Intelligent Reflecting Surface Aided NOMA for Millimeter-Wave Massive MIMO with Lens Antenna Array

Penglu Liu, *Student Member, IEEE,* Yong Li, Wei Cheng, Xiang Gao, and Xiaojing Huang, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a downlink intelligent reflecting surface (IRS) aided non-orthogonal multiple access (NOMA) for millimeter-wave (mmWave) massive MIMO with lens antenna array, i.e., IRS-aided mmWave beamspace NOMA, where the single-antenna users without direct-link but connected to the base station (BS) with the aid of the IRS are grouped as one NOMA group. Considering the power leakage problem in beamspace channel and the per-antenna power constraint, we propose two multi-beam selection strategies for the BS-IRS link under two channel models, i.e., 2-dimension (2D) channel model and 3-dimension (3D) channel model, respectively, where two corresponding RF chain configuration strategies are designed, respectively. Then, we formulate and solve the optimization problem for maximizing the weighted sum rate by jointly optimizing the active beamforming at the BS and the passive beamforming at the IRS, where we propose the alternating optimization (AO) method to solve the above joint optimization problem. Especially, different from the stochastic method, based on the beam-splitting technique, we propose the method to initialize the feasible solution for the proposed AO method, where the transmit power minimization problem is formulated and solved. Through simulations, the weighted sum rate performance of the proposed IRS-aided mmWave beamspace NOMA is verified.**

***Index Terms*—Intelligent reflecting surface, non-orthogonal multiple access, millimeter-wave, beamspace, beam-splitting.**

I. Introduction

DRIVEN by the rapid development of applications such as augmented reality (AR), virtual reality (VR), internet of things (IoT), internet of vehicles (IoV) and mobile-edge computing (MEC), next-generation wireless communication networks are required to support high capacity and massive connectivity [1]. Millimeter wave (mmWave) [2], massive multiple-input multiple-output (MIMO) [3–5] and non-orthogonal multiple access (NOMA) [6, 7] are three key technologies which can collectively provide massive connectivity, high spectral efficiency (SE) and energy efficiency (EE), and better interference control to meet such requirements.

In addition to the huge spectrum resource, the small wavelengths at mmWave allow more antennas in a limited physical space, which facilitates the generation of massive MIMO. However, for the conventional antenna architectures [8], i.e., full-digital and hybrid antenna architectures, it is difficult to realize the massive MIMO due to high transceiver complexity and energy consumption. To address the above problem, the antenna selection technique has been investigated in [9, 10]. However, the antenna selection may introduce obvious performance loss. Recently, a new MIMO architecture based on the high-resolution discrete lens array (DLA), i.e., continuous aperture phased (CAP) MIMO, is developed in [11], where the receiver uses a DLA architecture to map the analog spatial signals to signals in beamspace. Based on the lens array, through performing antenna/beam selection, the beamspace MIMO can reduce the number of required radio-frequency (RF) chains to fully exploit the advantages of the mmWave with reduced transceiver complexity and energy consumption [12, 13]. Furthermore, the obvious performance loss can be avoided.

Different from the conventional multiple access technologies, NOMA can serve multiple users within the same resource block, i.e., time, frequency, space and code, where the transmitter and the receiver perform superposition coding (SC) and successive interference cancellation (SIC), respectively. Such NOMA technologies proposed in [14–18] are based on the conventional antenna architectures. Making use of the advantages of the beamspace MIMO and the sparsity of mmWave channel, a new spectrum and energy-efficient mmWave transmission scheme that integrates the concept of NOMA with beamspace MIMO is proposed in [19], where the users within the same selected beam are served as one NOMA group. Based on the same beamspace NOMA scheme, the energy-efficient power allocation method is proposed in [20]. To support a larger number of groups of users and reduce the SIC complexity, based on reconfigurable antenna multiple access (RAMA) [21], the reconfigurable antenna NOMA (RA-NOMA) is proposed in [22]. Different from the single-beam beamspace NOMA scheme, a multi-beam NOMA using the mmWave massive MIMO with lens antenna array is proposed in [23] to reduce the total power consumption, where the joint user grouping and power allocation optimization problem is solved.

Note that all of the above NOMA systems were designed for the scenario where the users have the direct-link with the base station (BS). Considering the mmWave-band blockage,

This work is supported in part by the National Natural Science Foundation of China under Grant 61401360 and in part by Fundamental Research Funds for the Central Universities under Grant 3102017zy026. (*Corresponding Author: Yong Li.*)

Penglu Liu, Yong Li, Wei Cheng, and Xiang Gao are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (email: lpl@mail.nwpu.edu.cn; ruikel@nwpu.edu.cn; pupil_119@nwpu.edu.cn; 1486812648@mail.nwpu.edu.cn).

Xiaojing Huang is with the Global Big Data Technologies Centre, University of Technology Sydney, Sydney, NSW 2007, Australia (email: Xiaojing.Huang@uts.edu.au).

if there exists no direct-link between the BS and the user, the user must be served through the relay [24, 25]. In addition, considering the severe path loss of the mmWave channel, the relay can also be used to enhance the received SNR at the user. However, deploying a large number of relays results in the immense expenditure and power consumption, which is contrary to the Green 6G Network. Fortunately, by smartly reconfiguring the wireless propagation environment with the use of massive low-cost passive reflecting elements integrated on a planar surface, intelligent reflecting surface (IRS)/reconfigurable intelligent surface (RIS) can significantly improve the performance of wireless communication networks, which is a revolutionary technology for achieving spectrum and energy-efficient wireless communication cost-effectively in the future [26, 27].

In the IRS-aided communication systems, besides the cascaded-link channel estimation [28–30], the other key problem is how to solve the joint active beamforming at the BS and the passive beamforming at the IRS optimization problem. It is difficult to obtain the optimal solution to the above problem, due to the non-convex signal-to-interference-plus-noise ratio (SINR) costraints and the unit-modulus constraints imposed by the passive phase shifters. Based on the IRS-aided single-cell wireless system, the problem to minimize the total transmit power at an access point (AP) is formulated and solved through the designed alternating optimization (AO) method, where the semidefinite relaxation (SDR) method is used to transform the above two kinds of constraints into convex constraints [31]. To maximize the weighted sum rate for IRS-aided wireless networks, a joint transmit beamforming and reflect beamforming optimization method is designed in [32], where the original optimization problem is decoupled via Lagrangian dual transform, and two channel state information (CSI) setups, i.e., perfect CSI and imperfect CSI, were investigated, respectively. In [33], the intelligent reflecting surfaces (IRSs) are employed to enhance the physical layer security in a challenging radio environment, where a penalty-based approach is used to obatin the passive beamforming. In [34], the one-by-one (OBO) method is proposed to optimize the passive beamforming, in which each one of the phase coefficents are optimized in order by fixing the other coefficients as constant.

For the IRS-aided NOMA systems, [35–40] investigate the IRS-aided downlink NOMA communication. In [35], a simple design of IRS-NOMA transmission is proposed. The power minimization for IRS-aided downlink NOMA system is investigated in [36]. A three-step novel resource allocation algorithm for IRS-NOMA systems is proposed in [37], where the three steps include channel assignment, decoding order optimization, and joint power allocation and reflection coefficient design. In [38], the transmit power minimization problem and sum rate maximization problem are investigated, respectively, where the IRS is used to tune the wireless channels to satisfy the quasi-degradation constraint. The sum rate maximization problem is also investigated in [39], where two kinds of reflecting elements constraints were considered, i.e., ideal IRS case and non-ideal IRS case. Different from the signal enhancement based (SEB) design, where the IRS is used to boost the signal at the user side or at the BS side, a signal cancellation based (SCB) design is investigated in the RIS-aided NOMA networks [40]. Especially, the minimal required number of RISs for both the diffuse scattering and anomalous reflector scenarios are discussed. For the IRS-aided uplink NOMA, under the individual power constraint, the sum rate maximization problem is investigated in [41], where the investigated problem requires a joint power control at the users and beamforming design at the IRS. Nevertheless, to the best of our knowledge, there is no investigation into the IRS-aided NOMA for mmWave massive MIMO with lens antenna array, i.e., IRS-aided mmWave beamspace NOMA.

In this paper, we investigate a downlink IRS-aided mmWave beamspace NOMA communication system. We aim to maximize the weighted sum rate of all users by jointly optimizing the active beamforming at the BS and the passive beamforming at the IRS. In addition to the SIC constraint, in this paper, the minimum data rate requirement constraint, the total power consumption constraint, and the per-antenna power constraint are also considered, which play significant roles on the sum rate performance of the proposed system. Different from the conventional IRS-aided systems, in this proposed IRS-aided beamspace NOMA system, the main effect of per-antenna power constraint is reflected in how much power can be provided to the BS-IRS link. The main contributions of this paper can be summarized as follows:

- We design a novel IRS-aided mmWave beamspace NOMA communication system in which the users with direct-link are served through the traditional beamspace NOMA strategy, i.e., the users within the same selected beam are grouped as one NOMA group. However, the users without direct-link but connected to the BS with the aid of the IRS are grouped as a different NOMA group.
- Taking into account the power leakage problem in beamspace channel and the per-antenna power constraint, we propose two multi-beam selection strategies for the BS-IRS link under both 2-dimension (2D) channel model and 3-dimension (3D) channel model, respectively, where two corresponding RF chain configuration strategies are designed, respectively.
- We propose the AO method to solve the joint optimization of active beamforming at the BS and passive beamforming[1] at the IRS. For the first subproblem, i.e., the active beamforming optimization problem, we propose the SDR based method to achieve the active beamforming vector for each user, which is different from the Gaussian randomization method.
- In order to initialize a feasible solution for the proposed AO method, based on the beam-splitting technique, we propose a method to achieve an initial feasible solution, which is different from the conventional stochastic method.

[1] In this paper, we only consider the scenario that the reflecting elements of IRS have constant module values and continuous phase shifts. The method for solving the discrete phase shifts scenario is out of the scope of this paper. And we refer the readers to [39] for more details on how to achieve the corresponding passive beamforming.

The rest of this paper is organized as follows. In Section II, we introduce the system model. In section III, two multi-beam selection strategies are proposed. The weighted sum rate maximization (WSRM) problem is also formulated. In Section IV, the optimization problem is solved through the proposed AO method. We then propose the method to initialize the feasible solution for the proposed AO method. In Section V, the simulation results are provided. Finally, we conclude this paper in Section VI.

Notations: a, $\mathbf{a}$, $\mathbf{A}$, $\mathcal{A}$ denote a scalar, a vector, a matrix and a set. $|a|$ and $|\mathcal{A}|$ denote the module value of a complex scalar and the cardinality of a set, respectively. $[\mathbf{a}]_i$ and $\|\mathbf{a}\|$ denote the i-th element and the l_2-norm operation of $\mathbf{a}$, respectively. $\mathrm{diag}\left(\mathbf{a}\right)$ denotes the diagonal matrix of vector $\mathbf{a}$. $\mathbf{a}\otimes\mathbf{b}$ is the Kronecker product of vectors $\mathbf{a}$ and $\mathbf{b}$. The transpose and conjugate transpose of matrix $\mathbf{A}$ are denoted by $\mathbf{A}^T$ and $\mathbf{A}^H$, respectively. $\mathrm{rank}\left(\mathbf{A}\right)$ and $Tr\left(\mathbf{A}\right)$ denotes the rank and the trace of matrix $\mathbf{A}$. $\mathbf{A}\succeq\mathbf{0}$ represents $\mathbf{A}$ is a positive semidefinite matrix. $\mathbb{C}^{M\times N}$ denotes the set of all complex $M\times N$ matrices. All N-dimension complex Hermitian matrices are denoted by $\mathbb{H}^N$. $\mathrm{Re}\left(\cdot\right)$ denotes the real part of a complex variable. $\mathcal{CN}\left(\mu,\sigma^2\right)$ denotes the circularly symmetric complex Gaussian distribution with mean μ and covariance σ^2.

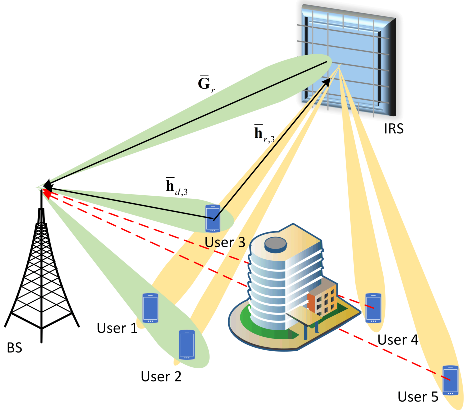


Fig. 1. IRS-aided mmWave beamspace NOMA system.

II. System Model

Consider a downlink IRS-aided mmWave beamspace NOMA communication system as shown in Fig. 1, where the BS is equipped with N antennas and N_{RF} chains, serves K single-antenna users with the aid of one IRS. Denote the user set as $\mathcal{K}$. Denote the set of users with direct-link as $\mathcal{K}'$. The IRS is equipped with M reflecting elements, i.e., antennas.

After performing channel estimation, the estimated up-link beamspace CSI vectors for the k-th user can be expressed as

$$\tilde{\mathbf{h}}_{d,k} = \mathbf{U}^H\bar{\mathbf{h}}_{d,k}, \tag{1a}$$

$$\begin{aligned}\tilde{\mathbf{H}}_{c,k} &= \mathbf{U}^H\bar{\mathbf{H}}_{c,k} = \mathbf{U}^H\left(\bar{\mathbf{G}}_r\mathrm{diag}\left(\bar{\mathbf{h}}_{r,k}\right)\right)\\ &= \tilde{\mathbf{G}}_r\mathrm{diag}\left(\bar{\mathbf{h}}_{r,k}\right),\end{aligned} \tag{1b}$$

where $\tilde{\mathbf{h}}_{d,k}$ denotes $N\times 1$ beamspace CSI vector of the BS-User link (direct-link), the matrix $\mathbf{U}$ of size $N\times N$ denotes the spatial discrete Fourier transformation achieved by the lens antenna array, $\bar{\mathbf{h}}_{d,k}$ denotes the $N\times 1$ spatial CSI vector of the direct-link, $\tilde{\mathbf{H}}_{c,k}$ denotes the $N\times M$ beamspace CSI matrix of the BS-IRS-User link (aided-link or cascaded-link), $\bar{\mathbf{H}}_{c,k}$ denotes the $N\times M$ spatial CSI matrix of the cascaded-link, $\bar{\mathbf{G}}_r$ denotes the spatial CSI matrix of size $N\times M$ between the BS and the IRS, $\bar{\mathbf{h}}_{r,k}$ denotes the spatial CSI vector of size $M\times 1$ between the IRS and the k-th user, and $\tilde{\mathbf{G}}_r$ denotes the beamspace[2] CSI matrix of size $N\times M$ between the BS and the IRS.

