}_r(\omega_u^{(l)}, \phi_u^{(l)}) \mathbf{a}_t(\theta_u^{(l)}, \psi_u^{(l)})^H, \quad (4)$$

where $\beta_u^{(l)}$, $l = 0, 1, \dots, L$ are the complex gains of LoS and NLoS paths, and $(\omega_u^{(l)}, \phi_u^{(l)})$ and $(\theta_u^{(l)}, \psi_u^{(l)})$ for $0 \leq l \leq L$ represent the physical angles of arrival (AoAs) and angles of departure (AoDs). In addition, the antenna array response at the u -th UE and the BS are $\mathbf{a}_r(\omega_u^{(l)}, \phi_u^{(l)})$ and $\mathbf{a}_t(\theta_u^{(l)}, \psi_u^{(l)})$ for both LoS and NLoS paths, respectively [6]. It is assumed that the spacing between the radiation elements of the HMIMOS at the BS and at all UEs along the x -axis and y -axis are both Δ_s . After taking the position of the HMIMOS at both the BS and the UEs into consideration, $\mathbf{a}_r(\omega_u^{(l)}, \phi_u^{(l)})$ and $\mathbf{a}_t(\theta_u^{(l)}, \psi_u^{(l)})$ can be given by

$$\begin{aligned} \mathbf{a}_r(\omega_u^{(l)}, \phi_u^{(l)}) \\ = [1, \dots, e^{jk_f[(i-1)\Delta_s \sin \omega_u^{(l)} \cos \phi_u^{(l)} + (j-1)\Delta_s \cos \phi_u^{(l)}]}, \\ \dots, e^{jk_f[(N_r-1)\Delta_s \sin \omega_u^{(l)} \cos \phi_u^{(l)} + (N_r-1)\Delta_s \cos \phi_u^{(l)}]}], \end{aligned} \quad (5)$$

$$\begin{aligned} \mathbf{a}_t(\theta_u^{(l)}, \psi_u^{(l)}) \\ = [1, \dots, e^{jk_f[(i-1)\Delta_s \sin \theta_u^{(l)} \cos \psi_u^{(l)} + (j-1)\Delta_s \cos \psi_u^{(l)}]}, \\ \dots, e^{jk_f[(N_t-1)\Delta_s \sin \theta_u^{(l)} \cos \psi_u^{(l)} + (N_t-1)\Delta_s \cos \psi_u^{(l)}]}], \end{aligned} \quad (6)$$

where $k_f = \frac{2\pi}{\lambda}$ represents the wave number of the incident EM wave.

Based on Eq. (1), the received signal-to-interference-plus-noise ratio (SINR) of the u -th UE can be expressed as

$$\text{SINR}_u = \frac{|\mathbf{v}_u^H \mathbf{B} \mathbf{H}_u^H \mathbf{A} \Theta \mathbf{w}_u|^2}{\sum_{u' \neq u} |\mathbf{v}_u^H \mathbf{B} \mathbf{H}_u^H \mathbf{A} \Theta \mathbf{w}_{u'}|^2 + \|\mathbf{v}_u^H \mathbf{B}\|^2 \sigma^2}, \quad (7)$$

and the spectral efficiency of the u -th UE can be therefore written as $\log_2(1 + \text{SINR}_u)$.

B. Problem Formulation

It is worth pointing out that in conventional beamforming methods, the beampattern is mainly designed based on the instantaneous CSI [6]–[8]. Nevertheless, with the increase of radiation elements on the HMIMOS, the size of the instantaneous CSI increases accordingly. As a result, obtaining instantaneous CSI in the HMIMOS-assisted systems entails high computational cost and signaling overhead. Therefore, we propose the TTS optimization method in the HMIMOS communication system in order to decompose the problem into two sub-problems with different timescales [10], which is also shown in Fig. 1.

