This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2023.3240718

Active RIS Assisted Rate-Splitting Multiple Access Network: Spectral and Energy Efficiency Tradeoff

Hehao Niu, Zhi Lin, Kang An, Jiangzhou Wang, *Fellow, IEEE*, Gan Zheng, *Fellow, IEEE*, Naofal Al-Dhahir, *Fellow, IEEE*, and Kai-Kit Wong, *Fellow, IEEE*

***Abstract*—With the increasing demand of high data rate and massive access in both ultra-dense and industrial Internet-of-things networks, spectral efficiency (SE) and energy efficiency (EE) are regarded as two important and inter-related performance metrics for future networks. In this paper, we investigate a novel integration of rate-splitting multiple access (RSMA) and reconfigurable intelligent surface (RIS) into cellular systems to achieve a desirable tradeoff between SE and EE. Different from the commonly used passive RIS, we adopt reflection elements with active load to improve a newly defined metric, called resource efficiency (RE), which is capable of striking a balance between SE and EE. This paper focuses on the RE optimization by jointly designing the base station (BS) transmit precoding and RIS beamforming (BF) while guaranteeing the transmit and forward power budgets of the BS and RIS, respectively. To efficiently tackle the challenges for solving the RE maximization problem due to its fractional objective function, coupled optimization variables, and discrete coefficient constraint, the formulated nonconvex problem is solved by proposing a two-stage optimization framework. For the outer stage problem, a quadratic transformation is used to recast the fractional objective into a linear form, and a closed-form solution is obtained by using auxiliary variables. For the inner stage problem, the system sum rate is approximated into a linear function. Then, an alternating optimization (AO) algorithm is proposed to optimize the BS precoding and RIS BF iteratively, by utilizing the penalty dual decomposition (PDD) method. Simulation results demonstrate the superiority of the proposed design compared to other benchmarks.**

Manuscript received May 25, 2022; revised October 9, 2022; accepted December 1, 2022. Date of publication XXX XX, 2023; date of current version XXX XX, 2023. This work was supported in part by the National Natural Science Foundation of China (61901490, 61901502, 62071352 and 62201592), in part by the Research Plan Project of NUDT under Grant ZK21-33, in part by the Young Elite Scientist Sponsorship Program of CAST under Grant 2021-JCJQ-QT-048, in part by the Macau Young Scholars Program under Grant AM2022011, and in part by the National Postdoctoral Program for Innovative Talents under Grant BX20200101. The work of N. Al-Dhahir was supported by an Erik Jonsson Distinguished Professorship at UT-Dallas.

(Corresponding author: Zhi Lin.)

Hehao Niu and Kang An are with the Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China (e-mail: niuhaonupt@foxmail.com; ankang89@nudt.edu.cn).

Zhi Lin is with the College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China, and also with the School of Computer Science and Engineering, Macau University of Science and Technology, Macau 999078, China (e-mail: linzhi945@163.com).

Jiangzhou Wang is with the School of Engineering, University of Kent, Canterbury CT2 7NZ, U.K. (e-mail: j.z.wang@kent.ac.uk).

Gan Zheng is with the School of Engineering, University of Warwick, Coventry, CV4 7AL, U.K. (e-mail: gan.zheng@warwick.ac.uk).

Naofal Al-Dhahir is with the Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson TX 75080, USA (e-mail: aldhahir@utdallas.edu).

Kai-Kit Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E, U.K., and also with the School of Integrated Technology, Yonsei University, Seoul 03722, Korea (e-mail: kai-kit.wong@ucl.ac.uk).

***Index Terms*—Rate-splitting multiple access, RIS, active load, tradeoff, resource efficiency.**

I. Introduction

Rate-splitting multiple access (RSMA) technique has been considered as a powerful multiple access strategy and interference management technique for the sixth generation (6G) wireless networks [1]. Particularly, the RSMA technique was firstly introduced in [2], which creatively separates the user messages into common and private parts, and encodes the former into common stream while encoding the latter into separate streams. Then, in [3] and [4], the authors utilized the RSMA in multiuser multiple-input single-output (MISO) networks and massive multiple-input multiple-output (MIMO) channels, respectively. The results of the above works confirmed that RSMA is a promising technique to reduce the inter-user interference and enhance the transmission robustness especially when channel state information (CSI) can not be fully attained at the transmitter in practice.

With these advantages, RSMA was systematically presented in [5], which is a more general case of the space-division multiple access (SDMA) or non-orthogonal multiple access (NOMA) scheme [6]. To be specific, RSMA utilizes linearly precoded rate-splitting (RS) at the transmitter and successive interference cancellation (SIC) at the receiver, which decodes part of the interference and treats the others as noise [7]. Existing works have revealed that RSMA outperforms the orthogonal multiple access (OMA), linear SDMA, and power-domain NOMA, in the aspects of the spectral efficiency (SE) [8]-[11], max-min fairness [12], [13], and energy efficiency (EE) [14], [15].

Meanwhile, following the breakthroughs on the fabrication of programmable metamaterials, reconfigurable intelligent surface (RIS) has emerged as an effective technique to enhance the SE, EE and network coverage, etc, [16]. RIS is a planar array with a large number of low-cost reflective elements, which reflect the incident signal to a desired direction via controlling the phase shifters [17]. Considering that RIS only reflects the received signal without decoding and regenerating signals process, the power consumption and hardware cost of RIS are much lower than those of the conventional radio frequency relay [18]. In addition, RIS can be easily deployed on the facades of buildings, ceilings of factories and indoor spaces [19]. Due to the above advantages, RIS has attracted a great deal of research attention.

For example, L. Wei *et al.* in [20] proposed a factor decomposition-aided channel estimation technique for RIS-assisted wireless networks. Then, the authors further developed a message passing algorithm to achieve joint channel estimation and signal recovery for RIS-aided networks in [21]. The authors in [22] studied the sum rate and fairness optimization in RIS-aided systems. Besides, [23] studied the fairness design in RIS-assisted multi-group multi-cast transmission, where a majorization-maximization (MM)-based method was developed. Then, the work [24] proposed a two-timescale beamforming (BF) scheme for an RIS-enhanced wireless network, where the passive BF was designed with the statistical CSI assumption, while the active BF was designed using the instantaneous CSI. Recently, [25] studied the sum rate maximization design for multiuser MISO networks using deep reinforcement learning, and the work has been extended in [26] when considering multiple cascaded RISs. Moreover, [27] investigated the new fully-connected and group-connected RIS architectures based on reconfigurable impedance networks that can alter both the magnitudes and phases of the incident signals. At present, RIS-aided transmission design has been widely studied in various aspect, such as RSMA networks [28], NOMA networks [29], symbiotic radio networks [30], the full-duplex communication scenario [31], the physical-layer security scenario [32], the millimeter-wave networks [33], hybrid terrestrial-aerial networks [34], and unmanned aerial vehicle networks [35], etc.

Typically, for passive or nearly-passive RIS, the reflected signals have to go through a cascaded channel composed of the transmitter-RIS and RIS-receiver links, leading to serious degradation of the system performance caused by the double fading attenuation. To overcome this shortcoming, the authors in [36] proposed an active RIS architecture in which active load impedance is used by each reflection element. By converting direct current bias power into radio frequency power, the active element can directly amplify the incident signal [37]. Then, in [38], the authors considered the effect of active RIS in improving the achievable secrecy rate of cognitive satellite-terrestrial network. In [39], the authors studied the active RIS-enabled energy-constrained wireless network and showed the superiorities of the active RIS in supporting multiple energy-limited devices. In [40], the authors demonstrated that active RIS achieves better EE performance than passive RIS. Recently, the work in [41] suggested that active RIS outperforms passive RIS in terms of the weighted sum secrecy rate. The authors in [42] compared whether active RIS is superior to passive RIS or not under the same power budget, and revealed that active RIS is superior if the RIS power budget exceeds certain value.

Among the above works, the SE is a commonly used performance metric and usually formulated as the objective of optimization problems [24], [31]. On the other hand, the EE, which is defined as the ratio of the signal data rate to the total power consumption, is one of the key performance metrics in green communication oriented networks [18], [43]. Specifically, [44] investigated the EE optimization in satellite-terrestrial networks with RSMA, where a penalty-based method was proposed. In [45], the authors investigated the EE maximization in RIS-assisted downlink transmission, where a multi-objective optimization (MOO) framework was proposed. Then, [46] studied the EE optimization in RIS-assisted uplink network. It is noted that the above works mainly focused on the SE or EE optimization. However, SE would still increase while EE remains stable with the increasing power in a high power region. Therefore, the SE and EE are not linearly correlated and can not simultaneously increase or decrease with the changing transmit power [47]. Thus, a fundamental tradeoff between the above two metrics needs to be investigated for meeting various communication requirements in future networks.

To be specific, [48] studied the SE-EE tradeoff optimization in a downlink RSMA network. Recently, [49] studied the MOO for the SE-EE tradeoff in the RIS-assisted cognitive radio network. The authors in [50] proposed a new metric called resource efficiency (RE) in RIS-enabled networks to obtain the tradeoff between EE and SE, where an MM-based approach was developed. In addition, [51] investigated the fully-connected RIS-aided RSMA scheme, where the authors proposed a weighted minimum mean square error (WMMSE)-based method to optimize the transmit beamformer and the scattering matrix of the RIS. Then, the authors of [52] proposed the group-connected RIS-based RSMA design to achieve the tradeoff between the beam controlling accuracy and hardware complexity of fully connected RIS. However, the RE optimization in RIS-assisted RSMA networks has not been studied, and it is worth investigating whether active RIS outperforms passive RIS in terms of the RE.

Motivated by the above facts, we focus on the RE optimization in the RIS-assisted RSMA network in this paper. To the best of the authors' knowledge, it is the first work to investigate the active RIS-enabled transmission design in RSMA system from the perspective of SE and EE tradeoff. Our main contributions are summarized as follows:

- We investigate a novel integration of RSMA and RIS into cellular systems in this paper, where the RIS with active reflection elements is deployed to mitigate the double fading effect and one-layer RSMA is adopted at the base station (BS) to suppress the inter-user interference. To achieve the SE-EE tradeoff, we optimize the RE metric by jointly designing the transmit precoding and reflecting BF, while guaranteeing transmit and forward power constraints of the BS and RIS, as well as the common message rate constraint.
- To efficiently tackle the non-convex problems due to its fractional objective, coupled optimization variables, and discrete coefficient constraint, we propose a two-stage approach for the one-layer RSMA transceiver structure. For the outer stage optimization, a quadratic transformation is used to recast the fractional objective into a linear form and obtain a closed-form solution. While for the inner stage problem, we tackle the non-concave common and private signal rate by the first-order Taylor expansion method, which approximates the logarithmic objective into a quadratic function. Then, an alternating optimization (AO) algorithm is developed to obtain the BS precoding and reflective BF iteratively, by resorting to the penalty dual decomposition (PDD) method.
- The proposed algorithm can be extended to the two-layer

RSMA scenario, which is rarely investigated in open literature, and can also handle the common message rate constraint and effectively solve quadratically constrained quadratic program (QCQP) problems with fast convergence. Besides, the proposed algorithm is also applicable to passive-RIS design, by setting the zero noise power at the RIS and normalizing the amplitude of each reflection element. Thus, the proposed scheme is actually an unified optimization framework, which is suitable for various RIS-assisted RSMA networks.

- Simulation results verified the effectiveness of the proposed scheme and provide some insightful analysis: 1) RE is an effective metric to tradeoff the SE and EE performance by adjusting the weight; 2) RSMA outperforms other multiple access techniques such as SDMA; 3) active RIS obtain better RE performance than passive RIS, and the discrete coefficient is a better choice compared the continuous coefficient in terms of RE due to the lower power consumption of former.

The rest of this paper is organized as follows. Section II presents the system model and problem. Section III investigates the joint BF and reflecting coefficient design. Section IV extends the proposed design. Section V illustrates the simulations and section VI is the conclusion.

Notations: Throughout the paper, the upper case boldface letters and lower case boldface letters are used to represent matrices and vectors, respectively. $\mathrm{Tr}\left(\mathbf{\Sigma}\right)$, $\mathbf{\Sigma}^H$, $\mathbf{\Sigma}^T$, and $\mathbf{\Sigma}^*$ denote the trace, the Hermitian transpose, the transpose, and the conjugate of matrix $\mathbf{\Sigma}$, respectively. The diagonal matrix with diagonal elements $\sigma_1,\ldots,\sigma_N$ is denoted by $\mathrm{Diag}\left(\sigma_1,\ldots,\sigma_N\right)$. $\mathrm{diag}\left(\mathbf{\Sigma}\right)$ denotes the main diagonal element of matrix $\mathbf{\Sigma}$. $\Re\{\cdot\}$, $|\cdot|$, and $\angle(\cdot)$ denote the real part, the modulus, and the angle of a complex number, respectively. $\mathbf{x} \sim \mathcal{CN}\left(\boldsymbol{\sigma},\mathbf{\Sigma}\right)$ denotes a circularly symmetric complex Gaussian (CSCG) random vector with mean $\boldsymbol{\sigma}$ and covariance matrix $\mathbf{\Sigma}$. In addition, $\|\cdot\|$ and $\|\cdot\|_F$ represent the Euclidean norm and the Frobenius norm, respectively. Besides, $\circ$ denotes the element-wise product.

II. System Model and Problem Formulation

We present the system model of the RSMA network and formulate the optimization problem.

