

An optimization framework for RIS-based energy-efficient multi-cell NOMA systems [☆]

A. Abdelaziz Salem ^{a,b,*}, Ahmed S. Ibrahim ^c, Mahmoud H. Ismail ^{a,d}

^a Department of Electrical Engineering, American University of Sharjah, Sharjah, 26666, United Arab Emirates

^b Department of Electronics and Electrical Communications Engineering, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Menoufia, Egypt

^c Electrical and Computer Engineering Department, Florida International University, Miami 33174, USA

^d Department of Electronics and Communications Engineering, Faculty of Engineering, Cairo University, Giza 12613, Egypt

ARTICLE INFO

Article history:

Received 18 May 2023

Received in revised form 13 July 2023

Accepted 2 August 2023

Available online 8 August 2023

Keywords:

Reconfigurable intelligent surfaces (RIS)

Alternating optimization

Energy efficiency (EE)

Non-orthogonal multiple access (NOMA)

ABSTRACT

Non-orthogonal multiple access (NOMA) is a highly promising multiple access strategy for wideband communication systems due to its high spectral efficiency. However, in multi-cell multi-band systems, interference issues can arise from resource sharing among multiple users, which limits the potential of NOMA. To tackle these issues, in this paper, we study a NOMA system where reconfigurable intelligent surfaces (RIS) are employed across multiple cells with multiple users clusters. Specifically, we propose an optimization framework to maximize the system energy efficiency (EE) through jointly designing the transmission beamformer at each base station, the RIS phase shifts and the corresponding binary selection matrix indicating which elements of the RIS should be used to serve a specific cell. Our optimization framework takes into account the minimum rate requirements of users, the successive interference cancellation (SIC) decoding-based rate constraint, fairness among users as well as the maximum allowed transmission power. The resulting non-convex problem is solved by using an alternating optimization algorithm, wherein semidefinite programming (SDP), successive convex approximation (SCA), and penalty-based difference-of-convex (DC) programming are used to reach a solution. Simulation results show that the proposed system outperforms conventional orthogonal access in terms of EE performance, the achievable rate, and power consumption. Also, the results revealed the critical transmission power level for NOMA beyond which the EE starts decaying. Moreover, the proposed NOMA system shows a limited EE performance compared to the conventional orthogonal access as the number of clusters exceeds a certain limit.

© 2023 Elsevier Inc. All rights reserved.

1. Introduction

For the past decade, smart devices have significantly penetrated the wireless market, which created a notable gap between the available radio resources and the demand for high data rate and low latency communications [1]. Moreover, the expected large-scale deployment of 6G networks in the near future will bring the challenges of hardware complexity, energy consumption, unprecedented growth of real-time services and ultra-wideband connectiv-

ity [2]. Therefore, research initiatives aiming at providing advanced spectral/energy efficient solutions, while limiting the interference impact have been well underway. Specifically, there has been great interest in using Non-orthogonal multiple access (NOMA) and reconfigurable intelligent surfaces (RIS) as candidate solutions for spectrum congestion and interference challenges [3–5]. NOMA has been selected as a multiple access candidate for 6G networks to try to strike the balance between spectrum utilization and user fairness [6]. The NOMA philosophy depends on applying superposition coding at the transmitters and successive interference cancellation (SIC) at the receivers to be able to share spectrum resources among all users in the power or code domains.

Benefiting from the innovative concept of RIS, the communication channels' conditions can be adjusted via optimizing the phase shifts and amplitude of the reflecting elements that comprise the RIS. Accordingly, the network performance can be enhanced through extending its coverage, improving the data rates experienced by the users, and guaranteeing fairness among them

[☆] The work of A. A. Salem and M. H. Ismail is supported by the American University of Sharjah through Faculty Research Grants number FRG20-M-E10 and FRG22-C-E13. The work of A. S. Ibrahim is supported in part by the National Science Foundation under Award CNS-2144297.

* Corresponding author.

E-mail addresses: ahmed.abdalaziz40@el-eng.menofia.edu.eg (A.A. Salem), aibrahim@fu.edu (A.S. Ibrahim), mhibrahim@aus.edu, mhiis@cu.edu.eg (M.H. Ismail).

[7]. Thanks to the passive nature of the reflecting elements, RIS-enabled networks are considered cost-effective from the hardware and energy consumption aspects. Moreover, by virtue of the RIS degrees of freedom (DoFs), it can be employed to assist multiple cells located in different directions [8]. Based on the above, the interplay between NOMA and RIS has motivated various research works to investigate the challenges that come with RIS-assisted NOMA systems [9].

1.1. Related work

1.1.1. RIS-NOMA single-cell networks

In the recent few years, RIS-aided networks have been investigated while focusing on single-cell deployments with single/multi-user and multi-input single (multi)-output (MISO)/MIMO, where spectral efficiency (SE) [10], energy efficiency (EE) [11], and sum rate [12,13] were the main concern. Moreover, the EE of single-cell NOMA networks has been examined in [14–19]. Specifically, with the objective of maximizing the EE, Fang et al. [15] exploited the difference of convex (DC) programming to optimize the power allocation and sub-channels assignment. Also, a dynamic power allocation algorithm was developed in [16] to balance between user fairness and system throughput via maximizing the weighted sum rate. In [17], branch and bound (BnB) and successive convex approximation (SCA) approaches were invoked to find the optimal power that can be allocated for a millimeter wave (mm-Wave) NOMA network. Furthermore, the outage probability was investigated by studying the impact of users pairing in [18], where the difference among users' channel gains is exploited to demonstrate that NOMA can provide better performance than conventional orthogonal access. In addition, the authors in [19] maximized the fairness of a MIMO-NOMA system by jointly optimizing the power allocation and receive beamforming, subject to a reliability constraint. Wang et al. investigated an uplink NOMA system in [20], wherein the base station (BS) supports parallel detection of users' groups to minimize the total energy consumption under the consideration of imperfect SIC. Also, in [21], Air-Ground Integrated clustered-NOMA (C-NOMA) Heterogeneous Power Internet of Things (PIoT) Networks were utilized to address the limitation of low computational resources and remote deployment of PIoT. The authors minimized the energy consumption by designing task offloading and resource allocation. Finally, motivated by the limited on-board energy storage of unmanned aerial vehicles (UAVs) and the finite battery life of IoT devices, C-NOMA-enabled integrated UAVs and high altitude platform stations (HAPs) were proposed in [22], where EE is maximized by optimizing the UAV trajectory plan and resource allocation using a two-step iterative approach.

1.1.2. RIS-NOMA multi-cell networks

Different from the single-cell case, RIS were employed in multi-cell NOMA networks in [23–28]. In this context, the resource allocation problem has been addressed in [25] in terms of EE maximization, wherein the matching theory and DC programming were adopted. Furthermore, motivated by reducing the overhead among BSs, Fu et al. minimized the total transmission power at the BSs by designing a distributed power control algorithm in [26], meanwhile the requirement of minimum target rates is satisfied for all users. Also, by considering both EE and user fairness, matching games and SCA are invoked by Zhao et al. in [28] to iteratively update the power and spectrum allocation. In addition, subject to SIC decoding and quality of service (QoS) requirements, the sum rate was maximized for a multi-cell RIS-NOMA network in [29]. In this work, the authors optimized the users' associations via matching theory. Meanwhile, they exploited convex upper bound substitution, SCA, and semi-definite relaxation (SDR) to iteratively optimize power

allocations, RIS phase shifts, and decoding order, respectively. Further, in [30], RIS are leveraged to minimize the uplink transmission power in two cells assisted by coordinated multi-point transmission (CoMP), where the power allocation and RIS reflection matrix are designed to meet the signal-to-interference plus noise ratio (SINR) requirement of the uplink users. Finally, with the objective of minimizing the total power in a downlink multi-cell single-input single-output (SISO) NOMA network, game theory and graph theory are exploited to study QoS-aware user grouping strategies in [31].

To reap the aforementioned advantages of both NOMA and RIS, namely, improving spectrum and energy efficiencies, respectively, these two metrics can be maximized via integrating them into one performance measure. This is imperative for NOMA-empowered multi-cell networks due to the following reasons. Firstly, the multi-clustered NOMA suffers from intra- and inter-cluster interference, which can be suppressed by properly designing the RIS phase shifts and transmit beamformers at each BS (spatial DoFs at BSs and RIS). Secondly, the NOMA signal is highly attenuated at cell-edges or severely faded at direct communication links, thus RIS provides coverage extension, while NOMA provides user fairness¹ and spectral efficiency. Finally, the decoding order of NOMA users can be effectively adjusted by tuning the RIS phase shifts, which enhances the network performance.

1.2. Motivation

It is clear that the majority of the aforementioned work focused on optimizing the RIS-aided multi-cell NOMA performance by creating power allocation algorithms, treating all users in a cell as one group. This is not practical as cells could have multiple clusters as well as different number of users per cluster and hence, multiple interference sources to deal with. Hence, we divide the users in a single cell into multiple clusters and optimize a dedicated beamforming vector for each user. Additionally, these studies did not consider the frequency-selective properties of RIS for the scenario of multiple cells operating over different frequency bands, where the RIS elements would show a different response for the impinging signals from different bands/cells. Accordingly, we design a binary selection matrix to ensure that there are no overlaps among the elements chosen for a specific NOMA band, as will be explained in Section 2.2. We show that the reflection coefficients of the RIS elements non-linearly depend on the operational frequency of the incident signal, which could further complicate the phase shift optimization process. Thus, we discuss how the binary indicator transforms this non-linearity into a simple linear Hadamard product of phase shift and index of the selected elements. Without loss of generality, this paper mainly investigates the joint design of active beamforming, applied for each user in all cells' clusters, and cell-RIS element(s) association as well as the passive beamforming of associated element(s).

We aim to design an energy-efficient RIS-aided multi-cell NOMA system, wherein a more practical RIS model is indirectly employed for multiple frequency bands multi-cell operation. Through this model, a unique set of RIS elements, or equivalently the electric circuits parameters of these elements, are independently chosen and phase-tuned to operate over a specific cell's assigned operating frequency. In order to achieve this goal, we firstly formulate the problem of NOMA EE maximization, in which we jointly design the BS transmit beamformers for each user to align the users' channels in each cluster, the RIS elements selection for each cell as well as their phase shifts. This is all subject

¹ NOMA fairness implies protecting the weak channel users from being overwhelmed by assigning most of the network resources to the strong channel users.

Table 1

Highlighting our distinctive contributions to the existing literature.

	Our paper	[30]-2020	[29]-2021	[31]-2021	[32]-2021	[33]-2022	[34]-2021	[35]-2021
Multi-cell and/or multi band	✓	✓		✓			✓	✓
Multi-clustered and/or Power domain NOMA	✓		✓	✓	✓	✓	✓	
Optimize element selection and phase of RIS	✓							✓
Per-user beamforming	✓					✓		
Multi-antenna BSs	✓					✓	✓	✓

to successful SIC decoding and satisfying minimum rate and fairness among users sharing the same resources. Unfortunately, this large-scale design brings challenges represented by the coupling between the design variables, specially the phase shift and RIS element selection. These challenges lead to non-convex problems, that cannot be solved directly in polynomial time.

To tackle these issues, the main problem is divided into three sub-problems. The first one seeks to optimize the transmit beamforming, while considering the maximum allowed transmission power, rate requirements, and fairness constraints. In the second one, the RIS phase shifts are designed for a given set of selected elements, where the rate and fairness requirements are ensured. While in the third problem, we design a binary selection matrix to ensure that there are no overlaps among the elements chosen for a specific NOMA band. To solve these problems, an efficient iterative alternating optimization technique based on semidefinite programming (SDP), SCA, and penalty-based DC programming are proposed. Finally, simulation results show that the EE performance of the proposed system outperforms that of conventional orthogonal multiple access (OMA) schemes, with and without RIS.

In summary, the main contributions of this paper compared to the state-of-the-art are effectively highlighted through a concise presentation in Table 1. To provide further details:

- Proposing a framework for optimizing the BS transmit beamformers, the RIS reflection coefficients for each cell, as well as the binary selection matrix associating a specific set of RIS elements to each cell, to maximize the EE of a multi-cluster multi-cell NOMA system. This is achieved under constraints of maximum allowed transmission power, fairness, and users' rate requirements.
- Introducing an efficient iterative optimization approach, which guarantees the convergence to the optimal solution of the three sub-problems.

1.3. Paper organization and notation

The remaining part of this paper is organized as follows. The system and signaling models are introduced in Section 2. The problem of EE maximization is formulated for practical RIS assisted NOMA system in Section 3 while considering QoS, fairness, and element selectivity constraints. Numerical and simulation results are then presented and discussed in Section 4 before the paper is finally concluded in Section 5.

