RIS-Assisted Grant-Free NOMA: User Pairing, RIS Assignment, and Phase Shift Alignment

Fatih Kilinc, *Graduate Student Member, IEEE*, Recep A. Tasci, *Graduate Student Member, IEEE*, Abdulkadir Celik, *Senior Member, IEEE*, Asmaa Abdallah, *Member, IEEE*, Ahmed M. Eltawil, *Senior Member, IEEE*, and Ertugrul Basar, *Fellow, IEEE*

***Abstract*—This paper introduces a reconfigurable intelligent surface (RIS)-assisted grant-free non-orthogonal multiple access (GF-NOMA) scheme. We propose a joint user equipment (UE) clustering and RIS assignment/alignment approach that jointly ensures the power reception disparity required by the power domain NOMA (PD-NOMA). The proposed approach maximizes the network sum rate by judiciously pairing UE with distinct channel gains and assigning RISs to proper clusters. To alleviate the computational complexity of the joint approach, we decouple UE clustering and RIS assignment/alignment subproblems, which reduces run times 80 times while attaining almost the same performance. Once the proposed approaches acknowledge UEs with the cluster index, UEs are allowed to access corresponding resource blocks (RBs) at any time requiring neither further grant acquisitions from the base station (BS) nor power control as all UEs are requested to transmit at the same power. In addition to passive RISs containing only passive elements and giving an 18% better performance, an active RIS structure that enhances the performance by 37% is also used to overcome the double path loss problem. The numerical results also investigate the impact of UE density, RIS deployment, RIS hardware specifications, and the fairness among the UEs in terms of bit-per-joule energy efficiency.**

***Index Terms*—Reconfigurable intelligent surface, power domain, grant-free, non-orthogonal multiple access, clustering, resource allocation.**

I. Introduction

NEXT-GENERATION communication networks are expected to meet the key performance requirements, which are ultra-reliable low latency communications (URRLC), enhanced mobile broadband (eMBB), and massive machine-type communications (mMTC). Although fifth-generation (5G) technology is a big step toward these goals, researchers continue to work on candidate technologies for beyond 5G networks to reach the expected limits for key performance metrics. For instance, reconfigurable intelligent surfaces (RISs) have emerged as a transformative technology envisioned for next-generation wireless networks to improve spectral efficiency (SE) with low power consumption and hardware costs [2]. RISs have low-cost and low-power hardware as they consist of a large number of passive elements. The RIS controller manipulates the reflection of incident electromagnetic waves towards the desired direction by intelligently controlling the phase shift of elements [2]. The passive beamforming nature of RISs has been shown to improve network capacity [2], [3], enhance the end-to-end performance of multi-hop communications [4], and deliver a performance comparable to traditional relaying schemes [5]. Furthermore, active RIS designs that include power amplifiers (PAs) are proposed to suppress the double path loss in passive RIS-assisted systems. In [6], a fully connected active RIS design is proposed where the gain of each amplifier can be controlled. In [7], the authors proposed a novel active RIS architecture with a single PA to boost energy efficiency.

Moreover, non-orthogonal multiple access (NOMA) has also been recognized as a promising candidate to enable enhanced mMTC by multiplexing several user equipments (UEs) on the same network resources [8], [9]. In particular, power domain (PD)-NOMA improves the SE by multiplexing UEs with different channel gain and transmit power such that successive interference cancellation (SIC) can be performed at the receiver. As the network size increases, PD-NOMA starts suffering from power control complexity, UE pairing and resource allocation overhead and ramifications of channel state information (CSI) acquisition [10]. NOMA has also been investigated with emerging technologies such as index modulation [11], cognitive radios [12], and unmanned aerial vehicles-aided wireless networks [13]. Recently, RISs have received attention thanks to their ability to create the reception power disparity required by PD-NOMA. While the RIS-aided NOMA is analyzed in [14], dynamic and static RIS configurations for multi-user NOMA schemes are considered in [15]. The authors of [16] have studied partitioning of RIS to facilitate PD-NOMA. Noting that all of the mentioned schemes are grant-based (GB), there is still a dire need to

Manuscript received 12 August 2022; revised 13 December 2022 and 26 April 2023; accepted 12 June 2023. Date of publication 21 June 2023; date of current version 6 October 2023. A conference version of this work has been submitted to IEEE ICC'23 [1]. The associate editor coordinating the review of this article and approving it for publication was K. Ota. *(Corresponding author: Fatih Kilinc.)*

Fatih Kilinc was with the Communications Research and Innovation Laboratory(CoreLab), Department of Electrical and Electronics Engineering, Koç University, 34450 Istanbul, Turkey. He is now with the Research and Development Department, ULAK Communications Inc., 06800 Ankara, Turkey (e-mail: fkilinc20@ku.edu.tr).

Recep A. Tasci and Ertugrul Basar are with the Communications Research and Innovation Laboratory, Department of Electrical and Electronics Engineering, Koç University, 34450 Istanbul, Turkey (e-mail: rtasci20@ku.edu.tr; ebasar@ku.edu.tr).

Abdulkadir Celik, Asmaa Abdallah, and Ahmed M. Eltawil are with the Computer, Electrical, and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia (e-mail: abdulkadir.celik@kaust.edu.sa; asmaa.abdallah@kaust.edu.sa; ahmed.eltawil@kaust.edu.sa).

Digital Object Identifier 10.1109/TCCN.2023.3288108

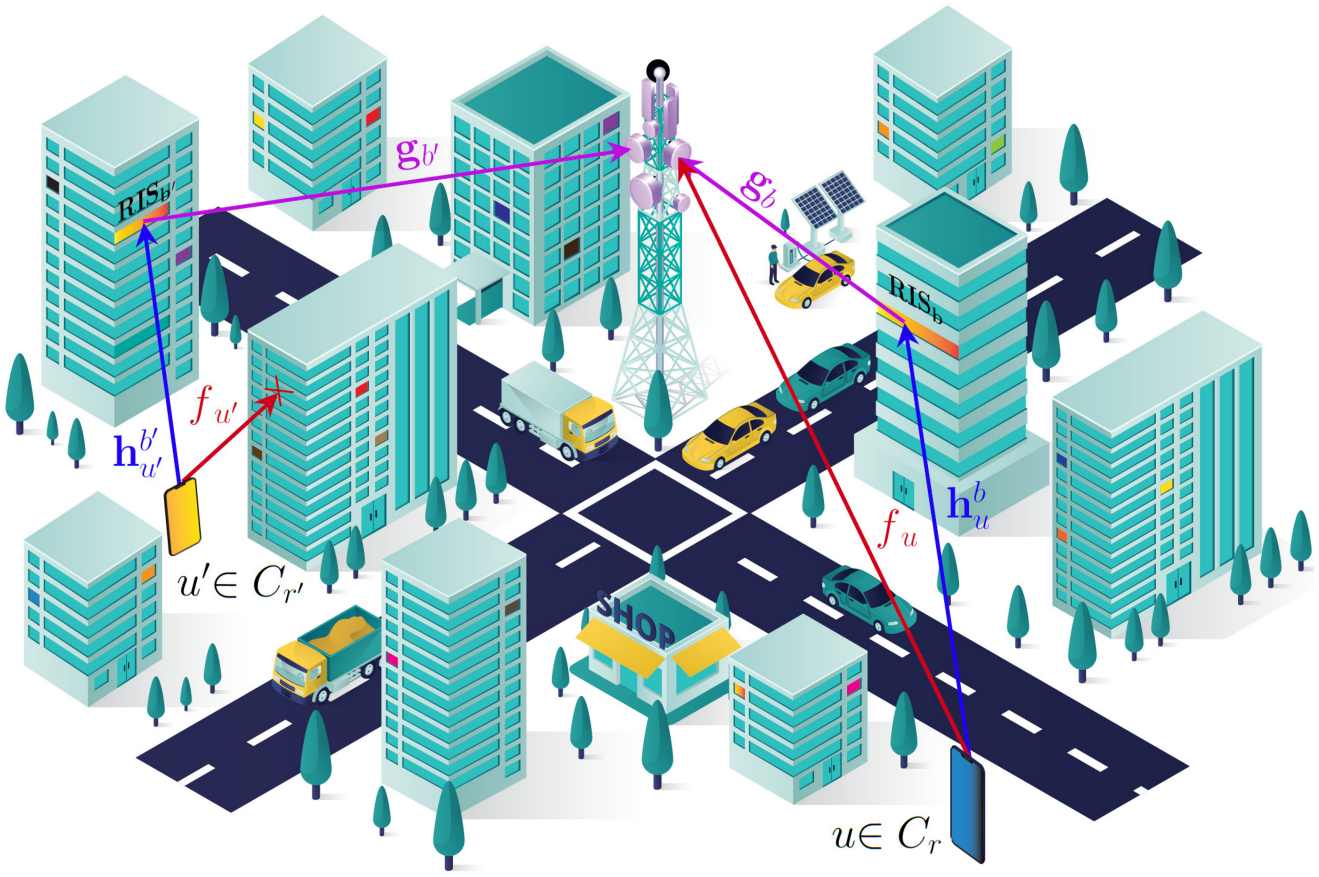

Fig. 1. Illustration of the RIS-aided GF-NOMA networks.

address the signaling overhead and computational complexity issues. For this reason, research on grant-free NOMA (GF-NOMA) schemes that UEs can operate without grant acquisition recently received much attention. In particular, power-domain GF-NOMA schemes aim to eliminate the need for power control and CSI acquisition at the transmitter side so that UEs can access allocated resource blocks (RBs) at any time without requiring grant acquisition. Therefore, this paper focuses on an RIS-aided GF-NOMA scheme where power reception disparity is ensured utilizing novel algorithms dealing with UE clustering, RIS assignment, and RIS phase shift alignment, considering a network consisting of a specific number of users and RISs distributed randomly over the cell coverage area as depicted in Fig. 1, where RIS blocks deployed on building facades/windows are highlighted in various colors.

A. Related Work

RIS-Assisted NOMA Schemes: RISs have received attention thanks to their ability to create the reception power disparity required by PD-NOMA [17]. The authors compare the RIS-assisted orthogonal multiple access (OMA) and NOMA technologies from different aspects and evaluate the performance of static and dynamic RISs in [15]. In [14], the authors design passive beamforming weights to boost the performance of the PD-NOMA and propose a priority-oriented NOMA scheme to increase SE. In [18], the authors aim to maximize the minimum signal-to-noise-plus-interference-ratio (SINR) threshold by deploying RISs in NOMA networks. Integration of the RIS technology to NOMA brings optimization problems such as joint optimal power allocation, RIS phase configuration, and UE allocation to obtain the best performance from the two technologies. To address these problems [19] uses a two-step machine learning approach for joint deployment, phase shift design, and power allocation. Also, [20] uses machine learning algorithms for UE partitioning and phase shifter network design. Reference [21] investigates the artificial intelligence empowered RIS-NOMA networks by using deep reinforcement learning. The existing literature also inspects RIS-NOMA schemes from the physical layer security perspective. In [22], authors propose novel RIS-assisted NOMA schemes for enhancing the physical layer security (PLS). The study in [23] proposes sub-optimal joint beamforming and power allocation scheme to increase the PLS as well. The RISs are also used for serving cell-edge UEs by integrating RIS technology into joint transmission-coordinated multipoint in a two-cell NOMA-based network in [24]. Using different RIS designs and partitioning RIS into sub-surfaces are also proposed to serve multiple UEs simultaneously to employ NOMA. The authors in [16] propose RIS partitioning to employ NOMA and maximize UE fairness. Furthermore, [25], [26] use simultaneously transmitting and receiving (STAR) RIS in NOMA networks and

evaluate from an error performance perspective. The hardware impairments of the RIS also considered in the literature, and the effects on the outage probability [27] and secrecy rate [28] is investigated.

GF-NOMA Schemes: GF access appears to be a savior technology for NOMA networks. GB-NOMA networks require the UEs in the same cluster to know the CSI of other members and align their transmission power accordingly. This process also causes a considerable signaling overhead and computational complexity on the UE side. Therefore scientists are looking for ways to exclude the computation on the UE side and enable them to access the base station (BS) without grant permission. Multiple access based, compute and forward-based, compressed sensing-based GF-NOMA schemes are presented, and future directions of GF-NOMA technology are given in [10]. The ALOHA-NOMA schemes are proposed in [29], [30] which provide high throughput and lower complexity. Semi-GF schemes where GB and GF UEs share the same spectrum are evaluated from the stochastic geometry perspective in [31], [32]. The authors analyze the compressing sensing-based GF-NOMA via stochastic geometry in [33]. A transmit power pool with the help of a deep reinforcement learning algorithm is created for UEs at certain distances sharing the same RB in [34]. In [35] a novel framework is proposed to avoid collusion in GF-NOMA networks when two UEs choose the same pilot sequence. Index modulation integrated GF-NOMA scheme for uplink (UL) transmission is presented in [36]. In [37], the author presents low complexity multi-level and single GF-NOMA schemes and an UE clustering approach exploiting the channel gains of the UEs.

