Beamforming Design for the Performance Optimization of Intelligent Reflecting Surface Assisted Multicast MIMO Networks

Songling Zhang, *Student Member, IEEE*, Zhaohui Yang, *Member, IEEE*, Mingzhe Chen, *Member, IEEE*, Danpu Liu, *Senior Member, IEEE*, Kai-Kit Wong, *Fellow, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—In this paper, the problem of maximizing the sum of data rates of all users in an intelligent reflecting surface (IRS)-assisted millimeter wave multicast multiple-input multiple-output communication system is studied. In the considered model, one IRS is deployed to assist the communication from a multi-antenna base station (BS) to the multi-antenna users that are clustered into several groups. Our goal is to maximize the sum rate of all users by jointly optimizing the transmit beamforming matrices of the BS, the receive beamforming matrices of the users, and the phase shifts of the IRS. To solve this non-convex problem, we first use a block diagonalization method to represent the beamforming matrices of the BS and the users by the phase shifts of the IRS. Then, substituting the expressions of the beamforming matrices of the BS and the users, the original sum-rate maximization problem can be transformed into a problem that only needs to optimize the phase shifts of the IRS. To solve the transformed problem, a manifold method is used. Simulation results show that the proposed scheme can achieve up to 28.6% gain in terms of the sum rate of all users compared to the algorithm that optimizes the hybrid beamforming matrices of the BS and the users using our proposed scheme and randomly determines the phase shifts of the IRS.**

I. Introduction

Millimeter wave (mmWave) communications, which utilizes the 30-300 GHz frequency band to achieve multi-gigabit data rates, is a promising technology for emerging and envisioned wireless systems [2]–[5]. However, mmWave suffers from severe path loss and is easily blocked by obstacles due to the short wavelengths [6]. To address these problems, massive multiple-input multiple-output (MIMO) and intelligent reflecting surfaces (IRSs) have been proposed [7]–[11]. However, deploying IRSs and massive MIMO in mmWave communication systems faces several challenges such as IRS deployment optimization, and joint active and passive beamforming design.

Recently, a number of works have studied important problems related to the deployment of IRSs in wireless networks. The work in [12] considered the maximization of the spectral efficiency by separately designing the passive beamforming matrix and active precoder. The authors in [13] jointly designed a hybrid precoder at a base station (BS) and a passive precoder at the IRS to maximize the average spectral efficiency in an IRS-assisted mmWave MIMO system. To maximize the end-to-end signal-to-noise ratio (SNR), the authors in [14] optimized the phase shifts of the IRS. The work in [15] maximized the received signal power by jointly optimizing a transmit precoding vector of the BS and the phase shift coefficients of an IRS. The authors in [16] maximized the spectral efficiency by jointly optimizing the reflection coefficients of the IRS, a hybrid precoder at the BS and a hybrid combiner at the end-user device. The work in [17] studied hybrid precoding design for an IRS aided multi-user mmWave communication system. In [18], a geometric mean decomposition-based beamforming scheme was proposed for IRS-assisted mmWave hybrid MIMO systems. In [19], the authors optimized a channel estimator in closed form while considering the signal reflection matrix of an IRS and an analog combiner at the receiver. The authors in [20] jointly optimized the transmit beamforming vectors of multiple BSs and the reflective beamforming vector of the IRS so as to maximize the minimum weighted signal-to-interference-plus-noise ratio (SINR) of users. In [21], the authors studied the deployment of multiple IRSs to improve the spatial diversity gain and designed a robust beamforming scheme based on stochastic optimization methods to minimize the maximum outage probability among multiple users. The work in [22] jointly designed a hybrid precoder at the BS and the passive precoders at the IRSs to maximize the spectral efficiency. The authors in [23] investigated the use of double IRSs to improve the spectral efficiency in a multi-user MIMO network operating in the mmWave band. The work in [24] studied a double IRS assisted multi-user communication system with a cooperative passive beamforming design. The work in [25] introduced the opportunities and key challenges in designing holographic MIMO surfaces-enabled wireless communication systems. The authors in [26] developped an IRS assisted uplink

S. Zhang and D. Liu are with the Beijing Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, Beijing, 100876, China (e-mail: slzhang@bupt.edu.cn; dpliu@bupt.edu.cn).

Z. Yang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and Zhejiang Provincial Key Lab of Information Processing, Communication and Networking (IP-CAN), Hangzhou 310007, China, and also with Zhejiang Lab, Hangzhou 31121, China. (e-mail: yang_zhaohui@zju.edu.cn)

M. Chen is with the Department of Electrical and Computer Engineering and Institute for Data Science and Computing, University of Miami, Coral Gables, FL, 33146 USA (e-mail: mingzhe.chen@miami.edu).

K. K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London, WC1E 6BT, UK (e-mail: kai-kit.wong@ucl.ac.uk).

H. V. Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, 08544, USA (e-mail: poor@princeton.edu).

This work is supported in part by the National Key Research and Development Program of China under Grant 2018YFB2201700, the National Natural Science Foundation of China under Grant 61971069 and 62271065, the Beijing Natural Science Foundation under Grant L202003, the Project of China Railway Corporation under Grant N2022G048, the U.S National Science Foundation under Grant CNS-2128448, Zhejiang Lab under grant K2023QA0AL02, and Zhejiang Science and Technology Program under grant 2023C01021. A preliminary version of this work [1] is accepted by the Proceedings of the 2022 IEEE International Global Communications Conference (GLOBECOM). *(Corresponding author: Danpu Liu.)*

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3297732

massive MIMO system, where multiple IRSs are introduced to improve the uplink transmission rates of all users. The work in [27] investigated passive beamforming for multi-IRS aided mmWave multi-user multiple-input single-output systems. The authors in [28] investigated covert communication in an IRS-assisted non-orthogonal multiple access (NOMA) system, where a legitimate transmitter applies NOMA for downlink and uplink transmissions with a covert user and a public user aided by an IRS. The authors in [29] investigated the uplink cascaded channel estimation for IRS-assisted multiuser multiple-input-single-output systems. However, most of these existing works [12]–[29] only consider the deployment of IRSs over unicast communication networks in which each BS transmits independent data streams to different users.

Compared to unicast which requires the use of NOMA or rate splitting multiple access (RSMA) to transmit different content to different users using one radio resource block, multicast enables the BS to cluster the users into several groups according to their requested content and transmit it to the user in a group using one radio resource block, thus improving the spectral and energy efficiency [30]–[32]. However, deploying IRSs over multicast communication systems faces several new challenges. First, users in a group that have different channel conditions need to be served by a coordinated beamforming matrix, thus complicating the design of the beamforming matrix of the transmitter. Moreover, in a multicast system, the data rate of a group is limited by the user with the worst channel gain. Therefore, in a multicast system, one must maximize the data rate of the user with the worst channel gain in each group.

Several existing studies [33]–[38] have considered the use of IRSs in multicast communication systems. In particular, the work in [33] studied a multicast system where a single-antenna transmitter sends a common message to multiple single-antenna users via an IRS. In [34], the authors maximized the sum rate of all multicasting groups by the joint optimization of the precoding matrix at the base station and the reflection coefficients at the IRS under both power and unit-modulus constraints. The authors in [35] considered an IRS assisted multicast transmission scenario, where a BS with multiple antennas multicasts a common message to multiple single-antenna users under the assistance of an IRS. The work in [36] optimized the energy efficiency of an IRS-assisted multicast communication network. The work in [37] improved the robustness of an IRS assisted wireless multi-group multicast system. The authors in [38] jointly maximized the transmit beamforming matrix and IRS phase shifts so as to maximize the sum rate of all users. However, these existing works [33]–[38] neither considered mmWave nor the use of hybrid beamforming at the BS and the users. Considering mmWave and the use of hybrid beamforming at the BS and the users in an IRS-assisted multicast communication system faces several challenges such as severe path loss, joint analog and digital precoder design and optimization for a BS that uses mmWave and multicast techniques to serve users, and jointly optimizing the transmit beamforming matrices of the BS, the receive beamforming matrices of the users, and the phase shifts of the IRS.

The main contribution of this paper is to develop a novel IRS assisted multigroup multicast MIMO system. To the best of our knowledge, this is the first work that studies the joint use of an IRS, the hybrid beamforming at the BS and the users, the mmWave band, multicast, and MIMO to service the users in several groups. The key contributions are summarized as follows:

- We consider an IRS-assisted mmWave multicast MIMO communication system. In the considered model, one IRS is used to assist the communication from a multi-antenna BS to multi-antenna users that are clustered into several groups. To maximize the sum rate of all the multicasting groups, we jointly optimize the transmit beamforming matrices of the BS, the receive beamforming matrices of the users, and the phase shifts of the IRS. We formulate an optimization problem with the objective of maximizing the sum rate of all the multicasting groups under amplitude constraints on radio frequency (RF) beamforming matrices, maximum transmit power constraint, and unit-modulus constraint of the IRS phase shifts.
- To solve this problem, we first use a block diagonalization (BD) method to represent the beamforming matrices of the BS and the users in terms of the phase shifts of the IRS. Then, we substitute the expressions for the beamforming matrices of the BS and the users into the original problem so as to transform it to a problem that only needs to optimize the phase shifts of the IRS. The transformed problem is then solved by a manifold method.

Simulation results show that the proposed scheme can achieve up to 28.6% gain in terms of the sum rate of all the multicasting groups compared to the algorithm that optimizes the hybrid beamforming matrices of the BS and the users using our proposed algorithm and randomly determines the phase shifts of the IRS.

The rest of this paper is organized as follows. The system model and problem formulation are described in Section II. The algorithm is introduced in Section III. Simulation results are presented in Section IV. Conclusions are drawn in Section V.

II. System Model And Problem Formulation

A. System Model

We consider an IRS-aided mmWave multigroup multicast MIMO communication system in which a BS is equipped with N^{B} antennas serving K users via an IRS, as shown in Fig. 1. The users are divided into H groups. We assume that the users in a group will request the same data streams and the data streams requested by the users in different groups are different. Here, a data stream refers to a sequence of data and it implies that a user continuously requests data. The set of user groups is denoted by $\mathcal{H} = \{1, 2, \ldots, H\}$. Meanwhile, the set of users in a group h is denoted as $\mathcal{H}_h$. We also assume that each user can only belong to one group, i.e., $\mathcal{H}_i \cap \mathcal{H}_j = \emptyset$, $\forall i, j \in \mathcal{H}, i \neq j$. In our model, the direct communication link between the BS and a user is blocked due to unfavorable propagation conditions. Each user is equipped with N^{U} antennas and M^{U}

TABLE I: List of Main Notation.

Notation	Description	Notation	Description
N^{B}	Number of antennas of the BS	K	Number of users
H	Number of user groups	$\mathcal{H}$	The set of user groups
N^{U}	Number of antennas of each user	M^{U}	Number of RF chains of each user
ζ	Data streams received by each user	M^{B}	Number of RF chains of the BS
$\boldsymbol{F}^{\mathrm{B}}$	Baseband transmit beamforming matrix	$\boldsymbol{F}_h^{\mathrm{B}}$	Transmit beamforming matrix of group h
$\boldsymbol{F}^{\mathrm{R}}$	RF transmit beamforming matrix	$\boldsymbol{W}_k^{\mathrm{R}}$	RF receive beamforming matrix of user k
$\boldsymbol{W}_k^{\mathrm{B}}$	Baseband receive beamfoming matrix of user k	$\boldsymbol{\Phi}$	Phase-shift matrix of the IRS
M	Number of reflecting elements at the IRS	ϕ_m	Phase shift introduced by element m of the IRS
N	Number of antennas in ULA	d	Interval between two antennas
λ	Signal wavelength	F_y	Number of elements in the horizontal directions
F_z	Number of elements in the vertical directions	$\boldsymbol{H}^{\mathrm{B}}$	BS-IRS channel
$\boldsymbol{H}_k^{\mathrm{R}}$	Channel from the IRS to user k	Y	Total number of paths between the BS and the IRS
θ_i^{A}	Azimuth angle of arrival of the IRS	L	Total number of paths between the IRS and user k
θ_i^{D}	Azimuth angle of departure of the IRS	η_i^{A}	Elevation angle of arrival of the IRS
η_i^{D}	Elevation angle of departure of the IRS	$r_{i,k}^{\mathrm{A}}$	Arrival angle of user k
r_i^{D}	Departure angle of the BS	$\hat{\boldsymbol{s}}_{k,h}$	Detected data of user k in group h
$\boldsymbol{s}$	Data streams to be transmitted to all users	$\boldsymbol{a}\left(r_i^{\mathrm{D}}\right)$	Normalized array response vectors of the BS
$\boldsymbol{a}\left(r_{i,k}^{\mathrm{A}}\right)$	Normalized array response vectors of the user k	$\boldsymbol{H}_{h,k}$	Effective channel from the BS to user k in group h
$\boldsymbol{s}_h$	ζ streams to be transmitted to each user in group h	$\boldsymbol{n}_k$	Additive white Gaussian noise vector of user k
$\hat{s}_{ik,h}$	Estimated data stream i received by user k in group h	$\xi_{ik,h}$	SINR of user k in group h receiving data stream i
$I_{ik,h}$	Interference from other streams of user k	$J_{ik,h}$	Interference from other groups
$R_{k,h}$	Achievable data rate of user k in group h	P	transmit power of the BS
$\boldsymbol{B}$	Fully digital transmit beamforming matrix	$\boldsymbol{J}_k$	Fully digital receive beamforming matrix of user k in group h
p_i	Transmit power of data stream i	$\boldsymbol{G}_h$	Power allocation matrix in group h
$\nabla f\left(\boldsymbol{\nu}_n\right)$	Euclidean gradient	$\mathcal{Q}$	Oblique manifold
$T_{\boldsymbol{\nu}_n}\mathcal{Q}$	Tangent space	$\mathcal{G}_{\boldsymbol{\nu}_n}\mathcal{Q}$	Riemannian gradient

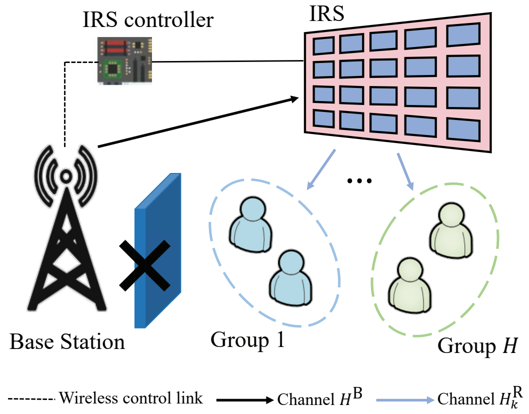


Fig. 1. An IRS-aided mmWave multigroup multicast MIMO communication system.