Specifically, we focus on a sufficient long time interval composed of T_s short time slots, where the statistical CSI remains unchanged in every long time interval, and the instantaneous CSI stays the same in each short time slot. In the long time interval, the long-term problem is solved based on the statistical CSI to optimize the holographic beampatterns at both BS and UE sides. The statistical CSI, e.g. the channel mean and associated AoAs/AoDs, can be acquired in the beginning of each long time interval, as shown in Fig. 1. Owing to the slow-varying statistical CSI, the long-term problem requires fewer channel acquisitions than its conventional counterparts. Therefore, the computational complexity and signaling overhead could be reduced dramatically. Moreover, in one short time slot, the instantaneous CSI is utilized to optimize the precoding matrix $\mathbf{W}$ at the BS. Similarly, the instantaneous CSI including the channel gains and the phase parameters are acquired in the beginning of each short time slot through instantaneous channel estimation. For accurate channel estimation in holographic MIMO scenarios, perfect knowledge of statistical CSI is utilized in [11] in order to propose a Gram-Schmidt decomposition channel estimation algorithm, while the deep-learning based approach is introduced in [12], which has high NMSE gains and reduces the pilot overhead considerably. The computational complexity and signaling overhead of the short-term optimization are typically related to the scale of the instantaneous CSI, which is in proportion to the number of feeds

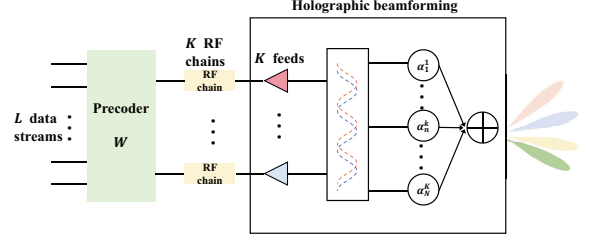


Fig. 2. Data Transmission Process [13].

K rather than the number of radiation elements N . Since K is much smaller than N , leveraging instantaneous CSI in the short-term optimization can ensure the overall system performance while avoiding high signaling overhead.

In the considered HMIMOS-assisted system, we mainly follow the transmission process in Fig. 2 [13], while the distinction is that we optimize the beampattern α , $\mathbf{V}$ and the precoding matrix $\mathbf{W}$ based on the CSI with different timescales in order to minimize the expected transmit power with guaranteed QoS for each UE, which can be written as

$$\begin{aligned} \min_{\{\alpha, \mathbf{V}\}, \mathbf{W}} \mathbb{E}\{\text{Tr}(\mathbf{A} \Theta \mathbf{W} \mathbf{W}^H \Theta^H \mathbf{A}^H)\} \\ \text{s.t. } \mathbb{E}\{\log_2(1 + \text{SINR}_u)\} \geq \delta_u, \\ \alpha(i), v_u(i) \in [0, 1], \forall u \in \mathcal{U}, \end{aligned} \quad (8)$$

where δ_u is the threshold of the u -th user's spectral efficiency to ensure the QoS for all UEs.

In the long-term problem, the statistical CSI is utilized to optimize the beampattern α and $\mathbf{V}$ at both the BS and the UE side for given $\widehat{\mathbf{W}}$. Accordingly, the long-term problem can be formulated

$$\begin{aligned} \min_{\alpha, \mathbf{V}} \mathbb{E}\{\text{Tr}(\mathbf{A} \Theta \widehat{\mathbf{W}} \widehat{\mathbf{W}}^H \Theta^H \mathbf{A}^H)\} \\ \text{s.t. } \mathbb{E}\{\log_2(1 + \text{SINR}_u)\} \geq \delta_u, \\ \alpha(i), v_u(i) \in [0, 1], \forall u \in \mathcal{U}. \end{aligned} \quad (9)$$

In the short-term problem, the precoding matrix $\mathbf{W}$ is optimized at each short time slot for given $\widehat{\alpha}$ and $\widehat{\mathbf{V}}$, where the instantaneous CSI is leveraged. The short-term problem can be expressed as

$$\begin{aligned} \min_{\mathbf{W}} \text{Tr}(\widehat{\mathbf{A}} \Theta \mathbf{W} \mathbf{W}^H \Theta^H \widehat{\mathbf{A}}^H) \\ \text{s.t. } \log_2(1 + \text{SINR}_u) \leq \delta_u, \\ \widehat{\alpha}(i), \widehat{v}_u(i) \in [0, 1], \forall u \in \mathcal{U}, \end{aligned} \quad (10)$$

where the expectation operation is removed since the instantaneous CSI is determined in each short time slot.

III. METHODOLOGY DESIGN

This section details the proposed TTS optimization method, where the CSSCA algorithm is leveraged to solve the long-term problem with randomly generated channel samples in the long time interval, while the short-term problem is optimized in each short time slot.