In this paper, we assume $\tilde{\mathbf{h}}_{d,k}$ and $\tilde{\mathbf{H}}_{c,k}$ are known by the BS. Note that there is no literature found about how to estimate the CSI in the IRS-aided beamspace system. For the conventional antenna architectures, as stated in [32], there are three main methods to estimate the spatial CSI. Different from the conventional channel estimation method in [28], the compressive-sensing (CS) based methods are proposed to estimate $\tilde{\mathbf{H}}_{c,k}$ in [29] and [30], respectively. Especially, the method proposed in [30] exploits the sparsity[3] of $\tilde{\mathbf{H}}_{c,k}$, which also makes it reasonable to perform beam selection for the cascaded-link. In addition, since the CSI of the cascaded-link in the IRS-aided beamspace system has inherent sparsity, there is no need to introduce the dictionary matrix $\mathbf{A}_{BS}\in{}^{N\times N}$ for the BS as shown in [30]. Then, by introducing the dictionary matrix $\mathbf{A}_R\in{}^{M\times M}$ for the IRS, new method can be designed to estimate the CSI of the cascaded-link in our proposed system, which will be investigated in our future work.

A. 2D Channel Model

For the 2D channel model, we assume that the BS and the IRS are equipped with uniform linear array[4] (ULA). Based on the sparsity of mmWave channel, which makes it reasonable to perform beam selection, we adopt the Saleh-Valenzuela model [13, 16, 23] as the mmWave channel model. Then, without the aid of IRS, the spatial CSI vector of direct-link for the k-th user can be given as

$$\bar{\mathbf{h}}_{d,k} = \alpha_{d,k,0}\mathbf{a}_{BS}\left(\theta_{d,k,0}\right) + \sum_{l=1}^{L}\alpha_{d,k,l}\mathbf{a}_{BS}\left(\theta_{d,k,l}\right), \tag{2}$$

where $\alpha_{d,k,0}$ and $\mathbf{a}_{BS}\left(\theta_{d,k,0}\right)$ denote the complex gain and the steering vector of the line-of-sight (LoS) component of the k-th user, $\alpha_{d,k,l}$ and $\mathbf{a}_{BS}\left(\theta_{d,k,l}\right)$ for $1\le l\le L$ denote the complex gain and the steering vector of the non-line-of-sight (NLoS) component of the k-th user, L denotes the number of NLoS components, $\theta_{d,k,0}$ and $\theta_{d,k,l}$ for $1\le l\le L$ denote the

[2] Without special instruction, in the remainder of this paper, the CSI refers to the beamspace CSI.

[3] Since both the BS and the IRS are always mounted a height, there are only a few paths between the BS and the IRS. As investigated in [30], besides the row sparsity, $\tilde{\mathbf{H}}_{c,k}$ also has column sparsity.

[4] For the IRS, the antennas are uniformly deployed in the linear array and the subsequent planar array. For the BS with lens antenna array, the antennas are uniformly deployed on the focal arc behind the lens [42–44]. In addition, if the IRS is equipped with the UPA, but the BS is equipped with the ULA, SNR loss may be incurred in the received uplink signal for the BS. If the IRS is equipped with the ULA, but the BS is equipped with the UPA, SNR loss may be incurred in the received downlink signal for the IRS. Therefore, in this paper, the BS and the IRS have the same dimension, i.e., either ULA or UPA.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3067938, IEEE Transactions on Vehicular Technology

spatial directions. The steering vector can be expressed as a column vector

$$\begin{aligned}\mathbf{a}_{BS}\left(\theta\right) &= \left[e^{-j2\pi p\theta}\right]_{p\in\mathcal{I}(N)} \\ &= \left[1,\ldots,e^{-j2\pi p\theta},\ldots,e^{-j2\pi(N-1)\theta}\right]^{T},\end{aligned} \tag{3}$$

where $\theta = \frac{d}{\lambda}\sin\left(\phi\right)$ denotes the spatial direction distributed in a range $[-0.5, 0.5]$, ϕ denotes the physical direction distributed in a range $[-\pi/2\ ,\pi/2\ ]$, λ is the wavelength, $d = \lambda/2$ is the antenna spacing, p belongs to $\mathcal{I}\left(N\right)$, i.e., $p \in \mathcal{I}\left(N\right)$, $\mathcal{I}\left(N\right) = \{0, 1, 2, \ldots, N-1\}$, and j is equal to $\sqrt{-1}$.

Similarly, $\bar{\mathbf{h}}_{r,k}$ can be given as

$$\bar{\mathbf{h}}_{r,k} = \alpha_{r,k,0}\mathbf{a}_{R}\left(\theta_{r,k,0}\right) + \sum_{l=1}^{L}\alpha_{r,k,l}\mathbf{a}_{R}\left(\theta_{r,k,l}\right), \tag{4}$$

where $\mathbf{a}_{R}\left(\theta\right)$ denotes the steering vector of the IRS, which has the same form as stated in (3) and can be given as $\mathbf{a}_{R}\left(\theta\right) = \left[e^{-j2\pi p\theta}\right]_{p\in\mathcal{I}(M)}$. Then, $\bar{\mathbf{G}}_{r}$ can be expressed as

$$\bar{\mathbf{G}}_{r} = \alpha_{0}\mathbf{a}_{BS}\left(\theta_{A,0}\right)\mathbf{a}_{R}(\theta_{D,0})^{H} + \sum_{l=1}^{L}\alpha_{l}\mathbf{a}_{BS}\left(\theta_{A,l}\right)\mathbf{a}_{R}(\theta_{D,l})^{H}, \tag{5}$$

where α_0 denotes the complex gain of the LoS path between the BS and the IRS, α_l for $1 \leq l \leq L$ denotes complex gain of the NLoS path, $\theta_{A,0}$ denotes the spatial direction of the angle of arrival (AoA) for the LoS path, $\theta_{A,l}$ for $1 \leq l \leq L$ denotes the spatial direction of the AoA for NLoS path, $\theta_{D,0}$ denotes the spatial direction of the angle of departure (AoD) for the LoS path, and $\theta_{D,l}$ for $1 \leq l \leq L$ denotes the spatial direction of the AoD for the NLoS path.

The matrix $\mathbf{U}$ contains the steering vectors of N directions which cover and equally divide the entire space as follows

$$\mathbf{U} = \frac{1}{\sqrt{N}}\left[\mathbf{a}_{BS}\left(\theta_{1}\right), \mathbf{a}_{BS}\left(\theta_{2}\right), \ldots, \mathbf{a}_{BS}\left(\theta_{N}\right)\right], \tag{6}$$

where $\theta_{q} = \frac{1}{N}\left(q - \frac{N+1}{2}\right)$ for $q = 1, 2, \ldots, N$ denotes the predefined orthogonal spatial direction uniformly distributed in a range $[-0.5, 0.5]$.

B. 3D Channel Model

For the 3D channel condition, we also adopt the Saleh-Valenzuela model [42]. The BS and the IRS are equipped with uniform planar array (UPA), respectively. The BS is equipped with N_1 horizontal antennas, N_2 vertical antennas, and $N_1 \times N_2 = N$. The IRS is equipped with M_1 horizontal antennas, M_2 vertical antennas, and $M_1 \times M_2 = M$. Then, the steering vector of the BS can be expressed as

$$\mathbf{a}_{UPA-BS}\left(\theta_{az}, \theta_{el}\right) = \mathbf{a}_{BS-az}\left(\theta_{az}\right) \otimes \mathbf{a}_{BS-el}\left(\theta_{el}\right). \tag{7}$$

where $\mathbf{a}_{BS-az}\left(\theta_{az}\right) = \left[e^{-j2\pi p\theta_{az}}\right]_{p\in\mathcal{I}(N_1)}$ is the azimuth steering vector with θ_{az} representing the azimuth spatial direction, $\mathbf{a}_{BS-el}\left(\theta_{el}\right) = \left[e^{-j2\pi p\theta_{el}}\right]_{p\in\mathcal{I}(N_2)}$ is the elevation steering vector with θ_{el} representing the elevation spatial direction. Similarly, the steering vector of the IRS can be expressed as

$$\mathbf{a}_{UPA-R}\left(\theta_{az}, \theta_{el}\right) = \mathbf{a}_{R-az}\left(\theta_{az}\right) \otimes \mathbf{a}_{R-el}\left(\theta_{el}\right), \tag{8}$$

where $\mathbf{a}_{R-az}\left(\theta_{az}\right) = \left[e^{-j2\pi p\theta_{az}}\right]_{p\in\mathcal{I}(M_1)}$ is the azimuth steering vector, $\mathbf{a}_{R-el}\left(\theta_{el}\right) = \left[e^{-j2\pi p\theta_{el}}\right]_{p\in\mathcal{I}(M_2)}$ is the elevation steering vector. Then, based on (2), (4) and (5), the spatial CSI vectors for the 3D channel model can be formulated, respectively. For concision, we do not dwell on them here. In addition, under the UPA scenario, the spatial discrete Fourier transformation matrix $\mathbf{U}$ for the BS can be given as

$$\begin{aligned}&\mathbf{U} = \tfrac{1}{\sqrt{N}}\left[\mathbf{a}_{UPA-BS}\left(\theta_{az,1}, \theta_{el,1}\right), \ldots,\right. \\ &\left.\mathbf{a}_{UPA-BS}\left(\theta_{az,q_1}, \theta_{el,q_2}\right), \ldots, \mathbf{a}_{UPA-BS}\left(\theta_{az,N_1}, \theta_{el,N_2}\right)\right],\end{aligned} \tag{9}$$

where $\theta_{az,q_1} = \frac{1}{N_1}\left(q_1 - \frac{N_1+1}{2}\right)$ for $q_1 = 1, 2, \ldots, N_1$ denotes the predefined orthogonal azimuth spatial direction, $\theta_{az,q_2} = \frac{1}{N_2}\left(q_2 - \frac{N_2+1}{2}\right)$ for $q_2 = 1, 2, \ldots, N_2$ denotes the predefined orthogonal elevation spatial direction, θ_{az,q_1} and θ_{el,q_2} are uniformly distributed in a range $[-0.5, 0.5]$.

III. Beam selection Strategies and Problem Formulation

In this section, we first specify the single-beam selection strategy and user grouping strategy for this IRS-aided mmWave beamspace NOMA system. We then propose two multi-beam selection strategies for the cascaded-link under both 2D channel model and 3D channel model, respectively, where two corresponding RF chain configuration strategies are designed, respectively. Finally, the optimization problem for maximizing the weighted sum rate is formulated.

A. Single-beam Selection Strategy

According to the obtained $\tilde{\mathbf{h}}_{d,k}, k \in \mathcal{K}'$, under the 2D channel model and the 3D channel model, we select the beam with maximum channel gain for each user, which is the same as the beam selection strategy adopted in [19]. The index of selected beam for the k-th user can be written as

$$b_k = \max_{i}\left|\left[\tilde{\mathbf{h}}_{d,k}\right]_{i}\right|. \tag{10}$$

If the k-th user is blocked, i.e., $\tilde{\mathbf{h}}_{d,k} = \mathbf{0}$, which means that the user cannot be directly served by the BS, there is no beam selected for this blocked user. For the cascaded-link, based on the sparsity of the cascaded-link, we also select one beam according to the estimated CSI $\tilde{\mathbf{H}}_{c,k}, \forall k \in \mathcal{K}$ and $\tilde{\mathbf{H}}_{c,k} \neq \mathbf{0}$. The index of selected beam for the cascaded-link can be given as

$$b_c = \max_{i}\left|\left[\tilde{\mathbf{h}}_{c,k}\right]_{i}\right| = \max_{i}\left|\left[\sum_{m=1}^{M}\tilde{\mathbf{H}}_{c,k}\left(:,m\right)\right]_{i}\right|. \tag{11}$$

where $\tilde{\mathbf{h}}_{c,k}$ of size $N \times 1$ is equal to $\sum\limits_{m=1}^{M}\tilde{\mathbf{H}}_{c,k}\left(:,m\right)$. Note that any user served by the IRS has the same selected beam. Therefore, the above selected beam can be determined by any user served by the IRS. Considering the channel estimation error, we select the above beam according to the strongest $\tilde{\mathbf{h}}_{c,k}$.

After performing beam selection for all users and the cascaded-limk, the downlink[5] dimension-reduced CSI vector of size $1 \times |\Gamma|$ for the k-th user can be expressed as

$$\begin{aligned}\mathbf{h}_k &= \mathbf{h}_{d,k}^H + \mathbf{u}\mathbf{H}_{c,k}^H \\ &= \mathbf{h}_{d,k}^H + \mathbf{u}\left(\mathbf{G}_r \mathrm{diag}\left(\bar{\mathbf{h}}_{r,k}\right)\right)^H,\end{aligned} \tag{12}$$

where Γ denotes the set of selected beams for all users, $|\Gamma|$ denotes the number of selected beams and equals to the number of activated RF chains, i.e., $|\Gamma| = N_{RF}'$, $\mathbf{h}_{d,k} = \left[\tilde{\mathbf{h}}_{d,k}\right]_{i\in\Gamma}$ denotes the dimension-reduced direct-link CSI vector of size $|\Gamma| \times 1$, $\mathbf{u} = [u_1, u_2, \ldots, u_M]$ denotes the passive beamforming vector of size $1 \times M$ in the IRS, $\mathbf{H}_{c,k} = \tilde{\mathbf{H}}_{c,k}(i,:)_{i\in\Gamma}$ denotes the dimension-reduced cascaded-link CSI matrix of size $|\Gamma| \times M$, and $\mathbf{G}_r = \tilde{\mathbf{G}}_r(i,:)_{i\in\Gamma}$ denotes the dimension-reduced CSI matrix between the BS and the IRS.

