Flexible Rate-Splitting Multiple Access with Finite Blocklength

Yuan Wang, *Student Member, IEEE*, Vincent W.S. Wong, *Fellow, IEEE*, and Jiaheng Wang, *Senior Member, IEEE*

***Abstract*—Rate-splitting multiple access (RSMA) is a promising multiple access (MA) technique. It employs rate-splitting (RS) at the transmitter and successive interference cancellation (SIC) at the receiver. Most of the existing works on RSMA assume that all users use SIC to decode the common stream and the blocklength is infinite. The first assumption causes the data rate of the common stream to be limited by the user with the worst channel quality. The second assumption may lead to suboptimal performance in practical systems with finite blocklength. In this paper, we propose a flexible RSMA scheme, which allows the system to decide whether a user should use SIC to decode the common stream or not. We consider the effective throughput as the performance metric, which incorporates the data rate as well as the error performance of RSMA with finite blocklength. We first derive the effective throughput expression and then formulate an effective throughput maximization problem by jointly optimizing the beamforming vectors, transmission data rates, and RS-user selection. We develop an optimal algorithm as well as a low-complexity algorithm for beamforming design. We derive a semi-closed-form solution of the optimal data rates and propose an efficient algorithm for the RS-user selection. Numerical results demonstrate that the proposed algorithm obtains a higher effective throughput than space division multiple access (SDMA), non-orthogonal multiple access (NOMA), and two other RSMA schemes.**

***Index Terms*—Beamforming design, effective throughput, finite blocklength, flexible rate-splitting (RS), rate-splitting multiple access (RSMA).**

I. Introduction

With the increase in data traffic, the conventional orthogonal multiple access (OMA) cannot meet the demands on massive connectivity and high transmission data rates for future wireless communication systems. This has recently spurred significant interests in new multiple access (MA) schemes. Space division multiple access (SDMA) is a common MA scheme adopted in multi-antenna systems, which utilizes the spatial resources to increase the system capacity. However, the performance of SDMA drops quickly in overloaded scenarios due to the severe interference when the number of users is larger than the number of transmit antennas [1], [2]. Non-orthogonal multiple access (NOMA) is another MA candidate, which allows multiple users to share the same time-frequency resource block in the power or code domain [3]–[6]. NOMA can support more users and provide higher aggregate data rates than OMA [7]–[9]. However, since the user order of NOMA is based on single-antenna degraded channel, in multi-antenna systems, NOMA may lead to inefficient use of the spatial dimension and incur a high computational complexity [1], [2].

Rate-splitting multiple access (RSMA) has been proposed as a promising MA scheme [1], [2], [10]. At the transmitter of a downlink RSMA system, the data streams of the selected users are split and encoded into a common stream and multiple private streams. At the receiver, the common stream is decoded first by the selected users. Then, the private streams are decoded by the corresponding users using successive interference cancellation (SIC). Through partially decoding interference and partially treating interference as noise, RSMA can mitigate interference and achieve a higher spectral efficiency and degrees of freedom (DoF) than NOMA and SDMA [1], [2]. Furthermore, RSMA is able to serve a large number of users with heterogeneous quality of service (QoS) requirements, and is robust to different network loads and imperfect channel state information (CSI).

Recently, numerous works have investigated RSMA and its applications. The authors of [1] propose a general framework of multi-layer RSMA and compare RSMA with NOMA and SDMA in terms of the weighted sum rate. In [11], a weighted minimum mean square error (WMMSE) algorithm is proposed to solve the max-min fairness problem of RSMA. In [12], precoding algorithms for RSMA are designed based on WMMSE and successive convex approximation (SCA) to maximize the weighted sum rate and energy efficiency, respectively. RSMA for intelligent reconfigurable surface (IRS)-aided system is investigated in [13] to jointly optimize the precoders, message splitting, and phase shifts of the IRS. In [14]–[16], RSMA and multigroup multicasting are employed to guarantee the diverse requirements of different user groups. The sum rate performance of RSMA is evaluated under imperfect CSI in [17], [18].

In most of the existing works of RSMA, e.g., [2], [10]–[15], [17], [18], there are two common assumptions. The first

The work of Y. Wang and J. Wang was supported by the National Key R&D Program of China under Grant 2021YFB2900300, the National Natural Science Foundation of China under Grants U22B2006, 61971130 and 61720106003, the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grant BK20212001, the Key Technologies R&D Program of Jiangsu (Prospective and Key Technologies for Industry) under Grants BE2022067 and BE2022067-5, and the Southeast University-China Mobile Research Institute Joint Innovation Center. The work of V. W.S. Wong was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Digital Research Alliance of Canada. *(Corresponding author: Jiaheng Wang.)*

Y. Wang and J. Wang are with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210018, China, and also with Purple Mountain Laboratories, Nanjing 210018, China (e-mail: wang_yuan@seu.edu.cn; jhwang@seu.edu.cn).

V. W.S. Wong is with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada (e-mail: vincentw@ece.ubc.ca).

assumption is that all the users have to use SIC to decode the common stream. In this case, the data rate of the common stream is limited by the user with the worst channel quality [19], which is adverse to reaping the benefits of RSMA. The second assumption is that the blocklength is infinite and the information rate based on Shannon capacity is often used as the performance metric. This may lead to suboptimal performance in practical systems with finite blocklength.

Some of the recent works have considered finite blocklength or allowed a subset of users not to use SIC. In [16], [20]–[22], practical physical layer coding with finite blocklength, e.g., polar codes, is employed to evaluate the performance of RSMA via simulations, while the beamforming design is still optimized based on infinite blocklength. In [23], the authors aim to maximize the sum throughput for RSMA with finite blocklength in a two-user single-antenna uplink system. In [24]–[26], finite blocklength is considered for downlink multi-antenna RSMA. However, the block error rates (BLERs) of the common stream and private streams are assumed to be the same, which may not be valid in practical systems. Furthermore, in [16], [20]–[26], all the users are assumed to use SIC to decode the common stream. In [1], [19], the authors introduce the generalized RS and hierarchical RS schemes, where a subset of users decode the common stream via SIC. However, the selection of the subset of users is assumed to be fixed. To the best of our knowledge, there is no prior work which has considered the case where the system can decide whether a user should use SIC to decode the common stream or not.

In this work, we investigate the downlink multiple-input single-output (MISO) RSMA systems. We consider a flexible RSMA scheme, where some users, denoted as RS-users, are selected to use SIC to decode the common stream, whereas other users only decode their own private streams by treating the common stream as noise. We also consider finite blocklength and use the effective throughput to evaluate the system performance. The effective throughput incorporates the data rate as well as the error performance [16], [22], [27]. To achieve the maximum effective throughput, we jointly optimize the beamforming vectors, transmission data rates, and RS-user selection. The resulting problem belongs to the class of mixed integer programs (MIPs), which is proved to be NP-hard [28], [29]. We decompose the original problem into three subproblems and propose efficient algorithms to solve them. The main contributions of this paper are summarized as follows.

- We propose a flexible RSMA scheme for the downlink MISO systems. Instead of having all the users to use SIC to decode the common stream, the proposed scheme allows the system to decide whether each user should use SIC or not.
- We consider the effective throughput as the performance metric for RSMA with finite blocklength. The expression of the effective throughput is derived as a function of the channel gains, beamforming vectors, transmission data rates, and RS-user selection variables.
- We formulate an effective throughput maximization problem by jointly optimizing the beamforming vectors, transmission data rates, and RS-user selection. The formulated problem is NP-hard and we decompose it into three subproblems.
- For beamforming design, we develop an optimal algorithm using monotonic optimization. We also propose a low-complexity algorithm to obtain a near-optimal solution. We explore the hidden concavity of the effective throughput with respect to the data rates and derive the semi-closed-form solution of the optimal data rates. An efficient algorithm is proposed to optimize the binary RS-user selection variables, which achieves near-optimal performance with a low computational complexity.
- Simulation results show that the proposed RSMA scheme achieves a higher effective throughput than the SDMA and NOMA schemes. It also outperforms the existing RSMA schemes that assume infinite blocklength or do not consider flexible RS-user selection in terms of the effective throughput.

The rest of this paper is organized as follows. In Section II, we introduce the signal model and formulate an effective throughput maximization problem. In Section III, we propose two algorithms to optimize the beamforming design. In Section IV, we derive the optimal data rates, solve the RS-user selection subproblem, and present a joint algorithm. Numerical results are presented in Section V. Conclusions are drawn in Section VI.

Notations: In this paper, vectors and matrices are expressed in bold type. $\mathbf{1}$ is the all-ones column vector or matrix. $\mathbf{0}$ is the all-zeros vector or matrix. $\mathbf{i}_i$ is the unit vector where the ith entry is equal to 1 and all other entries are equal to 0. $\mathbf{v}^T$ and $\mathbf{v}^H$ denote the transpose and conjugate transpose of vector $\mathbf{v}$, respectively. $\mathbb{R}^M$ and $\mathbb{R}_+^M$ denote the sets of M-dimensional real-valued and non-negative real-valued column vectors, respectively. $\mathbb{C}$ denotes the set of complex numbers. $\mathbb{C}^M$ denotes the set of M-dimensional complex-valued column vectors. $\mathbb{E}\{\cdot\}$ refers to the statistical expectation. $\text{tr}(\cdot)$ and $\text{rank}(\cdot)$ return the trace and rank of a matrix, respectively. $|a|$ denotes the absolute value of scalar a. $|\mathcal{V}|$ denotes the cardinality of set $\mathcal{V}$. $||\mathbf{v}||$ denotes the Euclidean norm of vector $\mathbf{v}$. $\emptyset$ denotes the empty set. The function $Q(x)$ is defined as $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{t^2}{2}) dt$.

II. System Model

A. Signal Model

Consider a downlink MISO system where a base station (BS) is equipped with N_t antennas serving K single-antenna users indexed by set $\mathcal{K} \triangleq \{1, \cdots, K\}$. We assume that the BS has full knowledge of the CSI. Different from most of the existing RSMA schemes where all users have to decode the common stream, in this work, the system can decide a subset of users to decode the common stream via SIC.[1] Depending

[1] When a user employs SIC to decode the common stream, the data rate of the common stream will be limited by the user's channel quality and SIC may cause error propagation, which may lead to the performance loss. When a user does not decode the common stream via SIC, the interference caused by the common stream may degrade the performance. We investigate the tradeoff of using SIC for each user via the flexible RS-user selection.

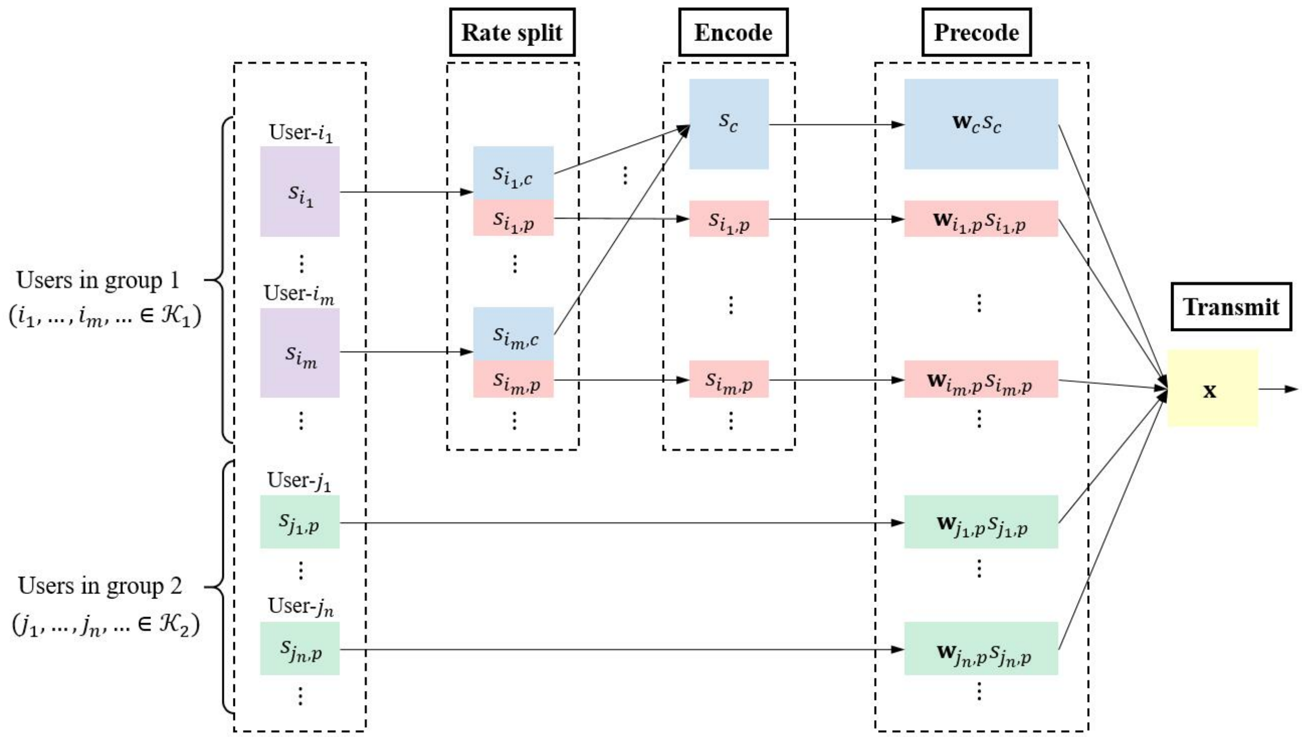


Figure 1: Transmitter of the downlink MISO RSMA system.