Regarding notation, lowercase boldface letters are used to denote vectors, while the uppercase boldface ones are utilized for matrices. Normal face letters are used for scalars. Mathematical fonts, e.g., $\mathcal{B}$ are dedicated for set notation. $\mathbb{H}^{N \times N}$, $\mathbb{C}^{N \times M}$, $\text{diag}(\cdot)$ and $\text{Tr}(\cdot)$ denote the $N \times N$ Hermitian matrix, complex space with dimension $N \times M$, the square diagonal matrix with the elements of the input vector on its main diagonal, and the trace operator, respectively. The notations $\mathbf{X} \succeq 0$, $\mathbf{X}^T$, $\mathbf{X}^*$, and $\mathbf{X}^H$ denotes Hermitian positive semidefinite property, transpose, conjugate, and the conjugate transpose of the matrix $\mathbf{X}$, respectively. $\Re\{\cdot\}$ refers to the real parts of. Finally, $\mathbf{I}_M$ denotes the $M \times M$ unit matrix and $\mathbf{1}$ denotes the all-ones matrix/vector.

2. System model

We consider a NOMA multi-cell downlink communication system, wherein a single RIS of M elements is dedicated to improve the EE of L cells. In the l^{th} cell, the BS transmits a composite signal by exploiting N_t transmit antennas towards a set of NOMA-clustered users, as shown in Fig. 1. We assume that users per cell l are grouped into K_l clusters, such that $N_t \geq K_l$, where the users' signals are multiplexed using NOMA within each cluster k_l and all clusters' signals are MIMO multiplexed within each cell l . The number of users in each cluster k_l is U_k^l . Without loss of generality, the sets $\mathcal{L} \triangleq \{1, \dots, L\}$, $\mathcal{K}_l \triangleq \{1, \dots, K_l\}$, $\mathcal{M} = \{1, \dots, M\}$, and $\mathcal{U}_k^l \triangleq \{1, \dots, U_k^l\}$ define the sets of BSs, NOMA clusters per cell l , RIS reflecting elements, and the single antenna users per cluster k and cell l , respectively. Moreover, we assume that each cell is allocated a specific bandwidth $B_l = W/L$, where W denotes the total bandwidth of the cellular system.

2.1. Transmitted signal

Assuming that BS l transmits a composite NOMA signal $\mathbf{x}_l$ by enabling superposition coding. Then, the broadcast NOMA signal is

$$\mathbf{x}_l = \sum_{k_l=1}^{K_l} \sum_{u_k^l=1}^{U_k^l} \mathbf{w}_{l,k_l,u_k^l} s_{l,k_l,u_k^l}, \quad (1)$$

where $\mathbf{w}_{l,k_l,u_k^l} \in \mathbb{C}^{N_t \times 1}$ and s_{l,k_l,u_k^l} represent the digital beamforming vector and transmitted signal for user $u_k^l \in \mathcal{U}_k^l$ respectively, where $\mathbb{E} \left[|s_{l,k_l,u_k^l}|^2 \right] = 1$. Accordingly, the received signal at user u_k^l is expressed as in (2) on top of the next page, where $\mathbf{h}_{d,u_k^l} \in \mathbb{C}^{N_t \times 1}$, $\mathbf{h}_{r,u_k^l} \in \mathbb{C}^{M \times 1}$ and $\mathbf{G}_l \in \mathbb{C}^{M \times N_t}$ represent the direct channel between BS l and u_k^l , the reflection channel between the RIS and user u_k^l , and the channel between BS l and RIS. $n_{u_k^l}$ denotes the additive white Gaussian noise (AWGN) with zero mean and variance σ^2 , i.e., $n_{u_k^l} \sim \mathcal{CN}(0, \sigma^2)$.

We assume that all communication links follow a quasi-static Rayleigh flat-fading, where the channel state information (CSI) is assumed known at the BS through deploying real-time tracking and channel estimation techniques [36–38]. Moreover, the RIS is controlled via a specific control link to adjust the reflection coefficient matrix Θ_l for each operational frequency of BS l , where $\Theta_l = \text{diag}\{\theta_l\} \in \mathbb{C}^{M \times M}$ and $\theta_l = [\theta_{l,1}, \dots, \theta_{l,M}]^T$.

2.2. Practical RIS model

The RIS response varies with the frequency of the incident signal,² where an RIS element is equivalent to a parallel resonance

² We assume that the amplitude degradation of a reflecting element, due to different frequency band operation, can be compensated by the transmitting BS [39]. Hence, we can neglect the amplitude response of RIS elements without affecting the system performance.

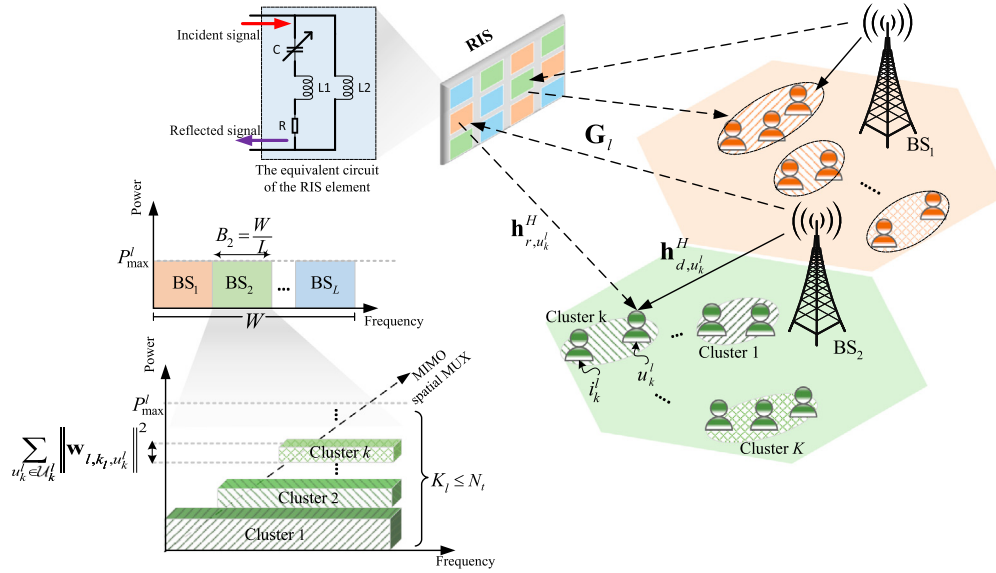


Fig. 1. An RIS-enabled multi-cell NOMA system. The color of the RIS elements indicates that these elements are tuned according to a specific cell's operational frequency.

$$y_{u_k^l} = \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \mathbf{\Theta}_l \mathbf{G}_l \right) \mathbf{x}_l + n_{u_k^l}, \quad (2a)$$

$$= \underbrace{\left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \mathbf{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,k_l,u_k^l} s_{l,k_l,u_k^l}}_{\text{intended signal}} + \underbrace{\sum_{n_k^l \neq u_k^l} \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \mathbf{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,k_l,n_k^l} s_{l,k_l,n_k^l}}_{\text{intra-cluster interference}} + \underbrace{\sum_{c_l \neq k_l} \sum_{x_c^l \in \mathcal{U}_c^l} \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \mathbf{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,c_l,x_c^l} s_{l,c_l,x_c^l}}_{\text{inter-cluster interference}} + n_{u_k^l}, \quad (2b)$$

circuit, see Fig. 1, that can be tuned to that specific frequency. Specifically, the equivalent impedance of an RIS element is denoted by [40] $Z(f, c) = \frac{Z_{L1}(Z_{L2} + Z_C + R)}{Z_{L1} + Z_{L2} + Z_C + R}$, where $Z_L = j2\pi fL$ and $Z_C = 1/j2\pi fC$ denote the equivalent impedance of an inductor and capacitor, respectively. Hence, the reflection coefficient of the RIS element is given by $\theta(f, c) = \frac{Z(f, c) - Z_0}{Z(f, c) + Z_0}$, where $Z_0 = 377 \Omega$ denotes the impedance of free space. However, the relation between the frequency and RIS phase shift is a non-linear relationship, which could further complicate the phase shift optimization process [35,40,41]. Moreover, adopting multiple RISs for multiple cells operating on different bands brings additional complexity in terms of optimal deployment, routing reflections and user association. These issues can be alleviated if the RIS elements are treated as network resources, enabling them to be partitioned and allocated optimally as will be adopted in this work.

Hereby, we introduce the binary variable $\mathbf{a}_l \triangleq [a_{l,1}, a_{l,2}, \dots, a_{l,M}]^T$, where $a_{l,m} \in \{0, 1\} \forall l, m$ which indicates that the cell l can be served by a set of unique reflecting elements as shown in Fig. 1. Specifically, $a_{l,m} = 1$ means that an element $m \in \mathcal{M}$ is allocated for BS l with a fully tunable phase shift $\boldsymbol{\varphi}_l = [\varphi_{l,1}, \dots, \varphi_{l,M}]^T$, where $\varphi_{l,m} \in (0, 2\pi] \forall l, m$. Otherwise, the phase shifts are chosen as $\varphi_{l,m} \triangleq \{0, 2\pi\}$, $\forall l \neq \tilde{l}, \tilde{l} \in \mathcal{L}$. Therefore, the actual phase shift, of the cell l , is determined by $\boldsymbol{\theta}_l = \exp(j\boldsymbol{\varphi}_l \odot \mathbf{a}_l)$, where $\odot$ defines element-wise or Hadamard multiplication. It is worth mentioning that the allocated RIS elements of the l^{th} BS cannot be reused for another BS $\tilde{l} \neq l, \tilde{l} \in \mathcal{L}$, since the resonance circuit of an RIS element can be adjusted for a single frequency of an incident signal.

To model this fact, the binary allocation vector of the m^{th} element, i.e., $\tilde{\mathbf{a}}_m = [\tilde{a}_{1,m}, \tilde{a}_{2,m}, \dots, \tilde{a}_{L,m}]^T$ should satisfy the following constraint,

$$\|\tilde{\mathbf{a}}_m\|_0 \leq 1. \quad (3)$$

Then, the binary allocation matrix for the combination of L cells and M elements can be defined by the $L \times M$ matrix $\mathbf{A} = [\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \dots, \tilde{\mathbf{a}}_M] = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L]^T$.

2.3. NOMA signaling

The user $u_k^l \in \mathcal{U}_k^l$ deploys the SIC process to remove the experienced co-channel interference from the other users being served in the same bandwidth B_l . In a given cluster k_l , a user with higher priority of detection should be firstly decoded by users of higher decoding order. After that, the high decoding order user subtracts the decoded signal(s) to extract their own. Accordingly, the SIC process complexity increases with increasing the number of users in each cluster.

Moreover, grouping NOMA users in each cell is not an easy task, where it requires an exhaustive search through all the possible combinations of users' clustering and different decoding orders ($U_k^l!$). In our work, we have adopted a simple clustering approach, which depends on grouping the users in a NOMA cluster based on a heuristic search method that guarantees differences between the cascaded channel gains between the users within the same NOMA cluster. This clustering approach has been proposed in [42]. Since, the clustering approach is highly coupled with the RIS phase

shifts, then the optimal SIC decoding order is updated based on the obtained optimal selection matrix $\mathbf{A}$ and phase shift Θ_l through the iterative process of alternating optimization, where the combined channel gains are reordered ascendingly for each BS as in Algorithm 4 of [29]. Another approach of exhaustive search can be applied to determine the optimal decoding order based on the RIS phase shift matrix [33]. Without loss of generality, we assume that the decoding order of cluster k_l in cell l is defined as $\dots > \pi(j_k^l) > \pi(u_k^l) > \pi(i_k^l) > \dots$, where $\pi(u_k^l)$ is the decoding order of user u_k^l , where $\pi(u_k^l) = u_k^l$ means the u^{th} signal of user u is the signal to be decoded in cluster k_l . To this end, u_k^l has to decode user i_k^l 's signal, before extracting their own signal if and only if $\pi(u_k^l) > \pi(i_k^l)$, while the user j_k^l , with $\pi(j_k^l) > \pi(u_k^l)$ is treated as an interferer [43].

