We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

6,500
Open access books available

176,000
International authors and editors

190M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

RS-Based MIMO-NOMA Systems in Multicast Framework

Sareh Majidi Ivari, Mohammad Reza Soleymani and Yousef R. Shayan

Abstract

This chapter presents a novel scheme that integrates the rate-splitting (RS) technique in Multiple Input Multiple Output (MIMO) systems with non-orthogonal multiple access (NOMA) to improve performance and capacity in wireless communication systems under imperfect channel state information at the transmitter (CSIT) and in overloaded regimes. The proposed approach addresses a general and realistic scenario, incorporating both unicast and multicast users, aiming to increase system throughput through the optimization of precoding vectors and power allocation. A generic power allocation optimization technique is introduced, which can be employed for maximizing both the minimum-rate and sum-rate, focusing on the rate of the weakest user within each group per cluster. To tackle the non-convex nature of the problems, the proposed technique leverages the WMMSE-rate relationship and an alternating optimization (AO) algorithm, transforming the problem into a convex one. The chapter provides a comprehensive analysis of the proposed scheme, offering a tutorial background and presenting novel insights for an enhanced understanding.

Keywords: MIMO, RS, NOMA, fairness, sum-rate

1. Introduction

In recent years, the demand for high-speed wireless communication has grown significantly, driven by the widespread use of smartphones, tablets, and other wireless devices [1]. Users now expect constant internet connectivity and access to high-quality voice and video services, putting tremendous pressure on wireless communication networks to keep up with the increasing demand.

One major challenge faced by wireless communication systems is the limited availability of radio spectrum. As more devices and users come online, the demand for radio spectrum increases. To address this challenge, new technologies have been developed to utilize the available spectrum more efficiently, such as the MIMO-NOMA scheme [2].

MIMO-NOMA is a promising technology that integrates multiple antenna technology (MIMO) with NOMA to enhance the efficiency and capacity of wireless communication systems [2]. MIMO techniques leverage the spatial dimension by transmitting multiple data streams simultaneously over the same frequency-time resource. In the NOMA scheme, the transmitter sends a superposition of messages for

multiple users, and users apply successive interference cancelation (SIC) to remove messages intended for weaker users and decode their own message [3]. By combining MIMO and NOMA, the MIMO-NOMA scheme brings together the advantages of both technologies, allowing multiple users to transmit and receive data concurrently using multiple antennas, non-orthogonal power allocation, and advanced signal processing techniques [4].

The integration of MIMO and NOMA is particularly useful in scenarios where multiple users are located in the same spatial direction but at distinct propagation distances, such as urban areas, stadiums, or office buildings [5]. In such scenarios, traditional wireless communication techniques like Orthogonal Multiple Access (OMA) may not provide sufficient capacity to serve all users [6]. In contrast, MIMO-NOMA can serve multiple users using the same resources, thereby improving the overall system capacity [2]. For instance, in a crowded stadium, many users may want to use their mobile devices to access the internet or stream videos simultaneously. MIMO-NOMA can serve these users using the same frequency band and time slot, whereas traditional techniques would require each user to be served in a separate time slot or frequency band.

In MIMO-NOMA systems, the transmitter employs interference cancelation techniques to form spatially orthogonal beams, with each beam carrying information for multiple users or groups of users [7]. The conventional linear precoding techniques such as Zero-forcing Beamforming (ZFBF) and Minimum Mean Square Error (MMSE) are commonly used to achieve spatial orthogonality [8, 9]. In the conventional linear precoding, the interference is canceled at the transmitter, and the receiver treats it as background noise [10]. These techniques play a crucial role in enhancing capacity, improving spectral efficiency, and enhancing overall system performance in wireless communication systems. In this chapter, the MIMO-NOMA scheme based on these conventional linear precoding is denoted as Conv-based MIMO-NOMA.

However, the implementation of MIMO-NOMA faces challenges, especially in the presence of imperfect CSIT and overloaded regime [10]. Imperfect CSIT can arise due to various factors, including channel estimation errors, quantization, and feedback delays [11, 12]. The accuracy of the channel information plays a vital role in interference cancelation techniques, and imperfect CSIT can degrade the performance of linear precoding methods.

Furthermore, wireless networks often comprise a combination of unicast and multicast users, which poses additional challenges for interference cancelation methods [13]. Accommodating both unicast and multicast users requires efficient resource allocation and power control strategies to optimize system performance and ensure fairness among users. Additionally, the performance of linear precoding techniques can deteriorate in overloaded regimes, where the number of users or groups of users exceeds the available resources. This further emphasizes the need for advanced techniques that can overcome the limitations of traditional linear precoding methods and improve system performance in realistic scenarios.

To address these challenges, the rate-splitting (RS) technique has emerged as a generic and powerful solution for interference cancelation in MIMO-NOMA systems [13]. RS decomposes the transmitted signal into two parts: a common part decoded by all users and a private part intended for the specific user. This enables the base station to exploit multiuser interference and achieve higher spectral efficiency [14].

RS has demonstrated significant potential for improving the sum-rate in multiuser MIMO systems under perfect CSIT conditions [15, 16]. It allows the base station to

exploit multiuser interference and achieve higher spectral efficiency by decomposing the signal into common and private parts. Several studies have shown that RS can increase system capacity, reduce interference, and improve MMF rate performance in scenarios with perfect CSIT [17, 18]. However, the assumption of perfect CSIT may not hold in real-world scenarios.

Researchers have examined the effects of imperfect CSIT on the sum-rate and MMF rate performance of RS-based MIMO systems and proposed robust RS algorithms to counteract the impacts of imperfect CSIT [16, 19]. However, most of these studies focus on RS in the unicast framework and do not consider realistic scenarios with both unicast and multicast users.

In addition to unicast transmission, RS has been studied in the context of multicast transmission in MIMO systems [20]. Multicast transmission presents unique challenges, as it involves simultaneously transmitting the same information to multiple users. Therefore, it has received more attention. The performance of RS has been investigated in terms of MMF rate in [20]. RS has also been explored in multibeam multicast satellite communication systems in terms of MMF rate [21, 22].

RS-based MIMO-NOMA in the uplink has been investigated in [23]. The MMF rate is optimized for the proposed system in the unicast framework. However, none of the aforementioned works consider RS in MIMO-NOMA in the downlink with realistic scenarios, considering both unicast and multicast users under imperfect CSIT.

This chapter investigates the use of RS in MIMO-NOMA in downlink under imperfect CSI and both unicast and multicast users. The objective is to investigate the potential of RS-based MIMO-NOMA to improve system throughput and user fairness in realistic scenarios with imperfect CSIT. The chapter provides a comprehensive guide for researchers, engineers, and students interested in understanding the principles and applications of RS-based MIMO-NOMA for future wireless communication systems.

1.1 Contributions and organization of the chapter

This chapter explores the use of RS and NOMA in multiuser MIMO systems in downlink and presents a comprehensive analysis of the proposed scheme while investigating the challenges and trade-offs involved in its implementation. The main contributions of this chapter include the first application of RS in multiuser MIMO-NOMA systems under imperfect CSIT assumption, where RS is used to cancel interference and combat the effects of imperfect CSIT.

The chapter also covers the derivation of achievable data rates for both the common and private parts of user groups in the proposed RS-based MIMO-NOMA system. Precoding vectors are designed for both the common and private parts to enhance performance. The common part's precoding vector is optimized to maximize the rate of the common message, while the private part's precoding vectors are designed to cancel interference in both unicast and multicast frameworks. A low-complexity technique for designing the private precoding vectors is proposed in multicast transmission to address the lack of spatial degrees of freedom. The proposed technique builds upon unicast linear precoding methods and employs a Singular Value Decomposition (SVD) mapper.

Furthermore, the chapter formulates the max-min fairness MMF rate and sum-rate optimization problems for the RS-based MIMO-NOMA system under imperfect CSIT using the Average Rate (AR) framework. The weighted minimum mean square error (WMMSE) approach is employed to transform the formulated MMF and sum-rate problems into convex problems. First, the chapter derives a rate-WMMSE relationship, and then, using this relationship and a low-complexity solution based on

alternating optimization (AO), the problems are transformed into equivalent convex problems.

Overall, this chapter provides a comprehensive analysis of the RS-based MIMO-NOMA system under imperfect CSIT, and the proposed solutions and derivations of achievable data rates and optimization problems offer valuable insights into the design of future MIMO-NOMA systems.

The chapter is organized as follows. It begins with an introduction to the system model, including the signal and CSIT models. The design of precoding vectors for both common and private parts is discussed in Section 3. Section 4 focuses on power allocation optimization problems to maximize the minimum rate and sum-rate. The performance of the proposed technique is evaluated through simulations in Section 5. Finally, the chapter concludes with a summary of the findings and potential future research directions in Section 6.

Notations: Throughout this chapter, the following notations are used. Boldface capital letters, boldface lowercase letters, and ordinary letters represent matrices, column vectors, and scalars, respectively. The real component of a complex number x is denoted by $\Re(x)$. The operators $()^T$ and $()^H$ represent transposition and Hermitian transpose, respectively. $\|$ and $\|$ are abbreviations for absolute value and Euclidean norm, respectively. $\mathbb{E}(.)$ represents the expected value of a random variable.