A. Active RIS Model

The reflecting coefficient matrix of the active RIS is given by $\mathbf{\Phi} = \mathrm{Diag}\left(\phi_1,\ldots,\phi_{N_r}\right) \in \mathbb{C}^{N_r\times N_r}$, where the reflecting coefficient of the n-th element is denoted by $\phi_n = \alpha_n e^{j\theta_n}, n=1,\ldots,N_r$, with α_n and θ_n being the amplitude and the phase within the intervals $\alpha_n \in [0,\alpha_{n,\max}]$, and $\theta_n \in [0,2\pi)$, respectively. Here, $\alpha_{n,\max}$ denotes the predetermined maximum amplitude of the active load for the n-th element. Different from the commonly used passive RIS, $\alpha_{n,\max}$ is not less than 1 for active RIS. Besides, active RIS only utilizes power amplifiers and phase-shift circuits that control and reflect the signals. No dedicated digital-to-analog converters (DACs), analog-to-digital converters (ADCs) and mixers are required. In contrast, relays are usually equipped with these mentioned electronic components for transmission, and low-noise amplifiers for reception, which leads to higher hardware cost and power consumption than active RIS.

Due to the practical hardware conditions, α_n and θ_n can only take discrete values. Let Q_α and Q_θ denote the quantization bits for α_n and θ_n, respectively. Then we have

$$\phi_n \in \mathcal{X}_d \triangleq \left\{\phi_n \left| \phi_n = \alpha_n e^{j\theta_n}, \alpha_n \in \mathcal{S}_\alpha, \theta_n \in \mathcal{S}_\theta \right.\right\}, \quad (1)$$

where $\mathcal{S}_\alpha \triangleq \left[0, \frac{\alpha_{n,\max}}{2^{Q_\alpha}-2},\ldots,\frac{\left(2^{Q_\alpha}-3\right)\alpha_{n,\max}}{2^{Q_\alpha}-2},\alpha_{n,\max}\right]$ denotes the amplitude set, i.e., uniformly values 2^{Q_α} points in $[0,\alpha_{n,\max}]$, and $\mathcal{S}_\theta \triangleq \left\{0,\frac{2\pi}{2^{Q_\theta}},\ldots,\frac{2\pi\left(2^{Q_\theta}-1\right)}{2^{Q_\theta}}\right\}$ denotes the phase set, i.e., θ_n are equally valued in $[0,2\pi)$.

B. Signal Model

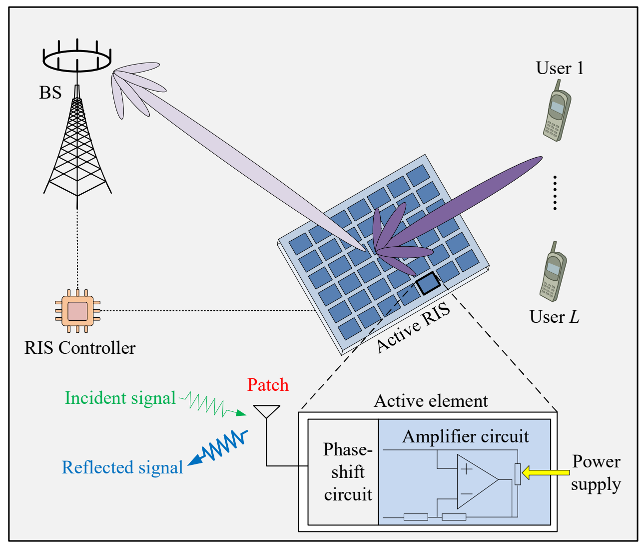


Fig. 1: System model.

As depicted in Fig. 1, we consider a multiuser MISO network which consists of a BS and L users, denoted as $\mathcal{L} \triangleq \{1,\ldots,L\}$. The BS and RIS are equipped with N_s antennas and N_r elements, respectively, while the users are single antenna node. The channel from BS to RIS, from RIS to the l-th user and from BS to the l-th user are denoted by $\mathbf{F} \in \mathbb{C}^{N_r\times N_s}$, $\mathbf{h}_l^H \in \mathbb{C}^{1\times N_r}$, and $\mathbf{g}_l^H \in \mathbb{C}^{1\times N_s}$, respectively. The instantaneous CSI for these links are available at the BS and RIS. For more details about the channel estimation technique for RIS-assisted networks, readers can refer to [20], [21]. In addition, there exists a RIS controller to adjust the amplitudes and phases of the reflection elements [17].

Inspired by [7], one-layer RSMA is an alternative and effective multiple access technique to manage the inter-user interference among these users with relatively low implementation complexity. Hence, we adopt the one-layer RSMA at the BS, where the transceiver architecture is shown in Fig. 2. Particularly, the BS adopts the message combiner and linear precoding to split the message M_l into two sections, namely, a common part $M_{c,l}$ and a private part $M_{p,l}$. $M_{c,1},\ldots,M_{c,L}$ are combined into a common message M_c and encoded into a common signal s_c using a codebook shared by all users. On the contrary, $M_{p,1},\ldots,M_{p,L}$ are separately

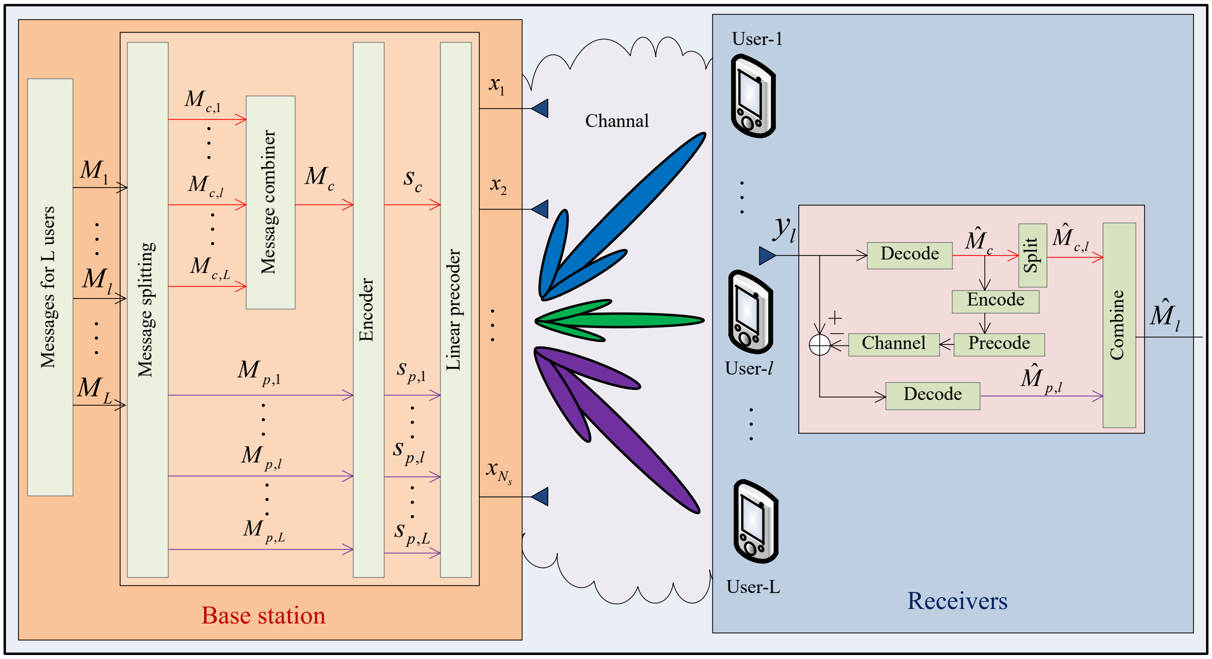


Fig. 2: Transceiver architecture of one-layer RSMA.

encoded into the private signals $s_1, \ldots, s_L$, which are only decoded by the specified user. Therefore, the whole signal $\mathbf{s} = [s_c, s_1, \ldots, s_L]^T \in \mathbb{C}^{(L+1)\times 1}$ is generated for RSMA transmission. Then, $\mathbf{s}$ is linearly precoded via a precoding matrix $\mathbf{W} = [\mathbf{w}_c, \mathbf{w}_1, \ldots, \mathbf{w}_L]^T \in \mathbb{C}^{N_s \times (L+1)}$ with $\mathbf{w}_c$ and $\mathbf{w}_l \in \mathbb{C}^{N_s \times 1}, \forall l \in \mathcal{L}$ being the corresponding precoding vector for s_c and s_l, respectively. Thus, the transmitted signal is given as

$$\mathbf{x} = \mathbf{w}_c s_c + \sum_{l=1}^{L} \mathbf{w}_l s_l. \quad (2)$$

Then, the received signal at the l-th user, $l \in \mathcal{L}$ is given by

$$y_l = \left(\mathbf{g}_l^H + \mathbf{h}_l^H \mathbf{\Phi} \mathbf{F}\right) \mathbf{x} + \mathbf{h}_l^H \mathbf{\Phi} \mathbf{n}_r + n_l, \quad (3)$$

where $\mathbf{n}_r \sim \mathcal{CN}\left(\mathbf{0}, \sigma_r^2 \mathbf{I}\right)$ and $n_l \sim \mathcal{CN}\left(0, \sigma_l^2\right)$ denote the effective noise including the self-interference and the additive white Gaussian noise (AWGN) at the RIS and the AWGN at the l-th user, respectively.

The l-th user first decodes s_c into $\hat{M}_c$ by treating $s_l, \forall l \in \mathcal{L}$ as noise. Therefore, the signal-to-interference-plus-noise ratio (SINR) to decode s_c at the l-th user is

$$\Gamma_{c,l} = \frac{\left|\bar{\mathbf{h}}_l^H \mathbf{w}_c\right|^2}{\sum\limits_{i=1}^{L} \left|\bar{\mathbf{h}}_l^H \mathbf{w}_i\right|^2 + \left\|\mathbf{h}_l^H \mathbf{\Phi}\right\|^2 \sigma_r^2 + \sigma_l^2}, \quad (4)$$

where $\bar{\mathbf{h}}_l = \left(\mathbf{g}_l^H + \mathbf{h}_l^H \mathbf{\Phi} \mathbf{F}\right)^H$ is the equivalent channel from BS to the l-th user.

After decoding and subtracting s_c from y_l, the l-th user decodes s_l by regarding other private signals $s_j, j \neq l, \forall j \in \mathcal{L}$ as the interference. Thus, the SINR to decode s_l at the l-th user is expressed as

$$\Gamma_{p,l} = \frac{\left|\bar{\mathbf{h}}_l^H \mathbf{w}_l\right|^2}{\sum\limits_{i=1, i\neq l}^{L} \left|\bar{\mathbf{h}}_l^H \mathbf{w}_i\right|^2 + \left\|\mathbf{h}_l^H \mathbf{\Phi}\right\|^2 \sigma_r^2 + \sigma_l^2}. \quad (5)$$

C. System SE and EE

The corresponding achievable rates of s_c and s_l at the l-th user are $R_{c,l} = \log_2\left(1 + \Gamma_{c,l}\right)$ and $R_{p,l} = \log_2\left(1 + \Gamma_{p,l}\right)$, respectively. To guarantee that all users can successfully decode s_c, R_c must satisfy $R_c = \min\left\{R_{c,1}, \ldots, R_{c,L}\right\}$.

According to [7], the overall achievable rate of the network is the sum of R_c and $R_{p,l}$, and is given by

$$\eta_{SE}\left(\mathbf{W}, \mathbf{\Phi}\right) = R_c + \sum_{l=1}^{L} R_{p,l} \quad (\text{bits/s/Hz}). \quad (6)$$

Besides, the total power consumption is given by

$$P_{tot}\left(\mathbf{W}, \mathbf{\Phi}\right) = \varepsilon_s \left\|\mathbf{W}\right\|_F^2 + \varepsilon_r \left(\left\|\mathbf{\Phi}\mathbf{F}\mathbf{W}\right\|_F^2 + \sigma_r^2 \left\|\mathbf{\Phi}\right\|_F^2\right) + P_c, \quad (7)$$

where ε_s and ε_r are the inverses of the power amplification coefficients at the BS and RIS, respectively, and P_c denotes the total static power consumption which is independent of $\{\mathbf{W}, \mathbf{\Phi}\}$ and given by

$$P_c = N_s P_a + P_s + N_r \left(P_r + P_{DC}\right), \quad (8)$$

where P_a denotes the power dissipation per antenna at the BS, P_s is the static circuit power consumption at the BS, P_r incorporates the power consumption of the circuit at each reflection element [50]. According to [18], P_r are 1.5, 4.5, 6.0, and 7.8 mW for 3-, 4-, 5-, and 6-bit quantization of each element, respectively. Besides, P_{DC} is the direct current biasing power consumption of a single reflection element [36].

Therefore, the EE is given by

$$\eta_{EE}\left(\mathbf{W}, \mathbf{\Phi}\right) = B \frac{\eta_{SE}\left(\mathbf{W}, \mathbf{\Phi}\right)}{P_{tot}\left(\mathbf{W}, \mathbf{\Phi}\right)} \quad (\text{bits/Joule}), \quad (9)$$

where B is the system bandwidth.

D. Problem Formulation

Rather than considering SE or EE as the objective, we aim to obtain the SE-EE tradeoff by using a weighted sum representation, i.e., maximizing $(1-\delta)\eta_{EE} + \delta\eta_{SE}$, $0 \leq \delta \leq 1$.

Nevertheless, it is improper to directly combine η_{EE} and η_{SE} because they have different units. Instead, we study a system indicator called RE, which is given in [47], [50]:

$$\eta_{RE}\left(\mathbf{W},\mathbf{\Phi}\right) \triangleq \frac{\eta_{EE}\left(\mathbf{W},\mathbf{\Phi}\right)}{B} + \frac{\beta\eta_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)}{P_{sum}} \quad (\text{bits/J/Hz}), \tag{10}$$

where $\beta > 0$ is the weight used to denote the priorities of EE and SE. The denominator P_{sum} in the second term is fixed and given by

$$P_{sum} = \varepsilon_s P_{s,\max} + \varepsilon_r P_{r,\max} + P_c. \tag{11}$$

P_{sum} denotes the whole power budget of the considered network. By multiplying η_{EE} and η_{SE} with $1/B$ and $1/P_{sum}$, the units of the two terms in (10) are both normalized into bits/Joule/Hz. Moreover, letting $\beta\frac{B}{P_{tot}} \triangleq \delta/(1-\delta)$ and substituting it into (10), the maximization of η_{RE} is equivalent to that of $(1-\delta)\,\eta_{EE} + \delta\eta_{SE}, 0 \le \delta \le 1$.

