

STAR-RIS-Assisted Hybrid NOMA mmWave Communication: Optimization and Performance Analysis

Muhammad Faraz Ul Abrar, Muhammad Talha, Rafay Iqbal Ansari, *Senior Member, IEEE*, Syed Ali Hassan, *Senior Member, IEEE*, and Haejoon Jung, *Senior Member, IEEE*

Abstract—Simultaneously reflecting and transmitting reconfigurable intelligent surfaces (STAR-RIS) has recently emerged out as prominent technology that exploits the transmissive property of RIS to mitigate the half-space coverage limitation of conventional RIS operating on millimeter-wave (mmWave). In this paper, we study a downlink STAR-RIS-based multi-user multiple-input single-output (MU-MISO) mmWave hybrid non-orthogonal multiple access (H-NOMA) wireless network, where a sum-rate maximization problem has been formulated. The design of active and passive beamforming vectors, time and power allocation for H-NOMA is a highly coupled non-convex problem. To handle the problem, we propose an optimization framework based on alternating optimization (AO) that iteratively solves active and passive beamforming sub-problems. Channel correlations and channel strength-based techniques have been proposed for a specific case of two-user optimal clustering and decoding order assignment, respectively, for which analytical solutions to joint power and time allocation for H-NOMA have also been derived. Simulation results show that: 1) the proposed framework leveraging H-NOMA outperforms conventional OMA and NOMA to maximize the achievable sum-rate; 2) using the proposed framework, the supported number of clusters for the given design constraints can be increased considerably; 3) through STAR-RIS, the number of elements can be significantly reduced as compared to conventional RIS to ensure a similar quality-of-service (QoS).

Index Terms—STAR-RIS, reconfigurable intelligent surfaces (RIS), mmWave communication, hybrid non-orthogonal multiple access, performance analysis, sum-rate maximization, beyond 5G.

I. INTRODUCTION

The development of various advanced applications such as augmented reality, high quality broadcast demands strict requirements for fifth-generation and beyond (B5G) networks with regards to ultra reliable low latency communications (URLLC) [1]. To achieve such requirements, millimeter-wave (mmWave) technology is of prime interest among researchers and industry, making it a promising technology for driving B5G networks [2]. However, higher sensitivity to blockages due to very short wavelength of mmWave frequency spectrum is a major drawback in achieving such high gains and low latency requirements [3].

M. Faraz, M. Talha and S. A. Hassan are with the School of Electrical Engineering and Computer Science (SEECs), National University of Sciences and Technology (NUST), 44000, Islamabad, Pakistan (email: {mabrar.bee17seecs, mtalha.bee17seecs, ali.hassan}@seecs.edu.pk)

R. I. Ansari is with the Department of Computer and Information Sciences, Northumbria University Newcastle, UK. (email: rafay.ansari@northumbria.ac.uk)

H. Jung is with the Department of Electronic Engineering, Kyung Hee University, Yongin 17104, Korea (e-mail: haejoonjung@khu.ac.kr)

To minimize the effect of channel impairments such as random small-scale fading, path loss and blockages, technologies such as multiple input multiple output (MIMO) and massive MIMO are proposed to be highly beneficial, however, [4]–[6] have shown that the usage of reconfigurable intelligent surfaces (RISs) can significantly enhance energy efficiency (EE) by creating virtual links between base station (BS) and users and achieving higher reflect beamforming gains. RIS can efficiently modify the properties of a large number of passive, low-cost and programmable reflective elements to reflect the incoming signal in the desired direction with the help of a controller [7]. Because of their nearly passive nature, RISs can enhance the communication performance without the need of radio frequency chains in contrast to active relays, and provide low self interference and hardware costs [7]. Moreover, in case of blockage of the direct link between the BS and mobile user, RIS can also be used to realize a smart tunable wireless environment by dynamically modifying the effective blocked end-to-end channel between the BS and users [8]. However, all these studies have considered RIS in reflection mode only. This half-space coverage limits the usage and potential benefits of RIS-assisted communication networks.

Non-orthogonal multiple access (NOMA) is considered to be a key candidate for interference-limited future wireless networks to provide high data rates [9], whose basic principle is to serve multiple users over the same time-frequency resource block via a common beamforming vector. NOMA has been studied in conjunction with RIS networks to increase coverage range, user fairness and EE of communication networks [10]. The integration of NOMA and RIS can further improve the system performance, as RIS can assist NOMA to dynamically tune the channels of the paired users yielding significant NOMA gains, and likewise, RIS-assisted systems operating on NOMA can achieve much higher SE than OMA. With the aim to overcome the half-space coverage constraint of RIS along with limited possible BS-user association in severe blockage conditions, researchers have recently proposed simultaneously reflecting and transmitting reconfigurable intelligent surfaces (STAR-RIS) to enhance the existing RIS capabilities to serve users omni-directionally, which can transmit and reflect the incoming signal simultaneously by introducing the equivalent surface electric and magnetic currents in the model [11], thus enabling full-space network coverage around RIS.

A STAR-RIS can operate in three different modes namely, energy splitting (ES), mode switching (MS) and time switching (TS) depending upon use case requirements [12]. For the ES mode STAR-RIS, each element may have a different reflection-transmission amplitude coefficient ratio, also known

as ES ratio that determines the proportion of signal being reflected or transmitted, respectively. The overall reflection or transmission of signal impinging on STAR-RIS appears as a net effect of the ES ratio of all the elements of the STAR-RIS. Another key feature is that the phases corresponding to reflection and transmission are completely independent, which can be exploited to maximize overall user coverage and performance gains of randomly scattered users.

Although NOMA assists RIS systems in increasing the spectral efficiency (SE) significantly, however, serving all users simultaneously on NOMA via a unified beamformer may not help achieve optimal sum-rate depending upon user-channel correlations. Therefore, we propose an alternative multiple access scheme H-NOMA for RIS and in particular, STAR-RIS systems, in which users are grouped into several clusters. While all the clusters are served using OMA, all the users belonging to a cluster are served simultaneously using NOMA. In the next section, we discuss prior studies related to the conventional RIS and STAR-RIS-based systems.

A. Related Works

Several works have been done to exploit the benefits of RIS-enabled communication by optimizing joint beamforming, EE and sum-rate maximization in RIS-assisted uplink network [13]–[17]. Due to the passive nature of RIS, channel estimation of the RIS-assisted communication systems is crucial, and therefore, there have been several studies in [18]–[23] to obtain channel state information (CSI) to maximize achievable gains and network coverage. In addition, the authors in [10], [24], [25] show that RIS-NOMA is a promising technique to achieve higher EE, SE and user fairness compared to OMA-assisted RIS systems for future wireless communication networks. However, all of the above-mentioned works consider RIS in reflection mode only, i.e., users are located in only half-coverage space of the RIS.

In contrast, recent studies of the STAR-RIS-empowered communication networks present that by deploying STAR-RIS, the coverage and SE are significantly improved as compared to conventional RIS-assisted system (i.e., with reflection mode only) [26]. The authors in [12] propose a power consumption minimization algorithm for ES, TS and MS protocols and prove that TS and ES operating protocols are generally preferable for uni-cast and multi-cast transmissions. Further, [27] shows that using NOMA with STAR-RIS can significantly increase the user coverage range as compared to OMA-assisted STAR-RIS system. All these studies have considered either single-user or two-user (one served via reflection and other through transmission of STAR-RIS) system model. Moreover, the authors in [28] propose an efficient algorithm for uplink CSI acquisition for the STAR-RIS systems, whereas the work in [29] shows how to implement the STAR-RIS system in practice. A recent work on STAR-RIS in [30] solves a similar sum-rate maximization problem, while serving multiple conventional NOMA clusters simultaneously, however, such a choice of multiple access scheme can provide higher gains only when the served users have similar channel conditions.

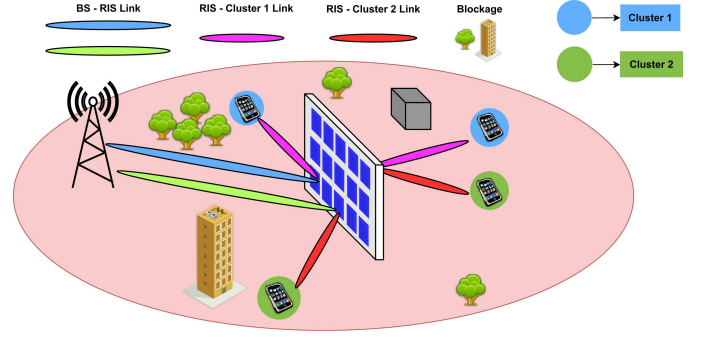


Figure 1. A STAR-RIS-assisted H-NOMA system serving four blocked users with $C = 2$, $K_c = 2$ through reflection and transmission.

B. Motivation and Contribution

Considering its unique features over conventional RIS-assisted systems, a STAR-RIS can enhance the wireless connectivity in 360° (i.e., full-space) coverage by introducing the independent transmissive and reflective beams. In order to fully exploit the advantages of STAR-RIS, joint active and passive beamforming optimization needs to be done for multiple users randomly scattered surrounding the STAR-RIS. The literature on STAR-RIS is in its initial phases, and to the best of our knowledge, optimal multiple access scheme and a detailed performance analysis for STAR-RIS systems with a design perspective showing the effect of active, passive beamforming vectors and resource allocation scheme for sum-rate maximization has not been presented yet. This motivates us to develop a complete optimization framework and analyze the impact of each design parameter of a STAR-RIS-assisted system in achieving optimal sum-rate. The main contributions of our work can be summarized as follows:

- For a multi-cluster hybrid NOMA (H-NOMA)-based STAR-RIS system, we propose optimal user pairing and decoding order strategies using user-channel correlations and channel strengths for two-user NOMA clustering, where each cluster consists of two users (one served via reflection and the other via transmission through the STAR-RIS).
- We propose an alternating optimization (AO)-based algorithm to determine the active and passive beamforming vectors, time slot allocation and power allocation. We also show the impact of each of these individual design parameters on the achievable sum-rate.
- We provide a detailed performance analysis and compare the proposed algorithm with OMA-assisted STAR-RIS and conventional NOMA-assisted STAR-RIS systems. The results show that because of additional degree of freedom in H-NOMA, it outperforms conventional NOMA-based STAR-RIS systems in [30].

C. Paper Organization and Notation

The rest of the paper is organized as follows. Section II presents the system model including signal model, channel model and multiple access schemes. Section III presents the problem formulation and the proposed solution, whereas in

Section III-A, we present optimal solutions to user pairing and decoding order assignment, active and passive beamforming optimizations and joint H-NOMA power coefficient and time-slot allocation problems. Section IV presents the performance analysis of the considered STAR-RIS system and other simulation results using the proposed algorithm. Finally, Section V concludes the paper.

Lower-case and upper-case boldface letters denote vectors and matrices, respectively. Further, $\mathbb{C}^{N \times 1}$ denotes a N dimensional complex vector, and $\mathbb{C}^{N \times K}$ represents a $N \times K$ dimensional complex matrix. Also, $(\cdot)^H$ denotes the Hermitian (conjugate transpose), and $\text{vec}(\cdot)$ and $\text{diag}(\cdot)$ corresponds to vectorization and diagonalization, respectively. $[\mathbf{A}]_{i,j}$ returns the entry of the input matrix corresponding to the i -th row and j -th column, whereas $\mathbf{A} \circ \mathbf{B}$ represents Hadamard (element-wise) product of the two matrices \mathbf{A} and \mathbf{B} . Also, $|\cdot|$ corresponds to the magnitude of a complex number or cardinality of the set if the input is a set. In addition, $\|\cdot\|_2$ denotes the ℓ_2 -norm, and $\text{Real}(\cdot)$, $\text{angle}(\cdot)$ returns the real component, and phase of a complex argument, respectively.