on whether SIC is used to decode the common stream, the users are divided into two user groups indexed by $\mathcal{K}_1$ and $\mathcal{K}_2$. Users in group 1, i.e., $\mathcal{K}_1$, decode the common stream via SIC, whereas users in group 2, i.e., $\mathcal{K}_2$, do not employ SIC. Thus, we have $\mathcal{K}_1 \cup \mathcal{K}_2 = \mathcal{K}$ and $\mathcal{K}_1 \cap \mathcal{K}_2 = \emptyset$. For convenience, we define the RS-user selection variable a_k to indicate whether user k is in group 1 or 2:

$$a_k \triangleq \begin{cases} 1, & \text{if } k \in \mathcal{K}_1, \\ 0, & \text{if } k \in \mathcal{K}_2. \end{cases} \tag{1}$$

The intended data stream of user k in group 1 is denoted by $s_k \in \mathbb{C}$. According to the 1-layer RSMA scheme [1], [2], [10], s_k is split into two parts, i.e., the common part $s_{k,c} \in \mathbb{C}$ with data rate $R_{k,c}$ and the private part $s_{k,p} \in \mathbb{C}$ with data rate $R_{k,p}$. The common parts of all the users in group 1 are encoded together into a common data stream s_c using a codebook shared by the users in group 1. The data rate of the common stream s_c is $R_c = \sum_{k\in\mathcal{K}} a_k R_{k,c}$. Users in group 2 do not have the common data part. The intended data stream of user k in group 2 is the private stream $s_{k,p}$ with data rate $R_{k,p}$. The overall data stream to be transmitted is $\mathbf{s} = [s_c \, s_{1,p} \cdots s_{K,p}]^T \in \mathbb{C}^{K+1}$. We denote the beamforming vector for the common stream s_c as $\mathbf{w}_c \in \mathbb{C}^{N_t}$ and the beamforming vector for the private stream $s_{k,p}$ as $\mathbf{w}_{k,p} \in \mathbb{C}^{N_t}$ for $k \in \mathcal{K}$. Thus, the resulting transmit signal after superposition is given by

$$\mathbf{x} = \mathbf{w}_c s_c + \sum_{k\in\mathcal{K}} \mathbf{w}_{k,p} s_{k,p} \in \mathbb{C}^{N_t}. \tag{2}$$

Assume that $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}$. The total power constraint is $|\mathbf{w}_c^H \mathbf{w}_c|^2 + \sum_{k\in\mathcal{K}} |\mathbf{w}_{k,p}^H \mathbf{w}_{k,p}|^2 \leq P$, where P is the total power budget. The received signal at user k is given by

$$y_k = \mathbf{h}_k^H \mathbf{x} + z_k, \; k \in \mathcal{K}, \tag{3}$$

where $\mathbf{h}_k \in \mathbb{C}^{N_t}$ is the channel vector between the BS and user k. $z_k \in \mathbb{C}$ is the additive white Gaussian noise (AWGN) of user k with zero mean. Within the limited bandwidth, we assume that noise z_k has variance σ_k^2. A block diagram of the transmitter at the BS is shown in Fig. 1.

At the receiver of user k in group 1, SIC is employed to decode the intended data streams from y_k. The common data stream s_c is first decoded by treating the interference from all the other data streams as noise. The signal-to-interference-plus-noise ratio (SINR) of decoding the common stream at user k in group 1 is given by

$$\gamma_{k,c} = \frac{|\mathbf{h}_k^H \mathbf{w}_c|^2}{\sum_{i\in\mathcal{K}} |\mathbf{h}_k^H \mathbf{w}_{i,p}|^2 + \sigma_k^2}, \; k \in \mathcal{K}_1. \tag{4}$$

Then, the common part $s_{k,c}$ can be recovered from the common stream s_c using the shared codebook. After SIC, if the common stream is successfully decoded, user k in group 1 will decode its private data stream by treating the private streams of other users as noise. The SINR of decoding the private stream at user k in group 1 is given by

$$\gamma_{k,p,1} = \frac{|\mathbf{h}_k^H \mathbf{w}_{k,p}|^2}{\sum_{i\in\mathcal{K}\setminus\{k\}} |\mathbf{h}_k^H \mathbf{w}_{i,p}|^2 + \sigma_k^2}, \; k \in \mathcal{K}_1. \tag{5}$$

User k in group 2 directly decodes its private stream by treating all other data streams as noise. The SINR of decoding

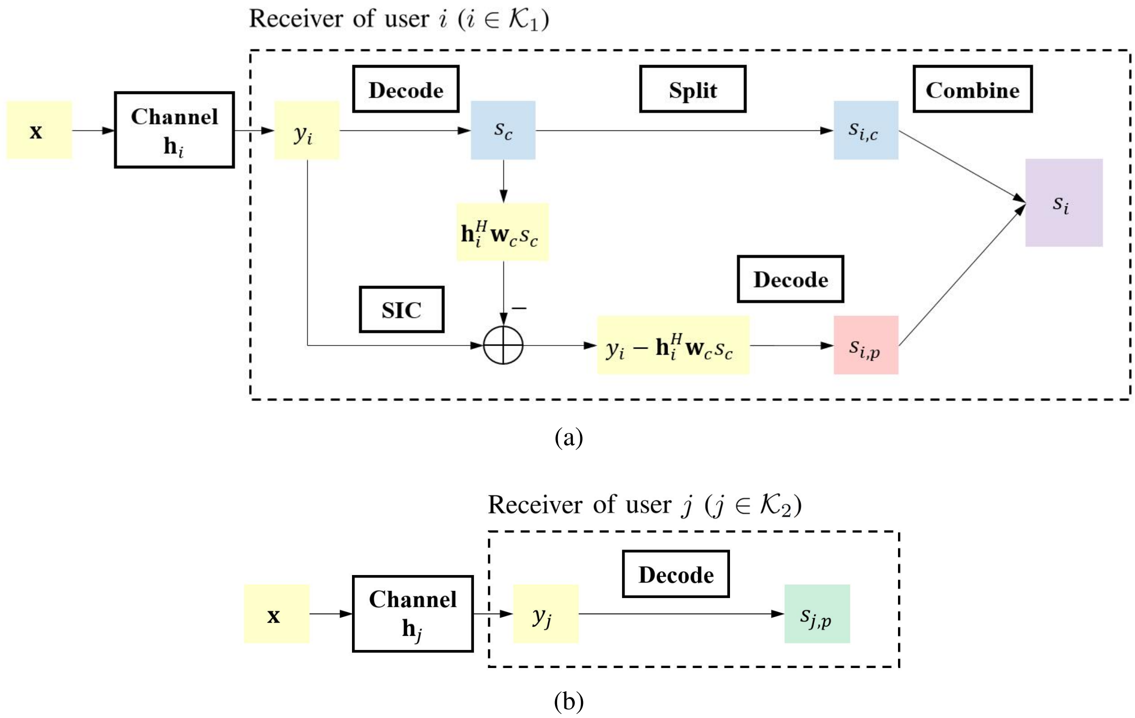


Figure 2: Receiver of the downlink MISO RSMA system. (a) User i in group 1 and (b) user j in group 2.

the private stream at user k in group 2 is given by

$$\gamma_{k,p,2} = \frac{|\mathbf{h}_k^H \mathbf{w}_{k,p}|^2}{|\mathbf{h}_k^H \mathbf{w}_c|^2 + \sum_{i\in\mathcal{K}\setminus\{k\}} |\mathbf{h}_k^H \mathbf{w}_{i,p}|^2 + \sigma_k^2},\ k \in \mathcal{K}_2. \tag{6}$$

From (1), we can combine (5)–(6) and denote the SINR of decoding the private stream at user k as

$$\gamma_{k,p} = \frac{|\mathbf{h}_k^H \mathbf{w}_{k,p}|^2}{(1-a_k)|\mathbf{h}_k^H \mathbf{w}_c|^2 + \sum_{i\in\mathcal{K}\setminus\{k\}} |\mathbf{h}_k^H \mathbf{w}_{i,p}|^2 + \sigma_k^2},\quad k \in \mathcal{K}. \tag{7}$$

A block diagram of the receiver is shown in Fig. 2.

B. Effective Throughput

In most of the existing works of RSMA, the Shannon capacity equation is used to evaluate the performance under the assumption of infinite blocklength. In practice, however, the blocklength is finite. In this paper, we consider finite blocklength and introduce the effective throughput as the performance metric. In this subsection, we derive the expression of the effective throughput.

Given the SINR γ, transmission data rate R, and blocklength L, the BLER in the finite blocklength regime can be approximated as [30]

$$\epsilon \approx Q\left(f(\gamma, R, L)\right), \tag{8}$$

where $f(\gamma,R,L) \triangleq \sqrt{\frac{(\ln 2)^2 L}{1-(1+\gamma)^{-2}}}\left(\log_2(1+\gamma) - R\right)$. Then, the achievable data rate is equal to R with probability $1-\epsilon$ and is equal to zero with probability ϵ. We introduce the effective throughput [16], [22], [27] as the expected achievable data rate:

$$T \triangleq R(1-\epsilon) + 0 \times \epsilon = R(1-\epsilon), \tag{9}$$

which takes into account both the transmission data rate and the BLER.

To derive the expression of the effective throughput, we first analyze the BLER of the users in the two groups. From (8), the BLER of decoding the common stream at user k in group 1 can be approximated as

$$\epsilon_{k,c} \approx Q\left(f(\gamma_{k,c}, R_c, L)\right),\ k \in \mathcal{K}_1, \tag{10}$$

where $\gamma_{k,c}$ is the SINR of the common stream at user k. Note that in the RSMA system, the common stream and private streams are superposed before transmission as shown in (2). Thus, the blocklength should be the same [23]–[26], which is equal to L. The BLER of decoding the private stream at user k in group 1 is given by

$$\begin{aligned}\epsilon_{k,p} &= \mathbb{P}(\mathrm{SIC}_k\,\mathrm{fail})\mathbb{P}(\mathrm{E}_{k,p}\,|\,\mathrm{SIC}_k\,\mathrm{fail}) \\ &\quad + \mathbb{P}(\mathrm{SIC}_k\,\mathrm{succ})\mathbb{P}(\mathrm{E}_{k,p}\,|\,\mathrm{SIC}_k\,\mathrm{succ}) \\ &= \epsilon_{k,c}\mathbb{P}(\mathrm{E}_{k,p}\,|\,\mathrm{SIC}_k\,\mathrm{fail}) + (1-\epsilon_{k,c})\mathbb{P}(\mathrm{E}_{k,p}\,|\,\mathrm{SIC}_k\,\mathrm{succ}),\end{aligned} \tag{11}$$

where $\mathbb{P}(\mathrm{SIC}_k\,\mathrm{fail})$ is the probability that SIC fails at user k, i.e., the common stream is decoded incorrectly. $\mathbb{P}(\mathrm{SIC}_k\,\mathrm{succ})$ is the probability that SIC is successful at user k, i.e., the common stream is decoded correctly. Thus, we have $\mathbb{P}(\mathrm{SIC}_k\,\mathrm{fail}) = \epsilon_{k,c}$ and $\mathbb{P}(\mathrm{SIC}_k\,\mathrm{succ}) = 1 - \epsilon_{k,c}$. $\mathbb{P}(\mathrm{E}_{k,p}\,|\,\mathrm{SIC}_k\,\mathrm{fail})$ is the BLER of the private stream at user k conditioned on failed SIC. When the common stream is decoded incorrectly at user k, we assume that the private stream will also be decoded incorrectly due to the high level of interference, i.e., $\mathbb{P}(\mathrm{E}_{k,p}\,|\,\mathrm{SIC}_k\,\mathrm{fail}) \approx 1$ [23]. $\mathbb{P}(\mathrm{E}_{k,p}\,|\,\mathrm{SIC}_k\,\mathrm{succ})$

is the BLER of the private stream at user k conditioned on successful SIC, which is given by

$$\mathbb{P}(\mathrm{E}_{k,p} \,|\, \mathrm{SIC}_k \,\mathrm{succ}) \approx Q\left(f(\gamma_{k,p}, R_{k,p}, L)\right), \tag{12}$$

where $\gamma_{k,p}$ is the SINR of the private stream at user k if SIC is successful. Thus, the BLER in (11) can be approximated as

$$\begin{aligned}\epsilon_{k,p} &\approx \epsilon_{k,c} + (1-\epsilon_{k,c})\mathbb{P}(\mathrm{E}_{k,p} \,|\, \mathrm{SIC}_k \,\mathrm{succ}) \\ &= Q\left(f(\gamma_{k,c}, R_c, L)\right) + Q\left(f(\gamma_{k,p}, R_{k,p}, L)\right) \\ &\quad - Q\left(f(\gamma_{k,c}, R_c, L)\right) Q\left(f(\gamma_{k,p}, R_{k,p}, L)\right), \; k \in \mathcal{K}_1.\end{aligned} \tag{13}$$

The BLER of decoding the private stream at user k in group 2 can be approximated as

$$\epsilon_{k,p} \approx Q\left(f(\gamma_{k,p}, R_{k,p}, L)\right), \; k \in \mathcal{K}_2. \tag{14}$$

From (9), the aggregate effective throughput of the system T, which is given by (15) shown at the bottom of this page. The aggregate effective throughput is a function of the beamforming vectors, transmission data rates, and RS-user selection. Since we take into account finite blocklength and the RS-user selection, this function is nonconvex.