Mathematically, the RE optimization is formulated as:

$$\max_{\mathbf{W},\mathbf{\Phi}} \ \eta_{RE}\left(\mathbf{W},\mathbf{\Phi}\right) \tag{12a}$$

$$\text{s.t.} \ \ \|\mathbf{W}\|_F^2 \le P_{s,\max}, \tag{12b}$$

$$\|\mathbf{\Phi}\mathbf{F}\mathbf{W}\|_F^2 + \sigma_r^2 \|\mathbf{\Phi}\|_F^2 \le P_{r,\max}, \tag{12c}$$

$$\left|[\mathbf{\Phi}]_n\right| \le \alpha_{n,\max}, \phi_n \in \mathcal{X}_d, \forall n, \tag{12d}$$

$$R_c \ge R_{c,\min}, \tag{12e}$$

where $P_{s,\max}$ and $P_{r,\max}$ are the transmit and forward power budgets at the BS and RIS, respectively. Besides, $[\mathbf{\Phi}]_n$ denotes the n-th element of $\mathbf{\Phi}$. It is known that $[\mathbf{\Phi}]_n = \phi_n$. In addition, $R_{c,\min}$ is the pre-determined minimum common message rate. In fact, different SE-EE tradeoff designs can be realized by changing β. For example, η_{RE} relaxes to η_{SE} when $\beta \to \infty$ and relaxes to η_{EE} when $\beta = 0$ with bandwidth normalization.

Unfortunately, problem (12) is challenging to solve where the major difficulties can be summarized as follows: 1) It is complicated to tackle the coupled variables $\mathbf{W}$ and $\mathbf{\Phi}$ jointly, especially for the case of large N_s or N_r; 2) η_{RE} is a transformation from η_{EE}, which means that the optimization of the former is much more complex than the latter; 3) The optimization of $\mathbf{\Phi}$ with a discrete coefficient constraint is a mixed integer program, and is also non-convex.

In the next section, we aim to develop an efficient approach to handle (12).

III. Joint BS Precoding and RIS BF Design

We first linearize the RE objective by the quadratic transformation method proposed in [54]. Next, an AO algorithm is proposed to do the optimization iteratively.

A. Quadratic Transformation to Fractional Programming

To be specifical, by introducing a slack variable $\kappa \in \mathbb{R}$, we equivalently recast (12) into a non-fractional formulation as:

$$\max_{\mathbf{W},\mathbf{\Phi},\kappa} \ f\left(\mathbf{W},\mathbf{\Phi},\kappa\right) = 2\kappa\sqrt{\eta_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)} - \kappa^2 P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right) + \frac{\beta\eta_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)}{P_{sum}} \tag{13a}$$

$$\text{s.t.} \ \ (12\text{b}) - (12\text{e}), \tag{13b}$$

To solve (13), we optimize the prime variables $\{\mathbf{W},\mathbf{\Phi}\}$ and the slack variable κ in an iterative manner. According to [54], given $\{\mathbf{W},\mathbf{\Phi}\}$, the optimal κ can be directly obtained as

$$\kappa = \frac{\sqrt{\eta_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)}}{P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)}. \tag{14}$$

In the following part, we deal with the optimization of $\{\mathbf{W},\mathbf{\Phi}\}$ with a given κ.

B. SE Approximation

In the previous subsection, we have transformed the fractional programming to a relatively simple formulation. However, given κ, (13) is still unsolvable since η_{SE} is non-concave with respect to (w.r.t.) $\mathbf{W}$ and $\mathbf{\Phi}$. Besides, the discrete coefficient constraint is also non-convex. To overcome this, in the following, we design a lower bound of η_{SE}. Specifically, at the t-th iteration, around the given point $\left\{\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right\}$, the following lemma is useful to approximate $R_{c,l}$ and $R_{p,l}$.

Lemma 1 [56]: For any δ and γ, we have

$$\begin{aligned}\log_2\left(1+\frac{|\delta|^2}{\gamma}\right) \ge{}& \log_2\left(1+\frac{\left|\bar{\delta}\right|^2}{\bar{\gamma}}\right) - \frac{\left|\bar{\delta}\right|^2}{\bar{\gamma}\ln 2} \\ &+ \frac{2\Re\{\bar{\delta}^*\delta\}}{\bar{\gamma}\ln 2} - \frac{\left|\bar{\delta}\right|^2\left(\gamma+|\delta|^2\right)}{\bar{\gamma}\left(\bar{\gamma}+\left|\bar{\delta}\right|^2\right)\ln 2},\end{aligned} \tag{15}$$

where $\left\{\bar{\delta},\bar{\gamma}\right\}$ are fixed points.

Based on Lemma 1, $R_{c,l}$ can be lower bounded as

$$\begin{aligned}R_{c,l} \ge{}& \log_2\left(1+\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}}\right) - \frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}\ln 2} \\ &+ \frac{2\Re\left\{\left(a_l^{(t)}\right)^* a_l\right\}}{b_l^{(t)}\ln 2} - \frac{\left|a_l^{(t)}\right|^2\left(b_l+|a_l|^2\right)}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2},\end{aligned} \tag{16}$$

where $a_l^{(t)} = \left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_c^{(t)}$, $a_l = \left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_c$, $b_l^{(t)} = \sum_{i=1}^L \left|\left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_i^{(t)}\right|^2 + \left\|\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\right\|^2\sigma_r^2 + \sigma_l^2$, and $b_l = \sum_{i=1}^L \left|\left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_i\right|^2 + \left\|\mathbf{h}_l^H\mathbf{\Phi}\right\|^2\sigma_r^2 + \sigma_l^2$, respectively.

Similarly, $R_{p,l}$ can be lower bounded as

$$\begin{aligned}R_{p,l} ={}& \log_2\left(1+\frac{\left|\left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_l\right|^2}{b_l - \left|\left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_l\right|^2}\right) \\ \ge{}& \log_2\left(1+\frac{\left|c_l^{(t)}\right|^2}{b_l^{(t)} - |c_l^t|^2}\right) - \frac{|c_l^t|^2}{\left(b_l^{(t)} - \left|c_l^{(t)}\right|^2\right)\ln 2} \\ &+ \frac{2\Re\left\{\left(c_l^{(t)}\right)^* c_l\right\}}{\left(b_l^{(t)} - \left|c_l^{(t)}\right|^2\right)\ln 2} - \frac{\left|c_l^{(t)}\right|^2 b_l}{b_l^{(t)}\left(b_l^{(t)} - \left|c_l^{(t)}\right|^2\right)\ln 2},\end{aligned} \tag{17}$$

where $c_l^{(t)} = \left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_l^{(t)}$, and $c_l = \left(\mathbf{g}_t^H + \mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_l$, respectively.

Based on the above reformulation, an approximated version of $\eta_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)$, which is denoted as $\tilde{\eta}_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)$, is achieved by solving the following problem

$$\max_{\mathbf{W},\mathbf{\Phi}} \ R_c+\sum_{l=1}^{L}\left\{\frac{2\Re\left\{\left(c_l^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_l\right\}}{\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}+\log_2\left(1+\frac{\left|c_l^{(t)}\right|^2}{b_l^{(t)}-\left|c_l^{(t)}\right|^2}\right)-\frac{\left|c_l^{(t)}\right|^2}{\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}\left(1+\frac{\sigma_l^2}{b_l^{(t)}}\right)-\frac{\left|c_l^{(t)}\right|^2\left(\sum_{i=1}^{L}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_i\right|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}\right\|^2\sigma_r^2\right)}{b_l^{(t)}\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}\right\} \tag{18a}$$

$$\text{s.t.} \ R_c\le\frac{2\Re\left\{\left(a_l^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_c\right\}}{b_l^{(t)}\ln 2}+\log_2\left(1+\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}}\right)-\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}\ln 2}\left(1+\frac{\sigma_l^2}{b_l^{(t)}+\left|a_l^{(t)}\right|^2}\right)-\frac{\left|a_l^{(t)}\right|^2\left(\left\|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{W}\right\|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}\right\|^2\sigma_r^2\right)}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2},\forall l, \tag{18b}$$

$$(12\text{b})-(12\text{e}). \tag{18c}$$

In fact, an effective way to handle the SE optimization of the RSMA network is by solving (18). Next, we will tackle the RE optimization with the assistance of $\tilde{\eta}_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)$. However, (18) is not jointly convex w.r.t. $\mathbf{W}$ and $\mathbf{\Phi}$. Fortunately, (18) can be decoupled into two subproblems. Then, the solution to (18) can be obtained in an alternating method. Next, we will focus on the formulation and solution of the subproblems.

C. Precoding Optimization

Here, we optimize $\mathbf{W}$ with fixed $\mathbf{\Phi}$. According to (18) and after some mathematical operations, the subproblem w.r.t. $\mathbf{W}$ is given as

$$\max_{\mathbf{W}} \ 2\kappa\sqrt{R_c+\sum_{l=1}^{L}\left\{2\Re\left\{\mathbf{p}_l^H\mathbf{w}_l\right\}-\mathbf{w}_l^H\mathbf{P}\mathbf{w}_l+p_l\right\}}-\kappa^2P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)+\frac{\beta\left(R_c+\sum_{l=1}^{L}\left\{2\Re\left\{\mathbf{p}_l^H\mathbf{w}_l\right\}-\mathbf{w}_l^H\mathbf{P}\mathbf{w}_l+p_l\right\}\right)}{P_{sum}} \tag{19a}$$

$$\text{s.t.} \ R_c\le-\mathrm{Tr}\left(\mathbf{Q}_l\mathbf{W}\mathbf{W}^H\right)+2\Re\left\{\mathbf{q}_l^H\mathbf{w}_c\right\}+q_l,\forall l, \tag{19b}$$

$$(12\text{b}),(12\text{c}),(12\text{e}), \tag{19c}$$

where $\{\mathbf{P},\mathbf{p}_l,p_l,\mathbf{Q},\mathbf{q}_l,q_l\}$ are respectively, given by (20) in the next page.

Note that with fixed $\mathbf{\Phi}$, $P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)$ is convex w.r.t. $\mathbf{W}$, and $R_c+\sum_{l=1}^{L}\left\{2\Re\{\mathbf{p}_l^H\mathbf{w}_l\}-\mathbf{w}_l^H\mathbf{P}\mathbf{w}_l+p_l\right\}$ is a concave function w.r.t. $\mathbf{W}$, so that the square-root is also concave. Thus, with fixed κ and $\mathbf{\Phi}$, (19) is convex w.r.t. $\mathbf{W}$, which can be solved by the optimization toolbox CVX [55].

D. Reflecting Coefficient Optimization

Now, we handle the optimization of $\mathbf{\Phi}$. By expending the relevant terms in (18) w.r.t. $\mathbf{\Phi}$ and omitting the irrelevant terms, we have the following problem w.r.t. $\mathbf{\Phi}$

$$\max_{\mathbf{\Phi}} \ R_c+2\Re\left\{\mathrm{Tr}\left(\mathbf{T}\mathbf{\Phi}\right)\right\}+u-\mathrm{Tr}\left(\mathbf{U}\mathbf{\Phi}\left(\mathbf{F}\sum_{i=1}^{L}\mathbf{w}_i\mathbf{w}_i^H\mathbf{F}^H+\sigma_r^2\mathbf{I}\right)\mathbf{\Phi}^H\right) \tag{21a}$$

$$\text{s.t.} \ R_c\le 2\Re\left\{\mathrm{Tr}\left(\mathbf{D}_l\mathbf{\Phi}\right)\right\}+d_l-\mathrm{Tr}\left(\mathbf{V}_l\mathbf{\Phi}\left(\mathbf{F}\mathbf{W}\mathbf{W}^H\mathbf{F}^H+\sigma_r^2\mathbf{I}\right)\mathbf{\Phi}^H\right),\forall l, \tag{21b}$$

$$\mathrm{Tr}\left(\mathbf{\Phi}\left(\mathbf{F}\mathbf{W}\mathbf{W}^H\mathbf{F}^H+\sigma_r^2\mathbf{I}\right)\mathbf{\Phi}^H\right)\le P_{r,\max}, \tag{21c}$$

$$(12\text{d}),(12\text{e}), \tag{21d}$$

where $\{\mathbf{T},u,\mathbf{U},\mathbf{D}_l,d_l,\mathbf{V}_l\}$ are respectively, given by (22) in the next page.

Next, we utilize the following lemma to handle (21).