Denoting the equivalent cascaded channel of user u_k^l as $\mathbf{h}_{u_k^l}^H = \mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \Theta_l \mathbf{G}_l$, the achievable SINR of u_k^l , while decoding their own signal, according to (2), is expressed as

$$\text{SINR}_{u_k^l \rightarrow u_k^l} = \frac{|\mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,u_k^l}|^2}{I_{u_k^l}}, \quad (4)$$

where $I_{u_k^l} = \sum_{\pi(i_k^l) > \pi(u_k^l)} |\mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,i_k^l}|^2 + \sum_{c_l \in \{\mathcal{K}_l \setminus k_l\}} \sum_{x_c^l \in \mathcal{U}_c^l} |\mathbf{h}_{u_k^l}^H \mathbf{w}_{l,c_l,x_c^l}|^2 + \sigma^2$ denotes the total interference plus noise experienced. Accordingly, the corresponding achievable rate is calculated as $R_{u_k^l \rightarrow u_k^l} = B_l \log_2(1 + \text{SINR}_{u_k^l \rightarrow u_k^l})$. Moreover, the SINR of decoding signal u_k^l at user j_k^l , given that $\pi(j_k^l) > \pi(u_k^l)$ is

$$\text{SINR}_{u_k^l \rightarrow j_k^l} = \frac{|\mathbf{h}_{j_k^l}^H \mathbf{w}_{l,k_l,u_k^l}|^2}{I_{u_k^l \rightarrow j_k^l}}, \quad (5)$$

where $I_{u_k^l \rightarrow j_k^l} = \sum_{\pi(i_k^l) > \pi(u_k^l)} |\mathbf{h}_{j_k^l}^H \mathbf{w}_{l,k_l,i_k^l}|^2 + \sum_{c_l \in \{\mathcal{K}_l \setminus k_l\}} \sum_{x_c^l \in \mathcal{U}_c^l} |\mathbf{h}_{j_k^l}^H \mathbf{w}_{l,c_l,x_c^l}|^2 + \sigma^2$ expresses the total interference plus noise experienced at user j_k^l while decoding u_k^l . Then, the corresponding rate is given by $R_{u_k^l \rightarrow j_k^l} = B_l \log_2(1 + \text{SINR}_{u_k^l \rightarrow j_k^l})$.

Moreover, the fairness among NOMA users in a given cluster $k_l \in \mathcal{K}_l$ can be achieved by guaranteeing that the user with high decoding order will not consume most of the cell resources [34]. Hence, the following condition should be enforced to ensure a reasonable communication rate for users with low decoding orders

$$\begin{aligned} |\mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,\pi(i_k^l)}|^2 &\leq |\mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,\pi(j_k^l)}|^2, \\ \forall \pi(i_k^l) > \pi(j_k^l), \{u_k^l, i_k^l, j_k^l\} &\in \mathcal{U}_k^l, l, k_l \end{aligned} \quad (6)$$

3. Problem formulation and proposed optimization framework

As mentioned earlier, we focus our attention on the total EE of the system defined as the ratio of the total achievable rate to the total consumed power for achieving this rate, viz.,

$$\eta_{\text{EE}}(\mathbf{w}_{l,k_l,u_k^l}, \boldsymbol{\theta}_l) = \frac{\sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} R_{u_k^l \rightarrow u_k^l}}{P_{\text{tot}}^l}, \quad (7)$$

where the total power consumption is denoted by P_{tot}^l . This mainly includes the circuit power consumption of all BSs $\sum_{l \in \mathcal{L}} P_{\text{cir}}^l$, and the total transmission power $\sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \|\mathbf{w}_{l,k_l,u_k^l}\|^2$. Hence, we have

$$P_{\text{tot}}^l = \sum_{l \in \mathcal{L}} P_{\text{cir}}^l + \sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \|\mathbf{w}_{l,k_l,u_k^l}\|^2. \quad (8)$$

EE provides a trade off between the maximum achieved sum rate and total power consumption, where increasing the transmission power improves the rate but increases the intra- and inter-cluster interference in NOMA.

We consider a constrained EE maximization problem via jointly designing the transmission beamformers $\mathbf{W} \triangleq \{\mathbf{w}_{l,k_l,u_k^l}\}_{u_k^l \in \mathcal{U}_k^l, k_l \in \mathcal{K}_l, l \in \mathcal{L}}$, RIS phase shifts $\Phi = [\varphi_1, \dots, \varphi_L]$ and the binary selection matrix $\mathbf{A}$, while ensuring the QoS constraints, perfect SIC decoding, as well as fairness among users. The corresponding optimization problem can be formulated as

$$\text{Maximize } \eta_{\text{EE}} \quad (9a)$$

subject to:

$$R_{u_k^l \rightarrow u_k^l} \geq R_{\min}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (9b)$$

$$\begin{aligned} |\mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,\pi(i_k^l)}|^2 &\leq |\mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,\pi(j_k^l)}|^2, \\ \forall \pi(i_k^l) > \pi(j_k^l), \{u_k^l, i_k^l, j_k^l\} &\in \mathcal{U}_k^l, l, k_l, \end{aligned} \quad (9c)$$

$$P_{\text{cir}}^l + \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \|\mathbf{w}_{l,k_l,u_k^l}\|^2 \leq P_{\max}^l, \forall l \in \mathcal{L}, \quad (9d)$$

$$\boldsymbol{\theta}_l = \exp(j\varphi_l \odot \mathbf{a}_l), \forall l, \quad (9e)$$

$$\varphi_{l,m} \in (0, 2\pi], \forall l, m, \quad (9f)$$

$$\|\tilde{\mathbf{a}}_m\|_0 \leq 1, a_{l,m} \in \{0, 1\}, \forall l, m, \quad (9g)$$

where the constraint in (9b) satisfies the target rate or QoS requirement for user $u_k^l \in \mathcal{U}_k^l$ while decoding their own signal. The constraint (9c) guarantees fairness through maintaining a reasonable communication rate at each user. The upper bound on the consumed power is represented in (9d). Moreover, (9e) and (9f) denote the actual phase shifts of the selected RIS elements and their feasible range, respectively and (9g) ensures the uniqueness of the selected RIS elements for a given operating band l .

Clearly, the problem in (9) is non-convex due to the coupling between the optimization variables in (9b) and (9c). Moreover, (9e)–(9g) bring additional complexity due to the binary selection variable $\mathbf{A}$. In order to solve these issues, we initialize our solution by assuming that the binary selection matrix is fixed as $\mathbf{A} = \mathbf{1}$ and the ideal phase shifts are randomly chosen according to a uniform distribution over $(0, 2\pi]$ to optimize both the BS beamformers and the ideal phase shifts via SDP and SCA for each cell l . Then, we solve the binary selection problem, which is formulated as a mixed integer nonlinear program (MINLP) using SCA and the penalty-based difference-of-convex approaches to determine the appropriate allocation of RIS elements per cells. Finally, we update the RIS phase shifts and transmit beamformers based on the optimum selection matrix, as shown in details in the following subsections.

$$R_{u_k^l \rightarrow u_k^l} = B_l \log_2 \left(1 + \frac{\left| \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \boldsymbol{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,k_l,u_k^l} \right|^2}{\sum_{i_k^l=u_k^l+1}^{U_k^l} \left| \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \boldsymbol{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,k_l,i_k^l} \right|^2 + \sum_{c_l \neq k_l} \sum_{x_c^l=1}^{U_c^l} \left| \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \boldsymbol{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,c_l,x_c^l} \right|^2 + \sigma^2} \right), \quad (11a)$$

$$= B_l \log_2 \left(1 + \frac{\text{Tr} \left(\mathbf{W}_{u_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l} \right)}{\sum_{i_k^l=u_k^l+1}^{U_k^l} \text{Tr} \left(\mathbf{W}_{i_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l} \right) + \sum_{c_l \neq k_l} \sum_{x_c^l=1}^{U_c^l} \text{Tr} \left(\mathbf{W}_{x_c^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l} \right) + \sigma^2} \right). \quad (11b)$$

3.1. Optimizing transmit beamformers $\mathbf{W}$

Given $\boldsymbol{\theta}_l$ and $\mathbf{A}$, the sub-problem for optimizing the transmit beamformers, $\mathbf{W}_l \triangleq \{\mathbf{w}_{l,k_l,u_k^l}\} \forall k_l \in \mathcal{K}_l, u_k^l \in \mathcal{U}_k^l$, for the l^{th} BS is expressed as

$$\text{Maximize}_{\mathbf{W}_l} \frac{\sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} R_{u_k^l \rightarrow u_k^l}}{P_{\text{cir}}^l + \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \|\mathbf{w}_{l,k_l,u_k^l}\|^2}, \quad (10a)$$

subject to:

$$R_{u_k^l \rightarrow u_k^l} \geq R_{\min}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (10b)$$

$$\left| \mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,\pi(i_k^l)} \right|^2 \leq \left| \mathbf{h}_{u_k^l}^H \mathbf{w}_{l,k_l,\pi(j_k^l)} \right|^2, \quad (10c)$$

$$\forall \pi(i_k^l) > \pi(j_k^l), \{u_k^l, i_k^l, j_k^l\} \in \mathcal{U}_k^l, l, k_l, \quad (10d)$$

$$\sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \|\mathbf{w}_{l,k_l,u_k^l}\|^2 \leq \bar{P}_{\max}^l, \forall l \in \mathcal{L}, \quad (10d)$$

where $\bar{P}_{\max}^l = P_{\max}^l - P_{\text{cir}}^l$, and the achievable rate of user $u_k^l \in \mathcal{U}_k^l$ can be given by (11a), on top of the page. The beamforming problem (10) is a non-convex problem due to the fractional objective (10a) and the non-convex constraints (10b) and (10c). To cope with this, we introduce the auxiliary variables $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_L\}$, $\forall l$ and $\boldsymbol{\beta} = \{\beta_{i_k^l}, \dots, \beta_{U_k^l}\}$, $\forall l, k_l, u_k^l \in \mathcal{U}_k^l$. Then, the transmit beamforming problem for each cell can be reformulated as

$$\text{Maximize}_{\mathbf{W}_l, \boldsymbol{\alpha}, \boldsymbol{\beta}} \alpha_l, \quad (12a)$$

subject to:

$$\frac{\sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} B_l \log_2 (1 + \beta_{u_k^l})}{P_{\text{cir}}^l + \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \|\mathbf{w}_{l,k_l,u_k^l}\|^2} \geq \alpha_l, \forall l, \quad (12b)$$

$$\text{SINR}_{u_k^l \rightarrow u_k^l} \geq \beta_{u_k^l}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (12c)$$

$$B_l \log_2 (1 + \beta_{u_k^l}) \geq R_{\min}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (12d)$$

$$(10c), (10d). \quad (12e)$$

Next, we define $\mathbf{W}_{u_k^l} = \mathbf{w}_{l,k_l,u_k^l} \mathbf{w}_{l,k_l,u_k^l}^H$, $\mathbf{W}_{u_k^l} \in \mathbb{H}^{N_t \times N_t}$, $\mathbf{v}_l =$

$$[e^{j\varphi_{l,1} a_{l,1}}, \dots, e^{j\varphi_{l,M} a_{l,M}}, 1]^T \in \mathbb{C}^{(M+1) \times 1}, \text{ and } \tilde{\mathbf{H}}_{u_k^l} = \begin{bmatrix} \text{diag}(\mathbf{h}_{r,u_k^l}^H) \mathbf{G}_l \\ \mathbf{h}_{d,u_k^l}^H \end{bmatrix}$$

$\in \mathbb{C}^{(M+1) \times N_t}$. Hence, the signal power term $\left| \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \boldsymbol{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,k_l,u_k^l} \right|^2$ in (11a), on top of the page, can be recast as

$$\left| \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \boldsymbol{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,k_l,u_k^l} \right|^2 = \left| \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l} \mathbf{w}_{l,k_l,u_k^l} \right|^2 = \text{Tr} \left(\mathbf{W}_{u_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l} \right). \quad (13)$$

Accordingly, the achievable rate of the user $u_k^l \in \mathcal{U}_k^l$ in (11a) is re-expressed as in (11b), on top of the page, and hence the transmit beamforming optimization problem can be reformulated in the following SDP form

$$\text{Maximize}_{\mathbf{W}_{u_k^l}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \alpha_l, \quad (14a)$$

subject to:

$$\frac{\sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} B_l \log_2 (1 + \beta_{u_k^l})}{P_{\text{cir}}^l + \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \text{Tr}(\mathbf{W}_{u_k^l})} \geq \alpha_l, \forall l, \quad (14b)$$

$$(12c), (12d) \quad (14c)$$

$$\text{Tr} \left(\mathbf{W}_{i_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l} \right) \leq \text{Tr} \left(\mathbf{W}_{j_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l} \right), \quad (14d)$$

$$\sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \text{Tr}(\mathbf{W}_{u_k^l}) \leq \bar{P}_{\max}^l, \forall l, \quad (14e)$$

$$\mathbf{W}_{u_k^l} \succeq 0, \quad \text{Rank}(\mathbf{W}_{u_k^l}) = 1, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (14f)$$

where the constraint (14f) is included to ensure that the optimum beamforming solution satisfies $\mathbf{W}_{u_k^l} = \mathbf{w}_{l,k_l,u_k^l} \mathbf{w}_{l,k_l,u_k^l}^H$. Although the objective of (14a) becomes a linear function, the overall problem is still non-convex due to the constraints (12c), (14b) and (14f). The constraint (14b) can be written as $\sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} R_{u_k^l \rightarrow u_k^l} \geq \alpha_l P_{\text{cir}}^l +$