A. Long-term Optimization

The long-term optimization is performed to obtain the optimal beampattern at the UE and the BS side with fixed precoding matrix at the beginning of every long time interval. However, the long-term problem (9) is non-convex and has an expectation constrained form, which can not be solved simply. Consequently, we use the CSSCA algorithm to tackle the difficulty. The basic idea of CSSCA is to iteratively solving a sequence of convex problems obtained by replacing the objective/constraint function in the original problem with a wide class of surrogate functions [14].

To be specific, at each iteration t , a set of new channel samples $\{\mathbf{H}_l^t\}$ is randomly generated from the known statistical CSI. Since the beampattern matrix $\mathbf{A} = \text{diag}(\boldsymbol{\alpha})$ at the UE side is a diagonal matrix, we can firstly utilize the equation $\text{Tr}(\mathbf{A}\boldsymbol{\Theta}\widehat{\mathbf{W}}\widehat{\mathbf{H}}^H\boldsymbol{\Theta}^H\mathbf{A}^H) = \sum_u \|\text{diag}(\boldsymbol{\Theta}\widehat{\mathbf{w}}_u)\boldsymbol{\alpha}\|^2$ to convert the problem (9) to

$$\min_{\boldsymbol{\alpha}, \mathbf{V}} f_0(\boldsymbol{\alpha}) = \mathbb{E}\{g_0(\boldsymbol{\alpha}, \widehat{\mathbf{W}}, \mathbf{H})\} \quad (11)$$

$$\text{s.t. } f_u(\boldsymbol{\alpha}, \mathbf{V}) = \mathbb{E}\{g_u(\boldsymbol{\alpha}, \mathbf{V}, \widehat{\mathbf{W}}, \mathbf{H})\} \leq 0, \forall u \in \mathcal{U},$$

where

$$g_0(\boldsymbol{\alpha}, \widehat{\mathbf{W}}, \mathbf{H}) = \sum_u \|\text{diag}(\boldsymbol{\Theta}\widehat{\mathbf{w}}_u)\boldsymbol{\alpha}\|^2, \quad (12a)$$

$$g_u(\boldsymbol{\alpha}, \mathbf{V}, \widehat{\mathbf{W}}, \mathbf{H}) = \delta_u - \log_2(1 + \text{SINR}_u), \quad (12b)$$

and

$$\text{SINR}_u = \frac{|\mathbf{v}_u^H \mathbf{B} \mathbf{H}_u^H \text{diag}(\boldsymbol{\Theta}\widehat{\mathbf{w}}_u)\boldsymbol{\alpha}|^2}{\sum_{u' \neq u} |\mathbf{v}_u^H \mathbf{B} \mathbf{H}_u^H \text{diag}(\boldsymbol{\Theta}\widehat{\mathbf{w}}_{u'})\boldsymbol{\alpha}|^2 + \|\mathbf{v}_u^H \mathbf{B}\|^2 \sigma^2}. \quad (13)$$

Then, according to the CSSCA algorithm, the non-convex stochastic functions $f_0(\boldsymbol{\alpha})$ and $f_u(\boldsymbol{\alpha}, \mathbf{V})$ can be approximated with two sequences of surrogate functions for all $1 \leq u \leq U$, i.e.,

$$\bar{f}_0^t(\boldsymbol{\alpha}) = f_0^t + (\mathbf{f}_\alpha^t)^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}) + \epsilon_0 \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}\|^2, \quad (14a)$$

$$\begin{aligned} \bar{f}_u^t(\boldsymbol{\alpha}, \mathbf{V}) &= f_u^t + (\mathbf{f}_{\alpha,u}^t)^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}) + (\mathbf{f}_{\mathbf{v},u}^t)^T(\mathbf{v}_u - \mathbf{v}_u^{t-1}) \\ &\quad + \epsilon_u (\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}\|^2 + \|\mathbf{v}_u - \mathbf{v}_u^{t-1}\|^2), \end{aligned} \quad (14b)$$

where ϵ_0 and ϵ_u are constant numbers to guarantee the convexity of the functions. f_0^t and f_u^t are the average approximation of $f_0(\boldsymbol{\alpha})$ and $f_u(\boldsymbol{\alpha}, \mathbf{V})$ through the channel sample set $\{\mathbf{H}_l\}$, which can be updated by