C. Problem Formulation

Recall that user k in group 1 receives the common data stream with data rate of $R_{k,c}$. For convenience, we define $R_{k,c} = 0$ for user k in group 2. We denote $\mathbf{R} = \{R_{k,c}, R_{k,p} \mid k \in \mathcal{K}\}$ as the collection of all the transmission data rates $R_{k,c}$ and $R_{k,p}$. We denote $\mathbf{w} = \{\mathbf{w}_c, \mathbf{w}_{k,p} \mid k \in \mathcal{K}\}$ as the collection of all the beamforming vectors and $\mathbf{a} = \{a_k \mid k \in \mathcal{K}\}$ as the collection of all the RS-user selection variables. Then, we formulate the aggregate effective throughput maximization problem to jointly optimize the beamforming vectors, transmission data rates, and RS-user selection:

$$\begin{aligned}
\mathcal{O}_1 : \underset{\mathbf{w}, \mathbf{R}, \mathbf{a}}{\text{maximize}} \quad & T && \text{(16a)}\\
\text{subject to} \quad & a_k \in \{0,1\}, \; k \in \mathcal{K} && \text{(16b)}\\
& |\mathbf{w}_c^H \mathbf{w}_c|^2 + \sum_{k\in\mathcal{K}} |\mathbf{w}_{k,p}^H \mathbf{w}_{k,p}|^2 \le P && \text{(16c)}\\
& a_k R_{k,c} \ge 0, \; k \in \mathcal{K} && \text{(16d)}\\
& (1-a_k) R_{k,c} = 0, \; k \in \mathcal{K} && \text{(16e)}\\
& a_k \left(\log_2(1+\gamma_{k,c}) - \sum_{i\in\mathcal{K}} a_i R_{i,c}\right) \ge 0, \\
& \qquad\qquad\qquad\qquad k \in \mathcal{K} && \text{(16f)}\\
& R_{k,p} \ge 0, \; k \in \mathcal{K} && \text{(16g)}\\
& R_{k,p} \le \log_2(1+\gamma_{k,p}), \; k \in \mathcal{K}. && \text{(16h)}
\end{aligned}$$

In problem $\mathcal{O}_1$, constraint (16c) is the total power constraint. Constraint (16e) guarantees that $R_{k,c} = 0$ for $k \in \mathcal{K}_2$. Constraints (16f) and (16h) guarantee that the transmission data rate does not exceed the Shannon capacity.[2] Note that the data rate of the common stream is limited by the channel capacity of the RS-users with the worst channel quality.

Problem $\mathcal{O}_1$ belongs to the class of MIPs and is NP-hard. To find the global optimal solution, one has to employ the exhaustive search method, which has a high computational complexity. To obtain an efficient solution, we decompose problem $\mathcal{O}_1$ into three subproblems: (a) optimizing the beamforming vectors for fixed RS-user selection and data rates; (b) optimizing the data rates for fixed RS-user selection and beamforming vectors; and (c) optimizing the RS-user selection. For each possible RS-user selection in subproblem (c), subproblems (a) and (b) are solved iteratively. Finally, the solution can be obtained. In the following sections, we first describe how to solve each of the subproblems and then present the joint algorithm.

III. Beamforming Design

In this section, we focus on optimizing the beamforming vectors to maximize the effective throughput. We first present an algorithm to determine the optimal beamforming vectors and then present a low-complexity algorithm which can obtain a close-to-optimal beamforming design.

A. Optimal Beamforming Design

Inspired by the idea of semidefinite relaxation (SDR) [31], we introduce rank-1 matrices $\mathbf{W}_c = \mathbf{w}_c \mathbf{w}_c^H$ and $\mathbf{W}_{k,p} = \mathbf{w}_{k,p}\mathbf{w}_{k,p}^H$ for $k \in \mathcal{K}$, and denote $\mathbf{W} = \{\mathbf{W}_c, \mathbf{W}_{k,p} \mid k \in \mathcal{K}\}$. Thus, the SINRs of the common and private streams of user k can be rewritten as

$$\gamma_{k,c} = \frac{\mathrm{tr}(\mathbf{H}_k \mathbf{W}_c)}{\sum_{i\in\mathcal{K}} \mathrm{tr}(\mathbf{H}_k \mathbf{W}_{i,p}) + \sigma_k^2}, \; k \in \mathcal{K}_1, \tag{17}$$

$$\gamma_{k,p} = \frac{\mathrm{tr}(\mathbf{H}_k \mathbf{W}_{k,p})}{(1-a_k)\mathrm{tr}(\mathbf{H}_k \mathbf{W}_c) + \sum_{i\in\mathcal{K}\setminus\{k\}} \mathrm{tr}(\mathbf{H}_k \mathbf{W}_{i,p}) + \sigma_k^2}, \quad k \in \mathcal{K}, \tag{18}$$

where $\mathbf{H}_k = \mathbf{h}_k \mathbf{h}_k^H$ for $k \in \mathcal{K}$. For the fixed RS-user selection and transmission data rates, problem $\mathcal{O}_1$ can be transformed into

$$\mathcal{O}_2 : \underset{\mathbf{W}}{\text{maximize}} \quad T \tag{19a}$$

[2]According to (8), $R \approx \log_2(1+\gamma) - Q^{-1}(\epsilon)\sqrt{\frac{1-(1+\gamma)^{-2}}{(\ln 2)^2 L}}$ is the channel capacity of finite blocklength, which should not exceed the Shannon capacity of infinite blocklength, i.e., $\log_2(1+\gamma)$. Otherwise, the term $Q^{-1}(\epsilon)$ will become negative, which is improper.

$$\begin{aligned}T &\triangleq \sum_{k\in\mathcal{K}_1} T_{k,c} + \sum_{k\in\mathcal{K}} T_{k,p} = \sum_{k\in\mathcal{K}_1} R_{k,c}(1-\epsilon_{k,c}) + \sum_{k\in\mathcal{K}} R_{k,p}(1-\epsilon_{k,p}) \\ &= \sum_{k\in\mathcal{K}} \left(a_k R_{k,c}\left(1 - Q\left(f(\gamma_{k,c}, R_c, L)\right)\right) + R_{k,p}\left(1 - a_k Q\left(f(\gamma_{k,c}, R_c, L)\right)\right)\left(1 - Q\left(f(\gamma_{k,p}, R_{k,p}, L)\right)\right)\right).\end{aligned} \tag{15}$$

$$\begin{aligned} \text{subject to} \quad & \mathrm{tr}(\mathbf{W}_c) + \sum_{k\in\mathcal{K}} \mathrm{tr}(\mathbf{W}_{k,p}) \le P && \text{(19b)} \\ & \mathbf{W}_c \succeq \mathbf{0} && \text{(19c)} \\ & \mathbf{W}_{k,p} \succeq \mathbf{0},\ k\in\mathcal{K} && \text{(19d)} \\ & \mathrm{rank}(\mathbf{W}_c) \le 1 && \text{(19e)} \\ & \mathrm{rank}(\mathbf{W}_{k,p}) \le 1,\ k\in\mathcal{K} && \text{(19f)} \\ & \sum_{i\in\mathcal{K}_1} R_{i,c} \le \log_2(1+\gamma_{k,c}),\ k\in\mathcal{K}_1 && \text{(19g)} \\ & R_{k,p} \le \log_2(1+\gamma_{k,p}),\ k\in\mathcal{K}, && \text{(19h)} \end{aligned}$$

where constraints (19c) and (19d) restrict matrices $\mathbf{W}_c$ and $\mathbf{W}_{k,p}$ to be positive semidefinite. Constraints (19e) and (19f) guarantee that $\mathbf{w}_c$ and $\mathbf{w}_{k,p}$ can be extracted from $\mathbf{W}_c$ and $\mathbf{W}_{k,p}$. Problem $\mathcal{O}_2$ belongs to the class of general quadratic fractional programs (GQFPs) and can further be transformed into a monotonic optimization problem. We next solve problem $\mathcal{O}_2$ via monotonic optimization and polyblock outer approximation algorithm [32]–[35].

First, we introduce the auxiliary variables $v_{i,c}$ for $i\in\mathcal{K}_1$ and $v_{k,p}$ for $k\in\mathcal{K}$ to bound the SINRs:

$$0 \le v_{i,c} \le \gamma_{i,c} = \frac{g_{i,c}(\mathbf{W})}{l_{i,c}(\mathbf{W})},\ i\in\mathcal{K}_1, \tag{20a}$$

$$0 \le v_{k,p} \le \gamma_{k,p} = \frac{g_{k,p}(\mathbf{W})}{l_{k,p}(\mathbf{W})},\ k\in\mathcal{K}, \tag{20b}$$

where $g_{i,c}(\mathbf{W})$ or $g_{k,p}(\mathbf{W})$ and $l_{i,c}(\mathbf{W})$ or $l_{k,p}(\mathbf{W})$ are the numerator and denominator of the SINRs in (4), (5), and (6):

$$g_{i,c}(\mathbf{W}) = \mathrm{tr}(\mathbf{H}_i\mathbf{W}_c),\ i\in\mathcal{K}_1, \tag{21}$$

$$g_{k,p}(\mathbf{W}) = \mathrm{tr}(\mathbf{H}_k\mathbf{W}_{k,p}),\ k\in\mathcal{K}, \tag{22}$$

$$l_{i,c}(\mathbf{W}) = \sum_{j\in\mathcal{K}} \mathrm{tr}(\mathbf{H}_i\mathbf{W}_{j,p}) + \sigma_i^2,\ i\in\mathcal{K}_1, \tag{23}$$

$$l_{k,p}(\mathbf{W}) = \begin{cases} \sum_{j\in\mathcal{K}\setminus\{k\}} \mathrm{tr}(\mathbf{H}_k\mathbf{W}_{j,p}) + \sigma_k^2, \\ \qquad\qquad \text{if } k\in\mathcal{K}_1, \\ \mathrm{tr}(\mathbf{H}_k\mathbf{W}_c) + \sum_{j\in\mathcal{K}\setminus\{k\}} \mathrm{tr}(\mathbf{H}_k\mathbf{W}_{j,p}) + \sigma_k^2, \\ \qquad\qquad \text{if } k\in\mathcal{K}_2. \end{cases} \tag{24}$$

We denote $\mathbf{v} = [v_{i,c}, i\in\mathcal{K}_1, v_{k,p}, k\in\mathcal{K}]^T \in \mathbb{R}^{|\mathcal{K}_1|+K}$ as the vector of the auxiliary variables. Then, problem $\mathcal{O}_2$ can be equivalently transformed into

$$\begin{aligned} \mathcal{O}_3: \underset{\mathbf{v},\mathbf{W}}{\text{maximize}} \quad & \Gamma(\mathbf{v}) && \text{(25a)} \\ \text{subject to} \quad & 0 \le v_{i,c} \le \frac{g_{i,c}(\mathbf{W})}{l_{i,c}(\mathbf{W})},\ i\in\mathcal{K}_1 && \text{(25b)} \\ & 0 \le v_{k,p} \le \frac{g_{k,p}(\mathbf{W})}{l_{k,p}(\mathbf{W})},\ k\in\mathcal{K} && \text{(25c)} \\ & \log_2(1+v_{i,c}) \ge \sum_{j\in\mathcal{K}_1} R_{j,c},\ i\in\mathcal{K}_1 && \text{(25d)} \\ & \log_2(1+v_{k,p}) \ge R_{k,p},\ k\in\mathcal{K} && \text{(25e)} \\ & \text{constraints (19b)} - \text{(19f)}, \end{aligned}$$

where $\Gamma(\mathbf{v})$ is given by

$$\begin{aligned} \Gamma(\mathbf{v}) = & \sum_{k\in\mathcal{K}_1} R_{k,c}\left(1 - Q\left(f(v_{k,c}, R_c, L)\right)\right) \\ & + \sum_{k\in\mathcal{K}_1} R_{k,p}\left(1 - Q\left(f(v_{k,c}, R_c, L)\right)\right) \\ & \times \left(1 - Q\left(f(v_{k,p}, R_{k,p}, L)\right)\right) \\ & + \sum_{k\in\mathcal{K}_2} R_{k,p}\left(1 - Q\left(f(v_{k,p}, R_{k,p}, L)\right)\right). \end{aligned} \tag{26}$$

The hidden monotonicity of problem $\mathcal{O}_3$ can be obtained from the following result.

Definition 1. [32]–[35] A function $F: \mathbb{R}_+^M \to \mathbb{R}$ is increasing if $F(\mathbf{x}) \le F(\mathbf{y})$ when $0 \preceq \mathbf{x} \preceq \mathbf{y}$.

