HAL
open science

Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access

Gang Yang, Xinyue Xu, Ying-Chang Liang, Marco Di Renzo

To cite this version:

Gang Yang, Xinyue Xu, Ying-Chang Liang, Marco Di Renzo. Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access. IEEE Transactions on Wireless Communications, 2021, 20 (5), pp.3137 - 3151. 10.1109/twc.2020.3047632 . hal-03377191

HAL Id: hal-03377191

https://hal.science/hal-03377191v1

Submitted on 14 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access

Gang Yang, *Member, IEEE*, Xinyue Xu, *Student Member, IEEE*, Ying-Chang Liang, *Fellow, IEEE*, and Marco Di Renzo, *Fellow, IEEE*

***Abstract*— Reconfigurable intelligent surface (RIS) is a revolutionary technology to achieve spectrum-, energy-, and cost-efficient wireless networks. This paper considers an RIS-assisted downlink non-orthogonal-multiple-access (NOMA) system. To optimize the rate performance and ensure user fairness, we maximize the minimum decoding signal-to-interference-plus-noise-ratio (equivalently the rate) of all users, by jointly optimizing the (active) transmit beamforming at the base station (BS) and the phase shifts (i.e., passive beamforming) at the RIS. A combined-channel-strength based user-ordering scheme for NOMA decoding is first proposed to decouple the user-ordering design and the joint beamforming design. Efficient algorithms are further proposed to solve the non-convex problem, by leveraging the block coordinated descent and semidefinite relaxation (SDR) techniques. For the single-antenna BS setup, the optimal power allocation at the BS and the asymptotically optimal phase shifts at the RIS are obtained in closed forms. For the multiple-antenna BS setup, it is shown that the rank of the SDR solution of the transmit beamforming design is upper bounded by two. Also, the proposed algorithms are analyzed in terms of convergence and complexity. Simulation results show that the RIS-assisted NOMA system can enhance the rate performance significantly, compared to traditional NOMA without RIS and traditional orthogonal multiple access with/without RIS.**

***Index Terms*— Non-orthogonal multiple access, reconfigurable intelligent surface, rate optimization, user fairness, iterative algorithm.**

Manuscript received April 26, 2020; revised October 10, 2020; accepted December 18, 2020. Date of publication January 8, 2021; date of current version May 10, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61631005, Grant 62071093, and Grant U1801261; in part by the National Key Research and Development Program of China under Grant 2018YFB1801105; in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2019Z022 and in part by the Programme of Introducing Talents of Discipline to Universities under Grant B20064. The work of Marco Di Renzo was supported in part by the European Commission through the H2020 ARIADNE Project under Grant 675806. This article was presented in part at the IEEE Wireless Communications and Networking Conference 2020. The associate editor coordinating the review of this article and approving it for publication was K. Kansanen. *(Corresponding author: Ying-Chang Liang.)*

Gang Yang and Xinyue Xu are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China, and also with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China (e-mail: yanggang@uestc.edu.cn; 201821220447@std.uestc.edu.cn).

Ying-Chang Liang is with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China (e-mail: liangyc@ieee.org).

Marco Di Renzo is with Laboratoire des Signaux et Systèmes, CNRS, CentraleSupélec, Université Paris-Saclay, 91192 Gif-sur-Yvette, France (e-mail: marco.di-renzo@universite-paris-saclay.fr).

I. Introduction

A. Motivation

NON-ORTHOGONAL multiple access (NOMA), which can serve multiple users in the same resource (e.g., time, frequency, code) block, has been recognized as a promising technology for future wireless communication systems, due to its appealing advantages such as enhanced spectrum efficiency (SE), massive wireless connectivity and low latency [2]. Specifically, power-domain NOMA exploits the difference in the channel gain among multiple users for multiplexing and relies on successive-interference-cancellation (SIC) for decoding multiple data flows. However, the application of NOMA has the following three main restrictions. First of all, it should be noticed that NOMA achieves significant SE gain than traditional orthogonal multiple access (OMA) only when the channel strengths of multiple users are quite different [3]. Secondly, NOMA does not always outperform OMA. For instance, when the channels of multiple users are mutually orthogonal in a downlink multiple-input-single-output (MISO) system, no gain can be obtained through NOMA and the traditional OMA (i.e., spatial division multiple access) is more preferable [4]. Finally, NOMA can achieve user fairness by proper power allocation, but the performance improvement of weak users may result in worse performance to users with good channel conditions [5].

On the other hand, reconfigurable intelligent surface (RIS) [6], also termed as intelligent reflecting surface (IRS) [7] or large intelligent surface (LIS) [8], has emerged as a new and transformative technology to design spectrum- and energy-efficient as well as cost-efficient wireless networks, thus has drawn fast-growing research interests [9], [10]. RIS consists of a large number of low-cost reflecting elements and each element can passively reflect a phase-shifted version of the incident electromagnetic field [11]. The reflected signal propagation can be smartly configured by digitally adjusting the phase shifts of all reflecting elements to achieve certain communication objectives such as interference mitigation [8] and received-signal power boosting [9].

RIS can be explored to enhance the performance of NOMA by intelligently reconfiguring the reflected signal propagations. RIS can provide additional channel paths to construct stronger combined channels with obvious strength difference and also re-align the users' combined channels in an artificial manner, to obtain NOMA gain in harsh scenarios (e.g., the channel strengths of different users are similar or the

users' channels are (close to) mutually orthogonal). Also, by adjusting the RIS's phase shifts, RIS-assisted NOMA systems can achieve better performance under user fairness guarantee. These advantages motivate us to study an RIS-assisted downlink NOMA system.

This work is an extension of the conference-version paper [1], which generalizes the system model from an RIS-assisted single-input-single-output (SISO) NOMA setup to an RIS-assisted MISO NOMA setup. In particular, we aim to maximize the minimum rate of NOMA users by jointly optimizing the transmit beamforming at the base station (BS) and the phase shifts at the RIS. This problem is challenging, since the phase shifts at the RIS change the strengths of all users' combined channels and thus affect the user ordering for the SIC-based NOMA decoding.

There exist a few recent works on RIS-assisted NOMA. Multiple RISs are employed to maximize the number of served users by effectively reconstructing the users' channels [12]. By jointly optimizing the BS's transmit beamforming and the RIS's phase shifts, the BS's total power is minimized [13], and the sum rate is maximized [14]. To our best knowledge, there is no existing work focusing on max-min rate optimization for RIS-assisted NOMA systems with user fairness guarantee.

B. Related Works

1) Literature on NOMA: NOMA was comprehensively studied in single-carrier communication systems. For single-antenna NOMA systems, the optimal transmission power allocation with user fairness guarantee was investigated in [5]. For downlink MISO NOMA systems, the beamforming design was studied in several prior works. For example, the weighted-sum rate was maximized under the decoding-order constraints and each user's quality-of-service constraint in [15]. The energy efficiency (EE) was maximized under the BS's transmission power constraint and the users' rate constraints in [16]. A minorization-maximization based algorithm was proposed to maximize the sum rate in [17]. In the context of multiple-input-multiple-output (MIMO) NOMA systems, a NOMA scheme based on intra-beam superposition coding and SIC was proposed in [3], and the power allocation was optimized to maximize the ergodic capacity in [18].

In addition, NOMA was studied in multiple-carrier communication systems. For a downlink multiple-carrier NOMA system, the subcarrier and power allocation were jointly optimized to maximize the EE and the weighted-sum throughput in [19] and [20], respectively. Also, NOMA was integrated with other communication technologies such as millimeter wave communications [21] and backscatter communications [22], [23]. In particular, a backscatter-NOMA system which integrates NOMA and a novel symbiotic radio paradigm (also termed as cooperative ambient backscatter communication (CABC) [24], [25]) was proposed in [23].

2) Literature on RIS-Assisted Communciations: RIS was recently shown to enhance the performances of various wireless communication systems. For an RIS-assisted multiuser MISO communication system, the power allocation at the BS and the phase shifts at the RIS were jointly optimized to maximize the system's sum rate in [26] and EE in [27], and the transmit beamforming at the BS and the phase shifts at the RIS were jointly optimized to maximize the weighted-sum rate in [28]. The minimum secrecy-rate was maximized for an RIS-assisted downlink MISO system with multiple eavesdroppers in [29]. Considering the practical case of a limited number of phase shifts, a hybrid beamforming scheme was proposed to maximize the sum rate of an RIS-assisted downlink multiuser MISO system in [30]. A deep-reinforcement-learning based algorithm was proposed to jointly design the BS's transmit beamforming and the RIS's phase shifts for an RIS-assisted multiuser MISO system in [31]. The RIS was shown to improve the channel rank of MIMO communications in [32]. The ergodic SE of an RIS-assisted large-scale antenna system was analyzed in [33].

RIS-assisted communication resembles but differs from backscatter communications. RIS operates in full-duplex mode without introducing self-interference, and provides additional paths for wireless signals without conveying its own information [34], while a backscatter device transmits its own information by modulating the incident signal from either a dedicated source or an ambient source [35]–[38]. Recently, the RIS was utilized to enhance backscatter communications and to assist the direct communication in an SR system [39], [40].

C. Contributions

This paper considers an RIS-assisted downlink NOMA system in which a BS transmits downlink superposed signals to multiple users with the assistance of an RIS. The contributions are as follows:

- In order to optimize the rate performance and ensure user fairness, we formulate a problem to maximize the minimum target decoding signal-to-interference-plus-noise-ratio (SINR) (i.e., equivalently the rate) of all users, by jointly optimizing the (active) transmit beamforming at the BS and the phase shifts (i.e., passive beamforming) at the RIS, subject to the BS's transmission power constraint, the RIS's phase-shift constraints, and the users' SINR constraints for SIC-based NOMA decoding. However, the formulated problem is challenging to be solved for two reasons: first, the user ordering and the optimization variables are closely coupled; second, the problem under a given user order is still non-convex and difficult to be solved optimally.
- To decouple the problem, an efficient combined-channel-strength (CCS) based user-ordering scheme is proposed for the RIS-assisted NOMA system, which orders the users according to their maximally achievable CCSs obtained by optimizing the RIS's phase shifts. Numerical results show that the designed user-ordering scheme achieves almost the same rate performance compared to the high-complexity scheme of exhaustive search over all possible user orders.
- Under the user order designed by the CCS-based scheme, we first solve the problem for the special setup of a single-antenna BS, in which the BS optimizes the power

allocation among data flows. We propose an efficient iterative algorithm based on the block coordinated descent (BCD) technique and the semidefinite relaxation (SDR) technique. In each iteration, the optimal power allocation for a set of given phase shifts at the RIS is derived in closed form, and for two-user NOMA, the asymptotically optimal phase shifts are obtained in closed form, in the regime of high BS transmission power. The complexity of the proposed algorithm is also studied.

- An extended iterative algorithm is further proposed for the general setup of a multiple-antenna BS. In each iteration, the SDR technique is applied to obtain a solution of the RIS's phase shifts for a given beamforming matrix at the BS, and a solution of the BS's beamforming matrix for a set of given phase shifts at the RIS. For the transmit beamforming optimization, it is shown that the rank of the SDR solution is no more than two, independent of the number of NOMA users. The convergence is proved and the complexity of the proposed algorithm is studied too.
- Numerical results show that our proposed design outperforms three benchmarks including the traditional NOMA, RIS-assisted OMA and traditional OMA. Moreover, a NOMA system assisted by a practical RIS with low-resolution phase shifters can approximate the best-achievable rate performance of a NOMA system assisted by an ideal RIS with infinite-resolution phase shifters. Also, there is a tradeoff between the number of NOMA users and the achievable sum rate, and good rate fairness can always be guaranteed as the number of NOMA users increases.

The rest of this paper is organized as follows. Section II presents the model for an RIS-assisted downlink NOMA system, and formulates the minimum-rate maximization problem. Section III describes the CCS-based user-ordering scheme. Section IV proposes solutions for the special setup of a single-antenna BS. Section V extends the solutions to the general setup of a multiple-antenna BS. Section VI presents the numerical results. Section VII concludes this paper.

The main notations and symbols are listed as follows: The lowersetup, boldface lowersetup, and boldface uppersetup letters, e.g., g, $\mathbf{g}$, and $\mathbf{G}$, denote a scalar, vector, and matrix, respectively. $|g|$ denotes the operation of taking the absolute value of a scalar g; $\|\mathbf{g}\|$ denotes the operation of taking the l_2-norm value of a vector $\mathbf{g}$; $[\mathbf{g}]^T$ and $[\mathbf{g}]^H$ denote the transpose and the hermitian of a vector $\mathbf{g}$, respectively. $\mathcal{CN}(\mu, \sigma^2)$ denotes the circularly symmetric complex Gaussian (CSCG) distribution with mean μ and variance σ^2. $\mathcal{C}$ denotes the set of complex numbers.