$$f_0^t = (1 - \rho^t)f_0^{t-1} + \rho^t \sum_{l=1}^{T_H} \frac{g_0(\boldsymbol{\alpha}^{t-1}, \mathbf{W}_l^t, \mathbf{H}_l^t)}{T_H}, \quad (15a)$$

$$f_u^t = (1 - \rho^t)f_u^{t-1} + \rho^t \sum_{l=1}^{T_H} \frac{g_u(\boldsymbol{\alpha}^{t-1}, \mathbf{V}^{t-1}, \mathbf{W}_l^t, \mathbf{H}_l^t)}{T_H}, \quad (15b)$$

where $\rho^t = \frac{1}{(1+t)^{2/3}}$ is a decreasing factor with the increase of the iteration number t . The intermediate variables $\mathbf{W}_l^t = \mathbf{W}(\boldsymbol{\alpha}^{t-1}; \mathbf{H}_l^t)$ are obtained by solving the short-term problem, while $\boldsymbol{\alpha}^{t-1}$ represents the beampattern at the UE side in the $(t-1)$ -th iteration, and $\mathbf{H}_l^t$ represents the l -th channel sample in the t -th iteration.

Algorithm 1 Long-term CSSCA algorithm

- 1: Initialize $\boldsymbol{\alpha}^0, \mathbf{V}^0$ randomly and set $t \leftarrow 1$.
 - 2: **repeat**
 - 3: Set $\rho^t = \frac{1}{(1+t)^{2/3}}, \gamma^t = \frac{2}{2+t}$.
 - 4: Generate a set of T_H channel samples $\{\mathbf{H}_l^t\}$ according to the statistical CSI.
 - 5: Obtain $\mathbf{W}_l^t$ based on the channel sample $\mathbf{H}_l^t$ and $\boldsymbol{\alpha}^{t-1}$ from the previous iteration.
 - 6: Update surrogate functions according to Eq. (14).
 - 7: Optimize $\bar{\boldsymbol{\alpha}}^t$ and $\bar{\mathbf{V}}^t$.
 - 8: Update $\boldsymbol{\alpha}^t$ and $\mathbf{V}^t$ based on Eq. (18).
 - 9: $t \leftarrow t + 1$
 - 10: **until** the number of iterations t reaches the upper limit of iteration N_{iter} .
-

Moreover, the function vectors $\mathbf{f}_\alpha^t, \mathbf{f}_{\alpha,u}^t$ and $\mathbf{f}_{\mathbf{v},u}^t$ are the approximations of the gradients $\nabla_{\boldsymbol{\alpha}} f_0(\boldsymbol{\alpha})$, $\nabla_{\boldsymbol{\alpha}} f_u(\boldsymbol{\alpha}, \mathbf{V})$ and $\nabla_{\mathbf{v}_u} f_u(\boldsymbol{\alpha}, \mathbf{V})$ respectively, which can be given by

$$\mathbf{f}_\alpha^t = (1 - \rho^t)\mathbf{f}_\alpha^{t-1} + \rho^t \sum_{l=1}^{T_H} \frac{\nabla_{\boldsymbol{\alpha}} g_0(\boldsymbol{\alpha}^{t-1}, \mathbf{W}_l^t, \mathbf{H}_l^t)}{T_H}, \quad (16a)$$

$$\mathbf{f}_{\alpha,u}^t = (1 - \rho^t)\mathbf{f}_{\alpha,u}^{t-1} + \rho^t \sum_{l=1}^{T_H} \frac{\nabla_{\boldsymbol{\alpha}} g_u(\boldsymbol{\alpha}^{t-1}, \mathbf{V}^{t-1}, \mathbf{W}_l^t, \mathbf{H}_l^t)}{T_H}, \quad (16b)$$

$$\mathbf{f}_{\mathbf{v},u}^t = (1 - \rho^t)\mathbf{f}_{\mathbf{v},u}^{t-1} + \rho^t \sum_{l=1}^{T_H} \frac{\nabla_{\mathbf{v}_u} g_u(\boldsymbol{\alpha}^{t-1}, \mathbf{V}^{t-1}, \mathbf{W}_l^t, \mathbf{H}_l^t)}{T_H}. \quad (16c)$$

The expressions of the gradients in (16) can be easily obtained [15], which is omitted in this paper.