2. System model

This chapter presents a comprehensive study of RS-based MIMO-NOMA for a realistic wireless communication framework that includes both multicast and unicast users under imperfect CSIT assumption. The system consists of a single base station with N_t antennas serving I single-antenna users, where $N_t \leq I$. The base station forms K clusters and generates one beam per cluster. The users that are in the same spatial direction but with distinct propagation distances are grouped into a cluster. This helps enhance the channel gain and combat inter-cluster interference. The distinctive propagation distances also facilitate SIC at mobile users. The k-th cluster contains G_k groups which has M_{g_k} users, where $M_{g_k} \geq 1$. If $M_{g_k} = 1$, group g_k contains only one unicast user, and if $M_{g_k} > 1$, it contains multicast users.

The system model notation is defined, where $\mathcal{I}$ represents the set of indices of all users, $\mathcal{K}$ represents the set of clusters, $\mathcal{I}_k$ represents the set of users in the k th cluster, and $\mathcal{G}_k$ represents the set of groups of users in the k th cluster. The proposed system model of the RS-based MIMO-NOMA is depicted in **Figure 1**. In this figure, the parameters are as follows, $K = 4, I_1 = 4, I_2 = 1, I_3 = 3, I_4 = 2$. In cluster 1 and cluster 3 there are multicast users, $G_1 = 2$ and $M_{1_1} = 2, M_{2_1} = 2$ $G_2 = 1, M_{1_2} = 1$, $G_3 = 3, M_{1_3} = 1, M_{2_3} = 1, M_{3_3} = 1, G_4 = 2, M_{1_4} = 1, M_{2_4} = 1$.

The base station uses MIMO-NOMA to transmit multiple data streams simultaneously to the I users by encoding I messages into K streams from a single data source. To enhance the system capacity and user fairness, the base station employs RS to divide the data of each user into two parts: a common part and a private part. The common stream is decoded by all users, while the private part is intended only for the specific user. The private part of the k-th message is further split into G_k sub-streams, and each sub-stream is assigned to a group of users in the k-th cluster using the principles of NOMA. This helps to cancel inter-cluster interference and enhance the spectral efficiency. The combination of RS and NOMA provides flexibility in the
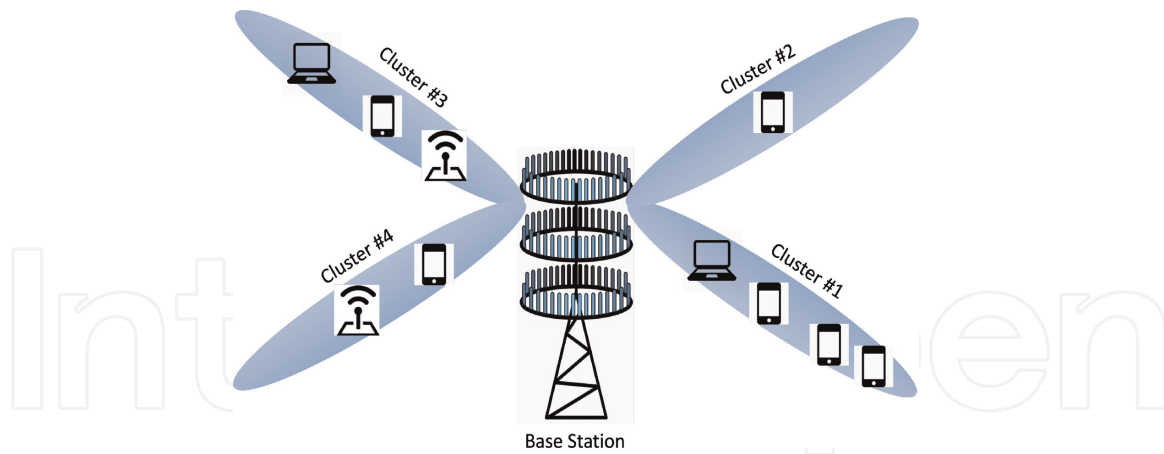
Figure 1.
System model of the proposed RS-based MIMO-NOMA with multicast and unicast users.

allocation of transmission power among users and the trade-off between system throughput and user fairness.

2.1 Signal model

In this section, we examine the transmitted and received signals to derive the signal-to-interference plus noise (SINR) and the achievable data rate. First, we need to discuss two main techniques of the proposed RS-based MIMO-NOMA: Rate Splitting and NOMA.

2.1.1 Rate-splitting approach

Generally in the L-layer RS, the transmitter splits the message of each cluster-k into L-sub-messages, $W_k^1, W_k^2, \ldots, W_k^L, \forall k \in \mathcal{K}$. Among the L messages, one message is shared by all users, which is called the common part. In this chapter, we consider 1-layer RS, in which a message is split into two parts: a common and a private messages. The common part of all messages $W_1^c, W_2^c, \ldots, W_K^c$ is packed together and encoded into a common stream s_c which is shared by all users. In the other hand, the private message of each message is encoded independently into private streams, $W_k^p \rightarrow s_k$.

As a result, the transmitted signal in time unit is $\mathbf{x}(t)$, where the time units are omitted for simplicity of expression. Therefore, the transmitted signal is

$$\mathbf{x} = \sqrt{p_c}\mathbf{w}_c s_c + \sum_{k=1}^{K} \mathbf{w}_k \sqrt{p_k} s_k \tag{1}$$

where $\mathbf{w}_c$ is the unit-norm precoding vector of the common message and $\mathbf{w}_k$ precodes the k-th message. p_c is the power allocated to the common part. p_k is the power allocated to the k-th cluster. The transmitted signal is constrained to

$$p_c + \sum_{k=1}^{K} p_k \|\mathbf{w}_k\|^2 \leq P_\mathrm{T} \tag{2}$$

where P_T is the maximum available power at the transmitter.

2.1.2 NOMA approach

Following the NOMA scheme in the power domain, different groups of users in a cluster are allocated different power levels according to their channel conditions to obtain the maximum gain in system performance. The transmitter sends all users information by sending the superposition of messages. Such power allocation is also beneficial to separate different groups of users. Therefore, users can apply SIC to cancel interference from the weaker groups of users in a cluster. However, the weak users perform single user detection (SUD) with considering the interference from the stronger users as the background noise. According to the NOMA scheme, the private stream can be contained information for more than one group of users. It means that the private stream s_k consists of

$$s_k = \sum_{g_k=1}^{G_k} \sqrt{\alpha_{g_k}} s_{k_g},$$ (3)

where α_{g_k} ($\sum_{g_k=1}^{G_k} \alpha_{g_k} = 1$) denote fraction of the power allocated to g-th group in cluster k.

The received signal at user-i is $y_i = \mathbf{h}_i \mathbf{x} + n_i, \forall i \in \mathcal{I}$. In terms of notation, $\mathbf{h}_i \in \mathbb{C}^{1 \times N_t}$ is the channel vector between the transmitter and i-th user. This chapetr defines $\mu(i)$ as mapping a user index to its corresponding cluster and group indices, $\mu : i \to (k, g_k)$. Therefore, the received signal by i-th user which belongs to k-th cluster and g_k-th group is expressed as

$$y_i = \sqrt{p_c} \mathbf{h}_i \mathbf{w}_c s_c + \sum_{k=1}^{K} \sqrt{p_k} \mathbf{h}_i \mathbf{w}_k s_k + n_i,$$ (4)

where $n_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the additive noise terms that contaminate the reception of i-th user. By substituting the Eq. (3) into the Eq. (4), the received signal is

$$y_i = \sqrt{p_c} \mathbf{h}_i \mathbf{w}_c s_c + \sqrt{\alpha_{g_k} p_k} \mathbf{h}_i \mathbf{w}_k s_{g_k} + \sum_{h_k > g_k}^{G_k} \sqrt{\alpha_{h_k} p_k} \mathbf{h}_i \mathbf{w}_k s_{h_k} + \sum_{j=1, j \neq k}^{K} \sqrt{p_j} \mathbf{h}_i \mathbf{w}_j s_j + n_i.$$

(5)

According to the RS technique, each user firstly decodes the common stream s_c and treats the private streams as noise. Therefore, the SINR of the common part of user-i is:

$$\gamma_{c,i} = \frac{p_c |\mathbf{h}_i \mathbf{w}_c|^2}{\sum_{j=1}^{K} p_j |\mathbf{h}_i \mathbf{w}_j|^2 + \sigma_i^2},$$ (6)

and its corresponding rate is $R_{c,i} = \log_2(1 + \gamma_{c,i})$. In the RS scheme, the common message, s_c, is shared among all beams and groups, and each user should be able to decode s_c. Therefore, the common rate is defined as

$$R_c = \min_{i \in \mathcal{I}} R_{c,i} \triangleq \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} C_{g_k},$$ (7)

where C_{g_k} denotes the portion of common rate of group g_k in the k-th cluster.

