Received 15 April 2021; revised 17 May 2021; accepted 20 May 2021. Date of current version 11 June 2021.

Digital Object Identifier 10.1109/OJCOMS.2021.3084799

Is NOMA Efficient in Multi-Antenna Networks? A Critical Look at Next Generation Multiple Access Techniques

BRUNO CLERCKX [1] (Senior Member, IEEE), YIJIE MAO [1] (Member, IEEE), ROBERT SCHOBER [2] (Fellow, IEEE), EDUARD A. JORSWIECK [3] (Fellow, IEEE), DAVID J. LOVE [4] (Fellow, IEEE), JINHONG YUAN [5] (Fellow, IEEE), LAJOS HANZO [6] (Fellow, IEEE), GEOFFREY YE LI [1] (Fellow, IEEE), ERIK G. LARSSON [7] (Fellow, IEEE), AND GIUSEPPE CAIRE [8] (Fellow, IEEE)

[1]Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K.

[2]Institute for Digital Communications, Friedrich-Alexander University Erlangen-Nürnberg, 91054 Erlangen, Germany

[3]Institute for Communications Technology, Technische Universität Braunschweig, 38106 Brunswick, Germany

[4]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

[5]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia

[6]School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

[7]Department of Electrical Engineering, Linköping University, 581 83 Linköping, Sweden

[8]Communications and Information Theory Group, Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany

CORRESPONDING AUTHOR: B. CLERCKX (e-mail: b.clerckx@imperial.ac.uk)

This work was supported in part by EPSRC of the U.K. under Grant EP/N015312/1 and Grant EP/R511547/1; in part by Australian Research Council Discovery Projects under Grant d DP190101363; in part by Linkage Project under Grant LP170101196; and in part by European Research Council's Advanced Fellow Grant QuantCom under Grant 789028.

ABSTRACT In the past few years, a large body of literature has been created on downlink Non-Orthogonal Multiple Access (NOMA), employing superposition coding and Successive Interference Cancellation (SIC), in multi-antenna wireless networks. Furthermore, the benefits of NOMA over Orthogonal Multiple Access (OMA) have been highlighted. In this paper, we take a critical and fresh look at the downlink Next Generation Multiple Access (NGMA) literature. Instead of contrasting NOMA with OMA, we contrast NOMA with two other multiple access baselines. The first is conventional Multi-User Linear Precoding (MU–LP), as used in Space-Division Multiple Access (SDMA) and multi-user Multiple-Input Multiple-Output (MIMO) in 4G and 5G. The second, called Rate-Splitting Multiple Access (RSMA), is based on multi-antenna Rate-Splitting (RS). It is also a non-orthogonal transmission strategy relying on SIC developed in the past few years in parallel and independently from NOMA. We show that there is some confusion about the benefits of NOMA, and we dispel the associated *misconceptions*. *First*, we highlight why NOMA is inefficient in multi-antenna settings based on basic multiplexing gain analysis. We stress that the issue lies in how the NOMA literature, originally developed for single-antenna setups, has been hastily applied to multi-antenna setups, resulting in a misuse of spatial dimensions and therefore loss in multiplexing gains and rate. *Second*, we show that NOMA incurs a severe multiplexing gain loss despite an increased receiver complexity due to an inefficient use of SIC receivers. *Third*, we emphasize that much of the merits of NOMA are due to the constant comparison to OMA instead of comparing it to MU–LP and RS baselines. We then expose the pivotal *design constraint* that multi-antenna NOMA requires one user to fully decode the messages of the other users. This design constraint is responsible for the multiplexing gain erosion, rate and spectral efficiency loss, ineffectiveness to serve a large number of users, and inefficient use of SIC receivers in multi-antenna settings. Our analysis and simulation results confirm that NOMA should not be applied blindly to multi-antenna settings, highlight the scenarios where MU–LP outperforms NOMA and vice versa, and demonstrate the inefficiency, performance loss,

and complexity disadvantages of NOMA compared to RSMA. The first takeaway message is that, while NOMA is suited for single-antenna settings (as originally intended), it is not efficient in most multi-antenna deployments. The second takeaway message is that another non-orthogonal transmission framework, based on RSMA, exists which fully exploits the multiplexing gain and the benefits of SIC to boost the rate and the number of users to serve in multi-antenna settings and outperforms both NOMA and MU–LP. Indeed, RSMA achieves higher multiplexing gains and rates, serves a larger number of users, is more robust to user deployments, network loads and inaccurate channel state information and has a lower receiver complexity than NOMA. Consequently, RSMA is a promising technology for NGMA and future networks such as 6G and beyond.

INDEX TERMS Multiple antennas, downlink, non-orthogonal multiple access, superposition coding, rate-splitting multiple access, broadcast channel, multiuser linear precoding, multiuser multiple-input multiple-output, space division multiple access, next generation multiple access.

I. INTRODUCTION

MULTIPLE access is a crucial part of any communication system and refers to techniques that make use of the resources (e.g., time, frequency, power, antenna, code) to serve multiple users, ideally in the most efficient way. In contrast to Orthogonal Multiple Access (OMA) that assigns users to orthogonal dimensions (e.g., Time-Division Multiple Access - TDMA, Frequency-Division Multiple Access - FDMA), (power-domain) Non-Orthogonal Multiple Access (NOMA)[1] superposes users in the same time-frequency resource and distinguishes them in the power domain [1]–[5]. By doing so, NOMA has been promoted as a solution for 5G and beyond to deal with the vast throughput, access (serving a large number of users), and Quality-of-Service (QoS) requirements that are projected to grow exponentially for the foreseeable future.

In the downlink, *NOMA refers to communication schemes where at least one user is forced to fully decode the message(s) of other co-scheduled user(s)*. This operation is commonly performed through the use of transmit-side Superposition Coding (SC) and receiver-side Successive Interference Cancellation (SIC) in downlink multi-user communications. Such techniques have been studied for years before being branded with the NOMA terminology. NOMA has indeed been known in the information theory and wireless communications literature for several decades, under the terminology of superposition coding with successive interference cancellation (denoted in short as SC–SIC), as the strategy that achieves (and has been used in achievability proofs for) the capacity region of the Single-Input Single-Output (SISO) (Gaussian) Broadcast Channel (BC) [6]. The superiority of NOMA over OMA was shown in the seminal paper by Cover in 1972. It is indeed well known that the capacity region of the SISO BC (achieved by NOMA) is larger than the rate region achieved by OMA (i.e., contains the achievable rate region of OMA as a subset) [6], [8], [9]. The use of SIC receivers is a major difference between NOMA and OMA, although it should be mentioned that SIC has also been studied for a long time in the 3G and 4G research phases in the context of interference cancellation and receiver designs [10]. Unfortunately, despite the existence of well-established textbooks on the topic in the past few decades [7]–[9], the recent literature on NOMA has been the subject of some confusion, misunderstandings, and misconceptions [11].

In today's wireless networks, access points commonly employ more than one antenna, which opens the door to the spatial domain and multi-antenna processing. The key building block of the downlink of multi-antenna networks is the multi-antenna (Gaussian) BC. Contrary to the SISO BC that is degraded and where users can be ordered based on their channel strengths, the multi-antenna BC is nondegraded and users cannot be ordered based on their channel strengths [8], [12]. This is why SC–SIC/NOMA is not capacity-achieving in this case,[2] and Dirty Paper Coding (DPC) is the only known strategy that achieves the capacity region of the multi-antenna (Gaussian) BC with perfect Channel State Information at the Transmitter (CSIT) [12]. Due to the high computational burden of DPC, linear precoding is often considered the most attractive alternative to simplify the transmitter design [13]–[17]. Interestingly, in a multi-antenna BC, Multi-User Linear Precoding (MU–LP) relying on treating the residual multi-user interference as noise, although suboptimal, is often very useful since the interference can be significantly reduced by spatial precoding. This is the reason why it has received significant attention in the past twenty years and it is the basic principle behind numerous 4G and 5G techniques such as Space-Division Multiple Access (SDMA) and multi-user (potentially massive) Multiple-Input Multiple-Output (MIMO) [17].

In view of the benefits of NOMA over OMA and multi-antenna over single-antenna, numerous attempts have been made in recent years to combine multi-antenna and

1. Although there is a broad range of NOMA schemes in the power and code domains, in this treatise, we focus only on power-domain NOMA and simply use NOMA to represent power-domain NOMA. Readers are referred to [2] for an overview of code-domain NOMA.

2. NOMA is capacity achieving if the user channels are aligned but this is not a realistic scenario in practice. Thus, throughout the paper we assume user channels are not aligned. This matter is further discussed in Section VIII-E.

NOMA schemes [1]–[5], [18]–[44] (and references therein). Although there are a few contributions comparing NOMA with MU–LP schemes, such as Zero-Forcing Beamforming (ZFBF) or DPC [29], [41]–[44], much emphasis is put on comparing (single/multi-antenna) NOMA and OMA, and showing that NOMA outperforms OMA. But there is a lack of emphasis on contrasting multi-antenna NOMA to other multi-user multi-antenna baselines developed for the multi-antenna BC, such as MU–LP (or other forms of multi-user MIMO techniques) and Rate-Splitting Multiple Access (RSMA) [45]. RSMA is a form of (power-domain) non-orthogonal transmission strategy based on multi-antenna Rate-Splitting (RS). RS designed for the multi-antenna BC also relies on SIC and has been developed in parallel and independently from NOMA [45]–[51]. Such a comparison is essential to assess the benefits and the efficiency of NOMA, since all these communication strategies can be viewed as different achievable schemes for the multi-antenna BC and all aim in their own way for the same objective, namely to meet the throughput, reliability, QoS, and connectivity requirements of beyond-5G multi-antenna wireless networks.

In this paper, we take a critical look at multi-antenna NOMA and Next Generation Multiple Access (NGMA) techniques for the downlink of communication systems and ask the important questions "*Is multi-antenna NOMA an efficient strategy?*" and "*What are the important design principles for NGMA techniques?*" To answer those questions, we go beyond the conventional NOMA vs. OMA comparison, and contrast multi-antenna NOMA with MU–LP and RS strategies. This allows us to highlight some misconceptions and shortcomings of multi-antenna NOMA. Explicitly, we show that in most scenarios the short answer to the first question is no, and demonstrate based on first principles and numerical performance evaluations why this is the case. Our discussions and results unveil the scenarios where MU–LP outperforms NOMA and vice versa, and demonstrate that multi-antenna NOMA is inefficient compared to RS. By contrasting multi-antenna NOMA to MU–LP and RS, we show that there is some confusion about multi-antenna NOMA and its merits and expose major misconceptions. Our results and discussions also reveal new insights and perspectives for the design of NGMA techniques.

The contributions of this paper are summarized as follows.

First, we analytically derive both the sum multiplexing gain as well as the max-min fair multiplexing gain of multi-antenna NOMA and compare them to those of MU–LP and RS. The scenarios considered are very general and include multi-antenna transmitter with single/multi-antenna receivers, perfect and imperfect CSIT, and underloaded and overloaded regimes. On the one hand, multi-antenna NOMA can achieve gains, but can also incur losses compared to MU–LP. On the other hand, multi-antenna NOMA *always* leads to a waste of multiplexing gain compared to RS. This multiplexing gain loss translates in a spectral efficiency loss and in an inability to serve a large number of users. The multiplexing gain analysis provides a firm theoretical ground to infer that multi-antenna NOMA is not as efficient as RS in exploiting the spatial dimensions and the available CSIT, and in serving a large number of users. This analysis is instrumental to identify the scenarios where the multiplexing gain gaps among NOMA, MU–LP, and RS are the smallest/largest, therefore highlighting deployments that are suitable/unsuitable for the different multiple access strategies.

Second, we show that multi-antenna NOMA leads to a high receiver complexity due to the inefficient use of SIC. For instance, we show that the higher the number of SIC operations (and therefore the higher the receiver complexity) in multi-antenna NOMA, the lower the sum multiplexing gain (and therefore the lower the sum-rate at high Signal-to-Noise Ratio SNR). Comparison with MU–LP and RS show that higher multiplexing gains can be achieved and a larger number of users can be served at a lower receiver complexity and a reduced number of SIC operations. Indeed, our results show that NOMA requires $K-1$ SIC layers to support K users with M transmit antennas, while RS can support $M-1+K$ users with only one SIC layer.

Third, we show that most of the *misconceptions behind NOMA are due to the prevalent comparison to OMA instead of comparing to MU–LP and RS*. We show and explain that the misconceptions, the multiplexing gain reduction, and the inefficient use of SIC receivers in both underloaded and overloaded multi-antenna settings relying on both perfect and imperfect CSIT originate from a *limitation of the multi-antenna NOMA design philosophy*, namely that *one user is forced to fully decode the messages of the other users*. Hence, while forcing a user to fully decode the messages of the other users is an efficient approach in single-antenna degraded BC, it may not be an efficient approach in multi-antenna networks.

Fourth, we stress that an efficient design of non-orthogonal transmission and multiple access/NGMA strategies ensures that the use of SIC never leads to a performance loss but rather leads to a performance gain over MU–LP. We show that such non-orthogonal solutions based on RS exist and truly benefit from the multi-antenna multiplexing gain and from the use of SIC receivers in both underloaded and overloaded regimes relying on perfect and imperfect CSIT. In fact, multi-antenna RS completely resolves the design limitations of multi-antenna NOMA. Consequently, RS with only one SIC layer can achieve higher spectral efficiency and support a larger number of users than NOMA with multiple SIC layers.

Fifth, we depart from the multiplexing gain analysis and design the transmit precoders to maximize the sum-rate and max-min rate for multi-antenna NOMA, followed by numerically comparing the sum-rate and the max-min fair rate of NOMA to those of MU–LP and RS. We show that the multiplexing gain analysis is accurate and instrumental to predict the rate performance of the multiple access strategies considered.

TABLE 1. Overview of the paper.

Section II. Two-User MISO NOMA with Perfect CSIT: The Basic Building Block	
II-A. System Model	II-B. Definition of Multiplexing Gain
II-C. Discussions	
Section III. K-User MISO NOMA with Perfect CSIT	
III-A. MISO NOMA System Model	III-B. Multiplexing Gains
Section IV. K-User MISO NOMA with Imperfect CSIT	
IV-A. CSIT Error Model	IV-B. Multiplexing Gains
Section V. MIMO NOMA	
Section VI. Baseline Scheme I: Conventional Multi-user Linear Precoding	
VI-A. MU–LP System Model	VI-B. Multiplexing Gains with Perfect CSIT
VI-C. Multiplexing Gains with Imperfect CSIT	
Section VII. Baseline Scheme II: Rate-Splitting	
VII-A. Rate-Splitting System Model	VII-B. Multiplexing Gains with Perfect CSIT
VII-C. Multiplexing Gains with Imperfect CSIT	
Section VIII. Shortcomings and Misconceptions of Multi-Antenna NOMA	
VIII-A. NOMA vs. Baseline I (MU–LP)	VIII-B. NOMA vs. Baseline II (RS)
VIII-C. Misconceptions of Multi-Antenna NOMA	VIII-D. Illustration of the Misconceptions with an Example
VIII-E. Shortcomings of Multi-Antenna NOMA	
Section IX. Numerical Results	
IX-A. Perfect CSIT	IX-B. Imperfect CSIT
IX-C. Discussions	
Section X. Conclusions and Future Works	

Sixth, our numerical simulations confirm the inefficiency of multi-antenna NOMA in general settings. Multi-antenna NOMA is shown to lead to performance gains over MU–LP in some settings but also to losses in other settings despite the use of SIC receivers and a higher receiver complexity. Our results also highlight the significant benefits, performance-wise and receiver complexity-wise, of RSMA and multi-antenna RS over multi-antenna NOMA. It is indeed possible to achieve a significantly better performance than MU–LP and NOMA with just one layer of SIC by adopting RS so as to partially decode messages of other users (instead of fully decoding them as in NOMA).

Organization: The remainder of this paper is organized as follows. Section II introduces two-user Multiple-Input Single-Output (MISO) NOMA (with single-antenna receivers) as a basic building block (and toy example) for our subsequent studies, compares to MU–LP, and raises some questions about the efficiency of NOMA. Section III studies the multiplexing gain of K-user MISO NOMA with perfect CSIT. Sections IV and V extend the discussion to imperfect CSIT and MIMO NOMA, respectively. Section VI and Section VII study the multiplexing gains of the baseline schemes considered, namely MU–LP and RS, respectively. Section VIII compares the multiplexing gains of all multiple access schemes considered and exposes the misconceptions and shortcomings of multi-antenna NOMA. Section IX provides simulation results. Section X concludes this paper, discusses future research and pathways to 6G standardization. An overview of the paper is shown in Table 1.

Notation: $|\cdot|$ refers to the absolute value of a scalar or to the cardinality of a set depending on the context. $\|\cdot\|$ refers to the l_2-norm of a vector. $\max\{a_1, \ldots, a_n\}$ refers to the maximum value between a_1 to a_n. $\mathbf{a}^H$ denotes the Hermitian transpose of vector $\mathbf{a}$. $\mathrm{Tr}(\mathbf{Q})$ refers to the trace of matrix $\mathbf{Q}$. $\mathbf{I}$ is the identity matrix. $P \nearrow$ means as P grows large. $\mathcal{CN}(0, \sigma^2)$ denotes the circularly symmetric complex Gaussian distribution with zero mean and variance σ^2. $\sim$ stands for "distributed as". $O(\cdot)$ refers to the big O notation. $\mathbb{E}\{\cdot\}$ denotes statistical expectation. $A \cap B$ and $A \cup B$ refer to the intersection (A and B have to be satisfied) and the union (A or B to be satisfied) of two sets/events A and B, respectively.

II. TWO-USER MISO NOMA WITH PERFECT CSIT: THE BASIC BUILDING BLOCK

We commence by studying two-user MISO NOMA and show that, by comparing NOMA to MU–LP instead of to OMA, the potential merits of NOMA are less obvious. Limited to two single-antenna users with perfect CSIT, this system model illustrates the simplest though fundamental building block of multi-antenna NOMA.

A. SYSTEM MODEL

We consider a downlink single-cell multi-user multi-antenna scenario with $K = 2$ users, also known as two-user MISO

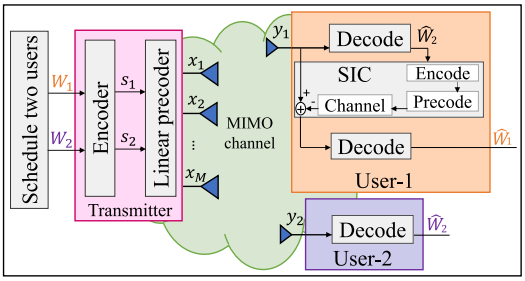


FIGURE 1. Two-user system architecture with NOMA (decoding order: user-2→user-1).

BC, consisting of one transmitter with $M \geq 2$ antennas[3] communicating with two single-antenna users. The transmitter aims to transmit simultaneously two messages W_1 and W_2 intended for user-1 and user-2, respectively.

The transmitter adopts the so-called multi-antenna NOMA or MISO NOMA strategy, illustrated in Fig. 1, that encodes one of the two messages using a codebook shared by both users[4] so that it can be decoded and cancelled from the received signal at the co-scheduled user (following the same principle as superposition coding for the degraded BC). Consider W_2 is encoded into s_2 using the shared codebook and W_1 is encoded into s_1. The two streams are then linearly precoded by $M \times 1$ precoders[5] $\mathbf{p}_1$ and $\mathbf{p}_2$ and superposed at the transmitter so that the transmit signal is given by

$$\mathbf{x} = \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2. \tag{1}$$

Defining $\mathbf{s} = [s_1, s_2]^T$ and assuming that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}$, the average transmit (sum) power constraint is written as $P_1 + P_2 \leq P$ where $P_k = \|\mathbf{p}_k\|^2$ with $k = 1, 2$.

The channel vector for user k is denoted by $\mathbf{h}_k$, and the received signal at user k can be written as $y_k = \mathbf{h}_k^H \mathbf{x} + n_k$, $k = 1, 2$, where $n_k \sim \mathcal{CN}(0, 1)$ is Additive White Gaussian Noise (AWGN). We assume perfect CSIT and perfect Channel State Information at the Receivers (CSIR).

At both users, stream s_2 is decoded first into[6] $\widehat{W}_2$ by treating the interference from s_1 as noise. Using SIC at user-1, $\widehat{W}_2$ is re-encoded, precoded, and subtracted from the received signal, such that user-1 can decode its stream s_1 into $\widehat{W}_1$. Assuming proper Gaussian signaling and perfect SIC,[7] the achievable rates of the two streams with MISO NOMA are given by[8]

$$R_1^{(\mathrm{N})} = \log_2\left(1 + \left|\mathbf{h}_1^H \mathbf{p}_1\right|^2\right), \tag{2}$$

$$R_2^{(\mathrm{N})} = \min\left(\log_2(1+A), \log_2(1+B)\right), \tag{3}$$

where

$$A = \frac{\left|\mathbf{h}_1^H \mathbf{p}_2\right|^2}{1 + \left|\mathbf{h}_1^H \mathbf{p}_1\right|^2}, B = \frac{\left|\mathbf{h}_2^H \mathbf{p}_2\right|^2}{1 + \left|\mathbf{h}_2^H \mathbf{p}_1\right|^2}. \tag{4}$$

In (3), $\log_2(1+A)$ is the rate supportable by the channel of user-1 when user-1 decodes s_2 and treats its own stream s_1 as noise. Similarly, $\log_2(1+B)$ is the rate supportable by the channel of user-2 when user-2 decodes its own stream s_2 while treating stream s_1 of user-1 as noise. The min in (3) is due to the fact that s_2, though carrying message W_2 intended to user-2, is decoded by both users and is therefore transmitted at a rate decodable by both users.

The most common performance metric of a multi-user system is the sum-rate. In this two-user MISO NOMA system model, the sum-rate is defined as $R_{\mathrm{s}}^{(\mathrm{N})} = R_1^{(\mathrm{N})} + R_2^{(\mathrm{N})}$ and can be upper bounded[9] as

$$\begin{aligned} R_{\mathrm{s}}^{(\mathrm{N})} &\leq \log_2\left(1 + \frac{\left|\mathbf{h}_1^H \mathbf{p}_2\right|^2}{1 + \left|\mathbf{h}_1^H \mathbf{p}_1\right|^2}\right) + \log_2\left(1 + \left|\mathbf{h}_1^H \mathbf{p}_1\right|^2\right), \\ &= \log_2\left(1 + \left|\mathbf{h}_1^H \mathbf{p}_2\right|^2 + \left|\mathbf{h}_1^H \mathbf{p}_1\right|^2\right). \end{aligned} \tag{5}$$

It is important to note that (5) can be interpreted as the sum-rate of a two-user multiple access channel (MAC) with a single-antenna receiver. Indeed, user-1 acts as the receiver of a two-user MAC whose effective SISO channels for both links are given by $\mathbf{h}_1^H \mathbf{p}_2$ and $\mathbf{h}_1^H \mathbf{p}_1$, respectively. This observation will be revisited in the next few sections, and will be shown very helpful to explain the performance of multi-antenna NOMA.

A drawback of the sum-rate is that it does not capture rate fairness among the users. Another popular system performance metric is the Max-Min Fair (MMF) rate or symmetric rate defined as $R_{\mathrm{mmf}}^{(\mathrm{N})} = \min_{k=1,2} R_k^{(\mathrm{N})}$. The MMF metric provides uniformly good QoS since it aims for maximizing the minimum rate among all users.

Throughout the manuscript, we will focus on the sum-rate and the MMF rate as two very different metrics to assess the system performance. We choose these two metrics as they are commonly used in wireless networks, and in the NOMA literature in particular (see, e.g., [20], [24], [30], [32], [33] for the sum-rate and [34], [35], [39], [53], [54] for the MMF rate). They are representative for two very

3. Throughout the paper, we will assume fully digital processing with M antennas and M RF chains. This is standard in communication theoretic studies but also in real multi-antenna deployments, even for massive mimo with sub 6GHz deployments (e.g., Ericsson AIR 6468). For millimeter-wave deployments, it is plausible that future systems will be fully digital too eventually [52].

4. This is not an issue in modern systems since, for example, in an LTE/5G NR system, all codebooks are shared since all users use the same family of modulation and coding schemes (MCS) specified in the standard.

5. The precoders $\mathbf{p}_1$ and $\mathbf{p}_2$ can be any vectors that satisfy the power constraint, though the best choice of precoders would depend on the objective function.

6. Though not expressed explicitly, $\widehat{W}_2$ is receiver dependent since both receivers decode s_2 and the same estimate is not necessarily obtained at both receivers. Hence, more rigorously, we could have written $\widehat{W}_{2,k}$, $k = 1, 2$ to refer to the estimate at user-k. For simplicity of presentation, we have nevertheless opted to drop the index k.

7. Note there is no error in the SIC operation since the chosen rates are achievable under Gaussian signaling and infinite block length.

8. Superscript (N) stands for NOMA. Similarly we will user (M) for MU–LP, (R) for Rate-Splitting, and $\star$ for the information theoretic optimum.

9. This is an upper bound since when $A < B$, it is achieved with equality, and when $B < A$, $\log_2(1+B) < \log_2(1+A)$ and it is a strict upper bound.

different operational regimes, with the former focusing on high system throughput and the latter on user fairness.