Based on (14)-(16), the long-term problem (11) is approximated by solving the following surrogate optimization problem.

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \mathbf{V}} \bar{f}_0^t(\boldsymbol{\alpha}) \\ \text{s.t. } \bar{f}_u^t(\boldsymbol{\alpha}, \mathbf{V}) \leq 0, \forall u \in \mathcal{U}. \end{aligned} \quad (17)$$

As is proven in [14], (17) is a convex problem.

Finally, with the optimal beampatterns $\bar{\boldsymbol{\alpha}}$ and $\bar{\mathbf{V}}$ derived from the convex optimization problem (17), $\boldsymbol{\alpha}$ and $\mathbf{V}$ at the t -th iteration are updated as

$$\boldsymbol{\alpha}^t = (1 - \gamma^t)\boldsymbol{\alpha}^{t-1} + \gamma^t \bar{\boldsymbol{\alpha}}^t, \quad (18a)$$

$$\mathbf{V}^t = (1 - \gamma^t)\mathbf{V}^{t-1} + \gamma^t \bar{\mathbf{V}}^t, \quad (18b)$$

where $\gamma^t = \frac{2}{2+t}$ is also a factor which decreases with t .

The above iterative long-term optimization algorithm is summarized in Algorithm 1.

B. Short-term Optimization

In order to optimize the precoding matrix at the BS, the short-term problem (10) is performed within every short time slot to guarantee the overall system performance without bringing much computational and signaling overhead. Through some variable substitutions, the short-term problem can be transformed into a SOCP problem [16].

TABLE I
SYSTEM PARAMETERS.

Parameters	Values
Number of radiation elements at the BS N	256
Number of radiation elements at each UE M	36
Number of feeds at the BS K	9
Central frequency f_c	30 GHz
Bandwidth W	100 MHz
Spacing between elements Δ_s	$\lambda/4 = 2.5$ mm
Noise power spectrum density	-169 dBm/Hz

$\text{Tr}(\hat{\mathbf{A}}\Theta\mathbf{W}\mathbf{W}^H\Theta^H\hat{\mathbf{A}}^H) = \sum_u \|\text{diag}(\Theta\mathbf{w}_u)\hat{\boldsymbol{\alpha}}\|^2$ is utilized again, and the short-term problem is therefore converted to

$$\begin{aligned} \min_{\mathbf{W}} \quad & \sum_u \|\text{diag}(\Theta\mathbf{w}_u)\hat{\boldsymbol{\alpha}}\|^2 \\ \text{s.t.} \quad & \frac{|\hat{\mathbf{v}}_u^H \mathbf{B} \mathbf{H}_u^H \hat{\mathbf{A}} \Theta \mathbf{w}_u|^2}{\sum_{u' \neq u} |\hat{\mathbf{v}}_{u'}^H \mathbf{B} \mathbf{H}_{u'}^H \hat{\mathbf{A}} \Theta \mathbf{w}_{u'}|^2 + \|\hat{\mathbf{v}}_u^H \mathbf{B}\|^2 \sigma^2} \geq 2^{\delta_u} - 1. \end{aligned} \quad (19)$$

Let $\bar{\mathbf{w}}_u = \text{diag}(\Theta\mathbf{w}_u)\hat{\boldsymbol{\alpha}}$, $\mathbf{h}_u = \hat{\mathbf{v}}_u^H \mathbf{B}^H \mathbf{H}_u$, $\eta_u = 2^{\delta_u} - 1$ and $\bar{\sigma}_u = \|\hat{\mathbf{v}}_u^H \mathbf{B}\|^2 \sigma^2$. The problem (19) is further simplified to

$$\begin{aligned} \min_{\mathbf{W}} \quad & \sum_u \|\bar{\mathbf{w}}_u\|^2 \\ \text{s.t.} \quad & f(u) = \frac{1}{\eta_u \bar{\sigma}_u} |\mathbf{h}_u^H \bar{\mathbf{w}}_u|^2 \geq \sum_{u' \neq u} \frac{1}{\bar{\sigma}_u} |\mathbf{h}_{u'}^H \bar{\mathbf{w}}_{u'}|^2 + 1. \end{aligned} \quad (20)$$