Lemma 2 [50]: Let $\mathbf{U}_1\in\mathbb{C}^{m\times m}$, $\mathbf{U}_2\in\mathbb{C}^{m\times m}$. Assuming that $\mathbf{\Sigma}=\mathrm{Diag}\left(\sigma_1,\ldots,\sigma_m\right)\in\mathbb{C}^{m\times m}$, $\boldsymbol{\sigma}=\mathrm{diag}\left(\mathbf{\Sigma}\right)$, and $\mathbf{u}_2=\mathrm{diag}\left(\mathbf{U}_2\right)$, then we have

$$\begin{aligned}&\mathrm{Tr}\left(\mathbf{\Sigma}^H\mathbf{U}_1\mathbf{\Sigma}\mathbf{U}_2\right)=\boldsymbol{\sigma}^H\left(\mathbf{U}_1\circ\mathbf{U}_2^T\right)\boldsymbol{\sigma},\\&\mathrm{Tr}\left(\mathbf{\Sigma}\mathbf{U}_2\right)=\boldsymbol{\sigma}^T\mathbf{u}_2.\end{aligned} \tag{23}$$

Then, by defining $\boldsymbol{\phi}=\left[\phi_1,\ldots,\phi_{N_r}\right]^T$, we turn (21) into the following problem

$$\max_{\boldsymbol{\phi}} \ 2\kappa\sqrt{R_c+u+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}}-\kappa^2P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)+\frac{\beta}{P_{sum}}\left(R_c+u+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}\right) \tag{24a}$$

$$\text{s.t.} \ R_c\le d_l+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{D}_l\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{V}}_l\boldsymbol{\phi},\forall l, \tag{24b}$$

$$\boldsymbol{\phi}^H\left(\mathbf{I}\circ\left(\mathbf{F}\mathbf{W}\mathbf{W}^H\mathbf{F}^H+\sigma_r^2\mathbf{I}\right)^T\right)\boldsymbol{\phi}\le P_{r,\max}, \tag{24c}$$

$$\left|\boldsymbol{\phi}_n\right|\le\alpha_{n,\max},\phi_n\in\mathcal{X}_d,\forall n, \tag{24d}$$

$$(12\text{e}), \tag{24e}$$

where $\bar{\mathbf{U}}=\mathbf{U}\circ\left(\mathbf{F}\sum_{i=1}^{L}\mathbf{w}_i\mathbf{w}_i^H\mathbf{F}^H+\sigma_r^2\mathbf{I}\right)^T$ and $\bar{\mathbf{V}}_l=\mathbf{V}_l\circ\left(\mathbf{F}\mathbf{W}\mathbf{W}^H\mathbf{F}^H+\sigma_r^2\mathbf{I}\right)^T$, respectively.

In fact, when the RIS has continuous coefficients, i.e., the constraint $\phi_n\in\mathcal{X}_d$ is absent, (24) is simplified to a convex problem and can be efficiently solved. However, when considering discrete coefficients, (24) can not be solved directly. Next, we propose a PDD-based algorithm to address this challenge. Specifically, we introduce the slack variable $\boldsymbol{\omega}=\left[\omega_1,\ldots,\omega_{N_r}\right]^T\in\mathbb{C}^{N_r\times 1}$ as a copy of the prime variable $\boldsymbol{\phi}$. Then, we reformulate (24) as

$$\max_{\boldsymbol{\phi},\boldsymbol{\omega}} \ 2\kappa\sqrt{R_c+u+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}}-\kappa^2P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)+\frac{\beta}{P_{sum}}\left(R_c+u+2\Re\left\{\phi^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}\right) \tag{25a}$$

$$\text{s.t.} \ \left|\boldsymbol{\phi}_n\right|\le\alpha_{n,\max},(24\text{b}),(24\text{c}),(24\text{e}), \tag{25b}$$

$$\boldsymbol{\phi}=\boldsymbol{\omega},\omega_n\in\mathcal{X}_d,\forall n. \tag{25c}$$

By penalizing the equality constraint $\boldsymbol{\phi}=\boldsymbol{\omega}$, the following

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2023.3240718

$$
\begin{aligned}
&\mathbf{P}=\sum_{l=1}^{L}\frac{\left|c_l^{(t)}\right|^2\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)^H\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)}{b_l^{(t)}\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2},\mathbf{p}_l^H=\frac{\left(c_l^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)}{\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2},\\
&p_l=\log_2\left(1+\frac{\left|c_l^{(t)}\right|^2}{b_l^{(t)}-\left|c_l^{(t)}\right|^2}\right)-\frac{\left|c_l^{(t)}\right|^2}{\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}-\frac{\left|c_l^{(t)}\right|^2\left(\left\|\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\right\|^2\sigma_r^2+\sigma_l^2\right)}{b_l^{(t)}\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2},\\
&\mathbf{Q}_l=\frac{\left|a_l^{(t)}\right|^2\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)^H\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)}{b_l^t\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2},\mathbf{q}_l^H=\frac{\left(a_l^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)}{b_l^{(t)}\ln 2},\\
&q_l=\log_2\left(1+\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}}\right)-\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}\ln 2}-\frac{\left|a_l^{(t)}\right|^2\left(\left\|\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\right\|^2\sigma_r^2+\sigma_l^2\right)}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2}.
\end{aligned}
\tag{20}
$$

$$
\begin{aligned}
&\mathbf{T}=\sum_{l=1}^{L}\frac{\left(c_l^{(t)}\right)^*}{\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}\mathbf{F}\mathbf{w}_l^{(t)}\mathbf{h}_l^H-\sum_{l=1}^{L}\frac{\left|c_l^{(t)}\right|^2\left(\mathbf{F}\sum\limits_{i=1}^{L}\mathbf{w}_i^{(t)}\left(\mathbf{w}_i^{(t)}\right)^H\mathbf{g}_l\mathbf{h}_l^H\right)}{b_l^{(t)}\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2},\mathbf{U}=\sum_{l=1}^{L}\frac{\left|c_l^{(t)}\right|^2\mathbf{h}_l\mathbf{h}_l^H}{b_l^{(t)}\left(b_l^{(t)}-\left|c_l^{t}\right|^2\right)\ln 2},\\
&u=\sum_{l=1}^{L}\left\{\log_2\left(1+\frac{\left|c_l^{(t)}\right|^2}{b_l^{(t)}-\left|c_l^{(t)}\right|^2}\right)-\frac{\left|c_l^{(t)}\right|^2}{\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}-\frac{\left|c_l^{(t)}\right|^2\left(\sigma_l^2+\sum\limits_{i=1}^{L}\left|\mathbf{g}_l^H\mathbf{w}_i^{(t)}\right|^2\right)}{b_l^{(t)}\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}+\frac{2\Re\left\{\left(c_l^{(t)}\right)^*\mathbf{g}_l^H\mathbf{w}_l^{(t)}\right\}}{\left(b_l^{(t)}-\left|c_l^{(t)}\right|^2\right)\ln 2}\right\},\\
&\mathbf{D}_l=\frac{\left(a_l^{(t)}\right)^*}{b_l^{(t)}\ln 2}\mathbf{F}\mathbf{w}_c^{(t)}\mathbf{h}_l^H-\frac{\left|a_l^{(t)}\right|^2\left(\mathbf{F}\mathbf{W}^{(t)}\left(\mathbf{W}^{(t)}\right)^H\mathbf{g}_l\mathbf{h}_l^H\right)}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2},\mathbf{V}_l=\frac{\left|a_l^{(t)}\right|^2\mathbf{h}_l\mathbf{h}_l^H}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2},\\
&d_l=\log_2\left(1+\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}}\right)-\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}\ln 2}-\frac{\left|a_l^{(t)}\right|^2\left(\sigma_l^2+\left\|\mathbf{g}_l^H\mathbf{W}^{(t)}\right\|^2\right)}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2}+\frac{2\Re\left\{\left(a_l^{(t)}\right)^*\mathbf{g}_l^H\mathbf{w}_c^{(t)}\right\}}{b_l^{(t)}\ln 2}.
\end{aligned}
\tag{22}
$$

augmented Lagrange (AL) of (25) can be obtained

$$
\begin{aligned}
\min_{\boldsymbol{\phi},\boldsymbol{\omega},\varphi,\boldsymbol{\lambda}}\ &-2\kappa\sqrt{R_c+u+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}}\\
&+\kappa^2P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)+\frac{1}{2\varphi}\left\|\boldsymbol{\phi}-\boldsymbol{\omega}+\varphi\boldsymbol{\lambda}\right\|^2 \qquad (26a)\\
&-\frac{\beta}{P_{sum}}\left(R_c+u+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}\right)\\
\text{s.t.}\ \ &(25b),\omega_n\in\mathcal{X}_d,\forall n, \qquad (26b)
\end{aligned}
$$

where $\varphi\geq 0$ is the penalty factor and $\boldsymbol{\lambda}\in\mathbb{C}^{M\times 1}$ is a dual variable associated with $\boldsymbol{\phi}=\boldsymbol{\omega}$, respectively. Actually, when $\varphi\leq\frac{1}{2\lambda_{\max}(\bar{\mathbf{U}})}$, (26) is guaranteed to converge [32].

The PDD procedure is composed of two stages. In the outer stage, we optimize φ and $\boldsymbol{\lambda}$, while in the inner stage, we decouple (26) into two problems and obtain $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ alternately. First, we optimize $\boldsymbol{\phi}$, given $\boldsymbol{\omega}$. The problem is given as

$$
\begin{aligned}
\min_{\boldsymbol{\phi}}\ &-2\kappa\sqrt{R_c+u+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}}\\
&+\kappa^2P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)+\frac{1}{2\varphi}\left\|\boldsymbol{\phi}-\boldsymbol{\omega}+\varphi\boldsymbol{\lambda}\right\|^2 \qquad (27a)\\
&-\frac{\beta}{P_{sum}}\left(R_c+u+2\Re\left\{\boldsymbol{\phi}^T\mathrm{diag}\left(\mathbf{T}\right)\right\}-\boldsymbol{\phi}^H\bar{\mathbf{U}}\boldsymbol{\phi}\right)\\
\text{s.t.}\ \ &(25b), \qquad (27b)
\end{aligned}
$$

which can be solved by CVX. Then, given $\boldsymbol{\phi}$, we optimize $\boldsymbol{\omega}$. The problem is written as

$$
\min_{\boldsymbol{\omega}}\left\|\boldsymbol{\phi}-\boldsymbol{\omega}+\varphi\boldsymbol{\lambda}\right\|^2 \tag{28a}
$$

$$
\text{s.t.}\ \omega_n\in\mathcal{X}_d,\forall n. \tag{28b}
$$

Since ω_n are decoupled from each other in (28), the optimal solution is $\omega_n^\star=\bar{\alpha}_n e^{j\angle\bar{\theta}_n}$, where $\bar{\theta}_n=\arg\min\limits_{\theta_n\in\mathcal{S}_\theta}\left|\theta_n-\angle\left(\phi_n+\varphi\lambda_n\right)\right|$ and $\bar{\alpha}_n=\arg\min\limits_{\angle\alpha_n\in\mathcal{S}_\alpha}\left|\alpha_n e^{j\angle\bar{\theta}_n}-\phi_n-\varphi\lambda_n\right|$, respectively.

The inner stage alternatively updates $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ until the stopping criterion is met. Then, for the outer stage iteration,

$\boldsymbol{\lambda}$ and φ are iterated by

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \frac{1}{\varphi}\left(\boldsymbol{\phi} - \boldsymbol{\omega}\right), \text{ and } \varphi \leftarrow \rho\varphi, \tag{29}$$

where $\rho < 1$ is a factor which scale φ at each iteration.

The PDD process is given in Algorithm 1, where ϵ_1 denotes the stopping threshold. According to [53], Algorithm 1 is guaranteed to converge whether in the continuous or discrete coefficients case. Readers can refer to [53] for more details.

Algorithm 1 The PDD Algorithm.

Initialize $\boldsymbol{\phi}^{(0)}$, $\boldsymbol{\omega}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, $\varphi^{(0)}$, and set $k=1$;
repeat
  Set $\boldsymbol{\phi}^{(k-1,\ell)} = \boldsymbol{\phi}^{(k-1)}$, $\boldsymbol{\omega}^{(k-1,\ell)} = \boldsymbol{\omega}^{(k-1)}$, and $\ell = 0$;
  repeat
    Obtain $\boldsymbol{\phi}^{(k-1,\ell+1)}$ via solving problem (27);
    Obtain $\boldsymbol{\omega}^{(k-1,\ell+1)}$ via solving problem (28);
    $\ell \leftarrow \ell + 1$;
  until Convergence.
  $\boldsymbol{\phi}^{(k)} \leftarrow \boldsymbol{\phi}^{(k-1,\ell)}$, $\boldsymbol{\omega}^{(k)} \leftarrow \boldsymbol{\omega}^{(k-1,\ell)}$;
  $\boldsymbol{\lambda}^{(k)} \leftarrow \boldsymbol{\lambda}^{(k-1)} + \frac{1}{\varphi^{(k)}}\left(\boldsymbol{\phi}^{(k)} - \boldsymbol{\omega}^{(k)}\right)$, $\varphi^{(k)} \leftarrow \rho\varphi^{(k-1)}$;
  $k \leftarrow k+1$;
until $\left\|\boldsymbol{\phi}^{(\ell)} - \boldsymbol{\omega}^{(\ell)}\right\| \leq \epsilon_1$ or the maximum number of iteration is met.
Output $\boldsymbol{\phi}^{\star}$.

E. Overall Algorithm and Analysis

Combining the proposed steps above, we obtain the integrated RE maximization approach in Algorithm 2, where ϵ_2 denotes the stopping threshold and the initial point $\left\{\mathbf{W}^{(0)}, \boldsymbol{\Phi}^{(0)}\right\}$ is set as the optimal solution of (32). In addition, we have the following Theorem.

Theorem 1: Algorithm 2 generates a convergent sequence of the objective values of (12).

Proof: Please refer to Appendix A. ■

Algorithm 2 Quadratic transformation algorithm for solving problem (12).