RIS-Assisted GF-NOMA Schemes: Another promising approach is using RIS to facilitate power reception disparity without requiring any power control at the UE side. Indeed, a recent survey on GF-NOMA has also pointed out that RIS-aided GF-NOMA is a promising technology but has yet to be explored in-depth [10]. To the best of the authors' knowledge, the RIS-assisted GF-NOMA has been studied only in [38], where authors exploit deep reinforcement learning to solve sub-carrier assignment, power control, and RIS phase-shift alignment. Unlike the single RIS assumption in [38], we consider a network of users and RISs distributed randomly over the cell coverage area, which necessitates a joint solution of UE clustering, RIS assignment, and RIS phase shift alignment problems.

B. Main Contributions

The main contributions of this paper can be summarized as follows:

- To the best of the authors' knowledge, this paper is first to introduce an RIS-assisted fully GF-NOMA scheme where the required power disparity is obtained by three levels of implicit power control: clustering UEs with different channel gains into the same cluster, properly assigning RISs to UE clusters to increase power reception disparity, and aligning the phase shift matrix of RISs to reach the highest possible network sum rate. In this way, the proposed approaches eliminate the need for power control at the UE side by performing an implicit over-the-air power control by judiciously pairing UEs, assigning RISs, and aligning phase shifts of RISs.
- Since optimizing UE clustering and RIS assignment/alignment for the maximum network sum rate is a non-deterministic polynomial-time (NP)-Hard problem, we propose two solution methodologies. In the former, we develop an iterative approach where each iteration exploits a three-dimensional axial assignment (3D-AA) problem for joint UE clustering, RIS assignment, and phase shift alignment. Since 3D cost matrix generation complexity increases with the number of UEs and RISs, the latter method decouples UE clustering and RIS assignment/alignment subproblems, which yields 80 times faster completion time while reaching almost the same performance as the joint approach. The proposed schemes are designed to be applicable for both passive and active RIS architectures. The active RIS architecture considers a fully connected design where each element is connected to a PA, which introduces an extra power budget constraint on the optimization problem.
- The network-level simulations show that the proposed RIS-assisted GF-NOMA schemes outperform optimal GB PD-NOMA and regular GF-NOMA schemes. The active and passive RIS-assisted schemes provide %34 and %15 higher sum rate in a network where 150 UEs sharing 25 RBs, respectively. The network size simulation investigates the deployment scenarios, number of available RIS blocks and number of RIS elements. Even though RISs are known to provide better performance if they are deployed nearby of transmitter or receiver, this is not directly obvious for distributed RIS networks. Shall we deploy RIS closer to the BS or spread them out to get closer to the UEs? The numerical results show that deploying the RISs closer to the BS can yield up to %15 higher performance than a uniform distribution of the RISs across the cell area. The fairness among the UEs sharing the same RB is compared in terms of bit-per-joule energy efficiency by obtaining Jain's fairness index. While the proposed active RIS-assisted scheme provides the fairest scheme up to a 10 dB transmit power level, the passive RIS-assisted scheme achieves the same fairness as regular GF-NOMA schemes, and they both enhance the network capacity.

C. Notations and Paper Organization

Throughout the paper, sets and their cardinality are denoted with calligraphic and regular uppercase letters (e.g., $|\mathcal{A}| = A$), respectively. Vectors and matrices are represented in lowercase and uppercase boldfaces (e.g., $\mathbf{a}$ and $\mathbf{A}$), respectively. The i^{th} member of a vector and set is denoted by a_i and $\mathcal{A}\{i\}$, respectively. Likewise, matrix $\mathbf{A}$'s entry on i^{th} row and j^{th} column is denoted by $\mathbf{A}[i, j]$. Subscripts b, r and u are used for indexing the RIS blocks, RB/cluster, and UEs, respectively. For clarity, the definitions of the parameters and notations are stated in Table I.

TABLE I
TABLE OF NOTATIONS

Not.	Description
W	RB bandwidth [Hz]
$\mathcal{U}$	Set of $\lvert\mathcal{U}\rvert = U$ UEs, indexed by $1 \leq u \leq U$
$\mathcal{R}$	Set of $\lvert\mathcal{R}\rvert = R$ RBs, indexed by $1 \leq r \leq R$
p_{id}	Identical transmit power of UEs
D	Cell radius
C	#Clusters in the network
K	UE density dependent maximum cluster size, $\lceil U/R \rceil$
M	#RISs in the network
D_{in}	Inner radius of the RIS area
D_{out}	Outer radius of the RIS area
G	#RIS blocks at each RIS
N	#Reflecting Elements on each RIS block
$\mathcal{B}$	Set of $\lvert\mathcal{B}\rvert = B = MG$ RIS blocks, indexed by $1 \leq b \leq B$
RB_r	rth RB
$\mathcal{C}_r$	Cluster set of UEs utilizing RB_r, where $\lvert\mathcal{C}_r\rvert \leq K$
χ_r^u	Binary UE clustering indicator, 1 if $\text{UE}_u \in \mathcal{C}_r$, otherwise 0.
RIS_b	bth RIS block
$\boldsymbol{\Delta}$	Binary RIS assignment matrix, $\boldsymbol{\Delta} \in \{0,1\}^{B\times U\times R}$
$\delta_{r,b}^u$	$\boldsymbol{\Delta}$ entries, 1 if RIS_b is assigned to $\mathcal{C}_r$ and configured based on the channels of UE_u, 0 otherwise
$\mathbf{h}_u^b$	$\text{UE}_u - \text{RIS}_b$ channel
$\mathbf{g}_b$	$\text{RIS}_b -$Tx channel
f_u	$\text{UE}_u -$Tx channel
$y_{r_{\text{pas}}}$	Received signal at the BS over RB_r with passive RISs
$\boldsymbol{\Phi}_{k,l}^m$	Phase shift matrix of RIS_l assigned to $\text{UE}_m \in \mathcal{C}_k$
$\phi_{k,l}^{m,n}$	Phase shift of the nth element of RIS_l assigned to $\text{UE}_m \in \mathcal{C}_k$
n_{Rx}	Additive white Gaussian noise at the UEs
σ_{Rx}^2	The variance of n_{Rx}
N_0	Thermal noise power spectral density
$\widetilde{\mathcal{C}_r}$	Descending order sorted $\mathcal{C}_r$ acc. the received signal power
o_u	The order of UE_u in $\widetilde{\mathcal{C}_r}$
$\widetilde{\mathcal{C}_r}\{o_z\}_{\lvert o_z > o_u}$	UE set with lower received signal power than UE_u
$\gamma_{r_{\text{pas}}}^u$	SINR of $\text{UE}_u \in \widetilde{\mathcal{C}_r}$ with passive RISs
$y_{r_{\text{act}}}$	Received signal at the BS over RB_r with active RISs
$v_{k,l}^{m,n}$	Noise at the nth active element of RIS_l which is configured as per $\text{UE}_m \in \mathcal{C}_k$, distributed as $\sim \mathcal{CN}(0, \sigma_{\text{RIS}}^2)$
$\mathbf{G}_{k,l}^m$	RIS_l amplification factor matrix configured as per $\text{UE}_m \in \mathcal{C}_k$ with $\text{diag}\left(\left[\alpha_{k,l}^{m,1}, \ldots, \alpha_{k,l}^{m,n}, \ldots, \alpha_{k,l}^{m,N}\right]\right) \in \mathbb{C}^{N\times N}$
$\alpha_{k,l}^{m,n}$	Amplification factor of the nth active element of RIS_l which is configured as per $\text{UE}_m \in \mathcal{C}_k$, distributed as $\sim \mathcal{CN}(0, \sigma_{\text{RIS}}^2)$
$\overline{P_{\text{out},n}}$	nth active element output power without the limitation of $P_{\max}$
$P_{\text{cons},n}$	Consumed power for the nth active element
$\eta_{\max}$	Maximum efficiency of each active component
$P_{\max}^{\text{act}}$	Maximum output power of each active element
P_{tot}	Total power limitation for each active RIS
ε	Amplifier hardware depending parameter
$P_{\text{out},n}$	Optimum output power for the nth active element
$\overline{G_{\text{opt},n}}$	nth active element gain without the limitation of $P_{\max}$
$G_{\max}$	Maximum gain of the active elements
$G_{\text{opt},n}$	Optimum gain value for the nth active element
γ_r^u	SINR of $\text{UE}_u \in \mathcal{C}_r$, $\gamma_r^u = \{\gamma_{r_{\text{pas}}}^u, \gamma_{r_{\text{act}}}^u\}$
q_u	QoS requirement
$\boldsymbol{\mathcal{X}}$	Binary UE clustering matrix with entries χ_r^u
$\mathcal{A}_k$	Admission awaiting UEs set, where $\max(\lvert\mathcal{A}_k\rvert) = U - K$
$\mathcal{T}_r$	Temporary cluster, where $\max(\lvert\mathcal{T}_r\rvert) = K$

The rest of this paper can be summarized as follows: Section II presents the considered network and channel model, the proposed signal model and the RIS phase alignment process. Section III provides the problem formulation and briefly explains the proposed solution methodologies. Section IV introduces proposed UE clustering, RIS assignment, and phase shift alignment approaches. Finally, Section V presents numerical results and the benchmark schemes are presented, and Section VI concludes the paper with a few remarks.

II. System Model

In this section, we introduce the considered system model in four subsections. In Section II-A, the architecture of the network is demonstrated by addressing the UE distribution as well as RIS deployment and partitioning. In Section II-B, the channel model is exhibited in detail. The signal models and SINRs of the passive and active RIS-assisted schemes are given in Sections II-C and II-D, respectively.

A. Network Model

We consider UL operation of a network consisting of a single BS serving U UEs over R RBs with W [Hz] bandwidth, whose sets are denoted by $\mathcal{U}$ and $\mathcal{R}$, respectively. All UEs operate at an identical power, p_{id}, which is broadcast by the BS over a control channel. Without loss of generality, UEs are assumed to be uniformly distributed over a cell area of radius D. To show the proposed GF-NOMA schemes' suitability for massive connectivity, we consider a dense network scenario (i.e., $U \gg R$) and group/cluster multiple UEs to utilize a single RB. That is, the number of clusters C is the same with R, and RB/cluster terms are used interchangeably throughout the paper. Even though R and C are fixed, U can dynamically change according to spatio-temporal characteristics of the network traffic. Hence, the UE density dependent maximum cluster size is set as $K = \lceil \frac{U}{R} \rceil$. Accordingly, the cluster set utilizing RB_r is defined by $\mathcal{C}_r = \{\text{UE}_u \mid \chi_r^u = 1, \forall u \in \mathcal{U}, \sum_{u\in\mathcal{U}} \chi_r^u \leq K\}$, where $\lvert\mathcal{C}_r\rvert \leq K$ and χ_r^u is the binary UE clustering indicator, i.e., $\chi_r^u = 1$ if UE_u belongs to rth cluster, otherwise $\chi_r^u = 0$.

In order to improve network performance and facilitate the GF-NOMA, M RISs are uniformly deployed in a disk having inner and outer radius of D_{in} and D_{out}, $D_{\text{in}} \leq D_{\text{out}} \leq D$, where D_{in} ensures that RISs are placed at the far-field of the BS as shown in Fig. 2. Each physical RIS is partitioned into G logical blocks, each with N elements, resulting in an overall $B = MG$ RIS blocks. Each block can be exploited by a member of a cluster, and a cluster cannot be assigned to more than one block. Accordingly, the set of RIS blocks is denoted by $\mathcal{B}$, and the bth RIS block is denoted by RIS_b. The binary RIS assignment matrix is denoted by $\boldsymbol{\Delta} \in \{0,1\}^{B\times U\times R}$ with entries $\delta_{r,b}^u$ such that $\delta_{r,b}^u = 1$ if RIS_b is assigned to the rth cluster and configured based on the channels of $\text{UE}_u \in \mathcal{C}_r$, $\delta_{r,b}^u = 0$ otherwise. Fig. 1 demonstrates the RIS assignment and phase shift alignment for two different type of UEs. Even though $\text{UE}_u \in \mathcal{C}_r$ has a LoS link to the BS, RIS_b is aligned to its channel to further improve the sum rate of $\mathcal{C}_r$. On the other hand, $\text{RIS}_{b'}$ is aligned to provide coverage to $\text{UE}_{u'} \in \mathcal{C}_{r'}$, whose LoS link to the BS is blocked by a building.