In the sequel, we introduce some useful definitions and then make some observations based on this two-user system model.

B. DEFINITION OF MULTIPLEXING GAIN

Throughout the paper, we will often refer to the multiplexing gain to quantify how well a communication strategy can exploit the available spatial dimensions. We define the multiplexing gain, also referred to as Degrees-of-Freedom (DoF), of user-k achieved with communication strategy[10] j as

$$d_k^{(j)} = \lim_{P\to\infty} \frac{R_k^{(j)}(P)}{\log_2(P)}, \tag{6}$$

and the sum multiplexing gain as

$$d_{\rm s}^{(j)} = \lim_{P\to\infty} \frac{R_{\rm s}^{(j)}(P)}{\log_2(P)} = \sum_{k=1}^{K} d_k^{(j)}, \tag{7}$$

where $R_{\rm s}^{(j)} = \sum_{k=1}^{K} R_k^{(j)}$ is the sum-rate. We also define the MMF multiplexing gain as

$$d_{\rm mmf}^{(j)} = \lim_{P\to\infty} \frac{R_{\rm mmf}^{(j)}(P)}{\log_2(P)} = \min_{k=1,\ldots,K} d_k^{(j)}, \tag{8}$$

where $R_{\rm mmf}^{(j)} = \min_{k=1,\ldots,K} R_k^{(j)}$ is the MMF rate.

The multiplexing gain $d_k^{(j)}$ is a first-order approximation of the rate of user-k at high SNR. $d_k^{(j)}$ can be viewed as the pre-log factor of the rate of user-k at high SNR and be interpreted as the number or fraction of interference-free stream(s) that can be simultaneously communicated to user-k by employing communication strategy j. The larger $d_k^{(j)}$, the faster the rate of user-k increases with the SNR. Hence, ideally a communication strategy should achieve the highest multiplexing gain possible.

The sum multiplexing gain $d_{\rm s}^{(j)}$ is a first-order approximation of the sum-rate at high SNR and therefore the pre-log factor of the sum-rate and can be interpreted as the total number of interference-free data streams that can be simultaneously communicated to all K users by employing communication strategy j. In other words, $R_{\rm s}^{(j)}$ scales as $d_{\rm s}^{(j)}\log_2(P)+\delta$ where δ is a term that scales slowly with SNR such that $\lim_{P\to\infty}\frac{\delta}{\log_2(P)} = 0$ (e.g., $O(1)$, $O(\log_2(\log_2(P)))$ or $O(\sqrt{\log_2(P)})$), and the larger $d_{\rm s}^{(j)}$, the faster the sum-rate increases with the SNR.

The MMF multiplexing gain $d_{\rm mmf}^{(j)}$, also referred to as symmetric multiplexing gain, corresponds to the maximum multiplexing gain that can be simultaneously achieved by all users, and reflects the pre-log factor of the MMF rate at high SNR. In other words, $R_{\rm mmf}^{(j)}$ scales as $d_{\rm mmf}^{(j)}\log_2(P)+\delta$, and the larger $d_{\rm mmf}^{(j)}$, the faster the MMF rate increases with the SNR.

10. Throughout this paper, j will be either N for NOMA, M for MU–LP, R for Rate-Splitting, or $\star$ for the information theoretic optimum, i.e., $j \in \{\text{N, M, R, }\star\}$.

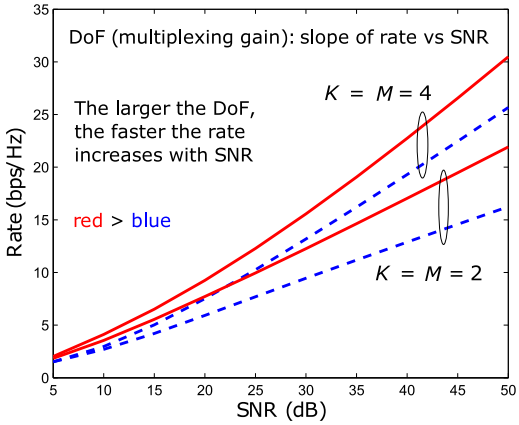


FIGURE 2. Illustration of the notion of multiplexing gain/DoF.

Remark 1: Much of the analysis and discussion in this paper emphasizes the (sum and MMF) multiplexing gain as a metric to assess the capability of a strategy to exploit multiple antennas. As it becomes plausible from its definition, the multiplexing gain is an asymptotic metric valid in the limit of high SNR, and hence, does not precisely reflect specific finite-SNR rates. Nevertheless, it provides firm theoretical grounds for performance comparisons and has been used in the MIMO literature for two decades [55]. Furthermore, the multiplexing gain also impacts the performance at finite SNRs as shown in numerous papers [47], [48], [56] and in our simulation results in Section IX. Moreover, it enables to gain deep insights into the performance limits and to guide the design of efficient communications strategies, as we will see throughout this paper. The notion of multiplexing gain is illustrated in Fig. 2 where the strategy characterized by the red curves is preferred over the other strategy in blue. A larger slope/multiplexing gain is indeed offered by the red strategy at high SNR which translates into rate gains at finite SNR.

Remark 2: In this manuscript, we will derive the sum multiplexing gain $d_{\rm s}^{(j)}$ and the MMF multiplexing gain $d_{\rm mmf}^{(j)}$ for strategy $j \in \{\text{N, M, R}\}$. The corresponding proofs rely on obtaining an upper bound (i.e., converse) on the (sum/MMF) multiplexing gain and then showing that this upper bound is tight since it is achievable by the strategy under study. In other words, we show that $d_{\rm s/mmf}^{(j)} \leq a$ (upper bound) and then $d_{\rm s/mmf}^{(j)} \geq a$ (achievability). Consequently, this paper characterizes the exact (sum/MMF) multiplexing gains achieved by each strategy, i.e., $d_{\rm s/mmf}^{(j)} = a$. We confirm the multiplexing gains by numerical simulations in evaluations.

C. DISCUSSIONS

Note that (2) and (3), respectively, suggest that s_1 is received interference-free at user-1, and that s_2 is always decoded in the presence of interference from s_1. We can now draw some important conclusions from (2), (3), and (5).

The sum-rate bound (5) of this two-user MISO NOMA strategy and user ordering user-2→user-1 can be further upper bounded as

$$R_{\mathrm{s}}^{(\mathrm{N})} \leq \log_2\Big(1+\|\mathbf{h}_1\|^2 P\Big), \tag{9}$$

where the equality in (9) is achieved (i.e., upper bound is tight) by choosing $\mathbf{p}_1 = \sqrt{P_1}\mathbf{h}_1/\|\mathbf{h}_1\|$ and $\mathbf{p}_2 = \sqrt{P_2}\mathbf{h}_1/\|\mathbf{h}_1\|$ with $P_1+P_2=P$. Note that the right hand side of (9) is the rate achieved by OMA when serving user-1. In other words, (9) is not just an upper bound on the sum-rate of MISO NOMA but is actually the maximum achievable sum-rate of MISO NOMA. This maximum achievable sum-rate of MISO NOMA is the same as that of OMA (when serving user-1).

Had we considered the other decoding order where the shared codebook is used to encode W_1 and user-2 decodes s_1, the role of user-1 and user-2 in Fig. 1 would have been switched (user-1→user-2) and we would have obtained

$$R_{\mathrm{s}}^{(\mathrm{N})} \leq \log_2\Big(1+\|\mathbf{h}_2\|^2 P\Big). \tag{10}$$

This sum-rate upper bound is achievable by choosing $\mathbf{p}_1 = \sqrt{P_1}\mathbf{h}_2/\|\mathbf{h}_2\|$ and $\mathbf{p}_2 = \sqrt{P_2}\mathbf{h}_2/\|\mathbf{h}_2\|$ with $P_1+P_2=P$ and the maximum achievable sum-rate of MISO NOMA with decoding order user-1→user-2 is the same as that of OMA (when serving user-2 only).

Hence, from (9) and (10), the sum-rate of MISO NOMA considering adaptive decoding order is upper bounded as

$$R_{\mathrm{s}}^{(\mathrm{N})} \leq \log_2\Big(1+\max\Big\{\|\mathbf{h}_1\|^2, \|\mathbf{h}_2\|^2\Big\}P\Big). \tag{11}$$

This sum-rate is again achievable and is the same as that of OMA when serving the strongest of the two users $\arg\max_{k=1,2}\|\mathbf{h}_k\|$.

Importantly, (9), (10), and (11) reveal the strong result that *the sum-rate of MISO NOMA is actually no higher than that of OMA for any SNR*! This fact is not surprising in the SISO case ($M=1$) since it is well known that to achieve the sum capacity of the SISO BC, one can simply transmit to the strongest user all the time (i.e., OMA) [57]. The above result shows that this also holds for the two-user MISO NOMA basic building block.

Considering the high SNR regime, (9), (10), (11) all scale at most as $\log_2(P)$, i.e.,

$$R_{\mathrm{s}}^{(\mathrm{N})} \overset{P\nearrow}{=} \log_2(P)+\delta, \tag{12}$$

which highlights that the sum multiplexing gain of two-user MISO NOMA (irrespectively of the decoding order) is (at most) one, i.e., $d_{\mathrm{s}}^{(\mathrm{N})}=1$. Hence, MISO NOMA limits the sum multiplexing gain to $d_{\mathrm{s}}^{(\mathrm{N})} = d_1^{(\mathrm{N})} + d_2^{(\mathrm{N})} = 1$, i.e., the same as OMA.

The sum multiplexing gain of one can be further split equally between the two users, which leads to an MMF multiplexing gain of two-user MISO NOMA given by $d_{\mathrm{mmf}}^{(\mathrm{N})} = \frac{1}{2}$. This is achieved by scaling the power allocated to user-1 as $O(P^{1/2})$ and that to user-2 as $O(P)$. In other words, the MMF rate of this two-user MISO NOMA scales at most as $\frac{1}{2}\log_2(P)$ at high SNR.

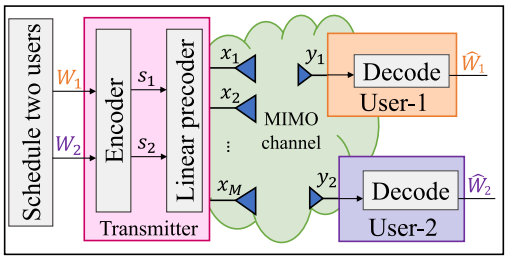


FIGURE 3. Two-user system architecture with MU–LP/SDMA.

The above contrasts with the optimal sum multiplexing gain $d_{\mathrm{s}}^{(\star)}$ of the two-user MISO BC, that is equal to 2, i.e., two interference-free streams can be transmitted.[11] This can be achieved by performing conventional MU–LP, illustrated in Fig. 3. Recall the MU–LP system model where W_1 and W_2 are independently encoded into streams s_1 and s_2 and respectively precoded by $\mathbf{p}_1$ and $\mathbf{p}_2$ such that the transmit signal is given by

$$\mathbf{x} = \mathbf{p}_1 s_1 + \mathbf{p}_2 s_2. \tag{13}$$

At the receivers, $y_k = \mathbf{h}_k^H\mathbf{x}+n_k$, $k=1,2$, and s_1 and s_2 are respectively decoded by user-1 and user-2 by treating any residual interference as noise, leading to MU–LP rates

$$R_1^{(\mathrm{M})} = \log_2(1+C),\ R_2^{(\mathrm{M})} = \log_2(1+B), \tag{14}$$

with

$$C = \frac{\left|\mathbf{h}_1^H\mathbf{p}_1\right|^2}{1+\left|\mathbf{h}_1^H\mathbf{p}_2\right|^2}, \tag{15}$$

and B as specified in (4). It is then indeed sufficient[12] to transmit two streams using uniform power allocation and Zero-Forcing Beamforming (ZFBF), so that $\mathbf{h}_1^H\mathbf{p}_2 = \mathbf{h}_2^H\mathbf{p}_1 = 0$, to reap the sum multiplexing gain $d_{\mathrm{s}}^{(\mathrm{M})} = d_{\mathrm{s}}^{(\star)} = 2$ and the MMF multiplexing gain $d_{\mathrm{mmf}}^{(\mathrm{M})} = d_{\mathrm{mmf}}^{(\star)} = 1$ (i.e., each user gets one full interference-free stream). Indeed, with MU–LP, the sum-rate scales as $2\log_2(P)$ and the MMF rate as $\log_2(P)$ at high SNR [58]–[60]. Such sum-rate and MMF rate would always strictly outperform that of NOMA (and OMA) at high SNR. Since both OMA and NOMA achieve only half the (sum/MMF) multiplexing gain of MU–LP in the two-user MISO BC considered, it is not clear whether (and under what conditions) multi-antenna NOMA can outperform MU–LP and other forms of multi-user multi-antenna communication strategies, and if it does, whether multi-antenna NOMA is

11. This assumes that the two channel directions are not aligned, or in other words, that the rank of the concatenated matrix $[\,\mathbf{h}_1 \quad \mathbf{h}_2\,]$ is equal to 2. Note that this condition on full-rank concatenated matrices is met in practice with probability one.

12. More complicated precoders (or communication strategies like non-linear precoding and DPC) can be used to enhance the rate performance, but the sum and MMF multiplexing gains will not improve in this 2-user setting.

worth the associated increase in receiver complexity. The above discussion exposes some weaknesses of multi-antenna NOMA and highlights the uncertainty regarding the potential benefits of multi-antenna NOMA. Hence, in the following sections, we derive the multiplexing gains of generalized K-user multi-antenna NOMA, so as to better assess its potential.

Remark 3: It appears from (1) and (13) that the transmit signal vectors for 2-user MISO NOMA and 2-user MU–LP are the same, therefore giving the impression that NOMA is the same as MU–LP. This is obviously incorrect. Recall the major differences in the encoding and the decoding of NOMA and MU–LP:

- *Encoding:* In NOMA, W_1 is encoded into s_1 and W_2 is encoded into s_2 at a rate such that s_2 is decodable by both users, while W_1 and W_2 are independently encoded into streams s_1 and s_2 in MU–LP.
- *Decoding:* User-1 decodes s_1 and s_2 and user-2 decodes s_2 by treating s_1 as noise in NOMA while s_1 is decoded by user-1 by treating s_2 as noise and s_2 is decoded by user-2 by treating s_1 as noise in MU–LP.

Consequently the rate expressions (2), (3) and (14) are different, which therefore suggests that the best pair of precoders $\mathbf{p}_1$ and $\mathbf{p}_2$ that maximizes a given objective function (e.g., sum-rate, MMF rate) would be different for NOMA and MU–LP. Choosing $\mathbf{p}_1$ and $\mathbf{p}_2$ according to ZFBF would commonly work reasonably well for MU–LP but would lead to $R_2^{(\mathrm{N})} = 0$ in (3) for NOMA. Nevertheless, the above discussion on multiplexing gain loss of MISO NOMA always holds, even in the event where MISO NOMA is implemented with the best choice of precoders, since the above analysis for MISO NOMA is based on an upper bound.

III. K-USER MISO NOMA WITH PERFECT CSIT

We now study K-user MISO NOMA relying on perfect CSIT and derive the sum and MMF multiplexing gains.

A. MISO NOMA SYSTEM MODEL

We consider a K-user MISO NOMA scenario where a single transmitter equipped with M transmit antennas serves K single-antenna users indexed by $\mathcal{K} = \{1, \ldots, K\}$. The K users are grouped into $1 \leq G < K$ groups[13] with groups indexed by $\mathcal{G} = \{1, \ldots, G\}$. There are g users per group, i.e., we therefore assume for simplicity that $K = gG$. Users in group i are indexed by $\mathcal{K}_i = \{ig - g + 1, \ldots, ig\}$. Hence, $\mathcal{K} = \bigcup_{i \in \mathcal{G}} \mathcal{K}_i$ and $|\mathcal{K}_i| = g$. Without loss of generality, we assume that users $1, g + 1, 2g + 1, \ldots, K - g + 1$ are the "strong users"[14] respectively in group 1 to G, and perform $g - 1$ layers of SIC to fully decode the messages (and therefore remove interference) from the other $g - 1$ users within the same group. Similarly, the second user in each group (i.e., $ig - g + 2$ in group i) performs $g - 2$ layers of SIC to fully decode messages from other $g - 2$ users within the same group, and so on. The two most popular MISO NOMA strategies employ either $G = 1$ [20]–[23] or $G = K/2$ [26]–[31] but we keep here the scenario general for any value of $1 \leq G < K$. The general architecture of MISO NOMA is illustrated in Fig. 4. The two-user building block in Section II can be viewed as a particular instance with $K = 2$ and $G = 1$.

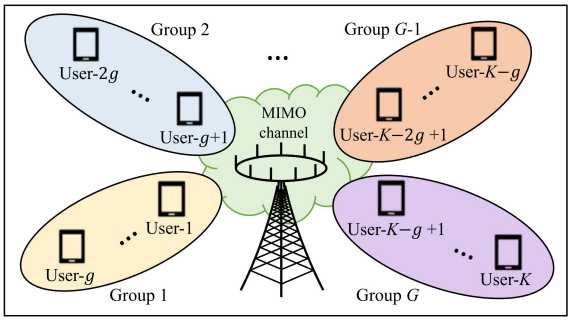


FIGURE 4. *K*-user system architecture with MISO NOMA (containing *G* user groups and *g* users within each group).

At the transmitter, the messages W_1 to W_K intended for user-1 to user-K, respectively, are encoded into s_1 to s_K. However, some of the messages in each group have to be encoded using codebooks shared by a subset of the users in that group so that they can be decoded and cancelled from the received signals at the co-scheduled users in that group. In particular, taking group 1 as an example, W_2 to W_g are encoded using codebooks shared with user-1 such that user-1 can decode all of these $g - 1$ messages. After encoding, the K streams are linearly precoded by precoders[15] $\mathbf{p}_1$ to $\mathbf{p}_K$, where $\mathbf{p}_k \in \mathbb{C}^M$ is the precoder of s_k, and superposed at the transmitter. The resulting transmit signal is

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{p}_k s_k. \tag{16}$$

Defining $\mathbf{s} = [s_1, \ldots, s_K]^T$ and assuming that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}$, the average transmit power constraint is written as $\sum_{k=1}^{K} P_k \leq P$, where $P_k = \|\mathbf{p}_k\|^2$.

13. Note that $1 \leq G < K$ is a widely considered option for MISO NOMA in which there exists (at least) one user decoding the message of (at least) one another user in each group. Importantly, $G = K$ corresponds to MU–LP as per Section VI and is not a MISO NOMA scheme since all K messages are independently encoded into K streams and residual interference is treated as noise at the receivers, i.e., there is no shared codebook and users therefore do not decode the messages of other users.

14. "Strong users" here refer to the users who decode the messages of other users in a group. Given the nondegraded nature of the multi-antenna BC, the strong users do not necessarily have to be the users with the largest channel vector norm. The multiplexing gain analysis is general and holds for any ordering. Nevertheless, following [22], [23], we consider in the simulation section the decoding order in each group to be the ascending order of users' channel strength such that "strong users" refer to the users with the largest channel vector norm respectively in group 1 to G.

15. Further constraints can be imposed on the precoder design such that the same precoder is used for all users in the same group. This constraint would however further reduce the optimization space and therefore the rate performance.

At the receiver side, the signal received at user-k is $y_k = \mathbf{h}_k^H \mathbf{x} + n_k$, $k \in \mathcal{K}$, where $\mathbf{h}_k$ is the channel vector[16] of user-k, perfectly known at the transmitter and that user, and $n_k \sim \mathcal{CN}(0,1)$ is the AWGN. By employing SIC, user-j in group i (i.e., $j \in \mathcal{K}_i$) decodes the messages of users-$\{k|k \geq j, k \in \mathcal{K}_i\}$ within the same user group in a descending order of the user index while treating the interference from users in different groups as noise. Under the assumption of Gaussian signaling and perfect SIC, the rate at user-$j, j \in \mathcal{K}_i$, to decode the message of user-$k, k \geq j, k \in \mathcal{K}_i$, is given by

$$R_{j,k} = \log_2\left(1 + \frac{\left|\mathbf{h}_j^H \mathbf{p}_k\right|^2}{I_{j,k}^{(in)} + I_{j,k}^{(ou)} + 1}\right), \tag{17}$$

where

$$I_{j,k}^{(in)} = \sum_{m<k, m\in\mathcal{K}_i} \left|\mathbf{h}_j^H \mathbf{p}_m\right|^2, I_{j,k}^{(ou)} = \sum_{l\neq i, l\in\mathcal{G}} \sum_{m\in\mathcal{K}_l} \left|\mathbf{h}_j^H \mathbf{p}_m\right|^2 \tag{18}$$

are the intra-group interference and inter-group interference received at user-k, respectively. As the message of user-$k, k \in \mathcal{K}_i$, has to be decoded by users-$\{j|j \leq k, j \in \mathcal{K}_i\}$, to ensure decodability, the rate of user-k should not exceed

$$R_k^{(\mathrm{N})} = \min_{j\leq k, j\in\mathcal{K}_i} R_{j,k}. \tag{19}$$

In the next subsection, we study the sum multiplexing gain and the MMF multiplexing gain of K-user MISO NOMA.

B. MULTIPLEXING GAINS

The following proposition provides the sum multiplexing gain of MISO NOMA for perfect CSIT.

Proposition 1: The sum multiplexing gain of K-user MISO NOMA with M transmit antennas, G groups of $g = K/G$ users, and perfect CSIT is $d_{\mathrm{s}}^{(\mathrm{N})} = \min(M, G)$.

Proof: The proof is obtained by showing that an upper bound on the sum multiplexing gain is achievable. The upper bound is obtained by applying the MAC argument (used in (5)) to the strong user in each group and noticing that the sum-rate in groups 1 to G is upper bounded as

$$\sum_{k=1}^{g} R_k^{(\mathrm{N})} \leq \log_2\left(1 + \sum_{k=1}^{g}\left|\mathbf{h}_1^H \mathbf{p}_k\right|^2\right), \tag{20}$$

$$\sum_{k=g+1}^{2g} R_k^{(\mathrm{N})} \leq \log_2\left(1 + \sum_{k=g+1}^{2g}\left|\mathbf{h}_{g+1}^H \mathbf{p}_k\right|^2\right), \tag{21}$$

$$\vdots$$

$$\sum_{k=K-g+1}^{K} R_k^{(\mathrm{N})} \leq \log_2\left(1 + \sum_{k=K-g+1}^{K}\left|\mathbf{h}_{K-g+1}^H \mathbf{p}_k\right|^2\right). \tag{22}$$

16. The rank of matrix $[\mathbf{h}_1 \ldots \mathbf{h}_K]$ is assumed equal to $\min\{M, K\}$ for simplicity. Note that this condition is met in practice and is motivated by practical deployments. Ranks strictly smaller than $\min\{M, K\}$ (due to, e.g., aligned channels) would not occur (zero probability) in real wireless deployments with fading channels and are therefore not of practical interest.

Note that the left-hand sides of (20), (21), and (22) refer to the sum of the rates of the messages in group 1, 2, and G, respectively, but can also be viewed as the total rate to be decoded by user 1, $g+1$, and $K-g+1$ (since those users decode all the messages in their respective group). We now notice that the right-hand sides of (20), (21), and (22) scale as $\log_2(P) + \delta$ for large P (following the same argument as in the two-user case). This implies that each group i achieves at most a (group) sum multiplexing gain $d_{\mathrm{s},i}^{(\mathrm{N})} = \sum_{k=ig-g+1}^{ig} d_k^{(N)}$ of 1, i.e., at most one interference-free stream can be transmitted to each group. Summing up all inequalities, we obtain in the limit of large P that

$$R_{\mathrm{s}}^{(\mathrm{N})} = \sum_{k=1}^{K} R_k^{(\mathrm{N})} \overset{P\nearrow}{\leq} G\log_2(P) + \delta, \tag{23}$$

which shows that $d_{\mathrm{s}}^{(\mathrm{N})} = \sum_{i=1}^{G} d_{\mathrm{s},i}^{(\mathrm{N})} \leq G$. Additionally, since $d_{\mathrm{s}}^{(\mathrm{N})} \leq d_{\mathrm{s}}^{\star} = \min(M, K)$, we have $d_{\mathrm{s}}^{(\mathrm{N})} \leq \min(M, G)$.

The achievability part shows that $d_{\mathrm{s}}^{(\mathrm{N})} \geq \min(M, G)$. To this end, it is indeed sufficient to perform ZFBF and transmit $\min(M, G)$ interference-free streams to $\min(M, G)$ of the G "strong users". Combining the upper bound and achievability leads to the conclusion that $d_{\mathrm{s}}^{(\mathrm{N})} = \min(M, G)$. ■

The following result derives the MMF multiplexing gain of MISO NOMA with perfect CSIT.

Proposition 2: The MMF multiplexing gain of K-user MISO NOMA with M transmit antennas, G groups of $g = K/G$ users and perfect CSIT is

$$d_{\mathrm{mmf}}^{(\mathrm{N})} = \begin{cases} \frac{1}{g}, & M \geq K-g+1, \\ 0, & M < K-g+1. \end{cases} \tag{24}$$

For $G = 1$, i.e., $g = K$, $d_{\mathrm{mmf}}^{(\mathrm{N})} = \frac{1}{K}$.