II. System Model and Problem Formulation

A. System Model

As illustrated in Fig. 1, we consider an RIS-assisted downlink NOMA communication system, in which a BS equipped with N $(N \geq 1)$ antennas transmits superposed signals to K $(K \geq 2)$ single-antenna users in the same time and frequency block with the assistance of an RIS. The RIS consists

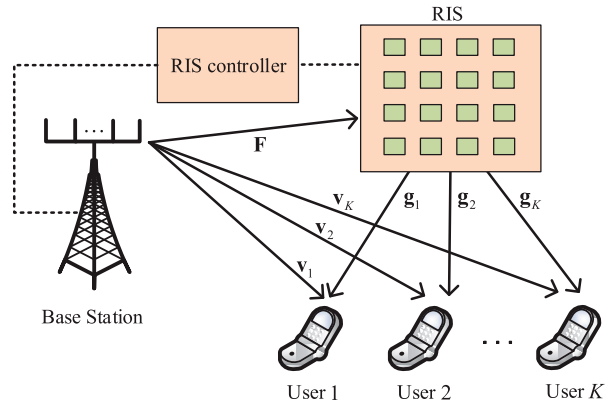


Fig. 1. Illustration of an RIS-assisted NOMA system.

of M $(M \geq 1)$ passive reflecting elements, and each element can reflect a phase-shifted version of its incident signal. A smart controller connected to the RIS can intelligently adjust the RIS's phase shifts to assist the NOMA transmission.

The channel between the BS and user j, $j = 1, \ldots, K$, is denoted as $\mathbf{v}_j \in \mathcal{C}^{N\times 1}$. Since the line-of-sight (LoS) path may be blocked, all BS-to-user channels $\mathbf{v}_j$'s are assumed to be mutually independent and Rayleigh fading distributed. The elements of $\mathbf{v}_j$ are independent and each element follows the distribution $\mathcal{CN}(0, \rho_j)$, where ρ_j denotes the large-scale pathloss from the BS to user j. Notice that the RIS is typically pre-deployed such that it is in LoS with the BS. Hence, we use Rician fading to model the channel matrix $\mathbf{F} \in \mathcal{C}^{M\times N}$ between the BS and the reflecting elements of the RIS, i.e.,

$$\mathbf{F} = \sqrt{\frac{K_1\kappa}{K_1+1}}\bar{\mathbf{F}} + \sqrt{\frac{\kappa}{K_1+1}}\tilde{\mathbf{F}}, \tag{1}$$

where κ denotes the large-scale pathloss from the BS to the RIS, K_1 is the Rician factor of the BS-to-RIS fading channels, $\bar{\mathbf{F}} \in \mathcal{C}^{M\times N}$ and $\tilde{\mathbf{F}} \in \mathcal{C}^{M\times N}$ are the LoS component and non-LoS (NLoS) component, respectively. The elements of $\tilde{\mathbf{F}}$ are independent and each element follows the distribution $\mathcal{CN}(0, 1)$. Since the RIS is typically deployed close to the users to enhance their performances, the channel vector between the RIS and each user j is modeled as

$$\mathbf{g}_j = \sqrt{\frac{K_2\beta_j}{K_2+1}}\bar{\mathbf{g}}_j + \sqrt{\frac{\beta_j}{K_2+1}}\tilde{\mathbf{g}}_j, \tag{2}$$

where β_j denotes the large-scale pathloss from the RIS to the user j, K_2 is the Rician factor of each RIS-to-user fading channel, $\bar{\mathbf{g}}_j \in \mathcal{C}^{M\times 1}$ and $\tilde{\mathbf{g}}_j \in \mathcal{C}^{M\times 1}$ are the LoS component and the NLoS component, respectively, which are generated similar to $\bar{\mathbf{F}}$ and $\tilde{\mathbf{F}}$.

The design of RIS-assisted NOMA systems requires the channel state information (CSI) of the BS-to-RIS link and RIS-to-user links, besides CSI of the BS-to-user links. The reflecting link is given by the cascade of the BS-to-RIS channel and the RIS-to-user channel. This feature makes the estimation of the individual channels challenging, since the RIS typically does not support signal processing capabilities for estimating the BS-to-RIS channel. Nevertheless, some

efficient algorithms have been proposed to estimate the RIS-related channels.[1] Since this paper focuses on an optimal system design, we assume that the required CSI is perfectly known.

The BS transmits a superposition of K data flows, each of which is assigned with one dedicated beamforming vector. That is, the transmitted complex baseband signal can be written as

$$\mathbf{x} = \sum_{j=1}^{K} \boldsymbol{\omega}_j x_j, \tag{3}$$

where $x_j \sim \mathcal{CN}(0,1)$ is the data flow intended to user j, and $\boldsymbol{\omega}_j$ is the corresponding beamforming vector. The signal received at user j is then given by

$$y_j = (\mathbf{g}_j^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_j^H)\mathbf{x} + n_j, \tag{4}$$

where $\boldsymbol{\Theta} = \mathrm{diag}\{e^{j\theta_1}, \cdots, e^{j\theta_M}\}$ is the RIS's diagonal phase-shift matrix with $\theta_m \in [0, 2\pi)$ denoting the phase shift of the m-th reflecting element, and $n_j \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) at user j.

The users in downlink NOMA systems employ the SIC technique to decode the signals. Preliminary, it is necessary to order the users according to their effective channel gains [2], [3], [5]. However, for RIS-assisted NOMA, since the combined channel $\mathbf{h}_j \triangleq \mathbf{g}_j^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_j^H$ also depends on the phase-shift values $\boldsymbol{\Theta}$, the optimal user order may be any one of all the $K!$ different user orders denoted by the set Π. We will propose an efficient user-ordering scheme in Section III.

Without loss of generality, it is assumed that after user-ordering mapping operation denoted by π, the users are ordered as $0 < \|\mathbf{h}_1\|^2 \leq \|\mathbf{h}_2\|^2 \leq \ldots \leq \|\mathbf{h}_K\|^2$, where the k-th ordered user corresponds to the k-th weakest combined channel. The k-th user is always able to sequentially decode the signal of the t-th user, for $t = 1, \ldots, k-1$, and then extract them from the received signal. The corresponding SINR for the k-th user decoding the signal intended to the t-th user is

$$\gamma_{t\to k} = \frac{|(\mathbf{g}_k^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_t|^2}{\sum\limits_{i=t+1}^{K} |(\mathbf{g}_k^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_i|^2 + \sigma^2}, \ 1 \leq t < k \leq K. \tag{5}$$

After cancelling the interference signals from all weaker users, the k-th user decodes its own signal by treating the signals from the other users as interference. The SINR for the k-th user decoding its own signal is expressed as

$$\gamma_{k\to k} = \frac{|(\mathbf{g}_k^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_k|^2}{\sum\limits_{i=k+1}^{K} |(\mathbf{g}_k^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_i|^2 + \sigma^2}, \ 1 \leq k \leq K. \tag{6}$$

In order to ensure that the k-th user can decode the t-th data flow correctly and cancel its resulting interference, the SINR for the k-th user decoding the t-th data flow (i.e., $\gamma_{t\to k}$) needs to be no smaller than the target SINR of the t-th user [17], [43], denoted by γ_t^{tar}, i.e., $\gamma_{t\to k} \geq \gamma_t^{\mathsf{tar}}, \forall t < k$. Then the target SINR of the t-th user is expressed as follows

$$\gamma_t^{\mathsf{tar}} = \min\{\gamma_{t\to t}, \gamma_{t\to t+1}, \ldots, \gamma_{t\to K}\}, \quad \forall t. \tag{7}$$

Thus, the corresponding rate for the t-th user is given by

$$R_t = \log_2\left(1 + \gamma_t^{\mathsf{tar}}\right), \quad \forall t. \tag{8}$$

From (8), the use of an RIS is beneficial for downlink NOMA systems, since it provides not only additional channel paths (i.e., $\mathbf{g}_k^H \boldsymbol{\Theta}\mathbf{F}$) but additional degree of freedom for system design (i.e., $\boldsymbol{\Theta}$). By jointly designing the beamforming vectors $\{\boldsymbol{\omega}_k\}$ at the BS and the phase shifts $\boldsymbol{\Theta}$ at the RIS, the rate performance can be improved.

B. Problem Formulation

To maximize the system's rate performance while ensuring the fairness among users, as in [5] [44], we maximize the minimum target SINR in (7) (equivalently the achievable rate in (8)) of the users by jointly optimizing the beamforming vectors $\{\boldsymbol{\omega}_k\}$ at the BS and the phase shifts $\boldsymbol{\Theta}$ at the RIS. The optimization problem can be formulated as

$$\text{(P1):} \max_{\{\boldsymbol{\omega}_k\}, \boldsymbol{\Theta}, Q, \pi} \quad Q \tag{9a}$$

$$\text{s.t.} \quad \frac{|(\mathbf{g}_k^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_t|^2}{\sum\limits_{i=t+1}^{K} |(\mathbf{g}_k^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_i|^2 + \sigma^2} \geq Q,$$

$$1 \leq t \leq k \leq K \tag{9b}$$

$$\|\mathbf{g}_K^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_K^H\|^2 \geq \|\mathbf{g}_{K-1}^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_{K-1}^H\|^2 \geq \cdots \geq \|\mathbf{g}_1^H \boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_1^H\|^2 \tag{9c}$$

$$\sum_{k=1}^{K} \|\boldsymbol{\omega}_k\|^2 \leq P \tag{9d}$$

$$0 \leq \theta_m < 2\pi, \ \forall m \tag{9e}$$

$$\pi \in \Pi. \tag{9f}$$

The constraint (9b) ensures that the SINR for the k-th ($k \geq t$) user decoding the t-th data flow exceeds Q, where Q is a slack variable denoting the target SINR of all users to be maximized; (9c) is the CCS constraint under the determined user order; (9d) is the constraint of the BS's transmission power; (9e) is the phase-shift constraints of the RIS's reflecting elements; (9f) is the user-order constraint.

The optimization problem (P1) is challenging to be solved directly. On the one hand, there are multiple coupled variables (i.e., $\{\boldsymbol{\omega}_k\}, \boldsymbol{\Theta}, Q$) and the non-convex constraints (9b). On the other hand, the user order needs to be determined. In order to solve (P1) effectively, we first propose an efficient user-ordering scheme in Section III, and then study the special setup of a single-antenna BS in Section IV and the general setup of a multiple-antenna BS in Section V, respectively.

III. CCS-Based User Ordering Design

User ordering is essential for SIC-based decoding and performance optimization in RIS-assisted NOMA systems. By using the exhaustive search scheme adopted in [14], one

[1] In [41], for instance, the BS-to-RIS channel and the RIS-to-user channel are estimated by using sparse matrix factorization and matrix completion, respectively. In [42], by deploying a few active antenna elements on the RIS, a compressive sensing and deep learning approach was proposed to estimate the RIS-related channels.

needs to solve $K!$ complex sum-rate maximization problems each of which is for a particular user order, thus the complexity is high especially when the number of users K is relatively large. The BS-to-user channel strength based user-ordering scheme adopted in [13] is simple, but it provides poor performance, since the impact of the RIS is ignored.