$\alpha_l \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \text{Tr}(\mathbf{W}_{u_k^l})$, where the LHS is a concave function with respect to $\mathbf{W}_{u_k^l}$, meanwhile the second term in the RHS is jointly convex with respect to $\mathbf{W}_{u_k^l}$ and α_l . Therefore, we invoke the SCA approach to find the lower bound on this term for given feasible

points $(\mathbf{w}_{u_k}^{(t)}, \alpha_l^{(t)})$, in the t^{th} iteration of the SCA, using first-order Taylor approximation (FTA) as follows

$$\begin{aligned} \alpha_l \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \text{Tr}(\mathbf{w}_{u_k^l}^{(t)}) &\geq \alpha_l^{(t)} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \text{Tr}(\mathbf{w}_{u_k^l}^{(t)}) \\ &+ \alpha_l^{(t)} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \text{Tr}(\mathbf{w}_{u_k^l}^{(t)} - \mathbf{w}_{u_k^l}^{(t)}) \\ &+ (\alpha_l - \alpha_l^{(t)}) \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \text{Tr}(\mathbf{w}_{u_k^l}^{(t)}) \triangleq \hat{f}_l(\alpha_l, \mathbf{w}_{u_k^l}^{(t)}). \end{aligned} \quad (15)$$

Regarding constraint (12c), additional auxiliary variables $\{\chi_{u_k^l}\}$, $\forall l, k_l, u_k^l \in \mathcal{U}_k^l$ are introduced as an upper limit for the total interference-plus-noise experienced at the user u_k^l . Hence, the constraint (12c) can be re-written as

$$\text{Tr}(\mathbf{w}_{u_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l}) \geq \beta_{u_k^l} \chi_{u_k^l}, \quad (16)$$

$$\begin{aligned} &\sum_{i_k^l = u_k^l + 1}^{U_k^l} \text{Tr}(\mathbf{w}_{i_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l}) \\ &+ \sum_{c_l \neq k_l} \sum_{u_c^l = 1}^{U_c^l} \text{Tr}(\mathbf{w}_{u_c^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l}) + \sigma^2 \leq \chi_{u_k^l} \end{aligned} \quad (17)$$

Obviously, (16) is non-convex with respect to $\beta_{u_k^l}$ and $\chi_{u_k^l}$. Hence, we again exploit SCA to find a lower bound on the product $\beta_{u_k^l} \chi_{u_k^l}$ at feasible points $\{\beta_{u_k^l}^{(t)}, \chi_{u_k^l}^{(t)}\}$ in the t^{th} iteration as follows

$$\begin{aligned} \beta_{u_k^l} \chi_{u_k^l} &\geq \beta_{u_k^l}^{(t)} \chi_{u_k^l}^{(t)} + \beta_{u_k^l}^{(t)} (\chi_{u_k^l} - \chi_{u_k^l}^{(t)}) \\ &+ \chi_{u_k^l}^{(t)} (\beta_{u_k^l} - \beta_{u_k^l}^{(t)}) \triangleq \hat{f}(\beta_{u_k^l} \chi_{u_k^l}). \end{aligned} \quad (18)$$

Accordingly, the optimization problem in (14) can now be approximated by the following

$$\text{Maximize } \alpha_l, \quad (19a)$$

subject to:

$$\sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} B_l \log_2(1 + \beta_{u_k^l}) \geq \alpha_l P_{\text{cir}}^l + \hat{f}_l(\alpha_l, \mathbf{w}_{u_k^l}^{(t)}), \forall l, \quad (19b)$$

$$\text{Tr}(\mathbf{w}_{u_k^l} \tilde{\mathbf{H}}_{u_k^l}^H \mathbf{v}_l \mathbf{v}_l^H \tilde{\mathbf{H}}_{u_k^l}) \geq \hat{f}(\beta_{u_k^l} \chi_{u_k^l}), \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (19c)$$

$$(12d), (14d), (14e), (17), \quad (19d)$$

$$\mathbf{w}_{u_k^l} \geq 0, \forall l, k_l, u_k^l \in \mathcal{U}_k^l. \quad (19e)$$

Obviously, the problem in (19) is a convex SDP, with relaxed rank-one constraint that can be optimally solved by standard convex optimization tools such as CVX [44]. If the optimal solution of (19) is rank-one, then it will be equivalent to the optimal beamformers of the problem in (12) and can be completely obtained as $\mathbf{w}_{u_k^l} = [\mathbf{w}_{u_k^l}]_{N_t} / \sqrt{[\mathbf{w}_{u_k^l}]_{N_t, N_t}}$ since the matrix $\mathbf{w}_{u_k^l}$ is positive semidefinite, where $[\mathbf{w}_{u_k^l}]_{N_t}$ represents N_t -th column of matrix

$\mathbf{w}_{u_k^l}$ and $[\mathbf{w}_{u_k^l}]_{N_t, N_t}$ is the matrix entry at the N_t -th row and N_t -th column. Otherwise, the non rank-one solution can be approximated by employing the Gaussian randomization (GR) method [45,46].

3.2. RIS phase shifts optimization

In this section, we consider the joint design of the ideal RIS phase shifts Φ and the selection matrix $\mathbf{A}$ for a given transmit set of beamformers $\mathbf{W}$. The RIS phase shift optimization problem is given as

$$\text{Maximize}_{\Phi, \mathbf{A}} \frac{1}{P_{\text{tot}}^l} \sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} R_{u_k^l \rightarrow u_k^l}, \quad (20a)$$

subject to:

$$R_{u_k^l \rightarrow u_k^l} \geq R_{\min}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (20b)$$

$$\begin{aligned} &\left| \mathbf{h}_{u_k^l}^H \mathbf{w}_{l, k_l, \pi(i_k^l)} \right|^2 \leq \left| \mathbf{h}_{u_k^l}^H \mathbf{w}_{l, k_l, \pi(j_k^l)} \right|^2, \\ &\forall \pi(i_k^l) > \pi(j_k^l), \{u_k^l, i_k^l, j_k^l\} \in \mathcal{U}_k^l, l, k_l, \end{aligned} \quad (20c)$$

$$\theta_l = \exp(j\varphi_l \odot \mathbf{a}_l), \forall l, \quad (20d)$$

$$\varphi_{l,m} \in (0, 2\pi], \forall l, m, \quad (20e)$$

$$\|\tilde{\mathbf{a}}_m\|_0 \leq 1, a_{l,m} \in \{0, 1\}, \forall l, m. \quad (20f)$$

We first try to reformulate the optimization problem (20) in terms of θ_l and then decouple it into two sub-problems, which are solved for the tunable phase shifts φ_l and selection matrix $\mathbf{A}$. For convenience, let $\mathbf{h}_{r, u_k^l}^H \Theta_l \mathbf{G}_l \mathbf{w}_{l, k_l, u_k^l} = \theta_l^H \text{diag}\{\mathbf{h}_{r, u_k^l}^H\} \mathbf{G}_l \mathbf{w}_{l, k_l, u_k^l} \triangleq \theta_l^H \mathbf{f}_{r, u_k^l, u_k^l}$, where $\mathbf{f}_{r, u_k^l, u_k^l} = \text{diag}\{\mathbf{h}_{r, u_k^l}^H\} \mathbf{G}_l \mathbf{w}_{l, k_l, u_k^l}$ and $q_{d, u_k^l, u_k^l} = \mathbf{h}_{d, u_k^l}^H \mathbf{w}_{l, k_l, u_k^l}$. Hence, we can rewrite the achievable rate, total interference-plus-noise and fairness constraints in (20a) as (21a), (21b) and (21c), respectively, on top of the next page. Next, we introduce the auxiliary variables $\Gamma = \{\Gamma_{1^l}, \dots, \Gamma_{U^l}\}$ and $\zeta = \{\zeta_{1^l}, \dots, \zeta_{U^l}\}$. The optimization problem (20) can now be recast as

$$\text{Maximize}_{\theta_l, \Gamma, \zeta} \frac{1}{P_{\text{tot}}^l} \sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} B_l \log_2(1 + \Gamma_{u_k^l}), \quad (22a)$$

subject to:

$$\left| q_{d, u_k^l, u_k^l} + \theta_l^H \mathbf{f}_{r, u_k^l, u_k^l} \right|^2 \geq \Gamma_{u_k^l} \zeta_{u_k^l}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (22b)$$

$$B_l \log_2(1 + \Gamma_{u_k^l}) \geq R_{\min}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (22c)$$

$$(20d)-(20f), (21b), (21c). \quad (22d)$$

Problem (22) is non-convex due to the non-convexity of constraints (20d)–(20f), (21c) and (22b). Moreover, the coupling between Φ and $\mathbf{A}$ increases the complexity of the problem. To handle this issue, we seek to decouple (22) into two sub-problems. The first is concerned with solving for Φ and the second for determining $\mathbf{A}$.

3.2.1. Ideal phase shift optimization φ_l

Different from the conventional design of RIS, we observe that (20d), in (22), implies that the actual phase shift of RIS depends on $\varphi_{l,m}$ and $a_{l,m}$ for each cell l and RIS element m . Specifically, optimizing the ideal phase shift $\varphi_{l,m}$ does not impact the objective of (22) when $a_{l,m} = 0$, i.e., $\theta_{l,m} = 1$. Hence, the problem (22) can be solved for only the chosen set of reflecting elements to serve a

$$R_{u_k^l \rightarrow u_k^l} = B_l \log_2 \left(1 + \frac{|q_{d,u_k^l,u_k^l} + \theta_l^H \mathbf{f}_{r,u_k^l,u_k^l}|^2}{\sum_{i_k^l=u_k^l+1}^{U_k^l} (|q_{d,u_k^l,i_k^l} + \theta_l^H \mathbf{f}_{r,u_k^l,i_k^l}|^2) + \sum_{c_l \neq k_l} \sum_{x_c^l=1}^{U_c^l} (|q_{d,u_k^l,x_c^l} + \theta_l^H \mathbf{f}_{r,u_k^l,x_c^l}|^2) + \sigma^2} \right), \quad (21a)$$

$$I(\theta_l) \triangleq \sum_{i_k^l=u_k^l+1}^{U_k^l} (|q_{d,u_k^l,i_k^l} + \theta_l^H \mathbf{f}_{r,u_k^l,i_k^l}|^2) + \sum_{c_l \neq k_l} \sum_{x_c^l=1}^{U_c^l} (|q_{d,u_k^l,x_c^l} + \theta_l^H \mathbf{f}_{r,u_k^l,x_c^l}|^2) + \sigma^2 \leq \zeta_{u_k^l}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (21b)$$

$$|q_{d,u_k^l,\pi(i_k^l)} + \theta_l^H \mathbf{f}_{r,u_k^l,\pi(i_k^l)}|^2 \leq |q_{d,u_k^l,\pi(j_k^l)} + \theta_l^H \mathbf{f}_{r,u_k^l,\pi(j_k^l)}|^2, \forall \pi(i_k^l) > \pi(j_k^l), \{u_k^l, i_k^l, j_k^l\} \in \mathcal{U}_k^l, l, k_l. \quad (21c)$$

$$S(\tilde{\theta}_l) \triangleq |q_{d,u_k^l,u_k^l} + \theta_l^H \mathbf{f}_{r,u_k^l,u_k^l}|^2 = |\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l} + (\mathbf{1}^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l} + q_{d,u_k^l,u_k^l})|^2 = |\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l} + \vartheta_{r,d,u_k^l,u_k^l}|^2 \\ = |\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l}|^2 + |\vartheta_{r,d,u_k^l,u_k^l}|^2 + 2\Re \left\{ \tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l} \vartheta_{r,d,u_k^l,u_k^l}^* \right\} \geq \Gamma_{u_k^l} \zeta_{u_k^l}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (23a)$$

$$I(\tilde{\theta}_l) \triangleq \sum_{i_k^l=u_k^l+1}^{U_k^l} (|\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,i_k^l}|^2 + |\vartheta_{r,d,u_k^l,i_k^l}|^2 + 2\Re \left\{ \tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,i_k^l} \vartheta_{r,d,u_k^l,i_k^l}^* \right\}) \\ + \sum_{c_l \neq k_l} \sum_{x_c^l=1}^{U_c^l} (|\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,x_c^l}|^2 + |\vartheta_{r,d,u_k^l,x_c^l}|^2 + 2\Re \left\{ \tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,x_c^l} \vartheta_{r,d,u_k^l,x_c^l}^* \right\}) + \sigma^2 \leq \zeta_{u_k^l}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l. \quad (23b)$$

$$|\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,i_k^l}|^2 + 2\Re \left\{ \tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,i_k^l} \vartheta_{r,d,u_k^l,i_k^l}^* \right\} + |\vartheta_{r,d,u_k^l,i_k^l}|^2 \leq 2\Re \left\{ (\tilde{\theta}_l^{(t)})^H \tilde{\mathbf{f}}_{r,u_k^l,j_k^l} \tilde{\mathbf{f}}_{r,u_k^l,j_k^l}^H \tilde{\theta}_l \right\} + |\vartheta_{r,d,u_k^l,j_k^l}|^2 \\ + 2\Re \left\{ \tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,j_k^l} \vartheta_{r,d,u_k^l,j_k^l}^* \right\} - |(\tilde{\theta}_l^{(t)})^H \tilde{\mathbf{f}}_{r,u_k^l,j_k^l}|^2, \forall i_k^l > j_k^l, \{u_k^l, i_k^l, j_k^l\} \in \mathcal{U}_k^l, l, k_l, \quad (23c)$$

specific BS, i.e., when $a_{l,m} = 1$. Accordingly, the reflecting elements are split into an active set $\tilde{\mathcal{M}}_l = \{m | a_{l,m} = 1\}$ and idle set $\bar{\mathcal{M}}_l = \{m | a_{l,m} = 0\}$ and hence, the ideal phase shifts are divided into a fully tuned phase shifts' set $\{\tilde{\varphi}_l, \tilde{\theta}_l | \varphi_{l,m}, \theta_{l,m}, \forall m \in \tilde{\mathcal{M}}_l\}$ of the selected elements and the non-selected elements set with phase shifts $\{\bar{\varphi}_l, \bar{\theta}_l = 1 | \varphi_{l,m}, \theta_{l,m}, \forall m \in \bar{\mathcal{M}}_l\}$. Also, the vector $\mathbf{f}_{r,u_k^l,j_k^l}$, appearing in (22), is split into $\tilde{\mathbf{f}}_{r,u_k^l,j_k^l}$ and $\bar{\mathbf{f}}_{r,u_k^l,j_k^l}$. Hence, the problem (22) can be solved for the optimization variable $\{\tilde{\theta}_l\}$, where the corresponding phase shift vector is retrieved by considering $\tilde{\theta}_l = \exp(j\tilde{\varphi}_l)$, i.e. $\tilde{\varphi}_l = \angle \tilde{\theta}_l$.