After users decode and remove the common signal, s_c through SIC, then each users decodes its private message. According to the NOMA scheme, users in group g_k in cluster k, $\forall k \in \mathcal{K}$, $\forall g_k \in \mathcal{G}_k$, perform SIC to decode s_{h_k}, $\forall h_k < g_k$ and remove it from the received signal. Finally, users apply SUD to decode s_{g_k} by considering all the other interference streams as noise. Therefore, the SINR of i-th user is determined by

$$\gamma_i = \frac{\alpha_{g_k} p_k |\mathbf{h}_i \mathbf{w}_k|^2}{1 + \sum_{h_k > g_k}^{G_k} \alpha_{h_k} p_k |\mathbf{h}_i \mathbf{w}_k|^2 + \sum_{j=1, j \neq k}^{K} p_j |\mathbf{h}_i \mathbf{w}_j|^2}. \tag{8}$$

In the multicast transmission, to guarantee all users can decode their messages, the user with the lowest SINR within a group dictates the rate of the corresponding group. Therefore, the achievable rate of group g_k in cluster k, r_{g_k}, is defined by

$$r_{g_k} \triangleq \min_{i \in \mathcal{I}_{g_k}} R_i, \tag{9}$$

Therefore, the rate of users in group g_k are composed of C_{g_k} and r_{g_k} and written as

$$R_{g_k} = C_{g_k} + r_{g_k}, \tag{10}$$

and the sum-rate is $R_{\text{sum-rate}} = R_c + \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} r_{g_k}$.

2.2 CSIT uncertainty model

In this study, we assume that the receiver has perfect channel state information (CSI), while the transmitter has imperfect CSI due to limited feedback, such as quantized feedback with a fixed number of bits. The imperfect CSI of user i is modeled as

$$\mathbf{h}_i = \hat{\mathbf{h}}_i + \tilde{\mathbf{h}}_i \tag{11}$$

where $\hat{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i$ denote the estimated channel state and the corresponding channel estimation error at the transmitter, respectively. The uncertainty in CSIT (i.e., the channel estimation error) can be characterized by a conditional density $f\left(\mathbf{h}|\hat{\mathbf{h}}\right)$ that is known at the transmitter.

Consider i-th user and we define

$$\Upsilon_i = \mathbb{E}|\mathbf{h}_i|^2,$$
$$\hat{\Upsilon}_i = \mathbb{E}\left|\hat{\mathbf{h}}_i\right|^2, \tag{12}$$
$$\tilde{\Upsilon}_i = \mathbb{E}\left|\tilde{\mathbf{h}}_i\right|^2.$$

According to the orthogonal principles, $\hat{\mathbf{h}}_i$ and $\tilde{\mathbf{h}}_i$ are uncorrelated, and $\tilde{\mathbf{h}}_i$ has a zero mean. Therefore,

$$\Upsilon_i = \hat{\Upsilon}_i + \tilde{\Upsilon}_i. \tag{13}$$

We can consider

$$\tilde{\Upsilon}_i = \sigma_{e,i}^2 \Upsilon_i \tag{14}$$

where $\sigma_{e,i}^2 \in [0,1]$ is the normalized CSIT error variance [16, 24]. Therefore, we have

$$\hat{\Upsilon}_i = \left(1 - \sigma_{e,i}^2\right)\Upsilon_i \tag{15}$$

A value of $\sigma_{e,i}^2 = 1$ corresponds to no instantaneous CSIT, while a value of $\sigma_{e,i}^2 = 0$ represents perfect instantaneous CSIT. For simplicity, we assume that all users have identical normalized CSIT error variances, that is, $\sigma_{e,i}^2 = \sigma_e^2, \forall i \in \mathcal{I}$.

The CSIT error variance scales with the signal-to-noise ratio (SNR) as $\sigma_e^2 = P_T^{-\eta}$, where $\eta \in [0, \infty)$ is the CSIT quality parameter. η can be interpreted as a relation to the number of feedback bits, where $\eta = 0$ corresponds to a fixed number of feedback bits for all SNRs, and $\eta = \infty$ corresponds to an infinite number of feedback bits. The CSIT quality parameter is truncated such that $\beta \in [0, 1]$. In this context, $\eta = 1$ corresponds to perfect CSIT in the multiplexing gain sense [16, 24].

3. Precoder design

The proposed RS-based MIMO-NOMA system requires careful design of linear precoding vectors and power allocation to optimize performance and capacity. Firstly, in this section, we investigate the design of the linear precoding vectors for the private and common parts, denoted as $\mathbf{w}_k$ and $\mathbf{w}_c$, respectively. The linear precoding vectors $\mathbf{w}_k$ and $\mathbf{w}_c$ should be designed in a way that mitigates inter cluster interference and maximizes the achievable rate of the common message. In the following subsections, we investigate the design of the linear precoding vectors for the private and common parts.

3.1 Linear precoding vector of the private part, $\mathbf{w}_k$

Designing the linear precoding vector for the private part in the unicast framework under perfect CSIT is a relatively straightforward process. The optimal structure of w_k is a generalization of regularized zero-forcing (RZF) precoding. However, in the presence of imperfect CSIT, the optimal precoders for private messages are still unknown and must be optimized numerically, as shown in [25]. This optimization process becomes particularly complex in large-scale systems. Despite this, RZF based on the channel estimates $\hat{\mathbf{H}}$ can be a suitable strategy for precoders of private messages, based on the findings of [26].

In the context of multicast transmission, designing the linear precoding vectors $\mathbf{w}_k$ is particularly challenging due to the matrix characterization of each cluster rather than a vector. To address this challenge, we propose a novel approach based on singular value decomposition (SVD) mapping. Specifically, we use the SVD mapping to transform the multicast transmission scenario into a set of parallel unicast channels, where the optimization problem is simplified. We then use the RZF technique to

design the linear precoding vectors for the private messages in the unicast channels. This approach provides a low-complexity and efficient solution for designing the precoding vectors in the presence of imperfect CSIT.

The precoding vector using the ZBF is obtained as:

$$\mathbf{W_{RZF}} = \frac{1}{\sqrt{\gamma_{\mathbf{RZF}}}} \left(\left(\hat{\mathbf{G}}^H \hat{\mathbf{G}} + \frac{K}{P_T} \mathbf{I}K \right)^{-1} \hat{\mathbf{G}}^H \right), \tag{16}$$

where $\hat{\mathbf{G}}$ is the estimated composite channel matrix, and $\mathbf{I}_K$ is the K-dimensional identity matrix. To ensure that the power constraints are satisfied, the precoding matrix should be normalized by the factor $\gamma_{\mathbf{RZF}}$, which is defined as:

$$\gamma_{\mathbf{RZF}} = \max_k \left(\mathrm{diag} \left(\mathbf{W_{RZF}} (\mathbf{W_{RZF}})^H \right) \right). \tag{17}$$

Here, $\mathrm{diag}(\mathbf{A})$ denotes the diagonal elements of a matrix $\mathbf{A}$, and $(\mathbf{A})^H$ represents the conjugate transpose of $\mathbf{A}$.

The estimated composite channel matrix $\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \dots, \hat{\mathbf{g}}_K]$ is obtained using the SVD mapping per beam [27]. In the SVD mapper, the estimated channel matrix of users in cluster k, denoted by $\hat{\mathbf{C}}_k = \left[\hat{\mathbf{h}}_{\mathcal{I}_k(1)}^H, \dots, \hat{\mathbf{h}}_{\mathcal{I}_k(2M)}^H \right]$, is first subjected to SVD as follows:

$$\hat{\mathbf{C}}_k^H \hat{\mathbf{C}}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^H, \tag{18}$$

where Σ_k is the diagonal matrix of singular valusers, and $\mathbf{U}_k$ ($\mathbf{V}_k$) gathers the left-singular vectors (right-singular vectors) [27]. Then, the right or left singular vector corresponding to the highest singular value is selected, which constructs the $\hat{\mathbf{g}}_k$ vector. The SVD mapper improves the energy spread over the users and the robustness to the CSIT uncertainty.

3.2 Linear precoding vector of the common part, $\mathbf{w}_c$

The precoding vector of the common message, $\mathbf{w}_c$, is designed to maximize the achievable rate of the common message. Therefore, the optimization problem is defined as

$$\overline{\mathcal{D}}_1 : \max_{\mathbf{w}_c \in \mathcal{N}} \min_{i \in \mathcal{I}} \frac{p_c |\mathbf{h}_i \mathbf{w}_c|^2}{\sum_{j=1}^K p_j |\mathbf{h}_i \mathbf{w}_j|^2 + \sigma_i^2} \tag{19}$$

$$\text{s.t.} \quad \|\mathbf{w}_c\|^2 = 1 \tag{20}$$

Since there is no interference in receiving the common message, the precoding vector for the common part can be designed as a linear combination of the channel vectors of all users, for a realization of $n \in \mathcal{N}$, the precoder of the common message is designed as

$$\mathbf{w}_c = \sum_{i \in \mathcal{I}} a_i \hat{\mathbf{h}}_i^H. \tag{21}$$

where a_i is the weight for the channel vector of user i, and $\hat{\mathbf{h}}_i^H$ is the conjugate transpose of the normalized channel vector of user i. By assuming $\sigma_{e,i}^2 = \sigma_e^2$, $\|\mathbf{h}_i\|^2 = 1$, $\|\mathbf{h}_i\hat{\mathbf{h}}_j^H\|^2 = (1-\sigma_e^2)\varepsilon^2$, $\forall i \in \mathcal{I}, j \neq i$, and substituting (21) into (19) and (20), the problem $\overline{\mathcal{D}}_1$ is equivalently transformed to $\overline{\mathcal{D}}_2$

$$\overline{\mathcal{D}}_2: \max_{a_i} \min_{i \in \mathcal{I}} \pi_i(1-\sigma_e^2)a_i^2 + \pi_i(1-\sigma_e^2)\varepsilon^2 \sum_{n=1,\,n\neq i}^{I} a_n^2 \tag{22}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}} a_i^2 = \frac{1}{N_t} \tag{23}$$