Theorem 1. *The function $\Gamma(\mathbf{v})$ is an increasing function with respect to vector $\mathbf{v}$.*

Proof. The proof is provided in Appendix A. □

Therefore, problem $\mathcal{O}_3$ is a standard monotonic optimization problem. For convenience, we rewrite all the constraints in problem $\mathcal{O}_3$ by the feasible set $\mathbf{v}\in\mathcal{V} = \mathcal{A}\cap\mathcal{B}$, which is the intersection of normal set $\mathcal{A}$ and co-normal set $\mathcal{B}$. The normal set $\mathcal{A} = \mathcal{A}_c\cap\mathcal{A}_p$, where

$$\mathcal{A}_c = \left\{\mathbf{v} \mid 0 \le v_{i,c} \le \frac{g_{i,c}(\mathbf{W})}{l_{i,c}(\mathbf{W})},\ i\in\mathcal{K}_1,\ \text{(19b)} - \text{(19f)}\right\}, \tag{27}$$

$$\mathcal{A}_p = \left\{\mathbf{v} \mid 0 \le v_{k,p} \le \frac{g_{k,p}(\mathbf{W})}{l_{k,p}(\mathbf{W})},\ k\in\mathcal{K},\ \text{(19b)} - \text{(19f)}\right\}. \tag{28}$$

The co-normal set $\mathcal{B} = \mathcal{B}_c\cap\mathcal{B}_p$, where

$$\mathcal{B}_c = \left\{\mathbf{v} \mid \log_2(1+v_{i,c}) \ge R_c,\ i\in\mathcal{K}_1\right\}, \tag{29}$$

$$\mathcal{B}_p = \left\{\mathbf{v} \mid \log_2(1+v_{k,p}) \ge R_{k,p},\ k\in\mathcal{K}\right\}. \tag{30}$$

Since the objective function $\Gamma(\mathbf{v})$ is an increasing function, the global optimal solution is at the upper boundary of the feasible set $\mathcal{V}$. Therefore, problems $\mathcal{O}_2$ and $\mathcal{O}_3$ yield the same optimal solution of the beamforming vectors. However, the boundary is unknown. Inspired by [32]–[35], we approach the boundary through enclosing the feasible set $\mathcal{V}$ by a polyblock. Then, the search for the global optimal solution reduces to choosing the best vertex of the polyblock.

We consider an initial polyblock $\mathcal{U}^{(0)}$, which encloses the feasible set $\mathcal{V}$. The initial vertex is denoted as the vector $\mathbf{u}^{(0)\star} = [u_{i,c}^{(0)\star}, i\in\mathcal{K}_1, u_{k,p}^{(0)\star}, k\in\mathcal{K}]^T \in \mathbb{R}^{|\mathcal{K}_1|+K}$. The projection of $\mathbf{u}^{(0)\star}$ on the upper boundary of $\mathcal{V}$ is denoted as $\varphi(\mathbf{u}^{(0)\star})$, which can be found by the bisection search algorithm. According to [34], cutting off the cone $\{\mathbf{v} \mid \mathbf{v} > \varphi(\mathbf{u}^{(0)\star})\}$ from $\mathcal{U}^{(0)}$ will not exclude any point in set $\mathcal{V}$. Then, a new polyblock is constructed by replacing $\mathbf{u}^{(0)\star}$ with new vertices $\mathbf{u}_1^{(1)}, \dots, \mathbf{u}_{|\mathcal{K}_1|+K}^{(1)}$. We can further refine the enclosing polyblocks by removing those vertices that are not in set $\mathcal{B}$. The obtained polyblock $\mathcal{U}^{(1)}$ still encloses set $\mathcal{V}$. Denote the vertex in $\mathcal{U}^{(1)}$ which maximizes the objective function as $\mathbf{u}^{(1)\star}$. Then, we calculate the projection of $\mathbf{u}^{(1)\star}$ and repeat the above steps.

Following this procedure, we can construct a sequence of polyblocks to approximate the feasible set from the outer:

$$\mathcal{U}^{(0)} \supset \mathcal{U}^{(1)} \supset \mathcal{U}^{(2)} \supset \cdots \supset \mathcal{V}. \tag{31}$$

Algorithm 1 Beamforming design algorithm based on monotonic optimization

1) **Initialization**:
   Initialize the polyblock $\mathcal{U}^{(0)}$ with a vertex set $\mathcal{U}^{(0)} := \{\mathbf{u}^{(0)\star}\}$, where the elements of $\mathbf{u}^{(0)\star}$ are $u_{i,c}^{(0)\star} := \|\mathbf{h}_i\|^2 P/\sigma_i^2$ for $i \in \mathcal{K}_1$ and $u_{k,p}^{(0)\star} := \|\mathbf{h}_k\|^2 P/\sigma_k^2$ for $k \in \mathcal{K}$;
   Set error tolerance $0 < \delta_1 \ll 1$, iteration index $j := 0$, and the maximum number of iterations $j_{\max}$.
2) **Repeat**
3) Obtain the projection of $\mathbf{u}^{(j)\star}$ on the upper boundary of $\mathcal{V}$, i.e., $\varphi(\mathbf{u}^{(j)\star})$, from Algorithm 2. Cut off the cone $\{\mathbf{v} \mid \mathbf{v} > \varphi(\mathbf{u}^{(j)\star})\}$ from $\mathcal{U}^{(j)}$ and construct a new polyblock $\mathcal{U}^{(j+1)}$;
4) Replace the vertex $\mathbf{u}^{(j)\star}$ in $\mathcal{U}^{(j)}$ with $|\mathcal{K}_1| + K$ new vertices $\mathbf{u}_1^{(j+1)}, \cdots, \mathbf{u}_{|\mathcal{K}_1|+K}^{(j+1)}$ and obtain the new vertex set $\mathcal{U}^{(j+1)}$. New vertex $\mathbf{u}_m^{(j+1)}$ is generated by
$$\mathbf{u}_m^{(j+1)} := \mathbf{u}^{(j)\star} - \left(\left[\mathbf{u}^{(j)\star}\right]_m - \left[\varphi(\mathbf{u}^{(j)\star})\right]_m\right)\mathbf{i}_m, \quad m = 1, \cdots, |\mathcal{K}_1| + K,$$
   where $\left[\mathbf{u}^{(j)\star}\right]_m$ is the mth element of $\mathbf{u}^{(j)\star}$ and $\left[\varphi(\mathbf{u}^{(j)\star})\right]_m$ is the mth element of $\varphi(\mathbf{u}^{(j)\star})$.
5) Remove the vertices that are not in $\mathcal{B}$ from $\mathcal{U}^{(j+1)}$;
6) Find the vertex in the set $\mathcal{U}^{(j+1)}$, which maximizes the objective function of problem $\mathcal{O}_3$:
$$\mathbf{u}^{(j+1)\star} := \arg\max_{\mathbf{u}\in\mathcal{U}^{(j+1)}} \mathit{\Gamma}(\mathbf{u}).$$
7) Set $j := j + 1$;
8) **Until** $\frac{\|\mathbf{u}^{(j)\star} - \varphi(\mathbf{u}^{(j)\star})\|}{\|\mathbf{u}^{(j)\star}\|} \le \delta_1$ or $j = j_{\max}$.
9) Obtain the optimal beamforming matrices $\mathbf{W}^\star$ from Algorithm 2.
10) Extract the optimal beamforming vectors $\mathbf{w}^\star$ from $\mathbf{W}^\star$ via rank-1 approximation.
11) **Return** the optimal beamforming vectors $\mathbf{w}^\star$.

Algorithm 2 Projection algorithm based on bisection method

1) **Initialization**:
   Get $\mathbf{u}^{(j)\star}$;
   Set $\lambda_{\min} := 0$, $\lambda_{\max} := 1$, and error tolerance $0 < \delta_2 \ll 1$.
2) **While** $\lambda_{\max} - \lambda_{\min} \ge \delta_2$ **do**
3) $\lambda := (\lambda_{\max} + \lambda_{\min})/2$;
4) Check the feasibility of problem $\mathcal{O}_4$ via, e.g., CVX;
5) **If** problem $\mathcal{O}_4$ is feasible
6) Set $\lambda_{\min} := \lambda$;
7) **Else**
8) Set $\lambda_{\max} := \lambda$;
9) **End if**
10) **End while**
11) **Return** $\varphi(\mathbf{u}^{(j)\star}) := \lambda\mathbf{u}^{(j)\star}$ and beamforming matrices $\mathbf{W}$ obtained by solving problem $\mathcal{O}_4$.

The algorithm terminates at the jth iteration when $\|\mathbf{u}^{(j)\star} - \varphi(\mathbf{u}^{(j)\star})\|/\|\mathbf{u}^{(j)\star}\| \le \delta_1$ is satisfied, where $\mathbf{u}^{(j)\star}$ is the vertex in $\mathcal{U}^{(j)}$ which maximizes the objective of problem $\mathcal{O}_3$, or when j reaches the maximum number of iterations. δ_1 is a positive error tolerance which specifies the accuracy of the algorithm. The details are described in Algorithm 1.

Next, we describe how to find the projection of $\mathbf{u}^{(j)\star}$ on the upper boundary of $\mathcal{V}$, i.e., $\varphi(\mathbf{u}^{(j)\star})$. Mathematically, we have to find the maximum value of $\lambda \in [0, 1]$, which satisfies $\varphi(\mathbf{u}^{(j)\star}) = \lambda\mathbf{u}^{(j)\star} \in \mathcal{V}$. We denote the vector $\mathbf{u}^{(j)\star} = [u_{i,c}^{(j)\star}, i \in \mathcal{K}_1, u_{k,p}^{(j)\star}, k \in \mathcal{K}]^T \in \mathbb{R}^{|\mathcal{K}_1|+K}$. To check whether the condition $\lambda\mathbf{u}^{(j)\star} \in \mathcal{V}$ is satisfied, we solve the following feasibility problem:

$$\mathcal{O}_4 : \underset{\mathbf{W}}{\text{maximize}} \quad 0 \tag{32a}$$
$$\text{subject to} \quad \lambda u_{i,c}^{(j)\star} l_{i,c}(\mathbf{W}) \le g_{i,c}(\mathbf{W}),\ i \in \mathcal{K}_1 \tag{32b}$$
$$\lambda u_{k,p}^{(j)\star} l_{k,p}(\mathbf{W}) \le g_{k,p}(\mathbf{W}),\ k \in \mathcal{K} \tag{32c}$$
$$\text{constraints (19b)} - \text{(19f)},$$

which returns a value 0 if the constraints are feasible (i.e., $\lambda\mathbf{u}^{(j)\star} \in \mathcal{V}$) and $-\infty$ otherwise. The only difficulty of solving problem $\mathcal{O}_4$ is the nonconvex rank constraints (19e) and (19f). After dropping the rank constraints (19e) and (19f), problem $\mathcal{O}_4$ is a convex optimization problem, which can be solved using standard optimization tools such as CVX. Then, the bisection search method can be used to find $\varphi(\mathbf{u}^{(j)\star})$. Finally, the beamforming vectors $\mathbf{w}$ can be extracted from the optimal $\mathbf{W}$ via rank-1 approximation [31]. The details are described in Algorithm 2.

According to [32], [33], [35], the computational complexity of Algorithm 1 grows exponentially with the number of users, which is $O(2^K)$. The algorithm based on monotonic optimization obtains the optimal solution at the cost of high complexity. In the next subsection, we propose a low-complexity algorithm to obtain a close-to-optimal solution.

B. Low-complexity Beamforming Design

In the previous section, we present an optimal beamforming design, which, however, has a high computational complexity especially for a large number of users. In this subsection, we develop a low-complexity solution to problem $\mathcal{O}_3$.

We first transform the objective function $\mathit{\Gamma}(\mathbf{v})$ into a convex form. For any feasible point $\mathbf{v}^{(t)}$, $\mathit{\Gamma}(\mathbf{v})$ can be approximated with the first-order Taylor expansion:

$$\tilde{\mathit{\Gamma}}(\mathbf{v}) = \mathit{\Gamma}(\mathbf{v}^{(t)}) + \sum_{i\in\mathcal{K}_1} \frac{\partial \mathit{\Gamma}(\mathbf{v}^{(t)})}{\partial v_{i,c}}\left(v_{i,c} - v_{i,c}^{(t)}\right) + \sum_{k\in\mathcal{K}} \frac{\partial \mathit{\Gamma}(\mathbf{v}^{(t)})}{\partial v_{k,p}}\left(v_{k,p} - v_{k,p}^{(t)}\right), \tag{33}$$

where the gradients are given in (34) and (35), shown at the bottom of next page.