Proof: Let us first consider $M \geq K-g+1$. The MMF multiplexing gain is always upperbounded by ignoring the inter-group interference, i.e., the G groups are non-interfering. Following again the MAC argument, the sum multiplexing gain of one in each group can then be further split equally among the g users, which leads to an upper bound on the MMF multiplexing gain of $\frac{1}{g}$. Achievability is simply obtained by designing the precoders using ZFBF to eliminate all inter-group interference, and allocating the power similarly to Section II-C. Indeed let us consider group 1 for simplicity, and allocate the power to user $k = 1, \ldots, g$ as $O(P^{k/g})$. This leads to an SINR for user-k scaling as $O(P^{1/g})$ and to an achievable MMF multiplexing gain of $\frac{1}{g}$. For $G = 1$, one can simply allocate the power to user $k = 1, \ldots, K$ as $O(P^{k/K})$, which leads to an achievable MMF multiplexing gain of $\frac{1}{K}$.

Let us now consider $M < K-g+1$. Take $M = K-g$ (any smaller M cannot improve the multiplexing gain). Precoder $\mathbf{p}_k$ of any user-k can be made orthogonal to the channel of at most $K-g-1$ co-scheduled users and will therefore cause interference to at least one user in another group. As a result, the MMF multiplexing gain collapses to 0. ■

Remark 4: For the MMF multiplexing gain analysis, it should be noted that we consider one-shot transmission schemes with no time-sharing between strategies. This is suitable for systems with rigid scheduling and/or tight latency constraints, and also allows for simpler designs. This assumption is also commonly used in the NOMA literature [34], [35], [39], [53], [54].

IV. K-USER MISO NOMA WITH IMPERFECT CSIT

We now go one step further and extend the multiplexing gain analysis to the imperfect CSIT setting. The results in this section therefore generalize the results in the previous section (with perfect CSIT being a particular case of imperfect CSIT). In this section, the achievable rates are defined in the ergodic sense in a standard Shannon theoretic fashion, and the corresponding sum and MMF mutiplexing gains are defined similarly to Section II-B using ergodic rates. We first introduce the CSIT error model before deriving the multiplexing gains of MISO NOMA relying on imperfect CSIT.

A. CSIT ERROR MODEL

For each user, the transmitter acquires an imperfect estimate of the channel vector $\mathbf{h}_k$, denoted as $\hat{\mathbf{h}}_k$. The CSIT imperfection is modelled by

$$\mathbf{h}_k = \hat{\mathbf{h}}_k + \tilde{\mathbf{h}}_k, \tag{25}$$

where $\tilde{\mathbf{h}}_k$ denotes the corresponding channel estimation error at the transmitter. For compactness, we define $\mathbf{H} = [\mathbf{h}_1 \dots \mathbf{h}_K]$, $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1 \dots \hat{\mathbf{h}}_K]$, and $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1 \dots \tilde{\mathbf{h}}_K]$, which implies $\mathbf{H} = \hat{\mathbf{H}} + \tilde{\mathbf{H}}$. The joint fading process is characterized by the joint distribution $f_{\mathbf{H},\hat{\mathbf{H}}}(\mathbf{H}, \hat{\mathbf{H}})$ of $\{\mathbf{H}, \hat{\mathbf{H}}\}$, assumed to be stationary and ergodic . The joint distribution $f_{\mathbf{H},\hat{\mathbf{H}}}(\mathbf{H}, \hat{\mathbf{H}})$ is continuous and known to the transmitter. The ergodic rates capture the long-term performance over a long sequence of channel uses $\{\mathbf{H}, \hat{\mathbf{H}}\}$ spanning almost all possible joint channel states.

For each user-k, we define the average channel (power) gain as $\Gamma_k = \mathbb{E}\{\|\mathbf{h}_k\|^2\}$. Similarly, we define $\hat{\Gamma}_k = \mathbb{E}\{\|\hat{\mathbf{h}}_k\|^2\}$ and $\tilde{\Gamma}_k = \mathbb{E}\{\|\tilde{\mathbf{h}}_k\|^2\}$. For many CSIT acquisition mechanisms [61], $\hat{\mathbf{h}}_k$ and $\tilde{\mathbf{h}}_k$ are uncorrelated according to the orthogonality principle [62]. By further assuming that $\hat{\mathbf{h}}_k$ and $\tilde{\mathbf{h}}_k$ have zero means, we have $\Gamma_k = \hat{\Gamma}_k + \tilde{\Gamma}_k$, based on which we can write $\hat{\Gamma}_k = (1 - \sigma^2_{e,k})\Gamma_k$ and $\tilde{\Gamma}_k = \sigma^2_{e,k}\Gamma_k$ for some $\sigma^2_{e,k} \in [0, 1]$. Note that $\sigma^2_{e,k}$ is the normalized estimation error variance for user-k's CSIT, e.g., $\sigma^2_{e,k} = 1$ corresponds to no instantaneous CSIT, while $\sigma^2_{e,k} = 0$ represents perfect instantaneous CSIT.

For simplicity, we assume identical normalized CSIT error variances for all users, i.e., $\sigma^2_{e,k} = \sigma^2_e$ for all $k = 1, \dots, K$. To facilitate the multiplexing gain analysis, we assume that σ^2_e scales with SNR as $\sigma^2_e = P^{-\alpha}$ for some CSIT quality parameter $\alpha \in [0, \infty)$ [46], [47], [60], [63], [64]. This is a convenient and tractable model extensively used in the information theoretic literature that allows us to assess the performance of the system in a wide range of CSIT quality conditions. Indeed, the larger α, the faster the CSIT error decreases with the SNR. The two extreme cases, $\alpha = 0$ and $\alpha = \infty$, correspond to no or constant CSIT (i.e., that does not scale or improve with SNR) and perfect CSIT, respectively. As far as the multiplexing gain analysis is concerned, however, we may truncate the CSIT quality parameter as $\alpha \in [0, 1]$, where $\alpha = 1$ amounts to perfect CSIT in the multiplexing gain sense. The regime $\alpha \in (0, 1)$ corresponds to partial CSIT, resulting from imperfect CSI acquisition. The CSIT quality α can be interpreted in many different ways, but a plausible interpretation of α is related to the number of feedback bits, where $\alpha = 0$ corresponds to a fixed number of feedback bits for all SNRs, $\alpha = \infty$ corresponds to an infinite number of feedback bits, and $0 < \alpha < \infty$ reflects how quickly the number of feedback bits increases with the SNR. As a reference, a system like 4G and 5G use $\alpha = 0$ when limited feedback (or codebook-based feedback) is used to report the CSI, since the number of feedback bits is constant and does not scale with SNR, e.g., 4 bits of CSI feedback in 4G LTE for $M = 4$.

B. MULTIPLEXING GAINS

The following result quantifies the sum multiplexing gain of MISO NOMA for imperfect CSIT.

Proposition 3: The sum multiplexing gain of K-user MISO NOMA with M transmit antennas, G groups of $g = K/G$ users, and CSIT quality $0 \le \alpha \le 1$ is $d_{\mathrm{s}}^{(\mathrm{N})} = \max(1, \min(M, G)\alpha)$.

Proof: Similar to the proof of Proposition 1, let us look at the G strong users since they have to decode all messages. We recall that $d_{\mathrm{s},i}^{(\mathrm{N})} = \sum_{k=ig-g+1}^{ig} d_k^{(N)}$ reflects the multiplexing gain of the total rate to be decoded by the strong user $ig - g + 1$ in group i as a consequence that this user decodes all g messages in group i. Making use of the results of MU–LP in the G-user MISO BC with imperfect CSIT [47].[17] we obtain $d_{\mathrm{s}}^{(\mathrm{N})} = \sum_{i=1}^{G} d_{\mathrm{s},i}^{(\mathrm{N})} = \sum_{k=1}^{K} d_k^{(N)} \le \max(1, \min(M, G)\alpha)$.

The achievability part shows that $d_{\mathrm{s}}^{(\mathrm{N})} \ge \max(1, \min(M, G)\alpha)$. It is indeed sufficient to perform ZFBF and transmit $\min(M, G)$ streams, each at a power level of $P^{\alpha}/\min(M, G)$, to $\min(M, G)$ of the G "strong users". If $\min(M, G)\alpha < 1$, one can simply transmit a single stream (i.e., perform OMA) and reap a sum multiplexing gain of 1. Combining the upper bound and achievability leads to the conclusion that we have $d_{\mathrm{s}}^{(\mathrm{N})} = \max(1, \min(M, G)\alpha)$. ■

For $\alpha = 1$ (perfect CSIT from a multiplexing gain perspective), Proposition 3 boils down to the perfect CSIT result in Proposition 1.

The following proposition provides the MMF multiplexing gain of MISO NOMA with imperfect CSIT.

Proposition 4: The MMF multiplexing gain of K-user MISO NOMA with M transmit antennas, G groups of

17. See also Proposition 7.

$g = K/G$ users, and CSIT quality $0 \leq \alpha \leq 1$ is

$$d_{\text{mmf}}^{(\text{N})} = \begin{cases} \frac{\alpha}{g}, & G > 1 \text{ and } M \geq K - g + 1, \\ 0, & G > 1 \text{ and } M < K - g + 1, \\ \frac{1}{K}, & G = 1. \end{cases} \tag{26}$$

The proof is relegated to Appendix A.

It is interesting to note that the sensitivity of the multiplexing gain of MISO NOMA to the CSIT quality α is different for $G > 1$ and $G = 1$. Indeed the sum and MMF multiplexing gains of MISO NOMA with $G > 1$ decay as α decreases, while the multiplexing gains of MISO NOMA with $G = 1$ are not affected by α. This can be interpreted in two different ways. On the one hand, this suggests that MISO NOMA $G = 1$ is inherently robust to CSIT imperfection since the multiplexing gains are not affected by $\alpha < 1$. On the other hand, this also reveals that MISO NOMA with $G = 1$ is unable to exploit the presence of CSIT since its multiplexing gains are the same as in the absence of CSIT ($\alpha = 0$).

V. MIMO NOMA

We now consider multi-antenna receivers and extend the two-user MISO NOMA toy example of Section II to a two-user MIMO NOMA setting with perfect CSIT.

We consider a two-user MIMO BC, consisting of one transmitter with M antennas and two users equipped with N antennas each. The signal vector $\mathbf{y}_k \in \mathbb{C}^{N\times 1}$ received at user-k is written as $\mathbf{y}_k = \mathbf{H}_k^H \mathbf{x} + \mathbf{n}_k$, where $\mathbf{H}_k \in \mathbb{C}^{M\times N}$ is the channel matrix[18] of user-k and $\mathbf{n}_k$ is the AWGN vector at user-k. Following the NOMA principle, the transmit signal vector $\mathbf{x}$ is generated such that the messages intended for user-2 are encoded using a shared codebook so as to be decodable by user-1. Defining the transmit covariance matrix of user-k as $\mathbf{Q}_k$ subject to the average transmit power constraint $\text{Tr}(\mathbf{Q}_1) + \text{Tr}(\mathbf{Q}_2) \leq P$, and assuming Gaussian signaling, the achievable rates of both users are given by

$$R_1^{(\text{N})} = \log_2 \det\Big(\mathbf{I}_N + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1\Big), \tag{27}$$

$$R_2^{(\text{N})} = \min\Big(\log_2 \det(\mathbf{I}_N + \mathbf{R}_1), \log_2 \det(\mathbf{I}_N + \mathbf{R}_2)\Big), \tag{28}$$

where $\mathbf{R}_k = \mathbf{H}_k^H \mathbf{Q}_2 \mathbf{H}_k (\mathbf{I}_N + \mathbf{H}_k^H \mathbf{Q}_1 \mathbf{H}_k)^{-1}$, $k = 1, 2$.

The sum-rate $R_s^{(\text{N})}$ of the two-user MIMO NOMA can then be bounded as

$$\begin{aligned} R_{\text{s}}^{(\text{N})} &\leq \log_2 \det(\mathbf{I}_N + \mathbf{R}_1) + \log_2 \det\Big(\mathbf{I}_N + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1\Big), \\ &= \log_2 \det\Big(\mathbf{I}_N + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1 + \mathbf{H}_1^H \mathbf{Q}_2 \mathbf{H}_1\Big). \end{aligned} \tag{29}$$

The sum-rate bound achieved with this two-user MIMO NOMA strategy can be further upper bounded as

$$\begin{aligned} R_{\text{s}}^{(\text{N})} &\leq \log_2 \det\Big(\mathbf{I}_N + \mathbf{H}_1^H \mathbf{Q}_1^\star \mathbf{H}_1\Big), \\ &\overset{P\nearrow}{=} \min(M, N) \log_2(P) + O(1), \end{aligned} \tag{30}$$

where $\mathbf{Q}_1^\star$ refers to the optimal covariance matrix for user-1 in a single-user (OMA) setup with $\text{Tr}(\mathbf{Q}_1) = P$, i.e., obtained by transmitting along the dominant eigenvector of $\mathbf{H}_1 \mathbf{H}_1^H$ and allocating power P according to the water-filling solution.

Similarly to the MISO case, the other decoding order could also be considered and a similar analysis can be obtained. Ultimately, the sum-rate of MIMO NOMA (irrespective of the decoding order) is actually no larger than that of OMA at any SNR. The sum multiplexing gain is limited by $d_{\text{s}}^{(\text{N})} = \min(M, N)$, which is smaller than the optimal sum multiplexing gain of the MIMO BC $d_{\text{s}}^{(\star)} = \min(M, 2N)$, achieved by conventional MU-MIMO/MU–LP precoding [59]. This analysis highlights that MIMO NOMA incurs a sum multiplexing loss whenever $N < M$, i.e., when the number of receive antennas at each device is smaller than the number of transmit antennas at the base station, which would occur in most realistic and practical MIMO deployments. Similarly, the MMF multiplexing gain is also affected since $d_{\text{mmf}}^{(\text{N})} = \min(M, N)/2$, obtained by equally splitting the sum multiplexing gain amongst the two users, which again incurs a loss whenever $N < M$. Taking for instance $M = 6$ and $N = 4$ leads to $d_{\text{s}}^{(\text{N})} = 4$ and $d_{\text{mmf}}^{(\text{N})} = 2$, though one could easily transmit using multi-user MIMO (e.g., block diagonalization [65], [66]) 6 interference-free streams with 3 streams per user.

Recall that the above MIMO NOMA scheme and analysis were based on the principle that one user is forced to fully decode the messages of the other co-scheduled user. Nevertheless other MIMO NOMA schemes have recently appeared that may not satisfy this definition of MIMO NOMA and may therefore achieve different (and hopefully superior) sum and MMF multiplexing gains [67], [68].

VI. BASELINE SCHEME I: CONVENTIONAL MULTI-USER LINEAR PRECODING

The first baseline to assess the performance of multi-antenna NOMA is conventional Multi-User Linear Precoding. In the sequel, we recall the multiplexing gains achieved by MU–LP.

A. MU–LP SYSTEM MODEL

Following Section III-A, we consider a K-user MISO BC with one transmitter equipped with M transmit antennas and K single-antenna users. As per Fig. 5, the messages $W_1, \ldots, W_K$ respectively for user-1 to user-K are independently encoded into s_1 to s_K, which are then mapped to the transmit antennas through precoders $\mathbf{p}_1, \ldots, \mathbf{p}_K$. The resulting transmit signal is $\mathbf{x} = \sum_{k=1}^K \mathbf{p}_k s_k$.

The signal received at user-k is $y_k = \mathbf{h}_k^H \mathbf{x} + n_k$ with $n_k \sim \mathcal{CN}(0, 1)$. Each user-$k$ directly decodes the intended message W_k by treating the interference from the other users as noise. Under the assumption of Gaussian signaling, the rate of user-k for $k \in \mathcal{K}$ is given by

$$R_k^{(\text{M})} = \log_2\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{p}_k\right|^2}{1 + \sum_{q\neq k} \left|\mathbf{h}_k^H \mathbf{p}_q\right|^2}\right). \tag{31}$$

18. We assume for simplicity that $\mathbf{H}_k$ is full rank.

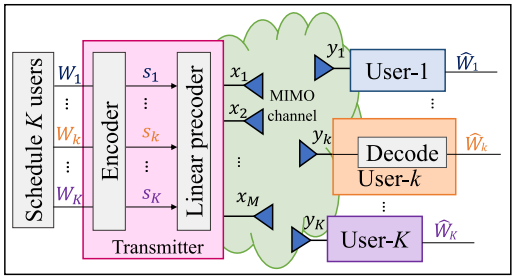


FIGURE 5. *K*-user system architecture with MU–LP. Receiver architecture is illustrated for user-*k* though the same applies to other users, i.e., all *K* users are equipped with a decoder that maps the received signal into an estimated message by treating residual interference as noise.

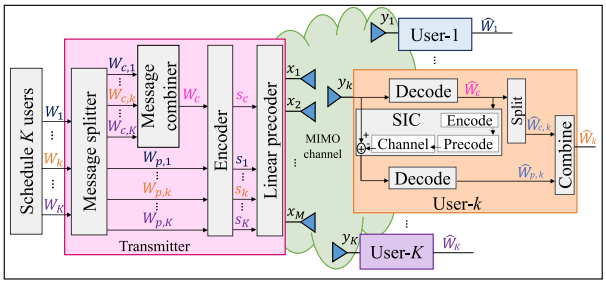


FIGURE 6. *K*-user system architecture with 1-layer rate-splitting. Receiver architecture is illustrated for user-*k* though the same applies to other users.

The sum-rate of MU–LP is therefore $R_{\mathrm{s}}^{(\mathrm{M})} = \sum_{k=1}^{K} R_k^{(\mathrm{M})}$, and the MMF rate of MU–LP is given as $R_{\mathrm{mmf}}^{(\mathrm{M})} = \min_{k=1,\ldots,K} R_k^{(\mathrm{M})}$.

B. MULTIPLEXING GAINS WITH PERFECT CSIT

We recall the sum multiplexing gain and the MMF multiplexing gain of MU–LP with perfect CSIT from [56] and [59], respectively.

Proposition 5: The sum multiplexing gain of K-user MU–LP with M transmit antennas and perfect CSIT is $d_{\mathrm{s}}^{(\mathrm{M})} = \min(M, K)$.

This result[19] is simply achieved by choosing the MU–LP precoders based on ZFBF and transmitting $\min(M, K)$ interference-free streams. Note that $\min(M, K)$ is also the optimal[20] sum multiplexing gain of the K-user MISO BC[21] [59]. In other words, $d_{\mathrm{s}}^{(\mathrm{M})} = d_{\mathrm{s}}^{(\star)} = \min(M, K)$.

Proposition 6: The MMF multiplexing gain of the K-user MU–LP with M transmit antennas and perfect CSIT is

$$d_{\mathrm{mmf}}^{(\mathrm{M})} = \begin{cases} 1, & M \geq K, \\ 0, & M < K. \end{cases} \tag{32}$$

When $M \geq K$, ZFBF can be used to fully eliminate interference. On the other hand, for $M < K$ interference cannot be eliminated anymore and $d_{\mathrm{mmf}}^{(\mathrm{M})}$ collapses, therefore leading to a rate saturation at high SNR.

C. MULTIPLEXING GAINS WITH IMPERFECT CSIT

We use the CSIT error model introduced in Section IV-A. We recall the sum multiplexing gain and the MMF multiplexing gain of MU–LP with imperfect CSIT from [47] and [48], [69], respectively.

19. It is implicitly assumed here that the coherence block is much larger than $\min(M, K)$ such that the resource needed to estimate the channel vanishes.

20. This is easily proved by showing that an upper bound on the sum multiplexing gain is equal to $\min(M, K)$, which is the same as the lower bound achieved by MU–LP. The upper bound is obtained by noticing that enabling full cooperation among receivers does not decrease the sum multiplexing gain and leads to an effective point-to-point MIMO channel with M transmit and K receive antennas, which has a sum multiplexing gain of $\min(M, K)$.

21. More generally, in MIMO BC, $d_{\mathrm{s}}^{(\mathrm{M})} = d_{\mathrm{s}}^{(\star)} = \min(M, KN)$ [59].

Proposition 7: The sum multiplexing gain of the K-user MU–LP with M transmit antennas and CSIT quality $0 \leq \alpha \leq 1$ is $d_{\mathrm{s}}^{(\mathrm{M})} = \max(1, \min(M, K)\alpha)$.

This result is simply achieved by choosing the MU–LP precoders based on ZFBF and transmitting $\min(M, K)$ streams, each with power level $P^{\alpha} / \min(M, K)$. This enables each stream to reap a multiplexing gain of α and therefore a sum multiplexing gain of $\min(M, K)\alpha$. If $\min(M, K)\alpha < 1$, one can simply transmit a single stream (i.e., perform OMA) and reap a sum multiplexing gain of 1.

Comparing Propositions 5 and 7, we note that imperfect CSIT leads to a reduction of the sum multiplexing gain. For $\alpha = 1$ (perfect CSIT in a multiplexing gain sense), Proposition 7 matches Proposition 5. Importantly, in contrast to the K-user MISO BC with perfect CSIT setting where MU–LP achieves the information theoretic optimal sum multiplexing gain $d_{\mathrm{s}}^{(\mathrm{M})} = d_{\mathrm{s}}^{(\star)}$, in the imperfect CSIT setting, MU–LP does not achieve the information theoretic optimal sum multiplexing gain [47], [63].

Proposition 8: The MMF multiplexing gain of the K-user MU–LP with M transmit antennas and CSIT quality $0 \leq \alpha \leq 1$ is

$$d_{\mathrm{mmf}}^{(\mathrm{M})} = \begin{cases} \alpha, & M \geq K, \\ 0, & M < K. \end{cases} \tag{33}$$

This is achieved by performing ZFBF when $M \geq K$. When $M < K$, rate saturation occurs (similarly to the perfect CSIT setting).

VII. BASELINE SCHEME II: RATE-SPLITTING

The second baseline to assess multi-antenna NOMA performance is Rate-Splitting Multiple Access (RSMA), based on multi-antenna Rate-Splitting (RS), for the multi-antenna BC [45]–[51]. This approach leverages and extends the concept of RS, originally developed in [70] for the two-user single-antenna interference channel, to design multi-antenna non-orthogonal transmission strategies for the multi-antenna BC.

A. RATE-SPLITTING SYSTEM MODEL

We consider again a MISO BC consisting of one transmitter with M antennas and K single-antenna users. As per Fig. 6, the architecture relies on rate-splitting of messages W_1 to

W_K intended for user-1 to user-K, respectively. To that end, message W_k of user-k is split into a common part $W_{c,k}$ and a private part $W_{p,k}$. The common parts $W_{c,1}, \ldots, W_{c,K}$ of all users are combined into the common message W_c, which is encoded into the common stream s_c using a codebook shared by all users. Hence, s_c is a common stream required to be decoded by all users and contains parts of messages W_1 to W_K intended for user-1 to user-K, respectively. The private parts $W_{p,1}, \ldots, W_{p,K}$, respectively containing the remaining parts of messages W_1 to W_K, are independently encoded into private stream s_1 for user-1 to s_K for user-K. From the K messages, $K+1$ streams $s_c, s_1, \ldots, s_K$ are therefore created. The streams are linearly precoded such that the transmit signal is given by

$$\mathbf{x} = \mathbf{p}_c s_c + \sum_{k=1}^{K} \mathbf{p}_k s_k. \tag{34}$$

Defining $\mathbf{s} = [s_c, s_1, \ldots, s_K]^T$ and assuming that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}$, the average transmit power constraint is written as $P_c + \sum_{k=1}^{K} P_k \leq P$, where $P_c = \|\mathbf{p}_c\|^2$ and $P_k = \|\mathbf{p}_k\|^2$.

At each user-k, the common stream s_c is first decoded into $\widehat{W}_c$ by treating the interference from the private streams as noise. Using SIC, $\widehat{W}_c$ is re-encoded, precoded, and subtracted from the received signal, such that user-k can decode its private stream s_k into $\widehat{W}_{p,k}$ by treating the remaining interference from the other private stream as noise. User-k reconstructs the original message by extracting $\widehat{W}_{c,k}$ from $\widehat{W}_c$, and combining $\widehat{W}_{c,k}$ with $\widehat{W}_{p,k}$ into $\widehat{W}_k$. Assuming proper Gaussian signaling, the rate of the common stream is given by

$$R_c = \min_{k=1,\ldots,K} \log_2\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{p}_c\right|^2}{1 + \sum_{q=1}^{K} \left|\mathbf{h}_k^H \mathbf{p}_q\right|^2}\right). \tag{35}$$

Assuming perfect SIC, the rates of the private streams are obtained as

$$R_k = \log_2\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{p}_k\right|^2}{1 + \sum_{q\neq k} \left|\mathbf{h}_k^H \mathbf{p}_q\right|^2}\right). \tag{36}$$

The rate of user-k is given by $R_k + R_{c,k}$ where $R_{c,k}$ is the rate of the common part of the kth user's message, i.e., $W_{c,k}$, and satisfies $\sum_{k=1}^{K} R_{c,k} = R_c$. The sum-rate is therefore simply written as $R_s^{(R)} = \sum_{k=1}^{K}(R_k + R_{c,k}) = R_c + \sum_{k=1}^{K} R_k$, and the MMF rate is written as $R_{mmf}^{(R)} = \min_{k=1,\ldots,K} R_k + R_{c,k}$.