RF chains to receive ζ data streams from the BS. The BS simultaneously transmits $H\zeta$ independent data streams to the users by M^{B} RF chains with $H\zeta \leq M^{\mathrm{B}} \leq N^{\mathrm{B}}$ and $\zeta \leq M^{\mathrm{U}} \leq N^{\mathrm{U}}$. The main notations used in this work are summarized in Table I.

At the BS, the transmitted data streams of H user groups are precoded by a baseband transmit beamforming matrix $\boldsymbol{F}^{\mathrm{B}} = \left[\boldsymbol{F}_1^{\mathrm{B}}, \boldsymbol{F}_2^{\mathrm{B}}, \ldots, \boldsymbol{F}_H^{\mathrm{B}}\right] \in \mathbb{C}^{M^{\mathrm{B}} \times H\zeta}$, with $\boldsymbol{F}_h^{\mathrm{B}}$ being the transmit beamforming matrix of group h. After that, each transmitted data stream of H user groups is precoded by an RF transmit beamforming matrix $\boldsymbol{F}^{\mathrm{R}} \in \mathbb{C}^{N^{\mathrm{B}} \times M^{\mathrm{B}}}$. The received data streams of user k in group h are first processed by an RF receive beamforming matrix $\boldsymbol{W}_k^{\mathrm{R}} \in \mathbb{C}^{N^{\mathrm{U}} \times M^{\mathrm{U}}}$. Then, user k uses a baseband receive beamfoming matrix $\boldsymbol{W}_k^{\mathrm{B}} \in \mathbb{C}^{M^{\mathrm{U}} \times \zeta}$ to recover ζ data streams. In our model, an IRS is used to enhance the received signal strength of users by reflecting signals from the BS to the users. We assume that the power of the signals that are reflected by the scatters more than once before reaching the IRS is ignored due to severe path loss. The phase-shift matrix of the IRS is $\boldsymbol{\Phi} = \operatorname{diag}\left(e^{j\phi_1}, \ldots, e^{j\phi_M}\right) \in \mathbb{C}^{M \times M}$, where $\operatorname{diag}\left(e^{j\phi_1}, \ldots, e^{j\phi_M}\right)$ is a diagonal matrix of $\left[e^{j\phi_1}, \ldots, e^{j\phi_M}\right]$, M is the number of reflecting elements at the IRS, and $\phi_m \in [0, 2\pi]$ is the phase shift introduced by element m of the IRS.

1) Channel Model: The BS and the users employ uniform linear arrays (ULAs), and the IRS uses a uniform planar array (UPA). The normalized array response vector for an ULA is

$$\boldsymbol{a}\left(r\right) = \frac{1}{\sqrt{N}}\left[1, \cdots, e^{j\frac{2\pi d}{\lambda}(n-1)\sin(r)}, \cdots, e^{j\frac{2\pi d}{\lambda}(N-1)\sin(r)}\right]^{\mathrm{T}}, \tag{1}$$

where r is the angle of arrival signal and N is the number of antennas in ULA, d is an interval between two antennas, and λ is the signal wavelength. The normalized array response vector of UPA is

$$\begin{aligned}\boldsymbol{a}\left(\theta, \eta\right) = & \frac{1}{\sqrt{F_y \times F_z}}[1, \cdots, e^{j\frac{2\pi d}{\lambda}((f_1-1)\cos(\eta)\sin(\theta)+(f_2-1)\sin(\eta))}, \\ & \cdots, e^{j\frac{2\pi d}{\lambda}((F_y-1)\cos(\eta)\sin(\theta)+(F_z-1)\sin(\eta))}]^{\mathrm{T}},\end{aligned} \tag{2}$$

where θ is the azimuth angle of arrival signals and η is the elevation angle of arrival signals, F_y and F_z are respectively the number of elements in the horizontal and vertical directions, and $F_y \times F_z$ is the number of elements in UPA. The BS-IRS channel $\boldsymbol{H}^{\mathrm{B}} \in \mathbb{C}^{M \times N^{\mathrm{B}}}$ and the channel $\boldsymbol{H}_k^{\mathrm{R}} \in \mathbb{C}^{N^{\mathrm{U}} \times M}$ from the IRS to user k in group h can be respectively given as

$$\boldsymbol{H}^{\mathrm{B}} = \sqrt{\frac{N^{\mathrm{B}}M}{Y}} \sum_{i=1}^{Y} \alpha_i \boldsymbol{a}\left(\theta_i^{\mathrm{A}}, \eta_i^{\mathrm{A}}\right) \left(\boldsymbol{a}\left(r_i^{\mathrm{D}}\right)\right)^{\mathrm{H}}, \tag{3}$$

$$\boldsymbol{H}_k^{\mathrm{R}} = \sqrt{\frac{MN^{\mathrm{U}}}{L}} \sum_{i=1}^{L} \beta_i \boldsymbol{a}\left(r_{i,k}^{\mathrm{A}}\right) \left(\boldsymbol{a}\left(\theta_i^{\mathrm{D}}, \eta_i^{\mathrm{D}}\right)\right)^{\mathrm{H}}, \tag{4}$$

where Y is the total number of paths (line-of-sight (LOS) and non-line-of-sight (NLOS)) between the BS and the IRS, L is the total number of paths (LOS and NLOS) between the IRS and user k, θ_i^{A} denotes the azimuth angle of arrival of the IRS, θ_i^{D} denotes the azimuth angle of departure of the IRS, η_i^{A} denotes the elevation angle of arrival of the IRS, η_i^{D} denotes the elevation angle of departure of the IRS, $r_{i,k}^{\mathrm{A}}$ represents the arrival angle of user k, r_i^{D} represents the departure angle of the BS, α_i and β_i are complex channel gains. $\boldsymbol{a}\left(r_i^{\mathrm{D}}\right)$ and $\boldsymbol{a}\left(r_{i,k}^{\mathrm{A}}\right)$ denote the normalized array response vectors of the BS and user k, respectively. $\left(\boldsymbol{a}\left(r_i^{\mathrm{D}}\right)\right)^{\mathrm{H}}$ is the Hermitian transpose of $\boldsymbol{a}\left(r_i^{\mathrm{D}}\right)$. $\boldsymbol{a}\left(\theta_i^{\mathrm{A}}, \eta_i^{\mathrm{A}}\right)$ represents the normalized array response vector of the IRS over the effective channel from the BS to the IRS. $\boldsymbol{a}\left(\theta_i^{\mathrm{D}}, \eta_i^{\mathrm{D}}\right)$ represents the normalized array response vector of the IRS over the effective channel from the IRS to user k. The effective channel from the BS to user k in group h is $\boldsymbol{H}_{h,k} = G_{\mathrm{t}} G_{\mathrm{r}} \boldsymbol{H}_k^{\mathrm{R}} \boldsymbol{\Phi} \boldsymbol{H}^{\mathrm{B}}$, where G_{t} and G_{r} are the antenna gains of the BS and each user, respectively.

2) Data Rate Model: The BS obtains channel state information (CSI) by channel estimation methods such as compressive sensing [39] and IRS-elements grouping method [40]. The BS is responsible for designing the reflection coefficients of the IRS. As a result, the detected data of user k in group h is given by

$$\hat{\boldsymbol{s}}_{k,h} = \left(\boldsymbol{W}_k^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{H}_{h,k} \boldsymbol{F}^{\mathrm{R}} \boldsymbol{F}^{\mathrm{B}} \boldsymbol{s} + \left(\boldsymbol{W}_k^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{n}_k, \tag{5}$$

where $\boldsymbol{s} = \left[\boldsymbol{s}_1^{\mathrm{T}}, \ldots, \boldsymbol{s}_H^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{H\zeta \times 1}$ represents the data streams to be transmitted to all users, with $\boldsymbol{s}_h = \left[s_{h,1}, \ldots, s_{h,\zeta}\right]^{\mathrm{T}} \in \mathbb{C}^{\zeta \times 1}$ being ζ streams that will be transmitted to each user in group h. $\boldsymbol{n}_k \in \mathbb{C}^{N^{\mathrm{U}} \times 1}$ is an additive white Gaussian noise vector of user k. Each element of $\boldsymbol{n}_k$ follows the independent and identically distributed complex Gaussian distribution with zero mean and variance σ^2. In (5), the first term represents the signal received by user k. The second term is the noise received by user k. The estimated data stream i received by user k in group h is

$$\begin{aligned} \hat{s}_{ik,h} = & \left(\boldsymbol{w}_{k,i}^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{H}_{h,k} \boldsymbol{F}^{\mathrm{R}} \bar{\boldsymbol{f}}_{h_i}^{\mathrm{B}} s_{h,i} \\ & + \sum_{j=1, j\neq i}^{\zeta} \left(\boldsymbol{w}_{k,i}^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{H}_{h,k} \boldsymbol{F}^{\mathrm{R}} \bar{\boldsymbol{f}}_{h_j}^{\mathrm{B}} s_{h,j} \\ & + \sum_{m=1, m\notin \mathcal{H}_h}^{H} \sum_{l=1}^{\zeta} \left(\boldsymbol{w}_{k,i}^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{H}_{h,k} \boldsymbol{F}^{\mathrm{R}} \bar{\boldsymbol{f}}_{m_l}^{\mathrm{B}} s_{m,l} \\ & + \left(\boldsymbol{w}_{k,i}^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{n}_k, \end{aligned} \tag{6}$$

where $h_i = (h-1)\zeta + i$, $\boldsymbol{w}_{k,i}^{\mathrm{B}}$ denotes row i of matrix $\boldsymbol{W}_k^{\mathrm{B}}$, and $\bar{\boldsymbol{f}}_{h_i}^{\mathrm{B}}$ denotes column h_i of matrix $\boldsymbol{F}^{\mathrm{B}}$. In (6), the first term represents the desired signal. The second term is the interference caused by other streams of user k. The third term is the interference caused by the users from other groups. The fourth term is the noise. The SINR of user k in group h receiving data stream i is

$$\xi_{ik,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}_h^{\mathrm{B}}\right) = \frac{\left|\left(\boldsymbol{w}_{k,i}^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{H}_{h,k} \boldsymbol{F}^{\mathrm{R}} \bar{\boldsymbol{f}}_{h_i}^{\mathrm{B}}\right|^2}{I_{ik,h} + J_{ik,h} + \sigma^2}, \tag{7}$$

where $I_{ik,h}$ is short for $I_{ik,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}_h^{\mathrm{B}}\right)$ and $I_{ik,h} = \sum_{j=1, j\neq i}^{\zeta} \left|\left(\boldsymbol{w}_{k,i}^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{H}_{h,k} \boldsymbol{F}^{\mathrm{R}} \bar{\boldsymbol{f}}_{h_j}^{\mathrm{B}}\right|^2$ represents the interference from other streams of user k, $J_{ik,h}$ is short for $J_{ik,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}_h^{\mathrm{B}}\right)$ and $J_{ik,h} = \sum_{m=1, m\notin \mathcal{H}_h}^{H} \sum_{l=1}^{\zeta} \left|\left(\boldsymbol{w}_{k,i}^{\mathrm{B}}\right)^{\mathrm{H}} \left(\boldsymbol{W}_k^{\mathrm{R}}\right)^{\mathrm{H}} \boldsymbol{H}_{h,k} \boldsymbol{F}^{\mathrm{R}} \bar{\boldsymbol{f}}_{m_l}^{\mathrm{B}}\right|^2$ represents the interference from other groups. The achievable data rate of user k in group h is given by

$$\begin{aligned} & R_{k,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}_h^{\mathrm{B}}\right) \\ & = W \sum_{i=1}^{\zeta} \log_2\left(1 + \xi_{ik,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}_h^{\mathrm{B}}\right)\right), \end{aligned} \tag{8}$$

where W is the bandwidth.