B. Channel Model

The channels are modeled according to the Urban Macro (UMa) scenario of the 3GPP standard specified for 0.5-100 GHz bands [39]. The channel between UE_u and RIS_b is denoted by $\mathbf{h}_u^b = [h_{u,b}^1, \ldots, h_{u,b}^n, \ldots, h_{u,b}^N] \in \mathbb{C}^{N\times 1}$,

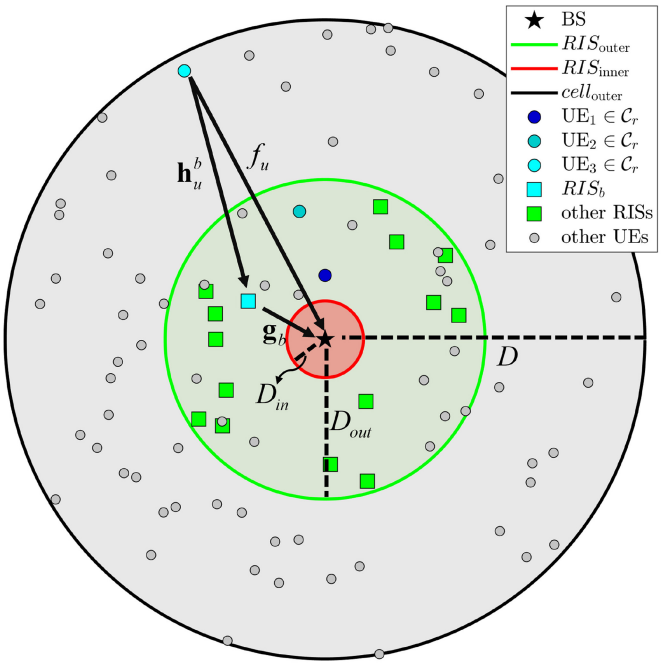

Fig. 2. Outer and inner cells of the network.

where $h_{u,b}^n$ is the channel coefficient from UE_u to nth reflecting element of the RIS_b, $n \in [1, N]$. Similarly, the channel vector between the BS and RIS_b is denoted by $\mathbf{g}_b = [g_b^1, \ldots, g_b^n, \ldots, g_b^N] \in \mathbb{C}^{N\times 1}$, where g_b^n is the channel coefficient between the BS and nth reflecting element of the RIS_b, $n \in [1, N]$. Moreover, the direct channel from UE_u to the BS is denoted by $f_u \in \mathbb{C}$. All these channels are assumed to experience Rayleigh and Rician fading under line-of-sight (LoS) and non-line-of-sight (NLoS) conditions, respectively. The channel coefficients are modeled as [40], [41]

$$h_{u,b}^n = \sqrt{\frac{1}{\lambda_h}}\left(\sqrt{\frac{K_h}{K_h+1}}h_{u,b}^{n,\text{LoS}} + \sqrt{\frac{1}{K_h+1}}h_{u,b}^{n,\text{NLoS}}\right), \quad (1)$$

$$g_b^n = \sqrt{\frac{1}{\lambda_g}}\left(\sqrt{\frac{K_g}{K_g+1}}g_b^{n,\text{LoS}} + \sqrt{\frac{1}{K_g+1}}g_b^{n,\text{NLoS}}\right), \quad (2)$$

$$f_u = \sqrt{\frac{1}{\lambda_f}}\left(\sqrt{\frac{K_f}{K_f+1}}f_u^{\text{LoS}} + \sqrt{\frac{1}{K_f+1}}f_u^{\text{NLoS}}\right), \quad (3)$$

where $K_h/K_g/K_f$ is the Rician factor for $\mathbf{h}_u^b/\mathbf{g}_b/f_u$; $\lambda_h/\lambda_g/\lambda_f$ is the path loss over $\mathbf{h}_u^b/\mathbf{g}_b/f_u$; $h_{u,b}^{n,\text{LoS}}/g_b^{n,\text{LoS}}/f_u^{\text{LoS}}$ is the LoS component; and $h_{u,b}^{n,\text{NLoS}}/g_b^{n,\text{NLoS}}/f_u^{\text{NLoS}} \sim \mathcal{CN}(0,1)$ is the NLoS component of $\mathbf{h}_u^b/\mathbf{g}_b/f_u$. If a channel does not include a LoS component, which mostly refers to Rayleigh fading, we consider the corresponding Rician factor as 0. λ_h, λ_g, and λ_f can be calculated as per LoS or NLoS links depending on the LoS probability given in [39, Table 7.2-2]. Additionally, the path loss models and the Rician factors are calculated and obtained according to the UMa scenario of the 3GPP standard [39, Table 7.2-1].

C. Passive RIS-Assisted NOMA System

Based on UE-RIS-BS cascaded channels and UE-BS channels defined in the previous subsection, the received complex baseband signal of the proposed system at the BS over RB_r when passive RISs are used can be expressed as

$$y_{r_{\text{pas}}}(\boldsymbol{\Delta}) = \sqrt{p_{\text{id}}}\sum_{\forall u\in\mathcal{C}_r}\sum_{\forall k,l,m}\left(\underbrace{\mathbf{g}_l^{\text{T}}\left(\boldsymbol{\Phi}_{k,l}^m\delta_{k,l}^m\right)\mathbf{h}_u^l s_u}_{\text{UEs-RIS-BS}} + \underbrace{f_u s_u}_{\text{UEs-BS}}\right) + n_{\text{Rx}}, \quad (4)$$

where $k \in \mathcal{R}$; $l \in \mathcal{B}$; $m \in \mathcal{U}$; $y_{r_{\text{pas}}}(\boldsymbol{\Delta})$ is the received signal from members of $\mathcal{C}_r$ at RB_r; s_u is the symbol transmitted by $\text{UE}_u \in \mathcal{C}_r$; $n_{\text{Rx}} \sim \mathcal{CN}(0, \sigma_{\text{Rx}}^2)$ is the additive white Gaussian noise with variance $\sigma_{\text{Rx}}^2 = N_0 W$; N_0 is the thermal noise power spectral density; $\boldsymbol{\Phi}_{k,l}^m = \text{diag}\left(\left[\phi_{k,l}^{m,1}, \ldots, \phi_{k,l}^{m,n}, \ldots, \phi_{k,l}^{m,N}\right]\right) \in \mathbb{C}^{N\times N}$ is the phase shift matrix of RIS_l assigned to $\text{UE}_m \in \mathcal{C}_k$; and $\phi_{k,l}^{m,n}$ is the phase shift of the nth element of the RIS_l configured as per $\text{UE}_m \in \mathcal{C}_k$. Assuming the CSI is acquired through accurate channel estimation methods [42], [43] and $\delta_{k,l}^m = 1$, the phase shifts of RIS_b are configured [41] as per the channel characteristics of $\text{UE}_m \in \mathcal{C}_k$

$$\phi_{k,l}^{m,n} = e^{j\angle f_m}\, e^{-j\angle g_l^n}\, e^{-j\angle h_{m,l}^n}, \quad (5)$$

where $\angle\cdot$ denotes the phase of complex numbers. Element phases cancel the overall channel phases incurred from UE-RIS-BS cascaded channel and UE-BS direct channel.

The SIC receiver at the BS iteratively decodes $y_{r_{\text{pas}}}(\boldsymbol{\Delta})$ in the descending order of received signal power. That is, the cluster member with the strongest/weakest reception power is decoded first/last such that the strongest/weakest cluster member observes all/no intra-cluster interference from other cluster members. Correspondingly, $\widetilde{\mathcal{C}_r}$ denotes the cluster set when the index set of the UEs is sorted in descending order according to the received signal power of the UEs. The order of UE_u in $\widetilde{\mathcal{C}_r}$ is denoted by o_u. Accordingly, following from (4), the SINR of $\text{UE}_u \in \widetilde{\mathcal{C}_r}$ is given by

$$\gamma_{r_{\text{pas}}}^u(\boldsymbol{\Delta}) = \frac{p_{\text{id}}\sum_{\forall k,l,m}\left(\left|\mathbf{g}_l^{\text{T}}\left(\boldsymbol{\Phi}_{k,l}^m\delta_{k,l}^m\right)\mathbf{h}_u^l\right|^2 + |f_u|^2\right)}{p_{\text{id}}\sum_{\substack{\forall z\in\widetilde{\mathcal{C}_r}\{o_z\}_{|o_z>o_u}\\ \forall k,l,m}}\left(\left|\mathbf{g}_l^{\text{T}}\left(\boldsymbol{\Phi}_{k,l}^m\delta_{k,l}^m\right)\mathbf{h}_z^l\right|^2 + |f_z|^2\right) + \sigma_{\text{Rx}}^2}, \quad (6)$$

where $\widetilde{\mathcal{C}_r}\{o_z\}_{|o_z>o_u}$ includes the set of UEs with lower received signal power than UE_u which are indexed from $o_z = o_u+1$ onwards. Even though the SINR expression in (6) accounts for $s_u, \forall u \in \mathcal{C}_r$ reflected from $\text{RIS}_b, \forall b \in \mathcal{B}$, regardless of their assignment, the non-coherent signals reflected from RIS blocks assigned to other clusters have a negligible impact on the received signal from $\mathcal{C}_r$, $y_r(\boldsymbol{\Delta})$. Accordingly, (6) can be simplified by merely considering the signal reflected from RIS_b assigned to $\mathcal{C}_r$ as follows:

$$\gamma^{u}_{r_{\text{pas}}}(\mathbf{\Delta}) = \frac{p_{\text{id}} \sum_{\forall m} \left(\left| \mathbf{g}_b^{\text{T}} \left(\mathbf{\Phi}^m_{r,b} \delta^m_{r,b} \right) \mathbf{h}^b_u \right|^2 + |f_u|^2 \right)}{p_{\text{id}} \sum_{\substack{\forall z \in \widetilde{\mathcal{C}}_r \{o_z\}_{|o_z > o_u}, \\ \forall m}} \left(\left| \mathbf{g}_b^{\text{T}} \left(\mathbf{\Phi}^m_{r,b} \delta^m_{r,b} \right) \mathbf{h}^b_z \right|^2 + |f_z|^2 \right) + \sigma^2_{\text{Rx}}}. \tag{7}$$

D. Active RIS-Assisted NOMA System

We further explore the utilization of active RIS elements to address the double-fading attenuation arising from the product of path losses on both links (i.e., BS-RIS and RIS-UE) and to enhance signal strength. Active RISs consist of a planar array of compact active elements capable of concurrently amplifying and reflecting incident signals with specific phase shifts, utilizing integrated amplifiers, phase shifters, and tunable filters [6], [44], [45]. The active components are responsible for signal amplification, while the matching network, through phase shifters, ensures the necessary phase shift and impedance matching to direct the amplified signal back toward the input port. Depending on the communication band of interest, tunable filters can be employed to select specific frequency bands using a range of analog or digital filtering technologies. The combination of these components leads to constructive interference between the reflected and incoming signals, ultimately generating an amplified signal within the desired frequency bands. Hence, the received complex baseband signal of the proposed system with active RIS is expressed as follows:

$$\begin{aligned} y_{r_{\text{act}}}(\mathbf{\Delta}) = \sqrt{p_{\text{id}}} \sum_{\forall u \in \mathcal{C}_r} \sum_{\forall k,l,m} & \times \left(\underbrace{\mathbf{g}_l^{\text{T}} \mathbf{G}^m_{k,l} \left(\mathbf{\Phi}^m_{k,l} \delta^m_{k,l} \right) \mathbf{h}^l_u s_u}_{\text{UEs-RIS-BS}} + \underbrace{f_u s_u}_{\text{UEs-BS}} \right) \\ & + \sum_{\forall k,l,m} \mathbf{g}_l^{\text{T}} \left(\mathbf{\Phi}^m_{k,l} \delta^m_{k,l} \right) \mathbf{v}^m_{k,l} + n_{\text{Rx}}, \end{aligned} \tag{8}$$

where $y_{r_{\text{act}}}(\mathbf{\Delta})$ is the received signal from members of $\mathcal{C}_r$ at RB_r; $\mathbf{v}^m_{k,l} = \left[v^{m,1}_{k,l}, \ldots, v^{m,n}_{k,l}, \ldots, v^{m,N}_{k,l} \right]^T \in \mathbb{C}^{N \times 1}$ and $v^{m,n}_{k,l} \sim \mathcal{CN}(0, \sigma^2_{\text{RIS}})$ is the additive white Gaussian noise at the active elements of the RISs with variance $\sigma^2_{\text{RIS}} = N_0 W$; and $\mathbf{G}^m_{k,l} = \text{diag}\left(\left[\alpha^{m,1}_{k,l}, \ldots, \alpha^{m,n}_{k,l}, \ldots, \alpha^{m,N}_{k,l} \right] \right) \in \mathbb{C}^{N \times N}$ holds for the amplification factor matrix of active RISs.