The above RS architecture is called 1-layer RS since it only relies on a single common stream and a single layer of SIC at each user as illustrated in Fig. 6.

B. MULTIPLEXING GAINS WITH PERFECT CSIT

We here summarize the sum and MMF multiplexing gains achieved by 1-layer RS with perfect CSIT.

Proposition 9: The sum multiplexing gain of K-user 1-layer RS with M transmit antennas and perfect CSIT is $d_s^{(R)} = \min(M, K)$.

Proof: Since MU–LP is a subscheme of 1-layer RS,[22] it is sufficient[23] to design the private precoders using ZFBF and allocate zero power to the common stream at high SNR. Note that $d_s^{(R)} = d_s^{(M)} = d_s^{(\star)} = \min(M, K)$. ■

Proposition 10: The MMF multiplexing gain of the K-user 1-layer RS with M transmit antennas and perfect CSIT is

$$d_{mmf}^{(R)} = \begin{cases} 1, & M \geq K \\ \frac{1}{1+K-M}, & M < K. \end{cases} \tag{37}$$

The MMF multiplexing gain of 1-layer RS was derived and proved in [56],[24] under the same assumption as in Remark 4. Readers are referred to [56] for more details of the proof of Proposition 10.

C. MULTIPLEXING GAINS WITH IMPERFECT CSIT

Again, we use the CSIT error model introduced in Section IV-A. We recall the sum multiplexing gain of RS with imperfect CSIT from [47].

Proposition 11: The sum multiplexing gain of K-user 1-layer RS with M transmit antennas and CSIT quality $0 \leq \alpha \leq 1$ is $d_s^{(R)} = 1 + (\min(M, K) - 1)\alpha$.

Achievability of $d_s^{(R)}$ in Proposition 11 is obtained by using random precoding to design $\mathbf{p}_c$ with power level $P_c = O(P)$, transmitting $\min(M, K)$ private streams and using ZFBF to design the precoders of those $\min(M, K)$ private streams, each with power level $P_k = O(P^\alpha)$. From the SINR expressions at the right-hand side of (35), it follows that the received SINR of the common stream at each user scales as $O(P^{1-\alpha})$, leading to the multiplexing gain of $1 - \alpha$ achieved by the common stream s_c. By performing ZFBF, the transmitter transmits $\min(M, K)$ interference-free private streams. The received SINR of each private stream scales as $O(P^\alpha)$ leading to multiplexing gain α. Hence, we obtain the sum multiplexing gain of $1 + (\min(M, K) - 1)\alpha$.

Importantly, for the underloaded regime $M \geq K$, 1-layer RS achieves the information theoretic optimal sum multiplexing gain $d_s^{(M)} = d_s^{(\star)}$ in the imperfect CSIT setting [47], [63]. Hence, 1-layer RS attains the optimal sum multiplexing gain for both perfect CSIT and imperfect CSIT (underloaded regime). Actually, for $M \geq K$, 1-layer RS is optimal, achieving the maximum multiplexing gain region

22. By allocating no power to the common stream, 1-layer RS boils down to MU-LP.

23. More complicated precoders for both the common and private streams can be used to enhance the rate performance, but the multiplexing gain will not improve.

24. The MMF multiplexing gain derived in [56] considers a more complex scenario involving the simultaneous transmission of distinct messages to multiple multicast groups (each message is intended for a group of users), known as multigroup multicasting. By considering the special case where there is a single user per group, we obtain the MMF multiplexing gain of 1-layer RS in this section.

of the underloaded K-user MISO BC[25] with imperfect CSIT [71], [72].

This optimality of RS (including 1-layer RS), shown through multiplexing gain analysis, is very significant since it implies that one cannot find any other scheme achieving a better multiplexing gain region in the multi-antenna BC. As a consequence of this optimality, MU–LP and multi-antenna NOMA will always incur a multiplexing gain loss or at best will achieve the same multiplexing gain as RS for both perfect and imperfect CSIT.

Proposition 12: The MMF multiplexing gain of K-user 1-layer RS with M transmit antennas and CSIT quality $0 \leq \alpha \leq 1$ is

$$d_{\text{mmf}}^{(\text{R})} = \begin{cases} \frac{1+(K-1)\alpha}{K}, & M \geq K \\ \frac{1+(M-1)\alpha}{K}, & M < K \text{ and } \alpha \leq \frac{1}{1+K-M} \\ \frac{1}{1+K-M}, & M < K \text{ and } \alpha > \frac{1}{1+K-M}. \end{cases} \tag{38}$$

The MMF multiplexing gain of 1-layer RS with imperfect CSIT was derived in [69] (by considering the specific case where there is a single user per group), under the same assumption as in Remark 4. Readers are referred to [69] for more details of the proof of Proposition 12.

This highlights that when $M < K$, the CSIT quality α can be reduced to $\frac{1}{1+K-M}$ without impacting the MMF multiplexing gain of 1-layer RS.

Following our discussion of Proposition 11, when $M \geq K$, the respective multiplexing gains of the common and each private streams are $1-\alpha$ and α. The MMF multiplexing gain when $M \geq K$ is achieved by evenly sharing the common stream among users and is the sum of the evenly allocated multiplexing gain of the common stream $\frac{1-\alpha}{K}$ and the multiplexing gain of one private stream α, yielding $\frac{1+(K-1)\alpha}{K}$.

When $M < K$, the achievability is obtained by partitioning users into two subsets $\mathcal{K}_1$ and $\mathcal{K}_2$ with set sizes of $|\mathcal{K}_1| = M$ and $|\mathcal{K}_2| = K - M$. Users in $\mathcal{K}_1$ are served via the common and private streams while users in $\mathcal{K}_2$ are served using the common stream only. Random precoding and ZFBF are respectively used for the common stream and the private streams with power allocation $P_c = O(P)$ and $P_k = O(P^{\beta}), \forall k \in \mathcal{K}_1$. It may be readily shown that the respective multiplexing gains of the common stream and each private stream are given by $1-\beta$ and $\min\{\alpha, \beta\}$, respectively. By further introducing a fraction $z \in [0, 1]$ to specify the fraction of the rate of the common stream allocated to the users in the two subsets, we obtain that the respective sum multiplexing gains of the common stream for the users in $\mathcal{K}_1$ and $\mathcal{K}_2$ are $z(1-\beta)$ and $(1-z)(1-\beta)$, respectively. By equally dividing the multiplexing gain of the common stream between the users in the two subsets, the multiplexing gain of each user in $\mathcal{K}_2$ is $d_{k,2} = \frac{(1-z)(1-\beta)}{K-M}$, and the multiplexing gain of each user in $\mathcal{K}_1$ is $d_{k,1} = \min\{\alpha, \beta\} + \frac{z(1-\beta)}{M}$. The MMF multiplexing gain of the users is $\max_z \min\{d_{k,1}, d_{k,2}\}$. When $\beta = \alpha$, the optimal rate allocation factor $z^{\star}$ is obtained when $\frac{(1-z)(1-\alpha)}{K-M} = \frac{z(1-\alpha)}{M} + \alpha$. We have $z^{\star} = \frac{(1-\alpha-\alpha K+\alpha M)M}{(1-\alpha)K}$ and the optimal MMF multiplexing gain is $\frac{1+(M-1)\alpha}{K}$. As $z^{\star} \in [0, 1]$, we have $1 - \alpha - \alpha K + \alpha M \geq 0$. Hence, when $\alpha \leq \frac{1}{1+K-M}$, $d_{\text{mmf}}^{(\text{R})} = \frac{1+(M-1)\alpha}{K}$. When $\beta < \alpha$ and $z = 0$, the optimal power allocation $\beta^{\star}$ is obtained when $\frac{1-\beta}{K-M} = \beta$. We have $\beta^{\star} = \frac{1}{1+K-M}$ and the optimal MMF multiplexing gain is $\frac{1}{1+K-M}$. Hence, when $\alpha > \frac{1}{1+K-M}$, $d_{\text{mmf}}^{(\text{R})} = \frac{1}{1+K-M}$.

For $\alpha = 1$, the results in Propositions 11 and 12 boil down to the perfect CSIT results in Propositions 9 and 10, respectively.

VIII. SHORTCOMINGS AND MISCONCEPTIONS OF MULTI-ANTENNA NOMA

In this section, we first compare the multiplexing gains of multi-antenna NOMA to those of the MU–LP and 1-layer RS baselines. The sum and MMF multiplexing gains of multi-antenna NOMA, MU–LP, and 1-layer RS for both perfect and imperfect CSIT are summarized in Table 2. The objective of this section is to identify under which conditions NOMA provides performance gains/losses over the two baselines. We then use these comparisons to reveal several misconceptions and shortcomings of multi-antenna NOMA.

A. NOMA VS. BASELINE I (MU–LP)

We show in the following corollaries that MISO NOMA can achieve a performance gain over MU–LP but it may also incur a performance loss, depending on the values of M, K, G, and α.

The performance (expressed in terms of multiplexing gain) gain/loss of multi-antenna NOMA vs. MU–LP is obtained by comparing Propositions 3 and 7 (for sum multiplexing gain), and Propositions 4 and 8 (for MMF multiplexing gain), and is summarized in Corollaries 1, and 2 ($G = 1$), and 3 ($G > 1$), respectively. For the MMF multiplexing gain with imperfect CSIT, we consider $G = 1$ and $G > 1$ in two different corollaries.

Corollary 1: The sum multiplexing gain comparison between MISO NOMA and MU–LP is summarized in (39), at the bottom of the next page. MISO NOMA never achieves a sum multiplexing gain higher than MU–LP.

Corollary 1 shows that MISO NOMA can achieve a lower or the same sum multiplexing gain compared to MU–LP, but cannot outperform MU–LP.

If $\alpha = 1$ (perfect CSIT), Corollary 1 boils down to $d_{\text{s}}^{(\text{N})} < d_{\text{s}}^{(\text{M})}$ whenever $M > G$, and $d_{\text{s}}^{(\text{N})} = d_{\text{s}}^{(\text{M})}$ whenever $M \leq G$. This is instrumental as it says that the slope of the sum-rate of MISO NOMA at high SNR will be strictly lower

25. The optimality of RS is not limited to MISO BC but also extends to MIMO BC. Indeed, a more complicated form of RS is multiplexing gain region-optimal for the two-user MIMO BC with imperfect CSIT in the general case of an asymmetric number of receive antennas [73], [74]. Following [73], in the symmetric MIMO setting with $M \geq KN$, the system model of RS can be extended as in [75] to the K-user scenario using $\mathbf{x} = \mathbf{P}_{\text{c}}\mathbf{s}_{\text{c}} + \sum_{k=1}^{K} \mathbf{P}_k \mathbf{s}_k$ where $\mathbf{s}_{\text{c}}, \mathbf{s}_k \in \mathbb{C}^{N\times 1}$ are vectors of common streams and private streams, respectively. $\mathbf{P}_{\text{c}}, \mathbf{P}_k \in \mathbb{C}^{M\times N}$ are the corresponding precoding matrices. The sum multiplexing gain of RS is $N(1-\alpha) + NK\alpha$ which contrasts with that of conventional MU-MIMO/MU–LP (obtained by turning off $\mathbf{s}_{\text{c}}$) given by $NK\alpha$ and that of MIMO NOMA ($G = 1$) given by $\min(M, N)$ [75]. Further comparisons between RS and MIMO NOMA are provided in [75].

TABLE 2. Comparison of sum and MMF multiplexing gains of different strategies with perfect and imperfect CSIT.

Strategy	Sum/MMF Multiplexing Gain	Perfect CSIT	Imperfect CSIT
MISO NOMA	$d_{\mathrm{s}}^{(\mathrm{N})}$	$\min(M,G)$	$\max(1,\min(M,G)\,\alpha)$
	$d_{\mathrm{mmf}}^{(\mathrm{N})}$	$\begin{cases}\frac{1}{g}, & M \geq K-g+1\\ 0, & M < K-g+1\end{cases}$	$\begin{cases}\frac{\alpha}{g}, & G>1 \text{ and } M \geq K-g+1\\ 0, & G>1 \text{ and } M < K-g+1\\ \frac{1}{K}, & G=1\end{cases}$
MU–LP	$d_{\mathrm{s}}^{(\mathrm{M})}$	$\min(M,K)$	$\max(1,\min(M,K)\,\alpha)$
	$d_{\mathrm{mmf}}^{(\mathrm{M})}$	$\begin{cases}1, & M \geq K\\ 0, & M < K\end{cases}$	$\begin{cases}\alpha, & M \geq K\\ 0, & M < K\end{cases}$
1-layer RS	$d_{\mathrm{s}}^{(\mathrm{R})}$	$\min(M,K)$	$1+(\min(M,K)-1)\,\alpha$
	$d_{\mathrm{mmf}}^{(\mathrm{R})}$	$\begin{cases}1, & M \geq K\\ \frac{1}{1+K-M}, & M < K\end{cases}$	$\begin{cases}\frac{1+(K-1)\alpha}{K}, & M \geq K\\ \frac{1+(M-1)\alpha}{K}, & M < K \text{ and } \alpha \leq \frac{1}{1+K-M}\\ \frac{1}{1+K-M}, & M < K \text{ and } \alpha > \frac{1}{1+K-M}\end{cases}$

than that of MU–LP (i.e., the sum-rate of MISO NOMA will grow more slowly than that of MU–LP) whenever the number of transmit antennas is larger than the number of groups, and hence in this case, MU–LP is guaranteed to outperform MISO NOMA at high SNR. Consequently, in the massive MIMO regime where M grows large, MISO NOMA would achieve a sum multiplexing gain strictly lower than MU–LP (and the role of NOMA in massive MIMO is therefore questionable as highlighted in [76]). If $G = 1$ as in, e.g., [20]–[23], MISO NOMA always incurs a sum multiplexing gain loss compared to MU–LP irrespective of M (except in single-antenna systems when $M = 1$). In other words, from a sum multiplexing gain perspective, one cannot find any multi-antenna configuration at the transmitter, i.e., any value of M, that would motivate the use MISO NOMA with $G = 1$ compared to MU–LP. If $G = K/2$ as in [26]–[30], MISO NOMA incurs a sum multiplexing gain loss compared to MU–LP whenever $M > K/2$. In other words, from a sum multiplexing gain perspective, the only multi-antenna deployments for which MISO NOMA with $G = K/2$ would not incur a multiplexing gain loss (but no improvement either) over MU–LP is when $M \leq K/2$. Note that these conclusions are not limited to MISO NOMA. From Section V, we note that two-user MIMO NOMA incurs a sum multiplexing gain loss compared to two-user MU–LP whenever $M > N$. If $M \leq N$, MIMO NOMA and MU–LP achieve the same sum multiplexing gain.

If $\alpha < 1$ (imperfect CSIT), a sum multiplexing gain loss of MISO NOMA over MU–LP occurs in two different scenarios: 1) medium CSIT quality setting with $\frac{1}{\min(M,K)} < \alpha < \frac{1}{\min(M,G)}$ or 2) sufficiently large number of antennas and high CSIT quality with $M > G$ and $\alpha \geq \frac{1}{\min(M,G)}$. In other scenarios where the CSIT quality is poor $\alpha \leq \frac{1}{\min(M,K)}$ or the CSIT quality is good $\alpha \geq \frac{1}{\min(M,G)}$ but the number of transmit antennas is low $M \leq G$, MISO NOMA and MU–LP achieve the same sum multiplexing gains.

Corollary 2: The MMF multiplexing gain comparison between MISO NOMA with $G = 1$ and MU–LP is summarized as follows

$$d_{\mathrm{mmf}}^{(\mathrm{N})} - d_{\mathrm{mmf}}^{(\mathrm{M})} \begin{cases} < 0, & \text{if } (M \geq K) \cap \left(\alpha > \frac{1}{K}\right) \\ = 0, & \text{if } (M \geq K) \cap \left(\alpha = \frac{1}{K}\right) \\ > 0, & \text{if } (M < K) \cup \left((M \geq K) \cap \left(\alpha < \frac{1}{K}\right)\right). \end{cases} \tag{40}$$

Corollary 3: The MMF multiplexing gain comparison between MISO NOMA with $G > 1$ and MU–LP is summarized as follows

$$d_{\mathrm{mmf}}^{(\mathrm{N})} - d_{\mathrm{mmf}}^{(\mathrm{M})} \begin{cases} < 0, & \text{if } M \geq K \\ = 0, & \text{if } M < K-g+1 \\ > 0, & \text{if } K > M \geq K-g+1. \end{cases} \tag{41}$$

Corollaries 2 and 3 show that MISO NOMA can achieve either a higher or a lower MMF multiplexing gain compared to MU–LP, depending on the values of M, G, K, and α.

If $\alpha = 1$ (perfect CSIT), with $G = 1$ as in, e.g., [20]–[23], $d_{\mathrm{mmf}}^{(\mathrm{N})} > d_{\mathrm{mmf}}^{(\mathrm{M})}$ whenever $M < K$, and incurs an MMF multiplexing loss otherwise ($M \geq K$). With $G = K/2$ as in [26]–[30], $d_{\mathrm{mmf}}^{(\mathrm{N})} < d_{\mathrm{mmf}}^{(\mathrm{M})}$ whenever $M \geq K$, and $d_{\mathrm{mmf}}^{(\mathrm{N})} > d_{\mathrm{mmf}}^{(\mathrm{M})}$ whenever $K > M \geq K-1$, and $d_{\mathrm{mmf}}^{(\mathrm{N})} = d_{\mathrm{mmf}}^{(\mathrm{M})}$

$$d_{\mathrm{s}}^{(\mathrm{N})} - d_{\mathrm{s}}^{(\mathrm{M})} \begin{cases} < 0, & \text{if } ([\min(M,G)\alpha < 1] \cap [\min(M,K)\alpha > 1]) \cup ([M > G] \cap [\min(M,G)\alpha \geq 1]) \\ = 0, & \text{if } (\min(M,K)\alpha \leq 1) \cup ([\min(M,G)\alpha \geq 1] \cap [M \leq G]) \end{cases} \tag{39}$$

whenever $M < K-1$. In other words, from an MMF multiplexing gain perspective, the multi-antenna deployments for which MISO NOMA with $G=1$ and $G=K/2$ can outperform or achieve the same performance as MU–LP when $M<K$.

If $\alpha < 1$ (imperfect CSIT), we note from Corollary 3, that for $G>1$, CSIT quality α does not affect the operational regimes where MISO NOMA outperforms/incurs a loss compared to MU–LP. This is different from $G=1$ where the condition for $d_{\text{mmf}}^{(\text{N})} < d_{\text{mmf}}^{(\text{M})}$ is a function of α in Corollary 2. MISO NOMA incurs an MMF multiplexing loss whenever the number of antenna and the CSIT quality are sufficiently large, i.e., $M \geq K$ and $\alpha > \frac{1}{K}$.

B. NOMA VS. BASELINE II (RS)

We show in the following corollaries that, for all M, K, α, 1-layer RS (that relies on a single SIC at each user) achieves the same or higher (sum and MMF) multiplexing gains than the best of the MISO NOMA schemes (i.e., whatever G and the number of SICs). In other words, 1-layer RS outperforms (multiplexing gain-wise) MISO NOMA and simultaneously requires fewer SICs (only one) than MISO NOMA. Hence, employing MISO NOMA over 1-layer RS can only cause a multiplexing gain loss and/or a complexity increase at the receiver.

The performance loss of MISO NOMA vs. RS is obtained by comparing Propositions 3 and 11 (for the sum multiplexing gain), and Propositions 4 and 12 (for the MMF multiplexing gain), and is summarized in Corollaries 4, and 5 ($G=1$), and 6 ($G>1$), respectively.

Corollary 4: The sum multiplexing gain comparison between MISO NOMA and 1-layer RS is summarized as follows

$$d_{\text{s}}^{(\text{N})} - d_{\text{s}}^{(\text{R})} \begin{cases} <0, & \text{if } (0<\alpha<1) \cup ([\alpha>0] \cap [M>G]) \\ =0, & \text{if } (\alpha=0) \cup ([\alpha=1] \cap [M \leq G]). \end{cases} \tag{42}$$

MISO NOMA never achieves a sum multiplexing gain higher than 1-layer RS.

If $\alpha=1$ (perfect CSIT), Corollary 4 boils down to $d_{\text{s}}^{(\text{N})} < d_{\text{s}}^{(\text{R})}$, whenever $M>G$, and $d_{\text{s}}^{(\text{N})} = d_{\text{s}}^{(\text{R})}$ whenever $M \leq G$.

Corollary 5: The MMF multiplexing gain comparison between MISO NOMA with $G=1$ and 1-layer RS is summarized as follows

$$d_{\text{mmf}}^{(\text{N})} - d_{\text{mmf}}^{(\text{R})} \begin{cases} <0, & \text{if } (\alpha>0) \cap (M>1) \\ =0, & \text{if } (\alpha=0) \cup (M=1). \end{cases} \tag{43}$$

MISO NOMA with $G=1$ never achieves an MMF multiplexing gain higher than 1-layer RS.

Corollary 6: The MMF multiplexing gain comparison between MISO NOMA with $G>1$ and 1-layer RS is summarized in (44) as shown at the bottom of the equation. MISO NOMA with $G>1$ never achieves an MMF multiplexing gain larger than 1-layer RS.

If $\alpha=1$ (perfect CSIT), Corollaries 5 and 6 simply boil down to $d_{\text{mmf}}^{(\text{N})} < d_{\text{mmf}}^{(\text{R})}$, whenever $M \neq K-g+1$, and $d_{\text{mmf}}^{(\text{N})} = d_{\text{mmf}}^{(\text{R})}$, whenever $M=K-g+1$.

We recall again from [71]–[74] that RS achieves the optimal multiplexing gain region in the multi-antenna BC with imperfect CSIT and multi-antenna NOMA (and MU–LP/MU-MIMO) will therefore always incur a multiplexing gain loss compared to RS.

C. MISCONCEPTIONS OF MULTI-ANTENNA NOMA

The comparisons with the MU–LP and 1-layer RS baselines reveal that depending on the particular setting NOMA may incur a multiplexing gain loss at the additional expense of an increased receiver complexity, as detailed in the following.

First, NOMA is an inefficient strategy to exploit the spatial dimensions. This issue could already be observed from the two-user MISO case with perfect CSIT, where NOMA limits the sum multiplexing gain to one, same as OMA, which is only half of the sum multiplexing gain obtained with MU–LP. Moreover, even when considering a fair metric such as MMF, NOMA limits the MMF multiplexing gain to $\frac{1}{2}$, which is again only half of the MMF multiplexing gain obtained by MU–LP. Similarly, in the two-user MIMO case, NOMA limits the sum multiplexing gain to $\min(M,N)$, again the same as OMA, and the MMF multiplexing gain to $\frac{\min(M,N)}{2}$, which are lower than what is achievable with MU–LP.

In the general K-user case, it is clear from Corollaries 1 and 4 that NOMA incurs a loss in sum multiplexing gain in most scenarios, and the best NOMA can achieve is the same sum multiplexing gain as the baselines in some specific configurations. NOMA with $G=1$ achieves $d_{\text{s}}^{(\text{N})}=1$ irrespectively of the number of transmit antennas M, i.e., it achieves the same sum multiplexing gain as OMA and the same as a single-antenna transmitter (hence, wasting the transmit antenna array). NOMA with $G=K/2$ achieves $d_{\text{s}}^{(\text{N})} = \min(M, K/2)$ with $\alpha=1$. On the other hand, MU–LP and 1-layer RS achieve the full sum multiplexing gain $d_{\text{s}}^{(\text{M})} = \min(M,K)$ with $\alpha=1$.

Considering the MMF multiplexing gain of the general K-user case, the situation appears to be better for NOMA. Assuming $\alpha=1$, from Corollaries 2 and 3, we observe that NOMA incurs a loss compared to MU–LP in the underloaded regime $M \geq K$ but outperforms MU–LP in the overloaded regime. In particular, NOMA with $G=1$ achieves a higher MMF multiplexing gain than NOMA with $G=K/2$ and MU–LP whenever $M<K-1$. Hence, though the receiver

$$d_{\text{mmf}}^{(\text{N})} - d_{\text{mmf}}^{(\text{R})} \begin{cases} <0, & \text{if } (M \neq K-g+1) \cup ([M=K-g+1] \cap [\alpha<1]) \\ =0, & \text{if } (M=K-g+1) \cap (\alpha=1) \end{cases} \tag{44}$$

complexity increase of NOMA does not pay off in the underloaded regime, it appears to pay off in the overloaded regime (since $G = 1$ with more SICs outperforms $G = K/2$ with fewer SICs). Nevertheless, the MMF multiplexing gain of NOMA with $G = 1$ is independent of M, suggesting again that the spatial dimensions are not properly exploited. This can indeed be seen from Corollary 5 where NOMA is consistently outperformed by 1-layer RS, i.e., the increase in MMF multiplexing gain attained by NOMA ($G = 1$) over MU–LP is actually marginal in light of the complexity increase, and is much lower than what can be achieved by 1-layer RS with just a single SIC operation. In other words, while NOMA has some merits over MU–LP in the overloaded regime, NOMA makes inefficient use of the multiple antennas, and fails to boost the MMF multiplexing gain compared to the 1-layer RS baseline.