Due to the nature of the multicast mechanism, the achievable data rate of group h depends on the user with minimum data rate, which is defined as

$$\min_{k \in \mathcal{H}_h} \left\{R_{k,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}_h^{\mathrm{B}}\right)\right\}. \tag{9}$$

B. Problem Formulation

Next, we introduce our optimization problem. Our goal is to maximize the sum rate of all the multicasting groups via jointly optimizing the transmit beamforming matrices $\boldsymbol{F}^{\mathrm{B}}$, $\boldsymbol{F}^{\mathrm{R}}$, the receive beamforming matrices $\boldsymbol{W}^{\mathrm{R}}$, $\boldsymbol{W}^{\mathrm{B}}$, and the phase shift $\boldsymbol{\nu}$ of the IRS. Mathematically, the optimization problem is formulated as

$$\max_{\boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}^{\mathrm{B}}, \boldsymbol{\nu}} \sum_{h=1}^{H} \min_{k \in \mathcal{H}_h} \left\{R_{k,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}_h^{\mathrm{B}}\right)\right\} \tag{10}$$

$$\text{s.t.} \quad \left\|\boldsymbol{F}^{\mathrm{R}} \boldsymbol{F}^{\mathrm{B}}\right\|_F^2 \leq P, \tag{10a}$$

$$\left|\boldsymbol{F}^{\mathrm{R}}(i,j)\right| = \left|\boldsymbol{W}_k^{\mathrm{R}}(i,j)\right| = 1, \forall i, j, \tag{10b}$$

$$0 \leq \phi_m \leq 2\pi, m = 1, \ldots, M, \tag{10c}$$

where P is the transmit power of the BS, $\left\|\boldsymbol{F}^{\mathrm{R}} \boldsymbol{F}^{\mathrm{B}}\right\|_F$ is the Frobenius norm of $\boldsymbol{F}^{\mathrm{R}} \boldsymbol{F}^{\mathrm{B}}$, $\boldsymbol{\nu} = \left[e^{j\phi_1}, \ldots, e^{j\phi_M}\right]^{\mathrm{H}}$, $\boldsymbol{F}^{\mathrm{R}}(i,j)$ denotes the element (i,j) of matrix $\boldsymbol{F}^{\mathrm{R}}$, with $\left|\boldsymbol{F}^{\mathrm{R}}(i,j)\right|$ being the amplitude of $\boldsymbol{F}^{\mathrm{R}}(i,j)$. The transmit power constraint of the BS is given in (10a). Constraint (10b) represents the amplitude constraints of the RF beamforming matrices of the BS and each user, while (10c) shows the phase shift limits of the IRS. Problem (10) cannot be solved directly due to two key challenges. First, the optimization variables (i.e., the digital beamforming vectors $\boldsymbol{F}^{\rm B}$ and $\boldsymbol{W}_k^{\rm B}$ and the analog beamforming vectors $\boldsymbol{F}^{\rm R}$ and $\boldsymbol{W}_k^{\rm R}$) in problem (10) are dependent. Second, the rate function $R_{k,h}\left(\boldsymbol{W}_k^{\rm R}, \boldsymbol{W}_k^{\rm B}, \boldsymbol{\nu}, \boldsymbol{F}^{\rm R}, \boldsymbol{F}_h^{\rm B}\right)$ in (10) is non-convex with respect to $\boldsymbol{F}^{\rm B}$, $\boldsymbol{W}_k^{\rm B}$, $\boldsymbol{F}^{\rm R}$, and $\boldsymbol{W}_k^{\rm R}$. Therefore, the complexity of using standard optimization algorithms to directly solve problem (10) is very high. Next, we introduce an efficient scheme to solve problem (10).

III. Proposed Scheme

Next, we first use the phase shift $\boldsymbol{\nu}$ of the IRS to represent the fully digital transmit beamforming matrix of the BS and receive beamforming matrix of the users. Then, we substitute them in (10) to transform problem (10). To solve the transformed problem, the phase shift $\boldsymbol{\nu}$ of the IRS is optimized by a manifold method. Finally, we introduce the entire algorithm used to solve problem (10).

A. Block Diagonalization Method

1) Simplification of Optimization Problem: Since $\boldsymbol{F}^{\rm R}\boldsymbol{F}^{\rm B}$ is a combination of baseband digital beamformer and analog beamformer, it represents the transmit beamforming matrix of the BS. Similarly, $\boldsymbol{W}_k^{\rm B}\boldsymbol{W}_k^{\rm R}$ is a combination of baseband digital combiner and analog combiner, and represents the effective receive beamforming matrix of user k. Therefore, if we consider $\boldsymbol{W}_k^{\rm B}\boldsymbol{W}_k^{\rm R}$ and $\boldsymbol{F}^{\rm R}\boldsymbol{F}^{\rm B}$ as a whole, problem (10) is a problem of fully digital beamforming. Once the fully digital transmit beamforming matrix and receive beamforming matrix are obtained, we can use the algorithm in [41] to find the hybrid transmit beamforming matrices and receive beamforming matrices to approximate the fully digital transmit beamforming matrix and receive beamforming matrix, as done in [42], [43]. The motivation for using hybrid beamforming algorithm instead of fully digital beamforming algorithm is that hybrid beamforming algorithms have lower hardware implementation complexity and power consumption compared to fully digital beamforming algorithms due to the specific design of the radio frequency chains and power amplifiers, and the use of simplified transmitter/receiver structures. To this end, the goal of our work is to design a novel hybrid beamforming algorithm that can reach the same performance as that of the fully digital beamforming algorithm but with lower hardware implementation complexity and power consumption. Let $\boldsymbol{B} = \boldsymbol{F}^{\rm R}\boldsymbol{F}^{\rm B} = [\boldsymbol{B}_1, \ldots, \boldsymbol{B}_H] \in \mathbb{C}^{N^{\rm B}\times H\zeta}$ be a fully digital transmit beamforming matrix and $\boldsymbol{J}_k = \boldsymbol{W}_k^{\rm R}\boldsymbol{W}_k^{\rm B} \in \mathbb{C}^{N^{\rm U}\times\zeta}$ be a fully digital receive beamforming matrix of user k in group h. Here, assuming $\boldsymbol{B} = \boldsymbol{F}^{\rm R}\boldsymbol{F}^{\rm B}$ can reduce the number of optimization variables in problem (10) and remove the coupling relationship between $\boldsymbol{F}^{\rm B}, \boldsymbol{W}_k^{\rm B}$, and $\boldsymbol{F}^{\rm R}, \boldsymbol{W}_k^{\rm R}$. Substituting $\boldsymbol{B}$ and $\boldsymbol{J}_k$ in (10), problem (10) can be transformed as

$$\max_{\boldsymbol{B},\boldsymbol{J},\boldsymbol{\nu}} \quad \sum_{h=1}^{H} \min_{k\in\mathcal{H}_h} \left\{R_{k,h}\left(\boldsymbol{B}_h, \boldsymbol{J}_k, \boldsymbol{\nu}\right)\right\} \tag{11}$$

$$\text{s.t.} \quad (10\text{c}),$$

$$\|\boldsymbol{B}\|_{\rm F}^2 \le P, \tag{11a}$$

where $\boldsymbol{J} = \operatorname{diag}\left(\boldsymbol{J}_1, \ldots, \boldsymbol{J}_K\right)$. In problem (11), we use a fully digital transmit beamforming matrix $\boldsymbol{B}$ to represent $\boldsymbol{F}^{\rm R}\boldsymbol{F}^{\rm B}$. Therefore, constraint (10a) in problem (10) is converted to constraint (11a) in problem (11). With regard to constraint (10b), since $\boldsymbol{B}$ is a fully digital transmit beamforming matrix, it does not have amplitude constraints and hence we remove constraint (10b) in problem (11).

2) Optimization of $\boldsymbol{B}$ and $\boldsymbol{J}$: Due to the low complexity of a BD method, we use it to find the relationship between $\boldsymbol{\nu}$ and the fully digital transmit beamforming matrix $\boldsymbol{B}$ of the BS as well as the receive beamforming matrix $\boldsymbol{J}$ of the users.

Lemma 1: Given $\boldsymbol{\nu}$ and the power allocation matrix $\boldsymbol{G}_h = \operatorname{diag}\left(p_1, \ldots, p_\zeta\right)$ in group h, where $p_i = \frac{P}{H\zeta}$ is the transmit power of data stream i, $\boldsymbol{B}$ and $\boldsymbol{J}$ can be given by

$$\boldsymbol{B}_h\left(\boldsymbol{\nu}\right) = \tilde{\boldsymbol{V}}_h^{(0)}\left(\frac{\boldsymbol{V}_1^{(1)} + \cdots + \boldsymbol{V}_K^{(1)}}{\sqrt{|\mathcal{H}_h|}}\right)\sqrt{\frac{P}{H\zeta}}, \tag{12}$$

$$\boldsymbol{J}_k\left(\boldsymbol{\nu}\right) = \boldsymbol{U}_k^{(1)}, \tag{13}$$

where $\tilde{\boldsymbol{V}}_h^{(0)} = \operatorname{null}\left(\tilde{\boldsymbol{H}}_h\right)$, $\boldsymbol{U}_k^{(1)}$, and $\boldsymbol{V}_k^{(1)}$ can be obtained by singular value decomposition (SVD) of $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}$ with $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)} = \left[\boldsymbol{U}_k^{(1)}, \boldsymbol{U}_k^{(2)}\right]\begin{bmatrix}\boldsymbol{\Sigma}_k^{(1)} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_k^{(2)}\end{bmatrix}\left[\boldsymbol{V}_k^{(1)}, \boldsymbol{V}_k^{(2)}\right]^{\rm H}$.

Proof: To prove Lemma 1, we first define $\tilde{\boldsymbol{H}}_h$ as

$$\tilde{\boldsymbol{H}}_h \triangleq \left[\bar{\boldsymbol{H}}_1, \ldots, \bar{\boldsymbol{H}}_{h-1}, \bar{\boldsymbol{H}}_{h+1}, \ldots, \bar{\boldsymbol{H}}_H\right]^{\rm T}, \tag{14}$$

where $\bar{\boldsymbol{H}}_{h-1} = \left[\boldsymbol{H}_{h-1,1}, \boldsymbol{H}_{h-1,2}, \ldots, \boldsymbol{H}_{h-1,E_{h-1}}\right]$ is a matrix of the effective channels of all users in group $h-1$ with E_{h-1} being the number of users in group $h-1$. We assume that the rank of $\tilde{\boldsymbol{H}}_h$ is $\tilde{L}_k$. Next, we introduce the use of BD method to represent the transmit beamforming matrices of the BS and the receive beamforming matrices of the users by the phase shifts of the IRS. To eliminate the inter-group interference, we define $\tilde{\boldsymbol{V}}_h^{(0)} \in \mathbb{C}^{N^{\rm B}\times\left(N^{\rm B}-\tilde{L}_k\right)}$ as

$$\tilde{\boldsymbol{V}}_h^{(0)} = \operatorname{null}\left(\tilde{\boldsymbol{H}}_h\right), \tag{15}$$

where $\operatorname{null}\left(\tilde{\boldsymbol{H}}_h\right)$ represents that $\tilde{\boldsymbol{V}}_h^{(0)}$ lies in the null space of $\tilde{\boldsymbol{H}}_h$. Hence, we have $\tilde{\boldsymbol{H}}_h\tilde{\boldsymbol{V}}_h^{(0)} = \boldsymbol{0}$. The interference among multiple streams of each user can be eliminated by the SVD of $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}$, which is

$$\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)} = \left[\boldsymbol{U}_k^{(1)}, \boldsymbol{U}_k^{(2)}\right]\begin{bmatrix}\boldsymbol{\Sigma}_k^{(1)} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_k^{(2)}\end{bmatrix}\left[\boldsymbol{V}_k^{(1)}, \boldsymbol{V}_k^{(2)}\right]^{\rm H}. \tag{16}$$

We assume that the rank of $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}$ is L_k, the column vectors of $\boldsymbol{U}_k^{(1)} \in \mathbb{C}^{N^{\rm U}\times\zeta}$, $\boldsymbol{U}_k^{(2)} \in \mathbb{C}^{N^{\rm U}\times(L_k-\zeta)}$, $\boldsymbol{V}_k^{(1)} \in \mathbb{C}^{\left(N^{\rm B}-\tilde{L}_k\right)\times\zeta}$, and $\boldsymbol{V}_k^{(2)} \in \mathbb{C}^{\left(N^{\rm B}-\tilde{L}_k\right)\times(L_k-\zeta)}$ can form orthonormal sets, $\boldsymbol{\Sigma}_k^{(1)} \in \mathbb{C}^{\zeta\times\zeta}$ and $\boldsymbol{\Sigma}_k^{(2)} \in \mathbb{C}^{(L_k-\zeta)\times(L_k-\zeta)}$ are diagonal matrices of singular values. $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ must be designed to cancel the inter-group interference. Thus, $\tilde{\boldsymbol{V}}_h^{(0)}$ and $\sum\limits_{i=1}^{K}\boldsymbol{V}_i^{(1)}$ must be included in $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$, which can be given by

$$\boldsymbol{B}_h\left(\boldsymbol{\nu}\right) = \tilde{\boldsymbol{V}}_h^{(0)}\left(\frac{\boldsymbol{V}_1^{(1)} + \cdots + \boldsymbol{V}_K^{(1)}}{\sqrt{|\mathcal{H}_h|}}\right)\boldsymbol{G}_h^{1/2}, \tag{17}$$

where $|\mathcal{H}_h|$ is the number of users in group h, $\frac{1}{\sqrt{|\mathcal{H}_h|}}$ ensures that the power of $\left(\frac{\boldsymbol{V}_1^{(1)}+\cdots+\boldsymbol{V}_K^{(1)}}{\sqrt{|\mathcal{H}_h|}}\right)$ is unit, and $\boldsymbol{G}_h$ is the power allocation matrix in group h. To eliminate the interference among multiple streams of group h, the fully digital receive beamforming matrix $\boldsymbol{J}_k$ of user k in group h is written as

$$\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)=\boldsymbol{U}_k^{(1)}. \tag{18}$$

Substituting $\boldsymbol{G}_h$ into (17), we have

$$\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)=\tilde{\boldsymbol{V}}_h^{(0)}\left(\frac{\boldsymbol{V}_1^{(1)}+\cdots+\boldsymbol{V}_K^{(1)}}{\sqrt{|\mathcal{H}_h|}}\right)\sqrt{\frac{P}{H\zeta}}. \tag{19}$$

This completes the proof. ■

From Lemma 1, we can see that $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ mainly depends on the orthogonal bases of the null space of users in other groups, the orthogonal bases of the subspace of users in group h, the maximum transmit power of the BS and number of groups, $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$ depends on the effective channel of user k and the orthogonal bases of the null space of users in other groups.