The goal is to find the optimal weights a_i that maximize the minimum SINR of all users, subject to the constraint that the sum of the squared weights is equal to $1/(N_t)$. The optimal solution of problem $\overline{\mathcal{D}}_2$ is obtained when all terms are equal [28], that is, $\pi_i a_i^2 + \pi_i \varepsilon^2 \sum_{n=1,\,n\neq i}^{I} a_n^2 = \pi_j a_j^2 + \pi_j \varepsilon^2 \sum_{n=1,\,n\neq j}^{I} a_n^2$, $\forall i \neq j$. Therefore, the optimal precoding vector is achieved when all users experience the same common part SINR (6). In this chapter for simplicity and in order to obtain a more insightful and tractable asymptotic performance, we consider that $\pi_i = \pi_j$, $\forall i \neq j$ and ε is very small, then the optimal a_i is equal to $a_i^* = 1/\sqrt{N_t I}$, where I is the total number of users.

4. Power allocation optimization

In this section, we examine the optimal power allocation for maximizing the MMF rate and sum-rate in the proposed RS-based MIMO-NOMA system under imperfect CSIT. To formulate the optimization problem under imperfect CSIT, we adopt a Stochastic Average Rate (AR) framework. Stochastic ARs are short-term metrics that represent the expected performance across the CSIT error distribution for a specific channel state estimate.

To define the AR framework, we first introduce three matrices: $\mathbf{H}$, $\hat{\mathbf{H}}$, and $\tilde{\mathbf{H}}$ which comprise the users' channel coefficients, users' channel coefficient estimations, and estimation errors. Given that the channel coefficients of users are independent and identically distributed (i.i.d.) and a sample index set $\mathcal{N} = \{1, 2, \ldots, N\}$, we construct a realization sample containing N i.i.d. realizations drawn from a conditional distribution $f(\mathbf{H}|\hat{\mathbf{H}})$. The realization sample can be expressed as:

$$\mathbb{H}^N \triangleq \left\{ \mathbf{H}^{(n)} = \hat{\mathbf{H}} + \tilde{\mathbf{H}}^{(n)} | \hat{\mathbf{H}}, n \in \mathcal{N} \right\}. \tag{24}$$

The realizations are accessible at the transmitter and can be utilized to approximate the ARs experienced by each user using Sample Average Functions (SAFs). As the number of samples (N) approaches infinity, $N \to \infty$, according to the strong law of large numbers, the ARs for user-i are as follows:

$$\overline{R}_{c,i} = \lim_{N\to\infty} \overline{R}_{c,i}^{(N)} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} R_{c,i}\left(\mathbf{H}^{(n)}\right), \tag{25}$$

$$\overline{R}_i = \lim_{N \to \infty} \overline{R}_i^{(N)} = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} R_i\left(\mathbf{H}^{(n)}\right) \tag{26}$$

where $R_{c,i}\left(\mathbf{H}^{(n)}\right)$ and $R_i\left(\mathbf{H}^{(n)}\right)$ are the achievable rates for i-th user based on the n-th realization in the sample set $\mathbb{H}^{(N)}$. The AR framework is then used to formulate the optimization problems for power allocation to maximize the MMF rate and sum-rate.

4.1 Problem statement

We define the optimization problems in this section. The AR framework is used to formulate the MMF and sum-rate optimization problems under imperfect CSIT.

4.1.1 Max-min fairness analysis

The MMF optimization problem using the AR framework can be formulated as

$$\overline{\mathcal{P}}_1 : \underset{\mathbf{p}, \alpha, \overline{\mathbf{c}}}{\arg\max} \quad \min_{k \in \mathcal{K}} \min_{g_k \in \mathcal{G}_k} \left\{ \overline{C}_{g_k} + \min_{i \in \mathcal{I}_{g_k}} \overline{R}_i^{(N)} \right\} \tag{27}$$

s.t.

$$\overline{R}_{c,i}^{(N)} \geq \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} \overline{C}_{g_k}, \forall i \in \mathcal{I} \tag{28}$$

$$\overline{C}_{g_k} \geq 0, \forall g_k \in \mathcal{G}_k, \forall k \in \mathcal{K} \tag{29}$$

$$\alpha_{g_k} \in [0,1], \sum_{g_k=1}^{G_k} \alpha_{g_k} = 1, \forall k \in K \tag{30}$$

$$p_c + \sum_{k=1}^{K} p_k \|\mathbf{w}_k\|^2 \leq P_{\mathrm{T}}, \forall k \in \mathcal{K} \tag{31}$$

here $\overline{\mathbf{c}} = \left[\overline{C}_{1,1}, \ldots, \overline{C}_{1,G}, \ldots, \overline{C}_{K,1}, \ldots, \overline{C}_{g_k} \right]$ is the vector of Average common-rate portions, and $\mathbf{p} = \{p_c, p_1, p_2, \ldots, p_K\}$, $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ are the vectors of powers and fraction of powers. The constraint (28) guarantees s_c to be decoded by each user since the definition of the Average common rate is $\overline{R}_c = \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} \overline{C}_{g_k} = \min_{i \in \mathcal{I}} \overline{R}_{c,i}$. Constraint (29) implies that each portion of the Average common rate is non-negative. Constraints (30) and (31) are the power constraint. By solving Problem $\overline{\mathcal{P}}_1$, variables $(\overline{\mathbf{c}}, \mathbf{p}, \alpha)$ are jointly optimized. Note that by fixing $p_c = 0$ and $\overline{\mathbf{c}} = 0$, the RS scheme turns into Conv-based MIMO-NOMA.

4.1.2 Sum-rate analysis

The sum-rate optimization is another problem which is addressed in this chapter. The sum-rate maximization under imperfect CSIT is also formulated using the AR framework as

$$\overline{\mathcal{S}}_1 : \underset{\overline{R}_c, \mathbf{p}, \alpha}{\text{argmax}} \quad \overline{R}_c + \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} \min_{i \in I_{g_k}} \overline{R}_i^{(N)} \tag{32}$$

s.t.

$$\overline{R}_{c,i}^{[N]} \geq \overline{R}_c, \forall i \in \mathcal{I} \tag{33}$$

$$(24d), (24e) \tag{34}$$

where $\overline{R}_c$ is an auxiliary variable. The constraint (33) guarantees that all users can decode s_c.

Problems $\overline{\mathcal{P}}_1$ and $\overline{\mathcal{S}}_1$ are non-convex problems that are very challenging to solve because they contain superimposed rate expressions.

The weighted mean squared error (WMMSE) approach is a powerful technique for solving non-convex optimization problems with superimposed rate expressions. The idea behind this approach is to replace the original rate expressions with a set of WMMSE expressions that are easier to handle mathematically. Using the WMMSE expressions, the original problems $\overline{\mathcal{P}}_1$ and $\overline{\mathcal{S}}_1$ can be reformulated as block-wise convex optimization problems, which can be solved iteratively using interior-point methods. Specifically, the reformulated problems involve optimizing the WMMSE variables and the power allocation coefficients, subject to some convex constraints. The optimization procedure involves iteratively updating the WMMSE variables and the power allocation coefficients until convergence is achieved.

4.2 Rate-WMMSE relationship

To define the achievable data rate with set of WMMSE expression, first we establish the Rate-WMMSE relationship. The mean square errors (MSEs) of the estimate $\hat{s}_{c,i}$ of the common signal s_c for user i is given by:

$$\varepsilon_{c,i} = \mathbb{E}\left\{ |\hat{s}_{c,i} - s_{c,i}|^2 \right\} = \mathbb{E}\left\{ \left| \hat{s}_{c,i} - q_{c,i} y_i \right|^2 \right\}, \tag{35}$$

where $q_{c,i}$ is a scalar equalizer. Since the transmitter sends the superposition of s_c and s_{g_k}, $\forall k \in \mathcal{K}, g_k \in \mathcal{G}_k$, user i first decodes and removes s_c from the received signal using SIC. Next, user i which belongs to cluster k and group g_k decodes and removes the signals intended for the weaker groups in cluster k, $h_k < g_k$, through the SIC. Therefore, the MSE of the estimate $\hat{s}_{g_k}$ of the private signal s_{g_k} for user i in group g_k of cluster k is given by:

$$\varepsilon_i = \mathbb{E}\left\{ |\hat{s}_i - s_i|^2 \right\} = \mathbb{E}\left\{ \left| \hat{s}_i - q_i \left(y_i - \sqrt{(1-t)P} \mathbf{h}_i \mathbf{w}_c s_c - \sum_{h_k=1}^{h_k < g_k} \sqrt{\alpha_{h_k} p_k} \mathbf{h}_i \mathbf{w}_k s_{h_k} \right) \right|^2 \right\} \tag{36}$$