In constraints (25b) and (25c), $v_{i,c} \ge 0$ and $v_{k,p} \ge 0$ can be dropped due to constraints (25d) and (25e). Then, constraints (25b) and (25c) are equivalent to $v_{i,c}l_{i,c}(\mathbf{W}) \le g_{i,c}(\mathbf{W})$ and $v_{k,p}l_{k,p}(\mathbf{W}) \le g_{k,p}(\mathbf{W})$, which contain the product of two convex functions and can be transformed into (36) and (37), shown at the bottom of next page. In (36) and (37), $I_{i,c}(\mathbf{W}) \triangleq l_{i,c}(\mathbf{W}) - \sigma_k^2$ for $i \in \mathcal{K}_1$ and $I_{k,p}(\mathbf{W}) \triangleq l_{k,p}(\mathbf{W}) - \sigma_k^2$ for $k \in \mathcal{K}$ [32]. After dropping the rank constraints (19e) and (19f), problem $\mathcal{O}_3$ can be transformed into

$$\mathcal{O}_5 : \underset{\mathbf{v},\mathbf{W}}{\text{maximize}} \quad \tilde{\mathit{\Gamma}}(\mathbf{v}) \tag{38a}$$

$$\text{subject to} \quad \log_2(1+v_{i,c}) \geq \sum_{k\in\mathcal{K}_1} R_{k,c},\, i\in\mathcal{K}_1 \quad (38b)$$

$$\log_2(1+v_{k,p}) \geq R_{k,p},\, k\in\mathcal{K} \quad (38c)$$

$$\text{constraints}\,(19\text{b})-(19\text{c}),\,(36),\,\text{and}\,(37).$$

Problem $\mathcal{O}_5$ is a convex optimization problem. However, it is challenging to find a feasible initial solution to problem $\mathcal{O}_5$ due to constraints (38b) and (38c). Thus, we introduce slack variables $\tau_{i,c}$ for $i\in\mathcal{K}_1$ and $\tau_{k,p}$ for $k\in\mathcal{K}$ such that constraints (38b) and (38c) can be relaxed to

$$\log_2(1+v_{i,c}) + \tau_{i,c} \geq \sum_{k\in\mathcal{K}_1} R_{k,c},\, i\in\mathcal{K}_1, \quad (39)$$

$$\log_2(1+v_{k,p}) + \tau_{k,p} \geq R_{k,p},\, k\in\mathcal{K}. \quad (40)$$

We denote $\boldsymbol{\tau} = \{\tau_{i,c}, \tau_{k,p} \mid i\in\mathcal{K}_1, k\in\mathcal{K}\}$ as the collection of all the slack variables. Then, problem $\mathcal{O}_5$ can be transformed into the following penalized problem:

$$\mathcal{O}_6: \underset{\mathbf{v},\mathbf{W},\boldsymbol{\tau}}{\text{maximize}} \quad \tilde{\varGamma}(\mathbf{v}) - \eta\left(\sum_{i\in\mathcal{K}_1}\tau_{i,c} + \sum_{k\in\mathcal{K}}\tau_{k,p}\right) \quad (41a)$$

$$\text{subject to} \quad \text{constraints}\,(19\text{b})-(19\text{c}),\,(36),\,(37),\,(39),\,\text{and}\,(40),$$

where η is a positive penalty coefficient. Problem $\mathcal{O}_6$ can be solved efficiently by standard convex program solvers such as CVX. The details are described in Algorithm 3. Note that we first choose a small penalty coefficient η such that the feasible set is large. After each iteration, we increase η by a factor β_1. Finally, the penalty coefficient η is sufficiently large such that $\tau_{i,c}$ for $i\in\mathcal{K}_1$ and $\tau_{k,p}$ for $k\in\mathcal{K}$ approach 0. After obtaining the optimal $\mathbf{W}$ from Algorithm 3, the beamforming vectors $\mathbf{w}$ can be extracted via rank-1 approximation [31]. The proposed Algorithm 3 converges to a locally optimal solution with a polynomial time computational complexity [32], [33], [35].

Algorithm 3 Beamforming design algorithm based on SCA

1) **Initialization**:
   Set the penalty coefficient η, the increasing factor $\beta_1 > 1$, the error tolerance δ_3, and the maximum number of iterations $t_{\max}$;
   Initialize the iteration index $t := 0$;
   Initialize $\mathbf{v}^{(0)}$ and $\mathbf{W}^{(0)}$.
2) **Repeat**
3) Solve problem $\mathcal{O}_6$ for given $\mathbf{v}^{(t)}$ and $\mathbf{W}^{(t)}$ and denote the obtained solution as $\mathbf{v}^{(t+1)}$ and $\mathbf{W}^{(t+1)}$.
4) Update $t := t+1$ and $\eta := \beta_1\eta$.
5) **Until** $\parallel \mathbf{W}^{(t)} - \mathbf{W}^{(t-1)} \parallel \leq \delta_3$ or $t = t_{\max}$;
6) **Return** the beamforming matrices $\mathbf{W}^{\star} := \mathbf{W}^{(t)}$.

IV. Data Rates Optimization and RS-User Selection

In the previous section, we present two algorithms to determine the beamforming vectors. In this section, we focus on optimizing the transmission data rates and the RS-user selection. Then, we present a joint optimization algorithm for RSMA.

$$\frac{\partial\varGamma(\mathbf{v}^{(t)})}{\partial v_{i,c}} = \sqrt{\frac{L}{2\pi}}\left(R_{i,c} + R_{i,p}\left(1 - Q\left(f(v_{i,p}^{(t)}, R_{i,p}, L)\right)\right)\right) \times \exp\left(-\frac{\left(f(v_{i,c}^{(t)}, R_c, L)\right)^2}{2}\right)\frac{(1+v_{i,c}^{(t)})^2 - 1 - \ln(1+v_{i,c}^{(t)}) + R_c\ln 2}{\left((1+v_{i,c}^{(t)})^2-1\right)^{\frac{3}{2}}},\, i\in\mathcal{K}_1, \quad (34)$$

$$\frac{\partial\varGamma(\mathbf{v}^{(t)})}{\partial v_{k,p}} = \begin{cases} \sqrt{\frac{L}{2\pi}}R_{k,p}\left(1 - Q\left(f(v_{k,c}^{(t)}, R_c, L)\right)\right) \\ \times\exp\left(-\frac{\left(f(v_{k,p}^{(t)}, R_{k,p}, L)\right)^2}{2}\right)\frac{(1+v_{k,p}^{(t)})^2-1-\ln(1+v_{k,p}^{(t)})+R_{k,p}\ln 2}{\left((1+v_{k,p}^{(t)})^2-1\right)^{\frac{3}{2}}}, & \text{if } k\in\mathcal{K}_1, \\ \sqrt{\frac{L}{2\pi}}R_{k,p}\exp\left(-\frac{\left(f(v_{k,p}^{(t)}, R_{k,p}, L)\right)^2}{2}\right)\frac{(1+v_{k,p}^{(t)})^2-1-\ln(1+v_{k,p}^{(t)})+R_{k,p}\ln 2}{\left((1+v_{k,p}^{(t)})^2-1\right)^{\frac{3}{2}}}, & \text{if } k\in\mathcal{K}_2. \end{cases} \quad (35)$$

$$\frac{(v_{i,c} + I_{i,c}(\mathbf{W}))^2}{2} - \frac{\left(v_{i,c}^{(t)}\right)^2 + \left(I_{i,c}(\mathbf{W}^{(t)})\right)^2}{2} - v_{i,c}^{(t)}\left(v_{i,c} - v_{i,c}^{(t)}\right) - I_{i,c}(\mathbf{W}^{(t)})\sum_{j\in\mathcal{K}}\text{tr}\left(\mathbf{H}_i\left(\mathbf{W}_{j,p} - \mathbf{W}_{j,p}^{(t)}\right)\right) \leq g_{i,c}(\mathbf{W}) - v_{i,c}\sigma_i^2,\, i\in\mathcal{K}_1, \quad (36)$$

$$\frac{(v_{k,p} + I_{k,p}(\mathbf{W}))^2}{2} - \frac{\left(v_{k,p}^{(t)}\right)^2 + \left(I_{k,p}(\mathbf{W}^{(t)})\right)^2}{2} - v_{k,p}^{(t)}\left(v_{k,p} - v_{k,p}^{(t)}\right) - I_{k,p}(\mathbf{W}^{(t)})\,(1-a_k)\,\text{tr}\left(\mathbf{H}_k\left(\mathbf{W}_c - \mathbf{W}_c^{(t)}\right)\right) - I_{k,p}(\mathbf{W}^{(t)})\sum_{j\in\mathcal{K}\setminus\{k\}}\text{tr}\left(\mathbf{H}_k\left(\mathbf{W}_{j,p} - \mathbf{W}_{j,p}^{(t)}\right)\right) \leq g_{k,p}(\mathbf{W}) - v_{k,p}\sigma_k^2,\, k\in\mathcal{K}, \quad (37)$$

A. Data Rates Optimization

For the fixed RS-user selection and beamforming vectors, problem $\mathcal{O}_1$ can be transformed into

$$\begin{aligned}\mathcal{O}_7 : \underset{\mathbf{R}}{\text{maximize}} \quad & T && \text{(42a)}\\ \text{subject to} \quad & R_{k,c} \geq 0,\, k \in \mathcal{K}_1 && \text{(42b)}\\ & \sum_{i\in\mathcal{K}_1} R_{i,c} \leq \log_2(1+\gamma_{k,c}),\, k \in \mathcal{K}_1 && \text{(42c)}\\ & 0 \leq R_{k,p} \leq \log_2(1+\gamma_{k,p}),\, k \in \mathcal{K}. && \text{(42d)}\end{aligned}$$

According to (15), the optimal transmission data rate $R_{k,p}$ can be obtained by solving the following subproblem:

$$\begin{aligned}\mathcal{O}_{8,k}\,(k\in\mathcal{K}) : \underset{R_{k,p}}{\text{maximize}} \quad & \Gamma_{k,p}(R_{k,p}) && \text{(43a)}\\ \text{subject to} \quad & 0 \leq R_{k,p} \leq \log_2(1+\gamma_{k,p}), && \text{(43b)}\end{aligned}$$

where $\Gamma_{k,p}(R_{k,p}) \triangleq R_{k,p}\left(1 - Q\left(f(\gamma_{k,p}, R_{k,p}, L)\right)\right)$. We next prove the convexity of problem $\mathcal{O}_{8,k}$ and present a semi-closed-form solution to problem $\mathcal{O}_{8,k}$.

Theorem 2. *Problem $\mathcal{O}_{8,k}$ is a convex optimization problem. The optimal solution is given by*

$$R^{\star}_{k,p} = \begin{cases} R^{\dagger}_{k,p}, & \text{if } \Phi_{k,p}\left(\log_2(1+\gamma_{k,p})\right) < 0 \\ \log_2(1+\gamma_{k,p}), & \text{if } \Phi_{k,p}\left(\log_2(1+\gamma_{k,p})\right) \geq 0, \end{cases} \tag{44}$$

where the constant $R^{\dagger}_{k,p}$ is the root of the function $\Phi_{k,p}(x) = 0$ and

$$\begin{aligned}&\Phi_{k,p}(x) \triangleq 1 - Q\left(f(\gamma_{k,p}, x, L)\right)\\ &- \frac{x\ln 2}{\sqrt{2\pi}} \exp\left(-\frac{\left(f(\gamma_{k,p}, x, L)\right)^2}{2}\right)\sqrt{\frac{L}{1-(1+\gamma_{k,p})^{-2}}}.\end{aligned} \tag{45}$$

Proof. The proof is provided in Appendix B. □

With the optimal transmission data rate $R^{\star}_{k,p}$, the problem of optimizing $R_{k,c}$ can be formulated as

$$\begin{aligned}\mathcal{O}_9 : \underset{\mathbf{R}_c}{\text{maximize}} \quad & T && \text{(46a)}\\ \text{subject to} \quad & R_{k,c} \geq 0,\, k \in \mathcal{K}_1 && \text{(46b)}\\ & \sum_{i\in\mathcal{K}_1} R_{i,c} \leq \log_2(1+\gamma_{k,c}),\, k \in \mathcal{K}_1, && \text{(46c)}\end{aligned}$$

where $\mathbf{R}_c \triangleq \{R_{i,c} \mid i \in \mathcal{K}_1\}$ is the collection of all the transmission data rates of the common stream. Note that it is difficult to decompose problem $\mathcal{O}_9$ since each term $Q\left(f(\gamma_{k,c}, R_c, L_c)\right)$ in the expression of the effective throughput T depends on all the data rates $R_{k,c}$ for $k \in \mathcal{K}_1$. Furthermore, problem $\mathcal{O}_9$ is a nonconvex optimization problem since the Hessian matrix of T with respect to $\mathbf{R}_c$ is neither positive semidefinite nor negative semidefinite. Nevertheless, we find that the effective throughput T is concave with respect to each $R_{k,c}$ for $k \in \mathcal{K}_1$. Next, we prove the hidden concavity and present a low-complexity algorithm.

Theorem 3. *The effective throughput function T is concave with respect to $R_{k,c}$ for $k \in \mathcal{K}_1$.*

Proof. The proof is provided in Appendix C. □

The solution to problem $\mathcal{O}_9$ can be obtained via alternating optimization. Specifically, when optimizing $R_{k,c}$ at the tth iteration, we fix all the other $R_{i,c}$ for $i \in \mathcal{K}_1 \setminus \{k\}$ by $R^{(t-1)}_{i,c}$, which is the optimization result in the $(t-1)$th iteration. Then, similar to Theorem 2, the value of $R_{k,c}$ in the tth iteration can be obtained by

$$R^{(t)}_{k,c} = \begin{cases} R^{\dagger}_{k,c}, & \text{if } \Phi_{k,c}\left(R^{\ddagger}_{k,c}\right) < 0 \\ R^{\ddagger}_{k,c}, & \text{if } \Phi_{k,c}\left(R^{\ddagger}_{k,c}\right) \geq 0, \end{cases} \tag{47}$$

where the constant $R^{\ddagger}_{k,c} \triangleq \min_{i\in\mathcal{K}_1}\{\log_2(1+\gamma_{i,c})\} - \sum_{i\in\mathcal{K}_1\setminus\{k\}} R^{(t-1)}_{i,c}$, the constant $R^{\dagger}_{k,c}$ is the root of the function $\Phi_{k,c}(x) = 0$, and the function $\Phi_{k,c}(x)$ is defined as (48), shown at the bottom of this page. Algorithm 4 shows the details of the data rates optimization algorithm. Assume that Algorithm 4 terminates after I_0 iterations. Then, the computational complexity is $O(KI_0)$.