In more detail, we express the quadratic terms in (22) as a function of $\tilde{\theta}_l$. Therefore, the constraints on the signal power and the interference-pulse-noise of (22b) and (21b), are, respectively, expressed as in (23a) and (23b) on top of the page, where $\vartheta_{r,d,u_k^l,u_k^l} = \mathbf{1}^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l} + q_{d,u_k^l,u_k^l}$. Hence, the optimization problem (22), for BS l , can now be expressed as

$$\text{Maximize}_{\tilde{\theta}_l, \Gamma, \zeta} \frac{1}{P_{\text{tot}}^l} \sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} B_l \log_2 (1 + \Gamma_{u_k^l}), \quad (24a)$$

subject to:

$$(22c), (23a), (23b), \quad (24b)$$

$$|\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,i_k^l}|^2 + 2\Re \left\{ \tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,i_k^l} \vartheta_{r,d,u_k^l,i_k^l}^* \right\} + |\vartheta_{r,d,u_k^l,i_k^l}|^2 \\ \leq |\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,j_k^l}|^2 + |\vartheta_{r,d,u_k^l,j_k^l}|^2 + 2\Re \left\{ \tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,j_k^l} \vartheta_{r,d,u_k^l,j_k^l}^* \right\},$$

$$\forall i_k^l > j_k^l, \{u_k^l, i_k^l, j_k^l\} \in \mathcal{U}_k^l, l, k_l, \quad (24c)$$

$$|\tilde{\theta}_{l,m}| = 1, \forall l, \forall m \in \tilde{\mathcal{M}}_l. \quad (24d)$$

Clearly, the non-convexity of the optimization problem in (24) originates from (23a), (24c) and (24d). To solve this issue, we apply the SCA approach. Specifically, the LHS and RHS of the constraint (23a) are lower bounded, via FTA, at given points $\{\tilde{\theta}_l^{(t)}\}$ and $\{\Gamma_{u_k^l}^{(t)}, \zeta_{u_k^l}^{(t)}\}$ in the iteration t as

$$|\tilde{\theta}_l^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l}|^2 \geq 2\Re \left\{ (\tilde{\theta}_l^{(t)})^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l} \tilde{\mathbf{f}}_{r,u_k^l,u_k^l}^H \tilde{\theta}_l \right\} \\ - |(\tilde{\theta}_l^{(t)})^H \tilde{\mathbf{f}}_{r,u_k^l,u_k^l}|^2 \triangleq \Omega(\tilde{\theta}_l), \quad (25)$$

$$\Gamma_{u_k^l} \zeta_{u_k^l} \geq \Gamma_{u_k^l}^{(t)} \zeta_{u_k^l}^{(t)} + \Gamma_{u_k^l}^{(t)} (\zeta_{u_k^l} - \zeta_{u_k^l}^{(t)}) \\ + \zeta_{u_k^l}^{(t)} (\Gamma_{u_k^l} - \Gamma_{u_k^l}^{(t)}) \triangleq \hat{f}(\Gamma_{u_k^l} \zeta_{u_k^l}), \quad (26)$$

respectively. Similar to (25), we apply FTA to the RHS of the constraint (24c) at given points $\{\tilde{\theta}_l^{(t)}\}$ as shown in (23c) on top of the page. Moreover, the equality constraint (24d) is relaxed to inequality, such that $|\tilde{\theta}_l| \leq 1$, where the optimal solution of $\tilde{\theta}_l$ is necessarily normalized to satisfy unit-modulus phase shifts of the active RIS elements. Hence, the optimization problem in (24) can finally be written as

$$\text{Maximize}_{\boldsymbol{\theta}_l, \boldsymbol{\Gamma}, \boldsymbol{\zeta}} \frac{1}{P_{\text{tot}}} \sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} B_l \log_2 (1 + \Gamma_{u_k^l}), \quad (27a)$$

subject to:

$$(23b), (22c), (23c) \quad (27b)$$

$$\left| \vartheta_{r,d,u_k^l, u_k^l} \right|^2 + 2\Re \left\{ \tilde{\boldsymbol{\theta}}_l^H \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \vartheta_{r,d,u_k^l, u_k^l}^* \right\} + \Omega(\tilde{\boldsymbol{\theta}}_l) \geq \hat{f}(\Gamma_{u_k^l} \zeta_{u_k^l}), \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (27c)$$

$$|\tilde{\boldsymbol{\theta}}_l| \leq 1, \forall l, m \in \tilde{\mathcal{M}}_l. \quad (27d)$$

The sub-problem (27) is a convex optimization problem, which can be solved by, e.g., CVX. The ideal phase shifts of the selected RIS elements are retrieved by firstly calculating the angle of the obtained phase shift matrix, and then the result is normalized as a compensation to the relaxed unit modulus constraint. Meanwhile, the phase shifts of the non-selected elements are chosen as $\{0, 2\pi\}$, i.e.,

$$\varphi_{l,m}^* = \begin{cases} \angle \frac{\tilde{\boldsymbol{\theta}}_l^*}{\|\tilde{\boldsymbol{\theta}}_l\|}, & a_{l,m} = 1, m \in \tilde{\mathcal{M}}_l, \\ \{0, 2\pi\}, & a_{l,m} = 0, m \in \overline{\mathcal{M}}_l. \end{cases} \quad (28)$$

3.2.2. Binary phase selection matrix design $\mathbf{A}$

Given the optimal transmit beamforming and ideal phase shifts for each BS, the binary selection matrix sub-problem is formulated as in (22) after appending the constraint (20f). For the sake of problem tractability, the optimization problem (22) is firstly reformulated in terms of $\mathbf{A}$ by substituting the constraint (20d) into the quadratic forms of (22), e.g., $\left| \left(\mathbf{h}_{d,u_k^l}^H + \mathbf{h}_{r,u_k^l}^H \boldsymbol{\Theta}_l \mathbf{G}_l \right) \mathbf{w}_{l,k_l,u_k^l} \right|^2$ as follows

$$\begin{aligned} & \left| q_{d,u_k^l, u_k^l} + \boldsymbol{\theta}_l^H \mathbf{f}_{r,u_k^l, u_k^l} \right|^2 \\ &= \left| q_{d,u_k^l, u_k^l} + \left(\mathbf{a}_l^T \odot e^{j\varphi_l} \right) \mathbf{f}_{r,u_k^l, u_k^l} + (\mathbf{1} - \mathbf{a}_l)^T \mathbf{f}_{r,u_k^l, u_k^l} \right|^2 \\ &= \left| q_{d,u_k^l, u_k^l} + \mathbf{1}^T \mathbf{f}_{r,u_k^l, u_k^l} + \mathbf{a}_l^T \text{diag}(e^{j\varphi_l} - \mathbf{1}) \mathbf{f}_{r,u_k^l, u_k^l} \right|^2 \\ &= \left| \tilde{q}_{d,u_k^l, u_k^l} + \mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \right|^2 \\ &= \left| \mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \right|^2 + 2\Re \left(\mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \tilde{q}_{d,u_k^l, u_k^l}^* \right) + \left| \tilde{q}_{d,u_k^l, u_k^l} \right|^2, \end{aligned} \quad (29)$$

where $\tilde{q}_{d,u_k^l, u_k^l} = q_{d,u_k^l, u_k^l} + \mathbf{1}^T \mathbf{f}_{r,u_k^l, u_k^l}$, $q_{d,u_k^l, u_k^l} = \mathbf{h}_{d,u_k^l}^H \mathbf{w}_{l,k_l,u_k^l}$, $\mathbf{f}_{r,u_k^l, u_k^l} = \text{diag}(\mathbf{h}_{r,u_k^l}^H) \mathbf{G}_l \mathbf{w}_{l,k_l,u_k^l}$ and $\tilde{\mathbf{f}}_{r,u_k^l, u_k^l} = \text{diag}(e^{j\varphi_l} - \mathbf{1}) \mathbf{f}_{r,u_k^l, u_k^l}$. Substituting (29) into the user signal power constraint (22b), the result becomes

$$\left| \mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \right|^2 + 2\Re \left(\mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \tilde{q}_{d,u_k^l, u_k^l}^* \right) + \left| \tilde{q}_{d,u_k^l, u_k^l} \right|^2 \geq \Gamma_{u_k^l} \zeta_{u_k^l}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (30)$$

which is non-convex with respect to $\mathbf{a}_l$. Hence, the LHS and RHS are approximated at given points $\{\mathbf{a}_l^{(t)}\}$ and $\{\Gamma_{u_k^l}^{(t)}, \zeta_{u_k^l}^{(t)}\}$, similar to (25) and (26), as

$$\left| \mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \right|^2 \geq 2\Re \left\{ \left(\mathbf{a}_l^{(t)} \right)^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \tilde{\mathbf{f}}_{r,u_k^l, u_k^l}^H \mathbf{a}_l \right\} - \left| \left(\mathbf{a}_l^{(t)} \right)^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \right|^2 \triangleq \Omega(\mathbf{a}_l), \quad (31)$$

$$\begin{aligned} \Gamma_{u_k^l} \zeta_{u_k^l} &\geq \Gamma_{u_k^l}^{(t)} \zeta_{u_k^l}^{(t)} + \Gamma_{u_k^l}^{(t)} \left(\zeta_{u_k^l} - \zeta_{u_k^l}^{(t)} \right) \\ &+ \zeta_{u_k^l}^{(t)} \left(\Gamma_{u_k^l} - \Gamma_{u_k^l}^{(t)} \right) \triangleq \hat{f}(\Gamma_{u_k^l} \zeta_{u_k^l}). \end{aligned} \quad (32)$$

Accordingly, the constraint (22b) can be reformulated as

$$\Omega(\mathbf{a}_l) + 2\Re \left(\mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, u_k^l} \tilde{q}_{d,u_k^l, u_k^l}^* \right) + \left| \tilde{q}_{d,u_k^l, u_k^l} \right|^2 \geq \hat{f}(\Gamma_{u_k^l} \zeta_{u_k^l}), \forall l, k_l, u_k^l \in \mathcal{U}_k^l, \quad (33)$$

which is convex. Similarly, the total interference-plus-noise constraint (21b) is reformulated in terms of $\mathbf{a}_l$ as in (34) on top of the next page. Meanwhile, the fairness constraint (21c) can be approximated using FTA as

$$\begin{aligned} & \left| \mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, i_k^l} \right|^2 + 2\Re \left(\mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, i_k^l} \tilde{q}_{d,u_k^l, i_k^l}^* \right) + \left| \tilde{q}_{d,u_k^l, i_k^l} \right|^2 \leq \\ & 2\Re \left\{ \left(\mathbf{a}_l^{(t)} \right)^T \tilde{\mathbf{f}}_{r,u_k^l, j_k^l} \tilde{\mathbf{f}}_{r,u_k^l, u_k^l}^H \mathbf{a}_l \right\} - \left| \left(\mathbf{a}_l^{(t)} \right)^T \tilde{\mathbf{f}}_{r,u_k^l, j_k^l} \right|^2 \\ & + 2\Re \left(\mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l, j_k^l} \tilde{q}_{d,u_k^l, j_k^l}^* \right) + \left| \tilde{q}_{d,u_k^l, j_k^l} \right|^2, \forall i_k^l > j_k^l, \\ & \left\{ u_k^l, i_k^l, j_k^l \right\} \in \mathcal{U}_k^l, l, k_l. \end{aligned} \quad (35)$$