This section designs an efficient and low-complexity CCS-based user-ordering scheme. That is, the user order is determined according to all users' maximally achievable CCSs each of which is obtained by optimizing the RIS's phase shifts. Specifically, the maximally achievable strength of the combined channel for the j-th user can be obtained by solving the following problem

$$\text{(P-Order)}: \max_{\boldsymbol{\Theta}} \quad \|\mathbf{g}_j^H \boldsymbol{\Theta} \mathbf{F} + \mathbf{v}_j^H\|^2 \tag{10a}$$

$$\text{s.t.} \quad 0 \leq \theta_m < 2\pi, \ \forall m. \tag{10b}$$

We denote $e_m = e^{j\theta_m}$, and $\mathbf{e} = [e_1, \ldots, e_M]^H$. Then the constraint (10b) is equivalent to $|e_m| = 1, \forall m$. Define $\mathbf{J}_j = \mathrm{diag}(\mathbf{g}_j^H)\mathbf{F}$. Then the term $\|\mathbf{g}_j^H \boldsymbol{\Theta} \mathbf{F} + \mathbf{v}_j^H\|^2$ can be rewritten as $\|\mathbf{e}^H \mathbf{J}_j + \mathbf{v}_j^H\|^2$. By introducing $\bar{\mathbf{e}} = [\mathbf{e}; 1]$ and

$$\mathbf{S}_j = \begin{bmatrix} \mathbf{J}_j \mathbf{J}_j^H & \mathbf{J}_j \mathbf{v}_j \\ \mathbf{v}_j^H \mathbf{J}_j^H & 0 \end{bmatrix}, \tag{11}$$

the term $\|\mathbf{e}^H \mathbf{J}_j + \mathbf{v}_j^H\|^2$ can be further transformed into $\bar{\mathbf{e}}^H \mathbf{S}_j \bar{\mathbf{e}} + \|\mathbf{v}_j^H\|^2$. Note that $\bar{\mathbf{e}}^H \mathbf{S}_j \bar{\mathbf{e}} = \mathrm{Tr}(\mathbf{S}_j \bar{\mathbf{e}} \bar{\mathbf{e}}^H)$. Define the matrix $\mathbf{E} = \bar{\mathbf{e}} \bar{\mathbf{e}}^H$, which needs to satisfy $\mathbf{E} \succeq 0$ and $\mathrm{rank}(\mathbf{E}) = 1$. Since the rank-one constraint is non-convex, we exploit the SDR technique [45] to relax (P-Order) as follows

$$\text{(P-Order-SDR)}: \quad \max_{\mathbf{E}} \quad \mathrm{Tr}(\mathbf{S}_j \mathbf{E}) + \|\mathbf{v}_j\|^2 \tag{12a}$$

$$\text{s.t.} \quad \mathbf{E} \succeq 0 \tag{12b}$$

$$[\mathbf{E}]_{m,m} = 1. \tag{12c}$$

The optimal $\mathbf{E}^\star$ obtained by solving (P-Order-SDR) generally does not satisfy the rank-one constraint, and a Gaussian randomization scheme can be applied to obtain a rank-one solution, which is described as follows. First, we obtain the eigenvalue decomposition of $\mathbf{E}^\star$ as $\mathbf{E}^\star = \mathbf{U} \boldsymbol{\Sigma} \mathbf{U}^H$. Define $\boldsymbol{\Sigma}^{\frac{1}{2}} \triangleq \mathrm{diag}\left\{\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_{M+1}}\right\}$. Random vectors are generated as follows

$$\tilde{\mathbf{e}} = \mathbf{U} \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{r}, \tag{13}$$

where $\mathbf{r} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{M+1})$ is a random vector. Then, we generate $\hat{\mathbf{e}}$ as $\hat{\mathbf{e}} = e^{j\angle\left([\tilde{\mathbf{e}}/\tilde{e}_{M+1}]_{(1:M)}\right)}$, where $[\mathbf{x}]_{(1:M)}$ denotes the vector containing the first M elements in $\mathbf{x}$. The objective value of (P-Order) is approximated as the maximal one achieved by the best $\hat{\mathbf{e}}$ among all $\mathbf{r}$'s. From Table II of [45], the SDR technique enhanced with such randomization scheme can guarantee at least an approximation accuracy $\frac{\pi}{4}$ of the optimal objective value of (P-Order).

After solving the K optimization problems (P-Order) independently, we obtain all users' maximally achievable CCSs according to which the users are ordered. That is, only the optimal objective values of K (P-Order) problems are used to decide the user order. Under the designed user order, the minimum-rate maximization problem needs to be solved just once. Numerical results in Section VI will show that compared to the exhaustive search scheme, the designed low-complexity user-ordering scheme achieves almost the same rate performance for in a two-user NOMA system, and suffers from slight rate-performance degradation when more users are multiplexed.

IV. Optimal Solution for Single-Antenna Base Station Setup

In this section, we solve the minimum rate maximization problem for the single-antenna BS setup, i.e., $N = 1$. The problem solved in this section is a special case of that in Section V, but the solution for this special case can provide useful insights for system design. The BS-to-RIS channel matrix $\mathbf{F} \in \mathcal{C}^{M \times N}$ reduces to the channel vector $\mathbf{f} \in \mathcal{C}^{M \times 1}$, the BS-to-user channel vectors $\mathbf{v}_k$'s reduce to the channel coefficients v_k's, and the beamforming vectors $\{\boldsymbol{\omega}_k\}$ reduce to the power allocation vector $\boldsymbol{\alpha} \in \mathcal{R}^{K \times 1}$. Thus, (P1) for the single-antenna BS setup is rewritten as follows

$$\text{(P2)}: \max_{\alpha, \boldsymbol{\Theta}, Q, \pi} \quad Q \tag{14a}$$

$$\text{s.t.} \quad \gamma_{t \to t} = \frac{\alpha_k P |\mathbf{g}_k^H \boldsymbol{\Theta} \mathbf{f} + v_k^H|^2}{\sum\limits_{i=k+1}^{K} \alpha_i P |\mathbf{g}_k^H \boldsymbol{\Theta} \mathbf{f} + v_k^H|^2 + \sigma^2} \geq Q, \ \forall k \tag{14b}$$

$$|\mathbf{g}_K^H \boldsymbol{\Theta} \mathbf{f} + v_K^H|^2 \geq |\mathbf{g}_{K-1}^H \boldsymbol{\Theta} \mathbf{f} + v_{K-1}^H|^2 \geq \cdots \geq |\mathbf{g}_1^H \boldsymbol{\Theta} \mathbf{f} + v_1^H|^2 \tag{14c}$$

$$\sum_{k=1}^{K} \alpha_k \leq 1 \tag{14d}$$

$$\alpha_k \geq 0, \ \forall k \tag{14e}$$

$$0 \leq \theta_m < 2\pi, \ \forall m. \tag{14f}$$

$$\pi \in \Pi. \tag{14g}$$

It can be easily verified that once the user-ordering constraint (14c) is satisfied, the inequality $\gamma_{t \to k} \geq \gamma_{t \to t}$ always holds. Thus, the constraint $\gamma_{t \to k} \geq Q$ contained in (9b) is omitted in the single-antenna BS setup. Equations (14d) and (14e) are the normalization constraint and non-negative constraints of the BS's power allocation coefficients.

It is difficult to solve (P2) due to the non-convex constraint (14b), the coupled variables and the user-ordering constraint (14g). In order to solve (P2), we use the following approach. First, the designed CCS-based user-ordering scheme in Section III is used to remove the constraint (14g). Then, we exploit the BCD (i.e., blocked coordinate descent) [46] and SDR (i.e., semidefinite-relaxation) techniques to solve it approximately. In each iteration $n = 1, 2, \ldots$, we optimize different blocks of variables alternatively. Therefore, (P2) is decoupled into two subproblems in each iteration, as described in Subsection IV-A and Subsection IV-B. In the sequel, $\boldsymbol{\alpha}^n$ and $\boldsymbol{\Theta}^n$ with superscript n indicate their values after the n-th algorithmic iteration.

Remark 1: Notice that the BCD technique is widely used to solve non-convex problems with multiple blocks of variables, and it can guarantee the convergence to a stationary point of

typical non-convex problems under some mild conditions on the properties of the problems. See details in [46].

A. Phase Shift Optimization

1) General Solution for $K \geq 2$: In each iteration n, given the power allocation coefficients $\boldsymbol{\alpha}^n$, the phase shifts $\boldsymbol{\Theta}$ can be optimized by solving the following problem

$$\text{(P2.1):} \quad \max_{\boldsymbol{\Theta},\, Q} \quad Q \tag{15a}$$

$$\text{s.t.} \quad (14b), (14c), (14f). \tag{15b}$$

Define $\mathbf{l}_k = \text{diag}(\mathbf{g}_k^H)\mathbf{f}_k$, then the term $|\mathbf{g}_k^H\boldsymbol{\Theta}\mathbf{f} + v_k|^2$ in (14b) and (14c) can be rewritten as $|\mathbf{e}^H\mathbf{l}_k + v_k|^2$, with $\mathbf{e} = [e^{j\theta_1}, \ldots, e^{j\theta_m}, \ldots, e^{j\theta_M}]^H$. Recall that $\mathbf{E} = \bar{\mathbf{e}}\bar{\mathbf{e}}^H$, with $\bar{\mathbf{e}} = [\mathbf{e}; 1]$. By introducing

$$\mathbf{R}_k = \begin{bmatrix} \mathbf{l}_k\mathbf{l}_k^H & \mathbf{l}_k v_k \\ v_k^H\mathbf{l}_k^H & 0 \end{bmatrix}, \tag{16}$$

we have $|\mathbf{g}_k^H\boldsymbol{\Theta}\mathbf{f} + v_k|^2 = \text{Tr}(\mathbf{R}_k\mathbf{E}) + |v_k|^2$. Problem (P2.1) can be transformed into the problem

$$\text{(P2.2):} \quad \max_{\mathbf{E}, Q} \quad Q \tag{17a}$$

$$\text{s.t. } P\left(\alpha_k^n - \sum_{i=k+1}^{K} \alpha_i^n Q\right) \times \left(\text{Tr}(\mathbf{R}_k\mathbf{E}) + |v_k|^2\right) \geq Q\sigma^2, \forall k \tag{17b}$$

$$\text{Tr}(\mathbf{R}_K\mathbf{E}) + |v_K|^2 \geq \text{Tr}(\mathbf{R}_{K-1}\mathbf{E}) + |v_{K-1}|^2 \geq \cdots \geq \text{Tr}(\mathbf{R}_1\mathbf{E}) + |v_1|^2 \tag{17c}$$

$$\mathbf{E} \succeq 0 \tag{17d}$$

$$[\mathbf{E}]_{m,m} = 1. \tag{17e}$$

However, (P2.2) is still non-convex due to the non-convex constraint (17b). It can be solved by using the following two-step procedure. First, the bisection-search method is used to decouple Q and $\boldsymbol{\Theta}$, and the SDR technique is used to obtain the optimal $\mathbf{E}$. Specifically, with certain $Q_{\max}$ and $Q_{\min}$, we replace Q in (P2.2) by $\frac{Q_{\max}+Q_{\min}}{2}$, and solve the resulting feasibility problem reduced from (P2.2). The update of $Q_{\max}$ and $Q_{\min}$ depends on whether a feasible $\mathbf{E}$ can be found. It can be checked that for sufficiently large $Q_{\max}$ and small $Q_{\min}$, the above bisection search over Q can give a globally optimal phase-shift-related matrix $\mathbf{E}^{n+1}$ in the n-th alterative iteration. Second, the Gaussian randomization technique described in Section III is applied to obtain a suboptimal rank-one solution $\mathbf{e}$.

2) Closed-Form Solution for $K = 2$: The above general procedure for optimizing $\boldsymbol{\Theta}$ involves the bisection search, SDR and randomization-based approximation. In this section, a closed-form solution is proposed for the two-user NOMA setup, i.e., $K = 2$.

For notational convenience, we rewrite the channel elements as $[\mathbf{g}_k^H]_i = |[\mathbf{g}_k^H]_i|e^{j\varphi_{k,i}}$, $[\mathbf{f}]_i = |[\mathbf{f}]_i|e^{j\psi_i}$ and $v_k = |v_k|e^{j\xi_k}$, for $i = 1, \ldots, M$, where $[\mathbf{g}_k^H]_i$ and $[\mathbf{f}]_i$ denote the channel from the k-th user to the i-th reflecting element of the RIS and the channel from the BS to the i-th reflecting element of the RIS, respectively. Define the combined channel of the k-th user as $h_k \triangleq \mathbf{g}_k^H\boldsymbol{\Theta}\mathbf{f} + v_k^H$. We have the following proposition on the optimal phase shifts $\boldsymbol{\Theta}$.

Proposition 1: For the two-user NOMA system, as the BS's transmission power P is sufficiently high such that $\alpha_2\, P|h_1|^2 \gg \sigma^2$, the RIS's asymptotically optimal phase shifts for (P2.1) are

$$\theta_i = \xi_2 - \varphi_{2,i} - \psi_i, \quad \textit{for } i = 1, \ldots, M. \tag{18}$$

Proof: See proof in Appendix A. ■

Remark 1: As shown in Appendix A, the above closed-form solution in Proposition 1 is asymptotically optimal, when the BS's transmission power P is sufficiently high such that the weak user's decoding interference power $\alpha_2\, P|h_1|^2$ dominates its noise power σ^2. This condition on the BS transmit power can be satisfied for typical small-cell topologies.[2] *However, this closed-form solution is suboptimal, when P is low or moderate such that σ^2 is not negligible compared to $\alpha_2\, P|h_1|^2$. As given in Appendix A, the SINRs $\gamma_{1\to1}$ in* (33) *and $\gamma_{2\to2}$ in* (34) *are monotonically increasing functions of $|h_1|^2$ and $|h_2|^2$. Hence, to maximize the minimum value between both SINRs, the optimal phase shifts $\boldsymbol{\Theta}$ should be designed to enhance both $|h_1|^2$ and $|h_2|^2$ in general. Numerical results show that, even for small or moderate P, the above closed-form solution suffers from slight rate performance degradation compared to the solution achieved by the general algorithm.*

B. Power Allocation Optimization

In each iteration n, for given phase shifts $\boldsymbol{\Theta}^n$, the power allocation $\boldsymbol{\alpha}$ can be optimized by solving

$$\text{(P2.3):} \quad \max_{\boldsymbol{\alpha}, Q} \quad Q \tag{19a}$$

$$\text{s.t.} \quad (14b), (14d), (14e). \tag{19b}$$

We have the following lemma on the optimal solution of (P2.3).