Solve (32) and obtain $\mathbf{W}^{(0)}$, $\boldsymbol{\Phi}^{(0)}$, initialize $\kappa^{(0)}$ and set $k=1$;
repeat
  Set $\mathbf{W}^{(k-1,t)} = \mathbf{W}^{(k-1)}$, $\boldsymbol{\Phi}^{(k-1,t)} = \boldsymbol{\Phi}^{(k-1)}$, and $t=0$;
  repeat
    Obtain $\mathbf{W}^{(k-1,t+1)}$ via solving problem (19);
    Obtain $\boldsymbol{\Phi}^{(k-1,t+1)}$ via solving problem (24);
    $t \leftarrow t+1$;
  until Convergence
  $\mathbf{W}^{(k)} \leftarrow \mathbf{W}^{(k-1,t)}$, $\boldsymbol{\Phi}^{(k)} \leftarrow \boldsymbol{\Phi}^{(k-1,t)}$;
  Update κ by (14);
  $k \leftarrow k+1$;
until $\kappa^{(k)} - \kappa^{(k-1)} \leq \epsilon_2$ or the maximum number of iteration is reached.
Output $\mathbf{W}^{\star}$, $\boldsymbol{\Phi}^{\star}$, and $\kappa^{\star}$.

Actually, by setting $\beta = 0$, Algorithm 2 can be directly simplified to the EE maximization design with normalized bandwidth. On the other hand, by setting $\varepsilon_s = 0$ and $\varepsilon_r = 0$, Algorithm 2 can tackle the SE optimization problem which is relaxed from RE optimization problem.

Next, we calculate the computational complexity of Algorithm 2. Since the complexities of (28) and (29) can be omitted when compared with those of (19) and (27), the complexity of Algorithm 2 is mainly determined by (19) and (27). According to [57], the complexity for solving a QCQP problem is given by $\mathcal{O}\left(\sqrt{m}\left(mn^2+n^3\right)\ln\left(\frac{2m}{\epsilon}\right)\right)$, where m denotes the number of variables and n denoted the number of constraints, and ϵ is the solution accuracy. Specifically, (19) has $(3L+2)$ quadratic constraints and the dimension of the variable is N_s, while (27) has (N_r+L+1) constraints and the dimension of the variable is N_r. Thus, the complexities of (19) and (27) are given by $\mathcal{O}\left(\sqrt{N_s}\left(9N_sL^2+27L^3\right)\ln\left(\frac{2N_s}{\epsilon}\right)\right)$ and $\mathcal{O}\left(\sqrt{N_r}\left(N_r(N_r+L)^2+(N_r+L)^3\right)\ln\left(\frac{2N_r}{\epsilon}\right)\right)$, respectively. Therefore, the overall computational complexity of Algorithm 2 is given by

$$\begin{aligned}\mathcal{C} = \mathcal{O}\Bigg(T_\kappa \max\Bigg\{&\sqrt{N_s}\left(9L^2\left(N_s+3L\right)\right)\ln\left(\frac{2N_s}{\epsilon}\right),\\ &\sqrt{N_r}\left(\left(2N_r+L\right)\left(N_r+L\right)^2\right)\ln\left(\frac{2N_r}{\epsilon}\right)\Bigg\}\Bigg)\end{aligned} \tag{30}$$

where T_κ denotes the search times for the updating of κ in the outer layer.

Thus, Algorithm 2 has polynomial time complexity, which is suitable for implementation.

F. Feasibility of (12)

Note that problem (12) with the common message rate constraint (12e) might be infeasible. It is necessary to study the feasibility of solving (12). Therefore, we solve the following problem to check the feasibility of (12):

$$\max_{\mathbf{W},\boldsymbol{\Phi}} \ \min_{l} R_{c,l} \tag{31a}$$

$$\text{s.t. } (12\text{b}) - (12\text{d}). \tag{31b}$$

If the obtained optimal value of (31) is larger than or equal to $R_{c,\min}$, then (12) is feasible to solve. Otherwise, (12) is infeasible. In fact, the previously proposed AO algorithm can be utilized to solve (31). Proceedings in a similar way as we obtained the lower bound of $R_{c,l}$, an approximated version of (31) is obtained as follows

$$\max_{\mathbf{W},\boldsymbol{\Phi},\tau} \ \tau \tag{32a}$$

$$\begin{aligned}\text{s.t. } \tau \leq\ & \frac{2\Re\left\{\left(a_l^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\boldsymbol{\Phi}\mathbf{F}\right)\mathbf{w}_c\right\}}{b_l^{(t)}\ln 2}\\ &-\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}\ln 2}\left(1+\frac{\sigma_l^2}{b_l^{(t)}+\left|a_l^{(t)}\right|^2}\right)+\log_2\left(1+\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}}\right)\\ &-\frac{\left|a_l^{(t)}\right|^2\left(\left\|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\boldsymbol{\Phi}\mathbf{F}\right)\mathbf{W}\right\|^2+\left\|\mathbf{h}_l^H\boldsymbol{\Phi}\right\|^2\sigma_r^2\right)}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2}, \forall l,\end{aligned} \tag{32b}$$

$$(12\text{b}) - (12\text{d}). \tag{32c}$$

Then, an AO procedure is used to solve (32), which is similar to that in Algorithm 2. Thus, the feasibility of (12) can be checked.

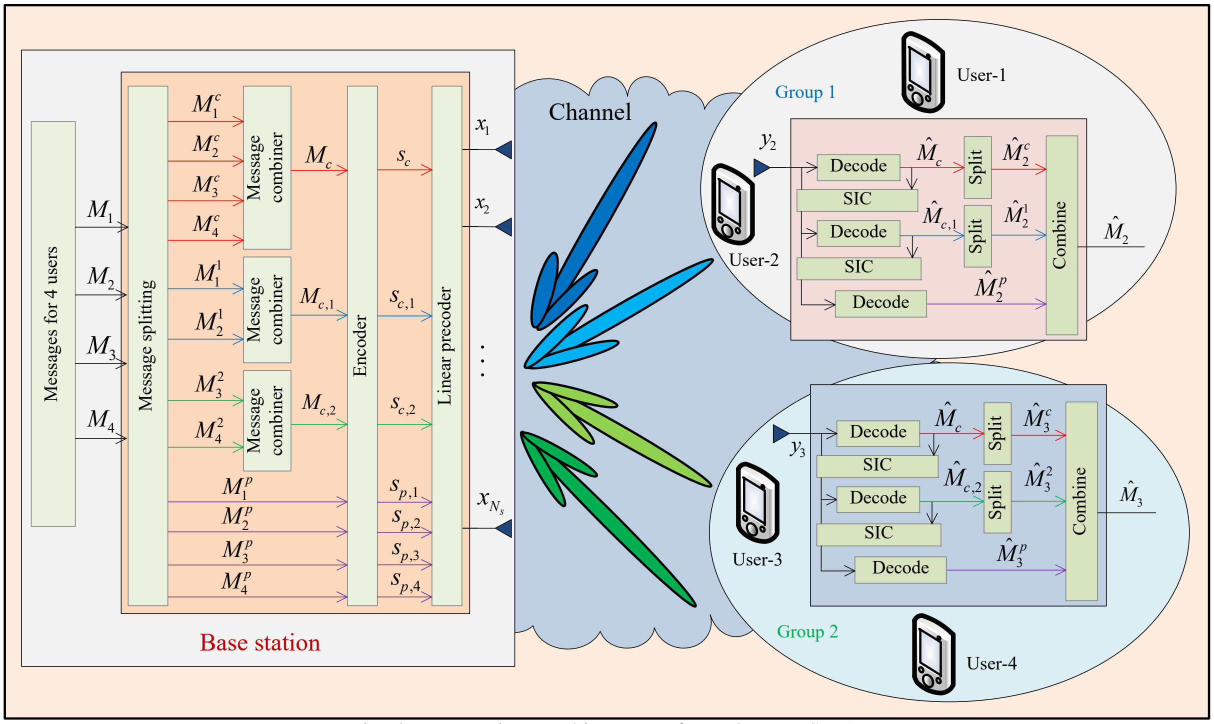


Fig. 3: Transceiver architecture of two-layer RSMA.

IV. Extension to Two-Layer RSMA

Here, we extend the proposed design to the two-layer RSMA scenario, which is commonly used in the multi-group multi-cast communication scenario.

In a two-layer RSMA network, the L users are separated into I individual groups denoted as $\mathcal{I} = \{1,\ldots,I\}$ and group-i contains $\mathcal{L}_i$ users with $\bigcup_{i\in\mathcal{I}}\mathcal{L}_i = \mathcal{L}$. User-$l$ splits its message M_l into three sections, namely, an inter-group component M_l^c, an inner-group component M_l^i, and a private component M_l^p. Then, $\{M_l^c\,|l\in\mathcal{L}\}$ are wrapped into a common message M_c, which is encoded into a common signal s_c using a codebook shared by all users and is decoded by all users. Then, $\left\{M_l^i\,|l\in\mathcal{L}_i\right\}$ are merged into a common message $M_{c,i}$. $M_{c,i}$ is encoded into an inner-group common signal $s_{c,i}$ using a codebook shared by the users in group-i and is decoded by these users. Finally, $\{M_l^p\,|l\in\mathcal{L}\}$ are independently encoded into L signals $s_{p,1},\ldots,s_{p,L}$, which are decoded by the specified users. The overall encoded streams $\mathbf{s} = [s_c, s_{c,1},\ldots,s_{c,I},s_{p,1},\ldots,s_{p,L}]^T \in \mathbb{C}^{(L+I+1)\times 1}$ are linearly precoded with the precoding matrix $\mathbf{W} = [\mathbf{w}_c,\mathbf{w}_{c,1},\ldots,\mathbf{w}_{c,I},\mathbf{w}_{p,1},\ldots,\mathbf{w}_{p,L}]^T \in \mathbb{C}^{N_s\times(L+I+1)}$. Hence, the signal sent from the BS is given as

$$\mathbf{x} = \mathbf{w}_c s_c + \sum_{i\in\mathcal{I}}\mathbf{w}_{c,i}s_{c,i} + \sum_{l\in\mathcal{L}}\mathbf{w}_{p,l}s_{p,l}. \tag{33}$$

Then, the received signal at user-$l, l\in\mathcal{L}$, is given as

$$y_l = \left(\mathbf{g}_l^H + \mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{x} + \mathbf{h}_l^H\mathbf{\Phi}\mathbf{n}_r + n_l. \tag{34}$$

User-l ($l\in\mathcal{L}_i$) employs two layers of SIC to sequentially decode s_c, $s_{c,i}$, and $s_{p,l}$ with s_c being decoded first, $s_{c,i}$ second, and followed by $s_{p,l}$. Then, the signal rates for decoding s_c, $s_{c,i}$, and $s_{p,l}$ at the l-th user are, respectively, given as

$$\begin{aligned}
R_l^c &= \log_2\left(1+\frac{\left|\bar{\mathbf{h}}_l^H\mathbf{w}_c\right|^2}{\sum\limits_{i\in\mathcal{I}}\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{c,i}\right|^2+\sum\limits_{j\in\mathcal{L}}\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{p,j}\right|^2+\tilde{\sigma}_l^2}\right),\\
R_l^{c,i} &= \log_2\left(1+\frac{\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{c,i}\right|^2}{\sum\limits_{i'\in\mathcal{I},i'\neq i}\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{c,i'}\right|^2+\sum\limits_{j\in\mathcal{L}}\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{p,j}\right|^2+\tilde{\sigma}_l^2}\right),\\
R_{p,l} &= \log_2\left(1+\frac{\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{p,l}\right|^2}{\sum\limits_{i'\in\mathcal{I},i'\neq i}\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{c,i'}\right|^2+\sum\limits_{j\in\mathcal{L},j\neq l}\left|\bar{\mathbf{h}}_l^H\mathbf{w}_{p,j}\right|^2+\tilde{\sigma}_l^2}\right),
\end{aligned} \tag{35}$$

where $\tilde{\sigma}_l^2 = \left\|\mathbf{h}_l^H\mathbf{\Phi}\right\|^2\sigma_r^2 + \sigma_l^2$.

Then, the message rates of s_c and $s_{c,i}$ are given by

$$\begin{aligned}
&R_c = \min\{R_l^c\,|l\in\mathcal{L}\},\\
&R_{c,i} = \min\left\{R_l^{c,i}\,|l\in\mathcal{L}_i\right\}, \forall i\in\mathcal{I}.
\end{aligned} \tag{36}$$

Thus, the total achievable rate of the two-layer RSMA network is $R_c + \sum_{i=1}^{I}R_{c,i} + \sum_{l=1}^{L}R_{p,l}$ [7].

A two-layer RSMA transceiver architecture with 4 users is plot in Fig. 3, where user-1/2 is in group-1, user-3/4 is in group-2, respectively. s_c is an inter-group common signal, while $s_{c,1/2}$ is the inner-group common signals for the users in group-1/2 only [7].