Moreover, the constraint (20f), in (20), has an l -0 norm in its LHS, which is equivalent to an l -1 norm due to the fact that the $\mathbf{A}$ is binary. Therefore, we have,

$$\|\tilde{\mathbf{a}}_m\|_0 = \|\tilde{\mathbf{a}}_m\|_1 = \mathbf{1}^T \tilde{\mathbf{a}}_m \leq 1, \forall m, \quad (36)$$

which is an affine constraint. However, $a_{l,m} \in \{0, 1\}$ is non-convex. Hence, we invoke the DC method [47] to recast it as

$$\sum_{l \in \mathcal{L}} \left(\mathbf{1}^T \mathbf{a}_l - \mathbf{a}_l^T \mathbf{a}_l \right) \leq 0, \quad (37)$$

$$0 \leq a_{l,m} \leq 1, \forall l, m. \quad (38)$$

It is obvious that constraint (37) is non-convex, which could be tackled by applying a penalty factor $\mu > 0$ and move it to the objective (20a). Doing so, we obtain the following objective

$$\begin{aligned} & \frac{1}{P_{\text{tot}}} \sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} B_l \log_2 (1 + \Gamma_{u_k^l}) \\ & - \mu \sum_{l \in \mathcal{L}} \left(\mathbf{1}^T \mathbf{a}_l - \mathbf{a}_l^T \mathbf{a}_l \right), \end{aligned} \quad (39)$$

which is still non-convex with respect to $\mathbf{a}_l$ due to the non-concavity of the penalized quadratic term. To resolve this, we employ FTA to find the lower bound on the term $\mu \sum_{l \in \mathcal{L}} (\mathbf{1}^T \mathbf{a}_l - \mathbf{a}_l^T \mathbf{a}_l)$

at given points $\{\mathbf{a}_l^{(t)}\}$ as

$$\mu \sum_{l \in \mathcal{L}} \left(\mathbf{1}^T \mathbf{a}_l - \mathbf{a}_l^T \mathbf{a}_l \right) \geq \mu \sum_{l \in \mathcal{L}} \bar{f}_l(\mathbf{a}_l), \quad (40)$$

where $\bar{f}_l(\mathbf{a}_l) \triangleq \mathbf{1}^T \mathbf{a}_l - \left(\mathbf{a}_l^{(t)} \right)^T \mathbf{a}_l^{(t)} - 2 \left(\mathbf{a}_l^{(t)} \right)^T (\mathbf{a}_l - \mathbf{a}_l^{(t)})$.

By plugging (33)–(36) and (38)–(40) into (20), the binary selection problem is finally approximated as

$$\begin{aligned}
I(\mathbf{a}_l) \triangleq & \sum_{i_k^l=u_k^l+1}^{U_k^l} \left(\left| \mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l,i_k^l} \right|^2 + 2\Re \left(\mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l,i_k^l} \tilde{\mathbf{q}}_{d,u_k^l,i_k^l}^* \right) + \left| \tilde{\mathbf{q}}_{d,u_k^l,i_k^l} \right|^2 \right) \\
& + \sum_{c_l \neq k_l} \sum_{x_c^l=1}^{U_c^l} \left(\left| \mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l,x_c^l} \right|^2 + 2\Re \left(\mathbf{a}_l^T \tilde{\mathbf{f}}_{r,u_k^l,x_c^l} \tilde{\mathbf{q}}_{d,u_k^l,x_c^l}^* \right) + \left| \tilde{\mathbf{q}}_{d,u_k^l,x_c^l} \right|^2 \right) + \sigma^2 \leq \zeta_{u_k^l}, \forall l, k_l, u_k^l \in \mathcal{U}_k^l
\end{aligned} \quad (34)$$

$$\begin{aligned}
& \text{Maximize}_{\mathbf{A}, \Gamma, \zeta} \sum_{l \in \mathcal{L}} \sum_{k_l \in \mathcal{K}_l} \sum_{u_k^l \in \mathcal{U}_k^l} \frac{B_l}{P_{\text{tot}}^l} \log_2 (1 + \Gamma_{u_k^l}) \\
& - \mu \sum_{l \in \mathcal{L}} \bar{f}_l(\mathbf{a}_l), \quad (41a)
\end{aligned}$$

subject to:

$$(22c), (33)–(36), (38), \quad (41b)$$

which is a convex optimization problem and can be solved using CVX.

3.3. Overall optimization framework

Algorithm 1 summarizes the proposed alternating optimization approach to find the optimal values of $\mathbf{W}_{u_k^l}$, ϕ_l , $\mathbf{A}$, given the initialization $\mathbf{A} = \mathbf{1}$, where $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(t)}$ denote the initial and the t -th iteration value of variable $\mathbf{x}$. Lines 2 to 6 in the algorithm solve the optimization problem corresponding to the transmit beamforming and RIS reflection coefficients for each cell l , followed by evaluating the binary variable of the allocated RIS elements. Based on the optimal $\mathbf{A}$, the variables $\mathbf{W}_{u_k^l}$, ϕ_l are updated one more time, as in line 8. Finally, the NOMA decoding order is updated based on the optimal θ_l as in line 9. The convergence of the iterative algorithm is guaranteed due to the non-increasing and bounded properties of the variables $\mathbf{W}_{u_k^l}$, ϕ_l , $\mathbf{A}$ [48] as will be demonstrated in the following section.

Algorithm 1 Joint optimization algorithm of the multi-cell RIS-assisted energy-efficient NOMA system.

- 1: Set the initial values $\mathbf{W}_{u_k^l}^{(0)}$, $\alpha^{(0)}$, $\beta^{(0)}$, $\tilde{\theta}_l^{(0)}$, $\Gamma^{(0)}$, $\zeta^{(0)}$ and iteration index $t = 1$.
- 2: **repeat**
- 3: **for each** $l \in \mathcal{L}$ **do**
- 4: Calculate $\mathbf{W}_{u_k^l}^{(t)}$ by solving (19).
- 5: Obtain $\tilde{\theta}_l^{(t)}$ by solving (27).
- 6: Construct $\phi_{l,m}^{(t)}$ using (28).
- 7: Find $\mathbf{A}^{(t)}$ by solving (41).
- 8: **end for**
- 9: Update NOMA decoding order as in [29, Eq. (33)].
- 10: $t = t + 1$.
- 11: **until** Convergence

3.4. Algorithm complexity analysis

Finally, it is important to study the computational complexity of the proposed algorithm. According to Algorithm 1, the main optimization problem in (9) can be efficiently solved through iteratively solving the three subproblems using the CVX toolbox in MATLAB until it reaches a preset accuracy ν . The complexity of applying the interior point method to solve the active beamforming optimization subproblem, in (19), is in the

order of $O \left(\sum_{l=1}^L \sum_{k_l=1}^{K_l} \sum_{u_k^l=1}^{U_k^l} (N_t l k_l u_k^l)^{4.5} \log_2(1/\nu) \right)$ [46]. Moreover,

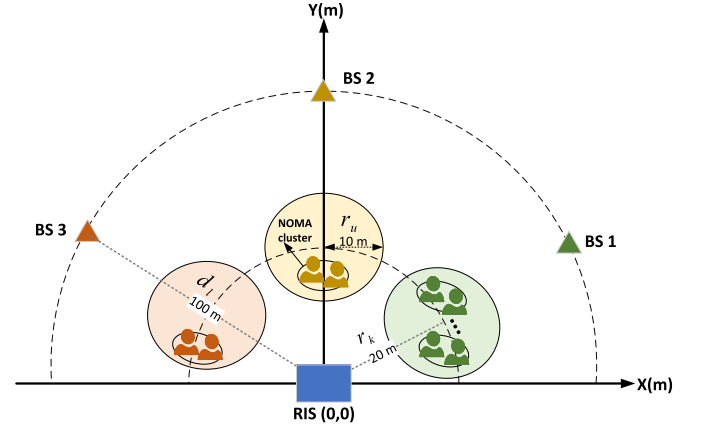


Fig. 2. A 2D illustration of the proposed model setup.

the complexity of applying SCA for solving the binary selection and ideal phase shift problems, in (41) and (27) is given as $O(2(LM)^{3.5} \log_2(1/\nu))$ [49], respectively. Hence, the overall complexity of our algorithm is in the order of

$$O \left(\left(\sum_{l=1}^L \sum_{k_l=1}^{K_l} \sum_{u_k^l=1}^{U_k^l} (N_t l k_l u_k^l)^{4.5} \right) + 2(LM)^{3.5} \right) \log^2 \left(\frac{1}{\nu} \right).$$

4. Simulation results

In this section, we examine the performance of the proposed Algorithm 1 via numerical simulations, where the EE performance is averaged over 500 Monte Carlo realizations. All the simulations are implemented on a computer with processor Ryzen 7 5800H, GPU NVIDIA RTX 3050 Ti 4GB, Matlab Version 2022b and CVX toolbox Version 2.2. We consider a 2D model of the multi-cell RIS-assisted NOMA system as shown in Fig. 2. In this scenario, the RIS is located at (0, 0) and equipped with $M = 20$ elements. The BSs' locations are assumed to follow the standard hexagonal deployment at a distance $d_l = 100$ measured from the origin. The users are randomly distributed in a circle of radius $r_u = 10$ m, which is located at $r_k = 20$ m away from the origin. Moreover, we assume that the system consists of 3 cells, where 4 NOMA users are clustered into 2 clusters. Each BS is equipped with $N_t = 6$ antennas, and each user has a single antenna. The noise power is assumed $\sigma^2 = -80$ dBm and the channel gain between two points, a and b follows a Rayleigh model with mean value $\mathbb{E}[|h_{a,b}|^2] = \varpi \left(\frac{d_{a,b}}{d_0} \right)^{\alpha_{a,b}}$ [50], where $\varpi = -20$ dB expresses the signal attenuation at the reference distance $d_0 = 1$ m and the $d_{a,b}$ denotes the distance between points a and b , and $\alpha_{a,b}$ is the pathloss exponent. The pathloss exponents of the BS-user, BS-RIS, and RIS-user links are assumed 3.6, 2, and 2, respectively [7,51] for each cell. The minimum target NOMA rate and the maximum power budget are set as $R_{\min}^l = 1$ bps/Hz, $\forall l, k_l, u_k^l \in \mathcal{U}_k^l$ and $P_{\max}^l = 25$ dBm, $\forall l$, respectively.

In the following, we compare the EE performance of the proposed system against the conventional OMA-based system. More-

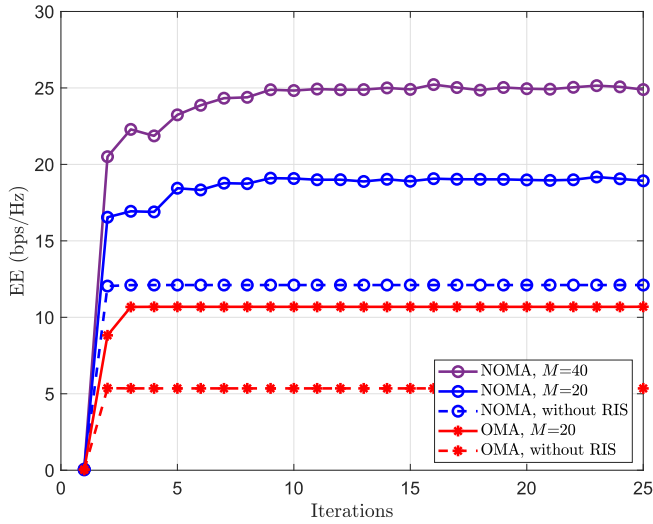


Fig. 3. Convergence of the proposed NOMA and OMA systems versus the number of iterations with the existence and absence of RIS at $P_{\max}^l = 25$ dBm, $P_{\text{cir}}^l = 18$ dBm and $N_t = 6$.

over, both access techniques will be investigated for two scenarios; with and without the RIS. It is worth mentioning that the OMA-based system follows a similar problem formulation as in (9) excluding the SIC decoding and clustering-based fairness NOMA constraints, where the total EE is calculated as follows:

$$\eta_{\text{EE}}^{\text{OMA}} = \frac{\sum_{l \in \mathcal{L}} \sum_{u_l \in \mathcal{U}_l} \delta_l \log_2 \left(1 + \frac{(\mathbf{h}_{d,u_l}^H + \mathbf{h}_{r,u_l}^H \mathbf{\Theta}_l \mathbf{G}_l) \mathbf{w}_{u_l}}{\delta_l \sigma^2} \right)}{\sum_{l \in \mathcal{L}} P_{\text{cir}}^l + \sum_{l \in \mathcal{L}} \sum_{u_l \in \mathcal{U}_l} \|\mathbf{w}_{u_l}\|^2},$$

where $\delta_l = \frac{B_l}{B}$ denotes the frequency allocation factor for each user u_l in cell l . Also, we deploy conventional OMA.