We note that the above observations hold for both the perfect and imperfect CSIT settings. Nevertheless, it is interesting to stress that the sensitivity to the CSIT quality α differs largely between MU–LP, NOMA with $G > 1$, NOMA with $G = 1$, and 1-layer RS. Indeed the sum and MMF multiplexing gains of MU–LP, NOMA with $G > 1$, and 1-layer RS decay as α decreases, while the multiplexing gains of NOMA with $G = 1$ are not affected by α. This can be interpreted in two different ways. On the one hand, this implies that NOMA with $G = 1$ is inherently robust to CSIT imperfections since the multiplexing gains are unchanged. On the other hand, this means that NOMA with $G = 1$ is unable to exploit the available CSIT since the resulting multiplexing gain is the same as in the absence of CSIT ($\alpha = 0$). One can indeed see from the above Propositions and Corollaries that the sum and MMF multiplexing gains for 1-layer RS with imperfect CSIT are clearly larger than those of MU–LP and NOMA. In other words, NOMA and MU–LP are inefficient in fully exploiting the available CSIT in multi-antenna settings.

We conclude from the theoretical results and above discussions that NOMA fails to efficiently exploit the multiplexing gain of the multi-antenna BC and is an inefficient strategy to exploit the spatial dimensions and the available CSIT, especially compared to the 1-layer RS baseline. *The first misconception behind NOMA is to believe that because NOMA is capacity achieving in the single-antenna BC, NOMA is an efficient strategy for multi-antenna settings. As a consequence, the single-antenna NOMA principle has been applied to multi-antenna settings without recognizing that such a strategy would waste the primary benefit of using multiple antennas, namely the capability of transmitting multiple interference-free streams.* In contrast to NOMA, other non-orthogonal transmission strategies such as 1-layer RS do not lead to any sum multiplexing gain loss. On the contrary, 1-layer RS achieves the information theoretic optimal sum multiplexing gain in both perfect and imperfect CSIT scenarios (and therefore has the capability of transmitting the optimal number of interference-free streams). 1-layer RS also achieves higher MMF multiplexing gains than NOMA and MU–LP.

Second, the multiplexing gain loss of NOMA is encountered despite the increased receiver complexity.[26] In the two-user MISO BC with perfect CSIT, MU–LP does not require any SIC receiver to achieve the optimal sum multiplexing gain of two (assuming $M > 1$) and an MMF multiplexing gain of one, while NOMA requires one SIC and only provides half the (sum and MMF) multiplexing gains of MU–LP. This is surprising since one would expect a performance gain from an increased architecture complexity. Here instead, NOMA causes a complexity increase at the receivers and a (sum and MMF) multiplexing gain loss compared to MU–LP, therefore highlighting that the SIC receiver is inefficiently exploited.

This inefficient use of SIC in NOMA also persists in the general K-user scenario. Recall that NOMA with G groups requires $g-1$ layers of SIC at the receivers. Among the two popular NOMA architectures $G = 1$ and $G = K/2$, the former requires an even higher number of SIC layers than the latter (namely $K-1$ for $G = 1$ and 1 for $G = K/2$) and has an even lower sum multiplexing gain ($d_{\mathrm{s}}^{(\mathrm{N})} = 1$ for $G = 1$ and $d_{\mathrm{s}}^{(\mathrm{N})} = \min(M, K/2)$ for $G = K/2$ with $\alpha = 1$). On the other hand, MU–LP achieves the full sum multiplexing gain $d_{\mathrm{s}}^{(\mathrm{M})} = \min(M, K)$ with $\alpha = 1$ without any need for SIC. This highlights the inefficient (and detrimental) use of SIC receivers in NOMA: the higher the number of SICs, the lower the sum multiplexing gain!

Comparing to the 1-layer RS baseline further highlights the inefficient use of SIC in NOMA. We note that 1-layer RS causes a complexity increase at the receivers (due to the one SIC needed) but also an increase in the (sum and MMF) multiplexing gains compared to MU–LP (i.e., it is easy to see from Propositions 7, 8, 11, and 12 that the sum and MMF multiplexing gains with RS are always either identical to or higher than those with MU–LP). Hence, in contrast to NOMA, the SIC in 1-layer RS is beneficial since it boosts the (sum and MMF) multiplexing gains and therefore introduces a performance gain compared to (or at least maintains the same performance as) MU–LP. Actually, 1-layer RS achieves the information theoretic optimal sum multiplexing gain for imperfect CSIT, and does so with a single SIC per user. This shows that to achieve the information theoretic optimality, it is sufficient to use a single SIC per user.[27] This is in contrast to NOMA whose sum multiplexing gain is far from optimal and for which the sum multiplexing gain decreases as the number of SICs increases. The inefficient use of SIC

26. Note that the hardware cost is the same for all schemes since we assume conventional digital processing with M antennas and M RF chains. The computational cost (digital processing) on the other hand is primarily related to the receiver complexity and is measured by the number of SIC layers.

27. Actually, though the analysis here is limited to 1-layer RS, all RS schemes (from 1-layer to generalized RS) in [51] guarantee the optimal sum multiplexing gain and a higher MMF multiplexing gain than MU–LP and NOMA, and provide an improved rate performance as the number of SIC increases [50], [51], [78].

in NOMA is also obvious from the MMF multiplexing gain. Indeed, from Propositions 2 and 10 and Corollary 5, the single SIC in 1-layer RS achieves a much larger MMF multiplexing gain than the $K-1$ layers of SIC needed for NOMA with $G=1$. This again illustrates how inefficient the use of SIC in NOMA often is. It also shows that there exists a non-orthogonal transmission strategy based on RS with better performance and lower receiver complexity requiring just a single SIC per user.

We conclude from the theoretical analysis and above discussion that NOMA often does not make efficient use of the SIC receivers compared to the considered baselines. *The second misconception regarding multi-antenna NOMA is to believe that adopting SIC receivers always boosts the rate since the interference is fully cancelled at the receiver.* Considering the two-user toy example, and comparing (2) and (14), the interference power term $|\mathbf{h}_1^H\mathbf{p}_2|^2$ appearing in the SINR of user-1 in the MU–LP rate has indeed disappeared in NOMA thanks to the SIC receiver, such that $R_{\mathrm{M},1} \leq R_{\mathrm{N},1}$. However, this comes at the cost of a reduced rate for user-2 since $R_{\mathrm{N},2} = \min(\log_2(1+A), R_{\mathrm{M},2}) \leq R_{\mathrm{M},2}$. In other words, for a given pair of precoders $\mathbf{p}_1$ and $\mathbf{p}_2$, NOMA increases the rate (or maintains the same rate) of user-1 but decreases the rate (or maintains the same rate) of user-2 compared to MU–LP.

Third, reflecting on the above two misconceptions, the NOMA design philosophy does not leverage the extensive research in multi-user MIMO, which has been fundamental to 4G and 5G in achieving the optimal sum multiplexing gain of the multi-antenna BC with perfect CSIT and low-complexity transmitter and receiver architectures. *The third misconception behind multi-antenna NOMA is to believe that, since NOMA is routinely compared to OMA in SISO BC, it is also sufficient to compare NOMA to OMA in multi-antenna settings to demonstrate its merits*. In fact, the Corollaries in Sections VIII-A and VIII-B show that NOMA is far from being an efficient strategy if NOMA is compared to alternative baselines. Unfortunately, simply comparing with OMA has led the NOMA literature to the misleading conclusion that multi-antenna NOMA is an efficient strategy. It should therefore be stressed that comparing NOMA to OMA does not demonstrate the merits of NOMA in multi-antenna settings and most importantly, the baseline for any multi-antenna NOMA design, optimization, and evaluation should be MU–LP and RS, not simply OMA[28]! In contrast to MISO NOMA, the gain of 1-layer RS over MU–LP is guaranteed, i.e., the rate of 1-layer RS is equal to or higher than that of MU–LP, since MU–LP is a particular instance of RS when no power is allocated to the common stream.

Fourth, the SISO BC is naturally overloaded (more users than the number of transmit antennas, namely one), and NOMA was therefore concluded to be suitable for overloaded scenarios. *The fourth misconception behind multi-antenna NOMA is to believe that MISO NOMA is an efficient strategy for overloaded regimes, namely whenever $K > M$.* The Corollaries in Sections VIII-A and VIII-B nevertheless expose that this is incorrect. It is clear that NOMA incurs a sum multiplexing gain erosion compared to MU–LP and 1-layer RS whenever $M > G$. Such a loss can occur also in the overloaded regime, namely whenever we have $K > M > G$. Moreover, NOMA incurs an MMF multiplexing gain loss compared to 1-layer RS whenever $M \neq K-g+1$. Here again, such a loss occurs also in the overloaded regime. In contrast to NOMA (and MU–LP), 1-layer RS is an efficient strategy for both the underloaded and overloaded regimes. Though NOMA with $G=1$ was shown in Proposition 2 to achieve a non-vanishing MMF multiplexing gain of $1/K$ in the overloaded regime, this MMF multiplexing gain is considerably smaller than that of 1-layer RS, therefore highlighting the inefficiency of NOMA in the overloaded regime. In particular, we note that the MMF multiplexing gain of 1-layer RS increases with M in contrast to that of NOMA with $G=1$ which is constant regardless of M.

TABLE 3. Sum multiplexing gain with $K=6$ - perfect CSIT.

M	regime	$d_{\mathrm{s}}^{(\mathrm{N})}$ $(G=1)$	$d_{\mathrm{s}}^{(\mathrm{N})}$ $(G=3)$	$d_{\mathrm{s}}^{(\mathrm{M})}, d_{\mathrm{s}}^{(\star)}, d_{\mathrm{s}}^{(\mathrm{R})}$
1	O	1	1	1
2	O	1	2	2
3	O	1	3	3
4	O	1	3	4
5	O	1	3	5
≥ 6	U	1	3	6

O: Overloaded ($K > M$), U: Underloaded ($K \leq M$)

D. ILLUSTRATION OF THE MISCONCEPTIONS WITH AN EXAMPLE

To illustrate the above discussion and make the statements more explicit based on numbers, we consider a MISO BC with $K=6$, and compare in Table 3 the sum multiplexing gains $d_{\mathrm{s}}^{(\mathrm{N})}$ of NOMA with $G=1$ and $G=3$ and the sum multiplexing gain of MU–LP $d_{\mathrm{s}}^{(\mathrm{M})}$ and 1-layer RS $d_{\mathrm{s}}^{(\mathrm{R})}$ (recall that $d_{\mathrm{s}}^{(\mathrm{M})} = d_{\mathrm{s}}^{(\mathrm{R})} = d_{\mathrm{s}}^{(\star)}$) as a function of M for perfect CSIT. We observe that NOMA incurs a sum multiplexing gain reduction (highlighted in red in Table 3) in the underloaded regime but also in the overloaded regime depending on the values of M and G. Specifically, in this example with $K=6$, $G=1$ incurs a sum multiplexing erosion compared to MU–LP and 1-layer RS whenever $M \geq 2$ and $G=3$ whenever $M \geq 4$. This shows that in an overloaded regime associated with $M < K$, although M is the limiting factor of the sum multiplexing gain in MU–LP and 1-layer RS, $\min(M,G)$ is the limiting factor in NOMA. Morever, Table 3 clearly illustrates that the higher the number of SICs in NOMA, the lower the sum multiplexing gain. NOMA with $G=1$ requires 5 layers of SIC to achieve a multiplexing gain $d_{\mathrm{s}}^{(\mathrm{N})}=1$, NOMA with $G=3$ requires 1 layer of SIC and achieves at most $d_{\mathrm{s}}^{(\mathrm{N})}=3$. On the other hand, MU–LP does not require any SIC and achieves the optimal sum

28. Recall also 4G and 5G are both based on MU–LP, and not simply on OMA.

TABLE 4. MMF multiplexing gain with $K = 6$ - perfect CSIT.

M	regime	$d_{\text{mmf}}^{(\text{N})}(G=1)$	$d_{\text{mmf}}^{(\text{N})}(G=3)$	$d_{\text{mmf}}^{(\text{M})}$	$d_{\text{mmf}}^{(\text{R})}$
1	O	1/6	0	0	1/6
2	O	1/6	0	0	1/5
3	O	1/6	0	0	1/4
4	O	1/6	0	0	1/3
5	O	1/6	1/2	0	1/2
≥ 6	U	1/6	1/2	1	1

O: Overloaded ($K > M$), U: Underloaded ($K \leq M$)

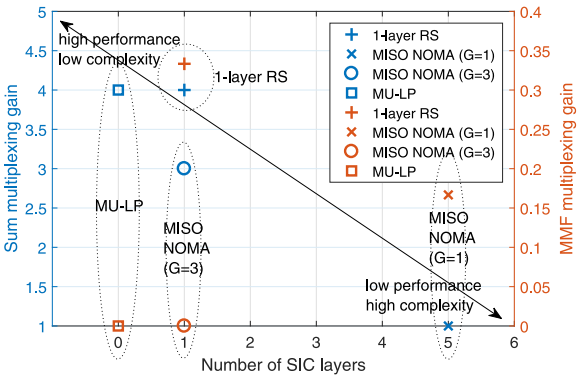


FIGURE 7. Multiplexing gains with single-antenna receivers and perfect CSIT vs. number of SIC layers for $M = 4$, $K = 6$.

multiplexing gain $d_{\text{s}}^{(\star)}$ (that can be as high as 6). 1-layer RS achieves the same (and optimal) sum multiplexing gain as MU–LP.

Table 4 highlights the MMF multiplexing gains of NOMA, MU–LP, and 1-layer RS for $K = 6$ with perfect CSIT and stresses the significant benefit of 1-layer RS over NOMA and MU–LP. The entries highlighted in red relate to configurations for which 1-layer RS provides a multiplexing gain strictly higher than that of NOMA and MU–LP. Recall that 1-layer RS provides these multiplexing gains with a single SIC per user!

In Fig. 7, we further illustrate the tradeoff between the multiplexing gains and the number of SIC layers for $M = 4$, $K = 6$ and perfect CSIT. We observe that 1-layer RS enables higher performance and lower receiver complexity compared to NOMA, stressing that the non-orthogonal transmission enabled by RS is much more efficient than NOMA. We see that NOMA with different G is suited for very different settings in this $M = 4$, $K = 6$ configuration, namely NOMA with $G = 3$ performs better in terms of sum multiplexing gain, whereas NOMA with $G = 1$ achieves a higher MMF multiplexing gain. The baseline 1-layer RS achieves a higher performance for both metrics and entails a lower receiver complexity.[29]

Though the above example was provided for perfect CSIT ($\alpha = 1$), it is easy to calculate from the above propositions

29. The reader is also invited to consult [51] for some more discussions on the complexity of RSMA, NOMA, and MU-LP.

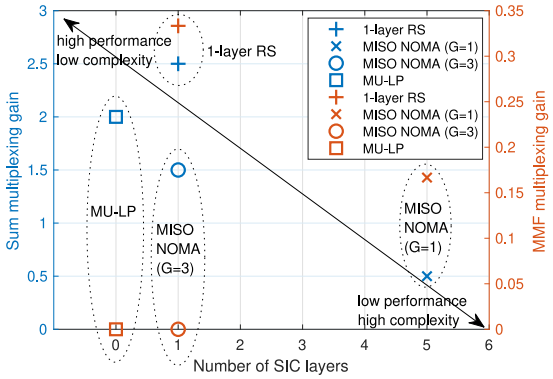
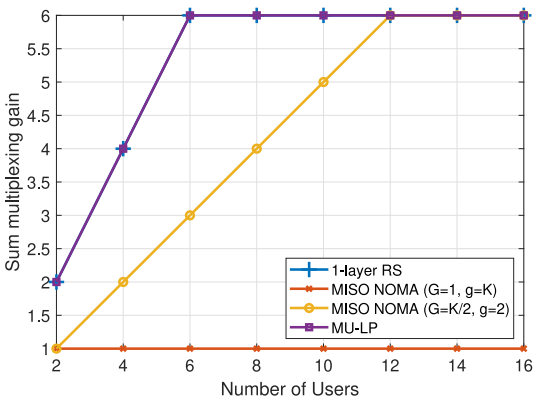
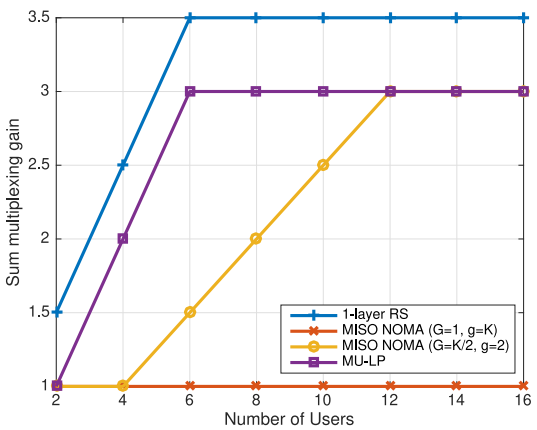


FIGURE 8. Multiplexing gains with single-antenna receivers and imperfect CSIT vs. number of SIC layers for $M = 4$, $K = 6$, $\alpha = 0.5$.



(a) Perfect CSIT

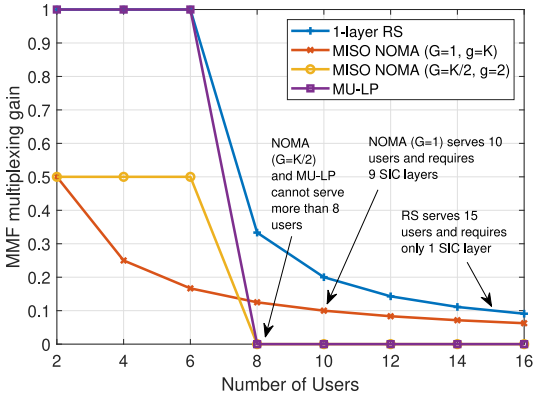


(b) Imperfect CSIT ($\alpha = 0.5$)

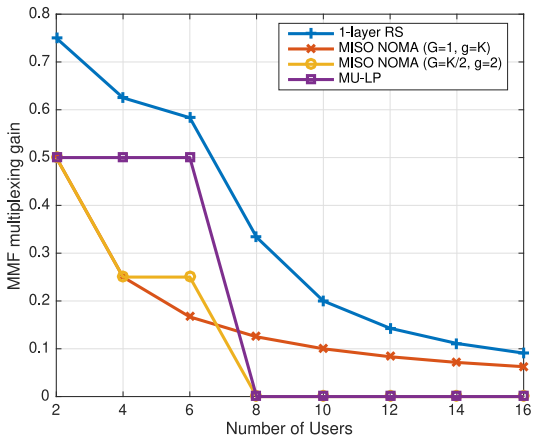
FIGURE 9. Sum multiplexing gain vs. number of users K for $M = 6$.

the multiplexing gains for the imperfect CSIT setting for a given CSIT quality α. For imperfect CSIT, the strict superiority of 1-layer RS over MU–LP and NOMA will become much more apparent, as illustrated in Fig. 8 for $\alpha = 0.5$.

In Figs. 9 and 10, the sum and MMF multiplexing gains are illustrated for $M = 6$ when we vary the number of users K under the assumption of perfect and imperfect CSIT.



(a) Perfect CSIT



(b) Imperfect CSIT ($\alpha = 0.5$)

FIGURE 10. MMF multiplexing gain vs. number of users K for $M = 6$.

Results here again confirm that NOMA achieves a lower sum multiplexing gain than MU-LP and 1-layer RS and a lower MMF multiplexing gain than 1-layer RS.

Recall that the MMF multiplexing gain reflects how fast the minimum rate among all K users increases with SNR. A zero MMF multiplexing gain means that the rate of the worst user does not scale with the SNR, which is something to avoid if one wants to simultaneously serve many users and maintain fairness and QoS among users. Hence, we can also interpret the results in Fig. 10 differently in terms of the number of users that a given strategy can serve while maintaining a target MMF multiplexing gain (and hence a certain QoS). We note from Fig. 10 that NOMA ($G = 1$) with its $K - 1$ SIC layers is more suitable than NOMA ($G = K/2$) and MU-LP to serve a large number of users when $K > M$. Indeed for $K \geq 8$, the MMF multiplexing gains of NOMA ($G = K/2$) and MU-LP collapse (are equal to 0), while that of NOMA ($G = 1$) is strictly positive. However, it is still outperformed by 1-layer RS which can support a larger number of users than any other strategy (and any combination thereof) despite using one single SIC layer. Indeed, assuming perfect CSIT and taking for instance a target MMF multiplexing gain of 0.1, NOMA ($G = 1$) can serve at most 10 users by using 9 SIC layers while 1-layer RS can serve 15 users with just 1 SIC layer. This can be indeed inferred from Table 2. Indeed, considering perfect CSIT and a target MMF multiplexing gain d_{mmf}, NOMA ($G = 1$) can serve $K = 1/d_{\text{mmf}}$ users while 1-layer RS can serve $K = M - 1 + 1/d_{\text{mmf}}$ users (assuming $M < K$). Hence, 1-layer RS with one SIC layer can serve $M - 1$ extra users compared to NOMA ($G = 1$) with $K - 1$ SIC layers while guaranteeing the same MMF multiplexing gain. As the target d_{mmf} decreases and both strategies can accommodate more users, NOMA requires an increasing number of SIC layers while 1-layer RS can still operate with a single SIC layer. In conclusion, *1-layer RS is significantly more efficient than NOMA since RS with only one SIC layer can support a larger number of users than NOMA with many SIC layers*. This demonstrates the inefficiency of NOMA to support a large number of users.

E. SHORTCOMINGS OF MULTI-ANTENNA NOMA

The previous subsections have highlighted that comparing multi-antenna NOMA to MU–LP and 1-layer RS, instead of OMA, provides a completely different picture of the actual merits of multi-antenna NOMA. In view of the previous results highlighting the waste of multiplexing gain and the inefficient use of the SIC receivers by multi-antenna NOMA, we can ask ourselves multiple questions, which help to pinpoint the shortcomings and limitations of the multi-antenna NOMA design philosophy.

The *first question* is "*What prevents multi-antenna NOMA from reaping the multiplexing gain of the system?*" The answer lies in (5), and similarly in (20), (21), and (22). Equation (5) can be interpreted as the sum-rate of a two-user MAC with a single antenna receiver. Indeed, in (5), user-1 acts as the receiver of a two-user MAC whose effective SISO channels of both links are given by $\mathbf{h}_1^H\mathbf{p}_2$ and $\mathbf{h}_1^H\mathbf{p}_1$. Similarly, in (20), user-1 acts as the receiver of a g-user MAC whose effective SISO channels of the g links are given by $\mathbf{h}_1^H\mathbf{p}_k$ for $k = 1, \ldots, g$. Such a MAC is well known to have a sum multiplexing gain of one [8], [17]. The multiplexing gain *losses* compared to the MU–LP and 1-layer RS baselines therefore *come from forcing one user to fully decode all streams in a group*, i.e., its intended stream and the co-scheduled streams in the group. This is radically different from MU–LP where streams are encoded independently and each receiver decodes its intended stream treating any residual interference as noise. By contrast, in 1-layer RS, no user is forced to fully decode the co-scheduled streams since all private streams are encoded independently and each receiver decodes its intended private stream treating any residual interference from the other private streams as noise.

The *second question* is "*Does an increase in the number of SICs always come with a reduction in the sum multiplexing gain?*" The answer is clearly no. This anomaly is deeply rooted in the way MISO NOMA was developed by applying

TABLE 5. Messages-to-streams mapping in two-user MISO BC.

	s_1	s_2	s_c
MU–LP	W_1	W_2	–
NOMA	W_1	–	W_2
OMA	W_1	–	–
Multicasting	–	–	W_1, W_2
RS	$W_{\mathrm{p},1}$	$W_{\mathrm{p},2}$	$W_{\mathrm{c},1}, W_{\mathrm{c},2}$
	decoded by its intended user and treated as noise by the other user		decoded by both users

the single-antenna NOMA principle to multi-antenna settings. The proof of Proposition 1 indeed tells us that *the fundamental principle of NOMA consisting in forcing one user in each group to fully decode the messages of* $g-1$ *co-scheduled users is an inefficient design in multi-antenna settings* that leads to a sum multiplexing gain reduction in each group.

The *third question* is "*Are non-orthogonal transmission strategies inefficient for multi-antenna settings?*" The answer is no. As we have seen, there exist frameworks of non-orthogonal transmission strategies also relying on SIC, such as RS, that do not incur the limitations of multi-antenna NOMA and make efficient use of the non-orthogonality and SIC receivers in multi-antenna settings. The key for the design of such non-orthogonal strategies is not to fall into the trap of blindly applying the SISO NOMA principle to multi-antenna settings, and therefore constraining the strategy to always fully decode the message of other users. Non-orthogonal transmission strategies and multiple access need to be re-thought for multi-antenna settings and one such strategy is based on the multi-antenna Rate-Splitting (RS) and Rate-Splitting Multiple Access (RSMA) literature for the multi-antenna BC.