3) Simplification of Problem (11): Based on the Lemma 1, the interference caused by other groups, $J_{ik,h}$ and other streams of user k, $I_{ik,h}$ can be eliminated by the fully digital transmit beamforming matrix $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ and receive beamforming matrix $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$. Substituting $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ and $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$ into (11), the achievable data rate of user k in group h can be rewritten as follows:

$$\begin{aligned}&R_{k,h}\left(\boldsymbol{B}_h\left(\boldsymbol{\nu}\right),\boldsymbol{J}_k\left(\boldsymbol{\nu}\right),\boldsymbol{\nu}\right)=\\&\quad W\sum_{i=1}^{\zeta}\log_2\left(1+\left|\left(\boldsymbol{j}_{k,i}\left(\boldsymbol{\nu}\right)\right)^{\mathrm{H}}\boldsymbol{H}_{h,k}\bar{\boldsymbol{b}}_{h,i}\left(\boldsymbol{\nu}\right)\right|^2/\sigma^2\right),\end{aligned} \tag{20}$$

Let $\frac{\left|\left(\boldsymbol{j}_{k,i}\left(\boldsymbol{\nu}\right)\right)^{\mathrm{H}}\boldsymbol{H}_{h,k}\bar{\boldsymbol{b}}_{h,i}\left(\boldsymbol{\nu}\right)\right|^2}{\sigma^2}=\lambda_i$, (20) can be rewritten by

$$\begin{aligned}&R_{k,h}\left(\boldsymbol{B}_h\left(\boldsymbol{\nu}\right),\boldsymbol{J}_k\left(\boldsymbol{\nu}\right),\boldsymbol{\nu}\right)\\&=W\sum_{i=1}^{\zeta}\log_2\left(1+\lambda_i\right),\\&=W\log_2\left(1+\lambda_1\right)+\ldots+\log_2\left(1+\lambda_\zeta\right),\\&=W\log_2\left(\left(1+\lambda_1\right)*\ldots*\left(1+\lambda_\zeta\right)\right),\\&=W\log_2\begin{vmatrix}1+\lambda_1&\cdots&0\\\vdots&\ddots&\vdots\\0&\cdots&1+\lambda_\zeta\end{vmatrix}.\end{aligned} \tag{21}$$

Therefore, we have

$$\begin{aligned}&R_{k,h}\left(\boldsymbol{B}_h\left(\boldsymbol{\nu}\right),\boldsymbol{J}_k\left(\boldsymbol{\nu}\right),\boldsymbol{\nu}\right)\\&=W\log_2\det\left(\boldsymbol{I}_\zeta+\left|\left(\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)\right)^{\mathrm{H}}\boldsymbol{H}_{h,k}\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)\right|^2/\sigma^2\right),\end{aligned} \tag{22}$$

where $\det(\cdot)$ represents the determinant of a square matrix, and $\boldsymbol{I}_\zeta$ is an $\zeta\times\zeta$ identity matrix. Substituting $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ and $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$ into (22), we have

$$\begin{aligned}&R_{k,h}\left(\boldsymbol{B}_h\left(\boldsymbol{\nu}\right),\boldsymbol{J}_k\left(\boldsymbol{\nu}\right),\boldsymbol{\nu}\right)\\&=W\log_2\det\left(\boldsymbol{I}_\zeta+\frac{\left|\left(\boldsymbol{U}_k^{(1)}\right)^{\mathrm{H}}\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}\left(\frac{\sum_{i=1}^{K}\boldsymbol{V}_i^{(1)}}{\sqrt{|\mathcal{H}_h|}}\right)\sqrt{\frac{P}{H\zeta}}\right|^2}{\sigma^2}\right).\end{aligned} \tag{23}$$

Since $\boldsymbol{H}_i\left(\boldsymbol{H}_{h,k}\right)^{\mathrm{H}}=\mathbf{0}$ $(i\neq k)$, we have $\boldsymbol{H}_{h,k}\left(\boldsymbol{V}_i^{(1)}\right)^{\mathrm{H}}=\mathbf{0}$ $(i\neq k)$. Substituting (16) into (23), we have

$$\begin{aligned}&R_{k,h}\left(\boldsymbol{B}_h\left(\boldsymbol{\nu}\right),\boldsymbol{J}_k\left(\boldsymbol{\nu}\right),\boldsymbol{\nu}\right)\\&=W\log_2\det\left(\boldsymbol{I}_\zeta+\frac{P}{|\mathcal{H}_h|H\zeta\sigma^2}\left(\boldsymbol{\Sigma}_k^{(1)}\right)^2\right),\end{aligned} \tag{24}$$

Then, the optimization problem in (11) can be transformed as

$$\begin{aligned}&\max_{\boldsymbol{\nu}}\sum_{h=1}^{H}\min_{k\in\mathcal{H}_h}\left\{W\log_2\det\left(\boldsymbol{I}_\zeta+\frac{P}{|\mathcal{H}_h|H\zeta\sigma^2}\left(\boldsymbol{\Sigma}_k^{(1)}\right)^2\right)\right\}\\&\quad\text{s.t.}\quad(10\text{c}).\end{aligned} \tag{25}$$

B. Phase Optimization with Manifold Method

1) Approximation of $\boldsymbol{\Sigma}_k^{(1)}$: Since $\boldsymbol{\Sigma}_k^{(1)}$ in (25) is unknown, we use the function of phase shift to represent $\boldsymbol{\Sigma}_k^{(1)}$, which is proved in Theorem 1.

Theorem 1: $\boldsymbol{\Sigma}_k^{(1)}(i,j)\approx\beta_i\alpha_j\boldsymbol{\nu}^{\mathrm{H}}\boldsymbol{c}^{ij}$, where $\boldsymbol{c}^{ij}=\left(\boldsymbol{a}\left(\theta_i^{\mathrm{D}},\eta_i^{\mathrm{D}}\right)\right)^*\circ\boldsymbol{a}\left(\theta_j^{\mathrm{A}},\eta_j^{\mathrm{A}}\right)$ with $\left(\boldsymbol{a}\left(\theta_i^{\mathrm{D}},\eta_i^{\mathrm{D}}\right)\right)^*$ being the conjugate of $\left(\boldsymbol{a}\left(\theta_i^{\mathrm{D}},\eta_i^{\mathrm{D}}\right)\right)$ and $\circ$ being the Hadamard product.

Proof: See Appendix A. ■

From Theorem 1, we can see that $\boldsymbol{\Sigma}_k^{(1)}$ depends on the distance α_j between the BS and the IRS, the distance β_i between the IRS and user k, the angle $\boldsymbol{a}\left(\theta_j^{\mathrm{A}},\eta_j^{\mathrm{A}}\right)$ from the BS to the IRS, the phase shifts of the IRS, and the angle $\boldsymbol{a}\left(\theta_i^{\mathrm{D}},\eta_i^{\mathrm{D}}\right)$ from the IRS to user k.

2) Problem Transformation: Based on Theorem 1, the optimization problem (25) can be rewritten as

$$\max_{\boldsymbol{\nu}}\sum_{h=1}^{H}\min_{k\in\mathcal{H}_h}W\left\{\sum_{i=1}^{\zeta}\log_2\left(1+\frac{P}{|\mathcal{H}_h|H\zeta\sigma^2}\left|\boldsymbol{D}_k\left(i,i\right)\right|^2\right)\right\} \tag{26}$$

$$\begin{aligned}&\text{s.t.}\quad(10\text{c}),\\&\qquad|d_{ij}|=\left|\boldsymbol{\nu}^H\boldsymbol{c}^{ij}\right|<\tau,\quad\forall i\neq j,\end{aligned} \tag{26a}$$

where $d_{ii}=\boldsymbol{\nu}^H\boldsymbol{c}^{ii}$, $\boldsymbol{D}_k\left(i,i\right)=\alpha_i\beta_i d_{ii}$ $(i\in\{1,\ldots,\zeta\})$, and τ is a small positive value. Constraint (26a) is to make sure that $\boldsymbol{D}_k$ is approximately a non-square diagonal matrix such that $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}=\boldsymbol{A}_k\boldsymbol{D}_k\left(\boldsymbol{A}\right)^{\mathrm{H}}\left[\boldsymbol{z}_{(K-|\mathcal{H}_h|)\zeta+1};\ldots;\boldsymbol{z}_{K\zeta}\right]$ can be treated as an approximation of the truncated SVD of $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}$, where $\boldsymbol{z}_{(K-|\mathcal{H}_h|)\zeta+1}$ denotes row $\left(K-|\mathcal{H}_h|\right)\zeta+1$ of matrix $\boldsymbol{Z}$. Constraint (26a) can be removed and this omission does not affect the validity of our proposed solution [16]. Hence, problem (26) can be rewritten as follows:

$$\begin{aligned} &\max_{\boldsymbol{\nu}} \sum_{h=1}^{H} \min_{k\in\mathcal{H}_h} W\left\{\sum_{i=1}^{\zeta}\log_2\left(1+\frac{P}{|\mathcal{H}_h|H\zeta\sigma^2}\left|\boldsymbol{D}_k\left(i,i\right)\right|^2\right)\right\} \\ &\text{s.t.} \quad (10\text{c})\,. \end{aligned} \tag{27}$$

Substituting $\boldsymbol{D}_k\left(i,i\right)=\alpha_i\beta_i\boldsymbol{\nu}^H\boldsymbol{c}^{ii}$ into (27), the problem (27) can be transformed as follows:

$$\begin{aligned} \max_{\boldsymbol{\nu}} \quad & \sum_{h=1}^{H}\min_{k\in\mathcal{H}_h} W\left\{\sum_{i=1}^{\zeta}\log_2\left(1+b_i\boldsymbol{\nu}^{\mathrm{H}}\boldsymbol{C}^{ii}\boldsymbol{\nu}\right)\right\} \\ \text{s.t.} \quad & (10\text{c})\,, \end{aligned} \tag{28}$$

where $\boldsymbol{C}^{ii}\triangleq\boldsymbol{c}^{ii}\left(\boldsymbol{c}^{ii}\right)^{\mathrm{H}}$ and $b_i\triangleq\frac{P}{|\mathcal{H}_h|H\zeta\sigma^2}\left|\alpha_i\beta_i\right|^2$.

3) Solution of Problem (28) : Since constraint (10c) has a manifold structure, problem (28) can be regarded as a manifold-constrained optimization problem. Standard gradient methods in Euclidean space cannot guarantee that the obtained solution is within the manifold. However, the Riemannian gradient method can extend the gradient method in Euclidean space to the manifold space. Next, we introduce the use of a manifold method [44] to solve problem (28). In particular, we first introduce the definition of a tangent space. Then, similar to the gradient in Euclidean space, we introduce the gradient on the manifold, called the Riemannian gradient. Finally, problem (28) is solved by an iterative method using the Riemannian gradient.

To solve problem (28), we first rewrite the objective function as

$$f\left(\boldsymbol{\nu}\right)\triangleq-\sum_{h=1}^{H}\min_{k\in\mathcal{H}_h} W\left\{\sum_{i=1}^{\zeta}\log_2\left(1+b_i\boldsymbol{\nu}^{\mathrm{H}}\boldsymbol{C}^{ii}\boldsymbol{\nu}\right)\right\}. \tag{29}$$

Let $\boldsymbol{\nu}_n$ be the value at iteration n. Based on (29), the Euclidean gradient of the objective function $f\left(\boldsymbol{\nu}\right)$ at point $\boldsymbol{\nu}_n$ is given by

$$\begin{aligned} \nabla f\left(\boldsymbol{\nu}_n\right) &= -\sum_{h=1}^{H}\min_{k\in\mathcal{H}_h}\left\{\nabla\left(R_{k,h}\left(\boldsymbol{B}_h\left(\boldsymbol{\nu}_n\right),\boldsymbol{J}_k\left(\boldsymbol{\nu}_n\right),\boldsymbol{\nu}_n\right)\right)\right\}, \\ &= -\sum_{h=1}^{H}\min_{k\in\mathcal{H}_h} W\left\{\sum_{i=1}^{\zeta}\frac{1}{\ln 2}\frac{2b_i\boldsymbol{C}^{ii}\boldsymbol{\nu}_n}{1+b_i\left(\boldsymbol{\nu}_n\right)^{\mathrm{H}}\boldsymbol{C}^{ii}\boldsymbol{\nu}_n}\right\}. \end{aligned} \tag{30}$$

To introduce the Riemannian gradient, we first define a tangent space of of an oblique manifold $\mathcal{Q}$ at point $\boldsymbol{\nu}_n$ as

$$T_{\boldsymbol{\nu}_n}\mathcal{Q}=\left\{\boldsymbol{u}\in\mathbb{C}^G\middle|\left[\boldsymbol{u}\boldsymbol{\nu}_n^{\mathrm{H}}\right]_{g,g}=0,\forall g\in\mathcal{G}=\{1,2,...,G\}\right\}. \tag{31}$$

From (31), we see that the tangent space containts all the tangent vectors of $\mathcal{Q}$ at $\boldsymbol{\nu}_n$.