The RE optimization in two-layer RSMA is more complicated to handle than the counterpart in the one-layer RSMA. Fortunately, the proposed method can be used here with modifications. Firstly, we use the quadratic transformation to recast the fractional programming into a linear programming. Then,

$$
\begin{aligned}
&\max_{\mathbf{W},\mathbf{\Phi}} \sum_{l=1}^{L}\Bigg\{\frac{2\Re\left\{\left(x_l^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_{p,l}\right\}}{y_l^{(t)}\ln 2}+\log_2\left(1+\frac{\left|x_l^{(t)}\right|^2}{y_l^{(t)}}\right)-\frac{\left|x_l^{(t)}\right|^2}{y_l^{(t)}\ln 2}\left(1+\frac{\sigma_l^2}{y_l^{(t)}+\left|x_l^{(t)}\right|^2}\right)\\
&-\frac{\left|x_l^{(t)}\right|^2\left(\sum\limits_{\substack{i'\in\mathcal{I},\\ i'\neq i}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_{c,i}\right|^2+\sum\limits_{j\in\mathcal{L}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_{p,j}\right|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}\right\|^2\sigma_r^2+\sigma_l^2\right)}{\left|x_l^{(t)}\right|^2\left(\left|x_l^{(t)}\right|^2+y_l^{(t)}\right)\ln 2}\Bigg\}+R_c+\sum_{i=1}^{I}R_{c,i}
\end{aligned}\tag{37a}
$$

$$
\begin{aligned}
&\text{s.t.} R_{c,i}\leq\frac{2\Re\left\{\left(c_{i,l}^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_{c,i}\right\}}{d_{i,l}^{(t)}\ln 2}+\log_2\left(1+\frac{\left|c_{i,l}^{(t)}\right|^2}{d_{i,l}^{(t)}}\right)-\frac{\left|c_{i,l}^{(t)}\right|^2}{d_{i,l}^{(t)}\ln 2}\left(1+\frac{\sigma_l^2}{d_{i,l}^{(t)}+\left|c_{i,l}^{(t)}\right|^2}\right)\\
&-\frac{\left|c_{i,l}^{(t)}\right|^2\left(\sum\limits_{i\in\mathcal{I}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_{c,i}\right|^2+\sum\limits_{j\in\mathcal{L}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_{p,j}\right|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}\right\|^2\sigma_r^2+\sigma_l^2\right)}{d_{i,l}^{(t)}\left(d_{i,l}^{(t)}+\left|c_{i,l}^{(t)}\right|^2\right)\ln 2},\forall i\in\mathcal{I},\forall l\in\mathcal{L}_i,
\end{aligned}\tag{37b}
$$

$$
\begin{aligned}
&R_c\leq\frac{2\Re\left\{\left(a_l^{(t)}\right)^*\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{w}_c\right\}}{b_l^{(t)}\ln 2}-\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}\ln 2}\left(1+\frac{\sigma_l^2}{b_l^{(t)}+\left|a_l^{(t)}\right|^2}\right)\\
&+\log_2\left(1+\frac{\left|a_l^{(t)}\right|^2}{b_l^{(t)}}\right)-\frac{\left|a_l^{(t)}\right|^2\left(\left\|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}\mathbf{F}\right)\mathbf{W}\right\|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}\right\|^2\sigma_r^2+\sigma_l^2\right)}{b_l^{(t)}\left(b_l^{(t)}+\left|a_l^{(t)}\right|^2\right)\ln 2},\forall l\in\mathcal{L}.
\end{aligned}\tag{37c}
$$

$$\text{(12b)}-\text{(12e)}.\tag{37d}$$

$$
\begin{aligned}
&a_l^{(t)}=\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_c^{(t)}, c_{i,l}^{(t)}=\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{c,i}^{(t)}, x_l^{(t)}=\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{p,l}^{(t)},\\
&b_l^{(t)}=\sum_{i\in\mathcal{I}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{c,i}^{(t)}\right|^2+\sum_{j\in\mathcal{L}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{p,j}^{(t)}\right|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\right\|^2\sigma_r^2+\sigma_l^2,\\
&d_{i,l}^{(t)}=\sum_{\substack{i'\in\mathcal{I},\\ i'\neq i}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{c,i'}^{(t)}\right|^2+\sum_{j\in\mathcal{L}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{p,j}^{(t)}\right|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\right\|^2\sigma_r^2+\sigma_l^2,\\
&y_l^{(t)}=\sum_{\substack{i'\in\mathcal{I},\\ i'\neq i}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{c,i'}^{(t)}\right|^2+\sum_{\substack{j\in\mathcal{L},\\ j\neq l}}\left|\left(\mathbf{g}_l^H+\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\mathbf{F}\right)\mathbf{w}_{p,j}^{(t)}\right|^2+\left\|\mathbf{h}_l^H\mathbf{\Phi}^{(t)}\right\|^2\sigma_r^2+\sigma_l^2.
\end{aligned}\tag{38}
$$

in the outer stage optimization, we update κ by using (14). While in the inner stage optimization, by utilizing Lemma 1 and introducing several slack variables, $\tilde{\eta}_{SE}(\mathbf{W},\mathbf{\Phi})$ can be obtained by solving problem (37), with the relevant constants given in (38). Then, we further decouple (37) into two subproblems and obtain $\mathbf{W}$ and $\mathbf{\Phi}$ using the corresponding methods. The whole procedure is similar to Algorithm 2. We omit the details due to space limitation.

V. Simulation Results

Here, we provide representative simulation results to verify the proposed design. As plot in Fig. 4, we assume one BS, one RIS, and 4 users, i.e., $L=4$, without loss of generality. The BS and RIS are deployed at $(10\text{ m}, 0\text{ m}, 10\text{ m})$ and $(0\text{ m}, 50\text{ m}, 10\text{ m})$, while all users are randomly located in a circle with radius 5 m and centered at $(10\text{ m}, 50\text{ m}, 1.5\text{ m})$, respectively.

The following settings are adopted unless specified otherwise: $N_s=8$, $N_r=40$, and the maximum amplitude of the active RIS is $\alpha_{n,\max}=10, \forall n$, [36]. The common rate constraint is 2 bits/s/Hz [8]. Besides, the bandwidth is set as $B=10$ MHz [47]. As for the power consumption model, we set $P_a=30$ dBm, $P_s=40$ dBm, $\varepsilon_s=1/0.9$ for the BS [47], and $P_r=-10$ dBm, $P_{DC}=-5$ dBm, $\varepsilon_r=1/0.8$ for the active RIS [36], respectively. Besides, the noise power is $\sigma_l^2=-80\text{ dBm}, \forall l$, and $\sigma_r^2=-80$ dBm [40]. The path loss is $\text{PL}=\text{PL}_0-10\nu\log_{10}\left(\frac{d}{d_0}\right)$, where d indicates the link distance, and ν means the path loss exponent. Here, we set $\text{PL}_0=-30$ dB and $d_0=1$ m. The exponents of the BS-users links and the RIS-related links are set as 4 and 2.2 [32]. Besides, $\mathbf{F}=\sqrt{\frac{r}{r+1}}\mathbf{F}^{\text{LoS}}+\sqrt{\frac{1}{r+1}}\mathbf{F}^{\text{NLoS}}$, with r being

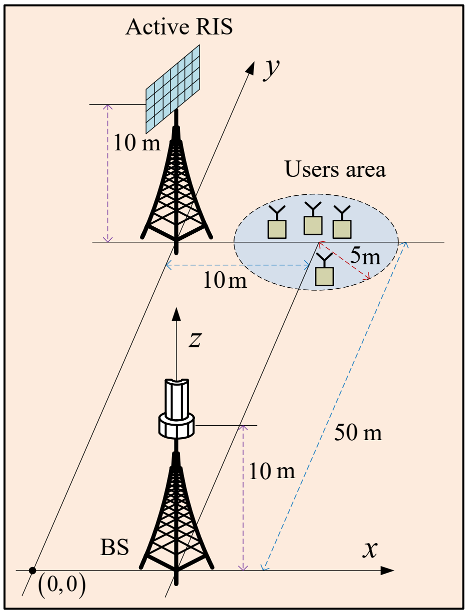


Fig. 4: Simulation scenario.

the Rician factor. Here, $\mathbf{F}^{\mathrm{LoS}}$ denotes the line-of-sight (LoS) component and is given by $\mathbf{F}^{\mathrm{LoS}} = \mathbf{a}_{\mathrm{r}}\mathbf{a}_{\mathrm{t}}^{H}$. When a uniform planar array is utilized, $\mathbf{a}_{\mathrm{t}}$ is given as

$$\begin{aligned}\mathbf{a}_{\mathrm{t}} = \frac{1}{\sqrt{MN}}\Big[1,\ldots,e^{j2\pi\varsigma\left(m\sin\left(\delta_q^{\mathrm{t}}\right)\sin\left(\psi_q^{\mathrm{t}}\right)+n\cos\left(\psi_q^{\mathrm{t}}\right)\right)},\\ \ldots,e^{j2\pi\varsigma\left((M-1)\sin\left(\delta_q^{\mathrm{t}}\right)\sin\left(\psi_q^{\mathrm{t}}\right)+(N-1)\cos\left(\psi_q^{\mathrm{t}}\right)\right)}\Big]^{T},\end{aligned} \tag{39}$$

where m and n are the element indices in horizontal and vertical directions, ς is the normalized distance between adjacent elements, and δ_q^{r} and δ_q^{t} represent the azimuth and elevation angles, respectively. $\mathbf{a}_{\mathrm{r}}$ can be obtained similarly. $\mathbf{F}^{\mathrm{NLoS}}$ denotes the non-LoS component and is modeled as the Rayleigh variable. Besides, the solution accuracy is set as $\epsilon_1 = \epsilon_2 = 10^{-3}$ and the scaling factor is set as $\rho = 0.85$ [24].

Here, we compare the proposed design with several benchmarks: 1) the active RIS with continuous coefficient; 2) conventional single-connected passive RIS; 3) the active RIS-assisted SDMA scheme; 4) the fully-connected RIS [51]; 5) the group-connected RIS [52]. These schemes are labelled as "Proposed scheme, 3 bit", "Proposed scheme, 4 bit", "Continuous scheme", "Passive RIS", "SDMA scheme", "Fully-connected RIS", and "Group-connected RIS", respectively, where 3 bit means $Q_{\alpha/\theta} = 3$ and 4 bit means $Q_{\alpha/\theta} = 4$, respectively.

A. Convergence Behaviour

Firstly, the convergence of Algorithm 1 for different values of N_s, N_r, and $Q_{\alpha/\theta}$ is examined in Fig. 5. From this figure, it can be seen that the RE always increases with the number of iterations, and gradually converges within 20 iterations for various parameters, which demonstrates the efficiency of the PDD method.

Then, we study the convergence of Algorithm 2 for different values of N_s, N_r, and $Q_{\alpha/\theta}$. From Fig. 6, it can be seen that no matter what values of these parameters are selected, the RE increases with the number of iterations, and gradually converges almost within 20 iterations, which verifies the convergence behaviour of Algorithm 2.

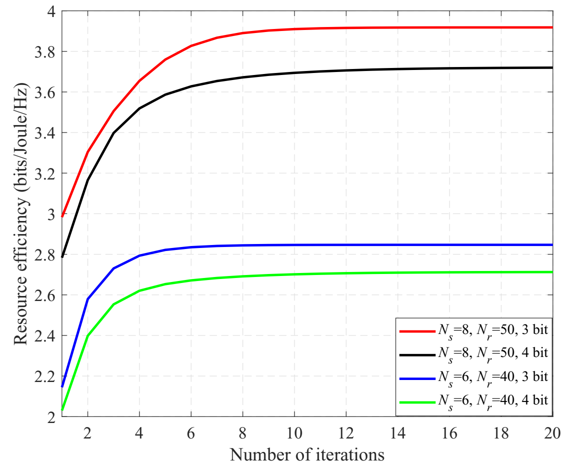


Fig. 5: Convergence of Algorithm 1.

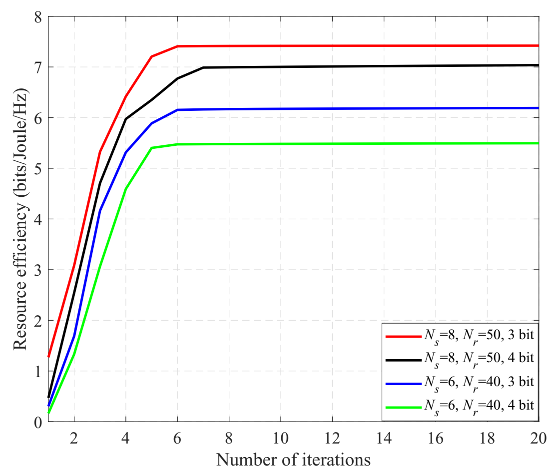


Fig. 6: Convergence of Algorithm 2.

It can be further seen from Figs. 5 and 6 that in the same channel realizations, using more bits to quantize α_n and θ_n is not beneficial to improve the RE, since a larger $Q_{\alpha/\theta}$ will indeed improve the SE, but the power consumption at the RIS will also increase, and thus the RE may be reduced. This observation will be further confirmed by the following simulation results.

B. Performance Evaluation

Now, we evaluate the performances of difference schemes. Here, we set the sum of $P_{s,\max}$ and $P_{r,\max}$, which is denoted by P_{sum}, as a constant equals 10dBW for active RIS, unless specified otherwise. In addition, since the passive RIS has no transmit power consumption and no direct current biasing power consumption, we add the term $N_r P_{DC} + P_{r,\max}/\varepsilon_r$ to the transmit power budget at the BS when using the passive RIS for fair comparison. Thus, we can ensure that P_{sum} is the same value regardless of the active RIS or passive RIS.

The SE performance achieved by Algorithm 2 is shown in Fig. 7, where we set $\delta = 1$. As expected, from Fig. 7, it can be seen that the SE performance always increases with $P_{s,\max}$,

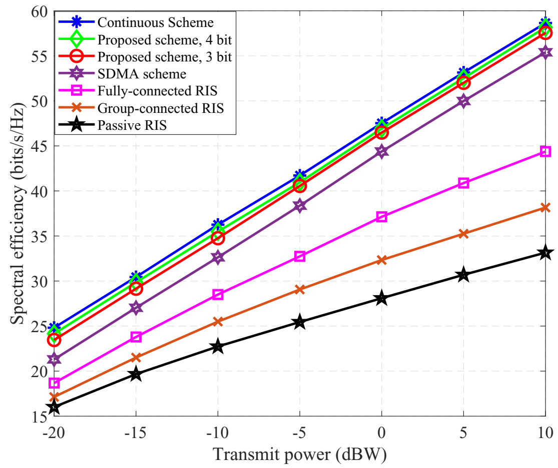


Fig. 7: SE versus $P_{s,\max}$.

and the active RIS with continuous coefficients attains the best performance. However, using the discrete coefficients with 3 or 4 bit resolutions performs very closely to the continuous coefficient case. In addition, it is seen that the active RIS significantly outperforms the other RIS-aided schemes due to its great capability of signal amplification. This observation indicates that active RIS design is effective to reduce the negative impact of the double fading, thus obtaining a higher SE. Besides, the RSMA scheme outperforms the SDMA scheme, since the RSMA technique splits the transmitted message in both the power domain and the spatial domain, and adjusts the split coefficient and power allocation to efficiently suppress the multiuser interference, and thus achieves better performance than the SDMA scheme.