The *fourth question* is "*Since NOMA and RS both rely on SIC, is there any relationship between NOMA and RS?*" The answer is yes in a two-user setting, but not necessarily in the general K-user case as it would depend on the specifc RS scheme used. In the two-user case, 1-layer RS is a superset of MU–LP, NOMA, and multicasting, i.e., MU–LP, NOMA, and multicasting are particular instances of 1-layer RS, as shown in [77] and in Table 5 and Fig. 11. Indeed, MU–LP is obtained as a special case from 1-layer RS by allocating no power to the common stream ($P_{\mathrm{c}}=0$) such that W_k is encoded directly into s_k. No interference is decoded at the receiver using the common message, and the interference between s_1 and s_2 is fully treated as noise. NOMA is obtained by encoding W_2 entirely into s_{c} (i.e., $W_{\mathrm{c}}=W_2$) and W_1 into s_1, and turning off s_2 ($P_2=0$).[30]

30. To better relate to the system model in Section II, note that NOMA also has a common message/stream, though commonly not denoted using such terminology. Indeed, the stream of the weakest user, namely s_2 in Section II, is a common stream since it is decoded by both users. s_2 in Section II carries information, namely W_2, intended for user-2 but is decoded by both user-1 and user-2. Hence, the common message is not a message that is originally intended for all users. It is required to be decoded by all users but is not necessarily intended for all users.

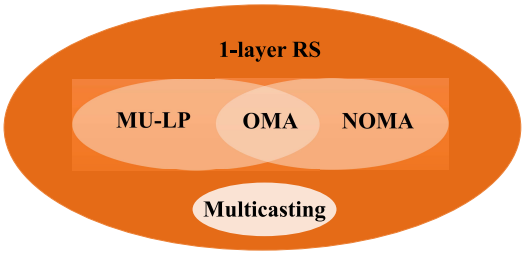


FIGURE 11. The relationship between existing strategies and 1-layer RS in two-user case. Each set illustrates the optimization space of the corresponding communication strategy. The optimization space of 1-layer RS is larger such that MU–LP, NOMA, and multicasting are just subsets.

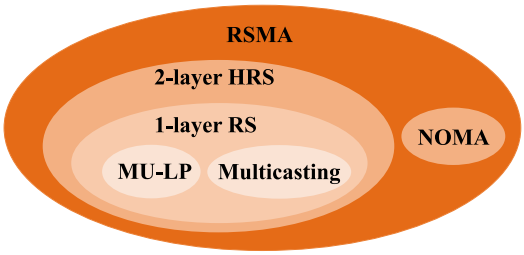


FIGURE 12. The relationship between existing strategies and the K-user RSMA framework.

In this way, user-1 fully decodes the interference created by the message of user-2. OMA is a sub-strategy of MU–LP and NOMA, which is encountered when only user-1 (with the stronger channel gain) is scheduled ($P_{\mathrm{c}}=0, P_2=0$). Multicasting is obtained when both W_1 and W_2 are entirely encoded into s_{c}. In the K-user case, 1-layer RS is a superset of MU–LP since by turning off (i.e., allocating no power to) the common stream, 1-layer RS boils down to MU–LP. On the other hand, 1-layer RS is *not* a superset of NOMA. 1-layer RS and NOMA are particular instances/schemes of the RSMA framework based on the generalized RS relying on multiple layers of SIC at each receiver [50], [51], [78], [89],[31] as illustrated in Fig. 12. As stated in the introduction, NOMA refers to communication schemes where at least one user is forced to fully decode the message(s) of other co-scheduled user(s). MU-LP and RSMA do not do that since they both do not force users to fully decode the messages of other co-scheduled users. MU-LP actually treats any residual interference as noise, and RSMA is built upon the principle of splitting the messages so as to partially treat interference as noise and partially decode the remaining interference. Consequently, RSMA is a superset of MU-LP and NOMA as per Fig. 12.

The *fifth question* is "*How does 1-layer RS achieve simultaneously higher multiplexing gains and a lower receiver*

31. 2-layer hierarchical RS (HRS) in Fig. 12 is proposed in [50] for massive MIMO. Besides one common message decoded by all users as in 1-layer RS, 2-layer HRS relies on multiple group-specific common messages being decoded by different groups of users to further manage inter-user interference. RSMA is a generalized framework that subsumes both 1-layer RS and 2-layer HRS as subschemes [51].

complexity than NOMA?" In view of the previous sections, the key is to build non-orthogonal transmission strategies upon MU–LP (and therefore SDMA/multi-user MIMO) such that the performance benefits (including sum multiplexing gain) of MU–LP are guaranteed but extra performance (e.g., in MMF multiplexing gain) is observed by the use of SIC receivers. Indeed, a performance gain over MU–LP should be expected from a more complex receiver architecture in the multi-antenna BC. To achieve this, one should enable the flexibility at the transmitter to encode messages such that parts of them can be decoded by all users using SIC while the remaining parts are decoded by their intended receivers and treated as noise by non-intended receivers. Hence, we provide the flexibility to partially decode interference and partially treat the remaining interference as noise. This contrasts with MU–LP where interference is always treated as noise, and with NOMA where interference is fully decoded. This flexibility is achieved by extending the concept of RS, originally developed in [70] for the two-user single-antenna interference channel, to the multi-antenna BC. To manage multi-user interference by partially decoding the interference and treating the remaining interference as noise, RS facilitates a complete message-to-streams mapping flexibility for each user to have part of its message transmitted in the common stream and the remaining part in one of the K private streams. By adjusting the power levels of the common and private streams, one can adjust the amount of interference caused to the private streams such that its level is weak enough to be treated as noise. This contrasts with MU–LP where the communication strategy is fundamentally constrained such that the messages are mapped to private streams only (i.e., there is no common stream, and multi-user interference between private streams is treated as noise even when its level is not weak enough to be treated as noise), and with NOMA where the constraint is that the entire message of one of the users is mapped onto a common stream (e.g., W_2 mapped to s_2 decoded by both user-1 and user-2 in Section II). These constraints imposed by MU–LP and NOMA are well illustrated by the message-to-stream mapping in Table 5 [77] and by the following example.

Example 1: To further illustrate the split of the messages and the flexibility of RS, let us consider a two-user scenario. Let us imagine that the message of user-1 $W_1 = (a_1\, a_2\, a_3\, a_4) \in \mathcal{W}_1 = \{0000, 0001, 0010, \ldots, 1111\}$, where $|\mathcal{W}_1| = 16$. Similarly, the message of user-2 is $W_2 = (b_1\, b_2\, b_3) \in \mathcal{W}_2 = \{000, 001, 010, \ldots, 111\}$, where $|\mathcal{W}_2| = 8$. In SDMA/MU–LP, W_1 would be encoded into s_1 and W_2 into s_2. In NOMA, W_1 would be encoded into s_1 and W_2 into s_c. In RS, we split user-1's message in, e.g., $W_{c,1} = (a_1\, a_2)$, $W_{p,1} = (a_3\, a_4)$, and user-2's message in, e.g., $W_{c,2} = (b_1)$, $W_{p,2} = (b_2\, b_3)$. The common message is then constructed as $W_c = (W_{c,1}\, W_{c,2}) = (a_1\, a_2\, b_1)$, which is then encoded into s_c. $W_{p,1}$ and $W_{p,2}$ are encoded into s_1 and s_2, respectively.

A consequence of the above flexibility is that by decreasing the amount of power allocated to the common stream, K-user 1-layer RS progressively converges to K-user MU–LP and in the limit where no power is allocated to the common stream, K-user 1-layer RS swiftly boils down to K-user MU–LP. Hence, 1-layer RS really builds upon MU–LP and MU–LP is a subscheme of 1-layer RS, which provides a guarantee to 1-layer RS that its rate and multiplexing gains are always the same or better than those of MU–LP. This is completely different from MISO NOMA. MISO NOMA does not build upon MU–LP. With G groups, K-user MISO NOMA can boil down to G-user MU–LP by turning off the power to the weaker users in each group, but K-user MISO NOMA can mathematically never boil down to K-user MU–LP (recall footnote 13). The rate/multiplexing gains of K-user MISO NOMA can therefore be worse than that of K-user MU–LP.

Another interpretation arises by noting that MU–LP (and other forms of multi-user MIMO), as one extreme, can be viewed as a full transmit-side interference management strategy. On the other extreme, NOMA can be seen as a full receiver-side interference cancellation strategy. In between stands RS that can be viewed as a smart combination of transmit-side and receive-side interference management/cancellation strategies where the contribution of the common stream is adjusted according to the level of interference that can be canceled by the receiver.

Consequently, RS is an enabler of a general class of communication strategies and can cover a wider set of communication strategies than SDMA and NOMA, which leads to significant multiplexing gain and complexity reduction benefits.

The *sixth question* is "*Can we use other types of receivers than SIC for NOMA and RS and would the multiplexing gains be improved?*" We can indeed use other types of receivers but the multiplexing gains will not improve. Instead of using stream-by-stream SIC, we can use any other joint (Maximum Likelihood) decoder. Hence, a strong user in NOMA could use a joint decoder to decode its intended stream jointly with all other streams intended for its co-scheduled users in the same group. The multiplexing gains would not improve since the strong user would still act as the receiver of an effective MAC (as discussed in relationship with (5), (20), (21), and (22) and the first question) which limits the multiplexing gains. Similarly, in 1-layer RS, each user could use a joint decoder to decode its private stream jointly with the common stream and the multiplexing gains would not improve (recall that 1-layer RS already achieves the information theoretic optimal multiplexing gain region, hence any other scheme, receiver or multi-layer RS would not increase the multiplexing gains any further).

The *seventh question* is "*When does it make sense to use NOMA?*" As we have seen from the multiplexing gain analysis, RS achieves the same or higher multiplexing gains than NOMA with a lower number of SIC layers. Hence, it is difficult to motivate the use of NOMA based on the above analysis. Nevertheless, recall that our analysis relies on having the concatenated matrix of the user channels being

full rank, or in other words that the user channels are not aligned, as per footnotes 11 and 16. Whenever the channels are aligned (though aligned channels are unlikely to occur in real wireless settings) and CSIT is perfect, NOMA can achieve the same performance as DPC, and could therefore be used as an alternative to DPC[32] in that outlier scenario [41], [42], [44]. This should not appear as a surprise since a multi-antenna setting with aligned channel vectors can effectively be seen as a SISO setting where users are distinguished only based on their channel strengths. In such a SISO setting (i.e., degraded BC), it is well-known that both NOMA and DPC are capacity achieving [8], [9], [17].

Once the channels are not aligned, our results show that NOMA generally incurs a multiplexing gain loss. This corroborates our previous results [77] that showed that in a 2-user MISO BC, RS always outperforms NOMA. In particular, RS was shown to boil down to NOMA and achieve the same rate performance as NOMA whenever the following conditions are met: 1) the SNR is low, 2) the channels are closely aligned, 3) there is a sufficiently large disparity of channel gains, and 4) the CSIT is perfect. In this regime, all NOMA, RS, and DPC schemes achieve very similar performance (if not the same performance). As we depart from that regime, NOMA incurs a loss over RS (and DPC) due to its inferior multiplexing gain.

IX. NUMERICAL RESULTS

Through numerical evaluation, we illustrate the misconceptions and the shortcomings of MISO NOMA. Moreover, we show that, by adopting 1-layer RS, the optimal sum multiplexing gain of the MISO BC is guaranteed in both underloaded and overloaded deployments for both perfect and imperfect CSIT scenarios. Furthermore, results also demonstrate that the MMF multiplexing gain (and MMF rate) is significantly enhanced when using 1-layer RS compared to MU–LP and MISO NOMA, and the complexity of the receivers is reduced compared to MISO NOMA. In other words, our evaluations show that 1-layer RS makes a more efficient use of the spatial dimensions (multiplexing gains) and of the SIC receivers than MISO NOMA, and it is more robust to CSIT inaccuracy.

The following two precoder optimization problems are solved in the simulation for the K-user MISO NOMA system model specified in Section III-A. The first problem is maximizing the sum-rate of MISO NOMA subject to the transmit power constraint, which is given by

$$\max_{\mathbf{P}} \ \sum_{k\in\mathcal{K}} R_k^{(\mathrm{N})} \tag{45a}$$

$$\text{s.t.} \ \ \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \leq P \tag{45b}$$

where $R_k^{(\mathrm{N})}$ is the rate of user-k in the MISO NOMA system as specified in (17)–(19). The second problem is maximizing the minimum rate subject to the transmit power constraint, which is formulated as

$$\max_{\mathbf{P}} \ \min_{k\in\mathcal{K}} R_k^{(\mathrm{N})} \tag{46a}$$

$$\text{s.t.} \ \ \mathrm{tr}(\mathbf{P}\mathbf{P}^H) \leq P. \tag{46b}$$

The Weighted Minimum Mean Square Error (WMMSE) optimization framework proposed in [80] (originally developed for MU–LP) is extended to solve both problems (45) and (46). The details of the algorithm are specified in Appendix B. The optimization problems requiring interior-point methods are solved using the CVX toolbox [79].

We will assume $K = 6$ in the simulations, so as to be able to relate the numerical results to the theoretical results of Tables 3 and 4. The channel $\mathbf{h}_k$ of user-k has i.i.d. complex Gaussian entries drawn from the distribution $\mathcal{CN}(0, \sigma_k^2)$. The presented results are averaged over 100 channel realizations.

The following five strategies are compared and analyzed for both perfect and imperfect CSIT:

- *MISO NOMA ($G = 3$):* MISO NOMA ($G = 3$) is the MISO NOMA strategy specified in Section III-A when $G = 3$. Each user requires $\frac{K}{3} - 1 = 1$ layer of SIC (since each user can be selected as the "strong user" in the corresponding user group). Ideally, the sum-rate (or max-min) rate is maximized by solving (45) (or (46)) for all possible user grouping methods and decoding orders within each group. Due to the high computational complexity of jointly optimizing the precoders, grouping, and decoding order, we assume that the user grouping is fixed[33] while the decoding order in each group i is the ascending order of users' channel strength $\|\mathbf{h}_k\|, \forall k \in \mathcal{K}_i$ in the following results. To keep aligned with the system model in Section III-A, user indices are updated within each group such that $\|\mathbf{h}_k\| \leq \|\mathbf{h}_j\|, \forall k < j$ and $k, j \in \mathcal{K}_i$. When the CSIT is imperfect, the decoding order follows the same method but based on $\|\hat{\mathbf{h}}_k\|, \forall k \in \mathcal{K}_i$.
- *MISO NOMA ($G = 1$):* MISO NOMA ($G = 1$) is the MISO NOMA strategy in Section III-A when $G = 1$. Each user requires $K - 1 = 5$ layers of SIC (since each user can potentially be selected as the "strong user"). There is no user grouping optimization issue at the transmitter since all users are assumed to be in the same user group. However, the decoding order at users should be jointly optimized with the precoders in order

32. Another detail missing and misleading in the comparison between NOMA and DPC is that the whole capacity region is achieved with DPC and time-sharing between the precoding orders [12]. In the NOMA literature [44], the optimality of NOMA is only shown with respect to one fixed precoding order in DPC. The true capacity is achieved with time-sharing between the precoding orders and is larger.

33. For a given K and G (with $g = \frac{K}{G}$), there are in total $\frac{1}{G!}\prod_{i=0}^{G-1}\binom{K-ig}{g}$ user grouping methods. When $K = 6$, the number of grouping methods for MISO NOMA ($G = 3$) is 15. To optimize the user grouping (for a fixed decoding order), the optimization problem (45) (or (46)) has to be solved 15 times. The computational complexity is 15-fold increase compared with MU–LP/1-layer RS/OMA. To consider the complexity fairness among all the studied strategies, we fix the grouping method to be user-1 and user-2 in Group 1, user-3 and user-4 in Group 2, and user-5 and user-6 in Group 3. Recall however that the multiplexing gain analysis is general and holds for any decoding order and any grouping method.

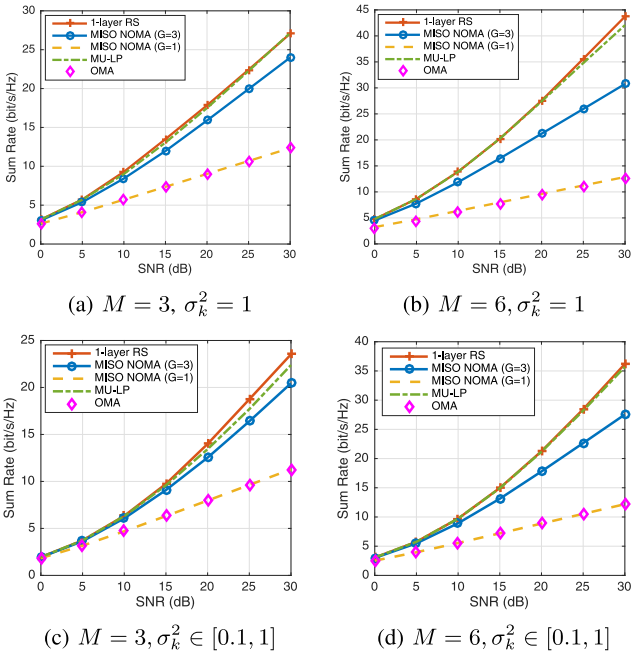
to maximize the sum-rate (or the max-min rate), which however, is computationally prohibitive. Following the literature of single-cell MISO NOMA [22], [23], we assume that the decoding order is the ascending order of the users' channel strength $\|\mathbf{h}_k\|, \forall k \in \mathcal{K}$. User indices are updated such that $\|\mathbf{h}_k\| \leq \|\mathbf{h}_j\|, \forall k < j$ and $k, j \in \mathcal{K}$. Similarly, the decoding order follows the same method but based on $\|\hat{\mathbf{h}}_k\|, \forall k \in \mathcal{K}$ when the CSIT is imperfect.

- *MU–LP:* MU–LP is the baseline strategy studied in Section VI. Each user directly decodes the intended stream by fully treating the interference as noise. The WMMSE algorithm specified in Appendix B can be applied and extended to solve the corresponding sum-rate and max-min problems of MU–LP [56], [80]. The transmitter and receiver complexity of MU–LP is low since there is no SIC deployed at each user and no user grouping and decoding order optimization issue at the transmitter.
- *Orthogonal Multiple Access (OMA):* This is the single-user transmission where only the user with the highest channel strength is served.
- *1-layer RS:* 1-layer RS is the RS strategy we specified in Section VII. The corresponding sum-rate and max-min rate maximization problems are solved by using the WMMSE algorithm proposed in [47], [56]. Compared with MISO NOMA, the transmitter and receiver complexities of 1-layer RS are much reduced. Similarly to MU–LP, no user grouping and decoding order optimization is needed. Each user only requires a single layer of SIC.

A. PERFECT CSIT

Following [47], the initialization of the precoding matrix $\mathbf{P}$ in Algorithm 1 is designed by using Maximum Ratio Transmission (MRT) combined with Singular Value Decomposition (SVD). Specifically, the precoder for the message to be decoded by a group of users is designed based on the SVD of the channel matrix formed by the channel vectors of the corresponding users while the precoder for the message to be decoded by a single user is designed based on MRT. For example, when considering MISO NOMA ($G = 3$), the message for user-$k, k \in \mathcal{K}_i$, is decoded by users-$\{j|j \leq k, j \in \mathcal{K}_i\}$. The precoders are initialized as $\mathbf{p}_k = \sqrt{p_k}\widehat{\mathbf{p}}_k$, where $\widehat{\mathbf{p}}_k$ is the largest left singular vector of the channel estimate $\mathbf{H}_k$ formed by channels $\{\mathbf{h}_j|j \leq k, j \in \mathcal{K}_i\}$. The precoder $\mathbf{p}_k$ of the stream to be decoded at last in each group is initialized as $\mathbf{p}_k = \sqrt{p_k}\frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}$, where p_k is the power allocated to the corresponding precoder $\mathbf{p}_k$ and it satisfies that $\sum_{k=1}^{K} p_k = P$.

Fig. 13 illustrates the sum-rate vs. SNR comparison of the five strategies considered when there are $K = 6$ users and the number of transmit antennas is $M = 3$ and $M = 6$. In Fig. 13(a) and Fig. 13(b), all users have equal channel variances, i.e., $\sigma_k^2 = 1, \forall k \in \mathcal{K}$ while the users' channel variances are randomly generated from $[0.1, 1]$ in Fig. 13(c) and Fig. 13(d), i.e., $\sigma_k^2 \in [0.1, 1], \forall k \in \mathcal{K}$. In other words,



(a) $M = 3, \sigma_k^2 = 1$ (b) $M = 6, \sigma_k^2 = 1$

(c) $M = 3, \sigma_k^2 \in [0.1, 1]$ (d) $M = 6, \sigma_k^2 \in [0.1, 1]$

FIGURE 13. Sum-rate vs. SNR comparison of different strategies, $K = 6$.

the average channel strength disparities among users are randomly generated between 0 and 10 dB[34] in Fig. 13(c) and Fig. 13(d). In the high SNR regime of each subfigure, the multiplexing gains of all strategies are found to match the theoretical sum multiplexing gains specified in Table 3. Specifically, when $M = 3, K = 6$, the sum multiplexing gains of 1-layer RS, MU–LP, and MISO NOMA ($G = 3$) in Fig. 13(a) and Fig. 13(c) approach $d_s^{(\star)} = 3$ (which is optimal). In Fig. 13(c) and Fig. 13(d) where $M = K = 6$, the sum multiplexing gains of 1-layer RS and MU–LP are $d_s^{(\star)} = 6$. The sum multiplexing gain of MISO NOMA ($G = 3$) remains 3. The sum multiplexing gains of MISO NOMA ($G = 1$) and OMA are limited to 1 in all subfigures of Fig. 13. Therefore, MISO NOMA has a reduced sum multiplexing gain, inefficiently makes use of the available multiple antennas, and incurs a significant rate loss, especially at medium and high SNRs. It is not an efficient strategy for multi-antenna settings. The first misconception behind multi-antenna NOMA is confirmed.

As pointed out earlier in this section, the complexity of MISO NOMA at both the transmitter and the receiver is the highest among all strategies studied in this work. At the transmitter, the scheduling complexity is high since the user grouping and decoding order are required to be jointly optimized with the precoders. At the receivers, each user

34. As a reference, at a carrier frequency of 2 GHz, the typical macro cell propagation model of [81] states that the path loss [dB] is equal to $128.1 + 37.6 \log_{10}(R)$ where R is the transmitter-receiver distance in km. Considering a macro cell deployment with an inter-site distance of 750m [81], a 0 to 10 dB channel gain disparity implies that users are distributed between, e.g., 160m to 300m or between 200m and 375m from their serving base station, i.e., a user located at 300m (375m) will experience 10dB extra path loss compared to a user at 160m (200m).

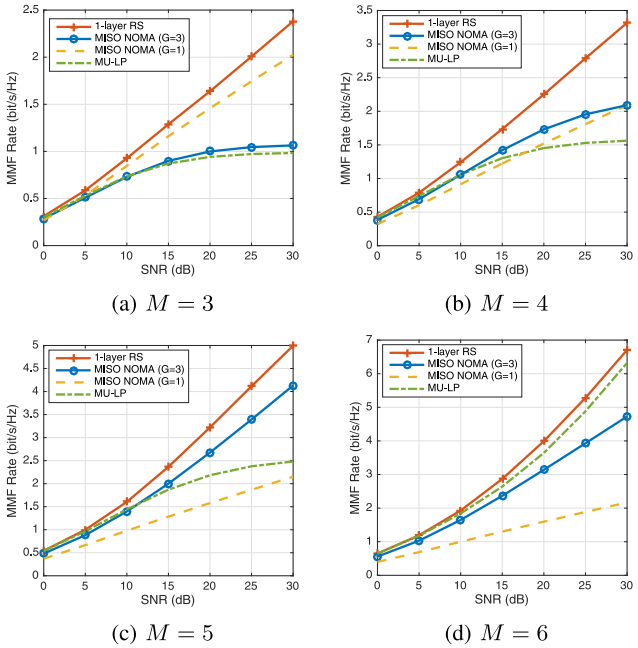


FIGURE 14. Max-min rate vs. SNR comparison of different strategies, $K = 6$, $\sigma_k^2 = 1, \forall k \in \mathcal{K}$.