II. SYSTEM MODEL

We consider a STAR-RIS-assisted mmWave downlink multi-user multiple-input single-output (MU-MISO) communication scenario with N_t BS antenna elements, K single-antenna users and M STAR-RIS passive units that reflect or transmit the signals coming from the BS towards the intended users. This model is particularly helpful in scenarios where the BS to user link is blocked. Fig. 1 shows the considered scenario. The STAR-RIS is assumed to operate in the ES mode, where it is able to both reflect and transmit the incoming BS signal simultaneously to create a link between the BS and the users. In addition, the K supported users are divided into C clusters, where the c -th cluster serves a total of K_c users and let $\mathbb{K}_c \triangleq \{1, 2, \dots, K_c\}$ represent the set of all K_c users in cluster c , where each user $k \in \{r, t\}$ represents the user being served in the reflecting and transmitting mode, respectively. Suppose that $\mathbb{C} \triangleq \{1, 2, \dots, C\}$ denotes the set of clusters, in which users are grouped and $m \in \mathbb{M} \triangleq \{1, 2, \dots, M\}$ denotes the set of STAR-RIS elements. Further, let T_{\max} denote the channel's coherence time, which is divided into C sub-time slots, where each $t_c, \forall c \in \mathbb{C}$, is the time allocated to serve all users of cluster c . Taking into account the practical hardware limitations, discrete STAR-RIS amplitudes and phases are assumed. To be specific, the m -th STAR-RIS element's phase corresponds to $\theta_{m,c}^x \in \Psi \triangleq \left\{0, \frac{2\pi}{2^{B_1}}, \dots, \frac{2\pi(2^{B_1}-1)}{2^{B_1}}\right\}$, while its amplitude is denoted by $\beta_{m,c}^x \in \Omega \triangleq \left\{0, \frac{1}{2^{B_2-1}}, \frac{2}{2^{B_2-1}}, \dots, \frac{2^{B_2-1}}{2^{B_2-1}}\right\}$, where B_1 and B_2 are finite available resolution bits for tuning phase shifts and amplitude coefficients of the RIS elements, respectively, and $\chi \in X \triangleq \{r, t\}$ for the reflection and transmission modes, respectively. For lossless operation, we assume for the m -th STAR-RIS element

$$\beta_{m,c}^x + \bar{\beta}_{m,c}^x = 1.$$

However, the phase shifts $\theta_{m,\chi}^x$ and $\bar{\theta}_{m,c}^x$ can be independently tuned $\forall m \in \mathbb{M}$.

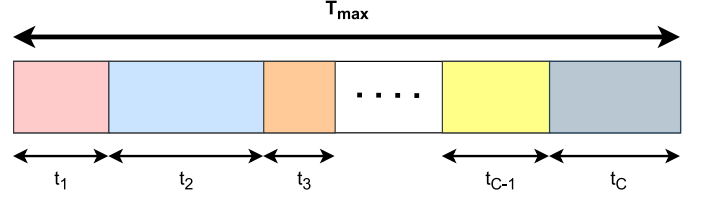


Figure 2. Unequal time-slot allocation for different clusters meeting QoS requirement of each within given channel coherence time T_{\max} .

A. Signal Model

Let $x_{k,c}$ be the transmitted signal from the BS in time-slot t_c intended for the k -th user belonging to the c -th cluster, where $\mathbb{E}\{|x_{k,c}|^2\} = 1$. The received signal $y_{k,c}$ for the user is

$$y_{k,c} = \underbrace{\mathbf{h}_{k,c} \mathbf{w}_c \sqrt{p_{k,c}} x_{k,c}}_{\text{desired information signal}} + \underbrace{\mathbf{h}_{k,c} \mathbf{w}_c \sum_{i \in \{\mathbb{K}_c \setminus k\}} \sqrt{p_{i,c}} x_{i,c}}_{\text{intra-cluster interference}} + \underbrace{n_{k,c}}_{\text{noise}}, \quad (1)$$

where $p_{k,c}$ represents the allocated power for the k -th user of cluster c , \mathbf{w}_c is the active beamforming vector associated with the cluster c . Also, $\mathbf{h}_{k,c} = \mathbf{g}_{k,c}^H \Phi_c^x \mathbf{H}$ is the end-to-end channel from the BS to the user, where $\mathbf{g}_{k,c} \in \mathbb{C}^{M \times 1}$ is the channel vector from the STAR-RIS to the k -th user of cluster c , $\mathbf{H} \in \mathbb{C}^{M \times N_t}$ is the BS-STAR-RIS channel matrix, $\Phi_c^x = \text{diag}\left(\left[\beta_{1,c}^x e^{j\theta_{1,c}^x}, \beta_{2,c}^x e^{j\theta_{2,c}^x}, \dots, \beta_{M,c}^x e^{j\theta_{M,c}^x}\right]\right) \in \mathbb{C}^{M \times M}$ represents diagonal phase shift matrix, in which $\beta_{m,c}^x \in [0, 1]$ and $\theta_{m,c}^x \in [0, 2\pi) \forall m \in M$, and $n_{k,c} \sim \mathcal{CN}(0, \sigma^2)$ is independent and identically distributed (i.i.d) additive white complex Gaussian noise at the receiver with zero mean and σ^2 variance. We define a function $\text{state}(k) \triangleq \chi^k$, which returns whether user k is being served through STAR-RIS reflection or transmission. Even though obtaining perfect CSI for both BS-STAR-RIS and STAR-RIS downlink channels is an open research problem, however, the recent attempts in [18]–[22] to estimate CSI for conventional RIS-empowered systems with low complexity techniques can also be extended to STAR-RIS-based systems.

B. Channel Model

We assume that the BS is serving users while operating at mmWave with $(N_t > K)$ antennas; therefore, the BS-STAR-RIS channel matrix is given as

$$\mathbf{H} = \mathbf{H}_{\text{LoS}} + \mathbf{H}_{\text{NLoS}}, \quad (2)$$

where \mathbf{H}_{LoS} and \mathbf{H}_{NLoS} represent the line-of-sight (LoS), and non-LoS components of the channel \mathbf{H} , respectively. We use the three-dimensional (3D) Saleh-Valenzuela geometry channel model, as in [31] for both BS-STAR-RIS and STAR-RIS to user LoS components. Thus, \mathbf{H}_{LoS} is given by

$$\mathbf{H}_{\text{LoS}} = \sqrt{PL(d_{BR})} \left(\sqrt{N_t M} h_{\text{LoS}} \mathbf{h}_r^T(\theta_{\text{LoS}}^r, \phi_{\text{LoS}}^r) \mathbf{h}_t^H(\theta_{\text{LoS}}^t, \phi_{\text{LoS}}^t) \right). \quad (3)$$

We assume the BS and STAR-RIS to be the rectangular uniform planar arrays (UPAs) with antennas and elements uniformly distributed over $y-z$ plane, i.e., $N_t = N_{ty} \times N_{tz}$, $M = M_y \times M_z$. Also, h_{LoS} is the random complex channel

gain of the varying channel at a particular time instant, whereas $\mathbf{h}_r(\theta_{\text{LoS}}^r, \phi_{\text{LoS}}^r)$ and $\mathbf{h}_t(\theta_{\text{LoS}}^t, \phi_{\text{LoS}}^t)$ denote the array steering response vectors at the BS and STAR-RIS, respectively, corresponding to the LoS paths, and are given by (4) and (5), where d is the inter-element spacing, which is $\frac{\lambda}{2}$ for both BS and STAR-RIS with λ being the carrier wavelength. In addition, $\theta_l^t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $\phi_l^t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ are the angles-of-departure (AoD) corresponding to the transmit antennas in the elevation and azimuth plane, respectively, for the l -th propagation path. Similarly, $\theta_l^r \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, $\phi_l^r \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ are the angles-of-arrival (AoA) of STAR-RIS elements for the elevation and azimuth plane, respectively.

Also, the NLoS component for the BS-STAR-RIS channel, $\mathbf{H}_{\text{NLoS}} \sim \mathcal{CN}(0, 10^{-0.1PL(d_{BR})})$, is modeled as a Rayleigh flat fading channel, where $PL(d_{BR})$ represents the path loss at a distance d_{BR} (the distance from the BS to the STAR-RIS) during signal propagation and, as in [32], is given by

$$PL(d) = PL_{do} + 10\eta \log\left(\frac{d}{d_o}\right) + \zeta \quad [\text{dB}], \quad (6)$$

where PL_{do} is path loss at a reference distance d_o , whereas η is the path loss exponent. Further, ζ represents the shadowing of the environment. Similarly, the STAR-RIS-user channel model is given by

$$\mathbf{g} = \mathbf{g}_{\text{LoS}} + \mathbf{g}_{\text{NLoS}}, \quad (7)$$

where $\mathbf{g}_{\text{NLoS}} \sim \mathcal{CN}(0, 10^{-0.1PL(d_{RU})})$ is the NLoS channel component of \mathbf{g} , while the LoS component \mathbf{g}_{LoS} is given by

$$\mathbf{g}_{\text{LoS}} = \sqrt{PL(d_{RU})} \left(\sqrt{N_t M} \mathbf{g}_{\text{LoS}} \mathbf{g}_t(\theta_{\text{LoS}}^t, \phi_{\text{LoS}}^t) \right), \quad (8)$$

where $PL(d_{RU})$ is the path loss at the distance d_{RU} (the distance between the STAR-RIS and the user). Also, \mathbf{g}_{LoS} represents the complex channel gain, whereas \mathbf{g}_t corresponds to the array steering response vector from the STAR-RIS to the users and is given by

$$\mathbf{g}_t(\theta_l^t, \phi_l^t) = \left[1, \dots, e^{j\frac{2\pi}{\lambda}((m_y-1)\cos(\theta_l^t)\sin(\phi_l^t))} \right], \quad (9)$$

where $\theta_l^t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\phi_l^t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

C. Multiple Access Schemes

NOMA works on the principle of superposition coding (SC) on the BS to serve multiple users in a cluster and performs successful successive interference cancellation (SIC) on users to remove the intra-cluster-interference (IntraCI). This intrinsic property of NOMA can be exploited to expand the rate region [33]. Consequently, optimal rates can be achieved if the BS serves the paired users via the STAR-RIS reflection and transmission of the combined signal via an active beamformer

\mathbf{w}_c having higher correlation with the paired user's channel vector $\mathbf{h}_{k,c}$ within the same time and frequency resources.

To the best of our knowledge, there is no substantial work on STAR-RIS that explores the choice of multiple access scheme for randomly scattered blocked users in addition to other design parameters. Although NOMA can be used to leverage higher gains through users having correlated channels, most of the practical communication situations might not demonstrate such behaviors, thereby, making NOMA inefficient in such scenarios. On the other hand, STAR-RIS-assisted OMA scheme also limits the achievable sum-rate. For example, time-division multiple access (TDMA) would serve every user in different time-slot. Similarly, frequency-division multiple access (FDMA), in addition to higher resource requirements, would also require higher external hardware control by partitioning the STAR-RIS into various sections in order to serve multiple BS beams simultaneously, where the elements in each section are controlled through their own controller to either reflect or transmit the signals to their associated users.

Therefore, to acquire a decent trade-off between achievable gains through NOMA and OMA, we pair the users into various clusters, and serve them on TDMA ensuring user quality of service (QoS) requirements. All the users in a cluster are served by the BS simultaneously with NOMA, by allocating different power allocation coefficients, in the time-slot allocated for that cluster, as shown in Fig. 2. This multiple access scheme has been introduced in [34] to maximize the EE of the system, and is referred to as H-NOMA. H-NOMA has been proven an ideal candidate to reduce the NOMA SIC complexity at the receiver of RIS-based systems without having performance degradation because of the increased degree of freedom in terms of the time-slot and power allocation [35]. The fundamental principles of NOMA, SC and SIC, require the right choice of user decoding order to obtain optimal rates. For single-input-single-output (SISO) systems, NOMA user decoding orders can be optimally decided via users' channel gains only. However, in the considered MU-MISO system, both active beamforming vector and STAR-RIS-user channel can change the user's effective signal-to-noise ratio (SNR), hence for such systems, choice of decoding order becomes an additional challenge, and any of the $K_c!$ order can be used to decode the signals intended for K_c users.