Here, $\mathbb{E}[\cdot]$ denotes the expectation operator, and $|\cdot|^2$ denotes the squared magnitude. With substituting the Eq. (5) into the Eq. (35) and (36), the MSEs of the common and private parts can be rewritten as

$$\varepsilon_{c,i} = \left|q_{c,i}\right|^2 T_{c,i} + 1 - 2\Re\left\{\sqrt{p_c}q_{c,i}\mathbf{h}_i\mathbf{w}_c\right\} \tag{37}$$

$$\varepsilon_i = \left|q_i\right|^2 T_i + 1 - 2\Re\left\{\sqrt{\alpha_{g_k}p_k}q_i\mathbf{h}_i\mathbf{w}_k\right\} \tag{38}$$

where

$$T_{c,i} = p_c|\mathbf{h}_i\mathbf{w}_c|^2 + \sum_{k=1}^K p_k|\mathbf{h}_i\mathbf{w}_k|^2 + \sigma_i^2, \tag{39}$$

$$T_i = p_k\alpha_{g_k}|\mathbf{h}_i\mathbf{w}_k|^2 + p_k\sum_{h_k>g_k}^{G_k}\alpha_{h_k}|\mathbf{h}_i\mathbf{w}_k|^2 + \sum_{j=1,j\neq k}^K p_j|\mathbf{h}_i\mathbf{w}_j|^2 + \sigma_i^2 \tag{40}$$

Moreover, we define the interference as

$$I_{c,i} = T_{c,i} - p_c|\mathbf{h}_i\mathbf{w}_c|^2, \tag{41}$$

$$I_i = T_i - p_k\alpha_{g_k}|\mathbf{h}_i\mathbf{w}_k|^2. \tag{42}$$

The optimum equalizers achieve by minimizing the MSEs over equalizers,

$$\frac{\partial\varepsilon_{c,i}}{q_{c,i}} = 0 \rightarrow q_{c,i}^{\mathrm{MMSE}} = \sqrt{p_c}\mathbf{h}_i\mathbf{w}_c T_i^{-1} \tag{43}$$

$$\frac{\partial\varepsilon_i}{q_i} = 0 \rightarrow q_i^{\mathrm{MMSE}} = \sqrt{p_k\alpha_{g_k}}\mathbf{h}_i\mathbf{w}_k T_i^{-1} \tag{44}$$

The minimum MSEs (MMSEs) with optimum equalizers are

$$\varepsilon_{c,i}^{\mathrm{MMSE}} = \min_{q_{c,i}}\varepsilon_{c,i} = T_{c,i}^{-1}I_{c,i}, \tag{45}$$

$$\varepsilon_i^{\mathrm{MMSE}} = \min_{q_i}\varepsilon_i = T_i^{-1}I_i. \tag{46}$$

Apparently, the SINRs can be expressed in the form of MMSEs, i.e., $\gamma = \left(1/\varepsilon^{\mathrm{MMSE}}\right) - 1$. Consequently, the corresponding rates are written as $R = -\log_2\left(\varepsilon^{\mathrm{MMSE}}\right)$.

Next, we define the augmented weighted MSEs (WMSEs) for the common and private parts. The term "augmented WMSE" is employed because it incorporates additional information or constraints into the standard WMSE, aiming to better capture the characteristics of the system under consideration, such as fairness or rate requirements, and facilitate the optimization process. This augmentation is particularly relevant in wireless communication system optimization problems, especially when dealing with RS or non-orthogonal multiple access techniques, to achieve more accurate and reliable results. The weighted WMSEs are given by:

$$\xi_{c,i} = u_{c,i}\varepsilon_{c,i} - \log_2(u_{c,i}), \xi_i = u_i\varepsilon_i - \log_2(u_i), \tag{47}$$

where $u_{c,i}, u_i > 0$ are weights associated with MSEs. In the following, we consider ξ s as WMSEs and, for simplicity, drop the "augmented". After defining the augmented

WMSEs, they are minimized with respect to both equalizers and weights, yielding the following conditions:

$$\frac{\partial \xi_{c,i}\left(q_{c,i}^{\text{MMSE}}\right)}{\partial q_{c,i}, u_{c,i}} = 0, \tag{48}$$

$$\frac{\partial \xi_i\left(q_i^{\text{MMSE}}\right)}{\partial q_i, u_i} = 0. \tag{49}$$

Then the optimal equalizers are substituted into the WMSEs, and we obtain

$$\xi_{c,i}\left(q_{c,i}^{\text{MMSE}}\right) = \min_{q_{c,i}} \xi_{c,i} = u_{c,i}\varepsilon_{c,i}^{\text{MMSE}} - \log_2(u_{c,i}) \tag{50}$$

$$\xi_i\left(q_i^{\text{MMSE}}\right) = \min_{q_i} \xi_i = u_i\varepsilon_i^{\text{MMSE}} - \log_2(u_i) \tag{51}$$

As a result, the optimum weights can be determined as:

$$u_{c,i} = \left(\varepsilon_{c,i}^{\text{MMSE}}\right)^{-1}, \tag{52}$$

$$u_i = \left(\varepsilon_i^{\text{MMSE}}\right)^{-1}. \tag{53}$$

We substitute (52) and (53) into (50), (51), leading to the Rate-WMMSE relationship

$$\xi_{c,i}^{\text{MMSE}} = \min_{q_{c,i}, u_{c,i}} \xi_{c,i} = 1 + \log_2 \varepsilon_{c,i}^{\text{MMSE}} = 1 - R_{c,i} \tag{54}$$

$$\xi_i^{\text{MMSE}} = \min_{q_i, u_i} \xi_i = 1 + \log_2 \varepsilon_i^{\text{MMSE}} = 1 - R_i. \tag{55}$$

With considering imperfect CSIT, a Stochastic Average Rate-WMMSE relationship is developed, and the average WMMSEs are given by:

$$\overline{\xi}_{c,i}^{\text{MMSE}(N)} = \frac{1}{N} \lim_{N \to \infty} \sum_{n=1}^{N} \xi_{c,i}^{\text{MMSE}(n)} = 1 - \overline{R}_{c,i}^{(N)}, \tag{56}$$

$$\overline{\xi}_i^{\text{MMSE}(N)} = \frac{1}{N} \lim_{N \to \infty} \sum_{n=1}^{N} \xi_i^{\text{MMSE}(n)} = 1 - \overline{R}_i^{(N)} \tag{57}$$

where $\xi_{c,i}^{\text{MMSE}(n)}$ and $\xi_i^{\text{MMSE}(n)}$ are associated with the n-th realization in $\mathbb{H}^{(N)}$. The sets of optimum MMSE equalizers associated with (56) and (57) are defined as

$$\mathbf{g}_{c,i}^{\text{MMSE}} = \left\{q_{c,i}^{\text{MMSE}(n)} | n \in \mathcal{N}\right\}, \tag{58}$$

$$\mathbf{g}_i^{\text{MMSE}} = \left\{q_i^{\text{MMSE}(n)} | n \in \mathcal{N}\right\}. \tag{59}$$

Moreover, the sets of optimum weights are

$$\mathbf{u}_{c,i}^{\text{MMSE}} = \left\{u_{c,i}^{\text{MMSE}(n)} | n \in \mathcal{N}\right\}, \tag{60}$$

$$\mathbf{u}_i^{\text{MMSE}} = \left\{ u_i^{\text{MMSE}(n)} | n \in \mathcal{N} \right\}. \tag{61}$$

Therefore, in each realization in $\mathbb{H}^{(N)}$, the optimum equalizer and weights are calculated. The composite set of optimum equalizer and weights are defined as

$$\mathbf{G}^{\text{MMSE}} = \left\{ \mathbf{g}_{c,i}^{\text{MMSE}}, \mathbf{g}_i^{\text{MMSE}} | i \in \mathcal{I}, \right\} \tag{62}$$

$$\mathbf{U}^{\text{MMSE}} = \left\{ \mathbf{u}_{c,i}^{\text{MMSE}}, \mathbf{u}_i^{\text{MMSE}} | i \in \mathcal{I} \right\} \tag{63}$$

Using the Rate-WMMSE relationship, the optimization problems are rewritten using the WMMSE variables in the following section.

4.3 WMMSE reformulation

In this section, we reformulate the optimization problems using the WMMSE expressions.