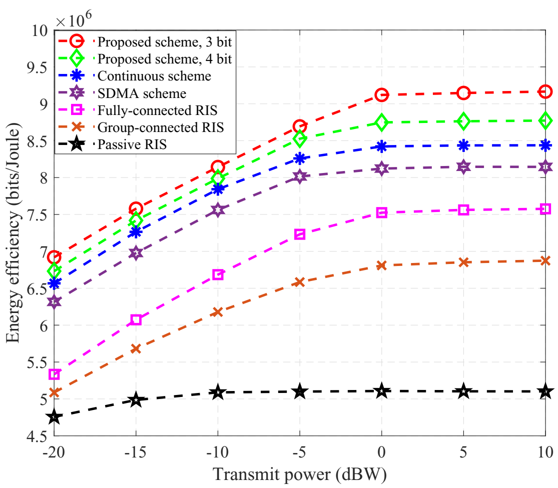


Fig. 8: EE versus $P_{s,\max}$.

Next, we show the corresponding EE performance achieved by Algorithm 2 in Fig. 8, by setting $\beta = 0$ and keeping the other parameters the same as those in Fig. 7. From Fig. 8, it can be seen that EE increases with $P_{s,\max}$ only when $P_{s,\max}$ is smaller than a threshold and then tends to be stable when $P_{s,\max}$ exceeds the threshold. Besides, it is worth noting that EE does not increase with $Q_{\alpha/\theta}$. Although employing the RIS with more quantization bits could achieve higher SE performance, as shown in Fig. 7, it also leads to a higher power consumption thus degrading the EE. Hence, using a RIS with discrete coefficients is more energy efficient than the counterpart with continuous coefficients.

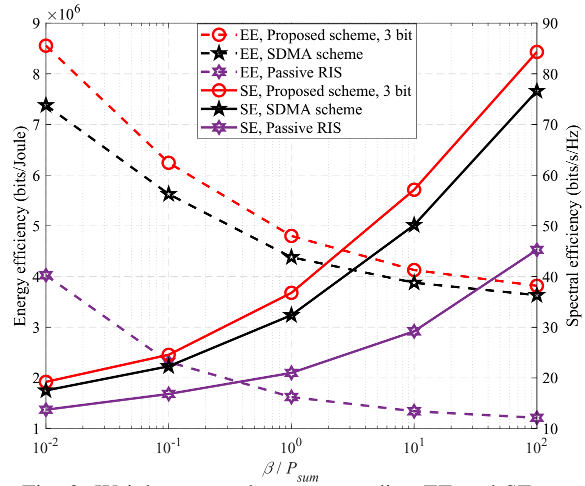


Fig. 9: Weight versus the corresponding EE and SE.

Moreover, Fig. 9 demonstrates the impact of the weight β on the corresponding system EE and SE, with $P_{s,\max} = P_{r,\max} = 0$ dBW. It can be seen that increasing β improves system SE but reduces system EE. This is because that a larger β puts a higher priority on SE and thus allocates more power to maximize SE. On the other hand, when reducing β, we obtain an improved EE but a reduced SE. Fig. 9 shows the ability of the proposed approach to balance the tradeoff between EE and SE by setting a proper β.

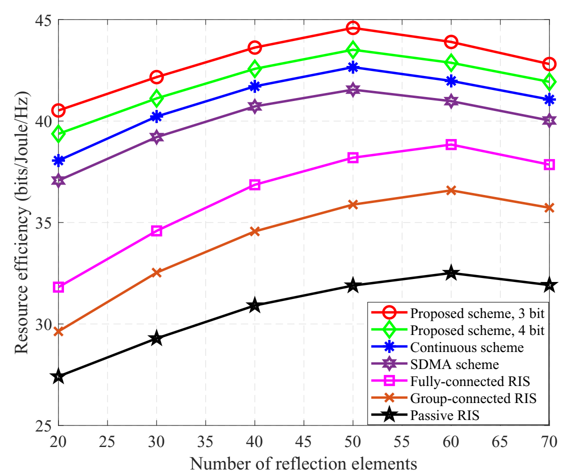


Fig. 10: RE versus N_r.

Besides, we show the RE performances of these schemes versus N_r in Fig. 10, where we set $\beta / P_{sum} = 1$ and $P_{s,\max} = P_{r,\max} = 0$ dBW. From this figure, it can be seen that the RE first increases with N_r then decreases, which is quite different from the results in most related works where larger N_r leads to higher SE performance. This is because a larger N_r may lead to more power consumption at the RIS. Thus there exists a tradeoff in the RE performance w.r.t. N_r.

Next, Fig. 11 plots the obtained RE versus BS-RIS distances, where we assume that the RIS moves along the y-axis from the BS to the users area. In this figure, it is seen that

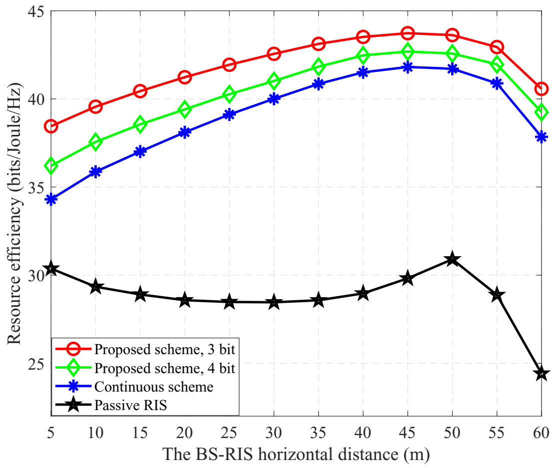


Fig. 11: RE versus the BS-RIS horizontal distance.

the active RIS scheme always outperforms the passive RIS scheme in the considered region. Moveover, for active RIS, the RE increases when RIS moves from the BS to the user area, while for passive RIS, the RE first decreases to a low point and then increases. Then, weather for active RIS or passive RIS, when RIS moves away from the user area, the RE decreases. Thus, deploying the active RIS near the users is beneficial to improve the RE.

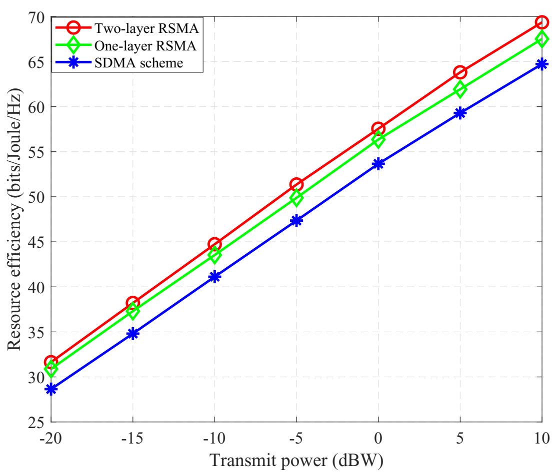


Fig. 12: RE versus $P_{s,\max}$.

Finally, we compare the RE performance of two-layer RSMA with the one-layer RSMA and SDMA schemes, where the 4 users are divided into 2 groups. The results are shown in Fig. 12, with $\beta/P_{sum}=1$ and the other parameters are same as those in Figs. 7 and 8. From this figure, it can be seen that the two-layer RSMA outperforms the other benchmarks in terms of the RE.

VI. Conclusion

In this paper, we proposed a novel infrastructure of active RIS-assisted downlink RSMA network. By adopting the RE as the performance metric to achieve a tradeoff between the SE and EE performances, we formulated the RE optimization problem. By applying a two-stage optimization scheme and proposing an AO algorithm, the non-convex RE maximization problem was decomposed into a two-stage optimization problem, and solved to obtain the BS precoding and reflective BF alternatively by using the PDD method. Simulation results demonstrated that through reconfiguring the wireless communication environment, the active RIS can achieve adjustable tradeoff between the SE and EE, and the proposed active RIS-assisted RSMA scheme outperforms the benchmark schemes.

Appendix A
Proof of Theorem 1

Firstly, for the convergence of the quadratic transformation, [54, Theorem 3] has proved that if problem (13) is nondecreasing and concave, then the original problem (12) is guaranteed to converge. Thus, we mainly focus on the convergence of (13). Specifically, we denote the corresponding objective value of $\eta_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)$ in (13) and the objective of (18) as $\mu\left(\mathbf{W},\mathbf{\Phi}\right)$ and $\mu^{(t)}\left(\mathbf{W},\mathbf{\Phi}\right)$, respectively, where t denotes the number of iterations. Then, for any $\mathbf{W}$ and $\mathbf{\Phi}$, we have

$$\begin{aligned}&\mu\left(\mathbf{W},\mathbf{\Phi}\right)\geq\mu^{(t)}\left(\mathbf{W},\mathbf{\Phi}\right),\\&\mu\left(\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right)=\mu^{(t)}\left(\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right),\end{aligned}\tag{40}$$

when $\{\mathbf{W},\mathbf{\Phi}\}\leftarrow\left\{\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right\}$. Based on this observation and the convexity of (19) and (24), we have

$$\begin{aligned}&\mu\left(\mathbf{W}^{(t+1)},\mathbf{\Phi}^{(t+1)}\right)\geq\mu^{(t)}\left(\mathbf{W}^{(t+1)},\mathbf{\Phi}^{(t+1)}\right)\\&>\mu^{(t)}\left(\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right)=\mu\left(\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right),\end{aligned}\tag{41}$$

where the second inequality holds true since both $\left\{\mathbf{W}^{(t+1)},\mathbf{\Phi}^{(t+1)}\right\}$ and $\left\{\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right\}$ are the optimal solutions and feasible points of (13). (41) suggests that $\left\{\mathbf{W}^{(t+1)},\mathbf{\Phi}^{(t+1)}\right\}$ is better to (13) than $\left\{\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right\}$. Moreover, the sequence $\left\{\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right\}$ is bounded by the constraints (12b)-(12d). Then, according to [56, Proposition 2], there exists a convergent subsequence $\left\{\mathbf{W}^{(t_\gamma)},\mathbf{\Phi}^{(t_\gamma)}\right\}$ with a limit point $\{\mathbf{W}^\star,\mathbf{\Phi}^\star\}$, i.e.,

$$\lim_{\gamma\to+\infty}\left[\mu\left(\mathbf{W}^{(t_\gamma)},\mathbf{\Phi}^{(t_\gamma)}\right)-\mu\left(\mathbf{W}^\star,\mathbf{\Phi}^\star\right)\right]=0.\tag{42}$$

For any t, there exists a γ such that $t_\gamma\leq t\leq t_{\gamma+1}$, then we have

$$\begin{aligned}0=&\lim_{\gamma\to+\infty}\left[\mu\left(\mathbf{W}^{(t_\gamma)},\mathbf{\Phi}^{(t_\gamma)}\right)-\mu\left(\mathbf{W}^\star,\mathbf{\Phi}^\star\right)\right]\\&\leq\lim_{t\to+\infty}\left[\mu\left(\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right)-\mu\left(\mathbf{W}^\star,\mathbf{\Phi}^\star\right)\right]\\&\leq\lim_{\gamma\to+\infty}\left[\mu\left(\mathbf{W}^{(t_{\gamma+1})},\mathbf{\Phi}^{(t_{\gamma+1})}\right)-\mu\left(\mathbf{W}^\star,\mathbf{\Phi}^\star\right)\right]=0,\end{aligned}\tag{43}$$

which means that $\lim_{t\to+\infty}\mu\left(\mathbf{W}^{(t)},\mathbf{\Phi}^{(t)}\right)=\mu\left(\mathbf{W}^\star,\mathbf{\Phi}^\star\right)$. Thus, the convergence of $\eta_{SE}\left(\mathbf{W},\mathbf{\Phi}\right)$ has been proved. Besides, since $P_{tot}\left(\mathbf{W},\mathbf{\Phi}\right)$ is bounded due to (12b)-(12d), (13) is guaranteed to converge, which completes the proof.

References

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

[2] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, “Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution,” *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May. 2016.

[3] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 64, no. 11, pp. 4847–4861, Nov. 2016.

[4] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A Rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4611–4624, Jul. 2016.

[5] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. & Net.*, vol. 1, no. 133, pp. 1–54, 2018.

[6] Z. Lin, M. Lin, J.-B. Wang, T. de Cola, and J. Wang, "Joint beamforming and power allocation for satellite-terrestrial integrated networks with non-orthogonal multiple access," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 657–670, June 2019.

[7] Y. Mao *et al.*, "Rate-splitting multiple access fundamentals, survey, and future research trends," *IEEE Commun. Surv. Tut.*, 2022, doi: 10.1109/COMST.2022.3191937.

[8] A. Mishra, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for downlink multiuser MIMO: Precoder optimization and PHY-layer design," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 874–890, Feb. 2022.

[9] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems – Part II: Joint chunk, power and bit allocation," *IEEE Trans. Commun*, vol. 60, no. 2, pp. 499–509, Feb. 2012.

[10] H. Fu, S. Feng, W. Tang, and D. W. K. Ng, "Robust secure beamforming design for two-user downlink MISO rate-splitting systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8351–8365, Dec. 2020.

[11] Z. Lin, M. Lin, T. de Cola, J.-B. Wang, W.-P. Zhu, and J. Cheng, "Supporting IoT with rate-splitting multiple access in satellite and aerial-integrated networks," *IEEE Inter. Things J.*, vol. 8, no. 14, pp. 11123–11134, Jul. 2021.