For the problem (20), we only need to consider the constraint function, as the objective function is already convex. Because the constraint function $f(u)$ only involves the absolute value of variable $\mathbf{h}_u^H \bar{\mathbf{w}}_u$, adding a phase factor $e^{j\theta}$ to $\bar{\mathbf{w}}_u$ will not cause any difference to the value of $f(u)$. In this way, $\mathbf{h}_u^H \bar{\mathbf{w}}_u$ can be considered to be real and positive by adjusting the phase factor added to $\bar{\mathbf{w}}_u$. Next, by taking square roots of both sides of the constraint function, we can express the constraint of the problem (20) as

$$\begin{aligned} \frac{1}{\sqrt{\eta_u \bar{\sigma}_u}} \sqrt{|\mathbf{h}_u^H \bar{\mathbf{w}}_u|^2} &= \frac{1}{\sqrt{\eta_u \bar{\sigma}_u}} \mathbf{h}_u^H \bar{\mathbf{w}}_u \geq \\ \sqrt{\sum_{u' \neq u} \frac{1}{\bar{\sigma}_u} |\mathbf{h}_{u'}^H \bar{\mathbf{w}}_{u'}|^2 + 1}, \quad \forall u \in \mathcal{U}, \end{aligned} \quad (21)$$

which is a SOCP constraint [10]. As a result, the short-term problem is converted to a SOCP optimization problem and can be solved with the CVX toolbox [17].

IV. SIMULATION RESULTS

In this section, a multi-user system with $U = 4$ users is considered, where the HMIMOS are utilized at both the BS and the UE side. The UEs are evenly distributed around the BS with a distance within 10 m, and the height of the BS and the UEs are set to 25 m and 1.5 m, respectively. As for the channel model, the UEs are all set to have $L = 2$ NLoS paths, and the complex gains of LoS paths are given by $\beta_u^{(0)} = 61.4 + 20 \log_{10}(d_{3D})$, while the gains of NLoS paths can be denoted by $\beta_u^{(l)} \sim \mathcal{CN}(0, 0.1(\beta_u^{(0)})^2)$ according to the settings in [18], where d_{3D} represents the spatial distance between the BS and UEs. Other corresponding parameters of the system are listed in Table I.

The following benchmarks are considered for comparisons. 1) *Alternative optimization (AO) with instantaneous CSI*: $\boldsymbol{\alpha}$, $\mathbf{V}$ and

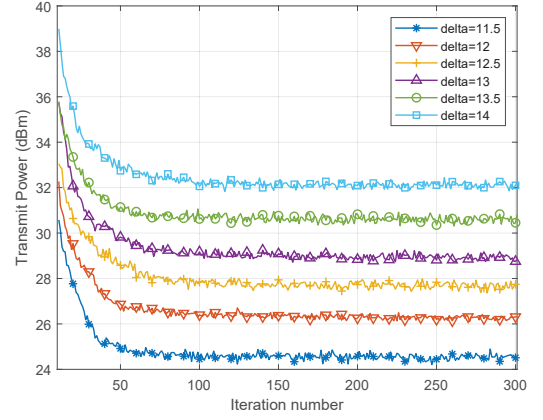


Fig. 3. Convergence behavior of the proposed algorithm under different QoS constraints.

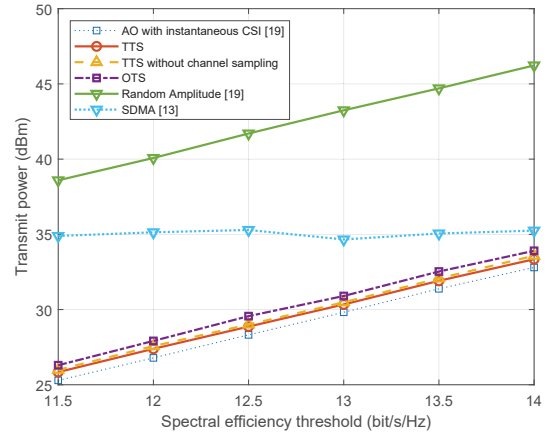


Fig. 4. Minimum transmit power P versus spectral efficiency threshold δ_u .