B. RS-user Selection

In this paper, a flexible RSMA scheme is considered, where the system can decide whether each user should use SIC to decode the common stream. To optimize the binary RS-user selection variables, we consider a heuristic RS-user

$$\begin{aligned}\Phi_{k,c}(x) \triangleq\, & 1 - Q\left(f(\gamma_{k,c}, x + \sum_{i\in\mathcal{K}_1\setminus\{k\}} R^{(t-1)}_{i,c}, L)\right) - \frac{\ln 2}{\sqrt{2\pi}} \sum_{j\in\mathcal{K}_1\setminus\{k\}} \sqrt{\frac{L}{1-(1+\gamma_{j,c})^{-2}}}\\ & \times \left(R^{(t-1)}_{j,c} + R^{\star}_{j,p}\left(1 - Q\left(f(\gamma_{j,p}, R^{\star}_{j,p}, L)\right)\right)\right)\exp\left(-\frac{\left(f(\gamma_{j,c}, x + \sum_{i\in\mathcal{K}_1\setminus\{k\}} R^{(t-1)}_{i,c}, L)\right)^2}{2}\right)\\ & - \frac{\ln 2}{\sqrt{2\pi}}\sqrt{\frac{L}{1-(1+\gamma_{k,c})^{-2}}}\left(x + R^{\star}_{k,p}\left(1 - Q\left(f(\gamma_{k,p}, R^{\star}_{k,p}, L)\right)\right)\right)\exp\left(-\frac{\left(f(\gamma_{k,c}, x + \sum_{i\in\mathcal{K}_1\setminus\{k\}} R^{(t-1)}_{i,c}, L)\right)^2}{2}\right).\end{aligned} \tag{48}$$

Algorithm 4 Transmission data rates optimization

1) **Optimizing the data rates** $R_{k,p}$ **for** $k \in \mathcal{K}$:
Obtain the optimal data rates $R^\star_{k,p}$for $k \in \mathcal{K}$ according to (44).
2) **Optimizing the data rates** $R_{i,c}$ **for** $i \in \mathcal{K}_1$:
Initialize the iteration index $t := 0$ and initialize $R^{(0)}_{i,c}$ for $i \in \mathcal{K}_1$;
Set the error tolerance δ_4 and the maximum number of iterations $t_{\max}$.
Repeat
  Update $t := t + 1$;
  Calculate $R^{(t)}_{i,c}$ for $i \in \mathcal{K}_1$ according to (47);
Until $\sum_{i\in\mathcal{K}_1} |R^{(t)}_{i,c} - R^{(t-1)}_{i,c}| \le \delta_4$ or $t = t_{\max}$.
3) **Return** the optimal transmission data rates $\mathbf{R}^\star := \left\{R^\star_{k,p}, R^{(t)}_{i,c} \mid k \in \mathcal{K}, i \in \mathcal{K}_1\right\}$.

selection algorithm based on simulated annealing. Then, we present a joint algorithm for beamforming design, data rates optimization, and RS-user selection.

Since each user is either in group 1 or 2, there are 2^K possible RS-user selection options in a K-user system. To obtain the optimal solution, one has to employ exhaustive search and calculate the effective throughput of all the 2^K RS-user selection options. This incurs a high computational complexity. A simple policy is to greedily decide each RS-user selection indicator a_k. In each step, we check if changing a_k to $1-a_k$ can increase the effective throughput. If the effective throughput increases, then we change a_k to $1-a_k$. Otherwise, a_k remains unchanged. However, this greedy policy is not guaranteed to converge to the global optimal solution.

To avoid this limitation, we propose a policy based on simulated annealing [36], which is able to escape from local optima via exploration. Specifically, in each step, we check whether changing a_k to $1 - a_k$ can increase the effective throughput. If the effective throughput increases, then we change a_k to $1 - a_k$, which is the same as the greedy policy. However, if the effective throughput does not increase, i.e., $T^{(\mathrm{new})} \le T^{(\mathrm{old})}$, we still change a_k to $1-a_k$ with probability

$$\mathbb{P} = \exp\left(\frac{T^{(\mathrm{new})} - T^{(\mathrm{old})}}{\theta}\right) \in (0, 1], \tag{49}$$

where $T^{(\mathrm{new})}$ and $T^{(\mathrm{old})}$ represent the effective throughput when the RS-user selection indicator of user k is $1-a_k$ and a_k, respectively. The parameter θ is used to control the percentage of changing a_k to $1 - a_k$. We set θ to a relatively large value at the beginning and then decrease it slowly as the number of iterations increases. With a sufficiently large number of iterations, the algorithm converges to the global optimum [36].

The details of the RS-user selection algorithm are described in Algorithm 5. Note that for a given RS-user selection scheme, the effective throughput is obtained via iteratively optimizing the beamforming vectors and transmission data rates according to Algorithm 6. Then, when the RS-user selection search terminates, the beamforming vectors, transmission data rates, and the RS-user selection scheme can be obtained jointly. Thus, Algorithm 5 is also the joint algorithm for the effective throughput maximization problem.

Algorithm 5 RS-user selection algorithm (Joint beamforming design, data rates optimization, and RS-user selection algorithm)

1) **Initialization**:
Generate the RS-user selection indicators randomly as $\mathbf{a}^{(0)} := \left\{a^{(0)}_k \mid k \in \mathcal{K}\right\}$;
Initialize the optimal effective throughput as $T^\star := 0$;
Initialize the old effective throughput as $T^{(\mathrm{old})} := 0$;
Set the precision $\delta_5 > 0$, the convergence parameter N_δ, and the reduction factor β_2;
Initialize the parameter θ and the iteration index $t := 0$.
2) **Repeat**
3) Set $\mathbf{a}^{(\mathrm{new})} := \mathbf{a}^{(t)}$.
4) Set $a^{(\mathrm{new})}_k := 1-a^{(\mathrm{new})}_k$, where $k := \mod(t, K)+1$.
5) Optimize the beamforming vectors, transmission data rates iteratively through Algorithm 6. Obtain the effective throughput $T^{(\mathrm{new})}$.
6) **If** $T^{(\mathrm{new})} > T^{(\mathrm{old})}$
7) Update $T^{(\mathrm{old})} := T^{(\mathrm{new})}$ and $\mathbf{a}^{(t+1)} := \mathbf{a}^{(\mathrm{new})}$;
8) **If** $T^{(\mathrm{new})} > T^\star$
9) Update $T^\star := T^{(\mathrm{new})}$;
10) **End if**
11) **Else**
12) Update $T^{(\mathrm{old})} := T^{(\mathrm{new})}$ and $\mathbf{a}^{(t+1)} := \mathbf{a}^{(\mathrm{new})}$ with probability $\exp\left(\frac{T^{(\mathrm{new})}-T^{(\mathrm{old})}}{\theta}\right)$;
13) Keep the RS-user selection indicators unchanged, i.e., $\mathbf{a}^{(t+1)} := \mathbf{a}^{(t)}$, with probability $1 - \exp\left(\frac{T^{(\mathrm{new})}-T^{(\mathrm{old})}}{\theta}\right)$.
14) **End if**
15) Update $t := t + 1$ and $\theta := \beta_2\theta$.
16) **Until** the change of $T^\star \le \delta_5$ for the last N_δ iterations.
17) **Return** the optimal RS-user selection scheme $\mathbf{a}^{(t)}$ and the corresponding effective throughput, beamforming vectors, and transmission data rates.

Algorithm 6 Iterative beamforming design and data rates optimization

1) **Initialization**:
Set $\mathbf{w}_c := \mathbf{w}_{k,p} := \sqrt{P/\left((K + 1)N_t\right)}\mathbf{1}$ for $k \in \mathcal{K}$.
Set the inner iteration index $q := 0$ and the precision $\delta_6 > 0$.
Initialize $T^{(0)} := 0$.
2) **Repeat**
3) Update $q := q + 1$;
4) Optimize the transmission data rates $\mathbf{R}$ according to Algorithm 4;
5) Optimize the beamforming vectors $\mathbf{w}$ according to Algorithm 1;
6) Calculate the effective throughput $T^{(q)}$ according to (15).
7) **Until** $|T^{(q)} - T^{(q-1)}| \le \delta_6$.
8) **Return** the beamforming vectors, transmission data rates, and the effective throughput $T^{(q)}$.

We next analyze the convergence and computational complexity of the proposed algorithms. In Algorithm 5, as the parameter θ decreases, the probability in (49) decreases accordingly and the search eventually converges [36]. In Algorithm 6, the effective throughput either increases or keeps unchanged in each step. Since the effective throughput is upper bounded

Table I: Proposed RSMA schemes and benchmarks.

	Scheme	Assumption of blocklength	RS-user selection / NOMA user grouping	Beamforming
flexible RSMA-SA-MO (proposed)	RSMA	Finite	Flexible: SA	MO
flexible RSMA-SA-SCA (proposed)	RSMA	Finite	Flexible: SA	SCA
flexible RSMA-ES-MO	RSMA	Finite	Flexible: ES	MO
fixed RSMA-SCA	RSMA	Finite	Fixed: set $a_k = 1$ for $k \in \mathcal{K}$	SCA
fixed RSMA-WMMSE	RSMA	Infinite	Fixed: set $a_k = 1$ for $k \in \mathcal{K}$	WMMSE
fixed SDMA-SCA	SDMA	Finite	Fixed: set $a_k = 0$ for $k \in \mathcal{K}$	SCA
fixed SDMA-WMMSE	SDMA	Infinite	Fixed: set $a_k = 0$ for $k \in \mathcal{K}$	WMMSE
NOMA	NOMA	Finite	Fixed: worst-best user pairing method	SCA

by a finite value, Algorithm 6 converges after a finite number of iterations.

The computational complexity of Algorithm 5 depends on three parts, i.e., the beamforming design in Algorithm 1, the data rates optimization in Algorithm 4, and the RS-user selection algorithm. According to the analysis in the previous sections, the complexity of Algorithms 1 and 4 are $O(2^K)$ and $O(KI_0)$, respectively, where I_0 is the number of iterations in Algorithm 4. Suppose Algorithm 6 converges after I_1 iterations and the RS-user selection search converges after I_2 iterations. The computational complexity of Algorithm 5 is $O\left(I_1 I_2 (2^K + KI_0)\right)$.

V. Performance Evaluation

In this section, we evaluate the performance of the proposed algorithm for RSMA. We consider that there are K users located uniformly in a cell with radius of 200 m and the BS is located in the center. We employ the channel model with the fading coefficient being a zero-mean unit-variance complex Gaussian variable and the path loss being $-(128.1 + 37.6 \log_{10} d_k)$ dB [37], where d_k is the distance between the BS and user k in kilometers. The noise power spectral density is assumed to be -174 dBm/Hz for all users. The following effective throughput is obtained by averaging 100 randomly-generated channel profiles. To compare the performance of the proposed RSMA scheme with other schemes, we consider the benchmarks in Table I.

The flexible RSMA-SA-MO and RSMA-SA-SCA schemes are our proposed RSMA schemes, where the beamforming design is optimized via monotonic optimization (MO) and SCA, respectively, and the RS-user selection is based on simulated annealing (SA). In the flexible RSMA-ES-MO scheme, exhaustive search (ES) is employed to obtain the optimal RS-user selection and MO is employed to obtain the optimal beamforming design. In the fixed RSMA-SCA and RSMA-WMMSE schemes, we consider the fixed 1-layer RSMA [1], [11], where all the users have to decode the common stream via SIC, i.e., setting $a_k = 1$ for $k \in \mathcal{K}$. SCA and WMMSE are employed to obtain the beamforming vectors under the assumptions of finite blocklength and infinite blocklength, respectively.

The SDMA system can be obtained by turning off the common stream of RSMA, i.e., setting $a_k = 0$ for $k \in \mathcal{K}$. In the fixed SDMA-SCA and SDMA-WMMSE schemes, SCA and WMMSE are employed to obtain the beamforming vectors under the assumptions of finite blocklength and infinite blocklength, respectively. We also consider a power-domain NOMA scheme, where we fix the user grouping and the decoding order via the worst-best user pairing method [38]. In each step, the user with the worst channel quality is paired with the one with the best channel quality to form a two-user group. At the receiver, SIC is used in each group to decode the intended data streams. In all the RSMA, SDMA, and NOMA schemes, the transmission data rates are obtained through the semi-closed-form expressions.