The Riemannian gradient of the objective function $f\left(\boldsymbol{\nu}\right)$ at point $\boldsymbol{\nu}_n$ can be obtained by orthogonally projecting the Euclidean gradient $\nabla f\left(\boldsymbol{\nu}_n\right)$ onto the tangent space $T_{\boldsymbol{\nu}_n}\mathcal{Q}$, which is given by

$$\mathcal{G}_{\boldsymbol{\nu}_n}\mathcal{Q}=\nabla f\left(\boldsymbol{\nu}_n\right)-\mathrm{Real}\left\{\nabla f\left(\boldsymbol{\nu}_n\right)\circ\left(\boldsymbol{\nu}_n^{\mathrm{T}}\right)^{\mathrm{H}}\right\}\circ\boldsymbol{\nu}_n, \tag{32}$$

Algorithm 1 Proposed Algorithm for Solving (28)

```
1: Initialize $\boldsymbol{\nu}_0 \in \mathcal{Q}$.
2: Obtain $d_{ij}$ according to (46).
3: Obtain $\boldsymbol{C}^{ii}$ and $b_i$ according to (28).
4: **repeat**
5:     Compute the Euclidean gradient using (30).
6:     Compute the Riemannian gradient by (32).
7:     Update $\boldsymbol{\nu}_{n+1}$ by (34).
8: **until** the objective function converges.
9: $\boldsymbol{\Phi} = \mathrm{diag}(\boldsymbol{\nu}^{\mathrm{H}})$.
```

where $\mathrm{Real}(\boldsymbol{M})$ is the real part of $\boldsymbol{M}$ and $\mathrm{Real}\left\{\nabla f\left(\boldsymbol{\nu}_n\right)\circ\left(\boldsymbol{\nu}_n^{\mathrm{T}}\right)^{\mathrm{H}}\right\}\circ\boldsymbol{\nu}_n$ is the projected gradient of Euclidean gradient $\nabla f\left(\boldsymbol{\nu}_n\right)$ on the tangent space $T_{\boldsymbol{\nu}_n}\mathcal{Q}$.

Given the Riemannian gradient, we use the optimization method in Euclidean space to solve the manifold-constrained optimization problem [45]. The update of $\boldsymbol{\nu}_n$ is

$$\bar{\boldsymbol{\nu}}_n=\boldsymbol{\nu}_n-\tilde{\lambda}_n\mathcal{G}_{\boldsymbol{\nu}_n}\mathcal{Q}, \tag{33}$$

where $\tilde{\lambda}_n$ is the step size. To ensure that the updated value of $\boldsymbol{\nu}_n$ lies in the feasible set, we have

$$\boldsymbol{\nu}_{n+1}=\bar{\boldsymbol{\nu}}_n\circ\frac{1}{\left|\bar{\boldsymbol{\nu}}_n\right|}. \tag{34}$$

The detailed process of using the manifold-based method to solve problem (28) is given in Algorithm 1. Given $\boldsymbol{\nu}$, we can use Lemma 1 to calculate $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ and $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$.

C. Optimization of the Transmit Beamforming Matrices of the BS and the Receive Beamforming Matrices of Users

Given $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ and $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$, we next introduce the use of the algorithm in [41] to optimize $\boldsymbol{F}^{\mathrm{B}}$, $\boldsymbol{F}^{\mathrm{R}}$, $\boldsymbol{W}_k^{\mathrm{B}}$, and $\boldsymbol{W}_k^{\mathrm{R}}$. The reason for using the algorithm in [41] is that the algorithm in [41] is a fast optimization algorithm on a complex oblique manifold which has lower complexity than that of the state-of-the-art algorithms [42] while keeping spectral efficiency near-optimal.

Since we have assumed that $\boldsymbol{B}=\left[\boldsymbol{B}_1,\ldots,\boldsymbol{B}_H\right]\in\mathbb{C}^{N^{\mathrm{B}}\times H\zeta}$ is a fully digital transmit beamforming matrix which has the same size as the hybrid transmit beamforming matrix $\boldsymbol{F}^{\mathrm{R}}\boldsymbol{F}^{\mathrm{B}}$ and we have also replaced $\boldsymbol{F}^{\mathrm{R}}\boldsymbol{F}^{\mathrm{B}}$ with $\boldsymbol{B}$ in problem (11), the problem of optimizing $\boldsymbol{F}^{\mathrm{B}}$ and $\boldsymbol{F}^{\mathrm{R}}$ can be formulated as

$$\min_{\boldsymbol{F}^{\mathrm{R}},\boldsymbol{F}^{\mathrm{B}}}\left\|\boldsymbol{B}-\boldsymbol{F}^{\mathrm{R}}\boldsymbol{F}^{\mathrm{B}}\right\|_F^2 \tag{35}$$

$$\left|\boldsymbol{F}^{\mathrm{R}}\left(i,j\right)\right|=1,\forall i,j, \tag{35a}$$

Due to the unit modulus constraints of (35a), vector $\boldsymbol{x}=\mathrm{v}\left(\boldsymbol{F}^{\mathrm{R}}\right)$ forms a complex circle manifold $\{\boldsymbol{x}\in\mathbb{C}^w:|\boldsymbol{x}_1|=|\boldsymbol{x}_2|=\cdots=|\boldsymbol{x}_w|=1\}$, where $\mathrm{v}\left(\boldsymbol{F}^{\mathrm{R}}\right)$ is the vectorization of $\boldsymbol{F}^{\mathrm{R}}$ and $w=N^{\mathrm{B}}M^{\mathrm{B}}$. Hence, problem (35) can be transformed to an unconstrained optimization problem on manifolds. The iterative algorithm used to solve problem (28) can be used to solve it. In particular, $\boldsymbol{x}^n$ can be updated using (33) and (34). Given the updated $\boldsymbol{x}$ with $\boldsymbol{x}=\mathrm{v}\left(\boldsymbol{F}^{\mathrm{R}}\right)$, the update of the RF transmit beamforming matrix at iteration n can be expressed as

$$\boldsymbol{F}_n^{\mathrm{R}} = \mathrm{v}^{-1}\left(\boldsymbol{x}^n\right), \tag{36}$$

where $\mathrm{v}^{-1}\left(\boldsymbol{x}^n\right)$ is the inverse-vectorization of $\boldsymbol{x}^n$. Given $\boldsymbol{F}_n^{\mathrm{R}}$, the optimization problem in (35) can be simplified as follows:

$$\min_{\boldsymbol{F}^{\mathrm{B}}} \left\|\boldsymbol{B} - \boldsymbol{F}_n^{\mathrm{R}}\boldsymbol{F}^{\mathrm{B}}\right\|_F^2. \tag{37}$$

Since problem (37) is a least-square optimization problem, $\boldsymbol{F}^{\mathrm{B}}$ at iteration n can be given by

$$\boldsymbol{F}_n^{\mathrm{B}} = \left(\boldsymbol{F}_n^{\mathrm{R}}\right)^{\dagger}\boldsymbol{B}, \tag{38}$$

where $\left(\boldsymbol{F}_n^{\mathrm{R}}\right)^{\dagger}$ is the Moore-Penrose pseudo inverse of $\boldsymbol{F}_n^{\mathrm{R}}$. At convergence, $\boldsymbol{F}^{\mathrm{R}}$ is expressed as

$$\boldsymbol{F}^{\mathrm{R}} = \mathrm{v}^{-1}\left(\boldsymbol{x}^{n+1}\right). \tag{39}$$

To satisfy constraint (10a), $\boldsymbol{F}^{\mathrm{B}}$ is expressed as

$$\boldsymbol{F}^{\mathrm{B}} = \frac{\sqrt{P}}{\left\|\boldsymbol{F}^{\mathrm{R}}\boldsymbol{F}_n^{\mathrm{B}}\right\|_F}\boldsymbol{F}_n^{\mathrm{B}}. \tag{40}$$

Similarly, given $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$, the problem of optimizing $\boldsymbol{W}_k^{\mathrm{B}}$ and $\boldsymbol{W}_k^{\mathrm{R}}$ can be given as follows:

$$\min_{\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}} \left\|\boldsymbol{J}_k - \boldsymbol{W}_k^{\mathrm{R}}\boldsymbol{W}_k^{\mathrm{B}}\right\|_F^2 \tag{41}$$

$$\left|\boldsymbol{W}_k^{\mathrm{R}}\left(i,j\right)\right| = 1, \forall i, j. \tag{41a}$$

Since this optimization problem is similar to the problem in (35), we can use the same method used to solve problem (35) to optimize $\boldsymbol{W}_k^{\mathrm{B}}$ and $\boldsymbol{W}_k^{\mathrm{R}}$. In future works, we plan to consider the extension of the proposed method to a wide band system with multiple carriers.

D. Complexity Analysis

The proposed algorithm for solving problem (10) is summarized in Algorithm 2. The complexity of Algorithm 2 lies in the calculation of (17), solving problem (28), and using the algorithm in [41] to find $\hat{\boldsymbol{F}}^{\mathrm{B}}, \hat{\boldsymbol{F}}^{\mathrm{R}}$, $\hat{\boldsymbol{W}}_k^{\mathrm{B}}$, and $\hat{\boldsymbol{W}}_k^{\mathrm{R}}$. The complexity of calculating (17) is $\mathcal{O}\left(\left(N^{\mathrm{B}}\right)^3 + \left|\mathcal{H}_h\right|\left(N^{\mathrm{B}}\right)^2\zeta\right)$. The complexity of solving problem (28) lies in computing the Euclidean gradient of the objective function in (28) at each iteration, which involves the complexity of $\mathcal{O}\left(HM^2\zeta\right)$. Hence, the total complexity of solving problem (28) is $\mathcal{O}\left(HM^2\zeta S_1\right)$, where S_1 is the number of iterations of using the manifold method to solve problem (28). The complexity of using the algorithm in [41] to find $\hat{\boldsymbol{F}}^{\mathrm{B}}, \hat{\boldsymbol{F}}^{\mathrm{R}}$, $\hat{\boldsymbol{W}}_k^{\mathrm{B}}$, and $\hat{\boldsymbol{W}}_k^{\mathrm{R}}$ is $\mathcal{O}\left(N^{\mathrm{B}}M^{\mathrm{B}}\zeta S_2 + N^{\mathrm{U}}M^{\mathrm{U}}\zeta S_2\right)$, where S_2 is the number of iterations required to converge. Hence, the total complexity of solving problem (10) is $\mathcal{O}\left(\left(N^{\mathrm{B}}\right)^3 + \left|\mathcal{H}_h\right|\left(N^{\mathrm{B}}\right)^2\zeta + HM^2\zeta S_1 + N^{\mathrm{B}}M^{\mathrm{B}}\zeta S_2 + N^{\mathrm{U}}M^{\mathrm{U}}\zeta S_2\right)$ $\approx O\left(\left(N^{\mathrm{B}}\right)^3 + HM^2\zeta S_1\right)$.

Algorithm 2 Proposed Scheme for Solving Problem (10)

```
Input: $\boldsymbol{H}^{\mathrm{B}}, \boldsymbol{H}_k^{\mathrm{R}}, \zeta, P, \sigma^2$.
Calculate $\boldsymbol{B}_h\left(\boldsymbol{\nu}\right)$ and $\boldsymbol{J}_k\left(\boldsymbol{\nu}\right)$ by Lemma 1.
Find the phase shift $\hat{\boldsymbol{\nu}}$ of the IRS by solving problem (28).
Obtain $\boldsymbol{B}_h\left(\hat{\boldsymbol{\nu}}\right)$ and $\boldsymbol{J}_k\left(\hat{\boldsymbol{\nu}}\right)$ by Lemma 1.
Calculate $\hat{\boldsymbol{F}}^{\mathrm{B}}, \hat{\boldsymbol{F}}^{\mathrm{R}}, \hat{\boldsymbol{W}}_k^{\mathrm{B}}$, and $\hat{\boldsymbol{W}}_k^{\mathrm{R}}$ by the algorithm in [41].
Output: $\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{F}}^{\mathrm{B}}, \hat{\boldsymbol{F}}^{\mathrm{R}}, \hat{\boldsymbol{W}}_k^{\mathrm{B}}, \hat{\boldsymbol{W}}_k^{\mathrm{R}}$.
```

E. Convergence Analysis

Theorem 2: Assume that there exists $\boldsymbol{\nu}^*$ such that $\mathcal{G}_{\boldsymbol{\nu}^*}\mathcal{Q} = \boldsymbol{0}$. Then, there exists a neighborhood $\mathcal{U}$ of $\boldsymbol{\nu}^*$ in $\mathbb{C}^G$, such that for all $\boldsymbol{\nu}_0 \in \mathcal{U}$, Algorithm 1 generates an infinite sequence $\{\boldsymbol{\nu}_n\}$ converging to $\boldsymbol{\nu}^*$.