Let $\vartheta(k_c)$ represent the decoding order of user k in cluster c . For convenience, we are dropping the subscript c . If $\vartheta(k) = k$ and $\vartheta(i) = i$, then the signal intended for each user i will be successively decoded before decoding the k -th user signal at user k , given that $(\vartheta(i) < \vartheta(k)) \forall i \in \mathbb{K}_c$. Let $\vartheta(K_c) \geq \vartheta(k) \geq \vartheta(1), \forall k \in \mathbb{K}_c$ be the general decoding order for the users of cluster c . Therefore, the achievable signal-to-interference noise ratio (SINR) at user k to decode

$$\mathbf{h}_t(\theta_l^t, \phi_l^t) = \frac{1}{\sqrt{N_t}} \text{vec} \left(\left[e^{j\frac{2\pi d}{\lambda}((m_y-1)\cos(\theta_l^t)\sin(\phi_l^t)) + (m_z-1)\sin(\theta_l^t)} \right]_{m_y, m_z} \right), \quad (4)$$

$$\mathbf{h}_r(\theta_l^r, \phi_l^r) = \frac{1}{\sqrt{M}} \text{vec} \left(\left[e^{j\frac{2\pi d}{\lambda}((m_y-1)\cos(\theta_l^r)\sin(\phi_l^r)) + (m_z-1)\sin(\theta_l^r)} \right]_{m_y, m_z} \right). \quad (5)$$

the i -th user signal is given as

$$\gamma_{k \rightarrow i}^c = \frac{|\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{i,c}}{\sum_{j=i+1}^{K_c} |\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{j,c} + \sigma^2}, \quad (10)$$

where $\sum_{j=i+1}^{K_c} p_{j,c}$ is the IntraCI, while decoding the i -th user's signal at user k , $\forall \vartheta(k) > \vartheta(i)$, coming from users with higher decoding order than that of user i . It is clear that for a given user k , whose signal is to be decoded, the IntraCI only depends on the assigned NOMA decoding order to the user k among all the K_c users of the cluster c . Thus, the IntraCI for this case can be expressed as

$$\mathcal{I}(k, c) = \sum_{j=k+1}^{K_c} p_{j,c} = \sum_{j: \vartheta(j) > \vartheta(k)} p_{j,c}.$$

Similarly, for decoding the k -th user signal at user k the SINR can be given as

$$\begin{aligned} \gamma_{k \rightarrow k}^c &= \frac{|\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{k,c}}{\sum_{j=k+1}^{K_c} |\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{j,c} + \sigma^2}, \\ \gamma_{k \rightarrow k}^c &= \frac{|\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{k,c}}{\mathcal{I}(k, c) + \sigma^2}. \end{aligned} \quad (11)$$

The corresponding decoding rate can be given as

$$R_{k \rightarrow i}^c = t_c (\log_2 (1 + \gamma_{k \rightarrow i}^c)). \quad (12)$$

To perform successful SIC to decode signals of all the users of a cluster, power allocation needs to be done carefully to meet QoS requirement of each user while also minimizing the IntraCI. Therefore, the two conditions in (13) need to be satisfied.

III. PROBLEM FORMULATION AND SOLUTION

We express the problem to maximize the sum-rate for all the users grouped into various clusters in the network. Therefore, we formulate a joint optimization problem consisting of active beamforming, reflection and transmission passive beamforming, user pairing, NOMA decoding order assignment, power allocation coefficients and cluster time-slot allocation. The given problem can be formulated as

$$\max_{\mathbf{w}_c, \theta_{m,c}^x, \beta_{m,c}^x, \{\mathbb{K}_c\}, \vartheta(k_c), p_k, t_c} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} R_{k,c} \quad (14a)$$

$$s.t. \|\mathbf{w}_c\|_2^2 \leq P_{max}, \quad (14b)$$

$$R_{k \rightarrow i}^c \geq R_{i,c}^{\min}, \forall i, k \in \mathbb{K}_c, \vartheta(i) \leq \vartheta(k), \quad (14c)$$

$$R_{k \rightarrow i}^c \geq R_{i \rightarrow i}^c, \forall i, k \in \mathbb{K}_c, \vartheta(i) \leq \vartheta(k), \quad (14d)$$

$$\sum_{c \in \mathbb{C}} t_c = 1, t_c \in [0, 1], \quad (14e)$$

$$\sum_{k \in K_c} p_k = 1, p_k \in [0, 1], \quad (14f)$$

$$\sum_{c \in \mathbb{C}} |\mathbb{K}_c| = K, \quad (14g)$$

$$1 \leq \vartheta(k) \leq |\mathbb{K}_c|, \quad (14h)$$

$$\beta_{m,c}^x + \beta_{m,c}^{\bar{x}} = 1, \quad (14i)$$

$$\theta_{m,c}^x \in \Psi, \beta_{m,c}^x \in \Omega, \quad (14j)$$

$$\forall c \in \mathbb{C}, m \in \mathbb{M}, \chi \in X,$$

where $R_{k,c} = R_{k \rightarrow k}^c$. The constraint (14b) represents the maximum power (P_{max}) constraint of transmitting antennas, whereas the constraint (14c) refers to the minimum QoS decoding and achievable rate requirement for each user to be served. The constraint (14d) refers to SIC constraint following a specific decoding order in a cluster, while (14e) refers to the cluster time-slot allocation constraint in terms of percentage of T_{max} , and (14f) represents the NOMA power allocation coefficients constraint for each user of a cluster. Further, (14g) refers to user pairing constraint such that each user gets served, and (14h) represents the NOMA decoding order assignment constraint. In addition, (14i) represents the STAR-RIS transmission and reflection amplitude coefficient constraint, whereas (14j) corresponds to the finite resolution bits constraint of STAR-RIS phases and amplitudes.

The given problem in (14) is a highly coupled non-convex optimization problem due to the strong coupling between: 1) NOMA power allocation coefficients, cluster time-slot allocation and STAR-RIS amplitude coefficients 2) active beamformer and the end-to-end BS-user channel due to the STAR-RIS phases-assisted passive beamformer. The non-convexity of the problem lies in the objective function (14a) as well as in the constraints (14c) and (14d).

A. Proposed Solutions

To solve the sum-rate maximization problem, we first perform an optimal user pairing and decoding order scheme, and then approach to solve it by dividing the given problem in multiple sub-problems, i.e., active beamforming, passive beamforming, NOMA power and time-slot allocation optimization problems, and then we solve them independently forming an iterative AO-based algorithm.

1) *Optimal User Pairing and Decoding Order Scheme*: We first deal with optimal user pairing of K users into C clusters, and later perform time-slot and power allocation problems. To maximize the sum-rate while also increasing EE, we approach the problem by incorporating two users $\{i, j\}$ in every cluster such that $|\mathbb{K}_c| = 2, \forall c \in \mathbb{C}$ and $state(i) \neq state(j)$. The user pairing with $|\mathbb{K}_c|$ as an optimization parameter has been set apart as a good future work. Therefore, for the scope of this paper, we present a low-complexity user pairing algorithm with even K and $|\mathbb{K}_c| = 2$ that presents a near optimal

$$\begin{cases} R_{k \rightarrow i}^c \geq R_{k-1 \rightarrow i}^c \geq \dots \geq R_{i+1 \rightarrow i}^c \geq R_{i \rightarrow i}^c, & \forall \vartheta(i) < \vartheta(k) \\ R_{k \rightarrow i}^c \geq R_{i,c}^{\min}, R_{k-1 \rightarrow i}^c \geq R_{i,c}^{\min}, \dots, R_{i \rightarrow i}^c \geq R_{i,c}^{\min}. \end{cases} \quad (13)$$

solution. Let \mathbf{Q} be a $K \times K$ pairing symmetric matrix such that

$$\begin{cases} [\mathbf{Q}]_{i,j} = \psi, & \forall i \neq j \text{ and } \text{state}(i) \neq \text{state}(j) \\ [\mathbf{Q}]_{i,j} = 0, & \forall i = j \text{ or } \text{state}(i) = \text{state}(j) \\ \sum_{i=1}^K [\mathbf{Q}]_{i,j} = 1, & \forall j \in [1, K] \\ \sum_{j=1}^K [\mathbf{Q}]_{i,j} = 1, & \forall i \in [1, K] \end{cases} \quad (15)$$

where $\psi \in \{1, 0\}$ represents paired and non-paired users, respectively. Obviously $[\mathbf{Q}]_{i,i} = [\mathbf{Q}]_{j,j} = 0$ corresponds to self-pairing. Then the given sub-problem can be modeled as

$$\max_{\mathbf{Q}} \sum_{i=1}^K \sum_{j=1}^K [\mathbf{Q}]_{i,j} R_{i,j}, \quad (16a)$$

$$\text{subject to (15),} \quad (16b)$$

where $R_{i,j}$ is given by (12). We handle the above user pairing sub-problem by segregating users in two groups \mathbf{B}_1 and \mathbf{B}_2 of equal size ($\frac{K}{2}$), on the basis of their channel strength $\|\mathbf{h}_k\|^2 \forall k \in [1, K]$, while maximizing channel correlations with feasible users consequently maximizing the sum-rate.

We also define users' channel correlation and users' state matrix \mathbf{C} and \mathbf{S} , respectively, such that $[\mathbf{C}]_{i,j} = |\mathbf{h}_i \mathbf{h}_j^H|^2$ where \mathbf{h}_i and \mathbf{h}_j corresponds to channels of group \mathbf{B}_1 and \mathbf{B}_2 , respectively. For the user state matrix, let $[\mathbf{S}]_{i,j} = 1$, if users i, j have different states allowing a feasible pair, otherwise 0. The details of solving (15) is summarized in **Algorithm 1** which transforms the original problem. Therefore, we proceed the problem by utilizing the obtained \mathbf{C} and solve the original problem as an assignment linear programming problem (LPP) given by

$$\max_{\mathbf{A}} \sum_{i=1}^{K/2} \sum_{j=1}^{K/2} [\mathbf{C}]_{i,j} [\mathbf{A}]_{i,j} \quad (17a)$$

$$\text{s.t. } \sum_{i=1}^{K/2} [\mathbf{A}]_{i,j} = 1, \quad \forall j = \{1, 2, \dots, \frac{K}{2}\}, \quad (17b)$$

$$\sum_{j=1}^{K/2} [\mathbf{A}]_{i,j} = 1, \quad \forall i = \{1, 2, \dots, \frac{K}{2}\}, \quad (17c)$$

$$[\mathbf{A}]_{i,j} \in \{0, 1\}. \quad (17d)$$

This LPP is a convex problem that can be easily solved by optimization solvers such as CVX, or Hungarian algorithm [36] to obtain \mathbf{A}^* , which can be utilized to obtain \mathbb{K}_c^* . The given algorithm performs user-pairing via channel correlation, however at a later stage, we present **Algorithm 2**, which optimizes the active beamformer such that $|\mathbf{h}_j \mathbf{h}_i^H|^2 \rightarrow n(|\mathbf{h}_j \mathbf{w}_c|^2)$, where n is a scaling constant. The proposed algorithm provides the optimal solution for the sum-rate maximization objective function for the considered case.

The optimal strategy to allocate NOMA user decoding order $\vartheta(k)$ for maximum sum-rate is to assign higher decoding order to the user with higher channel strength $\|\mathbf{h}_k\|^2$. Therefore, optimal decoding order is assigned such that $\vartheta(i) > \vartheta(j) \forall i \in \mathbf{B}_1, j \in \mathbf{B}_2$. The feasible user pairing matrix \mathbf{A} is a $\frac{K}{2} \times \frac{K}{2}$

Algorithm 1 User Pairing and Decoding Order Algorithm for (16)

Input : $\{\mathbf{h}_k\}, K$
Output : $\{\mathbb{K}_c^*\}, \{\vartheta(k_c)\}$

- 1: Initialize $\tilde{\mathbf{h}} : \{\|\mathbf{h}_k\|\} \forall k \in [1, K]$
- 2: Update $\tilde{\mathbf{h}} = \text{SORT}(\tilde{\mathbf{h}}) \ni \tilde{h}_1 \geq \tilde{h}_2 \geq \dots \geq \tilde{h}_K$
- 3: Set $\mathbf{B}_1 = \{i\}, \mathbf{B}_2 = \{j\} \forall i \in \{1, 2, 3, \dots, \frac{K}{2}\}$, and $j \in \{\frac{K}{2} + 1, \dots, K\}$ from $\tilde{\mathbf{h}}$
- 4: $[\mathbf{C}]_{i,j} \leftarrow \{|\mathbf{h}_i \mathbf{h}_j^H|^2\} \forall i \in \mathbf{B}_1, j \in \mathbf{B}_2$
- 5: $[\mathbf{S}]_{i,j} \leftarrow \{\text{state}(i) \oplus \text{state}(j)\} \forall i \in \mathbf{B}_1, j \in \mathbf{B}_2$
- 6: Update $\mathbf{C} \leftarrow \mathbf{C} \circ \mathbf{S}$
- 7: Solve (17) to obtain \mathbf{A}^*
- 8: Pair into $\mathbb{K}_c^* \leftarrow \{i, j\} \forall i, j \ni [\mathbf{A}^*]_{i,j} = 1$, where $\forall c \in \mathbb{C}$
- 9: Assign $\vartheta(i) \leftarrow |\mathbb{K}_c|$; $\vartheta(j) \leftarrow |\mathbb{K}_c| - 1 \iff i \in \mathbf{B}_1, j \in \mathbf{B}_2$
- 10: **return** $\{\mathbb{K}_c^*\}, \{\vartheta(k_c)\}$

Boolean matrix representing a feasible NOMA pair among two users, and hence it can easily be shown that by satisfying modified constraints (17b) and (17c), the original constraint in (14g) is satisfied. Similarly, the decoding order assignment in (17) satisfies (14h). The computational complexity of **Algorithm 1** is governed primarily by solving (17), which is a function of total supported users K , therefore, the overall complexity is given by $O(KN_t + (\frac{K}{2})^3)$.