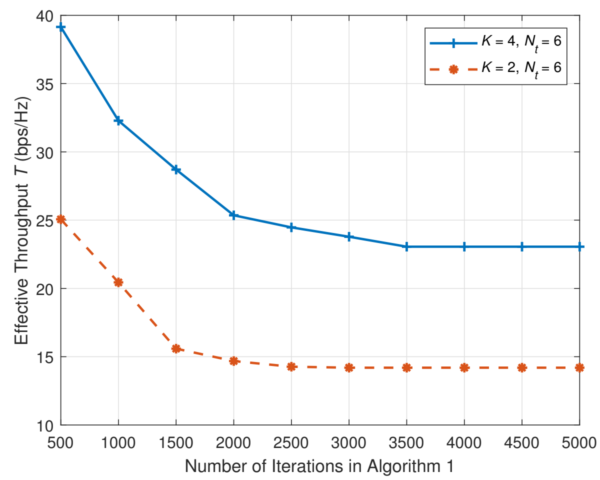


Figure 3: Convergence of the proposed beamforming design algorithm based on monotonic optimization (Algorithm 1).

The results of the convergence of the proposed beamforming design algorithms, i.e., Algorithms 1 and 3, are illustrated in Figs. 3 and 4, respectively. We set the number of antennas N_t to 6 and consider a system with either 2 or 4 users. In Algorithm 1, the feasible set is approximated via a set of outer polyblocks. Therefore, the maximum effective throughput is first achieved by an infeasible solution. After approximately 1500 and 3500 iterations for $K = 2$ and $K = 4$, respectively, the effective throughput converges. One can find that Algorithm 3 converges much faster after approximately 3 and 5 iterations for $K = 2$ and $K = 4$, respectively.

In Fig. 5, we show the convergence of the proposed joint algorithm, i.e., Algorithm 5, via plotting the effective throughput versus the number of iterations. Each iteration here represents a change of the RS-user selection indicator. From Fig. 5, one can find that Algorithm 5 achieves convergence quickly after no more than 14 iterations, i.e., searching 14 possible RS-

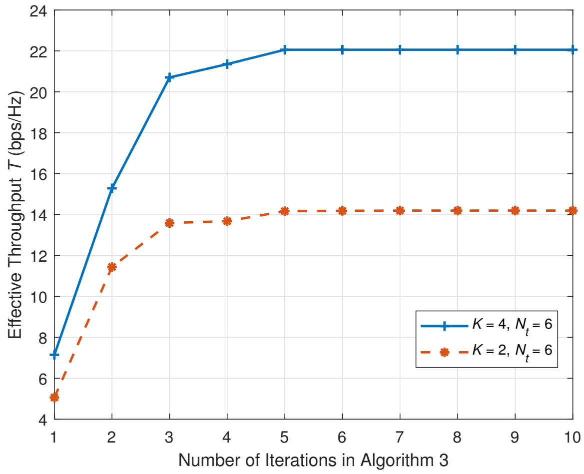


Figure 4: Convergence of the proposed beamforming design algorithm based on SCA (Algorithm 3).

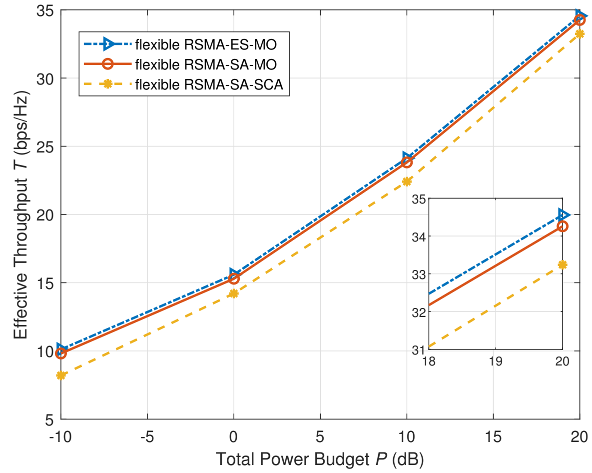


Figure 6: Comparison of the effective throughput versus the total power budget. We set $K = 4$, $N_t = 6$, and $L = 100$.

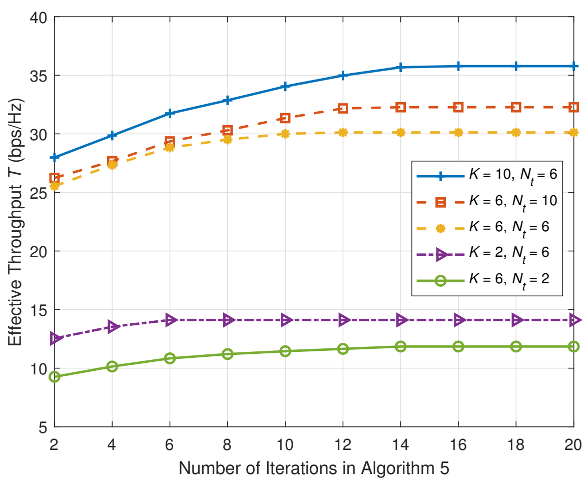


Figure 5: Convergence of the proposed joint algorithm (Algorithm 5).

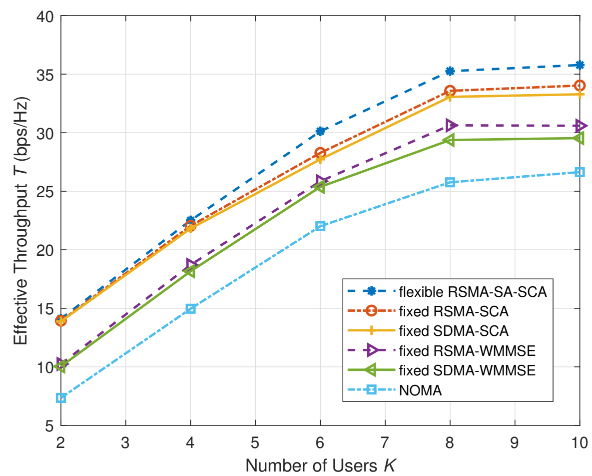


Figure 7: Comparison of the effective throughput versus the number of users. We set $P = 10$ dB, $N_t = 6$, and $L = 100$.

user selection options. For comparison, exhaustive search has to search 2^{10} possible RS-user selection options for a 10-user system, which leads to a relatively high computational complexity.

In Fig. 6, we compare the performance of the proposed flexible RSMA-SA-MO and RSMA-SA-SCA schemes with the benchmark, i.e., the flexible RSMA-ES-MO scheme. The effective throughput is plotted versus the total power budget P. We set the number of users K to 4, the number of antennas N_t to 6, and the blocklength L to 100. One can find that the RSMA-ES-MO scheme shows a minor gain over the two proposed algorithms since exhaustive search is employed to find the optimal RS-user selection. The RSMA-SA-MO scheme outperforms the RSMA-SA-SCA scheme slightly at the expense of a higher computational complexity due to monotonic optimization. Nevertheless, the gap between the RSMA-ES-MO scheme and RSMA-SA-SCA scheme is small. Furthermore, as the total power budget increases, the effective throughput increases due to lower BLER and higher transmission data rates.

Figs. 7 and 8 illustrate the effective throughput versus the number of users K and number of antennas N_t, respectively, with the total power budget P set to 10 dB and the blocklength L set to 100. The proposed flexible RSMA-SA-SCA scheme achieves a higher effective throughput than the fixed RSMA-SCA and RSMA-WMMSE schemes, where all the users have to decode the common stream via SIC. The fixed RSMA-SCA and SDMA-SCA schemes outperform the fixed RSMA-WMMSE and SDMA-WMMSE schemes, respectively, due to the consideration of finite blocklength. Furthermore, the performance of the RSMA schemes are better than SDMA and NOMA due to the superiority of RSMA.

From Fig. 7, one can observe that the effective throughput increases with respect to the number of users. The increasing trend gradually slows down when $K \geq N_t$. Similarly, Fig. 8 shows that, as the number of antennas increases, the effective

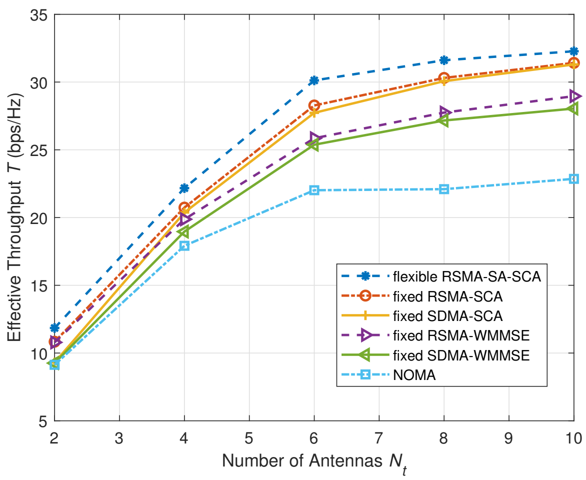


Figure 8: Comparison of the effective throughput versus the number of antennas. We set $P = 10$ dB, $K = 6$, and $L = 100$.

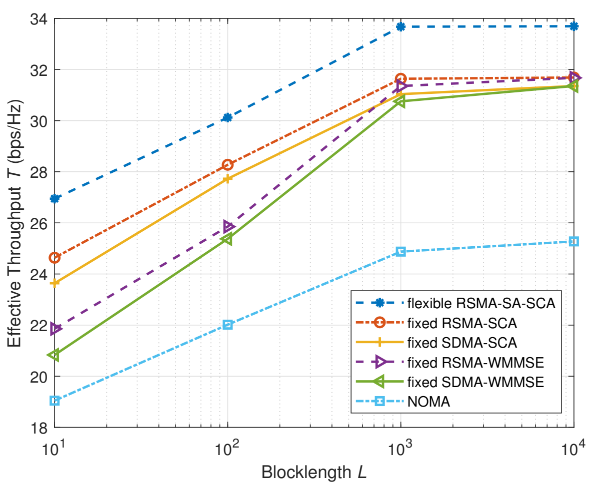


Figure 9: Comparison of the effective throughput versus the blocklength. We set $P = 10$ dB, $K = 6$, and $N_t = 6$.

throughput first has an increasing trend and then keeps almost unchanged when $N_t \geq K$, i.e., more antennas will not benefit the error performance. Furthermore, Figs. 7 and 8 reveal that the superiority of RSMA over SDMA is more obvious in an overloaded scenario, i.e., $K \geq N_t$. This is because SDMA requires more transmit antennas than users such that the multi-user interference can be cancelled, while RSMA can mitigate the interference more efficiently through rate-splitting.

Fig. 9 illustrates the effective throughput versus the blocklength. One can find that as the blocklength increases, the effective throughput increases monotonically due to the lower BLER. Furthermore, when the blocklength is large, e.g., $L = 10000$, the beamforming design based on the assumption of infinite blocklength converges to that considering finite blocklength. Thus, the performance gap between the fixed RSMA-SCA and RSMA-WMMSE schemes (also between the fixed SDMA-SCA and SDMA-WMMSE schemes) narrows.

VI. Conclusion

In this paper, we considered a flexible RSMA scheme for the downlink MISO system, which allows the system to decide whether a user should employ SIC or not. The effective throughput was proposed as the performance metric considering finite blocklength. We derived the expression of the effective throughput. To achieve the maximum effective throughput, we jointly optimized the beamforming vectors, transmission data rates, and RS-user selection. For beamforming design, we proposed an optimal algorithm as well as a low-complexity algorithm. A semi-closed-form solution of the optimal data rates was derived based on the hidden concavity. We also proposed an efficient algorithm for the RS-user selection. Numerical results were presented to show the advantage of the proposed RSMA scheme over SDMA, NOMA, and two other RSMA scheme. The extension of the proposed MISO RSMA scheme to multiple-input multiple-output (MIMO) RSMA schemes and multi-carrier scenarios are interesting directions for future research.

Appendix A
Proof of Theorem 1

For $i \in \mathcal{K}_1$, the first-order derivative of $\varGamma(\mathbf{v})$ with respect to $v_{i,c}$ is given by (50), shown at the bottom of the next page. Since $Q\left(f(v_{i,p}, R_{i,p}, L)\right) < 1$, we have $R_{i,c} + R_{i,p} - R_{i,p}Q\left(f(v_{i,p}, R_{i,p}, L)\right) \geq 0$. Define $G(v_{i,c}) = (1+v_{i,c})^2 - 1 - \ln(1+v_{i,c})$. Then, we have

$$\frac{dG(v_{i,c})}{dv_{i,c}} = 2(1+v_{i,c}) - \frac{1}{1+v_{i,c}} > 0. \tag{51}$$

Thus, $G(v_{i,c})$ is monotonically increasing with respect to $v_{i,c}$, which leads to $G(v_{i,c}) \geq G(0) = 0$ and $\partial\varGamma(\mathbf{v})/\partial v_{i,c} \geq 0$. The inequality $\partial\varGamma(\mathbf{v})/\partial v_{i,c} \geq 0$ holds with equality if and only if $v_{i,c} = 0$. Similarly, $\partial\varGamma(\mathbf{v})/\partial v_{k,p} \geq 0$ holds for $k \in \mathcal{K}$ with equality if and only if $v_{k,p} = 0$. Theorem 1 is proved.

Appendix B
Proof of Theorem 2

The first-order derivative of $\Gamma_{k,p}(R_{k,p})$ with respect to $R_{k,p}$ is given by (52), shown at the bottom of the next page. The second-order derivative of $\Gamma_{k,p}(R_{k,p})$ with respect to $R_{k,p}$ is given by (53), shown at the bottom of the next page, which satisfies $d^2\Gamma_{k,p}(R_{k,p})/dR_{k,p}^2 < 0$ when $0 \leq R_{k,p} \leq \log_2(1+\gamma_{k,p})$. Therefore, $\Gamma_{k,p}(R_{k,p})$ is a concave function and problem $\mathcal{O}_{4,k}$ is a convex optimization problem.