Similarly for the active RIS-assisted case, the SIC receiver at the BS iteratively decodes $y_{r_{\text{act}}}(\mathbf{\Delta})$ in the descending order of the received signal power. Hence, the SINR of o_uth ordered $\text{UE}_u \in \widetilde{\mathcal{C}_r}$ is given in (9) shown at the bottom of the page. For a similar reason as in the passive RIS-assisted scheme, (9) can also be simplified as given in (10), shown at the bottom of the page.

1) Power Limitation for the Active Components: In the proposed active RIS-assisted architecture, each reflecting element of RISs includes an active component, called fully-connected active RIS. According to [46], the power efficiency of these active elements can be modeled as

$$\frac{\overline{P_{\text{out},n}}}{P_{\text{cons},n}} = \eta_{\max} \left(\frac{\overline{P_{\text{out},n}}}{P^{\text{act}}_{\max}} \right)^{\varepsilon}, \tag{11}$$

where $P_{\text{cons},n}$, $\overline{P_{\text{out},n}}$, $\eta_{\max} \in (0, 1]$ and ε correspond to consumed power by the nth active element, the output power of the nth active component without the limitation of $P_{\max}$, the maximum efficiency of each active components, and a parameter depending on the amplifier hardware, respectively. The maximum output power for each active element is set to $P^{\text{act}}_{\max}$ to prevent it to operate in the non-linear region.

The proposed scheme has a power limitation for active RISs, so each can use a limited total power, P_{tot}, at most to amplify the incoming signal. Considering a fully-connected active RIS architecture and each RIS has N reflecting elements, there are also N active components for each RIS. The maximum power consumption for an ideal active element $P_{\text{cons},n}$ is obtained by reorganizing (11) assuming $\varepsilon = 0.5$ for more accurate modeling as follows:

$$P_{\text{cons},n} = \frac{1}{\eta_{\max}} \sqrt{\overline{P_{\text{out},n}} P^{\text{act}}_{\max}} = \frac{P_{\text{tot}}}{NG}, \tag{12}$$

where $\eta_{\max} \in (0, 1]$ is the maximum efficiency of the active components [46], [47]. Thus, $\overline{P_{\text{out},n}}$ can be extracted from (12) as

$$\overline{P_{\text{out},n}} = \left(\frac{P_{\text{tot}} \eta_{\max}}{NG} \right)^2 / P^{\text{act}}_{\max}. \tag{13}$$

Considering the maximum output power limitation of those active components, the optimum output power of the nth active component becomes

$$P_{\text{out},n} = \min\left(\overline{P_{\text{out},n}}, P^{\text{act}}_{\max} \right). \tag{14}$$

Moreover, the gain without the limitation of the maximum gain value that the nth active component can have is expressed as

$$\overline{G_{\text{opt},n}} = \frac{P_{\text{out},n}}{p_{\text{id}} |h^n_{u,b}|^2}. \tag{15}$$

$$\gamma^{u}_{r_{\text{act}}}(\mathbf{\Delta}) = \frac{p_{\text{id}} \sum_{\forall k,l,m} \left(|\mathbf{g}_l^{\text{T}} \mathbf{G}^m_{k,l} \left(\mathbf{\Phi}^m_{k,l} \delta^m_{k,l} \right) \mathbf{h}^l_u|^2 + |f_u|^2 \right)}{p_{\text{id}} \sum_{\substack{\forall z \in \widetilde{\mathcal{C}}_r \{o_z\}_{|o_z > o_u}, \\ \forall k,l,m}} \left(|\mathbf{g}_l^{\text{T}} \left(\mathbf{\Phi}^m_{k,l} \delta^m_{k,l} \right) \mathbf{h}^l_z|^2 + |f_z|^2 \right) + \sigma^2_{\text{RIS}} \sum_{\forall k,l,m} |\mathbf{g}_l^{\text{T}} \left(\mathbf{\Phi}^m_{k,l} \delta^m_{k,l} \right)|^2 + \sigma^2_{\text{Rx}}}. \tag{9}$$

$$\gamma^{u}_{r_{\text{act}}}(\mathbf{\Delta}) = \frac{p_{\text{id}} \sum_{\forall m} \left(|\mathbf{g}_b^{\text{T}} \mathbf{G}^m_{r,b} \left(\mathbf{\Phi}^m_{r,b} \delta^m_{r,b} \right) \mathbf{h}^b_u|^2 + |f_u|^2 \right)}{p_{\text{id}} \sum_{\substack{\forall z \in \widetilde{\mathcal{C}}_r \{o_z\}_{|o_z > o_u}, \\ \forall m}} \left(|\mathbf{g}_l^{\text{T}} \left(\mathbf{\Phi}^m_{r,b} \delta^m_{r,b} \right) \mathbf{h}^b_z|^2 + |f_z|^2 \right) + \sigma^2_{\text{RIS}} \sum_{\forall m} |\mathbf{g}_l^{\text{T}} \left(\mathbf{\Phi}^m_{r,b} \delta^m_{r,b} \right)|^2 + \sigma^2_{\text{Rx}}}. \tag{10}$$

Thus, the optimum gain value for the nth active component becomes

$$G_{\text{opt},n} = \min\left(\overline{G_{\text{opt},n}}, G_{\max}\right), \tag{16}$$

where $G_{\max}$ is the maximum gain of the active components. We note that $\alpha_{k,l}^{m,n}$s are calculated according to (16).

III. Problem Definition and Solution Methodology

In this section, we first provide joint UE clustering and RIS assignment formulation of the RIS-assisted GF-NOMA system, and then provide an overview of the solution methodology.

A. Problem Definition

The joint UE clustering and RIS assignment problem that maximizes the overall network sum rate can be formulated as follows:

$$\begin{aligned}
\mathbf{P_1}: \max_{\boldsymbol{\mathcal{X}},\boldsymbol{\Delta}} \; & \sum_{\forall r\in\mathcal{R}} \sum_{\forall u\in\mathcal{U}} W\log_2(1+\chi_r^u\gamma_r^u(\boldsymbol{\Delta})) \\
\text{s.t.} \; & \mathrm{C}_1: \gamma_r^u \geq 2^{q_u/B}-1, \forall u\in\mathcal{C}_r, \forall r\in\mathcal{R}, \\
& \mathrm{C}_2: \sum_{r\in\mathcal{R}} \chi_r^u = 1, \forall u\in\mathcal{U}, \\
& \mathrm{C}_3: \sum_{\forall u\in\mathcal{U}} \chi_r^u \leq K, \forall r\in\mathcal{R}, \\
& \mathrm{C}_4: \delta_{r,b}^u \leq \chi_r^u, \forall r\in\mathcal{R}, \forall b\in\mathcal{B}, \forall u\in\mathcal{U}, \\
& \mathrm{C}_5: \sum_{\forall u\in\mathcal{U}} \sum_{\forall b\in\mathcal{B}} \chi_r^u\delta_{r,b}^u \leq 1, \forall r\in\mathcal{R}, \\
& \mathrm{C}_6: \sum_{\forall u\in\mathcal{U}} \sum_{\forall r\in\mathcal{R}} \sum_{\forall b\in\mathcal{B}} \delta_{r,b}^u \leq B, \\
& \mathrm{C}_7: \chi_r^u\in\{0,1\}, \delta_{r,b}^u\in\{0,1\}, \forall r, \forall u, \forall b,
\end{aligned} \tag{17}$$

where γ_r^u is the SINR of $\mathrm{UE}_u \in \mathcal{C}_r$, $\boldsymbol{\mathcal{X}} \in \{0,1\}^{U\times R}$ is the binary UE clustering matrix with entries χ_r^u. In $\boldsymbol{P_1}$, C_1 ensures that UE_u is satisfied with the quality of service demand q_u [bps] $\forall u \in \mathcal{C}_r, \forall r \in \mathcal{R}$, C_2 assures each UE is admitted to a cluster, C_3 limits the cluster size by $K = \lceil U/R \rceil$, C_4 states that RIS_b cannot be configured as per UE_u if it is not a member of $\mathcal{C}_r$, C_5 guarantees each cluster is assisted by at most one RIS block, C_6 limits the total RIS assignments by the total number of RIS blocks, and C_7 specifies the domain and bounds on optimization variables. It is worth noting that $\mathbf{P_1}$ does not have any power control variable as the proposed approach performs implicit power control through user clustering and RIS assignment integer variables. $\mathbf{P_1}$ is a mixed-integer non-linear programming (MINLP) problem, which is non-convex due to the interference terms in the SINR expression defined in the next section. Finding an optimal solution to this NP-Hard problem is computationally prohibitive, even for moderate network size, and it necessitates a heuristic solution for real-life implementation, which is discussed next.

B. Solution Methodology

The proposed RIS-assisted GF-NOMA scheme allows UEs to access RBs at any time, requiring neither grant acquisition from the BS nor power control at the UE side. Instead, the power disparity requirement of PD-NOMA is satisfied by the UE clustering and the RIS assignment through a three-level implicit power control:

1) The first level of power control is obtained by pairing UEs with different channel gains on the same cluster/RB. Even if UEs operate at identical transmit power, the channel gain disparity introduces a reception power disparity to achieve a higher PD-NOMA gain.
2) The second level of power control is obtained by assigning a proper RIS block to a cluster. It is worth noting that RIS improves overall communication performance if it is deployed near the UE or the BS [4]. Therefore, RIS block assignment plays a crucial role in increasing the power reception disparity for an improved cluster sum rate.
3) The final level of power control is obtained by selecting the cluster member according to which assigned RIS is configured. Since reception power levels determine the SIC decoding order, the cluster member selection is also critical for improving the overall cluster sum rate.

Before delving into the details of the proposed UE clustering and RIS assignment solutions for the considered RIS-assisted GF-NOMA system, it is worth reminding that the optimal GB PD-NOMA scheme requires UEs with stronger channels to transmit at the higher powers to maximize the UL-NOMA cluster sum rate [37]. However, this approach is not fair in energy consumption and may result in reduced network lifetime, which is paramount for low-power mMTC. Alternatively, the proposed approaches improve network lifetime by requiring all UEs to transmit at identical power fairly and constitutes the required power disparity through iterative UE clustering and RIS assignments introduced in Section IV.

IV. UE Clustering and RIS Assignment Algorithms

In this section, we introduce the proposed UE clustering and RIS assignment schemes. Following cluster initialization, the proposed iterative approach admits a single UE to each cluster at each iteration until all UEs are admitted to a cluster. To this aim, the admissions are determined based on two different algorithms. Algorithm 1 jointly assigns UEs and RISs to clusters and executes a 3D-AA that yields the maximum network sum rate; therefore, it is called joint UE clustering and RIS assignment algorithm. Algorithm 2 first assigns UEs to clusters by using linear sum assignment (LSA). After that, it assigns RIS blocks to formed clusters by using LSA again. Since the second assignment first forms clusters and then makes RIS assignments, it follows a two-step process, unlike Algorithm 1, which is called consecutive UE clustering and RIS assignment algorithm.

A. Joint 3D UE Clustering and RIS Assignment Algorithm

Algorithm 1 follows an iterative joint 3D assignment approach. At each iteration, a 3D cost matrix is generated whose elements represent cluster rates for each combination of UE admission and RIS assignment status. UEs and RISs are then jointly assigned to clusters using a 3D-AA.