To optimize the remaining design parameters, the problem in (14) has been handled by solving the following three sub-problems given as

$$\max_{\mathbf{w}_c} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} R_{k,c}, \quad (18a)$$

$$\text{s.t. (14b), (14c), (14d).} \quad (18b)$$

$$\max_{\theta_{m,c}^x, \beta_{m,c}^x} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} R_{k,c}, \quad (19a)$$

$$\text{s.t. (14c), (14d), (14i), (14j).} \quad (19b)$$

$$\max_{p_{k,c}, t_c} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} R_{k,c}, \quad (20a)$$

$$\text{s.t. (14c), (14d), (14e), (14f).} \quad (20b)$$

where these problems correspond to active beamforming optimization, passive beamforming (STAR-RIS) optimization and the joint power and time-slot allocation optimization, respectively.

2) *Active Beamforming Optimization:* Since the objective function is non-convex with respect to \mathbf{w}_c , we modify the original sub-problem. Suppose $\mathbf{s}_{k,c} \triangleq \mathbf{g}_{k,c}^H \Phi_{\chi}^c \mathbf{H}$, $\mathbf{J}_{k,c} \triangleq \mathbf{s}_{k,c}^H \mathbf{s}_{k,c}$ and $\Upsilon_{k,c} \triangleq \{\gamma_{k,c}\} \forall k \in \mathbb{K}_c, c \in \mathbb{C}$, are the set of auxiliary variables. Let $r_{k,c}^{\min} = 2^{R_{k,c}^{\min}} - 1$, be the SNR corresponding to the minimum QoS rate requirement $R_{k,c}^{\min}$. The constraint (14c) can also be written as

$$r_{k,c}^{\min} \leq \gamma_{k \rightarrow i}^c$$

$$r_{k,c}^{\min} \leq \frac{|\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{i,c}}{\sum_{j=i+1}^{K_c} |\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{j,c} + \sigma^2}$$

$$r_{k,c}^{\min} \left(\sum_{j=i+1}^{K_c} |\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{j,c} + \sigma^2 \right) \leq |\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{i,c}$$

$$r_{k,c}^{\min} \sigma^2 \leq |\mathbf{h}_{k,c} \mathbf{w}_c|^2 \rho_{i,c}, \text{ where } \rho_{i,c} = p_{i,c} - r_{k,c}^{\min} \sum_{j=i+1}^{K_c} p_{j,c}.$$

Consequently, the problem (18a) can be re-written as

$$\max_{\mathbf{w}_{k,c}, \gamma_{k,c}} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} t_c (\log_2 (1 + \gamma_{k,c})), \quad (21a)$$

$$s.t. |\mathbf{s}_{k,c} \mathbf{w}_c|^2 \rho_{k,c} \geq r_{k,c}^{\min} \sigma^2, \quad (21b)$$

$$\frac{|\mathbf{s}_{k,c} \mathbf{w}_c|^2 p_{k,c}}{\sum_{j=k+1}^{K_c} |\mathbf{s}_{k,c} \mathbf{w}_c|^2 p_{j,c} + \sigma^2} \geq \gamma_{k,c}, \quad (21c)$$

$$(14b). \quad (21d)$$

The modified active beamforming sub-problem in (21), although having convex objective function, is still non-convex due to constraints (21b) and (21c). The sub-problem is further modified by defining auxiliary variables $\{\nu_{k,c}\} \forall k \in \mathbb{K}_c, c \in \mathbb{C}$. Applying successive convex approximation (SCA) method using first-order Taylor series for the non-convex term in (21b) around local point $\tilde{\mathbf{w}}_c$, it can be further re-written as

$$|\mathbf{s}_{k,c} \mathbf{w}_c|^2 \geq |\mathbf{s}_{k,c} \tilde{\mathbf{w}}_c|^2 + 2 \operatorname{Re} ((\tilde{\mathbf{w}}_c^H \mathbf{J}_{k,c}) (\mathbf{w}_c - \tilde{\mathbf{w}}_c)) \text{ or}$$

$$|\mathbf{s}_{k,c} \mathbf{w}_c|^2 \geq 2 \operatorname{Re} (\tilde{\mathbf{w}}_c^H \mathbf{J}_{k,c} \mathbf{w}_c) - |\mathbf{s}_{k,c} \tilde{\mathbf{w}}_c|^2$$

$$\geq \tau_{k,c} (\mathbf{w}_c). \quad (22)$$

Also, (21c) can be split into two sub-constraints given by

$$\frac{|\mathbf{s}_{k,c} \mathbf{w}_c|^2 p_{k,c}}{\nu_{k,c}} \geq \gamma_{k,c}, \quad (23)$$

$$\sum_{j=k+1}^{K_c} |\mathbf{s}_{k,c} \mathbf{w}_c|^2 p_{j,c} + \sigma^2 \leq \nu_{k,c}. \quad (24)$$

The constraint in (23) is also non-convex which is again written using first-order Taylor series for \mathbf{w}_c and $\nu_{k,c}$ around $\tilde{\mathbf{w}}_{k,c}$ and $\tilde{\nu}_{k,c}$, respectively, as

$$\frac{|\mathbf{s}_{k,c} \mathbf{w}_c|^2}{\nu_{k,c}} \geq \frac{2 \operatorname{Re} (\tilde{\mathbf{w}}_c^H \mathbf{J}_{k,c} \mathbf{w}_c)}{\tilde{\nu}_{k,c}} - \left(\frac{|\mathbf{s}_{k,c} \tilde{\mathbf{w}}_c|^2}{\tilde{\nu}_{k,c}} \right) \nu_{k,c} \quad (25)$$

$$\geq \tau_{k,c} (\mathbf{w}_c, \nu_{k,c}).$$

Using the modified constraints, the modified active beamforming optimization problem is given as

$$\max_{\mathbf{w}_{k,c}, \gamma_{k,c}} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} t_c (\log_2 (1 + \gamma_{k,c})), \quad (26a)$$

$$s.t. \tau_{k,c} (\mathbf{w}_c) \rho_{k,c} \geq r_{k,c}^{\min} \sigma^2, \quad (26b)$$

$$\tau_{k,c} (\mathbf{w}_c, \nu_{k,c}) p_{k,c} \geq \gamma_{k,c}, \quad (26c)$$

$$(24), (14b). \quad (26d)$$

Finally, the problem in (26) is a convex problem which provides a locally optimal solution around points $\tilde{\mathbf{w}}_{k,c}$ and

$\tilde{\nu}_{k,c}$. It is noted that the solution provided by (26) will be a lower-bound of the optimal solution to the original problem in (18).

3) *Reflection and Transmission Passive Beamforming Optimization*: To tackle the highly coupled STAR-RIS unit's reflection as well as the transmission phases and amplitudes, we take the sub-problem in (19a) and modify the original objective function by defining $\mathbf{u}_c^x \triangleq [\beta_{1,c}^x e^{j\theta_{1,c}^x}, \beta_{2,c}^x e^{j\theta_{2,c}^x}, \dots, \beta_{M,c}^x e^{j\theta_{M,c}^x}] \forall c \in \mathbb{C}$, where for convenience of calculation, $\beta_{m,c}^x \in [0, 1]$ and $\theta_{m,c}^x \in [0, 2\pi)$, $\forall m \in \mathbb{M}$, i.e., we relax the constraint (14j) temporarily by initially finding continuous valued optimal phases and amplitudes, and later on, find their discrete equivalents. We define the end-to-end channel $\mathbf{h}_{k,c}$ in terms of \mathbf{u}_c^x as, $\bar{\mathbf{h}}_{k,c} \triangleq \operatorname{diag}(\mathbf{g}_{k,c}^H) \mathbf{H} \mathbf{w}_c$, therefore, the SINR in (10) can also be expressed as

$$\text{SINR} = \frac{|\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 p_{k,c}}{\sum_{j=i+1}^{K_c} |\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 p_{j,c} + \sigma^2}. \quad (27)$$

The problem (19a) is modified in terms of SINR in (27) by defining set of auxiliary variables $\{\gamma_{k,c}\}$ and can be written as

$$\max_{\mathbf{u}_c^x, \gamma_{k,c}} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} t_c (\log_2 (1 + \gamma_{k,c})), \quad (28a)$$

$$s.t. |\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 \rho_{k,c} \geq r_{k,c}^{\min} \sigma^2, \quad (28b)$$

$$\frac{|\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 p_{k,c}}{\sum_{j=i+1}^{K_c} |\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 p_{j,c} + \sigma^2} \geq \gamma_{k,c}. \quad (28c)$$

The problem in (28) is still non-convex because of the non-convexity of the constraints (28b) and (28c). Auxiliary variables $\mathbf{Q}_{k,c} \triangleq \bar{\mathbf{h}}_{k,c} \bar{\mathbf{h}}_{k,c}^H$ and $\{\nu_{k,c}\}$ are introduced and the constraint (28c) is split into two sub-constraints as

$$\frac{|\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 p_{k,c}}{\nu_{k,c}} \geq \gamma_{k,c}, \quad (29)$$

$$\sum_{j=k+1}^{K_c} |\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 p_{j,c} + \sigma^2 \leq \nu_{k,c}. \quad (30)$$

Now, using a similar approach as done previously, we perform SCA assisting first-order Taylor approximations for left hand sides of the constraints (28b) and (29) for \mathbf{u}_c^x and $\nu_{k,c}$ around $\tilde{\mathbf{u}}_{k,c}$ and $\tilde{\nu}_{k,c}$ given by

$$|\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2 \geq 2 \operatorname{Re} (\tilde{\mathbf{u}}_c^x \mathbf{Q}_{k,c} (\mathbf{u}_c^x)^H) - |\tilde{\mathbf{u}}_c^x \bar{\mathbf{h}}_{k,c}|^2 = \tau_{k,c} (\mathbf{u}_c^x) \quad (31)$$

$$\frac{|\mathbf{u}_c^x \bar{\mathbf{h}}_{k,c}|^2}{\nu_{k,c}} \geq \frac{2 \operatorname{Re} (\tilde{\mathbf{u}}_c^x \mathbf{Q}_{k,c} (\mathbf{u}_c^x)^H)}{\tilde{\nu}_{k,c}} - \left(\frac{|\tilde{\mathbf{u}}_c^x \bar{\mathbf{h}}_{k,c}|^2}{\tilde{\nu}_{k,c}} \right) \nu_{k,c} = \tau_{k,c} (\mathbf{u}_c^x, \nu_{k,c}). \quad (32)$$

Finally, the problem in (28) can be expressed using (31) and (32) as

$$\max_{\mathbf{u}_c^x, \gamma_{k,c}} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} t_c (\log_2 (1 + \gamma_{k,c})), \quad (33a)$$

$$s.t. \tau_{k,c} (\mathbf{u}_c^x) \rho_{k,c} \geq r_{k,c}^{\min} \sigma^2, \quad (33b)$$

$$\tau_{k,c}(\mathbf{u}_c^X, v_{k,c}) p_{k,c} \geq \gamma_{k,c}, \quad (33c)$$

$$(30). \quad (33d)$$

The convex problem in (33) is used to obtain continuous \mathbf{u}_c^X , and then the feasible solution $\mathbf{u}_c^{X'}$ satisfying (14h) is obtained by a mean squared error (MSE) mapper (e.g., binary searched-based) as

$$\theta_m^{X'} = \arg \min_{\theta \in \Psi} |\theta - \text{angle}(\mathbf{u}_c^X)|^2, \quad (34a)$$

$$\beta_m^{X'} = \arg \min_{\beta \in \Omega} |\beta - |\mathbf{u}_c^X||^2. \quad (34b)$$