Lemma 1: The optimal power allocation of (P2.3), denoted by $\tilde{\boldsymbol{\alpha}} = [\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_K]$, is the unique solution of the following equations

$$\frac{\tilde{\alpha}_1 P|h_1|^2}{\sum_{i=2}^{K} \tilde{\alpha}_i P|h_k|^2 + \sigma^2} = \cdots = \frac{\tilde{\alpha}_k P|h_k|^2}{\sum_{i=k+1}^{K} \tilde{\alpha}_i P|h_k|^2 + \sigma^2} = \cdots = \frac{\tilde{\alpha}_K P|h_K|^2}{\sigma^2} = Q, \tag{20}$$

$$\sum_{k=1}^{K} \tilde{\alpha}_k = 1. \tag{21}$$

Proof: See proof in Appendix B. ■

From Lemma 1, the optimal solution $\tilde{\boldsymbol{\alpha}}$ can be obtained by solving the K nonlinear equations in (20) and (21). Fortunately, the optimal solution $\tilde{\boldsymbol{\alpha}}$ can be obtained in closed

[2]For example, we consider a topology where the BS-to-user distances are 245 m and 158 m, respectively, the BS-to-RIS distance is 175 m, and the RIS-to-user distances are 87.5 m and 35 m, respectively. By using pathloss model and parameters in Section VI, it is estimated that when the BS transmit power is 2.5 Watt, the received interference power at the weaker user is 20 times higher than its noise power, for $M = 100$, $\alpha_2 = 0.1$, and $\sigma^2 = -114$ dBm.

form. To this end, we define the length-K column vector with all-one elements as $\mathbf{1}$, the diagonal matrix $\mathbf{D} = \mathrm{diag}\{\frac{1}{|h_1|^2}, \frac{1}{|h_2|^2}, \ldots, \frac{1}{|h_K|^2}\}$, the upper-triangle matrix $\mathbf{\Psi}$ with elements $[\mathbf{\Psi}]_{ik} = |h_i|^2$ for $i < k$ and $[\mathbf{\Psi}]_{ik} = 0$ for $i \geq k$, and the following $(K+1)$-dimensional square matrix

$$\mathbf{\Pi} = \begin{bmatrix} \mathbf{D}\mathbf{\Psi} & \frac{\sigma^2}{P}\mathbf{D}\mathbf{1} \\ \mathbf{1}^T\mathbf{D}\mathbf{\Psi} & \frac{\sigma^2}{P}\mathbf{1}^T\mathbf{D}\mathbf{1} \end{bmatrix}. \tag{22}$$

The closed-form solution is given in the following Theorem 1.

Theorem 1: The optimal objective value of (P2.3) is

$$Q = \frac{1}{\lambda_{\max}(\mathbf{\Pi})}, \tag{23}$$

where $\lambda_{\max}(\mathbf{\Pi})$ denotes the largest eigenvalue of $\mathbf{\Pi}$, and the optimal power allocation vector $\tilde{\boldsymbol{\alpha}}$ is obtained as the first K components of $\mathbf{\Pi}$'s dominant eigenvector scaled by its last component.

Proof: See proof in Appendix C. ■

Remark 2: Recall $h_k \triangleq \mathbf{g}_K^H\mathbf{\Theta}\mathbf{f} + v_K^H$, for given $\mathbf{\Theta}$, the optimal solution of (P2.3) satisfies (20)*, thus the SINR of each user is balanced at the same value for the optimal $\boldsymbol{\alpha}$ and $\mathbf{\Theta}$. That is, for the single-antenna BS setup, the best rate fairness among users can be guaranteed.*

C. Overall Algorithm

The iterative algorithm for solving (P2) is summarized in Algorithm 1, wherein the $\boldsymbol{\alpha}$ and $\mathbf{\Theta}$ are alternatively optimized in each iteration. The bisection search and the SDR techniques are utilized to optimize $\mathbf{\Theta}$ from step 3 to step 23, and Theorem 1 is used to compute the optimal $\boldsymbol{\alpha}$ in step 24. The algorithm ends when the increase of the objective value is smaller than a small $\epsilon > 0$.

It can be shown that Algorithm 1 is guaranteed to converge, whose proof is similar to that of Algorithm 2 in Section V and thus omitted herein. But no global optimality can be assured for Algorithm 1. This is explained as follows. The BCD technique can guarantee the convergence to a stationary point, when each subproblem is solved optimally in each iteration [46]. However, the adopted method of bisection search and SDR in Algorithm 1 can only obtain the suboptimal solution of the subproblem (P2.1).

In Algorithm 1, the subproblems (P2.1) and (P2.3) are alteratively solved in each outer-layer BCD iteration, and the subproblem (P2.1) is solved by a bisection search inner-layer iteration in each of which an SDR problem (P2.2) is solved. Specifically, $\log_2(\frac{Q_{\max}-Q_{\min}}{\epsilon_b})$ iterations need to be executed to achieve an accuracy ϵ_{b} of bisection search over Q. Hence, from the complexity analysis of typical interior-point method like primal-dual path following method [45], the complexity of Algorithm 1 is obtained as $\mathcal{O}\big(I_{\mathsf{ite}}\log_2(\frac{Q_{\max}-Q_{\min}}{\epsilon_b})\max\{2K-1, M+1\}^4(M+1)^{\frac{1}{2}}\log\frac{1}{\epsilon_{\mathsf{s}}}\big)$, where I_{ite} denotes the number of outer-layer BCD iterations, and ϵ_{s} denotes the predefined accuracy of the SDR solution.

For the two-user NOMA system, the steps from 3 to 23 in Algorithm 1 can be replaced by using Proposition 1 to compute the phase-shift solution for given $\boldsymbol{\alpha}^n$. The resulting algorithm has a complexity of $\mathcal{O}(I_{\mathsf{ite}})$, thus termed as the low-complexity Algorithm 1 for two-user NOMA.

Algorithm 1 Iterative Algorithm for Solving (P2)

1: Initialize $\boldsymbol{\alpha}^0$, $\mathbf{\Theta}^0$, D (a large positive integer) and ϵ_{b} (a smaller positive value). Let $n = 0$.
2: **repeat**
3: Given $Q_{\max}$, $Q_{\min}$,
4: **while** $Q_{\max} - Q_{\min} \geq \epsilon_{\mathsf{b}}$, **do**
5: Solve the feasibility problem reduced from (P2.2) with given $Q = \frac{Q_{\max}+Q_{\min}}{2}$.
6: **if** the feasibility problem reduced from (P2.2) is solvable, **then**
7: $Q_{\min} = Q$, update $\mathbf{E}$.
8: **else**
9: $Q_{\max} = Q$.
10: **end if**
11: **end while**
12: **return** $\mathbf{E}^{n+1} = \mathbf{E}$.
13: Compute the eigenvalue decomposition of $\mathbf{E}^{n+1}$ as $\mathbf{E}^{n+1} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$.
14: Initialize $\mathcal{D} = \emptyset$.
15: **for** $d = 1, \ldots, D$ **do**
16: Generate a random vector $\tilde{\mathbf{e}}_d = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{r}_d$, where $\mathbf{r}_d \sim \mathcal{CN}(\mathbf{0}_{M+1}, \mathbf{I}_{M+1})$.
17: Compute $\hat{\mathbf{e}}_d = e^{j\angle([\frac{\tilde{\mathbf{e}}_d}{\tilde{\mathbf{e}}_{d,M+1}}]_{(1:M)})}$, and then obtain $\mathbf{\Theta}_d = \mathrm{diag}(\hat{\mathbf{e}}_d^H)$.
18: **if** $\mathbf{\Theta}_d$ is feasible for (P2.1), **then**
19: $\mathcal{D} = \mathcal{D}\bigcup d$
20: Obtain the objective value of (P2.1) as Q_d.
21: **end if**
22: **end for**
23: **return** $\mathbf{\Theta}^{n+1} = \arg\max_{d\in\mathcal{D}} Q_d$.
24: For given $\mathbf{\Theta}^{n+1}$, use Theorem 1 to compute the optimal power allocation $\boldsymbol{\alpha}^{n+1}$.
25: Update iteration index $n = n + 1$.
26: **until** The increase of the objective value is smaller than ϵ.
27: Return the optimal solution $\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}^n$, $\mathbf{\Theta}^\star = \mathbf{\Theta}^n$, and $Q^\star$.

V. Optimal Solution for Multiple-Antenna Base Station Setup

In this section, we consider the general setup of a multiple-antenna BS. The algorithm for the special setup of a single-antenna BS is generalized to solve (P1). The original problem is decoupled into two subproblems, which are described in the following subsections.

A. Phase Shift Optimization

For given beamforming vectors $\{\boldsymbol{\omega}_k^n\}$, the phase shifts $\mathbf{\Theta}$ can be optimized by solving

$$\text{(P1.1):} \quad \max_{\mathbf{\Theta}, Q} \quad Q \tag{24a}$$

$$\text{s.t.} \quad \text{(9b), (9c), (9e).} \tag{24b}$$

Recall $\mathbf{\Theta} = \mathrm{diag}\{e^{j\theta_1}, \ldots, e^{j\theta_M}\}$, with $m = 1, \ldots, M$. We denote $e_m = e^{j\theta_m}$, and $\mathbf{e} = [e_1, \ldots, e_M]^H$. By applying $\mathbf{l}_{k,t} = \mathrm{diag}(\mathbf{g}_k^H)\mathbf{F}\boldsymbol{\omega}_t^n$ and $v_{k,t} = \mathbf{v}_k^H\boldsymbol{\omega}_t^n$, the term $|(\mathbf{g}_k^H\mathbf{\Theta}\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_t^n|^2$ can be rewritten as $|\mathbf{e}^H\mathbf{l}_{k,t} + v_{k,t}|^2$. We further

reduce this term to $\bar{\mathbf{e}}^H\mathbf{R}_{k,t}\bar{\mathbf{e}} + |v_{k,t}|^2$, where $\bar{\mathbf{e}} = [\mathbf{e}; 1]^H$ and

$$\mathbf{R}_{k,t} = \begin{bmatrix} \mathbf{l}_{k,t}\mathbf{l}_{k,t}^H & \mathbf{l}_{k,t}v_{k,t}^H \\ \mathbf{l}_{k,t}^H v_{k,t} & 0 \end{bmatrix}. \tag{25}$$

Following similar manipulations as in Section III, the term $\|\mathbf{g}_k^H\mathbf{\Theta}\mathbf{F} + \mathbf{v}_k^H\|^2$ can be rewritten as $\bar{\mathbf{e}}^H\mathbf{S}_k\bar{\mathbf{e}} + \|\mathbf{v}_k^H\|^2$, where $\mathbf{S}_k$ is given in (11) with subscript replaced by k.