Proof: According to [44, Theorem 6.3.2], there exists γ_R such that

$$\|\boldsymbol{\nu}_{n+1} - \bar{\boldsymbol{\nu}}_n\| \leq \gamma_R\|\boldsymbol{\nu}_n - \boldsymbol{\nu}^*\|^2. \tag{42}$$

Based on Algorithm 1, we further have

$$\begin{aligned}\|\boldsymbol{\nu}_{n+1} - \boldsymbol{\nu}^*\| &\leq \|\boldsymbol{\nu}_{n+1} - \bar{\boldsymbol{\nu}}_n\| + \|\bar{\boldsymbol{\nu}}_n - \boldsymbol{\nu}^*\| \\ &\leq \|\boldsymbol{\nu}_n - \boldsymbol{\nu}^* - \tilde{\lambda}_n\mathcal{G}_{\boldsymbol{\nu}^*}\mathcal{Q}\| \\ &+ \gamma_R\|\boldsymbol{\nu}_n - \boldsymbol{\nu}^*\|^2 \\ &= \gamma_T\|\boldsymbol{\nu}_n - \boldsymbol{\nu}^*\|^2 + \gamma_R\|\boldsymbol{\nu}_n - \boldsymbol{\nu}^*\|^2\end{aligned} \tag{43}$$

where the last inequality follows from the Lipschitz-continuous differential of function $f(\boldsymbol{\nu})$ and $\gamma_T > 0$ is a parameter related to the step-size $\tilde{\lambda}_n$ [44, Theorem 6.3.2]. ■

Next, we analyze the convergence of the entire proposed algorithm. The proof is established by showing that the sum rate (10) is nondecreasing when the sequence $\left(\boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}^{\mathrm{B}}, \boldsymbol{\nu}\right)$ is updated. To prove the convergence of the proposed algorithm, we first define $Q\left(\boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}^{\mathrm{B}}, \boldsymbol{\nu}\right) = \sum_{h=1}^{H}\min_{k\in\mathcal{H}_h}\left\{R_{k,h}\left(\boldsymbol{W}_k^{\mathrm{R}}, \boldsymbol{W}_k^{\mathrm{B}}, \boldsymbol{\nu}, \boldsymbol{F}^{\mathrm{R}}, \boldsymbol{F}^{\mathrm{B}}\right)\right\}$ and t is the iteration index. Then, we have

$$\begin{aligned}&Q\left(\boldsymbol{W}_k^{\mathrm{B}(t-1)}, \boldsymbol{W}_k^{\mathrm{R}(t-1)}, \boldsymbol{F}^{\mathrm{R}(t-1)}, \boldsymbol{F}^{\mathrm{B}(t-1)}, \boldsymbol{\nu}^{(t-1)}\right) \\ &\overset{\mathrm{a}}{\leq} Q\left(\boldsymbol{J}_k\left(\boldsymbol{\nu}^{(t-1)}\right), \boldsymbol{B}\left(\boldsymbol{\nu}^{(t-1)}\right), \boldsymbol{\nu}^{(t-1)}\right) \\ &\overset{\mathrm{b}}{\leq} Q\left(\boldsymbol{J}_k\left(\boldsymbol{\nu}^{(t)}\right), \boldsymbol{B}\left(\boldsymbol{\nu}^{(t)}\right), \boldsymbol{\nu}^{(t)}\right) \\ &\overset{\mathrm{c}}{\approx} Q\left(\boldsymbol{W}_k^{\mathrm{B}(t)}, \boldsymbol{W}_k^{\mathrm{R}(t)}, \boldsymbol{F}^{\mathrm{R}(t)}, \boldsymbol{F}^{\mathrm{B}(t)}, \boldsymbol{\nu}^{(t)}\right)\end{aligned} \tag{44}$$

where the inequality (a) is due to the fact that we first transform problem (10) into a fully digital beamforming problem. Inequality (b) follows from the fact that $\boldsymbol{\nu}^{(t)}$ is one suboptimal IRS phase shift solution of problem (10). The approximation (c) is due to the fact that the effective transmit beamforming matrix $\boldsymbol{F}^{\mathrm{R}}\boldsymbol{F}^{\mathrm{B}}$ is very close to the fully digital transmit beamforming matrix $\boldsymbol{B}$, and the effective receive beamforming matrix $\boldsymbol{W}_k^{\mathrm{B}}\boldsymbol{W}_k^{\mathrm{R}}$ of user k is very close to the fully digital receive beamforming matrix $\boldsymbol{J}_k$ of user k.

IV. Simulation Results

In our simulations, the coordinates of the BS and the IRS are (2m, 0m, 10m) and (0m, 148m, 10m), respectively. All users are randomly distributed in a circle centered at (7m, 148m, 1.8m) with a radius being 10 m. The bandwidth is set to 251.1886 MHz [16]. The values of other parameters are defined in Table II. In the simulation, if the gap in terms of objective function value $\sum_{h=1}^{H} \min_{k \in \mathcal{H}_h} W \left\{ \sum_{i=1}^{\zeta} \log_2 \left(1 + b_i \boldsymbol{\nu}^{\mathrm{H}} \boldsymbol{C}^{ii} \boldsymbol{\nu}\right) \right\}$ between two consecutive iterations is below 0.1%, Algorithm 1 converges. For comparison purposes, we consider five baselines:

- Digital-BD-Manifold is an algorithm for solving problem (11) where the fully digital beamforming matrices of the BS and the users are optimized by a BD method, and the phase of each element of the IRS is optimized by a manifold method. The purpose of comparing the proposed algorithm with Digital-BD-Manifold is to determine whether the proposed algorithm can efficiently achieve the same performance as Digital-BD-Manifold.
- Hybrid-BD-Random angle is an algorithm for solving problem (10) where the hybrid beamforming matrices of the BS and the users are optimized by a BD method, the phase of each element of the IRS is randomly selected. Different from the Hybrid-BD-Random angle, the proposed algorithm uses a manifold method to optimize the phase shifts of the IRS. The purpose of comparing the proposed algorithm with the Hybrid-BD-Random angle is to show that the proposed algorithm can optimize the phase shifts of the IRS.
- Digital-BD-Random angle is an algorithm for solving problem (11) where the fully digital beamforming matrices of the BS and the users are optimized by a BD method, and the phase of each element of the IRS is randomly selected. Hence, Digital-BD-Random angle is also a fully digital beamforming algorithm and the phase of each element of the IRS is randomly selected. The purpose of comparing the proposed algorithm with Digital-BD-Random angle is to show that the proposed algorithm can optimize the phase shifts of the IRS.
- Hybrid-SVD-Manifold is an algorithm for solving problem (10) where the hybrid beamforming matrices of the BS and the users are determined by the algorithm in [16] and the phase of each element of the IRS is optimized by a manifold method. Different from Hybrid-SVD-Manifold, the proposed algorithm uses a block diagonalization method to optimize the hybrid beamforming matrices of the BS and the users. The purpose of comparing the proposed algorithm with Hybrid-SVD-Manifold is to show the proposed algorithm can optimize the hybrid beamforming matrices of the BS and the users.
- Digital-SVD-Manifold is an algorithm for solving problem (11) where the fully digital beamforming matrices of the BS and the users are determined by the algorithm in [16], and the phase of each element of the IRS is optimized by a manifold method. Hence, Digital-SVD-Manifold is a fully digital beamforming algorithm, and the fully digital beamforming matrices of the BS and the users are optimized by the algorithm in [16]. The purpose of comparing the proposed algorithm with Digital-SVD-Manifold is to show that the proposed algorithm can optimize the hybrid beamforming matrices of the BS and the users.

TABLE II: Simulation Parameters

Parameters	**Values**	**Parameters**	**Values**
M	16×16	N^{B}	64
N^{U}	64	M^{B}	8
M^{U}	4	ζ	4
Y	7	G_{t}	24.5 dBi
G_{r}	0 dBi	σ^2	-90 dBm
L	7	P	50 dBm

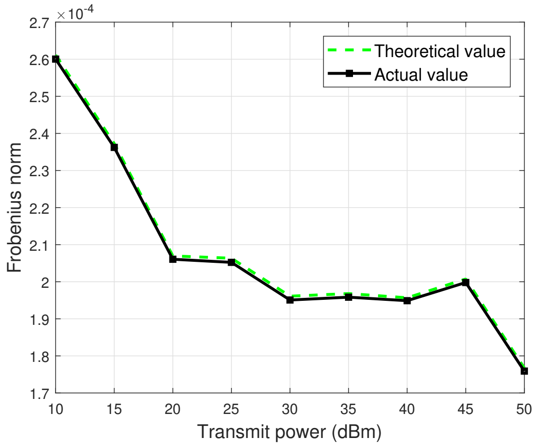


Fig. 2. The gap between the Frobenius norm of the matrix $\boldsymbol{\Sigma}_k^{(1)}(i,j)$ in Theorem 1 and $\beta_i \alpha_j \boldsymbol{\nu}^{\mathrm{H}} \boldsymbol{c}^{ij}$ obtained by our designed scheme.

Fig. 2 shows the gap between the Frobenius norm of the matrix $\boldsymbol{\Sigma}_k^{(1)}(i,j)$ in Theorem 1 and $\beta_i \alpha_j \boldsymbol{\nu}^{\mathrm{H}} \boldsymbol{c}^{ij}$ obtained by our designed scheme. In this figure, we randomly select one user to compare its Frobenius norm of $\beta_i \alpha_j \boldsymbol{\nu}^{\mathrm{H}} \boldsymbol{c}^{ij}$ with the F-norm of the theoretical value of diagonal matrix $\boldsymbol{\Sigma}_k^{(1)}(i,j)$ in Theorem 1. From this figure, we see that, the Frobenius norm of $\boldsymbol{\Sigma}_k^{(1)}(i,j)$ and $\beta_i \alpha_j \boldsymbol{\nu}^{\mathrm{H}} \boldsymbol{c}^{ij}$ are very close, which verifies the correctness of Theorem 1. From Fig. 2, we can also see that, the F-norm of $\boldsymbol{\Sigma}_k^{(1)}(i,j)$ and $\beta_i \alpha_j \boldsymbol{\nu}^{\mathrm{H}} \boldsymbol{c}^{ij}$ fluctuates up and down with the change of the transmit power. This implies that the singular values of the mmWave effective channel are not affected by the transmit power.

Fig. 3 shows the convergence of the proposed algorithm to solve problem (28). From Fig. 3, we observe that, the proposed algorithm can achieve up to 2x and 2x gains in terms of the sum of all users' data rates compared to Hybrid-SVD-Manifold and Digital-SVD-Manifold. This is due to the fact that, our proposed BD method can eliminate the inter-group interference. From Fig. 3, we can also see that, the number of iterations that the proposed algorithm needs to converge is similar to that of Digital-BD-Manifold. This implies that the proposed algorithm that uses hybrid beamforming can reduce energy consumption without increasing the complexity of finding suboptimal solution for serving users. Fig. 3 also shows that, the proposed scheme only needs seven iterations to solve problem (28), which further verifies the quick convergence

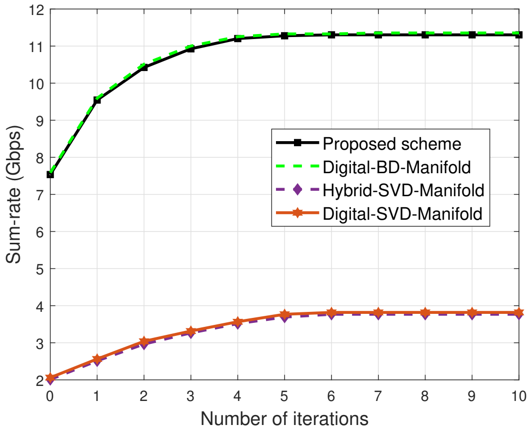


Fig. 3. Convergence of the considered algorithms for sovling problem (28).

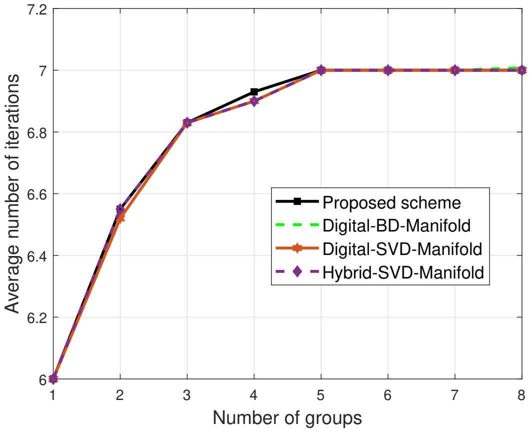


Fig. 4. The average number of iterations versus the number of groups.

speed of the designed scheme.

Fig. 4 shows how the average number of iterations that the considered algorithms need to converge changes as the number of groups varies. From Fig. 4, we see that, as the number of groups increases, the average number of iterations that the considered algorithms need to converge increases. This is due to the fact that, as the number of groups increases, the considered algorithms require more iterations to find the optimal phase shifts of the IRS. As the number of groups continues to increase, the average number of iterations for convergence remains constant. This is because the BS has enough user groups to determine the phase shifts of the IRS. From Fig. 4, we can also see that, the average number of iterations that the proposed algorithm needs to converge is similar to that of Digital-BD-Manifold and Digital-SVD-Manifold. This implies that the proposed algorithm can reduce the hardware implementation complexity without increasing the complexity of finding suboptimal solution for serving users.

Fig. 5 shows the cumulative distribution function (CDF) of the sum-rate of all users when the transmit power of the BS is 40 dBm. From Fig. 5, we can see that the proposed scheme improves the CDF of up to 81.8% and 78.33% gains compared to Hybrid-BD-Random angle and Digital-BD-Random angle when the sum of all users' data rates is 15 Gbps. This is because the phase shifts of the proposed scheme are optimized, resulting in high gain of the effective channel, which further facilitates the selection of the transmit beamforming matrices of the BS and the receive beamforming matrices of the users.

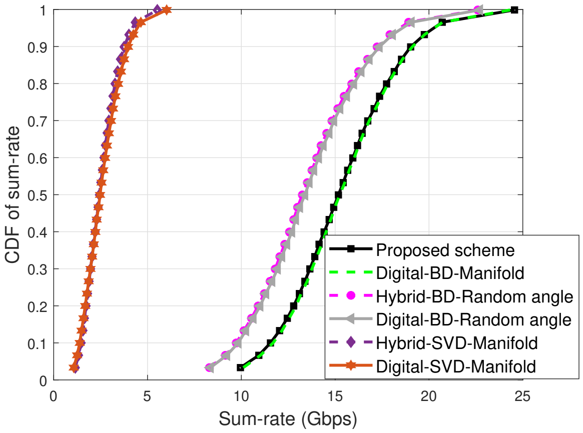
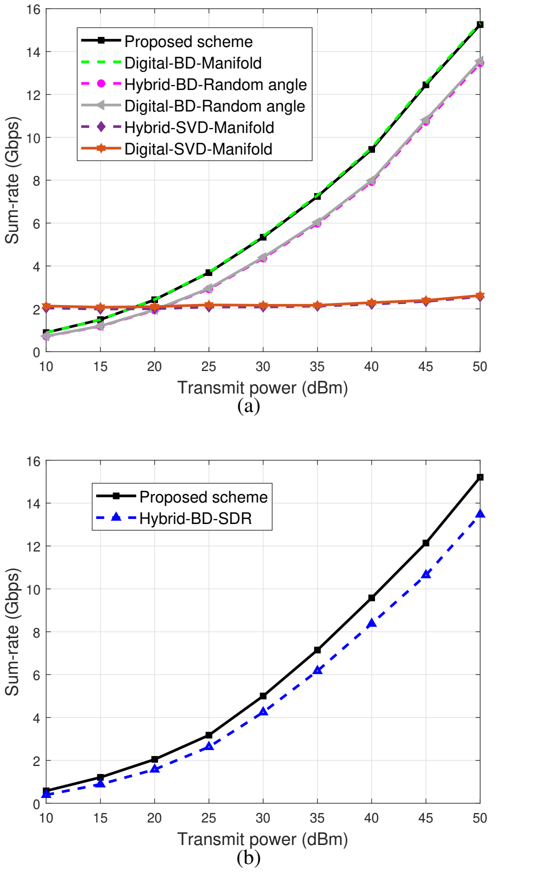


Fig. 5. CDF of sum-rate.