$\mathbf{W}$ are alternatively optimized based on the instantaneous CSI for each short time slot [19]. 2) *One time slot (OTS) optimization*: $\boldsymbol{\alpha}$, $\mathbf{V}$ and $\mathbf{W}$ are optimized based on the instantaneous CSI of the first short time slot and remain unchanged during the whole long time interval. 3) *TTS optimization without channel sampling*: $\boldsymbol{\alpha}$, $\mathbf{V}$ and $\mathbf{W}$ are optimized based on the TTS algorithm while the channel sample set remains unchanged during the whole long time interval. 4) *Random amplitude*: $\boldsymbol{\alpha}$ is randomly generated, while $\mathbf{V}$ and $\mathbf{W}$ are jointly optimized in each short time slot with instantaneous CSI and $\boldsymbol{\alpha}$ [19]. 5) *SDMA*: The optimization algorithm in [13], where the precoding matrix $\mathbf{W}$ is obtained through ZF precoding based on instantaneous CSI.

Prior to the performance comparison, we commence by showing the convergence behavior of CSSCA algorithm for the long-term problem as shown in Fig. 3. It can be seen that, our proposed method converges within 300 iterations. However, due to the non-convexity of the optimization problem and the stochastic nature of the proposed CSSCA algorithm, the curve of minimum power is not strictly monotonic. Moreover, with the increase of the user spectral efficiency threshold δ_u , the transmit power in the long-term optimization also converges to a higher value accordingly.

In Fig. 4, the minimum transmit power achieved by all methods is evaluated over the user spectral efficiency threshold δ_u with the size of channel sample set T_H fixed to 10, ϵ_0 and ϵ_u all set to 0.01 and the upper limit of iterations N_{iter} set to 300. It is clear that

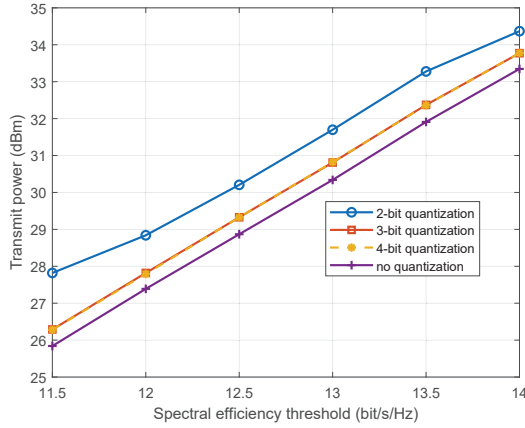


Fig. 5. Performance of TTS algorithm under different quantization bits.

our proposed method and all four other benchmarks outperform the *random amplitude* baseline significantly. Additionally, the transmit power achieved by the proposed TTS algorithm is also smaller than that of the *OTS optimization*. This is because the OTS benchmark only leverages instantaneous CSI in the first short time slot in the whole long time interval, while our proposed method utilizes both instantaneous CSI and statistical CSI in short-term and long-term optimization, respectively.

When compared to the *TTS optimization without channel sampling* benchmark, our proposed method shows a slight advantage, which illustrates the importance of the channel sampling step at the beginning of each iteration of CSSCA algorithm. Additionally, the proposed TTS algorithm also reveals a similar performance in comparison with the *AO with instantaneous CSI* benchmark, while our method has lower computational complexity and requires less signaling overhead than the benchmark, since the statistical CSI which is easier to acquire is leveraged to optimize the beam pattern instead of the instantaneous CSI.

As for the comparison with the *SDMA* benchmark in [13], our proposed method has better performance. Since the statistical CSI is utilized, the optimization problem (8) in our proposed TTS method has an expectation constrained form where the instantaneous spectral efficiency of users is not required to be larger than the threshold δ_u in every short time slot, allowing for efficient beamforming over different short time slots for the enhanced energy efficiency. In contrast, the SDMA benchmark is required to firstly satisfy the users' spectral efficiency constraints, leading to significant performance loss.

Moreover, due to the hardware resource constraints at both the BS and the UE side, the performance of our proposed method after μ -law quantization of the beam patterns α , $\mathbf{V}$ and the precoding matrix $\mathbf{W}$ with different numbers of quantization bits Q is also investigated in Fig. 5. We can conclude from Fig. 5 that our algorithm has good compability with the quantization operation. Specifically, when $Q = 2$, the quantization will lead to a considerable performance loss to the proposed algorithm. However, with an increase of Q , the performance gap between proposed scheme with or without quantization diminishes to a constant number. and is very close to that with no quantization.