Note that $\bar{\mathbf{e}}^H\mathbf{R}_{k,t}\bar{\mathbf{e}} = \mathrm{Tr}(\mathbf{R}_{k,t}\bar{\mathbf{e}}\bar{\mathbf{e}}^H)$, $\bar{\mathbf{e}}^H\mathbf{S}_k\bar{\mathbf{e}} = \mathrm{Tr}(\mathbf{S}_k\bar{\mathbf{e}}\bar{\mathbf{e}}^H)$. We define the matrix $\mathbf{E} = \bar{\mathbf{e}}\bar{\mathbf{e}}^H$, which needs to satisfy $\mathbf{E} \succeq 0$ and $\mathrm{rank}(\mathbf{E}) = 1$. Since the rank-one constraint is non-convex, we exploit the SDR technique to relax (P1.1) as follows

$$\text{(P1.2):} \quad \max_{\mathbf{E}, Q} \quad Q \tag{26a}$$

$$\text{s.t.} \quad \frac{\mathrm{Tr}(\mathbf{R}_{k,t}\mathbf{E}) + |v_{k,t}|^2}{\sum_{i=t+1}^{K}(\mathrm{Tr}(\mathbf{R}_{k,i}\mathbf{E}) + |v_{k,i}|^2) + \sigma^2} \geq Q,$$

$$1 \leq t \leq k \leq K \tag{26b}$$

$$\mathrm{Tr}(\mathbf{S}_K\mathbf{E}) + \|\mathbf{v}_K\|^2 \geq \mathrm{Tr}(\mathbf{S}_{K-1}\mathbf{E}) + \|\mathbf{v}_{K-1}\|^2 \geq \cdots \geq \mathrm{Tr}(\mathbf{S}_1\mathbf{E}) + \|\mathbf{v}_1\|^2 \tag{26c}$$

$$\mathbf{E} \succeq 0 \tag{26d}$$

$$[\mathbf{E}]_{m,m} = 1. \tag{26e}$$

Problem (P1.2) is still non-convex due to the non-convex constraint (26b). To tackle the coupled variables Q and $\mathbf{\Theta}$, we use the bisection search method similar to IV-A. Then the Gaussian randomization technique is applied to obtain an approximate solution.

B. Beamforming Matrix Optimization

For given phase shifts $\mathbf{\Theta}^n$, the beamforming vectors $\{\boldsymbol{\omega}_k\}$ can be optimized by solving

$$\text{(P1.3):} \quad \max_{\{\boldsymbol{\omega}_k\},\, Q} \quad Q \tag{27a}$$

$$\text{s.t.} \quad \text{(9b), (9d).} \tag{27b}$$

As (P1.3) is non-convex due to the non-convex constraint (9b), we adopt the SDR technique to obtain an efficient approximate solution. Denote the combined channel $\mathbf{h}_k^H = \mathbf{g}_k^H\mathbf{\Theta}^n\mathbf{F} + \mathbf{v}_k^H$. By introducing $\mathbf{H}_k = \mathbf{h}_k\mathbf{h}_k^H$ and $\mathbf{\Omega}_k = \boldsymbol{\omega}_k\boldsymbol{\omega}_k^H$, the term $|(\mathbf{g}_k^H\mathbf{\Theta}^n\mathbf{F} + \mathbf{v}_k^H)\boldsymbol{\omega}_k|^2$ can be written as $\mathrm{Tr}\,(\mathbf{H}_k\mathbf{\Omega}_k)$, where $\mathbf{\Omega}_k$ needs to satisfy $\mathbf{\Omega}_k \succeq 0$ and $\mathrm{rank}\,(\mathbf{\Omega}_k) = 1$. Since the constraint $\mathrm{rank}\,(\mathbf{\Omega}_k) = 1$ is non-convex, we relax this constraint and (P1.3) can be transformed as the following problem

$$\text{(P1.4):} \quad \max_{\{\mathbf{\Omega}_k\},\, Q} \quad Q \tag{28a}$$

$$\text{s.t.} \quad \frac{\mathrm{Tr}(\mathbf{H}_k\mathbf{\Omega}_t)}{\sum_{i=t+1}^{K}\mathrm{Tr}(\mathbf{H}_k\mathbf{\Omega}_i) + \sigma^2} \geq Q,$$

$$1 \leq t \leq k \leq K \tag{28b}$$

$$\mathbf{\Omega}_k \succeq 0, \forall k \tag{28c}$$

$$\sum_{k=1}^{K}\mathrm{Tr}(\mathbf{\Omega}_k) \leq P. \tag{28d}$$

Since (P1.4) is still non-convex due to the coupled variables Q and $\mathbf{\Omega}_i$'s in (28b), which is analogous to (P2.2), the method based on the bisection search and SDR technique similar to the approach described in Section IV-A can be applied to solve the problem. Furthermore, we have the following theorem on the SDR solution of (P1.4).

Theorem 2: The rank of the SDR solution $\{\mathbf{\Omega}_k^\star\}$ of (P1.4) with given Q is upper bounded by two.

Proof: See proof in Appendix D. ■

Remark 2: Theorem 2 provides a tight upper bound for the rank of SDR solution to (P1.4), which is important in the SDR methodology. The interior-point algorithms used in existing toolbox like CVX for solving semi-definite program (SDP) yield a solution whose rank is maximal among all optimal solutions [45]. Notice that (P1.4) with a given Q is a special quadratically constrained quadratic program (QCQP), in which the objective function reduces to a constant Q. From [47], an upper bound for the solution to (P1.4) is $\min\{\lfloor\frac{1}{2}\left(\sqrt{4K^2+4K+9}-1\right)\rfloor, N\}$, where $\lfloor\cdot\rfloor$ denotes the operation of taking the nearest integer no larger than the argument. For instance, when the number of NOMA users $K = 2, 3, 4, 5, 6$, the corresponding upper bound for the rank of SDR solution is $\min\{\{2, 3, 4, 5, 6\}, N\}$. Hence, we have obtained a tight upper bound as 2 for the rank of SDR solution, being independent of the system parameters like K and N.

When the SDR solution $\{\mathbf{\Omega}_k^\star\}$ is rank-one, the optimal transmit beamforming vector can be obtained through Cholesky decomposition as $\mathbf{\Omega}_k^\star = \boldsymbol{\omega}_k^\star(\boldsymbol{\omega}_k^\star)^H$, $\forall k$. If the SDR solution is not rank-one, the randomization-based technique can be similarly adopted as in Section III to obtain a rank-one solution. Specifically, we first obtain the eigenvalue decomposition of $\mathbf{\Omega}_k^{n+1}$ as $\mathbf{\Omega}_k^{n+1} = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{U}_k^H$, $k = 1, \ldots, K$. Then K random vectors are generated as follows

$$\tilde{\boldsymbol{\omega}}_k = \mathbf{U}_k\mathbf{\Sigma}_k^{\frac{1}{2}}\mathbf{r}_k, \tag{29}$$

where $\mathbf{r}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ is a random vector. Finally, the optimal transmit beamforming vector $\boldsymbol{\omega}_k^{n+1}$ of (P1.3) is approximated as the largest one achieved by the best $\tilde{\boldsymbol{\omega}}_k$ among all $\mathbf{r}_k$'s.

C. Overall Algorithm

The overall algorithm is summarized in Algorithm 2. As shown, the algorithm optimizes $\{\boldsymbol{\omega}_k\}$ and $\mathbf{\Theta}$ alternatively in the outer-layer iteration, and the SDR technique is adopted to obtain an approximate phase-shift solution and beamforming solution. The algorithm ends when the the incremental increase of the objective value is sufficiently small. The convergence of Algorithm 2 is proved as follows.

Theorem 3: Algorithm 2 is guaranteed to converge.

Proof: First, in step 3 of Algorithm 2, since the optimal solution $\mathbf{\Theta}^{n+1}$ is obtained for given $\{\boldsymbol{\omega}_k^n\}$, we have the following inequality on the minimum rate

$$Q(\mathbf{\Theta}^n, \boldsymbol{\omega}_k^n) \leq Q(\mathbf{\Theta}^{n+1}, \boldsymbol{\omega}_k^n). \tag{30}$$

Second, in step 4 of Algorithm 2, since $\{\boldsymbol{\omega}^n\}$ is the optimal solution of (P1.3), the following inequality holds

$$Q(\mathbf{\Theta}^{n+1}, \boldsymbol{\omega}_k^n) \leq Q(\mathbf{\Theta}^{n+1}, \boldsymbol{\omega}_k^{n+1}). \tag{31}$$

Algorithm 2 Iterative Algorithm for Solving (P1)

1: Initialize $\{\boldsymbol{\omega}_k^0\}$, $\boldsymbol{\Theta}^0$, D (a large positive integer) and ϵ (a smaller positive value). Let $n = 0$.
2: **repeat**
3: For given $\{\boldsymbol{\omega}_k^n\}$, obtain $\boldsymbol{\Theta}^{n+1}$ through a similar procedure as described in steps 3 to 23 of Algorithm 1 with (P2.2) replaced by (P1.2).
4: For given $\boldsymbol{\Theta}^{n+1}$, apply the bisection search and the SDR techniques to solve (P1.4), and obtain the optimal SDR solution as $\{\boldsymbol{\Omega}_k^{n+1}\}$, then apply Gaussian-randomization technique to obtain an approximate rank-one solution $\{\boldsymbol{\omega}_k^{n+1}\}$.
5: Update iteration index $n = n + 1$.
6: **until** The increase of the objective value is smaller than ϵ.
7: Return the optimal solution $\boldsymbol{\omega}_k^\star = \boldsymbol{\omega}_k^n$, $\boldsymbol{\Theta}^\star = \boldsymbol{\Theta}^n$, and $Q^\star$.

From (30) and (31), we further have

$$Q(\boldsymbol{\Theta}^n, \boldsymbol{\omega}_k^n) \leq Q(\boldsymbol{\Theta}^{n+1}, \boldsymbol{\omega}_k^{n+1}). \tag{32}$$

The inequality in (32) indicates that the objective value of (P1) is always non-decreasing after each iteration. On the other hand, since the objective is continuous over the compact feasible set of (P1), it is upper-bounded by some finite positive number [48]. Hence, the proposed Algorithm 2 is guaranteed to converge, which completes the proof. ■

Notice that the BCD-based Algorithm 2 only converge to a suboptimal solution in general. The reason is as follows. The (P1) is not jointly convex with respect to $\boldsymbol{\Theta}$, $\{\boldsymbol{\omega}_k\}$ and Q, and the SDR method together with randomization-based approximation for solving the subproblems (P1.2) and (P1.4) returns only suboptimal solution.

In Algorithm 2, the subproblems (P1.2) and (P1.4) are alteratively solved in each outer-layer BCD iteration, and each subproblem is solved by a bisection search inner-layer iteration in each of which an SDR problem is solved. Similar to the complexity analysis for Algorithm 1, the complexity of Algorithm 2 is obtained from [45] as $\mathcal{O}\big(I_{\text{ite}}\big(\log_2(\frac{Q_{\max}-Q_{\min}}{\epsilon_{\text{b}}})\max\{\frac{K^2+3K-2}{2}, M+1\}^4 (M+1)^{\frac{1}{2}} \log\frac{1}{\epsilon_{\text{s}}} + \log_2\left(\frac{Q_{\max}-Q_{\min}}{\epsilon_{\text{b}}}\right)\max\{\frac{(K+1)K}{2}, N\}^4 N^{\frac{1}{2}} \log\frac{1}{\epsilon_{\text{s}}}\big)\big)$.

VI. Numerical Results

Numerical results are provided in this section to evaluate the performance of the proposed design. As shown in Fig. 2, we consider a three-dimensional coordinate system with the axes (x, y, z). We assume that the altitudes of the BS and the RIS are set as 10 meters (m), and the horizontal locations of the BS and the RIS are (0, 0) and $(25\sqrt{2}, 25\sqrt{2})$. The users, whose altitudes are set as 1.5 m, are randomly and uniformly distributed in the rectangular area formed by the origin point and the point $(50, 25\sqrt{2})$ of Fig. 2. The coordinates of users are randomly generated and then fixed in all simulations. The operating frequency is assumed to be 2.5 GHz [9], which implies that the wavelength $\lambda = 0.12$m. We assume that the BS is equipped with a uniform linear array with antenna spacing $d_{\text{B}} = \lambda/2$. On the other hand, the RIS is equipped with a uniform rectangular array with reflecting-element spacing $d_{\text{I}} = \lambda/2$. The number of reflecting elements is set to be $M = 40$, if not specified locally.

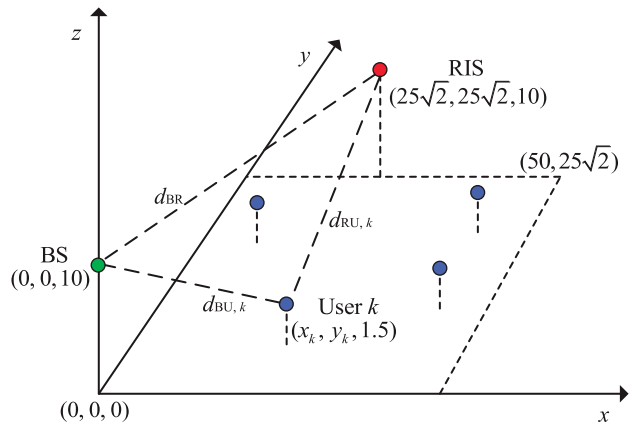


Fig. 2. Locations of an RIS-assisted NOMA system.