(a)

(b)

Fig. 6. Sum-rate changes as the transmit power of the BS.

Fig. 6 shows how the sum rate of all users changes as the transmit power of the BS varies. In this figure, Hybrid-BD-SDR is an algorithm for solving problem (10) where the hybrid beamforming matrices of the BS and the users are optimized by a BD method, and the phase of each element of the IRS is optimized by the semidefinite relaxation (SDR)

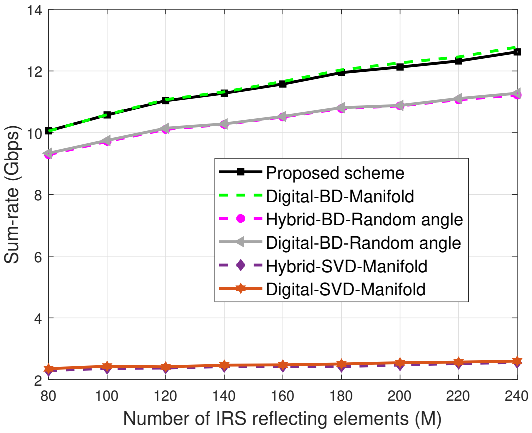


Fig. 7. Sum-rate versus the number of reflecting elements at the IRS.

algorithm [46]. From Fig. 6, we see that the gap in terms of the sum rate between Digital-BD-Manifold and the proposed scheme is only 0.8% when the transmit power of the BS is 50 dBm. This is because the proposed scheme can find the suboptimal hybrid beamforming matrices to represent the fully digital matrices. Fig. 6 also shows that compared to Hybrid-BD-Random angle, Hybrid-BD-SDR, and Hybrid-SVD-Manifold, the proposed scheme can achieve up to 13%, 13.81%, and 5x gains in terms of the sum rate of all users when P=50 dBm and M=256. This is because the proposed scheme optimizes the phase shifts of the IRS by a manifold method and eliminate the interference by the BD method. From Fig. 6, we can also see that, as the transmit power of the BS increases, the sum rates of Hybrid-SVD-Manifold and Digital-SVD-Manifold remain unchanged. This is due to the fact that Hybrid-SVD-Manifold and Digital-SVD-Manifold do not eliminate inter-group interference, which increases as the transmit power of the BS increases.

In Fig. 7, we show how the sum of all users' data rates changes as the number of reflecting elements at the IRS varies. Fig. 7 shows that the proposed scheme can achieve up to 13.3% and 4x gains in terms of the sum of all users' data rates compared to Hybrid-BD-Random angle and Hybrid-SVD-Manifold when M=240. This is due to the fact that the proposed scheme can align the angles of the cascaded channel and improve SINR. Fig. 7 also shows that as the number of reflection elements of the IRS increases, the performance of Hybrid-SVD-Manifold and Digital-SVD-Manifold remains unchanged. This is because Hybrid-SVD-Manifold and Digital-SVD-Manifold only eliminate the interference among multiple streams of each user without eliminating the inter-group interference.

In Fig. 8, we show how the sum of all users' data rates changes as the number of data streams changes. From this figure, we can see that, as the number of data streams increases, the sum-rate of all considered algorithms increases. This is due to the fact that the desired signal power increases as the number of data streams increases. Fig. 8 also shows that the proposed scheme can achieve up to 28.6% and 152.97% gains in terms of the sum of all users' data rates compared to Hybrid-BD-Random angle and Hybrid-SVD-Manifold when the number of data streams is 4. This is because our proposed BD method can eliminate the inter-group interference and the interference among multiple streams of each user, and the manifold method can find the suitable IRS phase shifts.

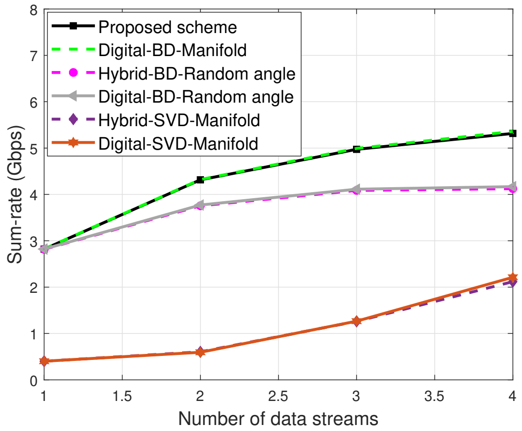


Fig. 8. Sum-rate versus number of data streams.

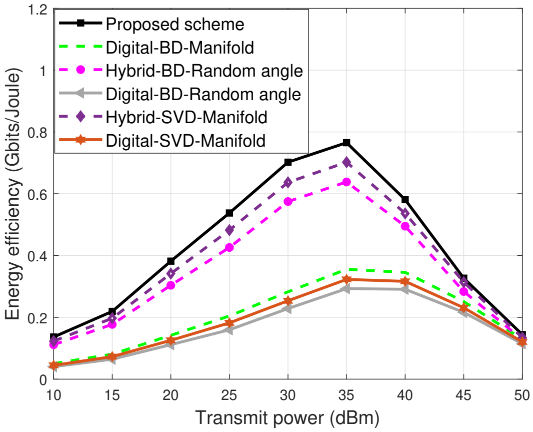


Fig. 9. Energy efficiency versus the transmit power of the BS.

Fig. 9 shows how the energy efficiency changes as the transmission power of the BS varies. In this figure, the energy efficiency is defined as the ratio of the sum-rate to the total power consumption of the system. The total power consumption of the system includes the transmission power at the BS, the hardware static power consumption at the BS, and the hardware static power consumption of the each reflecting element at the IRS. From Fig. 9, we can see that, the proposed scheme can achieve up to 19.92% gain in terms of energy efficiency compared to Hybrid-BD-Random angle. This is due to the fact that the proposed algorithm optimizes the phase shift to align the angle of the path from the BS to the users. From Fig. 9, we can also see that the proposed scheme can achieve up to 9% gain in terms of energy efficiency compared to Hybrid-SVD-Manifold. This is because the proposed BD method eliminates both the inter-group interference and the interference among multiple streams, while Hybrid-SVD-Manifold only eliminates the interference among multiple streams of each user. Fig. 9 also shows that the energy efficiency increases when the transmit power of the BS is less than 35 dBm. However, when the transmit power of the BS is higher than 35 dBm, the energy

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3297732

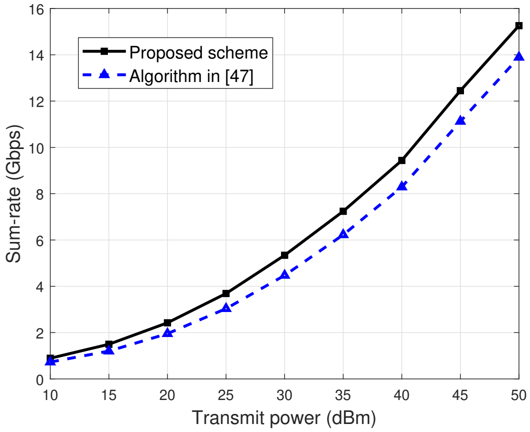


Fig. 10. Sum-rate versus the transmit power of the BS.

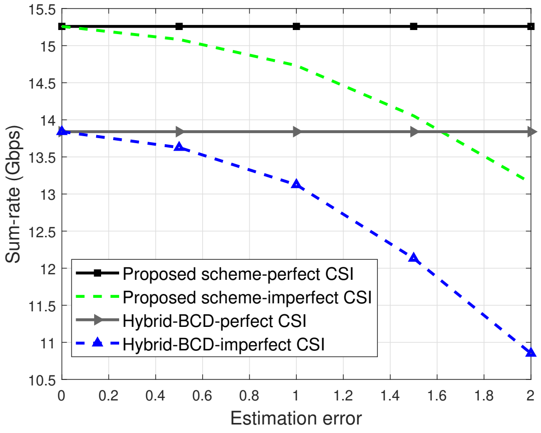


Fig. 11. Sum-rate versus the estimation error.

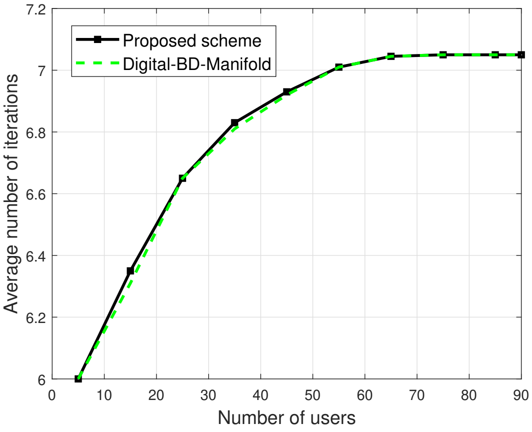


Fig. 12. The average number of iterations versus the number of users.

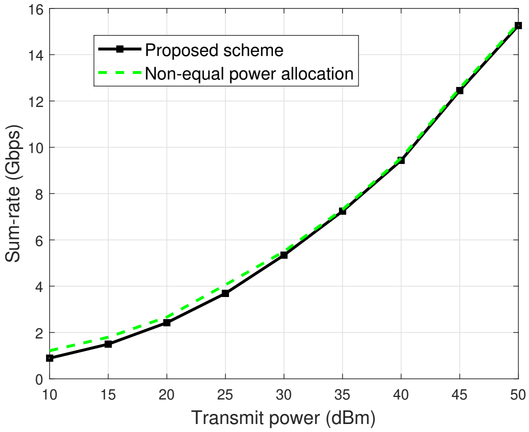


Fig. 13. Sum-rate versus the transmit power of the BS.

efficiency decreases. This is due to the fact that, as the transmit power of the BS increases, the total power consumption of the system increases.

Fig. 10 shows how the sum rate of all users changes as the transmit power of the BS varies. From Fig. 10, we can see that, compared to the algorithm in [47], the proposed scheme can achieve up to 11.09% gain in terms of the sum rate of all users when P=50 dBm and M=256. This is because the proposed BD method eliminates both the inter-group interference and the interference among multiple streams.

Fig. 11 shows how the sum rate of all users changes as the estimation error varies. In this figure, we compare the proposed scheme with the Hybrid-BCD which is a block coordinate descent (BCD) algorithm where the beamforming matrices of the BS and the IRS phase shifts are alternately optimized [47]. From Fig. 11, we can see that, with the presence of channel errors, the proposed algorithm has a significant performance gain compared to Hybrid-BCD. This is because our proposed BD method can eliminate interference, and the manifold method can find the suitable IRS phase shifts, thus improving SINR and reducing sensitivity to inaccurate CSI.

Fig. 12 shows how the average number of iterations that the considered algorithms need to converge changes as the number of users varies. From Fig. 12, we can see that, as the number of users increases, the average number of iterations that the considered algorithms need to converge increases. As the number of users continues to increase, the average number of iterations for convergence remains constant. This is because the BS has enough positions of users to determine the phase shifts of the IRS.

Fig. 13 shows how different power allocation schemes affect the sum data rate of all users. From Fig. 13, we see that, the proposed scheme is close to the non-equal power allocation algorithm. This is because the IRS and large number of antennas at the BS provide the massive array gain, which results in large effective SNR.

V. Conclusions

In this paper, we have developed a novel framework for an IRS-assisted mmWave multigroup multicast MIMO communication system. The transmit beamforming matrices of the BS, the receive beamforming matrices of the users, and the phase shifts of the IRS were jointly optimized to maximize the sum rate of all users. We have used a BD method to represent the beamforming matrices of the BS and the users in terms of the IRS phase shifts. Then, we have transformed the original problem to a problem that only needs to optimize the IRS phase shifts. The transformed problem is solved by a manifold method. Simulation results show that the proposed scheme can achieve significant performance gains compared to baselines.