For performing the subsequent mathematical transformation, we introduce two new variables as follows:

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{H}_{c,k}^H \\ \mathbf{h}_{d,k}^H \end{bmatrix} = \begin{bmatrix} \mathrm{diag}\left(\bar{\mathbf{h}}_{r,k}^H\right)\mathbf{G}_r^H \\ \mathbf{h}_{d,k}^H \end{bmatrix}, \tag{13}$$

$$\mathbf{v} = [\mathbf{u}, 1]^H. \tag{14}$$

Then, the downlink dimension-reduced CSI vector expressed in (12) can be rewritten as

$$\mathbf{h}_k = \mathbf{v}^H \mathbf{H}_k. \tag{15}$$

In the conventional beamspace NOMA communication system, only the users within the same beam are grouped as one NOMA group, as the user grouping strategy proposed in [19]. In our proposed IRS-aided beamspace NOMA communication system, the IRS can not only strengthen the channel state between the BS and the user, but also facilitate the NOMA transmission among the users without direct-link and within the different beams. Based on the beam selection results, the groups directly served by the BS can be given as

$$\mathcal{G}_B = \left\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{|\mathcal{G}_B|}\right\}, \tag{16}$$

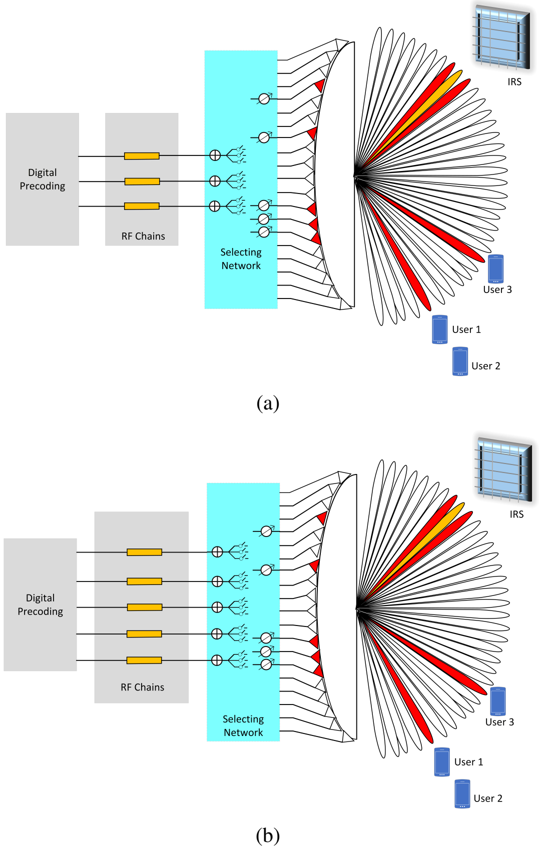
where $|\mathcal{G}_B|$ denotes the number of groups directly served by the BS, $\mathcal{G}_b$ for $1 \le b \le |\mathcal{G}_B|$ denotes the user set corresponding to the b-th group. In this paper, we only consider one IRS scenario, denote the users without direct-link but served by the BS with the aid of IRS as $\mathcal{G}_R$. Therefore, the user grouping solution for all users can be denoted by

$$\mathcal{G} = \{\mathcal{G}_B, \mathcal{G}_R\}, \tag{17}$$

where $|\mathcal{G}|$ denotes the number of groups and $|\mathcal{G}| = |\mathcal{G}_B| + 1$. For the single-beam selection strategy, under both 2D channel model and 3D channel model, the number of activated RF chains is equal to the number of NOMA groups, i.e., $N_{RF}' = |\mathcal{G}|$.

B. Multi-beam Selection Strategies

Note that the multi-beam selection strategy was originally proposed to solve the power leakage problem in the beamspace MIMO system [13]. In [13], based on the beam-aligning precoding method [45], one RF chain is activated for the multiple beams selected for one user, which is different from the single-beam selection strategy proposed in [12]. In this paper, considering the severe path loss of mmWave channel and the per-antenna power constraint, we propose two multi-beam selection strategies for the cascaded-link under 2D channel model and 3D channel model, respectively, where two corresponding RF chain configuration strategies are designed, respectively. The purpose of the RF chain configuration strategies is to provide more power to the IRS. For the user with direct-link, we only select one beam as stated in Subsection A. Next, we will design the two multi-beam selection strategies for the cascaded-link under 2D channel model and 3D channel model, respectively.



(a)

(b)

Fig. 2. Multi-beam selection strategy for 2D channel model.

1) For 2D channel model: For the 2D channel model where the BS and the IRS are equipped with the ULA, based on (11), we firstly select the beam which has the maximum gain as the reference beam for the cascaded-link. As shown in Fig. 2, according to the reference beam which is marked in yellow,

[5] We assume that there exists reciprocity between the uplink channel and the downlink channel.

two adjacent beams can be selected[6].

In the conventional multi-beam beamspace MIMO system [13], only one RF chain is selected. As shown in Fig. 2(a), the third RF chain is connected to the selected three adjacent beams. To increase the power provided to the cascaded-link, we activate three RF chains for the above three selected beams as shown in Fig. 2(b). Due to that one activated RF chain is connected to one selected beam, the beam-aligning can be achieved by performing digital beamforming, where the beam-aligning refers to aligning the channel gains of the selected beams towards the same direction for maximizing the received SNR at each user.

After performing beam selection for all users, the downlink dimension-reduced CSI vector of size $1 \times |\Gamma_{2D}|$ for the k-th user can also be expressed as (12), where Γ_{2D} denotes the set of selected beams for all users, $|\Gamma_{2D}| = |\mathcal{G}| + 2$ denotes the number of selected beams. As shown in Fig. 2(b), $|\Gamma_{2D}|$ is equal to the number of activated RF chains, i.e., $|\Gamma_{2D}| = N'_{RF}$. By performing the same mathematical transformations as stated in (13) and (14), the downlink dimension-reduced CSI vector $\mathbf{h}_k$ can also be rewritten as $\mathbf{h}_k = \mathbf{v}^H \mathbf{H}_k$.

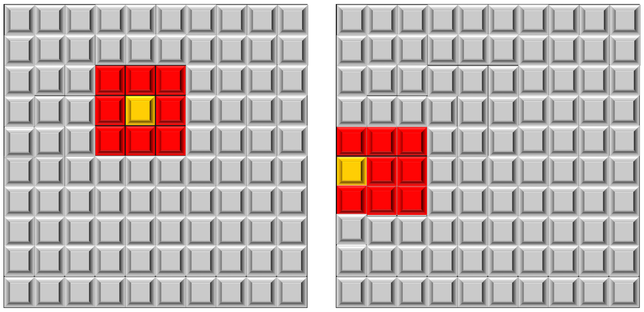

(a)
(b)

Fig. 3. Multi-beam selection strategy for 3D channel model.

2) *For 3D channel model:* For the 3D channel model where the BS and the IRS are equipped with the UPA, the reference beam is selected as stated in the 2D channel model. As shown in Fig. 3, according to the reference beam which is also marked in yellow, eight adjacent beams can be selected. Denote the set of selected beams for all users as Γ_{3D}. To increase the power provided to the cascaded-link, we also activate three RF chains for the above nine selected beams. Denote the number of activated RF chains for all users as N'_{RF}. For the above three activated RF chains, each RF chain is connected to three selected beams, which is different from the connection state as shown in Fig. 2(b). For achieving the beam-aligning, the analog beamforming should be designed, where we adopt the rotation-based precoding algorithm to design the analog beamforming [13].

After performing beam selection for all users, as expressed in (12), under the 3D channel model, the downlink dimension-reduced CSI vector of size $1 \times N'_{RF}$ for the k-th user can be expressed as

$$\begin{aligned}\mathbf{h}_k &= \left(\mathbf{h}_{d,k}^H + \mathbf{u}\mathbf{H}_{c,k}^H\right)\mathbf{W}_{RF} \\ &= \left(\mathbf{h}_{d,k}^H + \mathbf{u}\left(\mathbf{G}_r \mathrm{diag}\left(\bar{\mathbf{h}}_{r,k}\right)\right)^H\right)\mathbf{W}_{RF},\end{aligned} \tag{18}$$

where $\mathbf{W}_{RF}$ denotes the analog beamforming matrix of size $|\Gamma_{3D}| \times N'_{RF}$, $|\Gamma_{3D}|$ denotes the number of selected beams, and N'_{RF} satisfies $N'_{RF} = |\mathcal{G}| + 2 = |\Gamma_{3D}| - 6$. The analog beamforming matrix $\mathbf{W}_{RF}$ can be written as

$$\mathbf{W}_{RF} = \left[\mathbf{w}_1^{RF}, \mathbf{w}_2^{RF}, \ldots, \mathbf{w}_{|\mathcal{G}_B|}^{RF}, \mathbf{w}_{R,1}^{RF}, \mathbf{w}_{R,2}^{RF}, \mathbf{w}_{R,3}^{RF}\right], \tag{19}$$

where $\mathbf{w}_g^{RF}$ for $g = 1, 2, \ldots, |\mathcal{G}_B|$ denotes the analog beamforming vector for the g-th NOMA group, $\mathbf{w}_{R,1}^{RF}$, $\mathbf{w}_{R,2}^{RF}$ and $\mathbf{w}_{R,3}^{RF}$ denote the analog beamforming vectors for the cascaded-link or the NOMA group $\mathcal{G}_R$. Different from the 2D channel model, the analog beamforming within the 3D channel model is achieved through the phase shifting network (PSN), i.e., the selecting network displayed in Fig. 2.

Based on the beam-aligning precoding method proposed in [13], the beam-aligning procedure can be achieved through rotating the beams in $\mathbf{h}_{c,k}, \forall k \in \mathcal{K}$ and $\mathbf{h}_{c,k} \neq \mathbf{0}$, where $\mathbf{h}_{c,k}$ is given as

$$\mathbf{h}_{c,k} = \sum_{m=1}^{M} \mathbf{H}_{c,k}^H(m,:). \tag{20}$$

Then, according to the strongest $\mathbf{h}_{c,k}$, the analog beamforming vector for the i-th $(i = 1, 2, 3)$ RF chain activated for the cascaded-link can be expressed as

$$\frac{\left[\mathbf{w}_{R,j}^{RF}\right]_p}{\left[\mathbf{w}_{R,i}^{RF}\right]_q} = \left(\frac{\left[\mathbf{h}_{c,k}\right]_q}{\left[\mathbf{h}_{c,k}\right]_p}\right) \Bigg/ \left|\frac{\left[\mathbf{h}_{c,k}\right]_q}{\left[\mathbf{h}_{c,k}\right]_p}\right|, \forall q \in \mathcal{B}_i^c, \tag{21}$$

where $\mathcal{B}_i^c$ for $i = 1, 2, 3$ denotes the set of beams connected to the i-th RF chain and $\mathcal{B}^c = \mathcal{B}_1^c \bigcup \mathcal{B}_2^c \bigcup \mathcal{B}_3^c$, $\mathcal{B}_{i'}^c \bigcap \mathcal{B}_{i''}^c = \varnothing$ for $i' \neq i''$, $\mathcal{B}^c$ denotes the set of beams selected for the cascaded-link, $\mathbf{w}_{R,i}^{RF}$ denotes the analog beamforming vector for the i-th RF chain, p denotes the index of the reference beam and $p \in \mathcal{B}_j^c$, $\mathcal{B}_j^c$ denotes the beam set containing the reference beam, and $\left[\mathbf{w}_{R,j}^{RF}\right]_p$ is the reference element corresponding to the reference beam, which is set as 1. Particularly, $\mathbf{w}_{R,i}^{RF}$ for $i = 1, 2, 3$ should satisfy

$$\left|\left[\mathbf{w}_{R,i}^{RF}\right]_q\right| = \begin{cases} \frac{1}{\sqrt{|\mathcal{B}_i^c|}}, q \in \mathcal{B}_i^c, \\ 0, \text{otherwise}. \end{cases} \tag{22}$$

In practice, $\frac{1}{\sqrt{|\mathcal{B}_i^c|}}$ is the power-splitting factor for the i-th RF chain.

About the analog beamforming vector $\mathbf{w}_g^{RF}$, there is only one beam selected for the g-th NOMA group. Then, $\mathbf{w}_g^{RF}$ can be given as

$$\left[\mathbf{w}_g^{RF}\right]_p = \begin{cases} 1, p \in \mathcal{B}_g, \\ 0, \text{otherwise}, \end{cases} \tag{23}$$

where $\mathcal{B}_g$ for $g = 1, 2, \ldots, |\mathcal{G}_B|$ denotes the set of selected beams for the g-th NOMA group and satisfies $|\mathcal{B}_g| = 1$. Based on (13), (14) and (15), under the 3D channel model,

[6] We should mention that the optimality of the proposed multi-beam selection strategy cannot be guaranteed. For dealing with the power leakage problem, the suboptimal multi-beam selection strategy is proposed in [13]. According to the analyses in [13], if the power leakage problem exists, besides the reference beam, there is a great possibility to select two beams being adjacent to the reference beam. That's the reason why we adopt the multi-beam selection strategy as stated in our paper, even it is suboptimal.

the downlink dimension-reduced CSI vector $\mathbf{h}_k$ can also be rewritten as $\mathbf{h}_k = \mathbf{v}^H \mathbf{H}_k$, where $\mathbf{H}_k$ is given as

$$\mathbf{H}_k = \begin{bmatrix} \mathrm{diag}\left(\bar{\mathbf{h}}_{r,k}^H\right) \mathbf{G}_r^H \mathbf{W}_{RF} \\ \mathbf{h}_{d,k}^H \mathbf{W}_{RF} \end{bmatrix}. \tag{24}$$

C. Problem Formulation

In the above analyses, we propose the user grouping strategy for the IRS-aided mmWave beamspace NOMA communication system. Based on the different beam selection strategies and the different RF chain configuration strategies, we construct the downlink dimension-reduced CSI for the 2D channel model and 3D channel model, respectively. The optimization problem for maximizing the weighted sum rate of all users will be formulated next.