Finally, the $\mathbf{u}_c^{X'}$ is obtained using (34) as

$$\mathbf{u}_c^{X'} = [\beta_{1,c}^{X'} e^{j\theta_{1,c}^{X'}}, \beta_{2,c}^{X'} e^{j\theta_{2,c}^{X'}}, \dots, \beta_{M,c}^{X'} e^{j\theta_{M,c}^{X'}}]. \quad (35)$$

4) *Power and Time-slot Allocation*: The original problem of the power allocation is non-convex because of the objective function, and to maximize the system sum-rate both time-slot allocation and power allocation should be jointly optimized, as the following sub-problem:

$$\max_{t_c, p_{k,c}} \sum_{c \in \mathbb{C}} \sum_{k \in \mathbb{K}_c} t_c (\log_2 (1 + \gamma_{k,c})), \quad (36a)$$

$$s.t. (14c), (14d), (14e), (14f). \quad (36b)$$

Therefore, we derive analytical solutions for H-NOMA parameters such as user power allocation coefficients and cluster time-slot. First, optimal time-slot allocation and user power allocation are performed for all clusters except one with the highest aggregated user channel strength $\sum_{k \in \mathbb{K}_c} \|\mathbf{h}_{k,c}\|^2$, and then the remaining cluster's optimal allocations are handled separately. We solve a system of equations for all such clusters $c \in \{\mathbb{C} \setminus C\}$ in a way that their each user k is served at $R_{k,c}^{\min}$, which are given by

$$t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_1 \mathbf{w}_c|^2 p_{1,c}}{\sum_{j=1}^{K_c} |\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c} + \sigma^2} \right) \right) = R_{1,c}^{\min},$$

$$t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_2 \mathbf{w}_c|^2 p_{2,c}}{\sum_{j=2}^{K_c} |\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c} + \sigma^2} \right) \right) = R_{2,c}^{\min}, \quad (37)$$

$$\vdots$$

$$t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_{K_c} \mathbf{w}_c|^2 p_{K_c}}{\sigma^2} \right) \right) = R_{K_c}^{\min}, c \neq C.$$

Also, the sum-rate computed based on (37) can be given as

$$t_c \left(\sum_{k \in \mathbb{K}_c} \log_2 \left(1 + \frac{|\mathbf{h}_{k,c} \mathbf{w}_c|^2 p_{k,c}}{\sum_{j=k+1}^{K_c} |\mathbf{h}_{j,c} \mathbf{w}_c|^2 p_{j,c} + \sigma^2} \right) \right) = \sum_{k=1}^{K_c} R_{k,c}^{\min}. \quad (38)$$

For two users $\{i, j\} \in \mathbb{K}_c, \forall c \in \{\mathbb{C} \setminus C\}$ such that $\vartheta(i) < \vartheta(j)$, the following cluster sum-rate should therefore be satisfied

$$t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_i \mathbf{w}_c|^2 p_{i,c}}{p_{j,c} |\mathbf{h}_i \mathbf{w}_c|^2 + \sigma^2} \right) + \log_2 \left(1 + \frac{|\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c}}{\sigma^2} \right) \right)$$

$$= R_{i,c}^{\min} + R_{j,c}^{\min}.$$

The following systems of equations are solved to obtain optimal time-slot and power allocation

$$\begin{cases} t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_i \mathbf{w}_c|^2 p_{i,c}}{p_{j,c} |\mathbf{h}_i \mathbf{w}_c|^2 + \sigma^2} \right) \right) = R_{i,c}^{\min}, \\ t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c}}{\sigma^2} \right) \right) = R_{j,c}^{\min}. \end{cases} \quad (39)$$

Case 1: Paired users having similar QoS requirements ($R_{i,c}^{\min} \approx R_{j,c}^{\min}$)

Since the QoS requirements of the paired users $R_{i,c}^{\min}, R_{j,c}^{\min}$ are not different significantly, the closed-form solution to the power and time-slot allocations can be derived as

$$t_c = \frac{R_{j,c}^{\min}}{\log_2 \left(1 + \frac{|\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c}}{\sigma^2} \right)}$$

$$\frac{R_{j,c}^{\min}}{\log_2 \left(1 + \frac{|\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c}}{\sigma^2} \right)} \log_2 \left(1 + \frac{|\mathbf{h}_i \mathbf{w}_c|^2 (1 - p_{j,c})}{p_{j,c} |\mathbf{h}_i \mathbf{w}_c|^2 + \sigma^2} \right) = R_{i,c}^{\min}.$$

Since $(R_{j,c}^{\min}/R_{i,c}^{\min}) \rightarrow 1$, we approximate the QoS requirements of the two users by defining $R_{k,c}^{\min} \triangleq R_{i,c}^{\min} \approx R_{j,c}^{\min}$

$$\log_2 \left(1 + \frac{|\mathbf{h}_i \mathbf{w}_c|^2 (1 - p_{j,c})}{p_{j,c} |\mathbf{h}_i \mathbf{w}_c|^2 + \sigma^2} \right) = \log_2 \left(1 + \frac{|\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c}}{\sigma^2} \right)$$

$$\frac{|\mathbf{h}_i \mathbf{w}_c|^2 (1 - p_{j,c})}{p_{j,c} |\mathbf{h}_i \mathbf{w}_c|^2 + \sigma^2} = \frac{|\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c}}{\sigma^2}$$

$$p_{j,c}^2 (|\mathbf{h}_i \mathbf{w}_c|^2 |\mathbf{h}_j \mathbf{w}_c|^2) + p_{j,c} |\mathbf{h}_j \mathbf{w}_c|^2 \sigma^2 = |\mathbf{h}_i \mathbf{w}_c|^2 \sigma^2 - p_{j,c} |\mathbf{h}_i \mathbf{w}_c|^2 \sigma^2$$

$$p_{j,c}^2 (|\mathbf{h}_i \mathbf{w}_c|^2 |\mathbf{h}_j \mathbf{w}_c|^2) + p_{j,c} \sigma^2 (|\mathbf{h}_i \mathbf{w}_c|^2 + |\mathbf{h}_j \mathbf{w}_c|^2) = \sigma^2 |\mathbf{h}_i \mathbf{w}_c|^2. \quad (40)$$

The equation in (40) is feasibly solved for $0 \leq p_{j,c} \leq 1$ to obtain the optimal power allocation coefficients $p_{k,c}^*$, where $c \neq C$ given by (41) in the next page. Using $p_{j,c}^*$, the optimal power allocation coefficient for user i is calculated as $p_{i,c}^* = 1 - p_{j,c}^*$. Finally, the optimal time-slot allocation t_c^* substituting $p_{j,c}^*$ in (39) can be obtained, and is given by (42) in the next page.

Case 2: Paired users having different QoS requirements ($R_{i,c}^{\min} \neq R_{j,c}^{\min}$)

Finding the closed form solution analytically is dependent on the $R_{i,c}^{\min}/R_{j,c}^{\min}$, by which the computational complexity increases significantly for small changes in given ratio. Thus, for the considered case, it is feasible to find joint power and time-slot allocation by solving (39) numerically rather than analytically satisfying the constraints (14e) and (14f).

Now we derive the expression for optimal power allocation $p_{k,c}^*$ for cluster C with the highest sum of user channel strength. For the considered case of user i having lower decoding order than user j such that $\vartheta(i) < \vartheta(j)$, the optimal sum-rate maximization approach is to serve the user i with power coefficient satisfying its QoS requirement and allocate the remaining power coefficient to user j . The derivation to obtain $p_{k,c}^*$ based on this approach is as follows.

For the two-user case, since the IntraC at user i is given by $\mathcal{I}(i, c) = 1 - p_{i,c}$, the rate of user i can be written as

$$\begin{aligned}\gamma_{i \rightarrow i}^c &= \frac{|\mathbf{h}_i \mathbf{w}_c|^2 p_{i,c}}{|\mathbf{h}_i \mathbf{w}_c|^2 (1 - p_{i,c}) + \sigma^2}, \\ \gamma_{i \rightarrow i}^c |\mathbf{h}_i \mathbf{w}_c|^2 - \gamma_{i \rightarrow i}^c (p_{i,c}) |\mathbf{h}_i \mathbf{w}_c|^2 + \gamma_{i \rightarrow i}^c \sigma^2 &= |\mathbf{h}_i \mathbf{w}_c|^2 p_{i,c}, \\ \gamma_{i \rightarrow i}^c |\mathbf{h}_i \mathbf{w}_c|^2 + \gamma_{i \rightarrow i}^c \sigma^2 &= |\mathbf{h}_i \mathbf{w}_c|^2 p_{i,c} + \gamma_{i \rightarrow i}^c (p_{i,c}) |\mathbf{h}_i \mathbf{w}_c|^2, \\ \gamma_{i \rightarrow i}^c |\mathbf{h}_i \mathbf{w}_c|^2 + \gamma_{i \rightarrow i}^c \sigma^2 &= |\mathbf{h}_i \mathbf{w}_c|^2 p_{i,c} (1 + \gamma_{i \rightarrow i}^c), \\ p_{i,c} &= \frac{\gamma_{i \rightarrow i}^c}{(1 + \gamma_{i \rightarrow i}^c)} + \frac{\gamma_{i \rightarrow i}^c \sigma^2}{|\mathbf{h}_i \mathbf{w}_c|^2 (1 + \gamma_{i \rightarrow i}^c)}, \\ p_{i,c} &= \frac{\gamma_{i \rightarrow i}^c}{(1 + \gamma_{i \rightarrow i}^c)} \left(1 + \frac{\sigma^2}{|\mathbf{h}_i \mathbf{w}_c|^2} \right).\end{aligned}$$

Similarly, for user k having decoding order $\vartheta(k)$, power allocation coefficient $p_{k,c}$ can be generalized as

$$p_{k,c} = \frac{\gamma_{k \rightarrow k}^c}{(1 + \gamma_{k \rightarrow k}^c)} \left(1 + \frac{\sigma^2}{|\mathbf{h}_k \mathbf{w}_c|^2} - \sum_{i=1}^{k-1} p_{i,c} \right),$$

where it is supposed $\gamma_{k \rightarrow k}^c = 2^{R_{k,c}^{\min}/t_c} - 1$ with t_c being the allocated time-slot. Therefore, the generalized solution to optimal power allocation coefficient is given by (43).

Overall, the solutions to the user pairing and decoding order assignment as well as the active, passive beamforming and time and power allocation are summarized as a unified framework based on alternative optimization in **Algorithm 2**. To facilitate faster convergence of the proposed algorithm, active and passive beamformers can be feasibly initialized using the estimated CSI. For example, the set of initial active beamforming vectors can be found using the maximal-ratio-transmission (MRT) precoding corresponding to the strongest user of the cluster given by

$$\mathbf{w}_c^{(0)} = \frac{\mathbf{h}_{s,c}^{(0)}}{\|\mathbf{h}_{s,c}^{(0)}\|}, \text{ where } \mathbf{h}_{s,c}^{(0)} = \arg \max_{\mathbf{h}_{k,c} \in \{\mathbf{h}_{k,c}^{(0)}\}} \|\mathbf{h}_{k,c}\|^2. \quad (44)$$

Similarly, these initial active beamformers can be used to feasibly initialize passive beamforming vectors given by

$$\mathbf{u}_c^{\chi(0)} = \frac{\beta_m^{\chi(0)} \left(\text{diag} \left(\mathbf{g}_{k,c}^H \right) \mathbf{H} \mathbf{w}_c^{(0)} \right)^H}{\text{diag} \left(\mathbf{g}_{k,c}^H \right) \mathbf{H} \mathbf{w}_c^{(0)}}, \quad (45)$$

where $\beta_m^{\chi(0)}$, $p_{k,c}^{(0)}$, and $t_{k,c}^{(0)}$ can be initialized randomly satisfying the unit-bound constraint.

Lemma 1. For a cluster c with distinct users $i, j \in \mathbb{K}_c$, given the active beamforming vector \mathbf{w}_c , passive beamforming vector \mathbf{u}_c^{χ} , power allocation coefficients $\{p_{k,c}\}$ and time allocation t_c optimized through (26), (35), (41), (43) and (42), iff the decoding order through **Algorithm 1** is assigned such that $\vartheta(i) < \vartheta(j)$, then the following NOMA-SIC constraint is guaranteed

$$R_{j \rightarrow i} \geq R_{i \rightarrow i}.$$

Thus, without significant loss of generality, (14c) can be satisfied.