Algorithm 1 Joint UE Clustering and RIS Assignment

1: **Input:** $\mathcal{R}$, $\mathcal{U}$, $\mathcal{B}$, P
2: $\mathbf{h}_u^b/\mathbf{g}_b/f_u \leftarrow$ *Acquire CSI from SRS sent by* $UE_u, \forall u \in \mathcal{U}, \forall b \in \mathcal{B}$
3: $C \leftarrow R$ *// Determine number of clusters*
4: $K \leftarrow \left\lceil \frac{U}{R} \right\rceil$ *// Determine maximum cluster size*
5: $\tilde{\mathcal{U}} \leftarrow$ SORTDESCEND$(Rss, \forall u)$ *// UE ordering based on RSS of SRS signals*
6: $\mathcal{C}_r \leftarrow \tilde{\mathcal{U}}\{r\}, r \in [1, \ldots, R]$ *// Initialize clusters*
7: $\chi_r^u \leftarrow 1, u \in \mathcal{C}_r$ *//Initialize UE clustering matrix*
8: **for** $k = 1{:}K-1$ **do** *// Iterative UE admission and RIS Assignment starts*
9: $\mathcal{A}_k \leftarrow \tilde{\mathcal{U}} - \bigcup_{r=1}^{R} \mathcal{C}_r$ *// Initialize admission awaiting UEs*
10: $\boldsymbol{Q_k}, \boldsymbol{I_k} \leftarrow$ COST MATRIX$(\mathcal{C}_r, \mathcal{A}_k, \mathcal{B})$ *// Generating cost matrix*
11: $\boldsymbol{Y}_k \leftarrow$ 3D-AXIAL ASSIGNMENT$(\boldsymbol{Q_k}, \mathcal{A}_k)$
12: $\mathcal{C}_r \leftarrow \mathcal{C}_r \bigcup \mathcal{A}_k\{u\}, y_{r,b}^u = 1, \forall (r, u, b)$ *// Update clusters*
13: $\chi_r^u \leftarrow 1, u \in \mathcal{C}_r$ *// Update UE clustering matrix*
14: $\delta_{r,b}^{u'} \leftarrow 1$ iff $u' = \iota_{r,b}^u$, $u \in \mathcal{C}_r, \forall (r, b)$ *//Update RIS assignment matrix*
15: **end for**
16: **return** $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Delta}$

17: **procedure** COST MATRIX$(\mathcal{C}_r, \mathcal{A}_k, \mathcal{B})$
18: $\boldsymbol{Q}, \boldsymbol{I} \leftarrow \mathbf{0}^{R \times A \times B}$ *// Initialize* $\boldsymbol{Q}$ *and* $\boldsymbol{I}$ *to matrix of zeros*
19: **for** $r = 1{:}R$ **do**
20: **for** $b = 1{:}B$ **do**
21: **for** $u = 1{:}A_k$ **do**
22: $\mathcal{T}_r \leftarrow \mathcal{C}_r \bigcup \mathcal{A}_k\{u\}$ *// Temp. admit* u^{th} *UE of* $\mathcal{A}_k$ *to* $\mathcal{C}_r$
23: **for** $j \in T_r$ **do**
24: $\boldsymbol{\Phi}_{r,b}^j \leftarrow$ Align RIS_b to $\text{UE}_{\mathcal{T}_r\{j\}}$ as per (5)
25: $\gamma_r^j \leftarrow$ Obtain SINR as per (7)/(10) for pas./act. RIS
26: $\eta_j^\star \leftarrow \sum_{j \in \mathcal{T}_r} W \log_2(1 + \gamma_r^j)$
27: **end for**
28: $q_{r,b}^u \leftarrow \max(\eta_j^\star, j \in \mathcal{T}_r)$ *// Update cost matrix entries*
29: $\iota_{r,b}^u \leftarrow \text{argmax}(\eta_j^\star, j \in \mathcal{T}_r)$ *// Update* $\boldsymbol{I}$ *entries*
30: **end for**
31: **end for**
32: **end for**
33: **return** $\boldsymbol{Q}, \boldsymbol{I}$
34: **end procedure**

35: **procedure** 3D-AXIAL ASSIGNMENT$(\boldsymbol{Q}, \mathcal{A}_k)$
36: $\overset{\star}{\boldsymbol{Y}} \leftarrow \max_{\boldsymbol{Y}} \sum_{r \in \mathcal{R}} \sum_{b \in \mathcal{B}} \sum_{u \in \mathcal{U}} q_{r,b}^u y_{r,b}^u$
37: s.t. $\sum_{r \in \mathcal{R}} \sum_{b \in \mathcal{B}} y_{r,b}^u = 1, \forall u \in \mathcal{A}_k$
38: $\sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{U}} y_{r,b}^u = 1, \forall b \in \mathcal{B}$
39: $\sum_{b \in \mathcal{R}} \sum_{u \in \mathcal{U}} y_{r,b}^u = 1, \forall r \in \mathcal{R}$
40: **return** $\overset{\star}{\boldsymbol{Y}}$
41: **end procedure**

Algorithm 1 starts with receiving sounding reference signals (SRS), which is a Zadoff-Chu sequence transmitted by each UE separately from the Physical UL shared channel (PUSCH) and physical UL control channel (PUCCH). UEs can transmit SRS on any subcarriers in the last symbol of an UL subframe regardless of subcarriers assignments. We especially focus on time-division duplexing (TDD) mode for the sake of channel reciprocity such that SRS may also be sent in the last two symbols of the special subframe if UL pilot time slot (UpPTS) is configured to be in the long format. The SRS is imperative to estimate cascaded and direct channels [42] to determine RIS assignment and configure phase shift matrices accordingly. Notice that the CSI acquisition is an inherent part of any communication system and not specific to the proposed approach. After the cluster size is determined based on the available number of RBs in Line 3, the maximum cluster size is determined in Line 4 based on the UE density. For the sake of cluster initialization, the UEs are sorted in the descending order of received signal strength of SRS signals in Line 5. Then, the clusters are initialized by admitting the first R UEs into C clusters in Line 6. Then, the iterative UE admission and RIS assignment starts in Line 8, in each iteration one more UE is admitted to each cluster. Therefore, the first line of the for loop, Line 9, updates the set of UEs awaiting cluster admission, which is denoted by $\mathcal{A}_k$.

To execute 3D-AA, Line 10 calls cost matrix generation procedure given between Line 17 and Line 34, where cost matrix $\boldsymbol{Q} \in \mathbb{R}^{R \times A \times B}$ and temporary RIS alignment matrix $\boldsymbol{I} \in \mathbb{N}^{R \times A \times B}$ are formed over three nested for loops. In Line 22, the u-th element of admission awaiting UE set is temporarily admitted to $\mathcal{C}_r$. Then, the most inner loop between Line 23 and Line 27 determines the cluster member giving the maximum cluster sum rate if RIS_b is aligned as per its channel conditions. To this aim, Line 24 adjusts the phase shift matrix of RIS_b to $\text{UE}_{\mathcal{T}_r\{j\}}$, Line 25 calculates the cluster members' SINR based on adjusted the phase shift matrix as per (7)/(10) for passive/active RIS, and finally Line 26 records the cluster sum rate of $\mathcal{T}_r$ if RIS_b is aligned to its j-th member, $\eta_j^\star$. Accordingly, the cost matrix element $q_{r,u}^b$ is updated in Line 28 based on the UE cluster member giving the highest possible cluster rate if the phase shift matrix of RIS_b is aligned to itself, whose index matrix (representing the UE index) is also stored in a temporary RIS alignment matrix in Line 29.

Following the generation of the cost-matrix, Line 11 calls 3D-AXIAL ASSIGNMENT to find joint UE clustering and RIS assignment at k-th iteration. For instance, when 3D-AA returns $y_{3,4}^5 = 1$, it means that: 1) $\text{UE}_{\mathcal{A}_k\{5\}}$ is admitted to $\mathcal{C}_3$, 2) RIS_4 is assigned to $\mathcal{C}_3$, and 3) phase shift matrix of RIS_4 is aligned to assisted UE whose UE index stored in $u' = \iota_{3,4}^5$. Thereafter, the cluster members are updated as per the outcome of 3D-AA in Line 12, which is followed by update of UE clustering matrix $\boldsymbol{\mathcal{X}}$ and RIS assignment matrix $\boldsymbol{\Delta}$ in Line 13 and 14, respectively. The final UE clustering and RIS assignment matrices are returned at the end of $(K-1)$-th iteration. Notice in Line 14 that the phase alignment is obtained from the RIS alignment matrix $\boldsymbol{I}$ returned by the COST MATRIX procedure.

The time complexity of the proposed solution is mainly driven by COST MATRIX and 3D-AXIAL ASSIGNMENT procedures. The complexity of cost matrix computation is $\mathcal{O}(\sum_{k=1}^{K-1} RA_kB)$. Moreover, the 3D-AA is known to be an NP-Hard problem, whose matching theory-based approximate solutions can obtain results for a square cost matrix with complexity $\mathcal{O}(3MV^3)$ where M is the number of relaxations and V is the largest dimension of the cost matrix [48]. Hence, after $K-1$ iterations, the time complexity of the considered 3D-AA becomes $\mathcal{O}(3M \sum_{k=1}^{K-1} (V_k)^3)$, where $V_k = \max(R, A_k, B)$. Since the latter is more dominant, the overall complexity can be approximated as $\sim \mathcal{O}(3M \sum_{k=1}^{K-1} RA_kB + (V_k)^3)$.

B. Consecutive 2D UE Clustering and RIS Assignment Algorithm

Algorithm 2 proposes a faster and consecutive UE clustering and RIS assignment algorithm that achieves almost the same performance as Algorithm 1. Algorithm 2 mainly follows a similar iterative clustering approach but instead of jointly

Algorithm 2 Consecutive UE Clustering and RIS Assignment

1: **Input:** $\mathcal{R}, \mathcal{U}, \mathcal{B}$, P
2: $\mathbf{h}_u^b/\mathbf{g}_b/f_u \leftarrow$ *Acquire CSI from SRS sent by* $UE_u, \forall u \in \mathcal{U}, \forall b \in \mathcal{B}$
3: $C \leftarrow R$ *// Determine number of clusters*
4: $K \leftarrow \left\lceil \frac{U}{R} \right\rceil$ *// Determine maximum cluster size*
5: $\tilde{\mathcal{U}} \leftarrow$ SORTDESCEND$(Rss, \forall u)$ *// UE ordering based on RSS of SRS signals*
6: $\mathcal{C}_r \leftarrow \tilde{\mathcal{U}}\{r\}, r \in [1, \ldots, R]$ *// Initialize clusters*
7: $\chi_r^u \leftarrow 1, u \in \mathcal{C}_r$ *//Initialize UE clustering matrix*
8: **for** $k = 1{:}K-1$ **do** *// Iterative UE admission and RIS Assignment starts*
9: $\mathcal{A}_k \leftarrow \tilde{\mathcal{U}} - \bigcup_{r=1}^{R} \mathcal{C}_r$ *// Initialize admission awaiting UEs*
10: $\boldsymbol{Q_k} \leftarrow$ COST MATRIX UE$(\mathcal{C}_r, \mathcal{A}_k)$ *// CM for user clustering*
11: $\boldsymbol{Y}_k \leftarrow$ 2D-AXIAL ASSIGNMENT$(\boldsymbol{Q_k})$
12: $\mathcal{C}_r \leftarrow \mathcal{C}_r \bigcup \mathcal{A}_k\{u\}, y_r^u = 1, \forall (r, u)$ *// Update clusters*
13: $\chi_r^u \leftarrow 1, u \in \mathcal{C}_r$ *// Update UE clustering matrix*
14: **end for**
15: **return** $\boldsymbol{\mathcal{X}}$
16: $\boldsymbol{Q_m}, \boldsymbol{I_m} \leftarrow$ COST MATRIX RIS$(\mathcal{C}_r, \mathcal{B})$ *// CM for RIS assignment*
17: $\boldsymbol{Y}_m \leftarrow$ 2D-AXIAL ASSIGNMENT$(\boldsymbol{Q_m})$
18: $\delta_{r,b}^{u'} \leftarrow 1$ iff $u' = \iota_r^b, y_r^b = 1, u \in \mathcal{C}_r, \forall (r, b)$ *//Update RIS assignment matrix*
19: **return** $\boldsymbol{\Delta}$

20: **procedure** COST MATRIX UE$(\mathcal{C}_r, \mathcal{A}_k)$
21: $\boldsymbol{Q}, \boldsymbol{I} \leftarrow \mathbf{0}^{R \times A}$ *// Initialize* $\boldsymbol{Q}$ *and* $\boldsymbol{I}$ *to matrix of zeros*
22: **for** $r = 1{:}R$ **do**
23: **for** $u = 1{:}A_k$ **do**
24: $\mathcal{T}_r \leftarrow \mathcal{C}_r \bigcup \mathcal{A}_k\{u\}$ *// Temp. admit* u^{th} *UE of* $\mathcal{A}_k$ *to* $\mathcal{C}_r$
25: $\gamma_r^j \leftarrow$ Obtain SINR of $\text{UE}_j, \forall j \in \mathcal{T}_r$
26: $q_r^u \leftarrow \sum_{j \in \mathcal{T}_r} W \log_2(1 + \gamma_r^j)$
27: **end for**
28: **end for**
29: **return** $\boldsymbol{Q}$
30:
31: **end procedure**

32: **procedure** COST MATRIX RIS$(\mathcal{C}_r, \mathcal{B})$
33: $\boldsymbol{Q}, \boldsymbol{I} \leftarrow \mathbf{0}^{R \times B}$ *// Initialize* $\boldsymbol{Q}$ *and* $\boldsymbol{I}$ *to matrix of zeros*
34: **for** $r = 1{:}R$ **do**
35: **for** $b = 1{:}B$ **do**
36: **for** $j \in C_r$ **do**
37: $\boldsymbol{\Phi}_{r,b}^j \leftarrow$ Align RIS_b to $\text{UE}_{\mathcal{C}_r\{j\}}$ as per (5)
38: $\gamma_r^j \leftarrow$ Obtain SINR as per (7)/(10) for pas./act. RIS
39: $\eta_j^\star \leftarrow \sum_{j \in \mathcal{T}_r} W \log_2(1 + \gamma_r^j)$
40: **end for**
41: $q_r^b \leftarrow \max(\eta_j^\star, j \in \mathcal{C}_r)$ *// Update cost matrix entries*
42: $\iota_r^b \leftarrow \text{argmax}(\eta_j^\star, j \in \mathcal{C}_r)$ *// Update* $\boldsymbol{I}$ *entries*
43: **end for**
44: **end for**
45: **return** $\boldsymbol{Q}, \boldsymbol{I}$
46: **end procedure**

47: **procedure** 2D-AXIAL ASSIGNMENT$(\boldsymbol{Q})$
48: $\overset{\star}{\boldsymbol{Y}} \leftarrow \max_{\boldsymbol{Y}} \sum_{r \in \mathcal{R}} \sum_{z \in \mathcal{Z}} q_r^z y_r^z$
49: s.t. $\sum_{r \in \mathcal{R}} y_r^z = 1, \forall z \in \mathcal{Z}$
50: $\sum_{z \in \mathcal{Z}} y_r^z = 1, \forall r \in \mathcal{R}$
51: $z \in \{u, b\}, \mathcal{Z} \in \{\mathcal{U}, \mathcal{B}\}$
52: **return** $\overset{\star}{\boldsymbol{Y}}$
53: **end procedure**

assigning UEs and RIS blocks to cluster sets, it first assigns UEs to the clusters and then assigns RISs to previously formed clusters. Lines 2-6 performs setting obtaining CSI information, determining number of clusters and maximum cluster size, initializing clusters respectively, as in Algorithm 1.