Given the user grouping solution $\mathcal{G}$ as shown in (17), denote the digital precoding matrix of size $N'_{RF} \times K$ as

$$\mathbf{W} = \left[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K\right], \tag{25}$$

where $\mathbf{w}_k$ for $k = 1, 2, \ldots, K$ denotes the digital precoding vector of size $N'_{RF} \times 1$ for the k-th user. Without loss of generality, we assume that the users within the g-th NOMA group are indexed in the descending order as

$$\left|\mathbf{h}_{g,1}\mathbf{w}_{g,1}\right| \geq \left|\mathbf{h}_{g,2}\mathbf{w}_{g,2}\right| \geq \cdots \geq \left|\mathbf{h}_{g,|\mathcal{G}_g|}\mathbf{w}_{g,|\mathcal{G}_g|}\right|, \tag{26}$$

where $\mathbf{h}_{g,j}$ for $1 \leq g \leq |\mathcal{G}_g|$, $1 \leq j \leq |\mathcal{G}_g|$ denotes the achieved downlink dimension-reduced CSI of the j-th user within the g-th NOMA group, $\mathbf{w}_{g,j}$ denotes the corresponding digital precoding vector. Then, the received j-th user signal at l-th user within the g-th NOMA group can be represented as

$$\begin{aligned} y_{g,j}^l = & \underbrace{\mathbf{h}_{g,l}\mathbf{w}_{g,j}s_{g,j}}_{\text{desired signal}} + \underbrace{\mathbf{h}_{g,l}\sum_{n=1}^{j-1}\mathbf{w}_{g,n}s_{g,n} + \mathbf{h}_{g,l}\sum_{n=j+1}^{|\mathcal{G}_g|}\mathbf{w}_{g,n}s_{g,n}}_{\text{intra-group interference}} \\ & + \underbrace{\mathbf{h}_{g,l}\sum_{m=1,m\neq g}^{|\mathcal{G}|}\sum_{n=1}^{|\mathcal{G}_m|}\mathbf{w}_{m,n}s_{m,n}}_{\text{inter-group interference}} + v_{g,j}^l, \end{aligned} \tag{27}$$

where $v_{g,j}^l$ denotes the received noise signal at the l-th user within the g-th NOMA group, $s_{g,j}$ for $1 \leq g \leq |\mathcal{G}_g|$, $1 \leq j \leq |\mathcal{G}_g|$ denotes the transmitted signal and satisfies $E\left(s_{g,j}s_{g,j}^H\right) = 1$. Assume that the SIC can be successfully performed among the users. Then, the achievable data rate of decoding the j-th user signal at the l-th user wthin the g-th NOMA group can be written as

$$R_{g,j}^l = \log_2\left(1 + \gamma_{g,j}^l\right) = \log_2\left(1 + \frac{\left|\mathbf{h}_{g,l}\mathbf{w}_{g,j}\right|^2}{I_{g,j}^l + \hat{I}_{g,j}^l + \sigma^2}\right), \tag{28}$$

where $\gamma_{g,j}^l$ denotes the signal to interference and noise ratio (SINR), $I_{g,j}^l = \sum\limits_{n=1}^{j-1}\left|\mathbf{h}_{g,l}\mathbf{w}_{g,n}\right|^2$ denotes the intra-group interference power, $\hat{I}_{g,j}^l = \sum\limits_{m=1,m\neq g}^{|\mathcal{G}|}\sum\limits_{n=1}^{|\mathcal{G}_m|}\left|\mathbf{h}_{g,l}\mathbf{w}_{m,n}\right|^2$ denotes the inter-group interference power, and σ^2 denotes the received noise power.

To satisfy the data rate requirements of all users and ensure that SIC is performed successfully, the achievable data rate of the j-th user within the g-th NOMA group should satisfy $R_{g,j} \geq R_{g,j}^{\min}$, where $R_{g,j}^{\min}$ denotes the corresponding minimum data rate requirement, and $R_{g,j}$ is given as

$$R_{g,j} = \min\left\{R_{g,j}^j, R_{g,j}^{j-1}, \cdots, R_{g,j}^1\right\}. \tag{29}$$

Then, the WSRM problem can be formulated as follows[7]

$$\mathrm{OP1}: \quad \max_{\mathbf{W},\mathbf{v}} R_{sum} = \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\omega_{g,j}R_{g,j} \tag{30a}$$

$$\text{s.t.} \quad R_{g,j} \geq R_{g,j}^{\min}, \forall g, 1 \leq j \leq |\mathcal{G}_g|, \tag{30b}$$

$$\sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\left\|\mathbf{w}_{g,j}\right\|^2 \leq P_{bud}, \tag{30c}$$

$$\begin{aligned} & Tr\left(\mathbf{\Phi}_d\left(\sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\mathbf{w}_{g,j}\mathbf{w}_{g,j}^H\right)\right) \\ & \qquad \leq \rho, 1 \leq d \leq N'_{RF}, \end{aligned} \tag{30d}$$

$$\left|v_m\right|^2 = 1, 1 \leq m \leq M, v_{M+1} = 1. \tag{30e}$$

In OP1, $\omega_{g,j}$ denotes the priority of the j-th user within the g-th NOMA group, which is normalized by $\sum\limits_{g=1}^{|\mathcal{G}|}\sum\limits_{j=1}^{|\mathcal{G}_g|}\omega_{g,j}$, the constraint (30b) guarantees that the achievable data rate of each user should satisfy the minimum data rate requirement and that the SIC can be performed successfully, the constraint (30c) denotes that the total transmit power consumed by the communication system cannot exceed the power budget P_{bud}, the constraint (30d) denotes that the consumed transmit power by each RF chain cannot exceed the per-antenna power constraint, $\mathbf{\Phi}_d$ is given as $\mathbf{\Phi}_d = \mathbf{e}_d\mathbf{e}_d^H$, where the column vector $\mathbf{e}_d$ for $1 \leq d \leq N'_{RF}$ satisfies $\left[\mathbf{e}_d\right]_i = 1$ for $i = d$ and $\left[\mathbf{e}_d\right]_i = 0$ for all $i \neq d$, the constraint (30e) denotes that the module value of each element in the passive beamforming vector $\mathbf{u}$ is equal to 1. In the following section, we discuss how to solve the formulated WSRM problem.

IV. Solution of Weighted Sum Rate Maximization Problem

In this section, an alternating optimization (AO) mbethod is proposed to solve the WSRM problem. Firstly, by performing a series of mathematical transformations, the original problem OP1 can be transformed into a new optimization problem. Secondly, with $\mathbf{v}$ fixed, the optimization problem about $\mathbf{W}$ is formulated and solved through SDR method. Thirdly, with $\mathbf{W}$ fixed, the optimization problem about $\mathbf{v}$ is formulated and solved. Fourthly, based on the beam-splitting technique, we propose the algorithm to initialize the feasible solution for the proposed AO method, which is different from the conventional stochastic method. Finally, the convergence and complexity of the proposed AO method are analyzed.

[7] For the WSRM problem, there exists the optimal SIC decoding order for each NOMA group. However, in the IRS-aided NOMA system, except the traversal method, to our knowledge, there exists no other method to directly achieve the optimal SIC decoding order. Therefore, the formulated WSRM problem herein does not comprise the constraint about SIC decoding order.

A. Problem Transformation

Given the condition imposed by (29), we can find that the objective function and the constraint (30b) are complex. It is intractable to directly solve the original optimization problem OP1. However, by introducing serval auxiliary variables, the original problem can be transformed into a new optimization problem.

About the objective function, by introducing the auxiliary variable $\alpha_{g,j}$, which satisfies $R_{g,j} \geq \alpha_{g,j}$, the objective function can be rewritten as

$$R_{sum} = \sum_{g=1}^{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j} \alpha_{g,j}. \tag{31}$$

Then, the constraint (30b) can be transformed into two new constraints as follows:

$$R_{g,j} \geq \alpha_{g,j}, \tag{32a}$$

$$\alpha_{g,j} \geq R_{g,j}^{\min}. \tag{32b}$$

Based on (29), by introducing auxiliary variable $\eta_{g,j}^l$ and $\vartheta_{g,j}^l$, the constraint (32a) can be transformed into three new constraints as follows:

$$\log_2\left(1 + \frac{1}{\eta_{g,j}^l \vartheta_{g,j}^l}\right) \geq \alpha_{g,j}, \forall g, 1 \leq j \leq |\mathcal{G}_g|, l \leq j, \tag{33a}$$

$$|\mathbf{h}_{g,l}\mathbf{w}_{g,j}|^2 \geq \frac{1}{\eta_{g,j}^l}, \forall g, 1 \leq j \leq |\mathcal{G}_g|, l \leq j, \tag{33b}$$

$$I_{g,j}^l + \hat{I}_{g,j}^l + \sigma^2 \leq \vartheta_{g,j}^l, \forall g, 1 \leq j \leq |\mathcal{G}_g|, l \leq j. \tag{33c}$$

Since the left part of (33a) is a convex function, we can derive that

$$\begin{aligned} \log_2\left(1 + \tfrac{1}{\eta_{g,j}^l \vartheta_{g,j}^l}\right) \geq \psi_{g,j}^l = \log_2\left(1 + \tfrac{1}{\eta_{g,j}^{l,i} \vartheta_{g,j}^{l,i}}\right) \\ - \tfrac{\eta_{g,j}^l - \eta_{g,j}^{l,i}}{\ln(2)\left(\eta_{g,j}^{l,i} + \left(\eta_{g,j}^{l,i}\right)^2 \vartheta_{g,j}^{l,i}\right)} \\ - \tfrac{\vartheta_{g,j}^l - \vartheta_{g,j}^{l,i}}{\ln(2)\left(\vartheta_{g,j}^{l,i} + \eta_{g,j}^{l,i}\left(\vartheta_{g,j}^{l,i}\right)^2\right)} \end{aligned}, \tag{34}$$

where $\psi_{g,j}^l$ is the first order Taylor expansion of $\log_2\left(1 + \frac{1}{\eta_{g,j}^l \vartheta_{g,j}^l}\right)$ with respect to $\left(\eta_{g,j}^{l,i}, \vartheta_{g,j}^{l,i}\right)$. Then, (33a) can be rewritten as

$$\psi_{g,j}^l \geq \alpha_{g,j}, \forall g, 1 \leq j \leq |\mathcal{G}_g|, l \leq j. \tag{35}$$

Due to that $\psi_{g,j}^l$ is a lower bound of $\log_2\left(1 + \frac{1}{\eta_{g,j}^l \vartheta_{g,j}^l}\right)$, compared with (33a), (34) narrows the original feasible region.

According to the above analyses, the original optimization problem OP1 can be transformed into

$$\begin{aligned} \text{OP2}: \quad & \max_{\mathbf{W},\mathbf{v}} R_{sum} = \sum_{g=1}^{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j} \alpha_{g,j} \\ \text{s.t.} \quad & (35), (33b), (33c), (32b), \\ & (30c), (30d), (30e). \end{aligned} \tag{36}$$

Assume that the optimal objective function value of OP1 is $\phi(\mathbf{W}, \mathbf{v})$. Compared with the original problem OP1, due to the introduced Taylor expansion as shown in (34), we can derive that $\phi_{lb}(\mathbf{W}, \mathbf{v}) \leq \phi(\mathbf{W}, \mathbf{v})$, where $\phi_{lb}(\mathbf{W}, \mathbf{v})$ denotes tha optimal objective function value of OP2, the subscript 'lb' denotes the 'lower bound'. Next, we solve the optimization problem OP2 through AO method.

B. Active Beamforming Optimization

Given the passive beamforming vector $\mathbf{v}$, the optimization problem about $\mathbf{W}$ can be expressed as follows

$$\begin{aligned} \text{OP3}: \quad & \max_{\mathbf{W},\kappa} R_{sum} = \sum_{g=1}^{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j} \alpha_{g,j} \\ \text{s.t.} \quad & (35), (33b), (33c), (32b), \\ & (30c), (30d). \end{aligned} \tag{37}$$

In OP3, κ denotes the set of all introduced auxiliary variables. For solving OP3, we define the variable $\mathbf{W}_{g,j}$ for $1 \leq g \leq |\mathcal{G}|$, $1 \leq j \leq |\mathcal{G}_g|$, which satisfies $\mathbf{W}_{g,j} \succeq 0$ and $\text{rank}(\mathbf{W}_{g,j}) = 1$. That means $\mathbf{W}_{g,j}$ is equal to $\mathbf{w}_{g,j}\mathbf{w}_{g,j}^H$, i.e., $\mathbf{W}_{g,j} = \mathbf{w}_{g,j}\mathbf{w}_{g,j}^H$. Then, the optimization problem OP3 can be equivalently transformed into OP4 as follows:

$$\text{OP4}: \quad \max_{\{\mathbf{W}_{g,j}\},\kappa} R_{sum} = \sum_{g=1}^{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j} \alpha_{g,j} \tag{38a}$$

$$\begin{aligned} \text{s.t.} \quad & Tr\left(\mathbf{W}_{g,j}\mathbf{H}_{g,l}^H \mathbf{v}\mathbf{v}^H \mathbf{H}_{g.l}\right) \geq \tfrac{1}{\eta_{g,j}^l}, \\ & \forall g, 1 \leq j \leq |\mathcal{G}_g|, l \leq j, \end{aligned} \tag{38b}$$

$$\begin{aligned} & \textstyle\sum_{n=1}^{j-1} Tr\left(\mathbf{W}_{g,n}\mathbf{H}_{g,l}^H \mathbf{v}\mathbf{v}^H \mathbf{H}_{g,l}\right) + \\ & \textstyle\sum_{m=1,m\neq g}^{|\mathcal{G}|} \sum_{n=1}^{|\mathcal{G}_m|} Tr\left(\mathbf{W}_{m,n}\mathbf{H}_{g,l}^H \mathbf{v}\mathbf{v}^H \mathbf{H}_{g,l}\right) + \\ & \sigma^2 < \vartheta_{g,j}^l, \forall g, 1 \leq j \leq |\mathcal{G}_g|, l \leq j, \end{aligned} \tag{38c}$$

$$\sum_{g=1}^{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}_g|} Tr(\mathbf{W}_{g,j}) \leq P_{bud}, \tag{38d}$$

$$Tr\left(\mathbf{\Phi}_d \sum_{g=1}^{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}_g|} \mathbf{W}_{g,j}\right) \leq \rho, 1 \leq d \leq N_{RF}', \tag{38e}$$

$$\mathbf{W}_{g,j} \succeq 0, \mathbf{W}_{g,j} \in \mathbb{H}^{N_{RF}'}, \forall g, 1 \leq j \leq |\mathcal{G}_g|, \tag{38f}$$

$$\text{rank}(\mathbf{W}_{g,j}) = 1, \forall g, 1 \leq j \leq |\mathcal{G}_g|, \tag{38g}$$

$$(35), (32b). \tag{38h}$$

In OP4, any of the rank-one constraints in (38g) is a non-convex constraint. To deal with the rank-one constraint, we employ the SDR method to completely drop the rank-one constraint. The optimization problem OP4 can be reformulated as one SDP problem given as follows:

$$\begin{aligned} \text{OP5}: \quad & \max_{\{\mathbf{W}_{g,j}\},\kappa} R_{sum} = \sum_{g=1}^{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j} \alpha_{g,j} \\ \text{s.t.} \quad & (35), (38b), (38c), (32b), \\ & (38d), (38e), (38f). \end{aligned} \tag{39}$$

In OP5, all rank-one constraints have been dropped completely. Therefore, the rank of the achieved optimal solution $\mathbf{W}_{g,j}^{*}$ for $1 \leq g \leq |\mathcal{G}|$, $1 \leq j \leq |\mathcal{G}_g|$ by solving OP5 should be clarified. To tackle this issue, we give the following proposition.