4.3.1 Max-min fairness analysis

Using the Rate-WMMSE relationship, and auxiliary variables, $\overline{z}$, $\mathbf{G}$, $\mathbf{U}$, $\overline{r}_g = \left\{ \overline{r}_{1,g}, \dots, \overline{r}_{g_k} \right\}$, the problem $\overline{\mathcal{P}}_1$ can be transferred into an equivalent WMMSE problem, $\overline{\mathcal{P}}_2$:

$$\overline{\mathcal{P}}_2 : \underset{\mathbf{p}, \, \alpha, \, \overline{\mathbf{c}}, \, \overline{z}, \, \overline{r}_g}{\text{argmax}} \quad \overline{z} \tag{64}$$

s.t.

$$\overline{C}_{g_k} + \overline{r}_{g_k} \geq \overline{z}, \forall k \in \mathcal{K}, \forall g_k = \{1, \dots, G_k\} \tag{65}$$

$$1 - \overline{\xi}_i^{(N)} \geq \overline{r}_{g_k}, \forall i \in \mathcal{I}_{g_k}, \forall k \in \mathcal{K}, \forall g_k = \{1, \dots, G_k\} \tag{66}$$

$$1 - \overline{\xi}_{c,i}^{(N)} \geq \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} \overline{C}_{g_k}, \forall i \in \mathcal{I} \tag{67}$$

$$(24d), (24e) \tag{68}$$

where $\overline{\xi}_{c,i}$ and $\overline{\xi}_i$ are given in (47). It is worth to mention if $\left(\mathbf{p}^*, \alpha^*, \overline{\mathbf{c}}^*, \overline{z}^*, \mathbf{G}^*, \overline{r}_g^*, \mathbf{U}^* \right)$ satisfies the KKT optimality conditions of $\overline{\mathcal{P}}_2$, $(\mathbf{p}^*, \alpha^*, \overline{\mathbf{c}}^*)$ will satisfy the KKT optimality conditions of $\overline{\mathcal{P}}_1$.

4.3.2 Sum-rate analysis

Motivated by the Rate-WMMSE relationships given in (56), (57), and the auxiliary variable, $\left(\overline{\xi}_c, \mathbf{U}, \mathbf{G} \right)$, the problem $\overline{\mathcal{S}}_1$ is equivalently transferred into the problem $\overline{\mathcal{S}}_2$. The problem is reformulated as

$$\overline{\mathcal{S}}_2 : \underset{\overline{\xi}_c, \, \mathbf{p}, \, \alpha_k}{\text{argmin}} \quad \overline{\xi}_c + \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} \max_{i \in \mathcal{I}_{g_k}} \overline{\xi}_i^{(N)} \tag{69}$$

$$\text{s.t.}$$

$$\overline{\xi}_{c,i}^{[N]} \leq \overline{\xi}_c, \forall i \in \mathcal{I} \tag{70}$$

$$(24d), (24e) \tag{71}$$

where $\overline{\xi}_c$ refers to the common AWMSE. Noted problem $\overline{\mathcal{S}}_2$ and problem $\overline{\mathcal{S}}_1$ are equivalence. It means that for any point $\left(\mathbf{p}^*, \alpha^*, \overline{\xi}_c^*, \mathbf{G}^*, \mathbf{U}^*\right)$ satisfying the KKT optimality conditions of problem $\overline{\mathcal{S}}_2$, $(\mathbf{p}^*, \alpha^*)$ satisfies the KKT optimality conditions of problem $\overline{\mathcal{S}}_1$.

The problems $\overline{\mathcal{P}}_2$ and $\overline{\mathcal{S}}_2$ remain non-convex. However, they become convex when two out of the three variables, namely equalizer, weight, and power, are fixed. Taking into account this block-wise convexity property, we propose an Alternating Optimization algorithm to address the problems $\overline{\mathcal{P}}_2$ and $\overline{\mathcal{S}}_2$.

4.4 Alternating optimization algorithm

The problems $\overline{\mathcal{P}}_2$ and $\overline{\mathcal{S}}_2$ remain non-convex for the entire set of optimization variables, which include α, p, $\overline{\mathbf{c}}$, $\mathbf{U}$, and $\mathbf{G}$. However, they exhibit block-wise convexity, which can be leveraged to propose an alternating optimization algorithm. Each iteration of the algorithm consists of two steps: (1) updating $\mathbf{U}$ and $\mathbf{G}$ based on the value of $\mathbf{p}$ and α from the previous iteration, and (2) updating $\mathbf{p}$, α, and $\overline{\mathbf{c}}$ using $\mathbf{U}$ and $\mathbf{G}$ obtained in step 1. We now provide a detailed explanation of these two steps.

4.4.1 Step 1: Updating $\mathbf{G}$, $\mathbf{U}$

In l-th iteration, all the equalizers and weights are updated according to the $\mathbf{p}, \alpha$ form the previous round, $l - 1$, $\mathbf{G}\left(p^{[l-1]}, \alpha^{[l-1]}\right)$, $\mathbf{U}\left(p^{[l-1]}, \alpha^{[l-1]}\right)$. The corresponding SAFs $\overline{u}_{c,i}, \overline{u}_i, \overline{g}_{c,i}, \overline{g}_i$ are calculated by taking average over N realization. To facilitate the next step, we introduce a set of variables are

$$t_{c,i} = u_{c,i}^{(n)}\left|q_{c,i}^{(n)}\right|^2, \quad t_i = u_i^{(n)}\left|q_i^{(n)}\right|^2, \tag{72}$$

$$\Psi_{c,i}^{(n)} = t_{c,i}\mathbf{h}_i^{(n)H}\mathbf{h}_i^{(n)}, \quad \Psi_i^{(n)} = t_i\mathbf{h}_i^{(n)H}\mathbf{h}_i^{(n)}, \tag{73}$$

$$f_{c,i}^{(n)} = u_{c,i}^{(n)}q_{c,i}^{(n)}\mathbf{h}_i^{(n)}\mathbf{w}_c^{(n)}, \quad f_i^{(n)} = u_i^{(n)}q_i^{(n)}\mathbf{h}_i^{(n)}\mathbf{w}_k^{(n)} \tag{74}$$

$$v_{c,i}^{(n)} = \log_2(u_{c,i}), \quad v_i^{(n)} = \log_2(u_i) \tag{75}$$

and the corresponding SAFs are calculated in the same way,

$$t_{c,i}^{(N)}, \Psi_{c,i}^{(N)}, f_{c,i}^{(N)}, v_{c,i}^{(N)}, t_i, \Psi_i^{(N)}, f_i^{(N)}, v_i^{(N)} \tag{76}$$

4.4.2 Step 2: Updating $\mathbf{p}, \alpha$

In the l-th iteration up to this step, we fix $\mathbf{G}, \mathbf{U}$, and the other introduced variables, which are obtained using the updated valusers of $(\mathbf{U}, \mathbf{G})$. With these updated

variables, in this step, the problems $\overline{\mathcal{P}}_2$ and $\overline{\mathcal{S}}_2$ transform into problems $\overline{\mathcal{P}}_3^{[l]}$ and $\overline{\mathcal{S}}_3^{[l]}$, which are convex problems. These problems can be solved using interior-point methods, allowing for the optimization of $\mathbf{p}$, $\alpha_{\mathbf{k}}$, and the other auxiliary variables.

$$\overline{\mathcal{P}}_3^{[l]} : \underset{\mathbf{p}, \alpha, \overline{\mathbf{c}}, \overline{z}, \overline{r}_g,}{\text{argmax}} \ \overline{z} \tag{77}$$

$$\text{s.t.}$$

$$\overline{C}_{k_g} + \overline{r}_{k_g} \geq \overline{z}, \quad \forall k \in \mathcal{K}, \forall g_k = \{1, \dots, G_k\} \tag{78}$$

$$1 - \overline{r}_{g_k} \geq \sum_{j=1, j \neq k}^{K} p_j \overline{\mathbf{w}}_j^H \overline{\Psi}_i^{(N)} \overline{\mathbf{w}}_j^{(N)} - 2\mathcal{R}\left\{ \sqrt{\alpha_{g_k} p_k} \overline{f}_i^{(N)} \right\} + \overline{t}_i^{(N)} + \overline{u}_i^{(N)} - \overline{v}_i^{(N)}$$
$$+ \sum_{h \geq g} p_k \alpha_{h_k} \overline{\mathbf{w}}_k^H \overline{\Psi}_i^{(N)} \overline{\mathbf{w}}_k^{(N)}, \quad \forall i \in \mathcal{I}_{g_k}, \forall k \in \mathcal{K}, \forall g_k \in \mathcal{G}_k \tag{79}$$

$$1 - \sum_{k=1}^{K} \sum_{g_k=1}^{G_k} \overline{C}_{g_k} \geq p_c \overline{\mathbf{w}}_c^H \overline{\Psi}_{c,i}^{(N)} \overline{\mathbf{w}}_c + \sum_{k=1}^{K} p_k \overline{\mathbf{w}}_k^H \overline{\Psi}_{c,i}^{(N)} \overline{\mathbf{w}}_k + \overline{t}_{c,i}^{(N)} - 2\mathcal{R}\left\{ \sqrt{p} \overline{f}_{c,i}^{(N)} \right\}$$
$$+ \overline{u}_{c,i}^{(N)} - \overline{v}_{c,i}^{(N)}, \quad \forall i \in \mathcal{I} \tag{80}$$

$$(24d), (24e) \tag{81}$$

and

$$\overline{\mathcal{S}}_3^{[l]} : \underset{\overline{\xi}_c, \mathbf{p}, \alpha_{\mathbf{k}}}{\text{argmin}} \quad \overline{\xi}_c + \sum_{k=1}^{K} \sum_{g=1}^{G} \left\{ \max_{i \in \mathcal{I}_{g_k}} \overline{\xi}_i \right\} \tag{82}$$