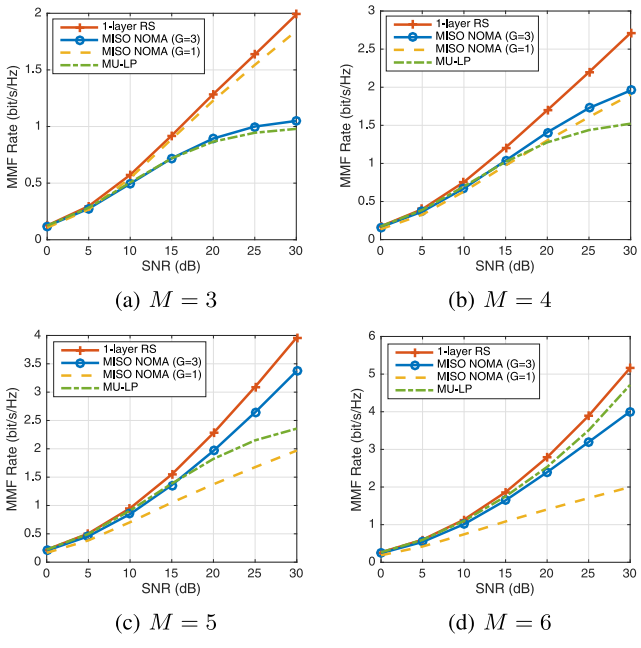


FIGURE 15. Max-min rate vs. SNR comparison of different strategies, $K = 6$, $\sigma_k^2 \in [0.1, 1], \forall k \in \mathcal{K}$.

requires multiple layers of SIC and the number of SIC layers at each user increases with the number of users K in the system. In addition to such a high complexity, as evident from Fig. 13, the sum-rate performance of MISO NOMA is worse than that of MU–LP[35] which exhibits a much lower complexity at the transmitter and each receiver. Adopting SIC receivers does not always boost the rate performance. On the contrary, an inefficient and inappropriate use of SIC as in MISO NOMA can make the rate performance worse than simply not using SIC (as in MU–LP). This illustrates the second misconception behind multi-antenna NOMA.

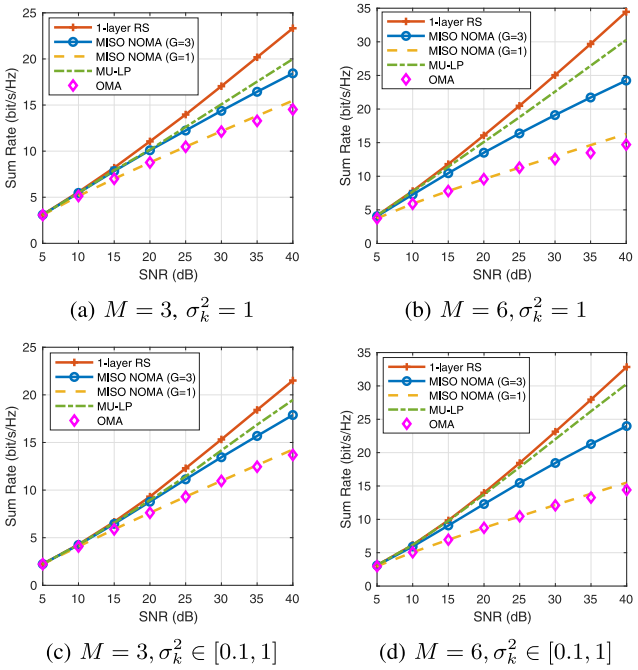
We also observe from Fig. 13 that the sum-rate performance of OMA and MISO NOMA ($G = 1$) is the worst, which is also reflected in their sum multiplexing gains. Hence, comparing MISO NOMA with OMA is not sufficient in multi-antenna settings. Both MU–LP and 1-layer RS should be considered as the baselines for all MISO NOMA schemes. This verifies the third misconception behind multi-antenna NOMA.

In Fig. 14 and Fig. 15, we focus on the MMF rate performance when there are $K = 6$ users and the number of transmit antennas is varied from $M = 3$ to $M = 6$. All users have equal channel variances in Fig. 14 while the users' channel variances are randomly generated from $[0.1, 1]$ in Fig. 15. The MMF multiplexing gains of all the strategies in both Fig. 14 and Fig. 15 match the corresponding theoretical MMF multiplexing gain results specified in Table 4. In the overloaded regime when $M = 3/4/5$, the corresponding MMF multiplexing gains of MISO NOMA ($G = 3$) and MISO NOMA ($G = 1$) are $d_{\text{mmf}}^{(\text{N,G=3})} = 0/0/\frac{1}{2}$, and $d_{\text{mmf}}^{(\text{N,G=1})} = \frac{1}{6}/\frac{1}{6}/\frac{1}{6}$, respectively. In contrast, the MMF multiplexing gain of 1-layer RS is $d_{\text{mmf}}^{(\text{R})} = \frac{1}{4}/\frac{1}{3}/\frac{1}{2}$ when $M = 3/4/5$, which is significantly higher. The low MMF multiplexing gains of the MISO NOMA strategy translates into a poor MMF rate performance as illustrated in Fig. 14 and Fig. 15. Though MISO NOMA has been promoted as a strategy to enhance user fairness and to deal with overloaded regimes, its MMF rate in the overloaded regime is actually worse than that of 1-layer RS. MISO NOMA is not an efficient strategy for overloaded regimes. This underscores the validity of the fourth misconception behind multi-antenna NOMA.

B. IMPERFECT CSIT

Let us now consider ergodic sum-rate and minimum ergodic rate maximization problems when the CSIT is imperfect. The two problems are solved by extending the WMMSE algorithm specified in Section X to the corresponding imperfect CSIT setting [47]. This is achieved by using the Sample Average Approximation (SAA) method [82] to transform the original ergodic problem to its deterministic counterpart and then using WMMSE to solve the corresponding deterministic problem. In the following results, for a given channel estimate $\hat{\mathbf{h}}_k, k \in \mathcal{K}$, $M = 1000$ channel samples are generated. The ergodic sum-rate or max-min ergodic rate is obtained by averaging over 100 channel estimates. The channel estimate $\hat{\mathbf{h}}_k$ and channel estimation error $\tilde{\mathbf{h}}_k$ have i.i.d. complex Gaussian entries respectively drawn from the distributions $\mathcal{CN}(0, \sigma_k^2 - \sigma_{e,k}^2)$, $\mathcal{CN}(0, \sigma_{e,k}^2)$, where $\sigma_{e,k}^2 = \sigma_k^2 P^{-\alpha}$. As

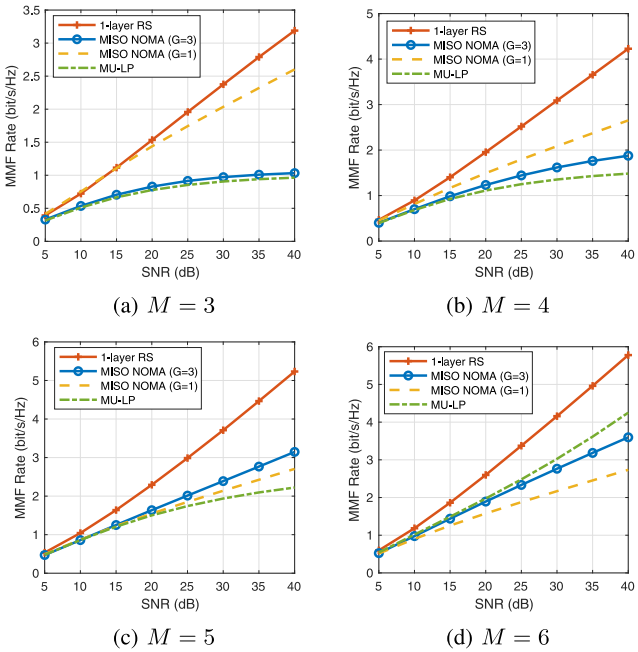
35. Though multiplexing gain analysis holds for any antenna configuration, simulations are here conducted for small MIMO systems. For larger antenna regimes, the same observation can be obtained and NOMA has an even less role to play as shown in [76] for massive MIMO.



(a) $M = 3$, $\sigma_k^2 = 1$ (b) $M = 6, \sigma_k^2 = 1$

(c) $M = 3, \sigma_k^2 \in [0.1, 1]$ (d) $M = 6, \sigma_k^2 \in [0.1, 1]$

FIGURE 16. Sum-rate vs. SNR comparison of different strategies with imperfect CSIT, $\alpha = 0.5$, $K = 6$.

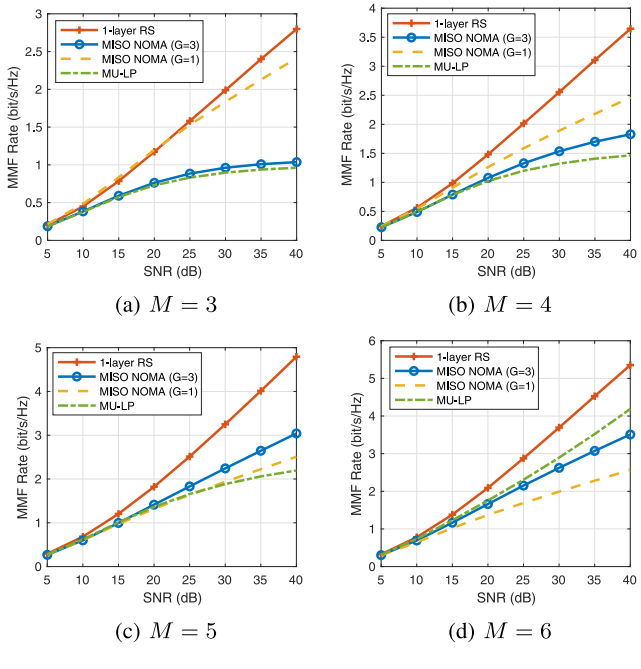


(a) $M = 3$ (b) $M = 4$

(c) $M = 5$ (d) $M = 6$

FIGURE 17. Max-min rate vs. SNR comparison of different strategies with imperfect CSIT, $\alpha = 0.5$, $K = 6$, $\sigma_k^2 = 1, \forall k \in \mathcal{K}$.

only channel estimate $\hat{\mathbf{h}}_k, k \in \mathcal{K}$, is known at the transmitter, the precoders are initialized using the same method as in the perfect CSIT scenario but based on realistic channel estimates. Figs. 16, 17, and 18 are the imperfect CSIT results corresponding to Fig. 13, 14, and 15, respectively. The unspecified parameters in this subsection remain the same as the corresponding ones used for perfect CSIT.



(a) $M = 3$ (b) $M = 4$

(c) $M = 5$ (d) $M = 6$

FIGURE 18. Max-min rate vs. SNR comparison of different strategies with imperfect CSIT, $\alpha = 0.5$, $K = 6$, $\sigma_k^2 \in [0.1, 1], \forall k \in \mathcal{K}$.

Fig. 16 illustrates the sum-rate vs. SNR comparison of the five strategies for imperfect CSIT. The sum multiplexing gains of all strategies in Fig. 16 match the theoretical sum multiplexing gains in Table 2. When $\alpha = 0.5$ and $M = 3/6$, the sum multiplexing gains of the five strategies are $d_{\mathrm{s}}^{(\mathrm{R})} = 2/3.5$ for 1-layer RS, $d_{\mathrm{s}}^{(\mathrm{M})} = 1.5/3$ for MU–LP, $d_{\mathrm{s}}^{(\mathrm{N,G=3})} = 1.5/1.5$ for MISO NOMA ($G = 3$), and $d_{\mathrm{s}}^{(\mathrm{N,G=1})} = d_{\mathrm{s}}^{(\mathrm{O})} = 1/1$ for MISO NOMA ($G = 1$) and OMA. As suggested by the multiplexing gain results, where MISO NOMA ($G = 1$) has the lowest multiplexing gain, we also observe from Fig. 16 that though MISO NOMA ($G = 1$) has the highest receiver complexity, its ergodic sum rate performance is the worst even in the preferred NOMA overloaded setting when the users have channel strength disparities. MISO NOMA ($G = 1$) always achieves a worse sum-rate than MU–LP. It is not beneficial for enhancing the sum-rate of multi-antenna scenarios regardless of whether perfect or imperfect CSIT is used. In comparison, 1-layer RS achieves explicit sum multiplexing gains and sum-rate improvement over all other strategies.

Figs. 17 and 18 illustrate the MMF ergodic rate results. In general, the MMF multiplexing gains of all strategies in both figures match the theoretical MMF multiplexing gain results specified in Table 2. When $M = 3/4/5/6$, the corresponding MMF multiplexing gains of MISO NOMA ($G = 3$) and MISO NOMA ($G = 1$) when $\alpha = 0.5$ are $d_{\mathrm{mmf}}^{(\mathrm{N,G=3})} = 0/0/\frac{1}{4}/\frac{1}{4}$ and $d_{\mathrm{mmf}}^{(\mathrm{N,G=1})} = \frac{1}{6}/\frac{1}{6}/\frac{1}{6}/\frac{1}{6}$, respectively, and the corresponding MMF multiplexing gain of MU–LP and RS are $d_{\mathrm{mmf}}^{(\mathrm{M})} = 0/0/0/\frac{1}{2}$, and $d_{\mathrm{mmf}}^{(\mathrm{R})} = \frac{1}{4}/\frac{1}{3}/\frac{1}{2}/\frac{1}{2}$. We observe that 1-layer RS achieves significantly higher multiplexing gains, which is also reflected in the MMF

ergodic rate performance in Figs. 17 and 18. In both the perfect and imperfect CSIT settings, user fairness cannot be improved by MISO NOMA. The MMF ergodic rate performance of MISO NOMA is much worse than that of 1-layer RS.

Therefore, the four misconceptions behind multi-antenna NOMA are further verified for imperfect CSIT. Higher sum-rate and MMF rate gaps between RS and MU–LP/multi-antenna NOMA are generally observed by comparing the corresponding perfect and imperfect CSIT results. By partially decoding the interference and treating the remaining interference as noise, 1-layer RS is more robust to CSIT inaccuracy. The large performance gain of RS makes it an appealing strategy for application in future communication networks.

C. DISCUSSIONS

The presented simulations fully validate the theoretical multiplexing gain analysis and confirm the inefficiency of MISO NOMA. We therefore conclude that the fundamental design principle of NOMA, namely forcing one user to decode the message(s) of other user(s), should be reconsidered or very carefully used for multi-antenna settings.

Thanks to its ability to partially decode interference and partially treat interference as noise, 1-layer RS achieves equal or higher sum-rate and MMF rate performance than all other strategies in both underloaded and overloaded regimes, especially when it comes to metrics that favor user fairness (e.g., MMF rate) in an overloaded regime. This is due to the fact that the inter-user interference becomes stronger in the setting when all users are active and the number of transmit antennas is limited. The superiority of 1-layer RS in managing multi-user interference becomes more pronounced when users suffer from stronger interference. Most importantly, 1-layer RS requires no user grouping and decoding order optimization at the transmitter and only one layer of SIC at each user. Compared with MISO NOMA, the sum-rate and MMF rate performance gain of RS comes at a much reduced transmitter and receiver complexity. 1-layer RS enables a better trade-off between the rate performance gains and the number of SIC layers. Hence, we conclude that 1-layer RS is a more powerful and promising strategy for multi-antenna networks.

Though the evaluations have been limited to 1-layer RS as the basic RSMA scheme, further rate enhancements over 1-layer RS can be obtained with multi-layer RS where the message of a user is split multiple times and multiple SIC layers are implemented at the receivers, as demonstrated in [50], [51], [78], [89], [90].

X. CONCLUSIONS, FUTURE RESEARCH, AND PATHWAYS TO 6G STANDARDIZATION

This paper provides a broad, different, and useful perspective on multi-antenna NOMA and non-orthogonal transmission to the community working on NOMA and multiple access, and to the future generations of researchers working on multi-user multi-antenna communications. Although NOMA in single-antenna settings has been well understood for a long time, the paper shows that the design of non-orthogonal transmission strategies for multi-antenna settings should be done with care so as to benefit from the multi-antenna dimensions and SIC receivers.

The paper showed in Section II that two-user multi-antenna NOMA increases the receiver complexity and at the same time incurs a loss in multiplexing gain (and therefore rate at high SNR) compared to conventional multiuser precoding (as in used in 4G and 5G), therefore raising concerns on the efficiency of multi-antenna NOMA. Subsequently, a general K-user setting with perfect CSIT and imperfect CSIT were studied in Section III and Section IV, respectively and various multiplexing gains of multi-antenna NOMA were derived. Then, we introduced two baseline schemes, namely K-user conventional multiuser precoding in Section VI and K-user multi-antenna rate-splitting in Section VII, and studied the multiplexing gains of those schemes. Section VIII compares the multiplexing gains of all considered schemes and provides strong theoretical grounds for performance comparisons among all schemes. In particular, it identifies the scenarios where NOMA incurs a gain and a loss compared to multiuser linear precoding and demonstrates how NOMA always leads to lower multiplexing gains than rate-splitting though it makes use of a larger number of SIC layers at the receivers. This section is instrumental and exposes various misconceptions and shortcomings of multi-antenna NOMA. Simulation results are then used in Section IX to confirm our findings and predictions from the multiplexing gain analysis.

Our results show that NOMA is not an efficient solution to cope with the high throughput, reliability, heterogeneity of QoS, and connectivity requirements of the downlink of future 5G and beyond multi-antenna wireless networks. This is due to the fact that the fundamental principle of NOMA consisting in forcing one user in each group to fully decode the messages of other co-scheduled users is an inefficient design in multi-antenna settings. Consequently, the benefits to the research community and future standards and networks of multi-antenna NOMA for downlink communications (e.g., MISO/MIMO techniques for NOMA, NOMA for massive MIMO and cell-free massive MIMO, multi-antenna NOMA for millimeter and terahertz communications, NOMA for multi-beam satellite communications, multi-antenna NOMA in reconfigurable intelligent surfaces, multi-antenna in Multiuser Superposition Transmission (MUST) in 3GPP, etc) are questionable and should be considered carefully in light of the results in this paper.

Instead, non-orthogonal transmission strategies for multi-antenna settings should be designed such that interference is partially decoded and partially treated as noise based on the rate-splitting (multiple access) literature so as to truly benefit from multi-antenna transmitters (and potentially multi-antenna receivers) and SIC receivers.

In this paper, we limited the multiplexing gain analysis and the numerical evaluations to two metrics, namely sum multiplexing gain/sum-rate and MMF multiplexing gain/MMF rate, and to the MISO BC. Nevertheless, the results can be extended to other metrics such as the weighted sum-rate (WSR) and to other scenarios. Readers are invited to check [51] that confirms the inefficiency of NOMA and the superiority of RS from a WSR perspective, and are encouraged to consult the growing literature on RS (and RSMA) demonstrating the superiority of RS over NOMA and MU-LP in numerous scenarios and applications, namely multi-user multi-antenna communications [51], [77], [83], multigroup multicast [39], [56], energy efficiency [78], [84], [85], multi-cell joint transmission [86], non-orthogonal unicast and multicast transmission [78], wireless information and power transfer [87], cooperative communication with user relaying [88], imperfect CSIT [89], [90], link-level simulations [91], C-RAN [92], secrecy rate [93], [94], aerial networks [85], [95], imperfect CSIT and CSIR [96], visible light communications [97], [98], network performance analysis [99], reconfigurable intelligent surface [100]. It would also be of interest for future work to understand how more recent MIMO NOMA schemes such as [67], [68] compare to RS [73], [75].

The emphasis of the paper was on downlink multi-user communications. Results suggest that future downlink multi-user multi-antenna communications would strongly benefits from RSMA. Indeed, RSMA achieves higher multiplexing gains and rates. It is capable of serving a larger number of users and is more robust to user deployments, network loads and inaccurate CSI. Moreover, RSMA has a lower receiver complexity than NOMA. RSMA is a gold mine of research problems for academia and industry with issues spanning numerous areas: RSMA to achieve the fundamental limits of wireless networks; RSMA for multi-user/multi-cell multi-antenna networks; RSMA-based robust interference management; RSMA in MU-MIMO, coordinated multi-point (CoMP), Massive MIMO, millimeter wave and higher frequency bands, relay, cognitive radio, caching, physical layer security, cooperative communications, cloud/fog-enabled platforms and Radio Access Networks (RAN) (such as cloud-RAN and fog-RAN), intelligent reflecting surfaces; RSMA to unify, generalize and outperform SDMA and NOMA; physical layer design of RSMA-based network; coding and modulation for RSMA; cross-layer design, optimization and performance analysis of RSMA; implementation and standardization of RSMA; RSMA in B5G services such as enhanced Mobile Broadband (eMBB), enhanced Ultra-Reliable Low Latency Communications (URLLC), enhanced Machine-Type Communications (MTC), massive MTC, massive Internet-of-Things (IoT), Vehicle-to-everything (V2X), cellular, Unmanned Aerial Vehicle (UAV) and satellite networks, wireless powered communications, integrated communications and sensing, etc.

RSMA can also be used in the uplink, as originally shown for single-antenna systems in [101]. The key benefit of RSMA in the uplink is its ability to achieve the capacity region of the MAC without the need for time sharing. Nevertheless, much is left to be done to identify the benefits of RSMA for general uplink multi-user multi-antenna communications. The performance benefits of RSMA vs. NOMA vs. OMA vs. other multiple access techniques in the uplink, beyond the existing NOMA vs. OMA comparison [102], is also much worth investigating. It should however be mentioned that thanks to the polymatroid structure of the Gaussian MAC capacity region, the solution to the max weighted sum rate problem is always at a vertex of the original region, i.e., RS is not needed.

Standardization is very important for a widespread adoption of multiple access techniques. MU-LP has been heavily discussed and standardized in 4G and 5G as part of MU-MIMO and Massive MIMO. NOMA was also investigated as part of a study item in 5G but was not considered any further in 5G because its gains compared to MU-MIMO were not found convincing [103]. Hence, in 5G New Radio (NR), NOMA was seen as a competing technology to MU-MIMO and an unwanted add-on technology. The standardization of RSMA has not been considered by 3GPP yet but is receiving a growing interest from academia and industry.[36] Parts of the features required by RSMA are already being studied, discussed and developed. Some current work items and features in 5G, i.e., MU-MIMO/Massive MIMO/CoMP, multiuser superposition transmission (MUST), network-assisted interference cancellation and suppression (NAICS), multicast functionality can be leveraged for RSMA. However, some more work is needed to realize the full potential of RSMA. Ongoing activities consist in investigating the potential benefits of RSMA for 6G [104] and demonstrating the significant benefits of RSMA over 5G NR design [105].

APPENDIX A
PROOF OF PROPOSITION 4

Let us first consider $G > 1$ and $M \geq K-g+1$. Recalling from the proof of Proposition 3 that the sum multiplexing gain of $G\alpha$ can be split equally among the G groups so that each group gets a (group) sum multiplexing gain of α, and following the MAC argument, the (group) sum multiplexing gain of α in each group can then be further split equally among the g users, which leads to an upper bound on the MMF multiplexing gain of $\frac{\alpha}{g}$. Achievability is obtained by designing precoders using ZFBF, and allocating power (consider group 1 for simplicity) to user $k = 1, \ldots, g$ as $O(P^{1-\frac{g-k}{g}\alpha})$, which causes the SINR for user-k to scale as $O(P^{\alpha/g})$ and an achievable MMF multiplexing gain of $\frac{\alpha}{g}$.

To illustrate the achievability in more detail, we consider a simple example for $K = 4$, $G = 2$, $g = 2$, and $M \geq 3$. First, we design the precoders $\mathbf{p}_1$ and $\mathbf{p}_2$ in group 1 to be orthogonal to the channel estimates $\hat{\mathbf{h}}_3$ and $\hat{\mathbf{h}}_4$ of users

36. See the special interest group on RSMA at https://sites.google.com/view/ieee-comsoc-wtc-sig-rsma/home.

3 and 4. Similarly, $\mathbf{p}_3$ and $\mathbf{p}_4$ in group 2 are made orthogonal to $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$. Second, allocate power $O(P^b)$ with $b = 1-\alpha/2$ to users 1 and 3, and $O(P-P^b) = O(P)$ to users 2 and 4. Using these precoders and power allocations, the received signals in group 1 can be written as

$$y_1 = \underbrace{\mathbf{h}_1^H\mathbf{p}_1 s_1}_{P^b} + \underbrace{\mathbf{h}_1^H\mathbf{p}_2 s_2}_{P} + \underbrace{\tilde{\mathbf{h}}_1^H\mathbf{p}_3 s_3}_{P^{b-\alpha}} + \underbrace{\tilde{\mathbf{h}}_1^H\mathbf{p}_4 s_4}_{P^{1-\alpha}} + \underbrace{n_1}_{P^0}, \quad (47)$$

$$y_2 = \underbrace{\mathbf{h}_2^H\mathbf{p}_1 s_1}_{P^b} + \underbrace{\mathbf{h}_2^H\mathbf{p}_2 s_2}_{P} + \underbrace{\tilde{\mathbf{h}}_2^H\mathbf{p}_3 s_3}_{P^{b-\alpha}} + \underbrace{\tilde{\mathbf{h}}_2^H\mathbf{p}_4 s_4}_{P^{1-\alpha}} + \underbrace{n_2}_{P^0}, \quad (48)$$

where the quantities under the brackets refer to how the power level of each term scales. From (47) and (48), s_2 can be decoded at an SINR level scaling as $\frac{P}{P^b+P^{1-\alpha}+P^{b-\alpha}+P^0} = \frac{P}{P^b} = P^{\alpha/2}$ (since $b \geq 1-\alpha \geq b-\alpha$ and $b \geq 0$). Using SIC, s_2 is cancelled in (47), and s_1 can be decoded at an SINR level scaling as $\frac{P^b}{P^{1-\alpha}+P^{b-\alpha}+P^0} = P^{\alpha/2}$. Similar expressions hold for group 2, and we note that all four streams have an SINR scaling as $P^{\alpha/2}$, therefore achieving an MMF multiplexing gain of $\frac{\alpha}{2}$.