The iterative UE admission starts in Line 8, in each iteration one more UE is admitted to each cluster. Therefore, the first line of the for loop, Line 9, updates the set of UEs awaiting cluster admission, which is denoted by $\mathcal{A}_k$. The iterative algorithm now only admits UEs to the clusters and starts in Line 8. In line 9 the set $\mathcal{A}_k$ is updated that denotes the admission awaiting UEs. In each iteration one more UE is admitted to each cluster from the set $\mathcal{A}_k$. Line 10 calls the cost matrix function for UE admission. The 2D Cost matrix $\boldsymbol{Q}_k \in \mathbb{R}^{R \times A}$ is generated between the Lines 20-31 over two nested loops. Line 24 admits the uth UE from the set $\mathcal{A}_k$ to $\mathcal{C}_r$ and forms a temporary cluster denoted by $\mathcal{T}_r$. The UE SINRs are calculated in Line 25 and the temporary cluster rate is calculated by summing the rates of UEs in $\mathcal{T}_r$. The cost matrix element q_r^u in Line 26 stands for the cluster rate if uth UE from the set $\mathcal{A}_k$ is assigned to C_r. Notice that the UE assignment merely depends on channel gain disparity among the UEs and does not consider RIS assistance and phase shift alignment.

The UE admission is performed according to LSA executed in Line 11 to find the assignment maximizing the network sum rate by using the generated cost matrix at kth iteration. For instance, when LSA returns $y_3^5 = 1 \in \boldsymbol{Y}_k$, it means that $\text{UE}_{\mathcal{A}_k\{5\}}$ is admitted to $\mathcal{C}_3$. In the final step, UE clustering matrix $\boldsymbol{\mathcal{X}}$ is updated. After finalizing the UE admission and forming clusters, RIS blocks needs to be assigned to generated clusters and assist specific UEs in each cluster to maximize the cluster rate. Line 16 calls the cost matrix function for RIS assignment and phase alignment purposes. The cost matrix $\boldsymbol{Q}_m \in \mathbb{R}^{R \times B}$ and a temporary RIS alignment matrix $\boldsymbol{I}_m \in \mathbb{N}^{R \times B}$ are formed over three nested for loops in procedure explained between lines 32-46. The first two loops in lines 34 and 35 assign $\text{RIS}_b \in \mathcal{B}$ to $\mathcal{C}_r$. The most inner loop finds the UE that maximizes the cluster sum rate when the corresponding RIS block's phase shifts are aligned to itself. Line 37 aligns the phase of the RIS_b to UE $j \in \mathcal{C}_r$. Following SINRs are calculated by considering the phase configuration of the RIS block in Line 38. Line 39 records the cluster sum rate if RIS_b is assigned to UE j in $\mathcal{C}_r$. In consequence, the cost matrix element q_r^u in Line 42 is obtained by considering the cluster member yielding the largest cluster rate when assisted with RIS_b and its UE index is stored in ι_r^b. Finally, LSA procedure is called again in Line 17 to find the RIS assignment and alignment at once. For instance, if LSA returns $y_2^4 = 1 \in \boldsymbol{Y}_k$ for user clustering, it means that RIS_4 is assigned to $\mathcal{C}_2$. Likewise, if LSA returns $y_2^4 = 1 \in \boldsymbol{Y}_m$ for RIS assignment/alignment, it means that the phase configuration of the RIS_4 is aligned for $\text{UE}_{u'}$ whose UE index is stored in ι_2^4. Thereafter, the elements of the RIS assignment matrix $\delta_{r,b}^{u'}$ is updated in Line 18 and returned the final assignment matrix $\boldsymbol{\Delta}$ in line 19.

The time complexity of the consecutive approach is mainly driven by COST MATRIX and LSA procedures. The complexity of cost matrix computation for UE clustering and RIS assignment are given by $\mathcal{O}(\sum_{k=1}^{K-1} RA_k)$ and $\mathcal{O}(RB)$, respectively. Considering the well known cubic complexity of LSA solutions, the overall complexity can be obtained as $\mathcal{O}(\sum_{k=1}^{K-1} RA_k + (\max(R, A_k))^3)$ and $\mathcal{O}(RB + (\max(R, B))^3)$, which constitutes the entire complexity of Algorithm 2 together.

TABLE II
TABLE OF THE SIMULATION PARAMETERS

Param.	Value	Param.	Value
U	75	N	256
R	25	q_u	$0.1Mbps$
M	25	f_c	5 GHz
G	1	p_{id}	21 dBm
C	25	$P_{\max}$	23 dBm
N_0	-174 dBm	$P_{\min}$	-40 dBm
D	250 m	$P_{\max}^{\text{act}}$	21 dBm
D_{in}	15 m	P_{tot}	1 W
D_{out}	50 m	$G_{\max}$	20 dB
W	180 kHz		

V. NUMERICAL RESULTS

In this section, we provide numerical results for the proposed schemes and compare their performances with three benchmark schemes for different system parameters. Achievable rates and fairness indexes for different schemes are presented by computer simulations. Although we consider the SINR expressions in (7)/(10) for passive/active RIS assignment in both algorithms, we consider the general SINR expression in (6)/(9) while obtaining the numerical results, so that the reflections from non-coherent RIS blocks are considered as well. We consider the default system parameters summarized in Table II, unless stated explicitly otherwise.

A. Benchmark Schemes

The proposed method significantly differs from regular power domain NOMA by eliminating the need for uplink power control through user clustering and RIS assistance. As a result, the natural comparison would be with traditional PD-NOMA approaches with various power control capabilities but without RIS assistance. In this way, the comparison will reveal the potential of RIS assistance on improving system performance and facilitating the power reception disparity required by PD-NOMA without the need of complex power control operations. To this aim, we have considered three main benchmark schemes [37]:

1) Grant based OPT PD-NOMA: This is the optimal power control scheme where UEs compute optimal transmit power levels based on the readily available CSI, and the maximum and minimum transmit power of UEs are $P_{\max}$ and $P_{\min}$, respectively. Throughout the simulations, optimal power weights are obtained by geometric programming solver of the CVX disciplinized convex optimization toolbox [49].
2) Multi-level MGF-NOMA: This is a multi-level fixed power allocation approach, where it clusters the set of UEs into $R = C$ groups based on their channel gains and divides power control range ($[p_{max} = 23, p_{min} - 40]$ dBm) into $R = C$ levels. MGF-NOMA requires UE partitions with higher channel gain to transmit at higher power levels. For instance, the UE partition consisting of UEs with the highest/lowest channel gains are required to transmit at the highest/lowest power levels.

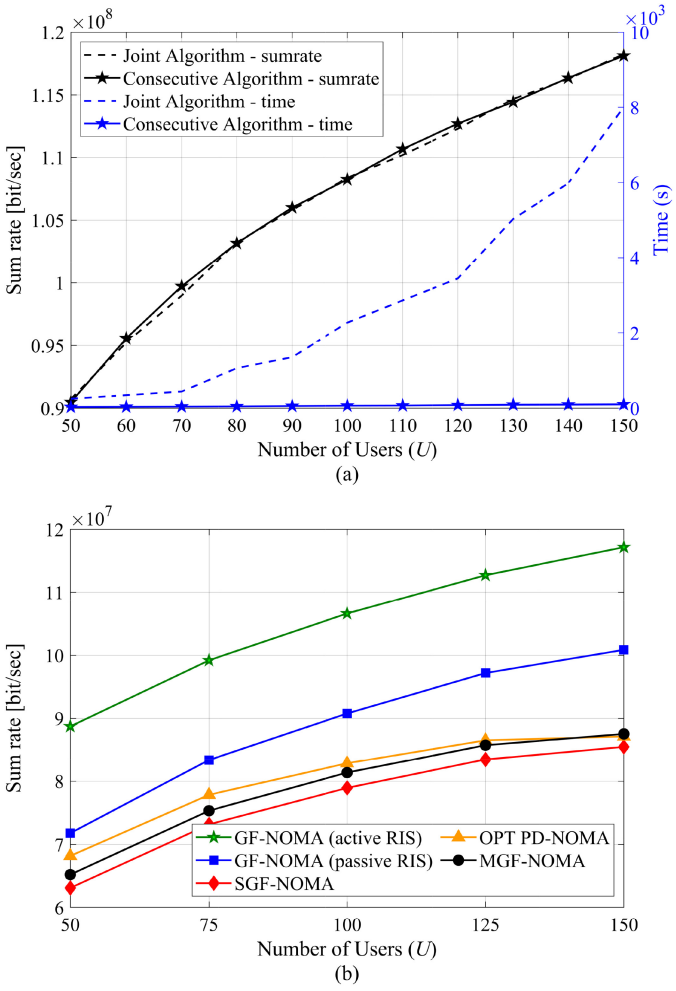

Fig. 3. (a) Sum rate and complexity of the proposed algorithms, and (b) sum rate of the proposed schemes.

3) Single-level SGF-NOMA: This scheme is similar to the proposed scheme since it does not apply power control allocations in which all UEs transmit at an identical power, p_{id}. However, this scheme do not leverage from the RIS assistance. Hence for comparison purposes, this scheme is used to show the potential of our proposed RIS-aided GF-NOMA systems versus the GF-NOMA systems with no RIS assistance.

We refer interested readers to [37] for a more detailed explanation of the benchmark schemes. For the sake of a fair comparison, the benchmark schemes follow the clustering approach presented in Algorithm 2. Additionally, for the same reason, the proposed schemes uses $p_{\text{id}} = 21$ dBm for all UEs, which is the average transmit power of UEs for OPT PD-NOMA. Noting that benchmark schemes do not benefit from RIS, the main difference between compared schemes is the underlying power control approach.