Proposition 1 Suppose that the problem OP5 is solvable. Then, there must exist the optimal solution $\mathbf{W}_{g,j}^{*}$ for $1 \leq g \leq |\mathcal{G}|$, $1 \leq j \leq |\mathcal{G}_g|$, which satisfies $\text{rank}\left(\mathbf{W}_{g,j}^{*}\right) = 1$, viz., $\mathbf{W}_{g,j}^{*} = \mathbf{w}_{g,j}^{*}\left(\mathbf{w}_{g,j}^{*}\right)^{H}$.

Proof See Appendix A.

The above proposition indicates that for the SDP problem OP5, there must exist the rank-one optimal solution $\mathbf{W}_{g,j}^{*}$ for $1 \leq g \leq |\mathcal{G}|$, $1 \leq j \leq |\mathcal{G}_g|$. Assume that the optimal solution $\bar{\mathbf{W}}_{g,j}^{*}$ and the corresponding optimal objective value $\bar{R}$ are obtained by solving the SDP problem OP5. If $\bar{\mathbf{W}}_{g,j}^{*}$ does not satisfy the rank-one constraint, i.e., $\text{rank}\left(\bar{\mathbf{W}}_{g,j}^{*}\right) \neq 1$, according to the obtained optimal objective function value $\bar{R}$, the rank-one optimal solution $\mathbf{W}_{g,j}^{*}$ can be achieved by solving the SDP problem OP10. Then, the optimal beamforming vector $\mathbf{w}_{g,j}^{*}$ can be decomposed from $\mathbf{W}_{g,j}^{*}$, i.e., $\mathbf{W}_{g,j}^{*} = \mathbf{w}_{g,j}^{*}\left(\mathbf{w}_{g,j}^{*}\right)^{H}$.

According to the above analyses, the proposed SDR based iterative algorithm for solving the active beamforming is given in **Algorithm 1**, where $I_{\max}$ and ξ_1 denote the permitted maximum number of iterations and the threshold of algorithm termination, respectively.

Algorithm 1 Iterative algorithm for solving the active beamforming

```
Initialization
Set $i = 0$, initialize feasible solution $\left(\mathbf{v}^{i}, \left\{\eta_{g,j}^{l,i}\right\}, \left\{\vartheta_{g,j}^{l,i}\right\}\right)$ and corresponding objective function value $R_i$.
repeat
    Solve the problem OP5, obtain the optimal solution $\left(\left\{\mathbf{W}_{g,j}^{*}\right\}, \left\{\eta_{g,j}^{l,*}\right\}, \left\{\vartheta_{g,j}^{l,*}\right\}\right)$ and $R^{*}$. Let $\eta_{g,j}^{l,i+1} = \eta_{g,j}^{l,*}$, $\vartheta_{g,j}^{l,i+1} = \vartheta_{g,j}^{l,*}$, $R_{i+1} = R^{*}$.
    $i = i + 1$.
until $i = I_{\max}$ or $|R_i - R_{i-1}| \leq \xi_1$.
if $\text{rank}\left(\mathbf{W}_{g,j}^{*}\right) \neq 1$ then
    Based on $R_i$ and $\left(\left\{\eta_{g,j}^{l,i}\right\}, \left\{\vartheta_{g,j}^{l,i}\right\}\right)$, achieve the rank-one optimal solution by solving the problem OP10.
end if
return The optimal active beamforming vector $\left\{\mathbf{w}_{g,j}^{*}\right\}$ through performing matrix decomposition.
```

C. Passive Beamforming Optimization

Given the active beamforming matrix $\mathbf{W}$, we then discuss how to solve the passive beamforming vector $\mathbf{u}$, which is incorporated in the defined variable $\mathbf{v}$ as shown in (14). The passive beamforming optimization problem is given as follows:

$$\text{OP6}: \quad \max_{\mathbf{v},\kappa} R_{sum} = \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j}\alpha_{g,j} \qquad \text{s.t.} \quad (35), (33b), (33c), (32b), (30e). \tag{40}$$

For solving the above problem, by introducing the variable $\mathbf{V} = \mathbf{v}\mathbf{v}^{H}$, which satisfies $\mathbf{V} \succeq 0$ and $\text{rank}\left(\mathbf{V}\right) = 1$, the problem OP6 can be equivalently transformed into

$$\text{OP7}: \quad \max_{\mathbf{V},\kappa} R_{sum} = \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j}\alpha_{g,j} \tag{41a}$$

$$\text{s.t.} \quad Tr\left(\mathbf{V}\mathbf{H}_{g,l}\mathbf{w}_{g,j}\mathbf{w}_{g,j}^{H}\mathbf{H}_{g,l}^{H}\right) \geq \frac{1}{\eta_{g,j}^{l}}, \quad \forall g, 1 \leq j \leq \left|\mathcal{G}_{\rm g}\right|, 1 \leq l \leq j, \tag{41b}$$

$$\sum_{n=1}^{j-1} Tr\left(\mathbf{V}\mathbf{H}_{g,l}\mathbf{w}_{g,n}\mathbf{w}_{g,n}^{H}\mathbf{H}_{g,l}^{H}\right) + \sum_{m=1,m\neq g}^{|\mathcal{G}|}\sum_{n=1}^{|\mathcal{G}_m|} Tr\left(\mathbf{V}\mathbf{H}_{g,l}\mathbf{w}_{m,n}\mathbf{w}_{m,n}^{H}\mathbf{H}_{g,l}^{H}\right) + \sigma^{2} \leq \vartheta_{g,j}^{l}, \forall g, 1 \leq j \leq \left|\mathcal{G}_{\rm g}\right|, 1 \leq l \leq j, \tag{41c}$$

$$Tr\left(\mathbf{\Phi}_m\mathbf{V}\right) = 1, m = 1, 2, \ldots, M+1, \tag{41d}$$

$$\mathbf{V} \succeq 0, \mathbf{V} \in \mathbb{H}^{M+1}, \tag{41e}$$

$$\text{rank}\left(\mathbf{V}\right) = 1, \tag{41f}$$

$$(35), (32b). \tag{41g}$$

For dealing with the non-convex rank-one constraint, instead of dropping the rank-one constraint completely, we adopt the sequential rank-one constraint relaxation (SROCR) method to solve the problem OP7. The SROCR method [39, 46] and penalty-based method [33] are based on the same theory. The penalty-based method in which the Taylor expansion is introduced to deal with the non-convex objective function, may generate suboptimal solution which is far away from the optimal solution. The solution of SROCR method is also suboptimal, but it can infinitely approximate the optimal rank-one solution from the upper bound. That is the reason why we adopt the SROCR method in this paper. The suboptimal rank-one solution can be achieved by iteratively solving the problem as follows:

$$\text{OP8}: \quad \max_{\mathbf{V},\kappa} R_{sum} = \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j}\alpha_{g,j} \tag{42a}$$

$$\text{s.t.} \quad \mathbf{u}_{\max}\left(\mathbf{V}^{i}\right)^{H}\mathbf{V}\mathbf{u}_{\max}\left(\mathbf{V}^{i}\right) \geq \beta^{i} Tr\left(\mathbf{V}\right), \tag{42b}$$

$$(35), (41b), (41c), (32b), (41d), (41e). \tag{42c}$$

In OP8, (42b) guarantees that the achieved solution satisfies the rank-one constraint step by step, where $\mathbf{V}^{i}$ denotes the value of $\mathbf{V}$ in the i-th iteration, $\mathbf{u}_{\max}\left(\mathbf{V}^{i}\right)$ denotes the eigenvector correponding to the maximum eigenvalue of $\mathbf{V}^{i}$, i.e., $\lambda_{\max}\left(\mathbf{V}^{i}\right)$, $\beta^{i} \in [0, 1]$ denotes a relaxation parameter that controls the largest eigenvalue to trace ratio of $\mathbf{V}$.

The iterative algorithm for solving passive beamforming is shown in **Algorithm 2**, where $I_{\max}'$, ξ_2 and ξ_2' denote the permitted maximum number of iterations, the permitted threshold of relaxation difference and the permitted threshold of data rate difference, respectively. Note that in Algorithm 2, the initial feasible solution $\left(\mathbf{V}^0, \left\{\eta_{g,j}^{l,0}\right\}, \left\{\vartheta_{g,j}^{l,0}\right\}\right)$ is the optimal solution achieved by solving the relaxed problem OP8 with $\beta^i = 0$. When we solve the above relaxed problem, the optimal solution achieved from OP5 is taken as initial solution. Before solving OP5, the key step is to achieve the initial feasible solution to problem OP5, which is addressed in next subsection.

Algorithm 2 Iterative algorithm for solving the passive beamforming

```
Initialization
Set $i = 0$, by solving the relaxed problem OP8 with $\beta^i = 0$, initialize feasible solution $\left(\mathbf{V}^i, \left\{\eta_{g,j}^{l,i}\right\}, \left\{\vartheta_{g,j}^{l,i}\right\}\right)$ and achieve the corresponding objective function value $R_i$. Set the initial step size $\delta^i = \left(0, 1-\lambda_{\max}\left(\mathbf{V}^i\right)/Tr\left(\mathbf{V}^i\right)\right]$. Set the initial relaxation parameter $\beta^i \in (0,1)$.
repeat
  Given $\left(\beta^i, \mathbf{V}^i, \left\{\eta_{g,j}^{l,i}\right\}, \left\{\vartheta_{g,j}^{l,i}\right\}\right)$, solve problem OP8.
  if problem OP8 is solvable then
    Obtain the optimal solution $\left(\mathbf{V}^i, \left\{\eta_{g,j}^{l,i}\right\}, \left\{\vartheta_{g,j}^{l,i}\right\}\right)$ and $R^*$. Let $\mathbf{V}^{i+1} = \mathbf{V}^*$, $\eta_{g,j}^{l,i+1} = \eta_{g,j}^{l,*}$, $\vartheta_{g,j}^{l,i+1} = \vartheta_{g,j}^{l,*}$, $R_{i+1} = R^*$, $\delta^{i+1} = \delta^i$.
  else
    Let $\mathbf{V}^{i+1} = \mathbf{V}^i$, $\eta_{g,j}^{l,i+1} = \eta_{g,j}^{l,i}$, $\vartheta_{g,j}^{l,i+1} = \vartheta_{g,j}^{l,i}$, $R_{i+1} = R_i$, $\delta^{i+1} = \delta^i/2$.
  end if
  $\beta^{i+1} = \min\left(1, \frac{\lambda_{\max}\left(\mathbf{V}^{i+1}\right)}{Tr\left(\mathbf{V}^{i+1}\right)} + \delta^{i+1}\right)$,
  $i = i + 1$.
until $\left|1-\beta^i\right| \le \xi_2$ and $\left|R_i - R_{i-1}\right| \le \xi_2'$ or $i = I_{\max}'$.
return $\mathbf{v}^*$.
```

D. Initialization of Feasible Solution

Let us consider how to achieve the initial feasible solution $\left(\mathbf{v}^0, \left\{\eta_{g,j}^{l,0}\right\}, \left\{\vartheta_{g,j}^{l,0}\right\}\right)$ to problem OP5, where the conditional[8] transmit power consumption minimization problem is formulated and solved.

Different from the stochastic method, based on the beam-splitting technique [18], we can make the IRS generate multiple beams for some specific users. The ‘specific user’ refers to the user without direct-link and served by the BS with the aid of IRS, i.e., the user within $\mathcal{G}_R$. If there is no such specific user, we can just choose one user within $\mathcal{G}_B$ as the specific user.

Assume that there exist two specific users, e.g., user k_1 and user k_2, for the ULA, the IRS can be divided into two adjacent subarrays for the above two users, where the number of antennas within the subarrays are M_{k_1} and M_{k_2}, respectively, and satisfy $M_{k_1} + M_{k_2} = M$. The two beamforming vectors for user k_1 and user k_2, respectively, are written as

$$\mathbf{w}\left(M_{k_1}, \theta_{k_1}\right) = \left[1, \ldots, e^{-j2\pi\theta_{k_1}\left(M_{k_1}-1\right)}\right]^T, \tag{43}$$

$$\mathbf{w}\left(M_{k_2}, \theta_{k_2}\right) = e^{-j2\pi M_{k_1}\theta_{k_1}}\left[1, \ldots, e^{-j2\pi\theta_{k_2}\left(M_{k_2}-1\right)}\right]^T, \tag{44}$$

where θ_{k_1} and θ_{k_2} denote the spatial direction of the strongest path for the above two users[9], respectively. Then, the beamforming vector of the IRS can be expressed as

$$\mathbf{w} = \left[\mathbf{w}^T\left(M_{k_1}, \theta_{k_1}\right), \mathbf{w}^T\left(M_{k_2}, \theta_{k_2}\right)\right]^T. \tag{45}$$

For multiple specific users scenario, if M is divisible by $|\mathcal{G}_R|$, each subarray has the same number of antennas, i.e., $\frac{M}{|\mathcal{G}_R|}$. Otherwise, except for the last user, each user has $\lfloor M/|\mathcal{G}_R| \rfloor$ antennas and the last user has $\lfloor M/|\mathcal{G}_R| \rfloor + \left(M - \left(|\mathcal{G}_R| - 1\right)\lfloor M/|\mathcal{G}_R| \rfloor\right)$ antennas. For the single specific user scenario, e.g., user k_3, all antennas are used to perform the analog beamforming. Then, the beamforming vector of the IRS can be expressed as

$$\mathbf{w} = [1, \ldots, e^{-j2\pi\theta_{k_3}(M-1)}]^T, \tag{46}$$

where θ_{k_3} denotes the spatial direction of the strongest path for the single specific user. As shown in Fig. 4(a), the beamforming gain of IRS with ULA for the two specific users and the single specific user are displayed, respectively, where $M = 64$, $\phi_{k_1} = -0.5806$, $\phi_{k_2} = 0.7961$, $\phi_{k_3} = -0.9433$. Note that ϕ denotes the physical direction corresponding to spatial direction θ.