Proof: From (11) and (12), the decoding rates $R_{j \rightarrow i}$ and $R_{i \rightarrow i}$ are given, respectively, as

$$\begin{aligned}R_{j \rightarrow i} &= t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_j \mathbf{w}_c|^2 p_{i,c}}{|\mathbf{h}_j \mathbf{w}_c|^2 p_{j,c} + \sigma^2} \right) \right), \\ R_{i \rightarrow i} &= t_c \left(\log_2 \left(1 + \frac{|\mathbf{h}_i \mathbf{w}_c|^2 p_{i,c}}{|\mathbf{h}_i \mathbf{w}_c|^2 p_{j,c} + \sigma^2} \right) \right).\end{aligned}$$

In the above expressions, the decoding rates are functions of terms $|\mathbf{h}_j \mathbf{w}_c|^2$ and $|\mathbf{h}_i \mathbf{w}_c|^2$ only as the terms $p_{i,c}$, $p_{j,c}$, and σ^2 are constants. In addition, the active beamformer \mathbf{w}_c in accordance with the given decoding order has been obtained such that the following inequality always holds true.

$$|\mathbf{h}_j \mathbf{w}_c|^2 \geq |\mathbf{h}_i \mathbf{w}_c|^2. \quad (47)$$

Now, we write decoding rates as

$$R_{j \rightarrow i} = t_c \left(\log_2 \left(1 + \frac{p_{i,c}}{p_{j,c} + \frac{\sigma^2}{|\mathbf{h}_j \mathbf{w}_c|^2}} \right) \right),$$

$$p_{j,c}^* = \frac{-\sigma^2 (|\mathbf{h}_i \mathbf{w}_c|^2 + |\mathbf{h}_j \mathbf{w}_c|^2) + \sigma \left(\sqrt{\sigma^2 (|\mathbf{h}_i \mathbf{w}_c|^2 + |\mathbf{h}_j \mathbf{w}_c|^2)^2 + 4 |\mathbf{h}_j \mathbf{w}_c|^2 |\mathbf{h}_i \mathbf{w}_c|^4} \right)}{2 |\mathbf{h}_i \mathbf{w}_c|^2 |\mathbf{h}_j \mathbf{w}_c|^2}. \quad (41)$$

$$t_c^* = \begin{cases} \frac{R_{k,c}^{\min}}{\log_2 \left(1 + \frac{-\sigma (|\mathbf{h}_i \mathbf{w}_c|^2 + |\mathbf{h}_j \mathbf{w}_c|^2) + \sqrt{\sigma^2 (|\mathbf{h}_i \mathbf{w}_c|^2 + |\mathbf{h}_j \mathbf{w}_c|^2)^2 + 4 |\mathbf{h}_j \mathbf{w}_c|^2 |\mathbf{h}_i \mathbf{w}_c|^4}}{2 \sigma |\mathbf{h}_i \mathbf{w}_c|^2} \right)}} & \text{if } c \in \{\mathbb{C} \setminus C\} \\ 1 - \sum_{c \in \{\mathbb{C} \setminus C\}} t_c & \text{if } c = C. \end{cases} \quad (42)$$

$$p_{k,C}^* = \begin{cases} \left(1 - \frac{1}{2^{R_{k,c}^{\min}/t_c}} \right) \left(1 + \frac{\sigma^2}{|\mathbf{h}_k \mathbf{w}_c|^2} - \sum_{i=1}^{k-1} p_{i,c} \right) & \text{if } \vartheta(K_c) \geq \vartheta(k) \\ 1 - p_{k,c} & \text{if } k = K_c. \end{cases} \quad (43)$$

Table I
SIMULATION PARAMETERS

Notation	Description	Value
P_{max}	BS transmit power	30 dBm
M	Number of STAR-RIS elements	16
N	Number of BS antenna elements	16
K	Number of users	6
C	Number of clusters	3
PL_{do}	Path loss at reference distance	60 dBm
$\eta_{(BS-STAR-RIS)}$	Path-loss exponent of BS-STAR-RIS link	2.2
$\eta_{(STAR-RIS-User)}$	Path-loss exponent of STAR-RIS-user link	2.8
ζ	Shadowing	5.8 dB
BW	Bandwidth	10 MHz
σ^2	Noise power	-104 dBm
T_{max}	Channel coherence time	650 μ sec

$$R_{i \rightarrow i} = t_c \left(\log_2 \left(1 + \frac{p_{i,c}}{p_{j,c} + \frac{\sigma^2}{|\mathbf{h}_i \mathbf{w}_c|^2}} \right) \right).$$

From (47), we can easily deduce that

$$R_{j \rightarrow i} \geq R_{i \rightarrow i}.$$

Proposition 1. For a cluster c with users $i, j \in \mathbb{K}_c$ such that $\vartheta(i) < \vartheta(j)$, the active beamforming vector \mathbf{w}_c , passive beamforming vector \mathbf{u}_c^* , power allocation coefficients $\{p_{k,c}\}$ and time allocation t_c optimized through (26), (35), (41), (43) and (42) are guaranteed to ensure QoS such that

$$R_{k \rightarrow k} \geq R_{k,c}^{min} \quad k \in \{i, j\}.$$

Using Lemma 1, we also obtain that

$$\begin{aligned} R_{j \rightarrow i} &\geq R_{i \rightarrow i} \geq R_{i,c}^{min} \\ \Rightarrow R_{j \rightarrow i} &\geq R_{i,c}^{min}, \end{aligned}$$

which without loss of generality satisfies (14d).

The proposed solutions to the aforementioned problems have polynomial-time complexity, and the individual computational complexity of the problem in (26) and (33) are $O(CK_c(2N_t^3 + 2K_c))$ and $O(CK_c(2M^3 + 2K_c))$, respectively, while the solution to (14e) and (14f) can be obtained analytically or numerically immediately.

IV. SIMULATION RESULTS

In this section, we present simulation results to evaluate the STAR-RIS mmWave communication system and validate the effectiveness of the proposed sum-rate maximization algorithm. We assume that the BS is located in a 3D space at $(x, y, z) = (0 \text{ m}, 0 \text{ m}, 20 \text{ m})$ and communicating at 28 GHz frequency with the STAR-RIS placed at $(45 \text{ m}, -22 \text{ m}, 0 \text{ m})$ to serve users randomly scattered around it (i.e., within a circular region of radius 50 m centered at the STAR-RIS). The parameters used for the simulation are given in Table I.

Fig. 3 shows the sum-rate optimized by the proposed **Algorithm 2** with the optimal user pairing for different values of M and P_{max} . The results are obtained from Monte Carlo simulations over 10^4 realizations using the maximum number of iterations $\mathcal{T}_{max} = 30$ and the convergence threshold $\xi = 0.1$

Algorithm 2 Proposed Algorithm for Sum-rate Maximization

- 1: **Input:** $\{\mathbf{h}_{k,c}\}, K, N_t, M, P_{max}, R_{k,c}^{min}, \sigma^2$, threshold ξ , and maximum iterations $\mathcal{T}_{max}, \forall k \in \mathbb{K}_c, c \in \mathbb{C}$.
- 2: Initialize with random user pairing to obtain $\{\mathbf{h}_{k,c}^{(0)}\}, \forall k \in \mathbb{K}_c, c \in \mathbb{C}$.
- 3: Perform user pairing and decoding order allocation using **Algorithm 1**($K, \{\mathbf{h}_k\}$) to obtain $\mathbb{K}_c^*, \vartheta(k_c)^*$.
- 4: Initialize the optimization parameters: $\{\mathbf{w}_c^{(0)}\}, \{\mathbf{u}_c^{(0)}\}, \{p_{k,c}^{(0)}\}, \{t_c^{(0)}\}$ and threshold ξ feasibly (e.g., using (44) and (45)).
- 5: **repeat**
- 6: Update $\mathbf{w}_c^{(i)}$ using $\mathbf{u}_c^{(i-1)}, p_{k,c}^{(i-1)}, t_c^{(i-1)}$ by solving the problem in (26);
- 7: Calculate $\mathbf{u}_c^{(i)}$ using $\mathbf{w}_c^{(i)}, p_{k,c}^{(i-1)}, t_c^{(i-1)}$ by solving the problem in (33);
- 8: Update $\mathbf{u}_c^{(i)}$ using (34) and (35);
- 9: Update $\mathbf{h}_{k,c}^{(i)} \leftarrow \mathbf{g}_{k,c}^H \Phi_c^{X'} \mathbf{H}$, where $\Phi_c^{X'} = \text{diag}(\mathbf{u}_c^{(i)})$;
- 10: Update $p_{k,c}^{(i-1)}$ and $t_c^{(i-1)}$ via joint optimization using $\mathbf{h}_{k,c}^{(i)}, \mathbf{w}_c^{(i)}$ through (41), (43) and (42);
- 11: **if** $|R_{sum} - R_{sum}^{(i-1)}|^2 \leq \xi$ or $i > \mathcal{T}_{max}$ **then**
- 12: $\{\mathbf{w}_c^*\} = \{\mathbf{w}_c^{(i)}\}, \{\mathbf{u}_c^{X*}\} = \{\mathbf{u}_c^{(i)}\}, \{p_{k,c}^*\} = \{p_{k,c}^{(i)}\}, \{t_c^*\} = \{t_c^{(i)}\}, R_{sum}^* = R_{sum}^{(i)}$;
- 13: **break**
- 14: **end if**
- 15: Set $i \leftarrow i + 1$;
- 16: **until** the value of the objective of the problem in (14) converges
- 17: **Output:** $R_{sum}^*, \{\mathbf{w}_c^*\}, \{\mathbf{u}_c^{X*}\}, \{p_{k,c}^*\}, \{t_c^*\}$.

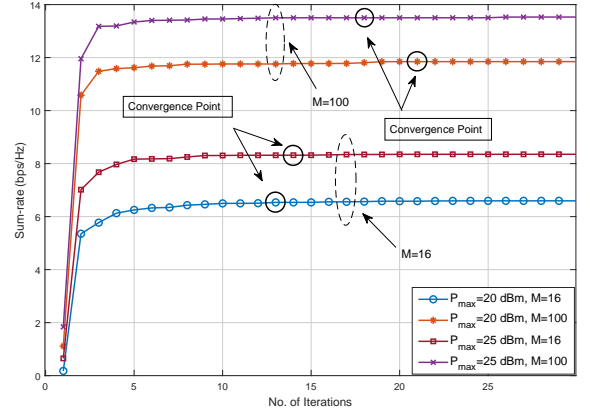


Figure 3. Sum-rate vs. the number of iterations showing the convergence of **Algorithm 2**

bps/Hz. First, it can be observed that we can achieve higher sum-rate, as M and P_{max} increase. It also shows that the proposed iterative algorithm converges for various values of parameters such as the STAR-RIS elements and BS power, i.e., $M \in \{16, 36\}$ and $P_{max} \in \{20, 30\}$ dBm. As shown in the figure, the algorithm, initialized feasibly, converges to the optimal sum-rate value in less than 25 iterations even for low values of ξ . While power and time allocation is performed

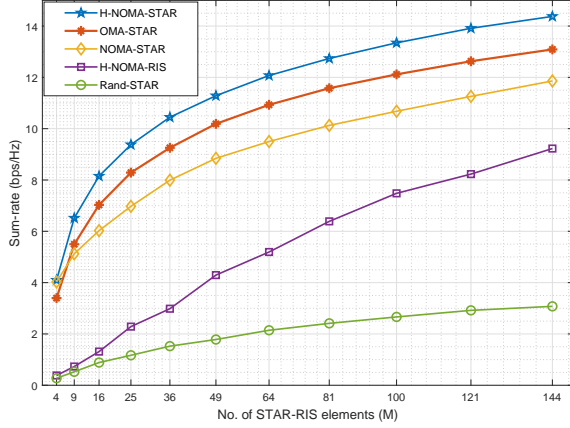


Figure 4. Sum-rate vs. the number of STAR-RIS elements M for various schemes

optimally in each iteration, the overall algorithm converges as the optimal active and passive beamformers \mathbf{w}_c and $\mathbf{u}_{k,c}^x$ are achieved in an iterative fashion utilizing the CSI. The result also demonstrates that a larger number of STAR-RIS elements M can be effectively used to obtain a higher sum-rate in an energy-constrained system because it is more likely to modify the blocked BS-STAR-RIS channels and improve the sum-rate by appropriately adjusting their amplitudes β_m and phases θ_m .