B. Impact of UE Density on Sumrate and Time Complexity

The number of UEs in the network is a crucial metric affecting both network performance and the complexity of the NOMA schemes. Fig. 3 shows run time complexity and

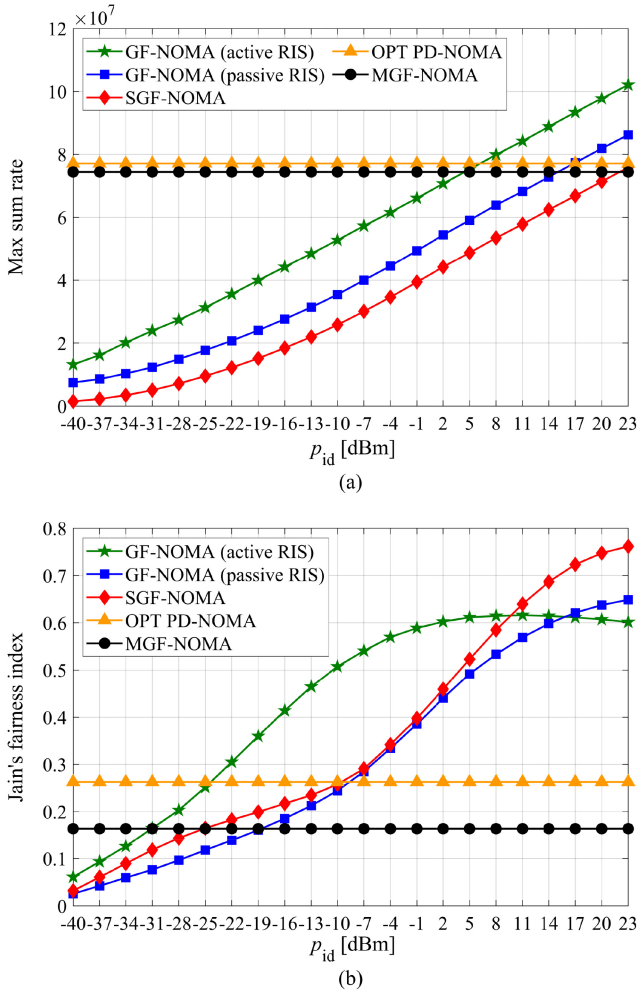

Fig. 4. Sum rate and Jain's fairness index for varying p_{id} for the proposed and benchmark schemes.

network achievable network sum rate for varying U. As seen in Fig. 3(a), the consecutive algorithm returns result up to 80 times faster than the joint Algorithm 1 while achieving almost the same network sum rate performance as shown in 3(b). Such a great run time complexity difference is mainly due to the time spent generating the 3D cost matrix for Algorithm 1, which increases as U increases. Since the consecutive approach provides almost the same performance with much-reduced running times, the rest of the simulations are carried out with Algorithm 2. It is important to note that both algorithms are guaranteed to converge as they do not include any while loops and only consist of nested for loops with fixed number of iteration.

As seen in Fig. 3(b), the sum rate of the proposed scheme increases with U because the number of UEs operating in a single RB increases, resulting in more efficient spectrum usage. The proposed scheme with passive RISs performs up to 15%, 15%, and 18% better than OPT PD-NOMA, MGF-NOMA, and SGF-NOMA schemes when $U = 150$, respectively. When active RIS architecture is employed, the proposed scheme performs up to 34%, 34%, and 37% better than OPT PD-NOMA, MGF-NOMA, and SGF-NOMA, respectively. Thus, active RISs provide 19% better performance than passive RISs. We note that the increase in the cluster size does not negatively affect the network performance. The proposed scheme performs better on denser networks, where $U = 150$ and the number of UEs per cluster is 6 and outperforms the benchmark schemes. The performance of the benchmark schemes starts to saturate after $U = 150$ while the proposed scheme continues to show an upwards trend.

C. The Impact of Identical Tx Power on Fairness

The network sum rate for varying identical transmit power levels, p_{id}, is presented in Fig. 4(a). As the p_{id} varies for SGF-NOMA and active/passive RIS-assisted GF-NOMA, transmit power levels are fixed in MGF-NOMA, and the average transmit power between UEs is obtained as 21 dBm for adaptively regulated OPT PD-NOMA via Monte Carlo simulations. While the SGF-NOMA achieves the same performance as MGF-NOMA and OPT GF-NOMA at the maximum transmit power level, the RIS-assisted schemes require less power at the UE side, resulting in more energy-efficient transmission. The passive RIS-assisted scheme requires 4 dB, and the active RIS-assisted scheme requires 16 dB less transmit power to achieve the same performance as OPT PD-NOMA and MGF-NOMA. The evaluation of the fairness in terms of the energy efficiencies of the UEs sharing the same cluster is presented in Fig. 4(b). The Jain's fairness index is obtained for each scheme as follows:

$$\mathcal{J} = \frac{\left(\sum_{\forall u \in \mathcal{C}_r} \eta_u\right)^2}{K\left(\sum_{\forall u \in \mathcal{C}_r} \eta_u\right)} \tag{18}$$

where $\eta_u = R_u / P_{\mathrm{t}}^u$, R_u, P_{t}^u are the bit per joule energy efficiency, the achievable rate, and transmit power of the UE$_u$, respectively. Since UEs transmit at identical power levels, the SGF-NOMA and RIS-assisted GF-NOMA schemes present higher fairness. While the SGF-NOMA and passive RIS-assisted GF-NOMA perform close fairness results, the active RIS-assisted scheme is the fairest scheme up to 8 dBm transmit power. Assisting low-rate UEs with active RIS yields a more fair energy usage. However, as the transmit power of the UEs increases, achievable rates of the assisted UEs become dominant, resulting in poorer fairness compared to the SGF-NOMA. Looking at Fig. 4, it can be concluded that RIS-assisted GF-NOMA schemes maintain fairness among UEs while providing enhanced capacity and sometimes even more fair structure.

Network sum rate for varying D_{out}, R, and N are presented in Fig. 5(a), 5(b), and 5(c), respectively. As seen in Fig. 5(a), the locations of the RISs substantially affect the system performance. It is a well-known fact that a basic RIS-assisted system exhibits its maximum performance when the RIS is located near the system's terminals. Similarly, the proposed design performs best when the RISs are close to the BS or UEs. The results show that deploying RISs near the BS performs better than the uniform deployment of RIS across the cell area. This is mainly because we cannot guarantee that the RISs will be placed close to the UEs since UEs are also randomly distributed. The proposed scheme performs 8%, 12%,

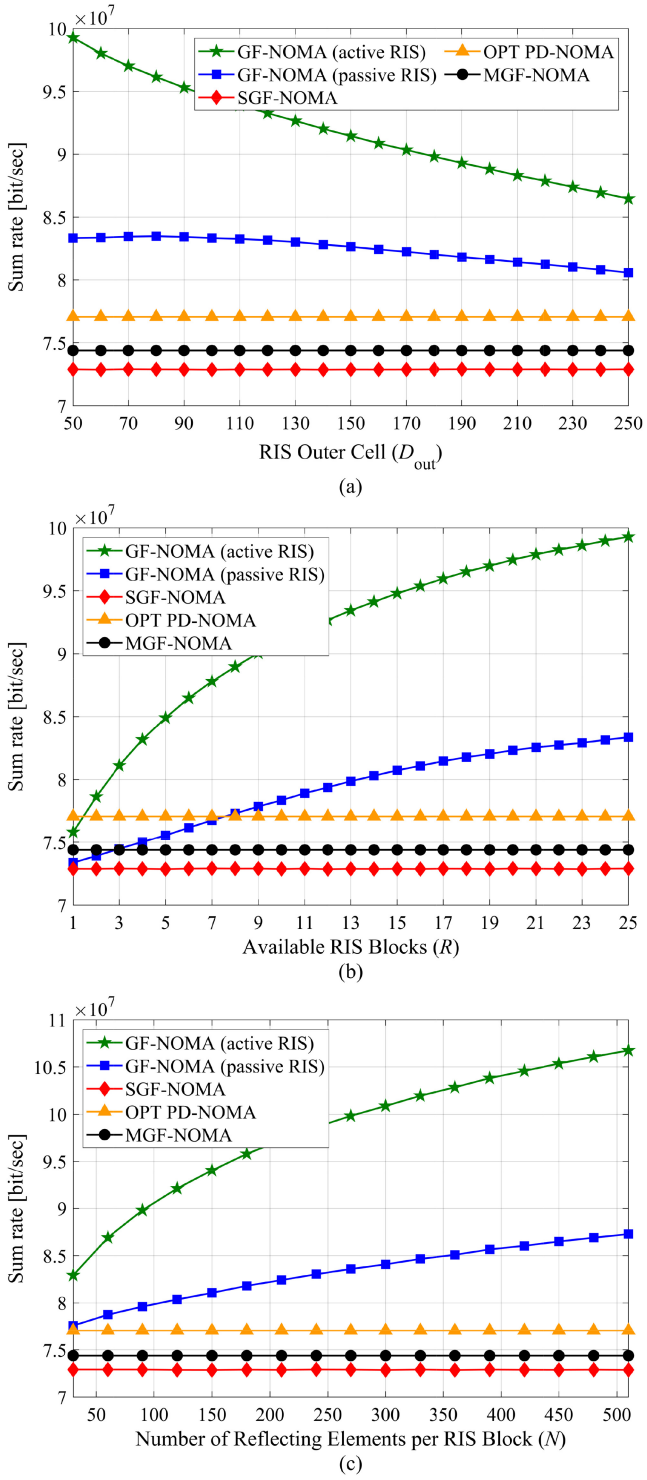

Fig. 5. Sum rate of the network for different varying RIS metrics: (a) D_{out}, (b)R and (c)N.

and 15% better than OPT PD-NOMA, MGF-NOMA, and SGF-NOMA schemes, respectively, when $D_{\text{out}} = 80$ m with passive RISs. When active RISs are used, the proposed scheme performs 25%, 30%, and 32% better than OPT PD-NOMA, MGF-NOMA, and SGF-NOMA, respectively. Thus, the active RIS scheme gives around 18% higher performance than the passive RIS scheme. As D_{out} increases, the network sum rate decreases because the RISs gradually become distant from the BS and most of the UEs. Nevertheless, the proposed scheme still performs 5%, 8%, and 11% better than OPT PD-NOMA, MGF-NOMA, and SGF-NOMA schemes, respectively, when $D_{\text{out}} = 250$ m with passive RISs. The enhancement percentages are 12%, 16%, and 19% for the case where active RISs are used, respectively. We previously mentioned that a RIS could assist the other UEs in different clusters, even if its phases are aligned with the dedicated UE. This condition also helps further improve the network sum rate when the RIS is close to the BS.

As observed in Fig. 5(b), the network performance increases with R because more spectrum are available and more cluster becomes able to use an RIS. Using passive RISs for the proposed algorithm gives 8%, 12%, and 14% better performances than OPT PD-NOMA, MGF-NOMA, and SGF-NOMA, respectively. Likewise, it enhances the network sum rate by 29%, 33%, and 36% for OPT PD-NOMA, MGF-NOMA, and SGF-NOMA, respectively, when active RISs are used. Additionally, using active RISs results in 19% better performance than using passive ones. Here, we note that using eight passive RISs gives nearly the same sum rate as using two active RISs.

Finally, Fig. 5(c) represents the sum rate of the network for varying N. The figure shows that the sum rate increases with N, similar to the case in Fig. 5(b). When passive RISs are used, the network's performance is enhanced by 13%, 17%, and 20% for OPT PD-NOMA, MGF-NOMA, and SGF-NOMA, respectively. For the other case that uses active RISs instead of the passive ones, the sum rate of the network is increased by 39%, 44% and 47% for OPT PD-NOMA, MGF-NOMA, and SGF-NOMA, respectively. In addition, a 22% increase in the network performance can be achieved by using active RISs instead of passive RISs. We emphasize that using $N = 30$ gives nearly the same sum rate as OPT PD-NOMA, meaning that we can achieve the same performance by using RISs with a very small number of reflecting elements. This amount is much lower when the active ones are used.

VI. Conclusion

This paper has proposed two iterative clustering algorithms with 3D and 2D assignment approaches for a RIS-assisted GF-NOMA system. The performances of both active and passive RIS have been considered to boost the capacity of a GF-NOMA system. Through the proposed algorithms, the network can leverage RIS assistance to create received power disparity among NOMA UEs without the need for complex power control. Also, Algorithm 2 has been found to be faster and superior to Algorithm 1 in terms of complexity. Moreover, assisting the system with RISs results in better SIC due to the power disparity introduced from the presence of the RISs, leading to improved achievable rates compared with the conventional GF-NOMA with no RIS assistance. Furthermore, it has also been shown that active RIS-assisted systems are superior to passive RIS-assisted systems in terms of maximizing the overall network sum rate and fairness among UEs, as they contain PAs in each element and enable us to exhibit a better SIC. Additionally, GF-NOMA schemes with identical transmit

power have been shown to provide more fairness among UEs in a cluster in terms of energy efficiency. Future works can consider assisting not only one UE but multiple UEs in a cluster and the multiple cell scenarios where UEs are also jointly assigned to the BSs, which can come up with some complex optimization problems.

REFERENCES

[1] R. A. Tasci, F. Kilinc, A. Celik, A. Abdallah, A. M. Eltawil, and E. Basar. "RIS-assisted grant-free NOMA." Jul. 2022. [Online]. Available: https://arxiv.org/abs/2207.11531

[2] 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. 116753–116773, 2019.

[3] E. Arslan, I. Yildirim, F. Kilinc, and E. Basar, "Over-the-air equalization with reconfigurable intelligent surfaces," *IET Commun.*, vol. 16, no. 33, pp. 1486–1497, 2022. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cmu2.12425

[4] F. Kilinc, I. Yildirim, and E. Basar, "Physical channel modeling for RIS-empowered wireless networks in sub-6 GHz bands: (Invited paper)," in *Proc. 55th Asilomar Conf. Signals Syst. Comput.*, Mar. 2021, pp. 704–708.

[5] M. Di Renzo et al., "Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 798–807, 2020.