For the UPA, the IRS can generate multiple beams through dividing the IRS into multiple sub-UPAs. In practice, most of the users are located on the ground, which means that most of the users can be spatially separated through the azimuth. Therefore, in this paper, the IRS will be divided into multiple sub-UPAs from the azimuth. The beamforming vectors for the two specific users k_1 and user k_2, respectively, can be given as

$$\begin{aligned}&\mathbf{w}\left(M_{1,k_1}, \theta_{az,k_1}, \theta_{el,k_1}\right) = \\ &\left[1, \ldots, e^{-j2\pi\theta_{az,k_1}\left(M_{1,k_1}-1\right)}\right]^T \otimes \mathbf{a}_{R-el}\left(\theta_{el,k_1}\right),\end{aligned} \tag{47}$$

$$\begin{aligned}&\mathbf{w}\left(M_{1,k_2}, \theta_{az,k_2}, \theta_{el,k_2}\right) = e^{-j2\pi M_{1,k_1}\theta_{az,k_1}} \\ &\left[1, \ldots, e^{-j2\pi\theta_{az,k_2}\left(M_{1,k_2}-1\right)}\right]^T \otimes \mathbf{a}_{R-el}\left(\theta_{el,k_2}\right),\end{aligned} \tag{48}$$

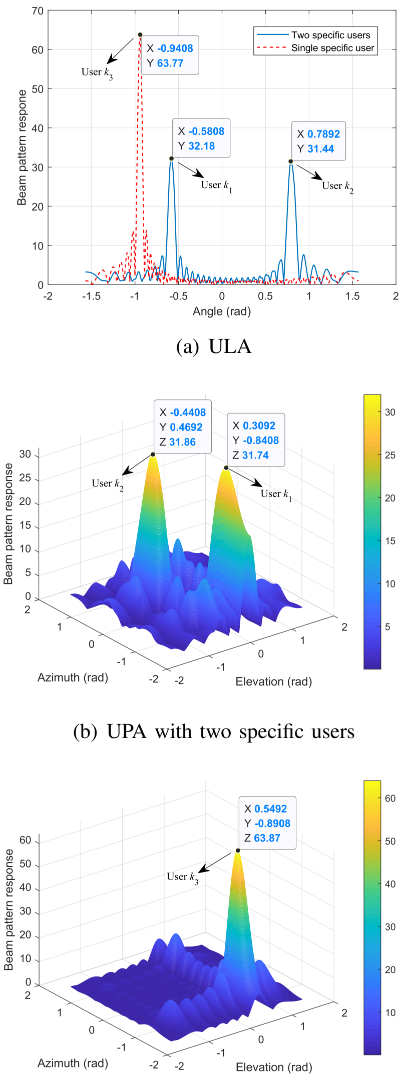
where M_{1,k_1} and M_{2,k_2} denote the number of antennas allocated to user k_1 and user k_2, respectively[10], and satisfy $M_{1,k_1} + M_{1,k_2} = M_1$, θ_{az,k_1} and θ_{az,k_2} denote the azimuth spatial direction of the strongest path for the above two users, respectively, θ_{el,k_1} and θ_{el,k_2} denote the elevation spatial direction of the strongest path for the above two users, respectively, and $\mathbf{a}_{R-el}$ is shown in (8). Then, the beamforming vector of the IRS can be expressed as

$$\mathbf{w} = \left[\mathbf{w}^T\left(M_{1,k_1}, \theta_{az,k_1}, \theta_{el,k_1}\right), \mathbf{w}^T\left(M_{1,k_2}, \theta_{az,k_2}, \theta_{el,k_2}\right)\right]^T. \tag{49}$$

[8] The ‘conditional’ refers to the case that the achieved solution is not optimal, which is restricted by the formulated transmit power consumption minimization problem.

[9] In the beamspace MIMO communication system, it is easy to obtain the angle information of the user-BS, user-IRS and IRS-BS by performing the downlink beam alignment. We refers the readers to [47] for more information about the beam alignment.

[10] For the multiple specific users scenario, the azimuth antenna allocation strategy is same as the antenna allocation strategy adopted for the ULA.



(a) ULA

(b) UPA with two specific users

(c) UPA with single specific user

Fig. 4. Antenna array beam pattern response of IRS.

For the single specific user scenario, e.g., user k_3, all azimuth ntennas are used to perform the azimuth beamforming. Then, the beamforming vector of the IRS can be expressed as

$$\mathbf{w} = \mathbf{a}_{UPA-R}\left(\theta_{az,k_3}, \theta_{el,k_3}\right), \tag{50}$$

where θ_{az,k_3} and θ_{el,k_3} denote the azimuth spatial direction and elevation spatial direction of the strongest path for the user k_3, respectively. The beamforming gain of IRS with UPA for the two specific users and the single specific use are displayed in Fig. 4(b) and Fig. 4(c), respectively, where $M_1 = M_2 = 8$, $\phi_{az,k_1} = -0.8322$, $\phi_{el,k_1} = 0.3048$, $\phi_{az,k_2} = 0.4625$, $\phi_{el,k_2} = -0.4275$, $\phi_{az,k_3} = -0.8943$, $\phi_{el,k_3} = 0.5390$.

The passive beamforming vector $\mathbf{u}$ is equal to $\mathbf{w}^T$, i.e., $\mathbf{u} = \mathbf{w}^T$. Then, as stated in Section III, the downlink dimension-reduced CSI vector $\mathbf{h}_k$ can be achieved. Given the user grouping solution $\mathcal{G}$ as shown in (17) and $\mathbf{h}_k$ for $k \in \mathcal{K}$, based on the maximal ratio transmission (MRT) precoding technique, the corresponding normalized digital precoding vector for the k-th user can be written as $\hat{\mathbf{w}}_k = \frac{\mathbf{h}_k^H}{\left\|\mathbf{h}_k^H\right\|}$. Then, the received signal as shown in (27) can be rewritten as

$$\begin{aligned} y_{g,j}^l =& \underbrace{\mathbf{h}_{g,l}\hat{\mathbf{w}}_{g,j}\sqrt{p_{g,j}}s_{g,j}}_{\text{desired signal}} \\ &+ \underbrace{\mathbf{h}_{g,l}\sum_{n=1}^{j-1}\hat{\mathbf{w}}_{g,n}\sqrt{p_{g,n}}s_{g,n} + \mathbf{h}_{g,l}\sum_{n=j+1}^{|\mathcal{G}_g|}\hat{\mathbf{w}}_{g,n}\sqrt{p_{g,n}}s_{g,n}}_{\text{intra-group interference}} \\ &+ \underbrace{\mathbf{h}_{g,l}\sum_{m=1,m\neq g}^{|\mathcal{G}|}\sum_{n=1}^{|\mathcal{G}_m|}\hat{\mathbf{w}}_{m,n}\sqrt{p_{m,n}}s_{m,n}}_{\text{inter-group interference}} + v_{g,j}^l, \end{aligned} \tag{51}$$

where $p_{g,j}$ denotes the power allocated to the j-th user within the g-th NOMA group. Then, the transmit power consumption minimization problem can be formulated as follows:

$$\text{OP9}: \quad \min_{\mathbf{p}} \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} p_{g,j} \tag{52a}$$

$$\text{s.t.} \quad R_{g,j} \geq R_{g,j}^{\min}, \forall g, 1 \leq j \leq |\mathcal{G}_g|, \tag{52b}$$

$$\sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} p_{g,j} \leq P_{bud}, \tag{52c}$$

$$\sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} \left|\left[\hat{\mathbf{w}}_{g,j}\right]_d\right|^2 p_{g,j} \leq \rho, 1 \leq d \leq N_{RF}'. \tag{52d}$$

In OP9, $\mathbf{p}$ denotes the power allocation vector containing the power allocation factors for all users, i.e., $p_{g,j}$ for $1 \leq g \leq |\mathcal{G}|$, $1 \leq j \leq |\mathcal{G}_g|$. Since the SIC constraint as shown in (52b) can be directly rewritten as a convex constraint [23], the problem OP9 is a convex optimization problem, which is easy to be solved.

Given the optimal solution $\mathbf{p}$, the digital precoding vector $\mathbf{w}_{g,j}$ as shown in (27) can be expressed as $\mathbf{w}_{g,j} = \hat{\mathbf{w}}_{g,j}\sqrt{p_{g,j}}$. Then, based on (33b) and (33c), the initial feasible solution $\eta_{g,j}^{l,0}$ and $\vartheta_{g,j}^{l,0}$ can be achieved as $\eta_{g,j}^{l,0} = 1/\left|\mathbf{h}_{g,l}\mathbf{w}_{g,j}\right|^2$ and $I_{g,j}^l + \hat{I}_{g,j}^l + \sigma^2 = \vartheta_{g,j}^{l,0}$. Therefore, the initial feasible solution $\left(\mathbf{v}^0, \left\{\eta_{g,j}^{l,0}\right\}, \left\{\vartheta_{g,j}^{l,0}\right\}\right)$ to problem OP5 is achieved, where $\mathbf{v}^0$ is equal to $[\mathbf{u}, 1]^H$, i.e., $\mathbf{v}^0 = [\mathbf{u}, 1]^H$, and $\mathbf{u} = \mathbf{w}^T$.

E. Algorithm Analysis

Based on the above proposed two iterative algorithms and the initialization of feasible solution, the proposed AO method for solving the joint active and passive beamforming optimization problem is summarized in **Algorithm 3**, where $T_{\max}$ and ξ_3 denote the permitted maximum number of iterations and the threshold of algorithm termination, respectively. In the proposed AO method, the Taylor expansion as shown in (34) is introduced in problem OP3 and problem OP6. In addition, for solving problem OP6, the SROCR method can approximate the optimal rank-one solution from the upper bound by introducing the rank-one relaxation constraint. However, it is easy to get the near rank-one solution. Therefore, for each subproblem, we just consider the effect of the introduced

Taylor expansion to optimality, which means the achieved objective function value is the lower bound of the optimal value. Assume that $\phi_{lb}^{\mathbf{v}}\left(\mathbf{v}^t,\left\{\mathbf{w}_{g,j}^t\right\}\right)$ is the achieved objective function value after performing the t-th iteration. In the $(t+1)$-th iteration, by performing the step 3 in Algorithm 3, the suboptimal objective function value $\phi_{lb}^{\mathbf{w}}\left(\mathbf{v}^t,\left\{\mathbf{w}_{g,j}^{t+1}\right\}\right)$ can be achieved and satisfies $\phi_{lb}^{\mathbf{v}}\left(\mathbf{v}^t,\left\{\mathbf{w}_{g,j}^t\right\}\right) \leq \phi_{lb}^{\mathbf{w}}\left(\mathbf{v}^t,\left\{\mathbf{w}_{g,j}^{t+1}\right\}\right)$. By performing the step 4 in Algorithm 3, the suboptimal objective function value $\phi_{lb}^{\mathbf{v}}\left(\mathbf{v}^{t+1},\left\{\mathbf{w}_{g,j}^{t+1}\right\}\right)$ can be obtained and satisfies $\phi_{lb}^{\mathbf{w}}\left(\mathbf{v}^t,\left\{\mathbf{w}_{g,j}^{t+1}\right\}\right) \leq \phi_{lb}^{\mathbf{v}}\left(\mathbf{v}^{t+1},\left\{\mathbf{w}_{g,j}^{t+1}\right\}\right)$. Then, we can derive that $\phi_{lb}^{\mathbf{v}}\left(\mathbf{v}^t,\left\{\mathbf{w}_{g,j}^t\right\}\right) \leq \phi_{lb}^{\mathbf{v}}\left(\mathbf{v}^{t+1},\left\{\mathbf{w}_{g,j}^{t+1}\right\}\right)$, i.e., the proposed AO method is convergent.

For evaluating the complexity of the proposed AO method, we assume that there only exist two users in each NOMA group, and N'_{RF} RF chains are activated to serve all K users. According to [39, 48], the main complexity of solving each subproblem with interior-point method are given as

$$C_{sub-1} = \mathcal{O}\left(\max\left\{\left(3K+1+N'_{RF}\right),\left(KN'_{RF}\right)\right\}^4 \right. \\ \left. * \sqrt{KN'_{RF}}\log\left(1/\xi_1\right)\right), \tag{53a}$$

$$C_{sub-2} = \mathcal{O}\left(\left(3K+M+1\right)^4\sqrt{M+1}\log\left(1/{\xi'}_2\right)\right. \\ \left. +\left(3K+M+2\right)^4\sqrt{M+1}\log\left(1/{\xi'}_2\right)\right). \tag{53b}$$

Then, the total complexity of the proposed AO method is given as $\mathcal{O}\left(T_{\max}\left(C_{sub-1}+C_{sub-2}\right)\right)$.

Algorithm 3 The proposed AO method

```
1: Initialization
   Set $t = 0$, by solving the problem OP9, initialize feasible solution $\left(\mathbf{v}^t,\left\{\eta_{g,j}^t\right\},\left\{\vartheta_{g,j}^t\right\}\right)$ and achieve the corresponding objective function value $R_t = \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\omega_{g,j}\alpha_{g,j}$.
2: repeat
3:     Given $\left(\mathbf{v}^t,\left\{\eta_{g,j}^{l,t}\right\},\left\{\vartheta_{g,j}^{l,t}\right\}\right)$, achieve the optimal solution $\left(\left\{\mathbf{w}_{g,j}^*\right\},\left\{\eta_{g,j}^{l,*}\right\},\left\{\vartheta_{g,j}^{l,*}\right\}\right)$ and $R^*$ by solving the problem OP3. Let $\mathbf{w}_{g,j}^{t+1} = \mathbf{w}_{g,j}^*$, $\eta_{g,j}^{l,t+1} = \eta_{g,j}^{l,*}$, $\vartheta_{g,j}^{l,t+1} = \vartheta_{g,j}^{l,*}$, $R^{t+1} = R^*$.
4:     Given $\left(\mathbf{v}^t,\left\{\mathbf{w}_{g,j}^{t+1}\right\},\left\{\eta_{g,j}^{l,t+1}\right\},\left\{\vartheta_{g,j}^{l,t+1}\right\}\right)$, achieve the optimal solution $\left(\mathbf{v}^*,\left\{\eta_{g,j}^{l,*}\right\},\left\{\vartheta_{g,j}^{l,*}\right\}\right)$ and $R^*$ by solving the problem OP6. Let $\mathbf{v}^{t+1} = \mathbf{v}^*$, $\eta_{g,j}^{l,t+1} = \eta_{g,j}^{l,*}$, $\vartheta_{g,j}^{l,t+1} = \vartheta_{g,j}^{l,*}$, $R^{t+1} = R^*$.
5:     $t = t + 1$.
6: until $t = T_{\max}$ or $\left|R^t - R^{t-1}\right| \leq \xi_3$.
7: return $\left\{\mathbf{w}_{g,j}^*\right\}$ and $\mathbf{v}^*$.
```

V. Simulations

In this section, we verify the performance of the proposed IRS-aided mmWave beamspace NOMA communication system via simulations. We consider a downlink mmWave massive MIMO with lens antenna array communication system, where the BS has $N = 256$ antennas and serves K single-antenna users, the IRS has M antennas. The number of users with the direct-link is K_1. The SNR is defined as $\frac{E_b}{\sigma^2}$, where E_b denotes the expected power of transmitted signal s and $E_b = 1$. The bandwidth[11] is set as 1. About the channel model, the channel for the user-IRS, user-BS and IRS-BS are assumed to contain four components: 1 LoS component and 3 NLoS components. As stated in [19, 23], the channel parameters conform to the conditions as follows: $\alpha_0 \sim \mathcal{CN}\left(0{,}1\right)$, $\alpha_l \sim \mathcal{CN}\left(0{,}0.1\right)$ for $1 \leq l \leq 3$, θ_0 and θ_l for $1 \leq l \leq 3$ follow the uniform distribution within $[-0.5, 0.5]$.