Fig. 3 shows the convergence of the proposed algorithm. For the case with RIS, the OMA-based system converges faster than the proposed NOMA one, where the OMA system requires 3 iterations compared with 10 iterations in case of NOMA. The reason is that the solution space of the proposed system is enlarged by the SIC decoding constraints. However, the proposed system outperforms the conventional OMA by 78% in case of deploying the RIS with $M = 20$. Also, the EE performance gap increases in the case of no RIS. Moreover, the NOMA- and OMA-based systems show faster convergence without deploying the RIS, where the EE maximization problem is solved without considering the actual phase shift constraints and only relies on the transmit beamforming design sub-problem.

The EE performance of both NOMA and OMA versus the number of RIS elements is then investigated in Fig. 4. It is obvious that the EE improves with increasing the number of RIS elements for both systems. Moreover, the RIS provide better EE compared with the baseline schemes with no RIS. Specifically, the NOMA system shows about 65% EE improvement against 100% in the OMA-based model when the number of RIS elements is increased from 20 to 70 elements. This is because as the number of RIS elements increases, the inter- and intra-cluster interference in NOMA restricts the passive RIS array gain and hence, limits the EE improvement. Moreover, the performance gap between NOMA and OMA starts at 80% for 20 elements and decreases to 47% when the number of RIS elements is increased to 70 since the RIS DoFs are fully exploited to enhance the users' signals for the OMA case. In the NOMA case, on the other hand, the trade off between limiting the total interference and energy consumption hinders the EE growth.

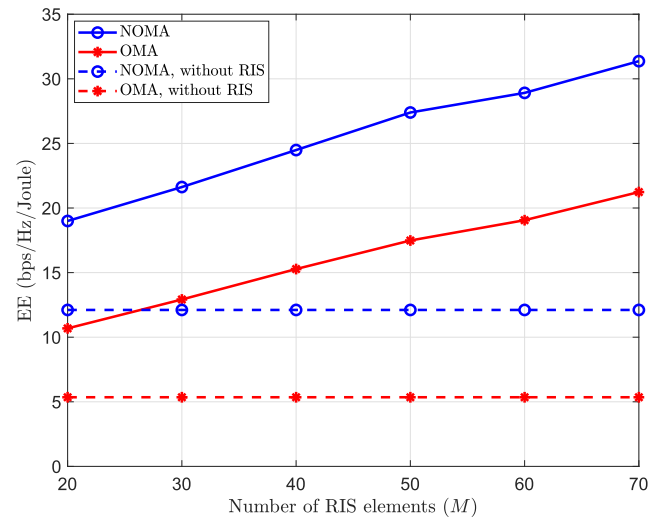


Fig. 4. EE performance versus the number of RIS elements M with $P_{\max}^l = 25$ dBm, $P_{\text{cir}}^l = 18$ dBm and $N_t = 6$.

However, a larger number of RIS elements is required for the OMA system to approach the NOMA performance, e.g., the OMA system requires 60 elements to approximately achieve the NOMA EE performance with 20 elements, which emphasizes the NOMA efficiency in multi-cell multi-band communications.

Next, Figs. 5a and 5b, respectively, demonstrate the SE, which is defined as the achievable rate per unit bandwidth and EE of the proposed NOMA system and the conventional OMA against the maximum allowed transmission power for different number of RIS elements. It is seen from Fig. 5a that the NOMA system can achieve better SE than OMA when the maximum transmission power is less than 31 dBm. However, the intra- and inter-cluster interference increases with $P_{\max}^l$, which hinders the SE improvement when the users share the same resources in NOMA. Specifically, NOMA is 25.4% more spectral efficient than OMA when $M = 40$ and $P_{\max}^l < 31$ dBm. On the other hand, OMA shows an SE improvement of 1.78 bps/Hz from that achieved by NOMA when $P_{\max}^l$ increases and reaches 35 dBm. Moreover, increasing the number of RIS elements, from 20 to 40, shows a smaller performance improvement in NOMA than in OMA along the range of $P_{\max}^l$. For instance, when $P_{\max}^l = 29$ dBm, NOMA's performance improvement is about 1.16 bps/Hz when M increases from 20 to 40 elements, meanwhile this gap is about 1.4 bps/Hz for the OMA scenario. This is because the intra- and inter-cluster interference limit the passive array gain of RIS in the NOMA scenario, as explained in Fig. 4. Moving to the EE performance in Fig. 5b, we can observe that EE improves with $P_{\max}^l$ and then decays with it in the case of NOMA. The reason is that the BSs exploit the maximum available transmission power to enhance the SE of NOMA users (in the numerator of the EE in (7)), where SE is an increasing function in the transmission powers $\|\mathbf{w}_{l,k_l,u_l}\|^2$. In Fig. 5a, it can be observed that as the allowed maximum transmission power increases to a specific value, such as $P_{\max}^l = 31$ dBm, the SE exhibits a logarithmic growth, which has a slower rate of increase compared to the power consumption in the denominator of (7). That is why the overall EE performance exhibits the observed behavior in Fig. 5b. On the other hand, the OMA EE is monotonically increasing with $P_{\max}^l$ as the SE grows exponentially for the same range of $P_{\max}^l$, as exhibited in Fig. 5a. Consequently, Fig. 5 reflects the optimal upper bound on each BS transmission power that maximizes the total EE of the proposed NOMA system.

The impact of the RIS-users distance r_k on the EE performance is then illustrated in Fig. 6. It can be noticed that when the cluster

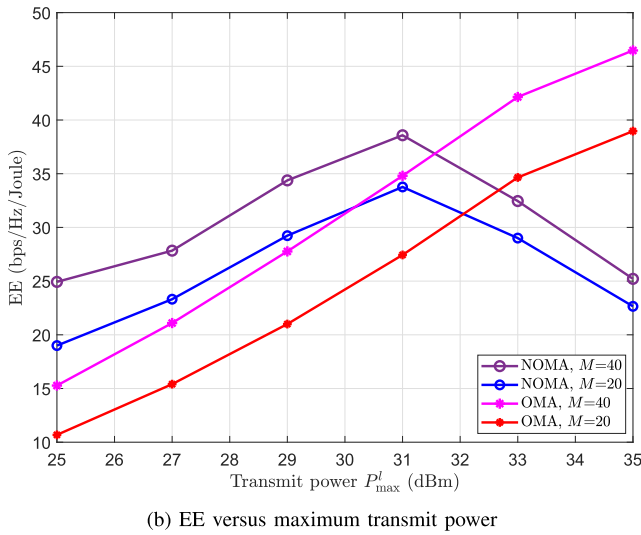
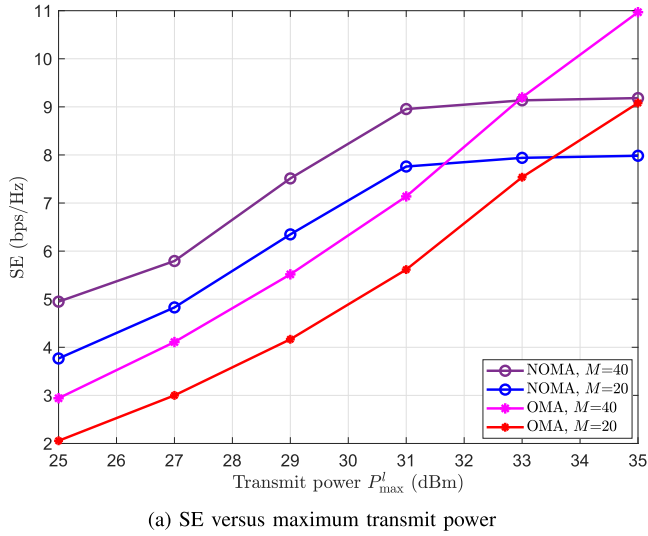


Fig. 5. SE and EE performance of the proposed NOMA and OMA systems versus the maximum allowed transmission power with different number of RIS elements, $P_{\text{cir}}^l = 18$ dBm, $P_{\text{max}}^l = 25$ dBm and $N_t = 6$.

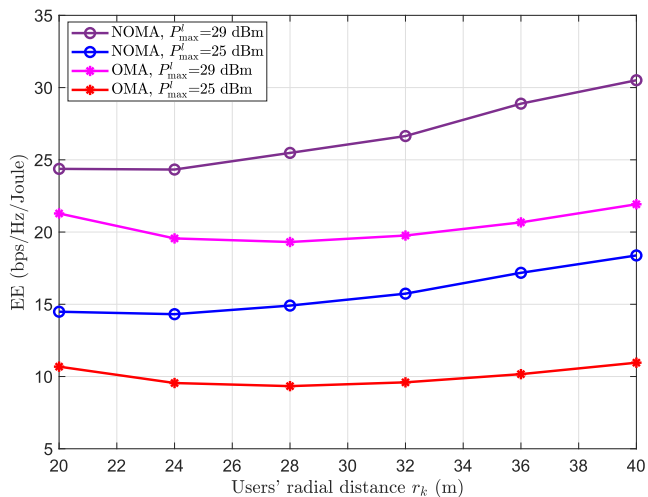


Fig. 6. EE performance versus the radial distance r_k at different P_{max}^l , $P_{\text{cir}}^l = 18$ dBm, $M = 20$ and $N_t = 6$.

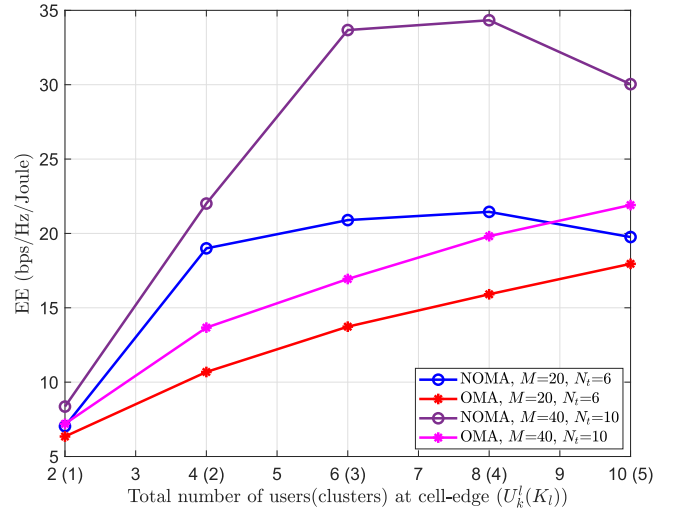


Fig. 7. EE performance versus number of OMA users/NOMA clusters with $P_{\text{cir}}^l = 18$ dBm and $P_{\text{max}}^l = 25$ dBm.

of users moves towards the BS, EE first decreases marginally, then increases with r_k . The reason is that the reflected signal by the RIS is attenuated when the users clusters are shifted away from the RIS, which weakens the RIS array gain. With further increment of r_k , the users become near to a BS, which enhances their received signals and hence, the corresponding EE.

Finally, Fig. 7 shows the effect of increasing the number of NOMA clusters and OMA users on the EE performance. We can observe the increasing EE performance gap that can be achieved by the proposed NOMA system over the OMA one due to the MIMO-NOMA multiplexing gain in the power domain. Specifically, increasing the number of transmitting antennas such that $K_l \leq N_t$ provides spatial DoFs to multiplex more NOMA clusters whereas the users within each cluster are multiplexed in the power domain. This achieves better EE and SE performance thanks to the two-dimensional (spatial and power domain) multiplexing of NOMA when compared with OMA that supports only one-dimension multiplexing of frequency/time. However, increasing the number of NOMA clusters beyond a certain limit, e.g., $K_l > 4$, reduces the EE performance due to the following facts. First, the inter-distance between the NOMA users, in the cell-edge area of radius r_u , shrinks, therefore the channel gain differences start to disappear, which leads to an inferior SIC decoding process. Second, the intra- and inter-cluster interference as well as the power consumption increase, which hinders the exponential growth of EE. On the other hand, increasing the users in OMA enhances the SE and EE as the multiplexed OMA users communicate independently in separate sub bands without inter- or intra-interference issues.

5. Conclusion

In this paper, we studied the impact of RIS deployment in NOMA-enabled multi-cell communications on energy efficiency. Based on this model, we designed the transmit beamforming at each BS, the reflection coefficient matrix of the RIS, and the selection matrix, which specifies unique RIS elements for each cell/BS, to maximize the total EE. Due to the coupling among the design variables, we adopted an efficient iterative algorithm, in which SDR was exploited to optimize the transmit beamforming and SCA as well as penalty-based DC programming were invoked to find the actual RIS phase shifts of the selected RIS elements. Numerical results showed a significant EE improvement in the proposed NOMA communication system when compared with the conventional OMA. Moreover, the results revealed the critical transmis-

sion power level for NOMA beyond which the EE starts decaying. Besides, the EE performance of the proposed NOMA system was shown to deteriorate after the number of clusters at each cell-edge exceeds a certain limit, where similar channel conditions are likely to hinder the NOMA potential.