V. CONCLUSION

In this paper, we focus on the co-design of the beam pattern and the precoding for the downlink multi-user HMIMOS-

assisted system. Since the conventional methods with instantaneous CSI suffers from high computational complexity and signaling overhead, we conceive a two-stage TTS algorithm, where the statistical CSI is used to optimize the beam patterns instead of the instantaneous CSI. However, both stages lead to intractable optimization problems, and thus the CSSCA algorithm is leveraged to solve the formulated non-convex optimization problem. Simulation results shows that our proposed method achieves a similar performance to the alternative optimization method leveraging the instantaneous CSI, and is attractive when compared to other baselines.

REFERENCES

- [1] R. Deng *et al.*, "Reconfigurable holographic surfaces for future wireless communications," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 126–131, Dec. 2021.
- [2] L. Wei *et al.*, "Multi-user holographic MIMO surfaces: channel modeling and spectral efficiency analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1112–1124, Aug. 2022.
- [3] B. Che, F. Meng, Y. Lyu, Y. Zhao, and Q. Wu, "Reconfigurable holographic antenna with low sidelobe level based on liquid crystals," *J. Phys. D, Appl. Phys.*, vol. 53, no. 31, pp. 302–315, May 2020.
- [4] O. Yurduseven, D. L. Marks, J. N. Gollub, and D. R. Smith, "Design and analysis of a reconfigurable holographic metasurface aperture for dynamic focusing in the Fresnel zone," *IEEE Access*, vol. 5, pp. 15055–15065, 2017.
- [5] X. Gan, C. Huang, Z. Yang, C. Zhong, and Z. Zhang, "Near-field localization for holographic RIS assisted mmWave systems," *IEEE Commun. Lett.*, vol. 27, no. 1, pp. 140–144, Jan. 2023.
- [6] R. Deng, B. Di, H. Zhang, Y. Tan, and L. Song, "Reconfigurable holographic surface-enabled multi-user wireless communications: amplitude-controlled holographic beamforming," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6003–6017, Aug. 2022.
- [7] R. Deng, B. Di, H. Zhang, H. V. Poor, and L. Song, "Holographic MIMO for LEO satellite communications aided by reconfigurable holographic surfaces," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 3071–3085, Oct. 2022.
- [8] X. Hu, R. Deng, B. Di, H. Zhang, and L. Song, "Holographic beamforming for ultra massive MIMO with limited radiation amplitudes: how many quantized bits do we need?," *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1403–1407, Jun. 2022.
- [9] Y. Chen, Y. Wang, and Z. Wang, "Reconfigurable intelligent surface aided high-mobility millimeter wave communications with dynamic dual-structured sparsity," *IEEE Trans. Wireless Commun.*, 2022.
- [10] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [11] I. Bahceci, M. Hasan, T. M. Duman and B. A. Cetiner, "Efficient channel estimation for reconfigurable MIMO antennas: training techniques and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 565–580, Jan. 2017.
- [12] M. Liang and A. Li, "Deep learning-based channel extrapolation for pattern reconfigurable massive MIMO," *arXiv preprint arXiv:2303.04432*, 2023.
- [13] R. Deng, B. Di, H. Zhang and L. Song, "HDMA: holographic-pattern division multiple access," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1317–1332, Apr. 2022.
- [14] A. Liu, V. K. N. Lau, and B. Kananian, "Stochastic successive convex approximation for non-convex constrained stochastic optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4189–4203, Aug. 2019.
- [15] X. -D. Zhang, *Matrix Analysis and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2017.
- [16] Y. Chen, Y. Wang, Z. Wang, and P. Zhang, "Robust beamforming for active reconfigurable intelligent omni-surface in vehicular communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 3086–3103, Oct. 2022.
- [17] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: <http://cvxr.com/cvx>.
- [18] Z. Sha, Y. Ming, C. Sun and Z. Wang, "Versatile resource management for millimeter-wave cellular network: near interference-free scheduling methodology," *IEEE Trans. Veh. Commun.*, early access, Feb. 03, 2023, doi: [10.1109/TVT.2023.3242069](https://doi.org/10.1109/TVT.2023.3242069).
- [19] M. -M. Zhao, A. Liu, Y. Wan, and R. Zhang, "Two-timescale beamforming optimization for intelligent reflecting surface aided multiuser communication with QoS constraints," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6179–6194, Sept. 2021.