A. Convergence Analysis

For evaluating the convergence of the proposed AO method, we set $M = 16$, K=5, $K_1 = 3$, $\rho = 50$ mW, $SNR = 20$ dB, $P_{bud} = \rho N'_{RF}$, $I_{\max} = I'_{\max} = T_{\max} = 100$. In this part, for indicating the convergence procedure, we set $\xi_1 = \xi_2 = \xi'_2 = \xi_3 = 0.001$.

From Fig. 5, we can find that the proposed AO method is convergent. Note that even though $T_{\max}$ is not reached, the iteration will terminate when the termination threshold is reached. For the proposed IRS-aided beamspace NOMA scheme, about the SE, we can find that for the ULA, the proposed Multi-beam scheme has higher SE performance than the Single-beam scheme. For the UPA, the same result can be found. The reasons are stated as follows. Firstly, considering the per-antenna power constraint, the proposed two multi-beam schemes for ULA and UPA, respectively, can provide more power to the cascaded-link than the corresponding Single-beam scheme. Secondly, the proposed two multi-beam schemes for ULA and UPA, respectively, can handle the power leakage problem in beamspace channel as stated in [13].

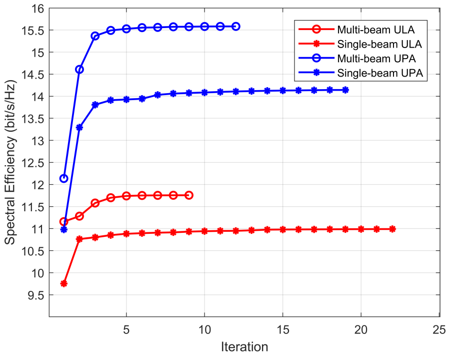


Fig. 5. The convergence of AO method.

B. Spectral Efficiency Versus Per-antenna Power Constraint and Performance Comparison with OMA

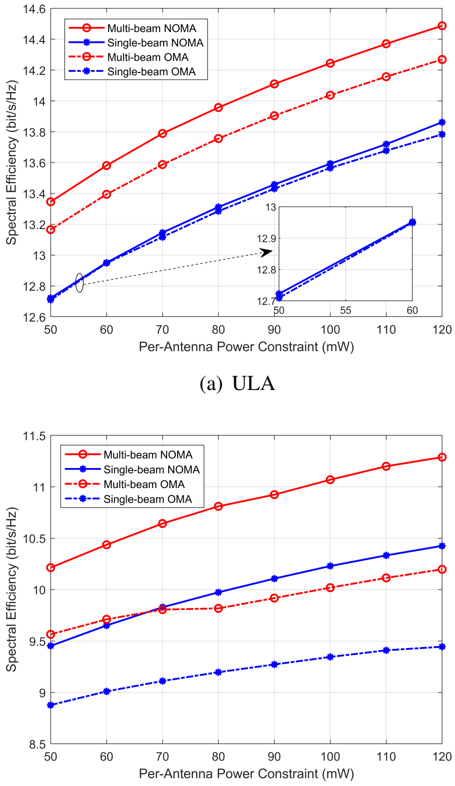
For evaluating the impact of the per-antenna power constraint and performing comparison with IRS-aided beamspace OMA scheme, the per-antenna power constraint is set ranging from 50 to 120 mW. The other parameters are set as in Subsection A. In the IRS-aided beamspace OMA scheme, the

[11] We consider the normalized bandwidth in this paper. Therefore, the achievable sum rate can also be defined as the spectral efficiency (bit/s/Hz).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3067938, IEEE Transactions on Vehicular Technology

users within one NOMA group are served through TDMA, where the equal time allocation strategy is adopted.

For the ULA, as shown in Fig. 6(a), under the same per-antenna power constraint, the proposed Multi-beam NOMA scheme has higher SE performance than the Multi-beam OMA scheme, and the Single-beam NOMA scheme also has the higher SE performance than the Single-beam OMA scheme under the same per-antenna power constraint. As the increase of the per-antenna power constraint, which means that more power can be provided to the cascaded-link, the SEs of above four schemes increase as shown in Fig. 6(a). Based on the same reason, under the same per-antenna power constraint, the proposed Multi-beam NOMA scheme has higher SE than the Single-beam NOMA scheme, which is same with the result shown in Fig. 5. And the Multi-beam OMA scheme also has higher SE than the Single-beam OMA scheme. For the UPA, according to Fig. 6(b), all above conclusions can also be drawn.



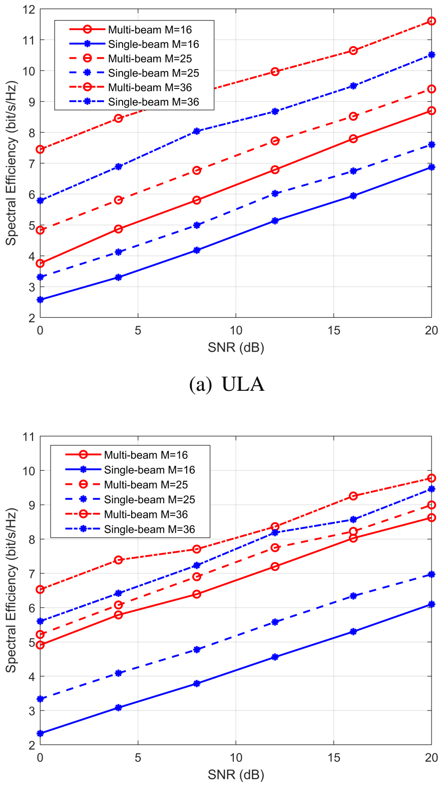
(a) ULA

(b) UPA

Fig. 6. Spectral efficiency versus per-antenna power constraint and performance comparison with OMA.

C. Spectral Efficiency Versus Number of Antennas in IRS

For evaluating the impact of number of antennas in IRS to SE, we set $K = 3$, $K_1 = 1$, and three different M values are set as 16, 25 and 36. The other parameters are set as in Subsection A. For the proposed IRS-aided beamspace NOMA scheme, from Fig. 7(a), we can find that for the ULA, with the same number of antennas in IRS, the proposed Multi-beam scheme has higher SE than the Single-beam scheme. For both Multi-beam scheme and Single-beam scheme, the higher SE can be achieved with larger number of antennas M, due to the beamforming gain provided by the IRS. For the UPA, according to Fig. 7(b), we can draw the same conclusions as stated before.



(a) ULA

(b) UPA

Fig. 7. Spectral efficiency versus number of antennas in IRS.

VI. Conclusion

In this paper, we have proposed the IRS-aided mmWave beamspace NOMA communication system. Considering the power leakage problem in beamspace channel and the per-antenna power constraint, two multi-beam selection strategies are proposed for the cascaded-link under both 2D channel model and 3D channel model, respectively, where two corresponding RF chain configuration strategies are designed, respectively. Based on the proposed communication system, the weighted sum rate maximization (WSRM) problem is formulated and solved through the proposed AO method. Especially, based on the beam-splitting technique, we propose the method to initialize the feasible solution for the AO method, in which the conditional transmit power minimization problem is formulated and solved. The weighted sum rate performance of the proposed communication system is evaluated by simulations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3067938, IEEE Transactions on Vehicular Technology

Appendix A
Proof of Proposition 1

We formulate a SDP problem for minimizing the transmit power consumption as follows:

$$\begin{aligned} \text{OP10}: \quad & \min_{\{\mathbf{W}_{g,j}\},\kappa} \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} Tr\left(\mathbf{W}_{g,j}\right) \\ \text{s.t.} \quad & \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} \omega_{g,j}\alpha_{g,j} \geq \bar{R}, \\ & (35), (38b), (38c), (38d), (32b), \\ & (38e), (38f). \end{aligned} \tag{54}$$

In OP10, $\bar{R}$ denotes the obtained optimal objective function value by solving the problem OP5. Next, we examine whether the optimal solution $\mathbf{W}_{g,j}^{*}$ achieved by solving the problem OP10 satisfies the rank-one constraint or not.

$$\begin{aligned} \mathcal{L}\left(\mathcal{X}\right) = & \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} Tr\left(\mathbf{W}_{g,j}\right) + \tau\left(\bar{R} - \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\omega_{g,j}\alpha_{g,j}\right) \\ & + \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\sum_{l=1}^{j}\mu_{g,j,l}\left(\alpha_{g,j} - \psi_{g,j}^{l}\right) \\ & + \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\sum_{l=1}^{j} r_{g,j,l}\left(\frac{1}{\eta_{g,j}^{l}} - Tr\left(\mathbf{W}_{g,j}\mathbf{H}_{g,l}^{H}\mathbf{v}\mathbf{v}^{H}\mathbf{H}_{g.l}\right)\right) \\ & + \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\sum_{l=1}^{j}\beta_{g,j,l}\left(Q_{g,j}^{l} - \vartheta_{g,j}^{l}\right) \\ & + \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\phi_{g,j}\left(R_{g,j}^{\min} - \alpha_{g,j}\right) \\ & + \nu\left(\sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|} Tr\left(\mathbf{W}_{g,j}\right) - P_{bud}\right) \\ & + \sum_{d=1}^{N'_{RF}}\varphi_d\left(Tr\left(\mathbf{\Phi}_d\sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\mathbf{W}_{g,j}\right) - \rho\right) \\ & + \sum_{g=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\mathbf{A}_{g,j}\mathbf{W}_{g,j}. \end{aligned} \tag{55}$$

Denote the left side of (38c) as $Q_{g,j}^{l}$. The Lagrange dual function of problem OP10 can be written as (55), where $\mathcal{X}$ denotes a collection of all the primal and dual variables of problem OP10, τ, $\mu_{g,j,l}$, $r_{g,j,l}$, $\beta_{g,j,l}$, $\phi_{g,j}$, ν, φ_d and $\mathbf{A}_{g,j}$ denote the Lagrange multipliers. The Karush-Kuhn-Tucker (KKT) conditions of problem OP10 for the optimal solution $\mathbf{W}_{q,e}^{*}$ for $1 \leq q \leq |\mathcal{G}|$, $1 \leq e \leq |\mathcal{G}_q|$ can be written as follows[12]:

$$\begin{aligned} & r_{g,j,l}^{*}, \beta_{g,j,l}^{*}, \nu^{*} \geq 0, \forall g, 1 \leq j \leq |\mathcal{G}_g|, 1 \leq l \leq j, \\ & \sum_{d=1}^{N'_{RF}}\varphi_d^{*} \geq 0, \mathbf{A}_{q,e}^{*} \succeq 0, \\ & \mathbf{A}_{q,e}^{*}\mathbf{W}_{q,e}^{*} = \mathbf{0}, \nabla_{\mathbf{W}_{q,e}}\mathcal{L}\left(\mathbf{W}_{q,e}^{*}\right) = \mathbf{0}, \end{aligned} \tag{56}$$

where $r_{g,j,l}^{*}$, $\beta_{g,j,l}^{*}$, ν^{*}, φ_d^{*} and $\mathbf{A}_{q,e}^{*}$ denote the optimal Lagrange multipliers, $\nabla_{\mathbf{W}_{q,e}}\mathcal{L}\left(\mathbf{W}_{q,e}^{*}\right)$ denotes the gradient of $\mathcal{L}\left(\mathcal{X}\right)$ with respect to $\mathbf{W}_{q,e}^{*}$. According to $\mathcal{L}\left(\mathcal{X}\right)$, $\nabla_{\mathbf{W}_{q,e}}\mathcal{L}\left(\mathbf{W}_{q,e}^{*}\right) = \mathbf{0}$ can be given as

$$\begin{aligned} & \mathbf{I} - \sum_{l=1}^{e} r_{q,e,l}^{*}\mathbf{H}_{q,l}^{H}\mathbf{v}\mathbf{v}^{H}\mathbf{H}_{q,l} + \sum_{j=e+1}^{|\mathcal{G}_q|}\sum_{l=1}^{j}\beta_{q,j,l}^{*}\mathbf{H}_{q,l}^{H}\mathbf{v}\mathbf{v}^{H}\mathbf{H}_{q,l} \\ & + \sum_{g=1,g\neq q}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\sum_{l=1}^{j}\beta_{g,j,l}^{*}\mathbf{H}_{g,l}^{H}\mathbf{v}\mathbf{v}^{H}\mathbf{H}_{g,l} + \nu^{*}\mathbf{I} \\ & + \sum_{d=1}^{N'_{RF}}\varphi_d^{*}\mathbf{\Phi}_d + \mathbf{A}_{q,e}^{*} = \mathbf{0}. \end{aligned} \tag{57}$$