[6] Z. Zhang et al., "Active RIS vs. passive RIS: Which will prevail in 6G?" *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1707–1725, Mar. 2023.

[7] R. A. Tasci, F. Kilinc, E. Basar, and G. C. Alexandropoulos, "A new RIS architecture with a single power amplifier: Energy efficiency and error performance analysis," *IEEE Access*, vol. 10, pp. 44804–44815, 2022.

[8] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf.*, Jun. 2013, pp. 1–5.

[9] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[10] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.

[11] E. Arslan, A. T. Dogukan, and E. Basar, "Index modulation-based flexible non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1942–1946, Nov. 2020.

[12] S. Arzykulov, G. Nauryzbayev, A. Celik, and A. M. Eltawil, "Hardware and interference limited cooperative CR-NOMA networks under imperfect SIC and CSI," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1473–1485, 2021.

[13] S. Arzykulov, A. Celik, G. Nauryzbayev, and A. M. Eltawil, "UAV-assisted cooperative & cognitive NOMA: Deployment, clustering, and resource allocation," *IEEE Trans. Cogn. Commun.*, vol. 8, no. 1, pp. 263–281, Mar. 2022.

[14] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, and L. Hanzo, "Reconfigurable intelligent surface aided NOMA networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2575–2588, Nov. 2020.

[15] Y. Liu, X. Mu, X. Liu, M. Di Renzo, Z. Ding, and R. Schober, "Reconfigurable intelligent surface-aided multi-user networks: Interplay between NOMA and RIS," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 169–176, Apr. 2022.

[16] A. Khaleel and E. Basar, "A novel NOMA solution with RIS partitioning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 1, pp. 70–81, Jan. 2022.

[17] C. Pan et al., "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 14–20, Jun. 2021.

[18] G. Yang, X. Xu, Y.-C. Liang, and M. D. Renzo, "Reconfigurable intelligent surface-assisted non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 3137–3151, May 2021.

[19] 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, Apr. 2021.

[20] Z. Yang, Y. Liu, Y. Chen, and N. Al-Dhahir, "Machine learning for user partitioning and phase shifters design in RIS-aided NOMA networks," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7414–7428, Nov. 2021.

[21] R. Zhong, Y. Liu, X. Mu, Y. Chen, and L. Song, "AI empowered RIS-assisted NOMA networks: Deep learning or reinforcement learning?" *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 182–196, Jan. 2022.

[22] Z. Tang, T. Hou, Y. Liu, J. Zhang, and C. Zhong, "A novel design of RIS for enhancing the physical layer security for RIS-aided NOMA networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2398–2401, Nov. 2021.

[23] Z. Zhang, C. Zhang, C. Jiang, F. Jia, J. Ge, and F. Gong, "Improving physical layer security for reconfigurable intelligent surface aided NOMA 6G networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4451–4463, May 2021.

[24] M. Elhattab, M.-A. Arfaoui, C. Assi, and A. Ghrayeb, "Reconfigurable intelligent surface assisted coordinated multipoint in downlink NOMA networks," *IEEE Commun. Lett.*, vol. 25, no. 2, pp. 632–636, Feb. 2021.

[25] M. Aldababsa, A. Khaleel, and E. Basar, "STAR-RIS-NOMA networks: An error performance perspective," *IEEE Commun. Lett.*, vol. 26, no. 8, pp. 1784–1788, Aug. 2022.

[26] M. Aldababsa, A. Khaleel, and E. Basar. "Simultaneous transmitting and reflecting intelligent surfaces-empowered NOMA networks." Oct. 2021. [Online]. Available: https://arxiv.org/abs/2110.05311

[27] A. Hemanth, K. Umamaheswari, A. C. Pogaku, D.-T. Do, and B. M. Lee, "Outage performance analysis of reconfigurable intelligent surfaces-aided NOMA under presence of hardware impairment," *IEEE Access*, vol. 8, pp. 212156–212165, 2020.

[28] Q. Chen, M. Li, X. Yang, R. Alturki, M. D. Alshehri, and F. Khan, "Impact of residual hardware impairment on the IoT secrecy performance of RIS-assisted NOMA networks," *IEEE Access*, vol. 9, pp. 42583–42592, 2021.

[29] E. Balevi, F. T. A. Rabee, and R. D. Gitlin, "ALOHA-NOMA for massive machine-to-machine IoT communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–5.

[30] M. Elkourdi, A. Mazin, E. Balevi, and R. D. Gitlin, "Enabling slotted aloha-NOMA for massive M2M communication in IoT networks," in *Proc. IEEE 19th Wireless Microw. Technol. Conf. (WAMICON)*, 2018, pp. 1–4.

[31] C. Zhang, Y. Liu, and Z. Ding, "Semi-grant-free NOMA: A stochastic geometry model," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1197–1213, Feb. 2022.

[32] C. Zhang, Z. Qin, Y. Liu, and K. K. Chai, "Semi-grant-free uplink NOMA with contention control: A stochastic geometry model," in *Proc. IEEE Int. Conf. Commun. (ICC Workshops)*, 2020, pp. 1–6.

[33] J. Liu, G. Wu, X. Zhang, S. Fang, and S. Li, "Modeling, analysis, and optimization of grant-free NOMA in massive MTC via stochastic geometry," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4389–4402, Mar. 2021.

[34] M. Fayaz, W. Yi, Y. Liu, and A. Nallanathan, "Transmit power pool design for grant-free NOMA-IoT networks via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7626–7641, Nov. 2021.

[35] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Mar. 2019.

[36] S. Doğan, A. Tusha, and H. Arslan, "NOMA with index modulation for uplink URLLC through grant-free access," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 6, pp. 1249–1257, Oct. 2019.

[37] A. Celik. "Grant-free NOMA: A low complexity power control through user clustering." Apr. 2021. [Online]. Available: https://www.techrxiv.org/articles/preprint/Grant-Free_NOMA_A_Low_Complexity_Power_Control_Through_User_Clustering/19688019/1

[38] J. Chen, L. Guo, J. Jia, J. Shang, and X. Wang, "Resource allocation for IRS assisted SGF NOMA transmission: A MADRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1302–1316, Apr. 2022.

[39] "Study on channel model for frequencies from 0.5 to 100 GHz, v16.1.0," 3GPP, Sophia Antipolis, France, Rep. TR 38.901, Dec. 2019.

[40] I. Yildirim, A. Uyrus, and E. Basar, "Modeling and analysis of reconfigurable intelligent surfaces for indoor and outdoor applications in future wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1290–1301, Feb. 2021.

[41] 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.

[42] A. Abdallah, A. Celik, M. M. Mansour, and A. M. Eltawil, "Deep learning-based channel estimation for wideband RIS-aided mmWave MIMO system with beam squint," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, 2022, pp. 1269–1275.
[43] A. Abdallah, A. Celik, M. M. Mansour, and A. M. Eltawil, "RIS-aided mmWave MIMO channel estimation using deep learning and compressive sensing," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 3503–3521, May 2023.
[44] K. K. Kishor and S. V. Hum, "An amplifying reconfigurable reflectarray antenna," *IEEE Trans. Antennas Propag.*, vol. 60, no. 1, pp. 197–205, Jan. 2012.
[45] R. Long, Y.-C. Liang, Y. Pei, and E. G. Larsson, "Active reconfigurable intelligent surface-aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 4962–4975, Aug. 2021.
[46] B. Makki, T. Svensson, T. Eriksson, and M. Nasiri-Kenari, "On the throughput and outage probability of multi-relay networks with imperfect power amplifiers," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 4994–5008, Sep. 2015.
[47] D. Persson, T. Eriksson, and E. G. Larsson, "Amplifier-aware multiple-input–multiple-output power allocation," *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1112–1115, Jun. 2013.
[48] K. Pattipati, S. Deb, Y. Bar-Shalom, and R. Washburn, "A new relaxation algorithm and passive sensor data association," *IEEE Trans. Autom. Control*, vol. 37, no. 2, pp. 198–213, Feb. 1992.
[49] M. Grant and S. Boyd. "CVX: MATLAB software for disciplined convex programming, version 2.1." Mar. 2014. [Online]. Available: http://cvxr.com/cvx

Fatih Kilinc (Graduate Student Member, IEEE) received the B.S. degree from Istanbul Medipol University in 2020, and the M.S. degree from Koç University in 2022. He is currently working as a Research and Development Engineer with ULAK Communications Inc. His research interests include channel modeling, intelligent surfaces, smart repeaters, and signal processing for wireless communications.

Recep A. Tasci (Graduate Student Member, IEEE) received the B.S. degree in electrical and electronics engineering from Istanbul Medipol University in 2020, and the M.S. degree in electrical and electronics engineering from Koç University in 2022, where he is currently pursuing the Ph.D. degree. He is a Research and Teaching Assistant with Koç University. His research interests include wireless communications, reconfigurable intelligent surfaces, thermal noise communication, channel modeling, and signal processing.

Abdulkadir Celik (Senior Member, IEEE) received the first M.S. degree in electrical engineering in 2013, the second M.S. degree in computer engineering in 2015, and the Ph.D. degree in co-majors of electrical engineering and computer engineering from Iowa State University, Ames, IA, USA, in 2016. He was a Postdoctoral Fellow with the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, from 2016 to 2020, where he is currently a Research Scientist with Communications and Computing Systems Laboratory. His research interests are in the areas of next-generation wireless communication systems and networks. He currently serves as an Editor for IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATION LETTERS, and *Frontiers in Communications and Networks*.

Asmaa Abdallah (Member, IEEE) received the B.S. (with High Distinction) and M.S. degrees in computer and communications engineering from Rafik Hariri University (RHU), Lebanon, in 2013 and 2015, respectively, and the Ph.D. degree in electrical and computer engineering from the American University of Beirut (AUB), Beirut, Lebanon, in 2020.

She is currently a Postdoctoral Fellow with the King Abdullah University of Science and Technology. Her research interests include communication theory, stochastic geometry for wireless communications, machine learning for wireless communications, array signal processing, with emphasis on energy and spectral efficient algorithms for emerging wireless communication technologies. Through her academic years, she was the recipient of a Scholarship from the Lebanese National Counsel for Scientific Research (CNRS-L/AUB) to support her doctoral studies and received the Academic Excellence Award at RHU in 2013 for ranking first in the graduating class. In 2022, she was selected as one of 15 leading innovators in the Middle East by MIT technology review under 35 years old. She also served as an Executive Committee Member of IEEE Young Professionals Lebanon's Section from 2016 to 2019.

Ahmed M. Eltawil (Senior Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees from Cairo University, Giza, Egypt, in 1999 and 1997, respectively, and the Ph.D. degree from the University of California at Los Angeles, Los Angeles, CA, USA, in 2003. He has been a Professor of Electrical Engineering and Computer Science with the University of California at Irvine (UCI), Irvine, CA, USA, since 2005. He is currently a Professor of Electrical and Computer Engineering with the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, where he joined the Division of Computer, Electrical and Mathematical Sciences and Engineering in 2019. At KAUST, he is the Founder and the Director of the Communication and Computing Systems Laboratory. His research is in the area of efficient architectures for computing and communications systems in general and wireless systems in particular, spanning the application domains of body area networks, low-power mobile systems, machine learning platforms, sensor networks, and critical infrastructure networks. He received several awards, including the NSF CAREER Award supporting his research in low-power computing and communication systems. He is a Distinguished Lecturer at the IEEE. He was selected as the "Innovator of the Year" for 2021 by the Henry Samueli School of Engineering at UCI. For his contributions to societal benefit through wireless innovations, he received two certificates of recognition from the United States Congress. He has been on the technical program committees and steering committees for numerous workshops, symposia, and conferences in the areas of low-power computing and wireless communication system design. He is a Senior Member of the National Academy of Inventors, USA.

Ertugrul Basar (Fellow, IEEE) received the Ph.D. degree from Istanbul Technical University in 2013. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, Koç University, Istanbul, Turkey, and the Director of Communications Research and Innovation Laboratory. He was a Mercator Fellow with Ruhr University Bochum, Germany, in 2022, and a Visiting Research Collaborator with Princeton University, USA, from 2011 to 2012. His primary research interests include beyond 5G and 6G wireless networks, communication theory and systems, reconfigurable intelligent surfaces, index modulation, waveform design, and signal processing for communications. In the past, he served as an Editor/Senior Editor for many journals, including IEEE COMMUNICATIONS LETTERS from 2016 to 2022, IEEE TRANSACTIONS ON COMMUNICATIONS from 2018 to 2022, *Physical Communication* from 2017 to 2020, and IEEE ACCESS from 2016 to 2018. He is currently an Editor of *Frontiers in Communications and Networks*. He is a Young Member of the Turkish Academy of Sciences.