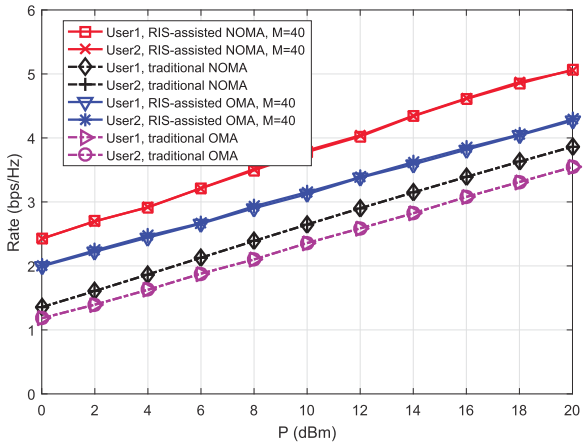
As in [33], we assume that the BS-to-user channels $\mathbf{v}_j$'s are Rayleigh fading distributed and the large-scale pathloss is $10^{-3}d^{-4}$, where d is the distance with unit of meter. Both the BS-to-RIS channel $\mathbf{F}$ and the RIS-to-user channels $\mathbf{g}_j$'s are assumed to be Rician fading distributed, and their pathloss are $10^{-3}d^{-2}$ and $10^{-3}d^{-2.5}$, respectively. The corresponding LoS components $\bar{\mathbf{F}}$ and $\bar{\mathbf{g}}_j$ are modeled by the steering vectors depending on the angle of arrival and angle of departure of particular LoS pathes [33]. We set the Rician factors $K_1 = K_2 = 10$. As in [49], we set $\sigma^2 = -114$ dBm. Let $D = 400$ and $\epsilon = 0.01$.

For performance comparison, we consider three benchmarks, i.e., traditional NOMA (without RIS but with optimized beamforming design at the BS), RIS-assisted OMA (with the optimized phase shifts at the RIS, and the optimized beamforming design as well as the allocation of degree of freedom (i.e., time/frequency resources) at the BS, and the traditional OMA (without RIS but with the optimized beamforming design as well as the allocation of degree of freedom at the BS). The achievable rate of each user for OMA is expressed as equation (8) in [50]. Simulation results are based on 1000 random channel realizations.

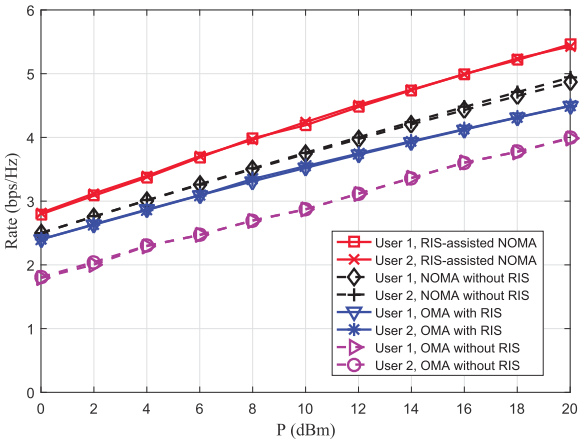
A. Rate Performance Evaluation for Proposed Solution

In this subsection, the numerical results for rate performance of the proposed RIS-assisted NOMA and three benchmarks are evaluated. We consider the setup of two users, i.e., $K = 2$. The coordinates of the two users are randomly generated as (32.52,23.48,1.5) and (48.45,19.55,1.5), respectively.

First, Fig. 3 (a) plots the per-user rate performance versus the BS's transmission power P for the proposed RIS-assisted NOMA and the three benchmarks, under the single-antenna BS setup (i.e., $N = 1$). In general, the proposed RIS-assisted NOMA achieves significant rate gains compared to the benchmarks. Specifically, for $P = 10$ dBm, the proposed RIS-assisted NOMA improves the rate performance by 53.2%,

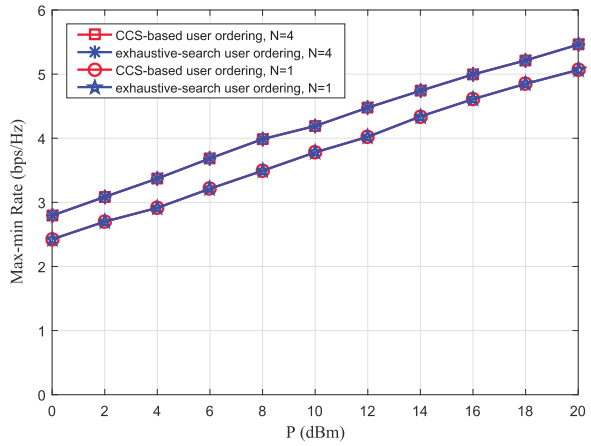


(a) Single-antenna BS setup.

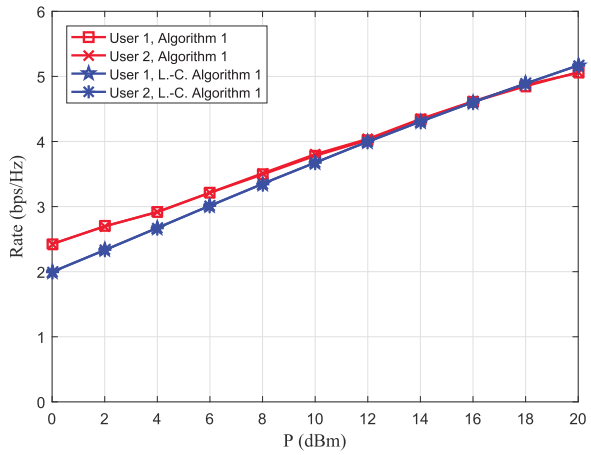


(b) Multiple-antenna BS setup.

Fig. 3. Rate comparison with different benchmarks.



(a) Comparison with exhaustive-search user ordering.



(b) Comparison with low-complexity solving scheme.

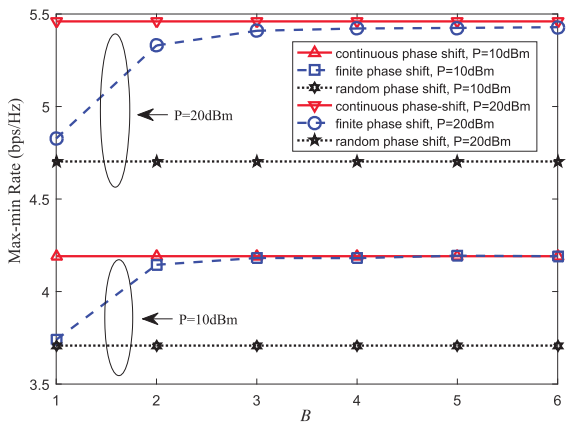
Fig. 4. Rate comparison with different solving schemes.

38.5%, and 14.3%, compared to the benchmarks of traditional OMA without RIS, traditional NOMA without RIS, and RIS-assisted OMA, respectively. The rate gain of RIS-assisted NOMA compared to the traditional NOMA, comes from the enhanced CCS and larger channel-strength difference introduced by the RIS. The rate gain of RIS-assisted NOMA compared to RIS-assisted OMA is due to the higher SE of NOMA relative to OMA. Notice that the superiority of NOMA compared to OMA still remains after introducing the RIS. Also, user 1 and user 2 achieve almost the same rate under each scheme, achieving best user fairness. The practical significance of this proposed RIS-assisted NOMA lies in that it enables the NOMA system to achieve higher rate and larger coverage than traditional NOMA and OMA systems while ensuring user fairness.

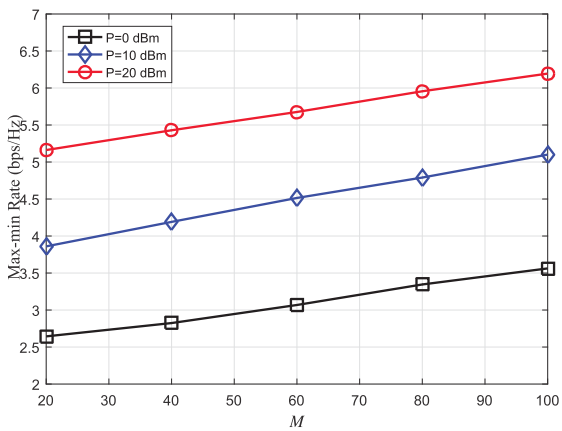
Fig. 3 (b) plots the per-user rate versus the BS's transmission power P for the proposed RIS-assisted NOMA and the three benchmarks, under the multiple-antenna BS setup. We set the number of BS antennas as $N = 4$. Similar to the single-antenna setup, it is observed that the RIS-assisted NOMA achieves significant rate gains compared to the three benchmarks, which verifies the enhanced SE of NOMA and the benefits of applying RIS to the downlink MISO-NOMA systems. Moreover, Fig. 3(a) for single-antenna BS setup reveals that RIS-OMA outperforms traditional NOMA, while Fig. 3(b) for multiple-antenna BS setup reveals that traditional NOMA outperforms RIS-OMA. Since traditional NOMA and RIS-OMA achieves rate gain through power-domain (non-orthogonal) multiplexing and reflecting link, respectively, the comparative relationship of rate performance depends on specific system parameters.

Then, Fig. 4 (a) compares the max-min rate performance of the proposed CCS-based user-ordering scheme with that of the exhaustive-search scheme. It is observed that the rate of the CCS-based user-ordering scheme achieves almost the same performance as the exhaustive-search scheme which needs to search all $K!$ possible user orders, for both setups of a single-antenna BS and a multiple-antenna BS. This numerically verifies the performance advantages of the proposed CCS-based user-ordering scheme, besides it low-complexity feature.

Fig. 4 (b) compares the per-user rate performance of the general Algorithm 1 with that of the low-complexity (L.-C.) Algorithm 1 for the RIS-assisted two-user NOMA system. Recall that the low-complexity Algorithm 1 uses Proposition 1 to compute the phase-shift solution for given $\boldsymbol{\alpha}^n$, instead of using the general procedure based on the bisection search and the SDR technique in Algorithm 1. When the BS's transmission power P increases, the low-

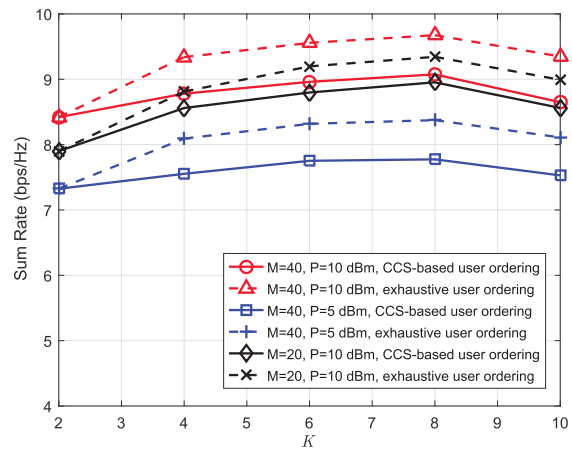


(a) Effect of phase-shift quantization bits B.

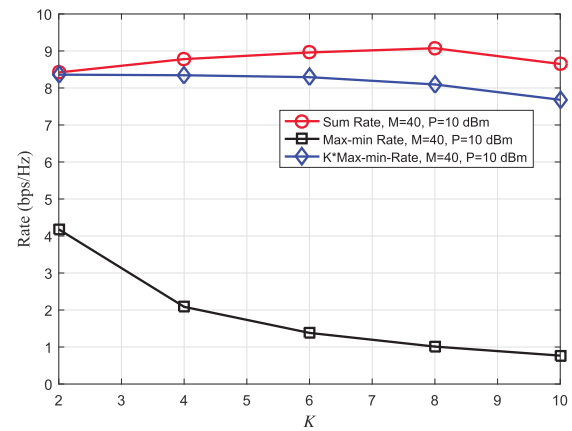


(b) Effect of number of reflecting elements M.

Fig. 5. Effect of RIS parameters on rate performance.



(a) Sum rate versus number of users K.



(b) Max-min rate versus number of users K.

Fig. 6. Rate performance versus number of users K.