$$\text{s.t.}$$

$$p_c \overline{\mathbf{w}}_c^H \overline{\Psi}_{c,i}^{(N)} \overline{\mathbf{w}}_c + \sum_{k=1}^{K} p_k \overline{\mathbf{w}}_k^H \overline{\Psi}_{c,i}^{(N)} \overline{\mathbf{w}}_k + \overline{t}_{c,i}^{(N)} - 2\mathcal{R}\left\{ \sqrt{p} \overline{f}_{c,i}^{(N)} \right\} + \overline{u}_{c,i}^{(N)} - \overline{v}_{c,i}^{(N)} \leq \overline{\xi}_c, \forall i \in \mathcal{I} \tag{83}$$

$$(24d), (24e) \tag{84}$$

where $\overline{\xi}_i$ is

$$\overline{\xi}_i = \sum_{j=1, j \neq k}^{K} p_j \overline{\mathbf{w}}_j^H \overline{\Psi}_i^{(N)} \overline{\mathbf{w}}_j^{(N)} + \overline{t}_i^{(N)} - 2\mathcal{R}\left\{ \sqrt{\alpha_{g_k} p_k} \overline{f}_i^{(N)} \right\} + p_k \sum_{h_k \geq g_k} \alpha_{h_k} \overline{\mathbf{w}}_k^H \overline{\Psi}_i^{(N)} \overline{\mathbf{w}}_k^{(N)}$$
$$+ \overline{u}_i^{(N)} - \overline{v}_i^{(N)}, \forall i \in \mathcal{I}_{g_k}, \forall k \in \mathcal{K}, \forall g_k \in \mathcal{G}_k$$

As the iteration procedure continusers, the objective function in $\mathcal{P}_3$ or $\mathcal{S}_3$ grows until convergence. The proposed alternating optimization approach alternately optimizes the variables of the corresponding WMMSE problem $\overline{\mathcal{P}}_3$ and $\overline{\mathcal{S}}_3$. The proposed algorithm is guaranteed to converge as the objective function is bounded above for the specified power limitations.

5. Illustrative results and discussions

In this section, we evaluate the performance of the proposed RS-based MIMO-NOMA scheme through numerical simulations and validate the effectiveness of the power allocation algorithm. Specifically, we investigate the achievable MMF rate and sum-rate in different scenarios by varying the SNR, the number of users per group per cluster (M) and the degree of CSIT uncertainty, η. We compare the performance of the proposed RS-based MIMO-NOMA with conventional MIMO-NOMA.

In the Conv-based MIMO-NOMA system, instead of the RS, the conventional linear precoding such as RZF is applied to cancel interbeam interference between clusters of users, and NOMA is applied to provide service for more than one group of users.

5.1 Simulation setup

To carry out our analysis, we consider a single-cell cellular network with a radius of 500 m, where the base station is located at the center. It is equipped with an array of $N_t = 64$ antennas and forms $K = 12$ clusters. Each cluster has two groups of users, $G_k = 2, \quad k \in \mathcal{K}$, and all groups have the same cardinality, M. Users are randomly and uniformly distributed throughout the cell, excluding an inner circle of radius 50 meters.

The large-scale fading coefficient for user i is expressed as $\beta_i = \frac{\overline{d}}{x_i^\nu}$, where x_i indicates the distance between the i-th user and the base station. Here, the constant $\overline{d} = 10^{-5}$ serves the role of regulating the channel attenuation at a distance of 50 m, and ν symbolizes the path loss exponent, which is assumed to be 3.76 for this study. The large-scale fading (β_i) in this context follows a log-normal distribution with a standard deviation of 8 dB.

Furthermore, in this chapter, we consider the noise variance to be set at 1. As a result, the SNR is defined by the peak power, denoted as p_{max}.

5.2 MMF rate analysis results

In this section, we aim to compare the performance of our proposed RS-based MIMO-NOMA scheme with that of the Conv-based MIMO-NOMA technique, specifically focusing on the MMF rate. The primary objective of this comparison is to maximize the minimum achievable rate by optimizing power allocation. We evaluate the MMF rate performance under varying SNR conditions, while simultaneously adjusting the number of users per group and the degrees of CSIT uncertainty.

Figure 2 presents a comparison of the MMF rate for the proposed RS-based MIMO-NOMA and conventional MIMO-NOMA systems as a function of SNR. **Figure 2a** compares the MMF rate for different numbers of users per group, considering cases with two and three users per group. The results reveal that the gain of the RS-based MIMO-NOMA over the conventional MIMO-NOMA systems expands as the number of users per group increases. This gain increases from 1.07 to 1.32 when the number of users per group increases from $M = 2$ to $M = 3$. Consequently, the RS-based MIMO-NOMA proves to be a more robust solution in overloaded regimes.

Figure 2b demonstrates the impact of CSIT uncertainty on the MMF rate performance. The results indicate that the MMF rate performance of the conventional MIMO-NOMA system degrades more significantly when CSIT transitions from perfect to imperfect with $\eta = 0.5$. Therefore, the RS-based MIMO-NOMA system is more robust to CSIT uncertainty fluctuations. The gap between MMF rates of the RS-based
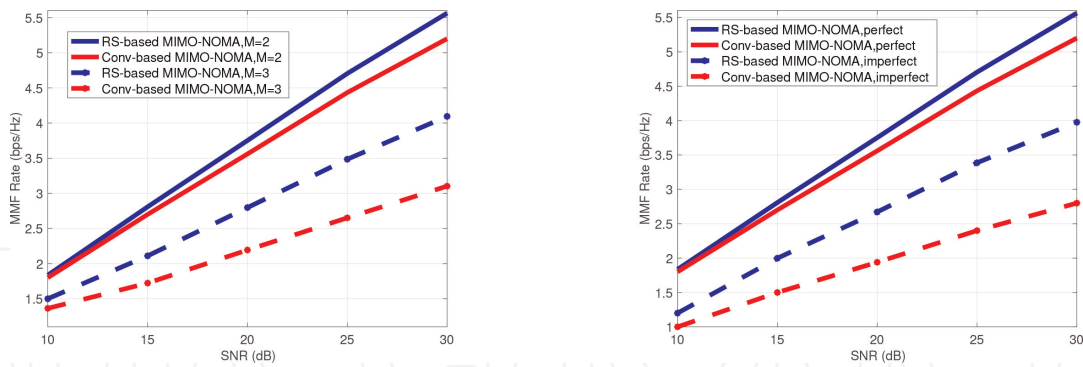
Figure 2.
Comparison of achievable MMF rate performance for RS-based MIMO-NOMA and Conv-based MIMO-NOMA.

MIMO-NOMA when CSIT changes from perfect to imperfect with $\eta = 0.5$ is 1.5880 bps/Hz. However, this gap is much higher in the conventional MIMO-NOMA system, amounting to 2.4 bps/Hz.

5.3 Sum-rate analysis results

This section investigate the performance of the proposed RS-based MIMO-NOMA scheme in terms of sum-rate. The objective is to maximize the overall system throughput by optimizing power allocation. The sum-rate performance is investigated under varying numbers of users per group and degrees of CSIT uncertainty.

Figure 3 illustrates the sum-rate versus SNR comparison of RS-based MIMO-NOMA and Conv-based MIMO-NOMA. **Figure 3a** compares the sum-rate for different numbers of users per group, considering cases with two and three users per group. The results show that increasing the number of users per group decreases the sum-rate in both cases. Moreover, the gain of the RS-based MIMO-NOMA over the Conv-MIMO-NOMA is not considerably high, even when the number of users increases from $M = 2$ to $M = 3$.

Figure 3b explores the impact of CSIT uncertainty on the sum-rate performance. The results indicate that the sum-rate performance of the conventional MIMO-NOMA system experiences a more significant decline when CSIT transitions from perfect to imperfect with $\eta = 0.5$. Therefore, the RS-based MIMO-NOMA system exhibits greater robustness against CSIT uncertainty fluctuations. The gap between sum-rates of the RS-based MIMO-NOMA when CSIT changes from perfect to imperfect with
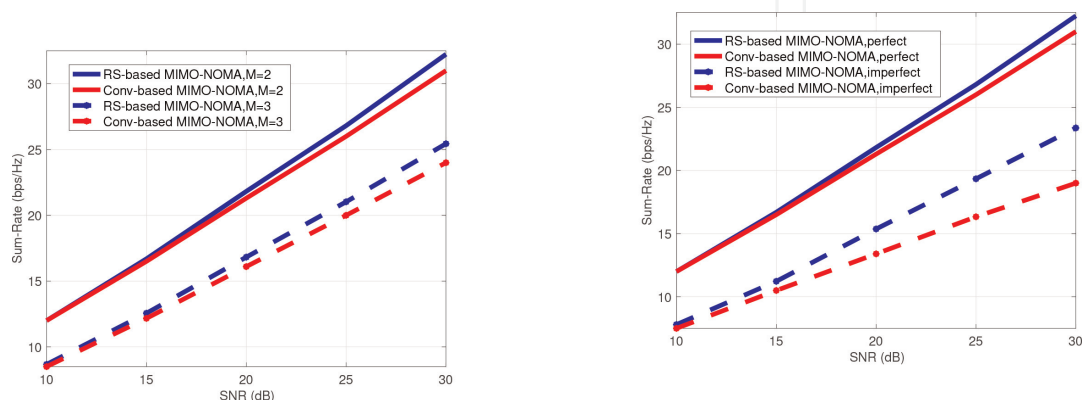

Figure 3.
Achievable sum-rate performance comparison for RS-based MIMO-NOMA and Conv-based MIMO-NOMA.