We next analyze the optimal solution to problem $\mathcal{O}_{4,k}$. Due to the concavity, $\Gamma_{k,p}(R_{k,p})$ is first increasing and then decreasing with a peak point $R_{k,p}^\dagger$ which satisfies $\Phi_{k,p}(R_{k,p}^\dagger) = 0$. Since $\Phi_{k,p}(0) = 1 - Q\left(f(\gamma_{k,p}, 0, L)\right) > 0$, we have $R_{k,p}^\dagger > 0$. If $\Phi_{k,p}\left(\log_2(1+\gamma_{k,p})\right) \geq 0$, the peak point $R_{k,p}^\dagger$ is greater than $\log_2(1+\gamma_{k,p})$, and thus, the optimal solution to problem $\mathcal{O}_{4,k}$ is $R_{k,p}^\star = \log_2(1+\gamma_{k,p})$. If $\Phi\left(\log_2(1+\gamma_{k,p})\right) < 0$, the peak point $R_{k,p}^\dagger$ is within the range $(0, \log_2(1+\gamma_{k,p}))$, and thus, the optimal solution to problem $\mathcal{O}_{4,k}$ is $R_{k,p}^\star = R_{k,p}^\dagger$. Theorem 2 is proved.

Appendix C
Proof of Theorem 3

The first-order derivative and the second-order derivative of T with respect to $R_{k,c}$ are given by (54) and (55), respectively, shown at the bottom of this page. The second-order derivative of T with respect to $R_{k,c}$ satisfies $\partial^2 T/\partial R_{k,c}^2 < 0$ when $0 \leq R_c \leq \log_2(1+\gamma_{i,c})$ for $i \in \mathcal{K}_1$. Thus, Theorem 3 is proved.

References

[1] 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. Netw.*, vol. 133, pp. 1–54, May 2018.

[2] B. Clerckx, Y. Mao, R. Schober, E. Jorswieck, and G. Caire, "Is NOMA efficient in multi-antenna networks? A critical look at next generation multiple access techniques," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1310–1343, Jun. 2021.

[3] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surveys & Tuts.*, vol. 22, no. 4, pp. 2192–2235, 4th Quart. 2020.

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

[5] V. W. S. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[6] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[7] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[8] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[9] Y. Wang, J. Wang, D. W. K. Ng, R. Schober, and X. Gao, "A minimum error probability NOMA design," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4221–4237, Jul. 2021.

[10] 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.

[11] H. Joudeh and B. Clerckx, "Robust transmission in downlink multiuser MISO systems: A rate-splitting approach," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6227–6242, Dec. 2016.

[12] Y. Mao, B. Clerckx, and V. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754–8770, Dec. 2019.

[13] Z. Yang, J. Shi, Z. Li, M. Chen, and M. Shikh-Bahaei, "Energy efficient rate splitting multiple access (RSMA) with reconfigurable

$$\begin{aligned}\frac{\partial \Gamma(\mathbf{v})}{\partial v_{k,c}} &= -\left(R_{k,c}+R_{k,p}\left(1-Q\left(f(v_{k,p},R_{k,p},L)\right)\right)\right)\frac{\partial Q\left(f(v_{k,c},R_c,L)\right)}{\partial v_{k,c}}\\ &= \sqrt{\frac{L}{2\pi}}\left(R_{k,c}+R_{k,p}\left(1-Q\left(f(v_{k,p},R_{k,p},L)\right)\right)\right)\exp\left(-\frac{\left(f(v_{k,c},R_c,L)\right)^2}{2}\right)\frac{(1+v_{k,c})^2-1-\ln(1+v_{k,c})+R_c\ln 2}{((1+v_{k,c})^2-1)^{\frac{3}{2}}}.\end{aligned} \tag{50}$$

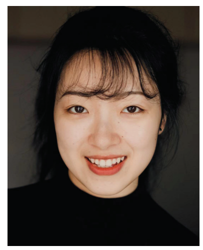
$$\Phi_{k,p}(R_{k,p}) \triangleq \frac{d\Gamma_{k,p}(R_{k,p})}{dR_{k,p}} = 1-Q\left(f(\gamma_{k,p},R_{k,p},L)\right)-\frac{\ln 2}{\sqrt{2\pi}}R_{k,p}\exp\left(-\frac{\left(f(\gamma_{k,p},R_{k,p},L)\right)^2}{2}\right)\sqrt{\frac{L}{1-(1+\gamma_{k,p})^{-2}}}. \tag{52}$$

$$\begin{aligned}\frac{d^2\Gamma_{k,p}(R_{k,p})}{dR_{k,p}^2} =& -\ln 2\sqrt{\frac{L}{2\pi\left(1-(1+\gamma_{k,p})^{-2}\right)}}\exp\left(-\frac{\left(f(\gamma_{k,p},R_{k,p},L)\right)^2}{2}\right)\\ &\times\left(\frac{R_{k,p}(\ln 2)^2L}{1-(1+\gamma_{k,p})^{-2}}\left(\log_2(1+\gamma_{k,p})-R_{k,p}\right)+2\right),\end{aligned} \tag{53}$$

$$\begin{aligned}\frac{\partial T}{\partial R_{k,c}} =& 1-Q\left(f(\gamma_{k,c},R_c,L)\right)-\sum_{i\in\mathcal{K}_1}\left(R_{i,c}+R_{i,p}^{\star}\left(1-Q\left(f(\gamma_{i,p},R_{i,p}^{\star},L)\right)\right)\right)\frac{\partial Q\left(f(\gamma_{i,c},R_c,L)\right)}{\partial R_{k,c}}\\ =& 1-Q\left(f(\gamma_{k,c},R_c,L)\right)-\frac{\ln 2}{\sqrt{2\pi}}\sum_{i\in\mathcal{K}_1}\sqrt{\frac{L}{1-(1+\gamma_{i,c})^{-2}}}\\ &\times\left(R_{i,c}+R_{i,p}^{\star}\left(1-Q\left(f(\gamma_{i,p},R_{i,p}^{\star},L)\right)\right)\right)\exp\left(-\frac{\left(f(\gamma_{i,c},R_c,L)\right)^2}{2}\right),\end{aligned} \tag{54}$$

$$\begin{aligned}\frac{\partial^2 T}{\partial R_{k,c}^2} =& -2\ln 2\sqrt{\frac{L}{2\pi\left(1-(1+\gamma_{k,c})^{-2}\right)}}\exp\left(-\frac{\left(f(\gamma_{k,c},R_c,L)\right)^2}{2}\right)-\frac{1}{\sqrt{2\pi}}\sum_{i\in\mathcal{K}_1}\left(\frac{(\ln 2)^3L^{3/2}}{\left(1-(1+\gamma_{i,c})^{-2}\right)^{3/2}}\right.\\ &\left.\left(R_{i,c}+R_{i,p}^{\star}\left(1-Q\left(f(\gamma_{i,p},R_{i,p}^{\star},L)\right)\right)\right)\left(\log_2(1+\gamma_{i,c})-R_c\right)\exp\left(-\frac{\left(f(\gamma_{i,c},R_c,L)\right)^2}{2}\right)\right).\end{aligned} \tag{55}$$

intelligent surface," in *Proc. IEEE Int. Conf. on Commun. Workshops (ICC Workshops)*, Jun. 2020.
[14] L. Yin and B. Clerckx, "Rate-splitting multiple access for multigroup multicast and multibeam satellite systems," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 976–990, Feb. 2021.
[15] H. Joudeh and B. Clerckx, "Rate-splitting for max-min fair multigroup multicast beamforming in overloaded systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7276–7289, Nov. 2017.
[16] H. Chen, D. Mi, T. Wang, Z. Chu, and P. Xiao, "Rate-splitting for multicarrier multigroup multicast: Precoder design and error performance," *IEEE Trans. Broadcasting*, vol. 67, no. 3, pp. 619–630, Sep. 2021.
[17] 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.
[18] M. Dai and B. Clerckx, "Multiuser millimeter wave beamforming strategies with quantized and statistical CSIT," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7025–7038, Nov. 2017.
[19] 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.
[20] 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.
[21] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access for downlink multi-antenna communications: Physical layer design and link-level simulations," in *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Aug./Sep. 2020.
[22] L. Yin, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for multigroup multicast cellular and satellite communications: PHY layer design and link-level simulations," in *Proc. IEEE Int. Conf. on Commun. Workshops (ICC Workshops)*, Jun. 2021.
[23] J. Xu, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for short-packet uplink communications: A finite blocklength error probability analysis," *arXiv preprint arXiv:2207.07782*, 2022.
[24] O. Dizdar, Y. Mao, Y. Xu, P. Zhu, and B. Clerckx, "Rate-splitting multiple access for enhanced URLLC and eMBB in 6G," in *Proc. IEEE Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2021.
[25] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access with finite blocklength for short-packet and low-latency downlink communications," *IEEE Trans. Veh. Technol.*, early access, 2022.
[26] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, "Max-min fairness of rate-splitting multiple access with finite blocklength communications," *arXiv preprint arXiv:2207.05614*, 2022.
[27] Y. Wang, J. Wang, V. W. S. Wong, and X. You, "Effective throughput maximization of NOMA with practical modulations," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1084–1100, Apr. 2022.
[28] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 9580–8594, Dec. 2016.
[29] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.
[30] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[31] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.
[32] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink MISO OFDMA-URLLC systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7184–7200, Nov. 2020.
[33] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.
[34] Y. J. Zhang, L. Qian, and J. Huang, "Monotonic optimization in communication and networking systems," *Found. Trends Netw.*, vol. 7, no. 1, pp. 1–75, 2012.
[35] Y. Sun, D. W. K. Ng, J. Zhu, and R. Schober, "Robust and secure resource allocation for full-duplex MISO multicarrier NOMA systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4119–4137, Sep. 2018.
[36] W. L. Goffe, G. D. Ferrier, and J. Rogers, "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, vol. 60, no. 1-2, pp. 65–99, 1994.
[37] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," Mar. 2010.
[38] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.



Yuan Wang (S'19) received the B.E. degree in information engineering from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2017, where she is currently pursuing the Ph.D. degree in information and communication engineering. From 2021 to 2022, she was a visiting Ph.D. student at the University of British Columbia (UBC), Vancouver, BC, Canada. Her research interests include next-generation multiple access, machine learning, and optimization theory for wireless communications.

Vincent W.S. Wong (S'94, M'00, SM'07, F'16) received the B.Sc. degree from the University of Manitoba, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Canada, in 1996, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, Canada, in 2000. From 2000 to 2001, he worked as a systems engineer at PMC-Sierra Inc. (now Microchip Technology Inc.). He joined the Department of Electrical and Computer Engineering at UBC in 2002 and is currently a Professor. His research areas include protocol design, optimization, and resource management of communication networks, with applications to 5G/6G wireless networks, Internet of things, mobile edge computing, smart grid, and energy systems. He received the Best Paper Award at the *IEEE ICC* 2022 and *IEEE GLOBECOM* 2020. Since January 2023, Dr. Wong is the Editor-in-Chief of *IEEE Transactions on Wireless Communications*. He has served as an Area Editor of *IEEE Transactions on Communications* and *IEEE Open Journal of the Communications Society*, an Associate Editor of *IEEE Transactions on Mobile Computing* and *IEEE Transactions on Vehicular Technology*, and a Guest Editor of *IEEE Journal on Selected Areas in Communications*, *IEEE Internet of Things Journal*, and *IEEE Wireless Communications*. He was a Tutorial Co-Chair of *IEEE GLOBECOM*'18, a Technical Program Co-chair of *IEEE VTC*2020-*Fall* and *IEEE SmartGridComm*'14, and a Symposium Co-chair of *IEEE ICC*'18, *IEEE SmartGridComm* ('13, '17) and *IEEE GLOBECOM*'13. He is the Chair of the IEEE Vancouver Joint Communications Chapter and has served as the Chair of the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications. He was an IEEE Communications Society Distinguished Lecturer (2019−2020).

Jiaheng Wang (M'10, SM'14) received the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2010, and the B.E. and M.S. degrees from the Southeast University, Nanjing, China, in 2001 and 2006, respectively.

He is currently a Full Professor at the National Mobile Communications Research Laboratory (NCRL), Southeast University, Nanjing, China. He is also with Purple Mountain Laboratories Nanjing, China. From 2010 to 2011, he was with the Signal Processing Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. He also held visiting positions at the Friedrich Alexander University Erlangen-Nürnberg, Nürnberg, Germany, and the University of Macau, Macau. His research interests are mainly on communication systems, wireless networks and blockchain technologies.

Dr. Wang has published more than 100 articles on international journals. He serves as an Associate Editor for the *IEEE Transactions on Wireless Communications* and *IEEE Transactions on Communications*, and a Senior Area Editor for the *IEEE Signal Processing Letters*. He was a recipient of the Humboldt Fellowship for Experienced Researchers and the best paper awards of *IEEE GLOBECOM* 2019, *ADHOCNETS* 2019, and *WCSP* 2022 and 2014.