In Fig. 4, to better evaluate our proposed **Algorithm 2**, we compare it with four baseline schemes, i) STAR-RIS-assisted OMA-STAR, ii) STAR-RIS-assisted NOMA (NOMA-STAR), and iii) hybrid-NOMA-RIS (H-NOMA-RIS), iv) Rand-STAR, with different numbers of STAR-RIS elements M . For NOMA-STAR and H-NOMA-RIS, the users are served with random user pairing and decoding orders, whereas each user is served in its respective sub-time-slot in OMA-STAR. The time allocation for NOMA-STAR and H-NOMA-RIS is done using (42), whereas for OMA-STAR the sub-time-slots allocation is made in a similar manner, as in (42). For NOMA-STAR, we reproduce the results of [30], where each cluster is served simultaneously with its own beamformer, thus introducing inter-cluster-interference (InterCI) in the system model. Therefore, the optimal active and passive beamformers \mathbf{w}_c and $\mathbf{u}_{k,c}^x$ are realized by incorporating the InterCI as in (31) and (33) in [30]. In the H-NOMA-RIS scheme, the STAR-RIS is set to operate only in reflection mode ($\beta_{m,c}^r \approx 1$), while the value of the transmission coefficient $\beta_{m,c}^t$ is negligible. An RIS is usually constructed as a metallic sheet, which has an intermediate copper back-plane to minimize the penetration of the impinging signals for maximum reflections [37]. In addition, a rectangular metallic sheet adds an average path loss of around 20 dBm at mmWave (e.g., 28 GHz) [38], which is also accounted for in the STAR-RIS transmission users' channels $\mathbf{g}_{k,c}^t$. For OMA-STAR and H-NOMA-RIS, the beamformers are obtained using our proposed algorithm with CSI. The optimal power allocation $p_{k,c}$ is performed using (41) and (43) for STAR-NOMA and H-NOMA-RIS. For OMA-STAR, since the users are served in orthogonal sub-time-slots, there is no need for power allocation. For the Rand-STAR

scheme, random user pairing, power allocation, active and passive beamforming are performed to compare the percentage increase in the sum-rate acquired. It can be observed from the figure that the achievable sum-rates of all the schemes (i.e., H-NOMA-STAR, OMA-STAR and NOMA-STAR) increase, as the number of STAR-RIS elements increases. This is due to the fact that the element array \mathbf{u}_c^x will have more potential to form a pointed beam towards the intended user (in either reflection or transmission) by strengthening the respective end-to-end channel gain $\mathbf{h}_{k,c}$, as indicated by the Rand-STAR's increase in achievable sum-rate with increasing M .

Fig. 4 also shows that the proposed scheme always outperforms other baseline schemes because operating STAR on H-NOMA not only leverages NOMA's effective resource allocation capability for higher SE, but also eliminates the InterCI to maximize the achievable sum-rate. For the H-NOMA-RIS scheme, as a result of the blocked BS-user (transmission) channel, highly unfair power allocation is essential to serve users, ensuring the QoS requirement $R_{k,c}^{\min}$ particularly for lower values of M . The result shows that for the cluster's mean $R_{k,c}^{\min} = 0.1$ bps/Hz, approximately $M > 64$ is required to serve all clusters satisfying the QoS constraint. While for a higher number of elements, i.e., $M > 64$, the majority of the sum-rate is achieved through reflection users via passive reflect beamformer $\mathbf{u}_{k,c}^r$ and the power allocation coefficients $p_{k,c}$ are adjusted to overcome RIS-blockage to serve transmission users. Therefore, to acquire a sum-rate equal to that of H-NOMA-STAR via H-NOMA-RIS, a very high value of M or P_{max} is required. For NOMA-STAR, the sum-rate is limited by the InterCI and hence, unless the users have uncorrelated channels, the proposed scheme yields greater sum-rate, whereas for OMA-STAR, further division of the resource (i.e., time-slots) reduces the sum-rate. Therefore, we can conclude that H-NOMA-STAR provides, in general, a flexible solution, whereas the other two schemes are highly subject to user-channel correlations and the resultant InterCI.

Fig. 5 illustrates the average increase in the sum-rate obtained through the user pairing and decoding order strategy proposed in **Algorithm 1**. The results are obtained with $M = 16$, $P_{max} = 30$ dBm, and mean cluster $R_{k,c}^{\min} = 0.1$ bps/Hz by finding the optimal H-NOMA resource allocation parameters (i.e., $t_c, p_{k,c}$), and averaging the gain achieved with respect to random user pairing and decoding order over 10^4 realizations. The result further presents three different cases proving that the gain achieved also depends on the number of clusters being served and the active and passive beamforming vectors:

- **Random Active and Passive Beamformers:** The average gain obtained through optimal pairing and decoding order allocation for random \mathbf{w}_c , \mathbf{u}_c^x is the lowest among the three cases because of the weak channel (non-optimized) correlations $|\mathbf{h}_j \mathbf{h}_i^H|^2$. It can also be observed that on a larger number of clusters C , serving users while guaranteeing their QoS would require allocating smaller time slots, and hence, despite optimal pairing and decoding order, relatively smaller time-slot allocated to the clusters limits the achievable sum-rate gain.
- **Optimized Active and Passive Beamformers:** In this

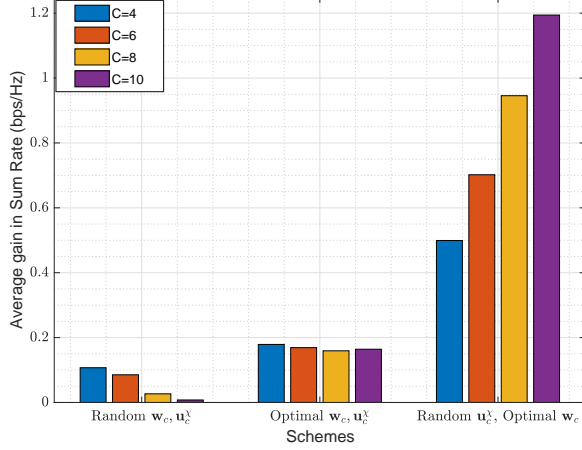


Figure 5. The average gain in sum-rate obtained via pairing and decoding order with **Algorithm 1**

case, since the optimal active and passive beamforming vectors w_c, u_c^x are used to serve each cluster and its users, except for the few highly blocked users with weaker channel strength $\|h_{k,c}\|$, each user can act as the sum-rate maximizer mainly depending on their STAR-RIS-user channels $g_{k,c}^H$, therefore, the proposed channel correlation-based pairing, on average, yields a constant gain in the sum-rate. For the given simulation parameters, a gain of around 0.2 bps/Hz for $C = 4, 6, 8, 10$ is achieved.

- **Random Passive and Optimized Active Beamformer:**

In this case, relatively more significant sum-rate gain is obtainable. This is due to the fact that obtaining optimal w_c can help increase a cluster's sum-rate even with random u_c^x (as highlighted in Fig. 7) by allocating longer time-slots to the cluster with the highest intrinsic sum-rate. Additionally, since the passive beamformers are not optimized, user-channel correlations $|h_j h_i^H|^2$ with different pairing and decoding order combination can have a noticeable difference in the sum-rate. Therefore, the optimal pairing and decoding order combination can be leveraged to have a decent increase in the sum rate, specifically at larger values of C due to the increase in the number of possible pairing combinations.

Fig. 6 depicts the expected number of clusters C that can be served by the BS-STAR-RIS system for a simulation setting with $P_{max} = 30 \text{ dBm}$, $M = 16$, such that all users are ensured their QoS requirements. We show that for various values of average $R_{k,c}^{\min}$, the expected number of clusters also largely depends upon w_c, u_c^x and the operating STAR-RIS ES-ratio $\beta_{m,c}^r/\beta_{m,c}^t$. For analysis purposes, the results have been simulated by allocating higher decoding order to every user i served via the STAR-RIS reflection such that $\vartheta(i) > \vartheta(j)$, where $\text{state}(i) = r^i$ and $\text{state}(j) = t^j$, and t_c and $p_{k,c}$ have been optimized using (41) and (43). We also observe that the system with random w_c, u_c^x is able to support, on average, no more than six clusters for $\beta_{m,c}^r/\beta_{m,c}^t = 0.05, 5$, and the number of the mean clusters supported obviously

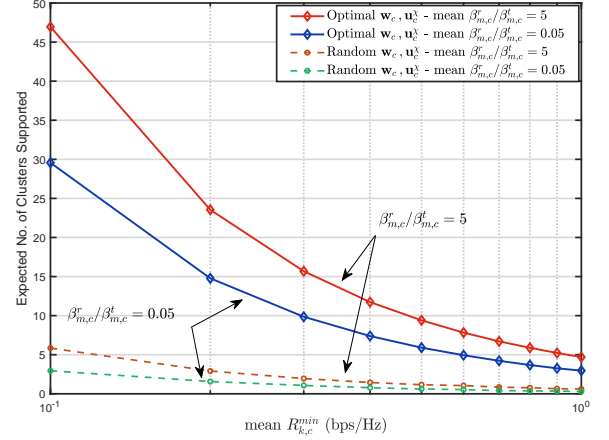


Figure 6. The average number of clusters C supported with $P_{max} = 30 \text{ dBm}$, $M = 16$ for different mean ES-ratio. $\beta_{m,c}^r/\beta_{m,c}^t$

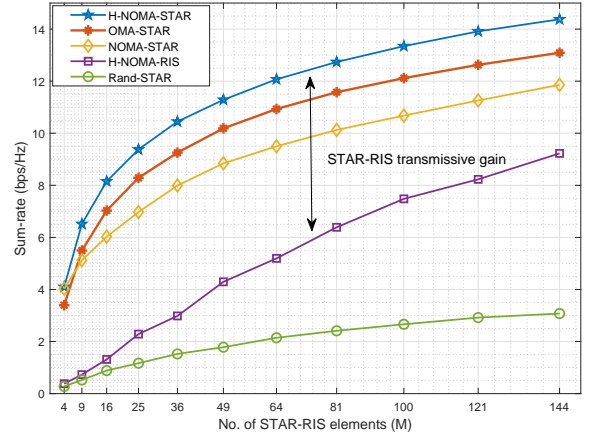


Figure 7. The contribution of $w_c, u_c^x, t_c, p_{k,c}$ to the sum-rate improvement with varying M , when $P_{max} = 25 \text{ dBm}$, and mean $R_{k,c}^{\min} = 0.1 \text{ bps/Hz}$.

decreases with increasing average $R_{k,c}^{\min}$ because of the increase in the average time-slot t_c allocated to a cluster. However, the optimal values of w_c, u_c^x greatly reduce t_c , while ensuring the same QoS requirements. Consequently, the spare time-slot could potentially be used to serve a relatively larger number of clusters by the system on a given $R_{k,c}^{\min}$. Moreover, by optimizing $\beta_{m,c}^r/\beta_{m,c}^t$ to achieve a higher sum-rate (e.g. $\beta_{m,c}^r/\beta_{m,c}^t = 5$, in this case), the average number of clusters supported can be further increased.