[12] Y. Mao, B. Clerckx, J. Zhang, V. O. K. Li, and M. A. Arafah, "Max min fairness of K-user cooperative rate-splitting in MISO broadcast channel with user relaying," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6362–6376, Oct. 2020.

[13] A. Yalcin and Y. Yapici, "Max-min fair beamforming for cooperative multigroup multicasting with rate-splitting," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 254–268, Jan. 2021.

[14] Z. Yang, M. Chen, W. Saad, and M. S. Bahaei, "Optimization of rate allocation and power control for rate splitting multiple access (RSMA)," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 4962–4975, Sep. 2021.

[15] Y. Mao, B. Clerckx, and V. O. K. Li, "Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA," *2018 IEEE Int. Symp. Wireless Commun. Sys.*, 2018, pp. 1–5.

[16] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May. 2021.

[17] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. D. Renzo, and N. A. Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surv. Tut.*, vol. 23, no. 3, pp. 1546–1577, 3rd Quart. 2021.

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

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

[20] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, Jun. 2021.

[21] L. Wei, C. Huang, Q. Guo, Z. Yang, Z. Zhang, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Joint channel estimation and signal recovery for RIS-empowered multiuser communications," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4640–4655, Jul. 2022.

[22] J. Woo, C. Song, and I. Lee, "Sum rate and fairness optimization for intelligent reflecting surface aided multiuser systems," *IEEE Trans. Veh. Tech.*, vol. 70, no. 12, pp. 13436–13440, Dec. 2021.

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

[24] M. M. Zhao *et al.*, "Intelligent reflecting surface enhanced wireless network: Two-timescale beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 2–17, Jan. 2021.

[25] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[26] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop RIS-empowered Terahertz communications: A DRL-based hybrid beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663–1677, Jun. 2021.

[27] S. Shen, B. Clerckx, and R. Murch, "Modeling and architecture design of reconfigurable intelligent surfaces using scattering parameter network analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1229–1243, 2022.

[28] Z. Yang, J. Shi, Z. Li, M. Chen, W. Xu and M. S. Bahaei, "Energy efficient rate splitting multiple access (RSMA) with reconfigurable intelligent surface," *2020 IEEE ICC Workshops*, 2020, pp. 1–6.

[29] Z. Li, W. Chen, Q. Wu, K. Wang, and J. Li, "Joint beamforming design and power splitting optimization in IRS-assisted SWIPT NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 2019–2033, Mar. 2022.

[30] M. Hua, Q. Wu, L. Yang, R. Schober, and H. V. Poor, "A novel wireless communication paradigm for intelligent reflecting surface based symbiotic radio systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 550–565, 2022.

[31] Y. Cai, M. M. Zhao, K. Xu, and R. Zhang, "Intelligent reflecting surface aided full-duplex communication: Passive beamforming and deployment design," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 383–397, Jan. 2022.

[32] H. Niu, Z. Chu, F. Zhou, Z. Zhu, M. Zhang, and K. K. Wong, "Weighted sum secrecy rate maximization using intelligent reflecting surface," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6170–6184, Sep. 2021.

[33] S. Hong, J. Park, S. Kim, and J. Choi, "Hybrid beamforming for intelligent reflecting surface aided millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7343–7357, Sep. 2022.

[34] Z. Lin, H. Niu, K. An, Y. Wang, G. Zheng, S. Chatzinotas, and Y. Hu, "Refracting RIS aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 4, pp. 3717–3724, Aug. 2022.

[35] X. Pang, N. Zhao, J. Tang, C. Wu, D. Niyato, and K. K. Wong, "IRS-assisted secure UAV transmission via joint trajectory and beamforming design," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1140–1152, Feb. 2022.

[36] 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.

[37] C. You and R. Zhang, "Wireless communication aided by intelligent reflecting surface: Active or passive?" *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2659–2663, Dec. 2021.

[38] H. Niu, Z. Lin, K. An, X. Liang, Y. Hu, D. Li, and G. Zheng, "Active RIS-assisted secure transmission for cognitive satellite terrestrial networks," *IEEE Trans. Veh. Tech.*, 2022, doi: 10.1109/TVT.2022.3208268.

[39] G. Chen, Q. Wu, C. He, W. Chen, J. Tang, and S. Jin, "Active IRS aided multiple access for energy-constrained IoT systems," *IEEE Trans. Wireless Commun.*, 2022, doi: 10.1109/TWC.2022.3206332.

[40] K. Liu, Z. Zhang, L. Dai, S. Xu, and F. Yang, "Active reconfigurable intelligent surface: Fully-connected or sub-connected?" *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 167–171, Jan. 2022.

[41] H. Niu, Z. Lin, Z. Chu, Z. Zhu, P. Xiao, H. X. Nguyen, I. Lee, and N. Al-Dhahir, "Joint beamforming design for secure RIS-assisted IoT networks," *IEEE Inter. Things J.*, early access, Sep. 2022, doi: 10.1109/JIOT.2022.3210115.

[42] K. Zhi, C. Pan, H. Ren, K. K. Chai, and M. Elkashlan, "Active RIS versus passive RIS: which is superior with the same power budget?" *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1150–1154, May 2022.

[43] Z. Lin, K. An, H. Niu, Y. Hu, S. Chatzinotas, G. Zheng, and J. Wang, "SLNR-based secure energy efficient beamforming in multibeam satellite systems," *IEEE Trans. Aerosp. Electron. Syst.*, early access, Jul. 2022, doi: 10.1109/TAES.2022.3190238.

[44] Z. Lin, M. Lin, B. Champagne, W. P. Zhu, and N. A. Dhahir, "Secure and energy efficient transmission for RSMA-based cognitive satellite terrestrial networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 251–255, Feb. 2021.

[45] Z. Yang, M. Chen, W. Saad, W. Xu, M. S. Bahaei, H. V. Poor, and S. Cui, "Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 665–679, Jan. 2022.

[46] L. You, J. Xiong, Y. Huang, D. W. K. Ng, C. Pan, W. Wang, and X. Gao, "Reconfigurable intelligent surfaces-assisted multiuser MIMO uplink transmission with partial CSI," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5613–5627, Sep. 2021.

[47] L. You *et al.*, "Spectral efficiency and energy efficiency tradeoff in massive MIMO downlink transmission with statistical CSIT," *IEEE Trans. Signal Process.*, vol. 68, pp. 2645–2659, 2020.

[48] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4816–4828, Jul. 2022.

[49] Y. Wu, F. Zhou, W. Wu, Q. Wu, R. Q. Hu, and K. K. Wong, "Multi-objective optimization for spectrum and energy efficiency tradeoff in IRS-assisted CRNs with NOMA," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6627–6642, Aug. 2022.

[50] L. You *et al.*, "Energy efficiency and spectral efficiency tradeoff in RIS-aided multiuser MIMO uplink transmission," *IEEE Trans. Signal Process.*, vol. 69, pp. 1407–1421, 2021.

[51] T. Fang, Y. Mao, S. Shen, Z. Zhu, and B. Clerckx, "Fully connected reconfigurable intelligent surface aided rate-splitting multiple access for multi-user multi-antenna transmission," *2022 IEEE ICC Workshops*, 2022, pp. 675–680.

[52] H. Li, Y. Mao, O. Dizdar and B. Clerckx, "Rate-splitting multiple access for 6G – Part III: Interplay with reconfigurable intelligent surfaces," *IEEE Commun. Lett.*, 2022, doi: 10.1109/LCOMM.2022.3192041.

[53] Q. Shi and M. Hong, "Penalty dual decomposition method for non-smooth nonconvex optimization–Part I: Algorithms and convergence analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 4108–4122, 2020.

[54] K. Shen and W. Yu, "Fractional programming for communication systems–Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 15, 2018.

[55] M. Grant and S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.0 beta, Sep. 2012. Available: http://cvxr.com/cvx.

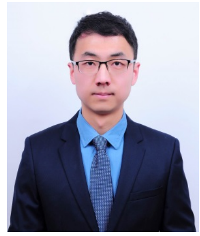
[56] A. A. Nasir *et al.*, "Secrecy rate beamforming for multicell networks with information and energy harvesting," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 677–689, Feb. 2017.

[57] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*. SIAM, 2004.

Hehao Niu is currently a engineer with the Sixty-third Research Institute, National University of Defense Technology, Nanjing, China. His current research interests include smart radio environments/smart reflecting surface, 5G/6G communication networks, unmanned aerial vehicle communication, machine learning, wireless physical layer security, and anti-jamming communication, etc.

Zhi Lin received the B.E. and M.E. degrees in information and communication engineering from the PLA University of Science and Technology and the Ph.D. degree in electronic science and technology from the Army Engineering University of PLA, Nanjing, China, in 2013, 2016, and 2020, respectively. From March 2019 to June 2020, he was a visiting Ph.D. student with the Department of Electrical and Computer Engineering, McGill University, Montréal, Canada. He is currently an Associate Professor with the College of Electronic Engineering, National University of Defense Technology, Hefei, China.

Dr. Lin's research interests include array signal processing, machine learning, physical layer security, reconfigurable intelligent surface and satellite-aerial-terrestrial integrated networks. He was the Symposium Co-Chair of IEEE WCSP'22 and TPC members of IEEE flagship conferences, including IEEE ICC, Globecom, Infocom, VTC, and so on. He was listed in the World's Top 2% Scientists identified by Stanford University in 2022. He was the recipient of the Outstanding Ph.D. Thesis Award of Chinese Institute of Electronics in 2022 and the Macao Young Scholars Fellowship in 2022.



Kang An received the B.E. degree in electronic engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2011, the M.E. degree in communication engineering from PLA University of Science and Technology, Nanjing, China, in 2014, and PhD degree in communication engineering from Army Engineering University, Nanjing, China, in 2017. Since Jan. 2018, he is with the National University of Defense Technology, Nanjing, China, where he is currently an associate professor. His current research interests include satellite communication, reconfigurable intelligent surface, anti-jamming communications, cooperative and cognitive communications, physical-layer security and signal processing for wireless communications.

Jiangzhou Wang (Fellow, IEEE) is a Professor with the University of Kent, U.K. He has published more than 400 papers and four books. His research focuses on mobile communications. He was a recipient of the 2022 IEEE Communications Society Leonard G. Abraham Prize and IEEE Globecom2012 Best Paper Award. He was the Technical Program Chair of the 2019 IEEE International Conference on Communications (ICC2019), Shanghai, Executive Chair of the IEEE ICC2015, London, and Technical Program Chair of the IEEE WCNC2013. He is/was the editor of a number of international journals, including IEEE Transactions on Communications from 1998 to 2013. Professor Wang is a Fellow of the Royal Academy of Engineering, U.K., Fellow of the IEEE, and Fellow of the IET.

Gan Zheng (Fellow, IEEE) received the BEng and the MEng from Tianjin University, Tianjin, China, in 2002 and 2004, respectively, both in Electronic and Information Engineering, and the PhD degree in Electrical and Electronic Engineering from The University of Hong Kong in 2008. He is currently a Professor in Connected Systems in the School of Engineering, University of Warwick, UK. His research interests include machine learning for wireless communications, UAV communications, mobile edge caching, full-duplex radio, and wireless power transfer. He is the first recipient for the 2013 IEEE Signal Processing Letters Best Paper Award, and he also received 2015 GLOBECOM Best Paper Award, and 2018 IEEE Technical Committee on Green Communications & Computing Best Paper Award. He was listed as a Highly Cited Researcher by Thomson Reuters/Clarivate Analytics in 2019. He currently serves as an Associate Editor for IEEE Wireless Communications Letters.

Naofal Al-Dhahir (Fellow, IEEE) is Erik Jonsson Distinguished Professor & ECE Associate Head at UT-Dallas. He earned his PhD degree from Stanford University and was a principal member of technical staff at GE Research Center and AT&T Shannon Laboratory from 1994 to 2003. He is co-inventor of 43 issued patents, co-author of about 520 papers and co-recipient of 5 IEEE best paper awards. He is an IEEE Fellow, AAIA Fellow, received 2019 IEEE COMSOC SPCC technical recognition award, 2021 Qualcomm faculty award, and 2022 IEEE COMSOC RCC technical recognition award. He served as Editor-in-Chief of IEEE Transactions on Communications from Jan. 2016 to Dec. 2019. He is a Fellow of the US National Academy of Inventors and a Member of the European Academy of Sciences and Arts.

Kai-Kit Wong (Fellow, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After graduation, he took up academic and research positions at The University of Hong Kong; Lucent Technologies, Bell-Labs, Holmdel; the Smart Antennas Research Group, Stanford University; and the University of Hull, U.K. He is the Chair of wireless communications with the Department of Electronic and Electrical Engineering, University College London, U.K.

His current research interests include around 5G and beyond mobile communications, including topics such as massive multiple-input multiple-output, full-duplex communications, millimeter-wave communications, edge caching and fog networking, physical layer security, wireless power transfer and mobile computing, V2X communications, and cognitive radios. There are also a few other unconventional research topics that he has set his heart on, for example, fluid antenna communications systems and remote ECG detection. He is a fellow of IET and is also on the editorial board of several international journals. He was a co-recipient of the 2013 IEEE SIGNAL PROCESSING LETTERS Best Paper Award; the 2000 IEEE VTS Japan Chapter Award from the IEEE Vehicular Technology Conference, Japan, in 2000; and a few other international best paper awards. He served as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2005 to 2011 and an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2009 to 2012. He was also a Guest Editor for the Special Issue on Virtual MIMO of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in 2013. He has been serving as a Senior Editor for the IEEE COMMUNICATIONS LETTERS since 2012 and the IEEE WIRELESS COMMUNICATIONS LETTERS since 2016. He has been an Area Editor for the Special Issue on Wireless Communication Theory and Systems-I of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018. He is a Guest Editor for the Special Issue on Physical Layer Security for 5G of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.