$\eta = 0.5$ is around 9 bps/Hz. In contrast, this gap is substantially larger in the conventional MIMO-NOMA system, amounting to 12 bps/Hz, roughly a 33% decline.

Figures 2 and **3** illustrate that the proposed RS-based MIMO-NOMA scheme effectively exploits the rate-splitting technique to enhance its performance in overloaded scenarios and under imperfect CSIT, particularly when compared to the conventional MIMO-NOMA system. This improvement can be attributed to the RS-based MIMO-NOMA's ability to mitigate interbeam interference and efficiently allocate power among users, thus providing superior service to a larger number of users within each group even under imperfect CSIT. Overall, these results emphasize the advantages of adopting the RS-based MIMO-NOMA framework in practical network deployments, particularly in high-density and overloaded scenarios and under imperfect CSIT.

6. Conclusion

In this chapter, we have presented a novel scheme that combines the RS technique in MIMO systems with NOMA scheme for wireless communication systems, aiming to improve performance and capacity under imperfect CSIT and overloaded regime. The proposed scheme has considered a general and realistic scenario with both unicast and multicast users, focusing on increasing system throughput and optimizing precoding vectors for enhanced performance.

Furthermore, we have introduced a technique that transforms a non-convex optimization problem into a convex problem. By employing the WMMSE-rate relationship and an AO algorithm, the proposed technique successfully tackles the non-convex problem, allowing for the maximization of both the minimum rate and sum-rate of the system, particularly concentrating on the rate of the weakest user in each group under imperfect CSIT.

The comprehensive analysis provided in this chapter covers both tutorial background and novel ideas, offering valuable insights into the design and performance of future MIMO-NOMA systems that employ RS techniqusers. The findings demonstrate the potential of the proposed RS-based MIMO-NOMA scheme in addressing the challenges posed by imperfect CSIT and overloaded regimes in realistic scenarios with unicast and multicast users.

Author details

Sareh Majidi Ivari*, Mohammad Reza Soleymani and Yousef R. Shayan
Concordia University, Montreal, Canada

*Address all correspondence to: sarehh.majidi@gmail.com

IntechOpen

References

[1] Cisco. Cisco Annual Internet Report (2018–2023) White Paper. Cisco; 2021. Available from: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] Ding Z, Adachi F, Poor HV. The application of MIMO to non-orthogonal multiple access. IEEE Transactions on Wireless Communications. 2016;**15**(1): 537-552. DOI: 10.1109/TWC.2015.2475746

[3] Huang Y, Zhang C, Wang J, Jing Y, Yang L, You X. Signal processing for MIMO-NOMA: Present and future challenges. IEEE Wireless Communications. 2018;**25**(2):32-38. DOI: 10.1109/MWC.2018.1700108

[4] Ali S, Hossain E, Kim DI. Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation. IEEE Access. 2017;**5**: 565-577. DOI: 10.1109/ACCESS.2016.2646183

[5] Chen X, Zhang Z, Zhong C, Jia R, Ng DWK. Fully non-orthogonal communication for massive access. IEEE Transactions on Communications. 2018; **66**(4):1717-1731. DOI: 10.1109/TCOMM.2017.2779150

[6] Senel K, Cheng HV, Björnson E, Larsson EG. What role can NOMA play in massive MIMO? IEEE Journal of Selected Topics in Signal Processing. 2019;**13**(3):597-611. DOI: 10.1109/JSTSP.2019.2899252

[7] Nguyen N, Zeng M, Dobre OA, Poor HV. Securing massive MIMO-NOMA networks with ZF beamforming and artificial noise. In: 2019 IEEE Global Communications Conference (GLOBECOM). Waikoloa, HI, USA: IEEE GLOBCOM; 2019. pp. 1-6

[8] Min K, Kim T, Jung M. Performance analysis of multiuser massive MIMO with multi-antenna users: Asymptotic data rate and its application. ICT Express. 2023. DOI: 10.1016/j.icte.2023.01.003

[9] Björnson E, Sanguinetti L, Debbah M. Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination. International Journal of Wireless Information Networks. 2014;**21**(2):133-149

[10] Mao Y, Clerckx B, Li VO. Rate-splitting multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA. Journal on Wireless Communications and Networking. 2018; **2018**:133. DOI: 10.1186/s13638-018-1104-7

[11] Love D, Heath R, Lau V, Gesbert D, Rao B, Andrews M. An overview of limited feedback in wireless communication systems. IEEE Journal on Selected Areas in Communications. 2008;**26**(8):1341-1365

[12] Turan N, Fesl B, Koller M, Joham M, Utschick W. A versatile low-complexity feedback scheme for FDD systems via generative modeling. arXiv. 2023

[13] Sadeghi M, Björnson E, Larsson EG, Yuen C, Marzetta T. Joint unicast and multi-group multicast transmission in massive MIMO systems. IEEE Transactions on Wireless Communications. 2018;**17**(10): 6375-6388. DOI: 10.1109/TWC.2018.2854554

[14] Joudeh H, Clerckx B. RS for MISO wireless networks: A promising PHY-

layer strategy for LTE evolution. IEEE Communications Magazine. 2016;**54**(5): 98-105

[15] Mao Y, Clerckx B. RS multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA. EURASIP Journal on Wireless Communications and Networking. 2018; **2018**(1):1-21

[16] Joudeh H, Clerckx B. Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A RS approach. IEEE Transactions on Communications. 2016; **64**(11):4847-4861

[17] Clerckx B, Kim I, Kim J, Zhang R, Poor HV. Is NOMA efficient in multi-antenna networks? A critical look at next generation multiple access Techniqusers. IEEE Communications Magazine. 2020; **58**(2):64-71

[18] Kim J, Kim I-M. Achievable rates of spatially coupled multiple-access channels with RS, superposition coding, and successive cancellation decoding. IEEE Transactions on Wireless Communications. 2018;**17**(10):6761-6775

[19] Lee B, Shin W. Max-min fairness precoder design for RS multiple access: Impact of imperfect channel knowledge. IEEE Transactions on Vehicular Technology. 2023;**72**(1):1355-1359. DOI: 10.1109/TVT.2022.3206808

[20] Joudeh H, Clerckx B. RS for max-min fair multigroup multicast beamforming in overloaded systems. IEEE Transactions on Wireless Communications. 2017;**16**(11): 7276-7289. DOI: 10.1109/TWC. 2017.2744629

[21] Yin L, Clerckx B. RS multiple access for multigroup multicast and multibeam

satellite systems. IEEE Transactions on Communications. 2021;**69**(2):976-990. DOI: 10.1109/TCOMM.2020.3037596

[22] Yin L, Dizdar O, Clerckx B. RS multiple access for multigroup multicast cellular and satellite communications: PHY layer design and link-level simulations. In: 2021 IEEE International Conference on Communications Workshops (ICC Workshops); 2021 Jun 14–18; Montreal, Canada. IEEE; 2021. pp. 1-6. DOI: 10.1109/ICCWorkshops 50388.2021.9473795

[23] Zeng J, Lv T, Ni W, Liu RP, Beaulieu NC, Guo YJ. Ensuring max–min fairness of UL SIMO-NOMA: A RS approach. IEEE Transactions on Vehicular Technology. 2019;**68**(11): 11080-11093. DOI: 10.1109/TVT.2019. 2943511

[24] Joudeh H, Clerckx B. Robust transmission in downlink multiuser MISO systems: A rate-splitting approach. IEEE Transactions on Signal Processing. 2016;**64**(23):6227-6242

[25] Joudeh H, Clerckx B. Sum rate maximization for MU-MISO with partial CSIT using joint multicasting and broadcasting. In: Proc. IEEE Int. Conf. Commun. Londan, UK: IEEE International Conference on Communications (ICC); 2015. pp. 4733-4738. DOI: 10.1109/ ICC.2015.7249071

[26] Hao C, Wu Y, Clerckx B. Rate analysis of two-receiver MISO Broadcast Channel with finite rate feedback: A rate-splitting approach. IEEE Transactions on Communications. 2015; **63**(9):3232-3246. DOI: 10.1109/ TCOMM.2015.2453270

[27] Ivari SM, Caus M, Vazquez MA, Soleymani MR, Shayan YR, Perez-Neira AI. Precoding and scheduling in multibeam multicast NOMA based

satellite communication systems. In: 2021 IEEE Int. Conf. Commun. Workshops (ICC Workshops). Montreal, Canada: IEEE International Conference on Communications (ICC); 2021. pp. 1-6. DOI: 10.1109/ ICCWorkshops50388.2021.9473484

[28] Xiang Z, Tao M, Wang X. Massive MIMO multicasting in noncooperative cellular networks. IEEE Journal on Selected Areas in Communications. 2014;**32**(6):1180-1193. DOI: 10.1109/ JSAC.2014.2328144