Appendix A

To prove Theorem 1, we first define the effective channel $\boldsymbol{H}_{h,k}$ as done in [16]:

$$\boldsymbol{H}_{h,k} = G_{\mathrm{t}} G_{\mathrm{r}} \boldsymbol{H}_k^{\mathrm{R}} \boldsymbol{\Phi} \boldsymbol{H}^{\mathrm{B}} = \boldsymbol{A}_k \boldsymbol{D}_k \left(\boldsymbol{A}\right)^{\mathrm{H}}, \tag{45}$$

where $\boldsymbol{A}_k = \left[\boldsymbol{a}\left(r_{1,k}^{\mathrm{A}}\right), \ldots, \boldsymbol{a}\left(r_{L,k}^{\mathrm{A}}\right)\right]$ is a array response matrix of user k, $\boldsymbol{A} = \left[\boldsymbol{a}\left(r_1^{\mathrm{D}}\right), \ldots, \boldsymbol{a}\left(r_Y^{\mathrm{D}}\right)\right]$ is a array response matrix of the BS, and $\boldsymbol{D}_k$ is an $Y \times L$ matrix with element $\boldsymbol{D}_k\left(i,j\right) = \beta_i \alpha_j d_{ij}$, where d_{ij} can be given by

$$\begin{aligned} d_{ij} &= \left(\boldsymbol{a}\left(\theta_i^{\mathrm{D}}, \eta_i^{\mathrm{D}}\right)\right)^{\mathrm{H}} \boldsymbol{\Phi} \boldsymbol{a}\left(\theta_j^{\mathrm{A}}, \eta_j^{\mathrm{A}}\right) \\ &= \boldsymbol{\nu}^{\mathrm{H}} \left(\left(\boldsymbol{a}\left(\theta_i^{\mathrm{D}}, \eta_i^{\mathrm{D}}\right)\right)^{*} \circ \boldsymbol{a}\left(\theta_j^{\mathrm{A}}, \eta_j^{\mathrm{A}}\right)\right) = \boldsymbol{\nu}^{\mathrm{H}} \boldsymbol{c}^{ij}. \end{aligned} \tag{46}$$

Given the effective channel $\boldsymbol{H}_{h,k}$, we define $\tilde{\boldsymbol{H}}_h$ as

$$\begin{aligned} \tilde{\boldsymbol{H}}_h &= \left[\bar{\boldsymbol{H}}_1, \ldots, \bar{\boldsymbol{H}}_{h-1}, \bar{\boldsymbol{H}}_{h+1}, \ldots, \bar{\boldsymbol{H}}_H\right]^{\mathrm{T}}, \\ &= \begin{bmatrix} \boldsymbol{A}_1 \boldsymbol{D}_1 (\boldsymbol{A})^{\mathrm{H}} \\ \boldsymbol{A}_2 \boldsymbol{D}_2 (\boldsymbol{A})^{\mathrm{H}} \\ \vdots \\ \boldsymbol{A}_K \boldsymbol{D}_K (\boldsymbol{A})^{\mathrm{H}} \end{bmatrix}, \\ &= \begin{bmatrix} \boldsymbol{A}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{A}_K \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{D}_K \end{bmatrix} \begin{bmatrix} (\boldsymbol{A})^{\mathrm{H}} \\ \vdots \\ (\boldsymbol{A})^{\mathrm{H}} \end{bmatrix}, \\ &= \begin{bmatrix} \boldsymbol{A}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{A}_K \end{bmatrix} \boldsymbol{P} \begin{bmatrix} \tilde{\boldsymbol{\Sigma}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{Q} \begin{bmatrix} (\boldsymbol{A})^{\mathrm{H}} \\ \vdots \\ (\boldsymbol{A})^{\mathrm{H}} \end{bmatrix}. \end{aligned} \tag{47}$$

We define $\boldsymbol{Q} \begin{bmatrix} (\boldsymbol{A})^{\mathrm{H}} \\ \vdots \\ (\boldsymbol{A})^{\mathrm{H}} \end{bmatrix}$ as $\boldsymbol{Z}$, hence, $\tilde{\boldsymbol{V}}_h^{(0)} = \left[\boldsymbol{z}_{(K-|\mathcal{H}_h|)\zeta+1}; \ldots; \boldsymbol{z}_{K\zeta}\right]$, with $\boldsymbol{z}_{(K-|\mathcal{H}_h|)\zeta+1}$ being row $\left(K - |\mathcal{H}_h|\right)\zeta + 1$ of matrix $\boldsymbol{Z}$. Based on (45), $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}$ is given by

$$\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)} = \boldsymbol{A}_k \boldsymbol{D}_k \left(\boldsymbol{A}\right)^{\mathrm{H}} \left[\boldsymbol{z}_{(K-|\mathcal{H}_h|)\zeta+1}; \ldots; \boldsymbol{z}_{K\zeta}\right]. \tag{48}$$

For ULA with N antennas, the column vectors of $\boldsymbol{A}_k$ and row vectors of $\left(\boldsymbol{A}\right)^{\mathrm{H}} \left[\boldsymbol{z}_{(K-|\mathcal{H}_h|)\zeta+1}; \ldots; \boldsymbol{z}_{K\zeta}\right]$ can form orthonormal sets [16]. Hence, $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)} = \boldsymbol{A}_k \boldsymbol{D}_k \left(\boldsymbol{A}\right)^{\mathrm{H}} \left[\boldsymbol{z}_{(K-|\mathcal{H}_h|)\zeta+1}; \ldots; \boldsymbol{z}_{K\zeta}\right]$ can be considered as an approximation of the truncated SVD of $\boldsymbol{H}_{h,k}\tilde{\boldsymbol{V}}_h^{(0)}$, and $\boldsymbol{D}_k$ can represent $\boldsymbol{\Sigma}_k^{(1)}$.

References

[1] S. Zhang, Z. Yang, M. Chen, D. Liu, K.-K. Wong, and H. V. Poor, “Performance optimization for intelligent reflecting surface assisted multicast MIMO networks,” in *Proc. IEEE Global Communications Conference*, Rio de Janeiro, Brazil, Dec. 2022, pp. 5838–5843.

[2] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, “Millimeter-wave massive MIMO: The next wireless revolution?” *IEEE Communications Magazine*, vol. 52, no. 9, pp. 56–62, Sep. 2014.

[3] J. Sun, M. Jia, Q. Guo, X. Gu, and Y. Gao, “Power distribution based beamspace channel estimation for mmWave massive MIMO system with lens antenna array,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 695–10 708, Dec. 2022.

[4] J. Wang, X. Zhang, X. Shi, and J. Song, “Higher spectral efficiency for mmWave MIMO: Enabling techniques and precoder designs,” *IEEE Communications Magazine*, vol. 59, no. 4, pp. 116–122, Apr. 2021.

[5] W. Wang and A. Leshem, “Non-convex generalized nash games for energy efficient power allocation and beamforming in mmWave networks,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 3193–3205, Jun. 2022.

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

[7] C. Qi, Q. Liu, X. Yu, and G. Y. Li, “Hybrid precoding for mixture use of phase shifters and switches in mmWave massive MIMO,” *IEEE Transactions on Communications*, vol. 70, no. 6, pp. 4121–4133, Jun. 2022.

[8] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, “A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives,” *IEEE Communications Surveys and Tutorials*, vol. 20, no. 4, pp. 3060–3097, Fourthquarter 2018.

[9] Z. Yang, M. Chen, W. Saad, W. Xu, M. Shikh-Bahaei, H. V. Poor, and S. Cui, “Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 665–679, Jan. 2022.

[10] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, “Reconfigurable intelligent surfaces for energy efficiency in wireless communication,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[11] S. Liu, Z. Gao, J. Zhang, M. D. Renzo, and M.-S. Alouini, “Deep denoising neural network assisted compressive channel estimation for mmWave intelligent reflecting surfaces,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 9223–9228, Aug. 2020.

[12] E. E. Bahingayi and K. Lee, “Low-complexity beamforming algorithms for IRS-aided single-user massive MIMO mmWave systems,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 9200–9211, Nov. 2022.

[13] F. Yang, J.-B. Wang, H. Zhang, M. Lin, and J. Cheng, “Intelligent reflecting surface assisted mmWave communication using mixed timescale channel state information,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 5673–5687, Jul. 2022.

[14] H. Du, J. Zhang, J. Cheng, and B. Ai, “Millimeter wave communications with reconfigurable intelligent surfaces: Performance analysis and optimization,” *IEEE Transactions on Communications*, vol. 69, no. 4, pp. 2752–2768, Apr. 2021.

[15] P. Wang, J. Fang, X. Yuan, Z. Chen, and H. Li, “Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 960–14 973, Dec. 2020.

[16] P. Wang, J. Fang, L. Dai, and H. Li, “Joint transceiver and large intelligent surface design for massive MIMO mmWave systems,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1052–1064, Feb. 2021.

[17] C. Pradhan, A. Li, L. Song, B. Vucetic, and Y. Li, “Hybrid precoding design for reconfigurable intelligent surface aided mmWave communication systems,” *IEEE Wireless Communications Letters*, vol. 9, no. 7, pp. 1041–1045, Jul. 2020.

[18] K. Ying, Z. Gao, S. Lyu, Y. Wu, H. Wang, and M.-S. Alouini, “GMD-based hybrid beamforming for large reconfigurable intelligent surface assisted millimeter-wave massive MIMO,” *IEEE Access*, vol. 8, pp. 19 530–19 539, Jan. 2020.

[19] W. Zhang, J. Xu, W. Xu, D. W. K. Ng, and H. Sun, “Cascaded channel estimation for IRS-assisted mmWave multi-antenna with quantized beamforming,” *IEEE Communications Letters*, vol. 25, no. 2, pp. 593–597, Feb. 2021.

[20] H. Xie, J. Xu, and Y.-F. Liu, “Max-min fairness in IRS-aided multi-cell MISO systems with joint transmit and reflective beamforming,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1379–1393, Feb. 2021.

[21] G. Zhou, C. Pan, H. Ren, K. Wang, and M. D. Renzo, “Fairness-oriented multiple RIS-aided mmWave transmission: Stochastic optimization methods,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 1402–1417, Mar. 2022.

[22] F. Yang, J.-B. Wang, H. Zhang, M. Lin, and J. Cheng, “Multi-IRS-assisted mmWave MIMO communication using twin-timescale channel state information,” *IEEE Transactions on Communications*, vol. 70, no. 9, pp. 6370–6384, Sept. 2022.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3297732

[23] H. Niu, Z. Chu, F. Zhou, C. Pan, D. W. K. Ng, and H. X. Nguyen, "Double intelligent reflecting surface-assisted multi-user MIMO mmWave systems with hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 1575–1587, Feb. 2022.

[24] B. Zheng, C. You, and R. Zhang, "Double-IRS assisted multi-user MIMO: Cooperative passive beamforming design," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4513–4526, Jul. 2021.

[25] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. D. Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[26] C. Zhang, H. Lu, and C. W. Chen, "Reconfigurable intelligent surfaces-enhanced uplink user-centric networks on energy efficiency optimization," *IEEE Transactions on Wireless Communications*, to appear, 2023.

[27] H. Huang, Y. Zhang, H. Zhang, Z. Zhao, C. Zhang, and Z. Han, "Multi-IRS-aided millimeter-wave multi-user MISO systems for power minimization using generalized benders decomposition," *IEEE Transactions on Wireless Communications*, to appear, 2023.

[28] L. Lv, Q. Wu, Z. Li, Z. Ding, N. Al-Dhahir, and J. Chen, "Covert communication in intelligent reflecting surface-assisted NOMA systems: Design, analysis, and optimization," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1735–1750, Mar. 2022.

[29] H. Guo and V. K. N. Lau, "Uplink cascaded channel estimation for intelligent reflecting surface assisted multiuser MISO systems," *IEEE Transactions on Signal Processing*, vol. 70, pp. 3964–3977, 2022.

[30] W. Huang, Y. Huang, S. He, and L. Yang, "Cloud and edge multicast beamforming for cache-enabled ultra-dense networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 3481–3485, Mar. 2020.

[31] W. Ci, C. Qi, G. Y. Li, and S. Mao, "Hybrid beamforming design for covert multicast mmWave massive MIMO communications," in *Proc. IEEE Global Communications Conference*, Madrid, Spain, Dec. 2021, pp. 1–6.

[32] Z. Zhang, Z. Ma, Y. Xiao, M. Xiao, G. K. Karagiannidis, and P. Fan, "Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 8, pp. 1794–1808, Aug. 2017.

[33] Q. Tao, S. Zhang, C. Zhong, and R. Zhang, "Intelligent reflecting surface aided multicasting with random passive beamforming," *IEEE Wireless Communications Letters*, vol. 10, no. 1, pp. 92–96, Jan. 2021.

[34] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Intelligent reflecting surface aided multigroup multicast MISO communication systems," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3236–3251, Apr. 2020.

[35] L. Du, S. Shao, G. Yang, J. Ma, Q. Liang, and Y. Tang, "Capacity characterization for reconfigurable intelligent surfaces assisted multiple-antenna multicast," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6940–6953, Oct. 2021.

[36] L. Du, W. Zhang, J. Ma, and Y. Tang, "Reconfigurable intelligent surfaces for energy efficiency in multicast transmissions," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 6266–6271, Jun. 2021.

[37] W. Jiang, P. Xiong, J. Nie, Z. Ding, C. Pan, and Z. Xiong, "Robust design of IRS-aided multi-group multicast system with imperfect CSI," *IEEE Transactions on Wireless Communications*, to appear, 2023.

[38] M. Farooq, V. Kumar, M. Juntti, and L.-N. Tran, "On the achievable rate of IRS-assisted multigroup multicast systems," in *Proc. IEEE Global Communications Conference*, Rio de Janeiro, Brazil, Dec. 2022, pp. 5844–5849.

[39] J. Chen, Y.-C. Liang, H. V. Cheng, and W. Yu, "Channel estimation for reconfigurable intelligent surface aided multi-user mmWave MIMO systems," *IEEE Transactions on Wireless Communications*, to appear, 2023.

[40] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4522–4535, Jul. 2020.

[41] H. Kasai, "Fast optimization algorithm on complex oblique manifold for hybrid precoding in millimeter wave MIMO systems," in *Proc. IEEE Global Conference on Signal and Information Processing*, Anaheim, USA, Nov. 2018, pp. 1266–1270.

[42] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[43] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[44] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[45] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober, "Resource allocation for secure IRS-assisted multiuser MISO systems," in *Proc. IEEE Global Communications Conference Workshops*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[46] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[47] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.