In Fig. 7, we highlight the impact of each design parameter, i.e. active and passive beamformers w_c, u_c^x , allocated resources $t_c, p_{k,c}$, and user-pairing and decoding order on the sum-rate. It can be observed that for $P_{max} = 30 \text{ dBm}$ and $C = 3$ although optimizing the active beamformer w_c can yield a decent gain in the sum-rate. However, with increasing M , it only increases with a rate similar to that of random w_c, u_c because of the constrained P_{max} and the unoptimized passive beamformer. While the impact of optimal w_c is greater compared to u_c^x for $M < 16$, a significant gain (about 30%)

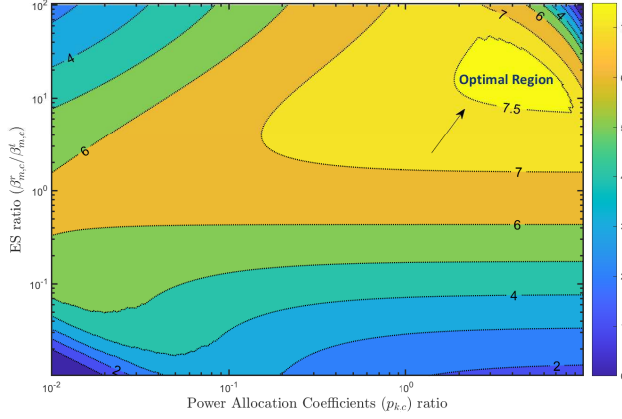


Figure 8. Achievable sum-rate for different Energy Splitting ratio ($\beta_{m,c}^r/\beta_{m,c}^t$) with $P_{max} = 25$ dBm, $C = 1$, $M = 16$, mean $R_{k,c}^{\min} = 0.1$ bps/Hz.

can be achieved by optimizing \mathbf{u}_c^X for higher values of M even with random \mathbf{w}_c . The result also gives a strong insight that the design of optimal active and passive beamformers is the most crucial one in a STAR-RIS-assisted system for higher sum-rate performance gains, as by only optimizing \mathbf{w}_c and \mathbf{u}_c , approximately 80% of the total achievable sum-rate (through a fully optimized system), on average, is obtainable. It is also worth noticing that H-NOMA using optimal t_c , $p_{k,c}$ can further enhance the sum-rate again verifying the effectiveness of H-NOMA scheme. As depicted more clearly in Fig. 5, it can be re-examined that through optimal user pairing and decoding order, only a constant gain enhancement (about 5-10%) is achievable.

The achievable sum-rate with optimal \mathbf{w}_c , \mathbf{u}_c^X for various STAR-RIS ES ($\beta_{m,c}^r/\beta_{m,c}^t$) and H-NOMA power allocation coefficients $p_{k,c}$ ratios is represented in Fig. 8. To delve into the impacts of the ES and power allocation coefficients on the sum-rate, similar to Fig. 6, we simulated two users i and j being served via STAR-RIS with decoding orders such that $\vartheta(i) > \vartheta(j)$, with $C = 1$. It can be clearly seen from the contour map that for extreme values of ES i.e. $\beta_{m,c}^r/\beta_{m,c}^t \gg 1$, and $\beta_{m,c}^r/\beta_{m,c}^t \ll 1$, the sum-rate drops from the optimal value because despite obtaining optimal STAR-RIS phases $\theta_{m,c}^X$, the imbalance in $\beta_{m,c}^r/\beta_{m,c}^t$ restricts the STAR-RIS QoS ensured coverage to the blocked (either reflection or transmission) side's users for the fixed $P_{max} = 25$ dBm. Likewise, allocating extreme power coefficients $p_{k,c}$ to the users also limits the sum-rate. For the given scenario, it can also be observed that because user j has the lower decoding order, therefore, due to the IntraCI at user j by user i , the optimal sum-rate can be achieved by allocating a relatively higher value of the power coefficient to user i than j such that $p_{i,c} > p_{j,c}$. It is also worth noticing that due to the IntraCI at j , since the reflection user i has higher potential to increase the sum-rate, the rate of increase in the sum-rate due to slight increase in $\beta_{m,c}^r$ is higher than that caused by $\beta_{m,c}^t$. Hence, to operate in optimal region, the ES and power coefficients' ratio slightly favors the user with better channel condition, which is

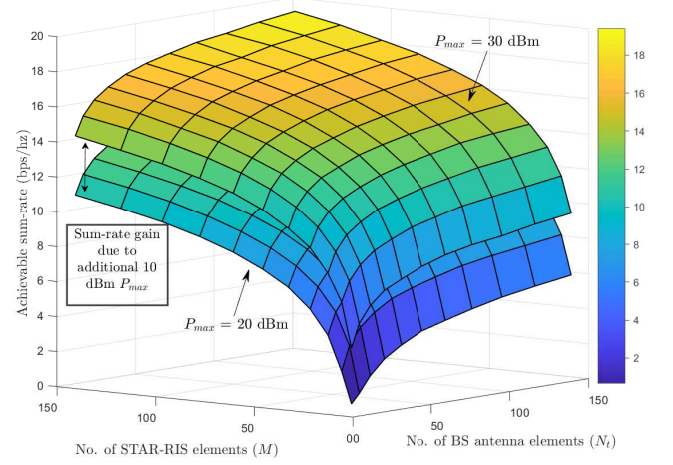


Figure 9. Achievable sum-rate against different BS transmit antennas N_t and STAR-RIS elements M with mean $R_{k,c}^{\min} = 0.1$ bps/Hz.

to be served with higher decoding order to obtain the optimal sum-rate.

Fig. 9 also examines the importance of two design parameters of the system model, i.e., the BS transmit antennas N_t and the number of STAR-RIS elements M , on the achievable sum-rate using the proposed **Algorithm 1** and **Algorithm 2**. For analysis purposes, the results are obtained on two different BS powers i.e., $P_{max} = 20$ dBm and $P_{max} = 30$ dBm. We observe that the achievable sum-rate is more sensitive to the STAR-RIS elements M compared to the BS transmit antennas N_t . This is because on increasing N_t (while keeping M constant), the end-to-end BS-user channel $\mathbf{h}_{k,c}$ will cause the major increase in the sum-rate as the active beamforming vector \mathbf{w}_c will have an approximately constant effect on the sum rate due to the maximum power constraint in (14b). On the other hand, increasing M will cause more elements to coherently contribute to increase the sum-rate using their corresponding tunable coefficients $\beta_{m,c}^X$ and independent phases $\theta_{m,c}^X$.

V. CONCLUSION

In this paper, we have investigated a STAR-RIS-assisted mmWave communication for sum-rate maximization and proposed a comprehensive optimization framework in the presence of blockages. We have handled the coupled non-convex optimization problem by solving sub-problems for the active and passive beamforming optimization. Moreover, we have also studied the appropriate multiple access scheme for the given scenario. Our results have confirmed that the proposed H-NOMA scheme outperforms the conventional NOMA and OMA. A channel correlation-based technique has been proposed to group a pair of reflective and transmissive users and find their NOMA decoding orders. Analytical solutions to the joint power-time allocation problem have also been derived for the optimal H-NOMA operation. Our results have shown that the STAR-RIS transmission property presents an ideal solution to the blocked communication coverage, since

a specific number of STAR-RIS elements can be leveraged to obtain significant gains over the conventional RIS regardless of the chosen multiple access scheme. To fully exploit the advantages of the STAR-RIS-aided H-NOMA communication, more works on the accurate downlink CSI acquisition through STAR-RIS is required to realize the enhanced beamforming gains in practice. Moreover, as suggested earlier, the user pairing algorithm with different cluster sizes is also worth studying to further increase the sum-rate. Finally, although several works have explored the optimal deployments of the conventional RISs, the deployment strategies of STAR-RISs to improve the existing networks in the light of the stochastic geometry can be included in the future extensions.

REFERENCES

- [1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.
- [2] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tutorials*, vol. 20, no. 2, pp. 836–869, 2018.
- [3] C. Seker, M. T. Gneser, and T. Ozturk, "A review of millimeter wave communication for 5G," in *2018 2nd Intl. Symposium on Multidisciplinary Studies and Innovative Tech. (ISMSIT)*, 2018, pp. 1–5.
- [4] Y. Wang, H. Lu, D. Zhao, and H. Sun, "Energy efficiency optimization in IRS-enhanced mmwave systems with lens antenna array," in *2020 IEEE Global Commun. Conference (GLOBECOM)*, 2020, pp. 1–6.
- [5] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, pp. 3313–3351, 2021.
- [6] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, 2019.
- [7] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, 2019.
- [8] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tutorials*, vol. 22, no. 4, pp. 2283–2314, 2020.
- [9] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Magazine*, vol. 53, no. 9, pp. 74–81, 2015.
- [10] Z. Ding and H. Vincent Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, 2020.
- [11] J. Xu, Y. Liu, X. Mu, and O. A. Dobre, "STAR-RISs: Simultaneous transmitting and reflecting reconfigurable intelligent surfaces," *IEEE Commun. Letters*, vol. 25, no. 9, pp. 3134–3138, 2021.
- [12] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2021.
- [13] 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, 2019.
- [14] 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, 2020.
- [15] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, 2019.
- [16] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Letters*, vol. 25, no. 1, pp. 234–238, 2021.
- [17] F. E. Bouanani, S. Muhaidat, P. C. Sofotasios, O. A. Dobre, and O. S. Badarneh, "Performance analysis of intelligent reflecting surface aided wireless networks with wireless power transfer," *IEEE Commun. Letters*, vol. 25, no. 3, pp. 793–797, 2021.
- [18] G. Zhou, C. Pan, H. Ren, P. Popovski, and A. L. Swindlehurst, "Channel estimation for RIS-aided multiuser millimeter-wave systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 1478–1492, 2022.
- [19] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications," in *IEEE Wireless Commun. and Networking Conference (WCNC)*, 2020, pp. 1–6.
- [20] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, 2021.
- [21] Z. Chen, J. Tang, H. Tang, X. Zhang, D. K. C. So, and K.-K. Wong, "Channel estimation of IRS-aided communication systems with hybrid multiobjective optimization," in *ICC 2021 - IEEE Intl. Conference on Commun.*, 2021, pp. 1–6.
- [22] S. Ma, W. Shen, X. Gao, and J. An, "Robust channel estimation for RIS-aided millimeter-wave system with RIS blockage," *IEEE Trans. Vehicular Tech.*, pp. 1–1, 2022.
- [23] X. Liu, Y. Liu, Y. Chen, and H. V. Poor, "RIS enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1057–1071, 2021.
- [24] Y. Liu, Z. Qin, M. El-kashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.
- [25] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Robust beamforming techniques for non-orthogonal multiple access systems with bounded channel uncertainties," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2033–2036, 2017.
- [26] S. Zhang, H. Zhang, B. Di, Y. Tan, Z. Han, and L. Song, "Beyond intelligent reflecting surfaces: Reflective-transmissive metasurface aided communications for full-dimensional coverage extension," *IEEE Trans. Vehicular Tech.*, vol. 69, no. 11, pp. 13 905–13 909, 2020.
- [27] C. Wu, Y. Liu, X. Mu, X. Gu, and O. A. Dobre, "Coverage characterization of STAR-RIS networks: NOMA and OMA," *IEEE Commun. Letters*, vol. 25, no. 9, pp. 3036–3040, 2021.
- [28] C. Wu, C. You, Y. Liu, X. Gu, and Y. Cai, "Channel estimation for STAR-RIS-aided wireless communication," *IEEE Commun. Letters*, vol. 26, no. 3, pp. 652–656, 2022.
- [29] J. XU, Y. Liu, X. Mu, T. Zhou, L. Song, V. Poor, and L. Hanzo, "Simultaneously transmitting and reflecting intelligent omni-surfaces: Modeling and implementation," *IEEE Vehicular Tech. Magazine*, pp. 2–10, 2022.
- [30] J. Zuo, Y. Liu, Z. Ding, and L. Song, "Simultaneously transmitting and reflecting (STAR) RIS assisted NOMA systems," in *2021 IEEE Global Commun. Conference (GLOBECOM)*, 2021, pp. 1–6.
- [31] V. Jamali, A. M. Tulino, G. Fischer, R. R. Miller, and R. Schober, "Intelligent surface-aided transmitter architectures for millimeter-wave ultra massive MIMO systems," *IEEE Open J. Commun. Society*, vol. 2, pp. 144–167, 2021.
- [32] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, 2014.
- [33] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "A new evaluation criterion for non-orthogonal multiple access in 5G software defined networks," *IEEE Access*, vol. 3, pp. 1633–1639, 2015.
- [34] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient joint user-RB association and power allocation for uplink hybrid NOMA-OMA," *IEEE IoT Journal*, vol. 6, no. 3, pp. 5119–5131, 2019.
- [35] D. Zhang, Q. Wu, M. Cui, G. Zhang, and D. Niyato, "Throughput maximization for IRS-assisted wireless powered hybrid NOMA and TDMA," *IEEE Wireless Commun. Letters*, vol. 10, no. 9, pp. 1944–1948, 2021.
- [36] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [37] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.
- [38] F. Alsaleem, J. S. Thompson, D. I. Laurenson, S. K. Podilchak, and C. A. Alistarh, "Small-size blockage measurements and modelling for mmwave communications systems," in *2020 IEEE 31st Ann. Intl. Symposium on Personal, Indoor and Mobile Radio Commun.*, 2020, pp. 1–6.