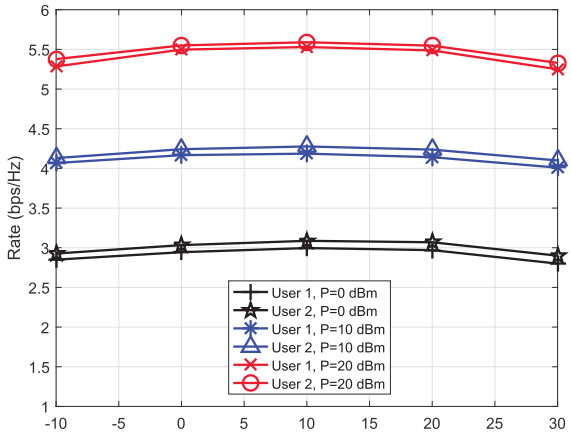
complexity Algorithm 1 suffers from slight rate degradation compared to Algorithm 1. Moreover, when P is higher than 16 dBm, the low-complexity Algorithm 1 even outperforms Algorithm 1. This is because that for low or moderate P, the low-complexity Algorithm 1 maximizes the combined channel of the stronger user, without strengthening the weaker user's combined channel which may result into relatively low rate for the weaker user; while for high P, the low-complexity Algorithm 1 which uses the closed-form phase-shift solution being proved to be asymptotically optimal, outperforms Algorithm 1. Therefore, the low-complexity Algorithm 1 is efficient for the two-user NOMA systems.

B. Effect of System Parameters on Rate Performance

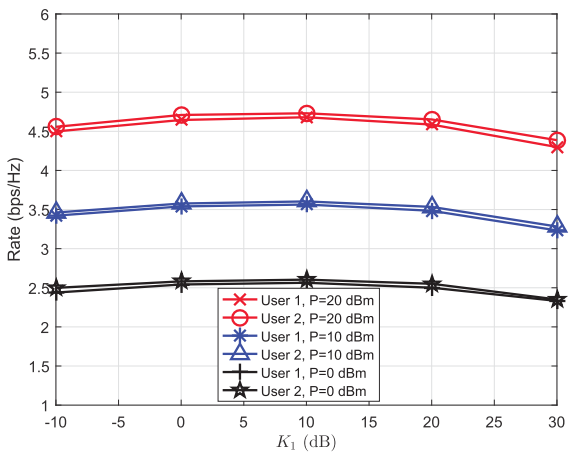
In this subsection, the effects of the main system parameters on the performance are investigated.

The RIS has two main parameters. One is the the number of phase-shift quantization bits denoted as B, which determines the resolution of the implementable phase shifts of the IRS, and the other is the number of reflecting elements M. Fig. 5 (a) plots the max-min rate versus B for the multiple-antenna BS setup, for the BS's transmission power $P = 10$ dBm and $P = 20$ dBm. Each optimized continuous phase shift θ_m is quantized to its nearest discrete value in the set $\{0, \frac{2\pi}{2^B}, \ldots, \frac{2\pi(2^B-1)}{2^B}\}$. It is observed that the RIS's finite-resolution phase shifters in general degrade the max-min rate compared to the ideal RIS with infinite-resolution phase shifters, but the rate performance degradation becomes negligible very quickly as B increases. Since the RIS-related channels may be difficult to obtain in practice, we have compared the max-min rate achieved by the optimized phase shift with that achieved by random phase shift. It is observed that the optimized phase shift can enhance the max-min rate. For instance, for case of $P = 10$ dBm, the optimized continuous phase shift design and 3-bit discretized phase shift design can increase the max-min rate by about 16.0% and 14.9%, respectively. Also, Fig. 5 (b) plots the max-min rate versus M, for the BS's transmission power $P = 10,\ 20$ and 30 dBm. As expected, the max-min rate increases with M.

Fig. 6 plots the rate performance versus the number of NOMA users K. Fig. 6(a) compares the sum rate performance for the proposed CCS-based user ordering scheme and the (high-complexity) exhaustive user ordering scheme. It is observed that as K increases, the sum rate increases first and then decreases, and achieves the maximum value for $K = 8$. This reveals that there is a tradeoff between the number of NOMA users and the sum rate. This can be explained as follows. For a small or moderate number of NOMA users, the overall SINR (thus the sum rate) can be



(a) Rate versus K_1 for RIS-NOMA with BS-to-user links.



(b) Rate versus K_1 for RIS-NOMA without BS-to-user links.

Fig. 7. Rate performance versus Rician factor K_1 of BS-to-RIS link.

improved by utilizing the joint active and passive beamforming design, thus the sum rate benefits from more multiplexing users. However, for a relatively large number of NOMA users, the joint beamforming design is not sufficient to mitigate the severe inter-user interference for decoding, thus the sum rate decays. Fig. 6(b) verifies the rate fairness for different NOMA users K, under the proposed CCS-based user ordering scheme. By comparing the circle-marker curve for sum rate and the diamond-marker curve for the K multiples of the max-min rate, we observe that both users for the two-user NOMA setup have almost the same rate and sufficiently good rate fairness among users can still be guaranteed for cases of more than two users. For instance, for the case of $K = 8$, the fluctuation among the rates of different users is no larger than 10%.

Fig. 7 plots the rate performance as the Rician factor K_1 of BS-to-RIS channel grows to infinity in two-user NOMA system. Set $M = 40$. Fig. 7(a) shows the rate performance for RIS-NOMA system with BS-to-user link as in previous simulations. It is observed that the rate of each user increases first and then decreases slightly as K_1 grows. The variation of rate performance comes from the characteristic of the reflecting-link channels. To validate this point, we simulate the rate performance for RIS-NOMA system without BS-to-user link (i.e., with only the reflecting link). Fig. 7(b) shows that its rate also increases first and then decreases as K_1 grows. The explanation is as follows. It can be shown that the correlation of the reflecting-link channel vectors for different users increases as K_1 grows, the proof of which is standard and thus omitted herein due to limited space. On the one hand, when K_1 approaches zero, the correlation tends to zero, for sufficiently large number of reflecting elements in practice. For less correlated user channels, the rate performance of NOMA will degrade, being even worse than OMA for case of orthogonal user channels [4]. On the other hand, when K_1 tends to infinity, the reflecting-link channel vectors become fully correlated due to the pinhole channel characteristic, i.e., the reflecting-link channel matrix becomes rank-one. This limits the reflecting-link channel's degrees of freedom to one, thus degrades the rate performance of NOMA transmission [51].

VII. Conclusion

This paper has investigated the problem of rate optimization for an RIS-assisted downlink NOMA system. The minimum SINR (i.e., equivalently the rate) of all users are maximized by jointly optimizing the BS's transmit beamforming and the RIS's phase shifts. Efficient algorithms are proposed to solve the formulated non-convex problem, by leveraging the block coordinated descent and semidefinite relaxation techniques. Numerical results show that the RIS-assisted downlink NOMA system can enhance the rate performance significantly, compared to traditional NOMA without RIS and traditional OMA with/without RIS, and a practical RIS with low-resolution phase shifters can approximate the best-achievable rate performance achieved by an ideal RIS with infinite-resolution phase shifters. Other interesting future work for RIS-assisted NOMA includes the outage performance analysis, rate performance under imperfect CSI, etc.

Appendix

A. Proof of Proposition 1

Proof: We have the following NOMA-decoding SINR for both users

$$\gamma_{1\to1} = \frac{\alpha_1 P|h_1|^2}{\alpha_2 P|h_1|^2 + \sigma^2}, \tag{33}$$

$$\gamma_{2\to2} = \frac{\alpha_2 P|h_2|^2}{\sigma^2}. \tag{34}$$

For sufficiently high transmission power P at the BS such that $\alpha_2\ P|h_1|^2 \gg \sigma^2$, we have $\gamma_{1\to1} \approx \frac{\alpha_1}{\alpha_2}$. Given α_1 and α_2, $\gamma_{1\to1}$ is determined, while $\gamma_{2\to2}$ is monotonically increasing with respect to $|h_2|^2$. To maximize the minimum value between $\gamma_{1\to1}$ and $\gamma_{2\to2}$, it suffices to maximize the 2-nd user's combined channel strength $|h_2|^2 = |\mathbf{g}_2^H\mathbf{\Theta}\mathbf{f} + v_2|^2$ by optimizing the phase shifts $\mathbf{\Theta}$.

Specifically, the combined channel strength $|h_2|^2$ can be rewritten as $|\mathbf{e}^H \operatorname{diag}(\mathbf{g}_2^H)\mathbf{f} + v_2|^2$ by some variable substitutions, which can be further expressed as $\left|\sum_{i=1}^{M}\left(|[\mathbf{g}_2^H]_i||[\mathbf{f}]_i|e^{j(\theta_i+\varphi_{2,i}+\psi_i)}\right) + |v_2|e^{j\xi_2}\right|^2$. Hence, it is standard to show that the optimal phase shifts that maximize $|h_2|^2$ are given by $\theta_i = \xi_2 - \varphi_{2,i} - \psi_i$, for $i = 1, \ldots, M$. ■

B. Proof of Lemma 1

Let $\boldsymbol{\alpha}^*$ denote the power allocation vector satisfying (20) (21) and γ^* denote the obtained equal SINR. The optimality and uniqueness of $\boldsymbol{\alpha}^*$ are proved in the sequel.

First, we prove $\boldsymbol{\alpha}^*$ is the optimal solution of (P2.1). Assume $\boldsymbol{\alpha}^{**} \neq \boldsymbol{\alpha}^*$ with corresponding max-min SINR γ^{**}, and $\gamma^{**} > \gamma^*$. Due to the constraint (14d), there must be an element of $\boldsymbol{\alpha}^{**}$ is smaller than that of $\boldsymbol{\alpha}^*$. As proved in the next paragraph, if any element of $\boldsymbol{\alpha}^*$ decrease, the max-min SINR would be smaller than γ^*, which contradicts with the previous assumption. Hence, $\boldsymbol{\alpha}^*$ and γ^* are the optimal solution and the optimal objective value, respectively.

For the K-th user, by using the fact that γ_K is a monotonically increasing function of α_K, the conclusion is clear. Therefore, in order to ensure $\gamma_K \geq \gamma^*$, α_K cannot be reduced. For the $K-1$ -th user, $\gamma_{K-1} = \frac{\alpha_{K-1}P|h_{K-1}|^2}{\alpha_K P|h_{K-1}|^2+\sigma^2}$, due to non-decreasing α_K, the reduction of α_{K-1} will directly result into smaller γ_{K-1} and thus α_{K-1} cannot be reduced either. The remaining $\gamma_{K-2}, \gamma_{K-3}, \ldots$, and γ_1 can be sequentially analyzed in the same manner.

Second, we prove that $\boldsymbol{\alpha}^*$ is the unique solution of the equations (20) and (21). From (20), we have the following recursive equations

$$\alpha_K(Q) = \frac{Q\sigma^2}{P|h_k|^2}, \tag{35}$$

$$\alpha_k(Q) = \frac{Q}{P|h_k|^2}\left(\sum\nolimits_{i=k+1}^{K} \alpha_i P|h_k|^2 + \sigma^2\right), \tag{36}$$

for $k = K-1, K-2, \ldots, 1$. It can be easily shown that each $\alpha_k(Q)$ is strictly and monotonically increasing with Q. Therefore, $\sum_{k=1}^{K} \alpha_k$ also strictly monotonically increases as Q increases. Thus, there exist a unique positive value Q^* which satisfies (20) and (21). The optimal power allocation coefficients $\boldsymbol{\alpha}^*$ are thus unique, and obtained as $\alpha_k^*(Q^*)$ for each k, which completes the proof.

C. Proof of Theorem 1

The equations in (20) can be written as

$$\tilde{\boldsymbol{\alpha}}\frac{1}{Q} = \mathbf{D}\boldsymbol{\Psi}\tilde{\boldsymbol{\alpha}} + \frac{\sigma^2}{P}\mathbf{D}\mathbf{1}. \tag{37}$$

Multiplexing both sides by $\mathbf{1}^T$ yields

$$\frac{1}{Q} = \mathbf{1}^T\mathbf{D}\boldsymbol{\Psi}\tilde{\boldsymbol{\alpha}} + \mathbf{1}^T\frac{\sigma^2}{P}\mathbf{D}\mathbf{1}. \tag{38}$$

Define $\tilde{\boldsymbol{\alpha}}_{\mathsf{ext}} = [\tilde{\boldsymbol{\alpha}}; 1]$. From (37) and (38), an eigenvalue system can be constructed as

$$\boldsymbol{\Pi}\tilde{\boldsymbol{\alpha}}_{\mathsf{ext}} = \frac{1}{Q}\tilde{\boldsymbol{\alpha}}_{\mathsf{ext}}, \tag{39}$$

where Q is a reciprocal eigenvalue of the nonnegative matrix $\boldsymbol{\Pi}$.