Then, (57) can be rewritten as

$$\begin{aligned} & \left(1+\nu^{*}\right)\mathbf{I} + \sum_{d=1}^{N'_{RF}}\varphi_d^{*}\mathbf{\Phi}_d + \mathbf{A}_{q,e}^{*} = \sum_{l=1}^{e} r_{q,e,l}^{*}\mathbf{H}_{q,l}^{H}\mathbf{v}\mathbf{v}^{H}\mathbf{H}_{q,l} \\ & - \sum_{j=e+1}^{|\mathcal{G}_q|}\sum_{l=1}^{j}\beta_{q,j,l}^{*}\mathbf{H}_{q,l}^{H}\mathbf{v}\mathbf{v}^{H}\mathbf{H}_{q,l} \\ & - \sum_{g=1,g\neq q}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\sum_{l=1}^{j}\beta_{g,j,l}^{*}\mathbf{H}_{g,l}^{H}\mathbf{v}\mathbf{v}^{H}\mathbf{H}_{g,l}. \end{aligned} \tag{58}$$

Denote the right part of (58) as ψ_R. Through postmultiplying both sides of (58) by $\mathbf{W}_{q,e}^{*}$, due to $\mathbf{A}_{q,e}^{*}\mathbf{W}_{q,e}^{*} = \mathbf{0}$, (58) can be transformed into

$$\left(\left(1+\nu^{*}\right)\mathbf{I} + \sum_{d=1}^{N'_{RF}}\varphi_d^{*}\mathbf{\Phi}_d\right)\mathbf{W}_{q,e}^{*} = \psi_R\mathbf{W}_{q,e}^{*}. \tag{59}$$

Since $\left(1+\nu^{*}\right)\mathbf{I} + \sum\limits_{d=1}^{N'_{RF}}\varphi_d^{*}\mathbf{\Phi}_d \succ 0$, the following relation holds

$$\operatorname{rank}\left(\mathbf{W}_{q,e}^{*}\right) = \operatorname{rank}\left(\left(\left(1+\nu^{*}\right)\mathbf{I} + \sum_{d=1}^{N'_{RF}}\varphi_d^{*}\mathbf{\Phi}_d\right)\mathbf{W}_{q,e}^{*}\right). \tag{60}$$

Based on the right part of (59), we can derive that

$$\begin{aligned} \operatorname{rank}\left(\psi_R\mathbf{W}_{q,e}^{*}\right) & \leq \min\left(\operatorname{rank}\left(\psi_R\right), \operatorname{rank}\left(\mathbf{W}_{q,e}^{*}\right)\right) \\ & \leq \operatorname{rank}\left(\psi_R\right). \end{aligned} \tag{61}$$

The rank of ψ_R can be given as

$$\operatorname{rank}\left(\psi_R\right) = \operatorname{rank}\left(\mathbf{H}_S\mathbf{v}\mathbf{v}^{H}\mathbf{H}_S^{H}\right) \leq 1, \tag{62}$$

[12] We only enumerate the KKT conditions which are necessary for the subsequent proof.

where $\mathbf{H}_S$ is given as

$$\begin{aligned}\mathbf{H}_S = \sum_{l=1}^{e}\sqrt{r^*_{q,e,l}}\mathbf{H}^H_{q,l} - \sum_{j=e+1}^{|\mathcal{G}_q|}\sum_{l=1}^{j}\sqrt{\beta^*_{q,j,l}}\mathbf{H}^H_{q,l} \\ - \sum_{g=1,g\neq q}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}_g|}\sum_{l=1}^{j}\sqrt{\beta^*_{g,j,l}}\mathbf{H}^H_{g,l}.\end{aligned} \tag{63}$$

Then, according to (59), (60), (61) and (62), we can derive that $\operatorname{rank}\left(\mathbf{W}^*_{q,e}\right) \leq 1$. In practice, the optimal solution $\mathbf{W}^*_{q,e}$ cannot be a zero matrix. Therefore, the rank of $\mathbf{W}^*_{q,e}$ must satisfy the rank-one constraint, viz., $\operatorname{rank}\left(\mathbf{W}^*_{q,e}\right)=1$ for $1 \leq q \leq |\mathcal{G}|$, $1 \leq e \leq |\mathcal{G}_q|$.

According to the above analyses, we can conclude that if the SDP problem OP5 is solvable, there must exist the optimal solution $\mathbf{W}^*_{g,j}$ for $1 \leq g \leq |\mathcal{G}|$, $1 \leq j \leq |\mathcal{G}_g|$, which satisfies $\operatorname{rank}\left(\mathbf{W}^*_{g,j}\right) = 1$, viz., $\mathbf{W}^*_{g,j} = \mathbf{w}^*_{g,j}\left(\mathbf{w}^*_{g,j}\right)^H$.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen, “A vision of 6G wireless systems: Applications, trends, technologies, and open research problems,” *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[2] N. Al-Falahy and O. Y. K. Alani, “Millimetre wave frequency band as a candidate spectrum for 5G network architecture: A survey,” *Phys. Commun.*, vol. 32, pp. 120–144, Feb. 2019.

[3] R. W. Heath, N. González-Prelcic, S. Rangan, and W. Roh, “An overview of signal processing techniques for millimeter wave MIMO systems,” *IEEE J. Sel. Top. Signal Process.*, Apr. 2016.

[4] M. Wang, F. Gao, S. Jin, and H. Lin, “An overview of enhanced massive MIMO with array signal processing techniques,” *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 5, pp. 886–901, Sep. 2019.

[5] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, “Prospective multiple antenna technologies for beyond 5G,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[6] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, “A survey of non-orthogonal multiple access for 5G,” *IEEE Commun. Surveys Tut.*, vol. 20, no. 3, pp. 2294–2323, Jul.–Sep. 2018.

[7] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, “Nonorthogonal multiple access for 5G and beyond,” *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

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

[9] S. Sanayei and A. Nosratinia, “Antenna selection in MIMO systems,” *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[10] ——, “Capacity of MIMO channels with antenna selection,” *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 4356–4362, Nov. 2007.

[11] A. Sayeed and N. Behdad, “Continuous aperture phased MIMO: Basic theory and applications,” in *2010 48th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2010, pp. 1196–1203.

[12] A. Sayeed and J. Brady, “Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies,” in *2013 IEEE Global Commun. Conf. (GLOBECOM)*, 2013, pp. 3679–3684.

[13] T. Xie, L. Dai, D. W. K. Ng, and C. Chae, “On the power leakage problem in millimeter-wave massive MIMO with lens antenna arrays,” *IEEE Trans. Signal Process.*, vol. 67, no. 18, pp. 4730–4744, Sep. 2019.

[14] H. M. Al-Obiedollah, K. Cumanan, J. Thiyagalingam, A. G. Burr, Z. Ding, and O. A. Dobre, “Energy efficient beamforming design for MISO non-orthogonal multiple access systems,” *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4117–4131, Jun. 2019.

[15] Y. Feng, S. Yan, Z. Yang, N. Yang, and J. Yuan, “Beamforming design and power allocation for secure transmission with NOMA,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2639–2651, May 2019.

[16] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, “Joint power control and beamforming for uplink non-orthogonal multiple access in 5G millimeter-wave communications,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6177–6189, Sep. 2018.

[17] ——, “Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.

[18] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, “Multi-beam NOMA for hybrid mmwave systems,” *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1705–1719, Feb. 2019.

[19] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, “Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[20] P. Liu, Y. Li, W. Cheng, W. Zhang, and X. Gao, “Energy-efficient power allocation for millimeter wave beamspace MIMO-NOMA systems,” *IEEE Access*, vol. 7, pp. 114 582–114 592, 2019.

[21] M. A. Almasi, H. Mehrpouyan, D. Matolak, C. Pan, and M. Elkashlan, “Reconfigurable antenna multiple access for 5G mmWave systems,” in *2018 IEEE Int. Conf. Commun. Workshops, ICC Workshops - Proc.*, 2018, pp. 1–6.

[22] M. A. Almasi, R. Amiri, M. Vaezi, and H. Mehrpouyan, “Lens-based millimeter wave reconfigurable antenna NOMA,” in *2019 IEEE Int. Conf. Commun. Workshops, ICC Workshops - Proc.*, 2019, pp. 1–5.

[23] P. Liu, Y. Li, W. Cheng, X. Gao, and W. Zhang, “Multi-beam NOMA for millimeter-wave massive MIMO with lens antenna array,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11 570–11 583, Oct. 2020.

[24] S. Abdel-Razeq, S. Zhou, R. Bansal, and M. Zhao, “Uplink NOMA transmissions in a cooperative relay network based on statistical channel state information,” *IET Commun.*, vol. 13, no. 4, pp. 371–378, Mar. 2019.

[25] Y. Wu, L. P. Qian, H. Mao, X. Yang, H. Zhou, and X. Shen, “Optimal power allocation and scheduling for non-orthogonal multiple access relay-assisted networks,” *IEEE Trans. Mob. Comput.*, vol. 17, no. 11, pp. 2591–2606, Nov. 2018.

[26] Q. Wu and R. Zhang, “Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network,” *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[27] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, Z. Qin, M. D. Renzo, and N. Al-Dhahir, “Reconfigurable intelligent surfaces: Principles and opportunities,” 2020. [Online]. Available: https://arxiv.org/abs/2007.03435

[28] Z. Wang, L. Liu, and S. Cui, “Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.

[29] Z. He and X. Yuan, “Cascaded channel estimation for large intelligent metasurface assisted massive MIMO,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[30] J. Chen, Y.-C. Liang, H. V. Cheng, and W. Yu, “Channel estimation for reconfigurable intelligent surface aided multi-user MIMO systems,” 2019. [Online]. Available: https://arxiv.org/abs/1912.03619

[31] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[32] H. Guo, Y. Liang, J. Chen, and E. G. Larsson, “Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064–3076, May 2020.

[33] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, “Robust and secure wireless communications via intelligent reflecting surfaces,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[34] L. Dong and H. M. Wang, “Enhancing secure MIMO transmission via intelligent reflecting surface,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7543–7556, Nov. 2020.

[35] Z. Ding and H. Vincent Poor, “A simple design of IRS-NOMA transmission,” *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, May 2020.

[36] H. Wang, C. Liu, Z. Shi, Y. Fu, and R. Song, “On power minimization for IRS-aided downlink NOMA systems,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1808–1811, Nov. 2020.

[37] J. Zuo, Y. Liu, Z. Qin, and N. Al-Dhahir, “Resource allocation in intelligent reflecting surface assisted NOMA systems,” *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7170–7183, Nov. 2020.

[38] J. Zhu, Y. Huang, J. Wang, K. Navaie, and Z. Ding, “Power efficient IRS-assisted NOMA,” *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 900–913, Feb. 2021.

[39] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, “Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6884–6898, Oct. 2020.

[40] T. Hou, Y. Liu, Z. Song, X. Sun, and Y. Chen, “MIMO-NOMA networks relying on reconfigurable intelligent surface: A signal cancellation-based design,” *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6932–6944, Nov. 2020.

[41] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, pp. 1–1, Jan. 2020.

[42] Y. Zeng and R. Zhang, "Millimeter wave MIMO with lens antenna array: A new path division multiplexing paradigm," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1557–1571, Apr. 2016.

[43] ——, "Cost-effective millimeter-wave communications with lens antenna array," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 81–87, Aug. 2017.

[44] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.

[45] J. Zhang, Y. Huang, J. Wang, B. Ottersten, and L. Yang, "Per-antenna constant envelope precoding and antenna subset selection: A geometric approach," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6089–6104, Dec. 2016.

[46] P. Cao, J. Thompson, and H. V. Poor, "A sequential constraint relaxation algorithm for rank-one constrained problems," in *2017 25th European Signal Process. Conf. (EUSIPCO)*, 2017, pp. 1060–1064.

[47] L. Yang, Y. Zeng, and R. Zhang, "Channel estimation for millimeter-wave MIMO communications with lens antenna arrays," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3239–3251, Apr. 2018.

[48] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

Penglu Liu received the B.S. degree in Electronic and Information Engineering from Zhengzhou University (SIAS International College), China, in 2010, and his M.S. degree in Signal and Information Processing from Xi'an University of Post & Telecommunications, China, in 2017. He is currently pursuing the Ph.D. degree in Information and Communication Engineering at Northwestern Polytechnical University from 2017. His current research interests focus on resource allocation for wireless communication networks, millimeter wave communication, intelligent reflecting surface aided communication, rate-splitting multiple access, convex optimization and its application.

Yong Li received the B.S. degree in Avionics Engineering, M.S. and Ph.D degrees in Circuits and Systems from Northwestern Polytechnical University, Xi'an, China, in 1983, 1988 and 2005, respectively. He joined School of Electronic Information, Northwestern Polytechnical University in 1993 and was promoted to professor in 2002. His research interests include digital signal processing and radar signal processing.

Wei Cheng received the B.S. degree in Electronic and Information Engineering, M.S., and Ph.D. degrees in Communication and Information System from Northwestern Polytechnical University, Xi'an, China, in 2003, 2006 and 2011, respectively. From April 2011 to April 2013, he worked in the post-doctoral research station of School of Electronic Information, Northwestern Polytechnical University. Since 2013, he has been a lecturer of School of Electronic Information, Northwestern Polytechnical University and was promoted to associate professor in 2015. His research interests include Wireless Sensor Networks and Ad Hoc Networks, Radar Signal Processing.

Xiang Gao received the B.S. degree in communication engineering from North University of China, Taiyuan, China, in 2017 and his M.S. degree in communication and information system from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently pursuing the Ph.D. degree in Information and Communication Engineering at Northwestern Polytechnical University. His research interests include physical layer security, signal processing, convex optimization and its applications.

Xiaojing Huang (M'99-SM'11) received the B.Eng., M.Eng., and Ph.D. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1983, 1986, and 1989, respectively. He was a Principal Research Engineer with the Motorola Australian Research Center, Botany, NSW, Australia, from 1998 to 2003, and an Associate Professor with the University of Wollongong, Wollongong, NSW, Australia, from 2004 to 2008. He had been a Principal Research Scientist with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, NSW, Australia, and the Project Leader of the CSIRO Microwave and mm-Wave Backhaul projects since 2009. He is currently a Professor of Information and Communications Technology with the School of Electrical and Data Engineering and the Program Leader for Mobile Sensing and Communications with the Global Big Data Technologies Center, University of Technology Sydney (UTS), Sydney, NSW, Australia. His research interests include high-speed wireless communications, digital and analog signal processing, and synthetic aperture radar imaging. With over 32 years of combined industrial, academic, and scientific research experience, he has authored over 330 book chapters, refereed journal and conference papers, major commercial research reports, and filed 31 patents. Prof. Huang was a recipient of the CSIRO Chairman's Medal and the Australian Engineering Innovation Award in 2012 for exceptional research achievements in multigigabit wireless communications.