Although we presented results for a specific scenario, the proposed model can be generalized and applied for different communication scenarios with different objectives but share the same assumptions and setup as our proposed model. For instance, minimizing the power consumption for NOMA-enabled IoT networks, where different operators deploy IoT devices communicating over different frequency bands to collect environmental or sensing information. Another example is coexistence in communication-radar systems, where specific RIS elements can be adjusted for maximizing the radar sensing functionality, while the required QoS of the communication users can be achieved by a different group of RIS elements. Hence, a different waveform can be independently designed based on operational frequencies.

Moreover, there are many issues of the proposed RIS-based model that are worth studying in future work. To illustrate, this model can be exploited in future fast channel estimation thanks to the associated low overhead corresponding to the partially selected RIS elements for each BS/cell. In particular, the number of reflected channels requiring estimation reduces significantly, which reduces the complexity of channel estimation as well as the errors in CSI. Another potential aspect for future research is the integration of machine learning techniques to dynamically update the optimization variables in real-time, considering the continuous variations in the associated channels, particularly in highly-correlated NOMA channel environments.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: A. Abdelaziz Salem reports financial support was provided by American University of Sharjah FRG20-M-E10 and FRG22-C-E13. A. Abdelaziz Salem reports financial support was provided by National Science Foundation CNS-2144297.

Data availability

Data will be made available on request.

References

- [1] J.G. Andrews, S. Buzzi, W. Choi, S.V. Hanly, A. Lozano, A.C. Soong, J.C. Zhang, What will 5G be?, *IEEE J. Sel. Areas Commun.* 32 (6) (2014) 1065–1082.
- [2] S. Dang, O. Amin, B. Shihada, M.-S. Alouini, What should 6G be?, *Nat. Electron.* 3 (1) (2020) 20–29.
- [3] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, Z. Wang, Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends, *IEEE Commun. Mag.* 53 (9) (2015) 74–81.
- [4] S.H. Chae, K. Lee, Cooperative communication for the rank-deficient MIMO interference channel with a reconfigurable intelligent surface, *IEEE Trans. Wirel. Commun.* (2022).
- [5] H. Lu, D. Zhao, Y. Wang, C. Kong, W. Chen, Joint power control and passive beamforming in reconfigurable intelligent surface assisted user-centric networks, *IEEE Trans. Commun.* (2022).
- [6] A. Yadav, C. Quan, P.K. Varshney, H.V. Poor, On performance comparison of multi-antenna HD-NOMA, SCMA, and PD-NOMA schemes, *IEEE Wirel. Commun. Lett.* 10 (4) (2020) 715–719.
- [7] Q. Wu, R. Zhang, Towards smart and reconfigurable environment: intelligent reflecting surface aided wireless network, *IEEE Commun. Mag.* 58 (1) (2019) 106–112.
- [8] C. Huang, S. Hu, G.C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. Di Renzo, M. Debbah, Holographic MIMO surfaces for 6G wireless networks: opportunities, challenges, and trends, *IEEE Wirel. Commun.* 27 (5) (2020) 118–125.
- [9] Y. Liu, X. Mu, X. Liu, M. Di Renzo, Z. Ding, R. Schober, Reconfigurable intelligent surface-aided multi-user networks: interplay between NOMA and RIS, *IEEE Wirel. Commun.* 29 (2) (2022) 169–176.
- [10] S. Zhou, W. Xu, K. Wang, M. Di Renzo, M.-S. Alouini, Spectral and energy efficiency of IRS-assisted MISO communication with hardware impairments, *IEEE Wirel. Commun. Lett.* 9 (9) (2020) 1366–1369.
- [11] C. Huang, A. Zappone, G.C. Alexandropoulos, M. Debbah, C. Yuen, Reconfigurable intelligent surfaces for energy efficiency in wireless communication, *IEEE Trans. Wirel. Commun.* 18 (8) (2019) 4157–4170.
- [12] Z. Li, M. Hua, Q. Wang, Q. Song, Weighted sum-rate maximization for multi-IRS aided cooperative transmission, *IEEE Wirel. Commun. Lett.* 9 (10) (2020) 1620–1624.
- [13] P. Wang, J. Fang, L. Dai, H. Li, Joint transceiver and large intelligent surface design for massive MIMO mmWave systems, *IEEE Trans. Wirel. Commun.* 20 (2) (2020) 1052–1064.
- [14] L. Lei, D. Yuan, C.K. Ho, S. Sun, Power and channel allocation for non-orthogonal multiple access in 5G systems: tractability and computation, *IEEE Trans. Wirel. Commun.* 15 (12) (2016) 8580–8594.
- [15] F. Fang, H. Zhang, J. Cheng, V.C. Leung, Energy-efficient resource allocation for downlink non-orthogonal multiple access network, *IEEE Trans. Commun.* 64 (9) (2016) 3722–3732.
- [16] F. Liu, M. Petrova, Dynamic power allocation for downlink multi-carrier NOMA systems, *IEEE Commun. Lett.* 22 (9) (2018) 1930–1933.
- [17] J. Cui, Y. Liu, Z. Ding, P. Fan, A. Nallanathan, Optimal user scheduling and power allocation for millimeter wave NOMA systems, *IEEE Trans. Wirel. Commun.* 17 (3) (2017) 1502–1517.
- [18] Z. Ding, P. Fan, H.V. Poor, Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions, *IEEE Trans. Veh. Technol.* 65 (8) (2015) 6010–6023.
- [19] J. Cui, Z. Ding, P. Fan, Outage probability constrained MIMO-NOMA designs under imperfect CSI, *IEEE Trans. Wirel. Commun.* 17 (12) (2018) 8239–8255.
- [20] H. Wang, Z. Shi, Y. Fu, S. Fu, On intelligent reflecting surface assisted NOMA uplinks with imperfect SIC, *IEEE Wirel. Commun. Lett.* (2022).
- [21] P. Qin, Y. Fu, X. Zhao, K. Wu, J. Liu, M. Wang, Optimal task offloading and resource allocation for C-NOMA heterogeneous air-ground integrated power Internet of things networks, *IEEE Trans. Wirel. Commun.* 21 (11) (2022) 9276–9292.
- [22] P. Qin, X. Wu, Z. Cai, X. Zhao, Y. Fu, M. Wang, S. Geng, Joint trajectory plan and resource allocation for UAV-enabled C-NOMA in air-ground integrated 6G heterogeneous network, *IEEE Trans. Netw. Sci. Eng.* (2023).
- [23] J. Cui, Y. Liu, Z. Ding, P. Fan, A. Nallanathan, QoS-based resource allocation for multi-cell NOMA networks, *IEEE Trans. Wirel. Commun.* 17 (9) (2018) 6160–6176.
- [24] L. You, D. Yuan, L. Lei, S. Sun, S. Chatzinotas, B. Ottersten, Resource optimization with load coupling in multi-cell NOMA, *IEEE Trans. Wirel. Commun.* 17 (7) (2018) 4735–4749.
- [25] Y. Liu, X. Li, F.R. Yu, H. Ji, H. Zhang, V.C. Leung, Grouping and cooperating among access points in user-centric ultra-dense networks with non-orthogonal multiple access, *IEEE J. Sel. Areas Commun.* 35 (10) (2017) 2295–2311.
- [26] Y. Fu, Y. Chen, C.W. Sung, Distributed power control for the downlink of multi-cell NOMA systems, *IEEE Trans. Wirel. Commun.* 16 (9) (2017) 6207–6220.
- [27] L. Lei, L. You, Y. Yang, D. Yuan, S. Chatzinotas, B. Ottersten, Load coupling and energy optimization in multi-cell and multi-carrier NOMA networks, *IEEE Trans. Veh. Technol.* 68 (11) (2019) 11323–11337.
- [28] J. Zhao, Y. Liu, K.K. Chai, A. Nallanathan, Y. Chen, Z. Han, Spectrum allocation and power control for non-orthogonal multiple access in HetNets, *IEEE Trans. Wirel. Commun.* 16 (9) (2017) 5825–5837.
- [29] W. Ni, X. Liu, Y. Liu, H. Tian, Y. Chen, Resource allocation for multi-cell IRS-aided NOMA networks, *IEEE Trans. Wirel. Commun.* 20 (7) (2021) 4253–4268.
- [30] H. Wang, C. Liu, Z. Shi, Y. Fu, R. Song, Power minimization for two-cell IRS-aided NOMA systems with joint detection, *IEEE Commun. Lett.* 25 (5) (2020) 1635–1639.
- [31] F. Guo, H. Lu, X. Jiang, M. Zhang, J. Wu, C.W. Chen, QoS-aware user grouping strategy for downlink multi-cell NOMA systems, *IEEE Trans. Wirel. Commun.* 20 (12) (2021) 7871–7887.
- [32] X. Xie, F. Fang, Z. Ding, Joint optimization of beamforming, phase-shifting and power allocation in a multi-cluster IRS-NOMA network, *IEEE Trans. Veh. Technol.* 70 (8) (2021) 7705–7717.
- [33] G. Li, H. Zhang, Y. Wang, Y. Xu, QoS guaranteed power minimization and beamforming for IRS-assisted NOMA systems, *IEEE Wirel. Commun. Lett.* (2022).
- [34] Y. Wu, F. Zhou, W. Wu, Q. Wu, R.Q. Hu, K.-K. Wong, Multi-objective optimization for spectrum and energy efficiency tradeoff in IRS-assisted CRNs with NOMA, *IEEE Trans. Wirel. Commun.* 21 (8) (2022) 6627–6642.
- [35] W. Cai, R. Liu, M. Li, Y. Liu, Q. Wu, Q. Liu, IRS-assisted multi-cell multi-band systems: practical reflection model and joint beamforming design, *IEEE Trans. Commun.* (2022).
- [36] S. Lin, B. Zheng, G.C. Alexandropoulos, M. Wen, F. Chen, et al., Adaptive transmission for reconfigurable intelligent surface-assisted OFDM wireless communications, *IEEE J. Sel. Areas Commun.* 38 (11) (2020) 2653–2665.

- [37] G.C. Alexandropoulos, E. Vlachos, A hardware architecture for reconfigurable intelligent surfaces with minimal active elements for explicit channel estimation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 9175–9179.
- [38] M. Jian, G.C. Alexandropoulos, E. Basar, C. Huang, R. Liu, Y. Liu, C. Yuen, Reconfigurable intelligent surfaces for wireless communications: overview of hardware designs, channel models, and estimation techniques, *Intell. Converg. Netw.* 3 (1) (2022) 1–32.
- [39] F. Yang, R. Deng, S. Xu, M. Li, Design and experiment of a near-zero-thickness high-gain transmit-reflect-array antenna using anisotropic metasurface, *IEEE Trans. Antennas Propag.* 66 (6) (2018) 2853–2861.
- [40] S. Abeywickrama, R. Zhang, Q. Wu, C. Yuen, Intelligent reflecting surface: practical phase shift model and beamforming optimization, *IEEE Trans. Commun.* 68 (9) (2020) 5849–5863.
- [41] W. Cai, H. Li, M. Li, Q. Liu, Practical modeling and beamforming for intelligent reflecting surface aided wideband systems, *IEEE Commun. Lett.* 24 (7) (2020) 1568–1571.
- [42] S. Ali, E. Hossain, D.I. Kim, Non-orthogonal multiple access (NOMA) for down-link multiuser MIMO systems: user clustering, beamforming, and power allocation, *IEEE Access* 5 (2016) 565–577.
- [43] Y. Liu, Z. Qin, M. El-kashlan, Z. Ding, A. Nallanathan, L. Hanzo, Non-orthogonal multiple access for 5G and beyond, *Proc. IEEE* 105 (12) (2017) 2347–2381.
- [44] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, 2014.
- [45] M. Cui, G. Zhang, R. Zhang, Secure wireless communication via intelligent reflecting surface, *IEEE Wirel. Commun. Lett.* 8 (5) (2019) 1410–1414.
- [46] Z.-Q. Luo, W.-K. Ma, A.M.-C. So, Y. Ye, S. Zhang, Semidefinite relaxation of quadratic optimization problems, *IEEE Signal Process. Mag.* 27 (3) (2010) 20–34.
- [47] H. Ding, K.-C. Leung, Cross-layer resource allocation in NOMA systems with dynamic traffic arrivals, in: 2020 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2020, pp. 1–6.
- [48] K. Shen, W. Yu, Fractional programming for communication systems—part I: power control and beamforming, *IEEE Trans. Signal Process.* 66 (10) (2018) 2616–2630.
- [49] M.S. Lobo, L. Vandenberghe, S. Boyd, H. Lebret, Applications of second-order cone programming, *Linear Algebra Appl.* 284 (1–3) (1998) 193–228.
- [50] H. Liu, K.J. Kim, K.S. Kwak, H.V. Poor, Power splitting-based SWIPT with decode-and-forward full-duplex relaying, *IEEE Trans. Wirel. Commun.* 15 (11) (2016) 7561–7577.
- [51] J. Zuo, Y. Liu, Z. Qin, N. Al-Dhahir, Resource allocation in intelligent reflecting surface assisted NOMA systems, *IEEE Trans. Commun.* 68 (11) (2020) 7170–7183.