It is obvious that $Q > 0, \tilde{\boldsymbol{\alpha}}_{\mathsf{ext}} \geq 0$ must be satisfied to represent physical meaning. According to Perron-Frobenius theory, for any nonnegative real matrix $\mathbf{B}_{K\times K} \geq 0$, whose spectral radius is $\rho(B)$, there exists a vector $\mathbf{y} \geq 0$ such that $\mathbf{B}\mathbf{y} = \rho(B)\mathbf{y}$, thus the maximal eigenvalue $\rho(B)$ and the corresponding eigenvector are always nonnegative. Therefore, the optimal solution of (P2.3) is

$$Q = \frac{1}{\lambda_{\max}(\boldsymbol{\Pi})}. \tag{40}$$

And the optimal power allocation vector $\boldsymbol{\alpha}$ is given by the first K components of the dominant eigenvector of $\boldsymbol{\Pi}$, which can be scaled such that its last component equals 1.

D. Proof of Theorem 2

Let $\mu_{t,k} \geq 0$ and $\nu \geq 0$ be the dual variables corresponding to the constraints given in (28b) and (28d), respectively, where $1 \leq t \leq k \leq K$. Let $\mathbf{S}_k \succeq 0$ be the dual variable corresponding constraints $\boldsymbol{\Omega}_k \succeq 0$ in (28c). The Lagrangian of (P1.4) is then written as

$$\begin{aligned} &L(\{\boldsymbol{\Omega}_k\}, \mu_{t,k}, \nu, \{\mathbf{S}_k\}) \\ &= -\sum_{t=1}^{K}\sum_{k=t}^{K}\mu_{t,k}\left[\mathrm{Tr}\left(\mathbf{H}_k\boldsymbol{\Omega}_t\right) - Q\sum_{i=t+1}^{K}\mathrm{Tr}\left(\mathbf{H}_k\boldsymbol{\Omega}_i\right) - Q\sigma^2\right] \\ &\quad + \nu\left[\sum_{k=1}^{K}\mathrm{Tr}\left(\boldsymbol{\Omega}_k\right) - P\right] - \sum_{k=1}^{K}\mathrm{Tr}\left(\mathbf{S}_k\boldsymbol{\Omega}_k\right). \end{aligned} \tag{41}$$

Let $\{\boldsymbol{\Omega}_k^*\}$, $\mu_{t,k}^*$, ν^* and $\{\mathbf{S}_k^*\}$ be the optimal primal and dual variables, respectively. Since (P1.4) is convex for given Q and satisfies the Slaters condition, the strong duality holds for this problem. Hence, the optimal primal and dual solutions should satisfy the Karush-Kuhn-Tucker conditions

$$\begin{aligned} &\nabla_{\boldsymbol{\Omega}_l} L(\{\boldsymbol{\Omega}_k^*\}, \mu_{t,k}^*, \nu^*, \{\mathbf{S}_k^*\}) \\ &\quad = -\sum_{k=l}^{K}\mu_{l,k}^*\mathbf{H}_k + Q\sum_{t=1}^{l-1}\sum_{k=t}^{K}\mu_{t,k}^*\mathbf{H}_k + \nu^*\mathbf{I} - \mathbf{S}_l^* = 0, \end{aligned} \tag{42}$$

$$\mathbf{S}_l^*\boldsymbol{\Omega}_l^* = 0. \tag{43}$$

By multiplying (42) by $\boldsymbol{\Omega}_l^*$ on both sides and substituting (43) into the obtained equation, we have

$$\sum_{k=l}^{K}\mu_{l,k}^*\mathbf{H}_k\boldsymbol{\Omega}_l^* - Q\sum_{t=1}^{l-1}\sum_{k=t}^{K}\mu_{t,k}^*\mathbf{H}_k\boldsymbol{\Omega}_l^* = \nu^*\boldsymbol{\Omega}_l^*. \tag{44}$$

Recall $\mathbf{H}_k = \mathbf{h}_k^H\mathbf{h}_k$ and $\mathbf{h}_k = \mathbf{g}_k^H\boldsymbol{\Theta}\mathbf{F} + \mathbf{v}_k^H = \mathbf{e}^H\,\mathrm{diag}\{\mathbf{g}_k^H\}\mathbf{F} + \mathbf{v}_k^H$, by introducing

$$\boldsymbol{\Gamma}_k = \begin{bmatrix}\mathrm{diag}\{\mathbf{g}_k^H\}\mathbf{F} \\ \mathbf{v}_k^H\end{bmatrix}, \tag{45}$$

the $\mathbf{h}_k$ can be rewritten as $\bar{\mathbf{e}}^H\boldsymbol{\Gamma}_k$. Thus, we have

$$\left(\sum_{k=l}^{K}\mu_{l,k}^*\boldsymbol{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\boldsymbol{\Gamma}_k - Q\sum_{t=1}^{l-1}\sum_{k=t}^{K}\mu_{t,k}^*\boldsymbol{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\boldsymbol{\Gamma}_k\right)\boldsymbol{\Omega}_l^* = \nu^*\boldsymbol{\Omega}_l^*. \tag{46}$$

Since $\sum_{k=l}^{K}\mu_{l,k}^*\boldsymbol{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\boldsymbol{\Gamma}_k = \mathbf{Z}_k\bar{\mathbf{e}}\bar{\mathbf{e}}^H\mathbf{Z}_k^H$, where $\mathbf{Z}_k = \sum_{k=l}^{K}\mu_{l,k}^*\boldsymbol{\Gamma}_k^H$, thus we have

$$\mathrm{rank}\left(\sum_{k=l}^{K}\mu_{l,k}^*\boldsymbol{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\boldsymbol{\Gamma}_k\right) \leq \mathrm{rank}\left(\bar{\mathbf{e}}\bar{\mathbf{e}}^H\right) = 1. \tag{47}$$

Similarly, $\text{rank}\left(\sum_{t=1}^{l-1}\sum_{k=t}^{K}\mu_{t,k}^{*}\mathbf{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\mathbf{\Gamma}_k\right) \leq 1$. The following derivations complete the proof

$$\begin{aligned}
&\text{rank}(\mathbf{\Omega}_l^*)\\
&= \text{rank}\left(\sum_{k=l}^{K}\mu_{l,k}^{*}\mathbf{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\mathbf{\Gamma}_k - Q\sum_{t=1}^{l-1}\sum_{k=t}^{K}\mu_{t,k}^{*}\mathbf{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\mathbf{\Gamma}_k\right)\mathbf{\Omega}_l^*\\
&\leq \text{rank}\left(\sum_{k=l}^{K}\mu_{l,k}^{*}\mathbf{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\mathbf{\Gamma}_k\right) + \ldots\\
&\qquad \text{rank}\left(\sum_{t=1}^{l-1}\sum_{k=t}^{K}\mu_{t,k}^{*}\mathbf{\Gamma}_k^H\bar{\mathbf{e}}\bar{\mathbf{e}}^H\mathbf{\Gamma}_k\right)\\
&\leq 2. \qquad (48)
\end{aligned}$$

References

[1] G. Yang, X. Xu, and Y.-C. Liang, "Intelligent reflecting surface assisted non-orthogonal multiple access," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Seoul, South Korea, May 2020, pp. 1–6.

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

[3] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, 2015.

[4] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174–6189, Dec. 2016.

[5] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[6] M. D. Renzo *et al.*, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–20, May 2019.

[7] O. Ozdogan, E. Bjornson, and E. G. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 581–585, May 2020, doi: 10.1109/LWC.2019.2960779.

[8] Y.-C. Liang, R. Long, Q. Zhang, J. Chen, H. V. Cheng, and H. Guo, "Large intelligent surface/antennas (LISA): Making reflective radios smart," *J. Commun. Inf. Netw.*, vol. 4, no. 2, pp. 40–50, Jun. 2019.

[9] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 162–169, Sep. 2018.

[10] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[11] T.-J. Cui, M.-Q. Qi, X. Wan, J. Zhao, and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light, Sci. Appl.*, vol. 3, no. 10, pp. 215–218, Oct. 2014.

[12] Z. Ding and H. V. Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, May 2020.

[13] M. Fu, Y. Zhou, and Y. Shi, "Intelligent reflecting surface for downlink non-orthogonal multiple access networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[14] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in multi-antenna aided NOMA systems," 2019, *arXiv:1910.13636*. [Online]. Available: http://arxiv.org/abs/1910.13636

[15] F. Zhu, Z. Lu, J. Zhu, J. Wang, and Y. Huang, "Beamforming design for downlink non-orthogonal multiple access systems," *IEEE Access*, vol. 6, pp. 10956–10965, Jun. 2018.

[16] H. M. Al-Obiedollah, K. Cumanan, J. Thiyagalingam, A. G. Burr, Z. Ding, and O. A. Dobre, "Energy efficient beamforming design for MISO non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4117–4131, Jun. 2019.

[17] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[18] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Aug. 2015.

[19] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[20] 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.

[21] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X.-G. Xia, "Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2961–2974, May 2018.

[22] G. Yang, X. Xu, and Y.-C. Liang, "Resource allocation in NOMA-enhanced backscatter communication networks for wireless powered IoT," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 117–120, Jan. 2020.

[23] Q. Zhang, L. Zhang, Y.-C. Liang, and P.-Y. Kam, "Backscatter-NOMA: A symbiotic system of cellular and Internet-of-Things networks," *IEEE Access*, vol. 7, pp. 20000–20013, Feb. 2019.

[24] G. Yang, Q. Zhang, and Y.-C. Liang, "Cooperative ambient backscatter communications for green Internet-of-Things," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1116–1130, Apr. 2018.

[25] R. Long, Y.-C. Liang, H. Guo, G. Yang, and R. Zhang, "Symbiotic radio: A new communication paradigm for passive Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1350–1363, Feb. 2020.

[26] C. Huang, A. Zappone, M. Debbah, and C. Yuen, "Achievable rate maximization by passive intelligent mirrors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canda, Apr. 2018, pp. 1–6.

[27] 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.

[28] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064–3076, May 2020.

[29] J. Chen, Y.-C. Liang, Y. Pei, and H. Guo, "Intelligent reflecting surface: A programmable wireless environment for physical layer security," *IEEE Access*, vol. 7, pp. 82599–82612, Jul. 2019.

[30] B. Di, H. Zhang, L. Li, L. Song, Y. Li, and Z. Han, "Practical hybrid beamforming with finite-resolution phase shifters for reconfigurable intelligent surface based multi-user communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4565–4570, Apr. 2020.

[31] 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.

[32] O. Ozdogan, E. Bjornson, and E. G. Larsson, "Using intelligent reflecting surfaces for rank improvement in MIMO communications," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 9160–9164.

[33] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238–8242, Aug. 2019.

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

[35] G. Yang, Y.-C. Liang, R. Zhang, and Y. Pei, "Modulation in the air: Backscatter communication over ambient OFDM carrier," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1219–1233, Mar. 2018.

[36] Q. Zhang, H. Guo, Y.-C. Liang, and X. Yuan, "Constellation learning-based signal detection for ambient backscatter communication systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 452–463, Feb. 2019.

[37] G. Yang, C. K. Ho, and Y. L. Guan, "Multi-antenna wireless energy transfer for backscatter communication systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2974–2987, Dec. 2015.

[38] G. Yang, R. Dai, and Y.-C. Liang, "Energy-efficient UAV backscatter communication with joint trajectory design and resource optimization," *IEEE Trans. Wireless Commun.*, early access, Oct. 13, 2020, doi: 10.1109/TWC.2020.3029225.

[39] Y.-C. Liang, Q. Zhang, E. G. Larsson, and G. Y. Li, "Symbiotic radio: Cognitive backscattering communications for future wireless networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 4, pp. 1242–1255, Dec. 2020.

[40] W. Zhao, G. Wang, S. Atapattu, T. A. Tsiftsis, and C. Tellambura, "Is backscatter link stronger than direct link in reconfigurable intelligent surface-assisted system?" *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1342–1346, Jun. 2020.

[41] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[42] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," 2019, *arXiv:1904.10136*. [Online]. Available: http://arxiv.org/abs/1904.10136

[43] Z. Liu, L. Lei, N. Zhang, G. Kang, and S. Chatzinotas, "Joint beamforming and power optimization with iterative user clustering for MISO-NOMA systems," *IEEE Access*, vol. 5, pp. 6872–6884, Jun. 2017.

[44] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1465–1468, Jul. 2016.

[45] Z.-Q. Luo, W.-K. Ma, A. 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.

[46] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, Jun. 2001.

[47] G. Pataki, "On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues," *Math. Oper. Res.*, vol. 23, no. 2, pp. 339–358, May 1998.

[48] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ., 2004.

[49] Z. Wei, D. W. K. Ng, and J. Yuan, "Joint pilot and payload power control for uplink MIMO-NOMA with MRC-SIC receivers," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 692–695, Apr. 2018.

[50] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[51] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7748–7764, Dec. 2020.