Let us now consider $G > 1$ and $M < K-g+1$. Since the MMF multiplexing gain collapses to 0 in the perfect CSIT setting, the same holds for imperfect CSIT.

Let us now consider $G = 1$. The situation here is the same as in the perfect CSIT setting. There is no inter-group interference and the sum multiplexing gain of one in the single group can be split equally among the K users, which leads to an upper bound on the MMF multiplexing gain of $\frac{1}{K}$. Achievability is obtained by choosing the powers of users $k = 1, \ldots, K$ as $O(P^{k/K})$, which causes the SINR of user-k to scale as $O(P^{1/K})$ and results in an achievable MMF multiplexing gain of $\frac{1}{K}$.

APPENDIX B
WMMSE OPTIMIZATION FRAMEWORK

The WMMSE optimization framework to solve both problems (45) and (46) is specified as follows.

At user-$j, j \in \mathcal{K}_i$, equalizer $g_{j,k}$ is employed to decode stream $s_k, k \in \{k | k \geq j, k \in \mathcal{K}_i\}$. The estimate of s_k at user-j is obtained as $\widehat{s}_{j,k} = g_{j,k}y_{j,k}$, where $y_{j,k} = \sum_{m\leq k, m\in\mathcal{K}_i} \mathbf{h}_j^H\mathbf{p}_m s_m + \sum_{l\neq i, l\in\mathcal{G}}\sum_{m\in\mathcal{K}_l}\mathbf{h}_j^H\mathbf{p}_m s_m + n_j$ is the signal received at user-j after removing the streams decoded before s_k. The corresponding Mean Square Error (MSE) is given by

$$\begin{aligned}\varepsilon_{j,k} &= \mathbb{E}\left\{\left|\widehat{s}_{j,k} - s_k\right|^2\right\} \\ &= \left|g_{j,k}\right|^2 T_{j,k} - 2\Re\left\{g_{j,k}\mathbf{h}_j^H\mathbf{p}_k\right\} + 1, \end{aligned} \quad (49)$$

where $T_{j,k} = |\mathbf{h}_j^H\mathbf{p}_k|^2 + I_{j,k}^{(in)} + I_{j,k}^{(ou)}$ is the power received at user-j when decoding s_k. Furthermore, $I_{j,k}^{(in)}$ and $I_{j,k}^{(ou)}$ are respectively the intra-group and inter-group interference power defined in (18).

By solving $\frac{\partial\varepsilon_{j,k}}{\partial g_{j,k}} = 0$, the optimal Minimum MSE (MMSE) equalizer is calculated as

$$g_{j,k}^{\text{MMSE}} = \mathbf{p}_k^H\mathbf{h}_j\left(T_{j,k}\right)^{-1}. \quad (50)$$

Substituting (50) back to (49), the corresponding MMSE is then obtained as

$$\varepsilon_{j,k}^{\text{MMSE}} = \min_{g_{j,k}} \varepsilon_{j,k} = T_{j,k}^{-1}\left(I_{j,k}^{(in)} + I_{j,k}^{(ou)}\right). \quad (51)$$

With the introduced $\varepsilon_{j,k}^{\text{MMSE}}$, the rate at user-$j$ to decode the message of user-k in (17) is equivalently written as $R_{j,k} = -\log_2(\varepsilon_{j,k}^{\text{MMSE}})$. Defining the Weighted MSE (WMSE) of $\varepsilon_{j,k}$ with a weight $u_{j,k} > 0$ as

$$\xi_{j,k} = u_{j,k}\varepsilon_{j,k} - \log_2\left(u_{j,k}\right), \quad (52)$$

and defining its Weighted MMSE (WMMSE) by minimizing $\xi_{j,k}$ over $u_{j,k}$ and $g_{j,k}$ as

$$\xi_{j,k}^{\text{MMSE}} = \min_{u_{j,k}, g_{j,k}} \xi_{j,k}, \quad (53)$$

we then establish the rate-WMMSE relationship, which is given by

$$\xi_{j,k}^{\text{MMSE}} = 1 - R_{j,k}. \quad (54)$$

The rate-WMMSE relation in (54) is obtained as follows. The optimum equalizer is calculated as $g_{j,k}^* = g_{j,k}^{\text{MMSE}}$ from $\frac{\partial\xi_{j,k}}{\partial g_{j,k}} = 0$. Substituting $g_{j,k}^{\text{MMSE}}$ back to (52) yields $\xi_{j,k}(g_{j,k}^{\text{MMSE}}) = u_{j,k}\varepsilon_{j,k}^{\text{MMSE}} - \log_2(u_{j,k})$. By solving $\frac{\partial\xi_{j,k}(g_{j,k}^{\text{MMSE}})}{\partial g_{j,k}} = 0$, we then obtain the optimal MMSE weight, which is given as

$$u_{j,k}^* = u_{j,k}^{\text{MMSE}} = \left(\varepsilon_{j,k}^{\text{MMSE}}\right)^{-1}. \quad (55)$$

Substituting $u_{j,k}^{\text{MMSE}}$ back to $\xi_{j,k}(g_{j,k}^{\text{MMSE}})$, we have $\min_{u_{j,k}, g_{j,k}} \xi_{j,k} = 1 - R_{j,k}$. Following (53), we obtain (54).

Motivated by the rate-WMMSE in (54), we find that the achievable rate of user-k in (19) is equal to $R_k = 1-\xi_k^{\text{MMSE}}$, where $\xi_k^{\text{MMSE}} = \max_{j\leq k, j\in\mathcal{K}_i} \xi_{j,k}^{\text{MMSE}}$. By defining the WMSE of user-k as

$$\xi_k = \max_{j\leq k, j\in\mathcal{K}_i} \xi_{j,k}, \quad (56)$$

and the respective set of equalizers and weights as $\mathbf{g} = \{g_{j,k} | j \leq k, k, j \in \mathcal{K}_i, i \in \mathcal{G}\}$, $\mathbf{u} = \{u_{j,k} | j \leq k, k, j \in \mathcal{K}_i, i \in \mathcal{G}\}$, the sum-rate WMMSE problem is formulated as

$$\min_{\mathbf{P},\mathbf{u},\mathbf{g}} \sum_{k\in\mathcal{K}} \xi_k \quad (57a)$$

$$\text{s.t. } \text{tr}\left(\mathbf{P}\mathbf{P}^H\right) \leq P. \quad (57b)$$

Following the proof of [47], we find that the MMSE solutions of the equalizers $\mathbf{g}^{\text{MMSE}} = \{g_{j,k}^{\text{MMSE}} | j \leq k, k, j \in \mathcal{K}_i, i \in \mathcal{G}\}$ and weights $\mathbf{u}^{\text{MMSE}} = \{u_{j,k}^{\text{MMSE}} | j \leq k, k, j \in \mathcal{K}_i, i \in \mathcal{G}\}$ satisfy the KKT optimality conditions of (57). Substituting $(\mathbf{g}^{\text{MMSE}}, \mathbf{u}^{\text{MMSE}})$ back to (57) with affine transformations applied to the objective function, (57) boils down to (45). In fact, for any point $(\mathbf{P}^*, \mathbf{u}^*, \mathbf{g}^*)$ satisfying the KKT optimality conditions of (57), the solution $\mathbf{P}^*$ satisfies the KKT optimality conditions of (45). Hence, (57) yields a solution for (45).

Algorithm 1: AO Algorithm

Initialize: $t \leftarrow 0$, $\mathbf{P}$;
repeat
  $t \leftarrow t+1$, $\mathbf{P}^{[t-1]} \leftarrow \mathbf{P}$;
  $\mathbf{g} \leftarrow \mathbf{g}^{\mathrm{MMSE}}(\mathbf{P}^{[t-1]})$; $\mathbf{u} \leftarrow \mathbf{u}^{\mathrm{MMSE}}(\mathbf{P}^{[t-1]})$;
  Substitute $(\mathbf{g}, \mathbf{u})$ back to (57) and update $\mathbf{P}$ by solving (57);
until *convergence*;

Although the transformed problem (57) is still non-convex, it is block-wise convex with respect to $\mathbf{P}$ and $(\mathbf{g}, \mathbf{u})$. For a given $\mathbf{P}$, the optimal solution of the weights and equalizers are $\mathbf{g}^{\mathrm{MMSE}}(\mathbf{P})$, $\mathbf{u}^{\mathrm{MMSE}}(\mathbf{P})$. When $(\mathbf{g}, \mathbf{u})$ are fixed, problem (57) becomes convex and can be solved by interior-point methods. Motivated by the block-wise convexity, we use the Alternating Optimization (AO) algorithm as illustrated in Algorithm 1 to solve (57). In each iteration, the equalizers and weights are first updated by $(\mathbf{g}^{\mathrm{MMSE}}(\mathbf{P}), \mathbf{u}^{\mathrm{MMSE}}(\mathbf{P}))$ for a given $\mathbf{P}$. The updated equalizers and weights $(\mathbf{g}^{\mathrm{MMSE}}(\mathbf{P}), \mathbf{u}^{\mathrm{MMSE}}(\mathbf{P}))$ are substituted back to (57). Precoder $\mathbf{P}$ is then updated by solving (57). $\mathbf{P}$ and $(\mathbf{g}, \mathbf{u})$ are updated in an alternating manner until the convergence of the sum-rate. Algorithm 1 is guaranteed to converge and it converges to the KKT solution of problem (45). Readers are referred to [47] for the proof.

Following the same procedure, we are able to obtain the transformed WMMSE problem for max-min rate maximization, which is given by

$$\min_{\mathbf{P},\mathbf{u},\mathbf{g}} \ \max_{k\in\mathcal{K}} \ \xi_k \tag{58a}$$

$$\text{s.t.} \ \ \mathrm{tr}\left(\mathbf{P}\mathbf{P}^H\right) \leq P. \tag{58b}$$

By substituting problem (57) in Algorithm 1 with problem (58), we obtain the corresponding AO Algorithm to achieve the KKT solution of the max-min rate problem (46).

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, 2013, pp. 1–5.

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

[3] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[4] 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.

[5] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[6] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 2–14, Jan. 1972.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[8] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[9] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[10] Q. Li *et al.*, "MIMO techniques in WiMAX and LTE: A feature overview," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 86–92, May 2010.

[11] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 174–180, Oct. 2019.

[12] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input–multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[13] Q. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multi-user MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 462–471, Feb. 2004.

[14] M. Stojnic, H. Vikalo, and B. Hassibi, "Rate maximization in multiantenna broadcast channels with linear preprocessing," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2338–2342, Sep. 2006.

[15] A. D. Dabbagh and D. J. Love, "Precoding for multiple antenna Gaussian broadcast channels with successive zero-forcing," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3837–3850, Jul. 2007.

[16] A. D. Dabbagh and D. J. Love, "Multiple antenna MMSE based downlink precoding with quantized feedback or channel mismatch," *IEEE Trans. Commun.*, vol. 56, no. 11, pp. 1859–1868, Nov. 2008.

[17] B. Clerckx and C. Oestges, *MIMO Wireless Networks: Channels, Techniques and Standards for Multi-Antenna, Multi-User and Multi-Cell Systems*. Cambridge, MA, USA: Academic, 2013.

[18] Y. Liu, H. Xing, C. Pan, A. Nallanathan, M. Elkashlan, and L. Hanzo, "Multiple-antenna-assisted non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 17–23, Apr. 2018.

[19] M. Vaezi and H. V. Poor, "NOMA: An information-theoretic perspective," in *Multiple Access Techniques for 5G Wireless Networks and Beyond*, M. Vaezi, Z. Ding, and H. Poor, Eds. Heidelberg, Germany: Springer, 2019.

[20] 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.

[21] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[22] 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.

[23] Q. Zhang, Q. Li, and J. Qin, "Robust beamforming for nonorthogonal multiple-access systems in MISO channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10231–10236, Dec. 2016.

[24] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.

[25] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[26] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[27] J. Choi, "On generalized downlink beamforming with NOMA," *J. Commun. Netw.*, vol. 19, no. 4, pp. 319–328, 2017.

[28] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 84–87, Jan. 2017.

[29] V. D. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O. S. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2681–2695, Dec. 2017.

[30] 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.

[31] X. Chen, Z. Zhang, C. Zhong, and D. W. K. Ng, "Exploiting multiple antenna techniques for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2207–2220, Oct. 2017.

[32] 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, 2018.

[33] C. Chen, W. Cai, X. Cheng, L. Yang, and Y. Jin, "Low complexity beamforming and user selection schemes for 5G MIMO-NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2708–2722, Dec. 2017.

[34] 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.

[35] F. Alavi, K. Cumanan, Z. Ding, and A. G. Burr, "Beamforming techniques for nonorthogonal multiple access in 5G cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9474–9487, Oct. 2018.

[36] Y. Jeong, C. Lee, and Y. H. Kim, "Power minimizing beamforming and power allocation for MISO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6187–6191, Jun. 2019.

[37] J. Zhang, Y. Zhu, S. Ma, X. Li, and K.-K. Wong, "Large system analysis of downlink MISO-NOMA system via regularized zero-forcing precoding With imperfect CSIT" *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2454–2458, Nov. 2020.

[38] J. Chu, X. Chen, C. Zhong, and Z. Zhang, "Robust design for NOMA-based multi-beam LEO satellite Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1959–1970, Feb. 2021.

[39] A. Z. Yalcin and M. Yuksel, "Max–min fair precoder design for non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, early access.

[40] Y. Liu, X. Mu, X. Liu, M. D. Renzo, Z. Ding, and R. Schober, "Reconfigurable intelligent surface (RIS) aided multi-user networks: Interplay between NOMA and RIS," 2020. [Online]. Available: arXiv:2011.13336.

[41] Z. Chen, Z. Ding, P. Xu, and X. Dai, "Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1263–1266, Jun. 2016.

[42] 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.

[43] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[44] J. Zhu, J. Wang, Y. Huang, K. Navaie, Z. Ding, and L. Yang, "On optimal beamforming design for downlink MISO NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3008–3020, Mar. 2020.

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

[46] S. Yang, M. Kobayashi, D. Gesbert, and X. Yi, "Degrees of freedom of time correlated MISO broadcast channel with delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 315–328, Jan. 2013.

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

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

[49] C. Hao, Y. Wu, and B. Clerckx, "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3232–3246, Sep. 2015.

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

[51] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, p. 133, May 2018.

[52] E. Bjornson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next? Five promising research directions for antenna arrays," *Digit. Signal Process.*, vol. 94, pp. 3–20, Nov. 2019.

[53] 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.

[54] J. Choi, "Power allocation for max–sum rate and max–min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.

[55] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[56] H. Joudeh and B. Clerckx, "Rate-splitting for max–min fair multigroup multicast beamforming in overloaded systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7276–7289, Nov. 2017.

[57] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 1995, pp. 1–42.

[58] G. Caire and S. S. Shitz, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[59] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1783–1794, May 2005.

[60] P. Ding, D. J. Love, and M. D. Zoltowski, "Multiple antenna broadcast channels with shape feedback and limited feedback," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3417–3428, Jul. 2007.

[61] M. Kobayashi, N. Jindal, and G. Caire, "Training and feedback optimization for multiuser MIMO downlink," *IEEE Trans. Commun.*, vol. 59, no. 8, pp. 2228–2240, Aug. 2011.

[62] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY, USA: Springer, 2013.

[63] A. G. Davoodi and S. A. Jafar, "Aligned image sets under channel uncertainty: Settling conjectures on the collapse of degrees of freedom under finite precision CSIT," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5603–5618, Oct. 2016.

[64] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

[65] L.-U. Choi and R. D. Murch, "A transmit preprocessing technique for multi user MIMO systems using a decomposition approach," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 20–24, Jan. 2004.

[66] Z. Pan, K. K. Wong, and T.-S. Ng, "Generalized multiuser orthogonal space-division multiplexing," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 1969–1973, Nov. 2004.

[67] A. Krishnamoorthy and R. Schober, "Uplink and downlink MIMO-NOMA with simultaneous triangularization," *IEEE Trans. Wireless Commun.*, early access, Jan. 13, 2021, doi: 10.1109/TWC.2021.3049594.

[68] A. Krishnamoorthy, Z. Ding, and R. Schober, "Precoder design and statistical power allocation for MIMO-NOMA via user-assisted simultaneous diagonalization," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 929–945, Feb. 2021.

[69] L. Yin and B. Clerckx, "Rate-splitting multiple access for multigroup multicast and multibeam satellite systems," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 976–990, Feb. 2021.

[70] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 1, pp. 49–60, Jan. 1981.

[71] E. Piovano and B. Clerckx, "Optimal DoF region of the K-user MISO BC with partial CSIT," *IEEE Commun. Lett.*, vol. 21, no. 11, pp. 2368–2371, Nov. 2017.

[72] H. Joudeh and B. Clerckx, "DoF region of the MISO BC with partial CSIT: Proof by inductive Fourier–Motzkin elimination," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2019, pp. 1–5.

[73] C. Hao, B. Rassouli, and B. Clerckx, "Achievable DoF regions of MIMO networks with imperfect CSIT," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6587–6606, Oct. 2017.

[74] A. G. Davoodi and S. Jafar, "Degrees of freedom region of the (M, N_1, N_2) MIMO broadcast channel with partial CSIT: An application of sum-set inequalities based on aligned image sets," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6256–6279, Oct. 2020.

[75] A. Mishra, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for downlink multiuser MIMO: Precoder optimization and phy-layer design," 2020. [Online]. Available: arxiv.abs/2105.07362.

[76] K. Senel, H. V. Cheng, E. Bjornson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 597–611, Jun. 2019.

[77] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 349–353, Mar. 2020.

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

[79] M. Grant, S. Boyd, and Y. Ye. (2008). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: http://www.stanford.edu/ boyd/cvx

[80] S. S. Christensen, R. Agarwal, E. D. Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

[81] "LTE; Evolved universal terrestrial aadio access (E-UTRA); radio frequency (RF) requirements for LTE pico node B," 3GPP, Sophia Antipolis, France, Rep. TR 36.931, May 2011.

[82] A. Shapiro, D. Dentcheva, and A. Ruszczynski, "Lectures on stochastic programming: Modeling and theory," in *Proc. Soc. Ind. Appl. Math.*, 2014, pp. 1–494.

[83] Z. Yang, M. Chen, W. Saad, and M. Shikh-Bahaei, "Optimization of rate allocation and power control for rate splitting multiple access (RSMA)," Accessed: Jun. 2, 2021. [Online]. Available: https://arxiv.org/abs/1903.08068.

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

[85] A. Rahmati, Y. Yapici, N. Rupasinghe, I. Guvenc, H. Dai, and A. Bhuyan, "Energy efficiency of RSMA and NOMA in cellular connected mmWave UAV networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[86] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for coordinated multi-point joint transmission," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[87] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting for multi-user multi-antenna wireless information and power transfer," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2019, pp. 1–5.

[88] J. Zhang, B. Clerckx, J. Ge, and Y. Mao, "Cooperative rate-splitting for MISO broadcast channel with user relaying, and performance benefits over cooperative NOMA," *IEEE Signal Process. Lett.*, vol. 26, no. 11, pp. 1678–1682, Nov. 2019.

[89] Y. Mao and B. Clerckx, "Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: A rate-splitting approach," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6775–6791, Nov. 2020.

[90] Y. Mao and B. Clerckx, "Dirty paper coded rate-splitting for non-orthogonal unicast and multicast transmission with partial CSIT," in *Proc. Asilomar Conf. Signals Syst. Comput.* 2020.

[91] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access for downlink multi-antenna communications: Physical layer design and link-level simulations," in *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2020, pp. 1–6.

[92] D. Yu, J. Kim, and S. Park, "An efficient rate-splitting multiple access scheme for the downlink of C-RAN systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1555–1558, Dec. 2019.

[93] P. Li, M. Chen, Y. Mao, Z. Yang, B. Clerckx, and M. Shikh-Bahaei, "Cooperative rate-splitting for secrecy sum-rate enhancement in multi-antenna broadcast channels," in *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2020, pp. 1–6.

[94] H. Fu, S. Feng, W. Tang, and D. W. K. Ng, "Robust secure resource allocation for downlink two-user MISO rate-splitting systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–7.

[95] W. Jaafar, S. Naser, S. Muhaidat, P. C. Sofotasios, and H. Yanikomeroglu, "Multiple access in aerial networks: From orthogonal and non-orthogonal to rate-splitting," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 372–392, Oct. 2020.

[96] J. An, O. Dizdar, B. Clerckx, and W. Shin, "Rate-splitting multiple Access for multi-antenna broadcast channel with imperfect CSIT and CSIR," in *Proc. IEEE Annu. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2020, pp. 1–7.

[97] S. Tao, H.-Y. Yu, L. Qing, Y. Tang, and D. Zhang, "One-layer rate-splitting multiple access with benefits over power-domain NOMA in indoor multi-cell VLC networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[98] S. Naser *et al.*, "Rate-splitting multiple access: Unifying NOMA and SDMA in MISO VLC channels," *IEEE Open J. Veh. Tech.*, vol. 1, pp. 393–413, 2020.

[99] E. Demarchou, C. Psomas, and I. Krikidis, "Channel statistics-based rate splitting with spatial randomness," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[100] Z. Yang, J. Shi, Z. Li, M. Chen, W. Xu, and M. Shikh-Bahaei, "Energy efficient rate splitting multiple access (RSMA) with reconfigurable intelligent surface," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[101] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996.

[102] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, Jan. 2020.

[103] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 179–189, 2020.

[104] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access: A New Frontier for the PHY layer of 6G," in *Proc. 91st Veh. Technol. Conf. (VTC) Spring*, 2020, pp. 1–7.

[105] O. Dizdar, Y. Mao, and B. Clerckx, "Rate-splitting multiple access to mitigate the curse of mobility in (massive) MIMO networks," 2021. [Online]. Available: arXiv:2102.06405.

BRUNO CLERCKX (Senior Member, IEEE) received the Ph.D. degree in 2005. He is a Professor, the Head of the Communications and Signal Processing Lab, and the Deputy Head of Communications and Signal Processing Group, within the Electrical and Electronic Engineering Department, Imperial College London, London, U.K. From 2006 to 2011, he was with Samsung Electronics, Suwon, South Korea, where he actively contributed to 4G (3GPP LTE/LTE-A and IEEE 802.16m) and acted as the Rapporteur for the 3GPP Coordinated Multi-Point (CoMP) Study Item. Since 2011, he has been with Imperial College London. From 2014 to 2016, he also was an Associate Professor with Korea University, Seoul, South Korea. He has authored two books on *MIMO Wireless Communications* and *MIMO Wireless Networks*, 200 peer-reviewed international research papers, and 150 standards contributions, and he is the Inventor of 80 issued or pending patents among which 15 have been adopted in the specifications of 4G standards and are used by billions of devices worldwide. He served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He has also been a (Lead) Guest Editor for special issues of the EURASIP *Journal on Wireless Communications and Networking*, IEEE ACCESS, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, and the PROCEEDINGS OF THE IEEE. He was an Editor for the 3GPP LTE-Advanced Standard Technical Report on CoMP. He is an IEEE Communications Society Distinguished Lecturer from 2021 to 2022. He was an Elected Member of the IEEE Signal Processing Society SPCOM Technical Committee.

YIJIE MAO (Member, IEEE) received the B.Eng. degree from the Beijing University of Posts and Telecommunications, the B.Eng. degree (Hons.) from the Queen Mary University of London, London, U.K., in 2014, and the Ph.D. degree from the Electrical and Electronic Engineering Department, The University of Hong Kong, Hong Kong, in 2018. She was a Postdoctoral Research Fellow with The University of Hong Kong from October 2018 to July 2019. Since August 2019, she has been a Postdoctoral Research Associate with the Communications and Signal Processing Group, Department of the Electrical and Electronic Engineering, Imperial College London, London. Her research interests include multiple-input multiple-output communication networks, rate splitting, and non-orthogonal multiple access. She served as the Co-Chair for the 2020 IEEE International Conference on Communications and the 2020 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications on the workshops of Rate-Splitting and Robust Interference Management for Beyond 5G.

ROBERT SCHOBER (Fellow, IEEE) received the Diploma and the Ph.D. degrees in electrical engineering from the Friedrich-Alexander University of Erlangen-Nuremberg (FAU), Germany, in 1997 and 2000, respectively. From 2002 to 2011, he was a Professor and a Canada Research Chair with the University of British Columbia (UBC), Vancouver, Canada. Since January 2012, he has been an Alexander von Humboldt Professor and the Chair for Digital Communication with FAU. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing. He several awards for his work, including the 2002 Heinz Maier–Leibnitz Award of the German Science Foundation, the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, the 2011 Alexander von Humboldt Professorship, the 2012 NSERC E. W. R. Stacie Fellowship, and the 2017 Wireless Communications Recognition Award by the IEEE Wireless Communications Technical Committee. Since 2017, he has been listed as a Highly Cited Researcher by the Web of Science. From 2012 to 2015, he served as the Editor-in-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS. He currently serves as member of the Editorial Board of the PROCEEDINGS OF THE IEEE and as a VP Publications for the IEEE Communication Society. He is a Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada, and a member of the German National Academy of Science and Engineering.

EDUARD A. JORSWIECK (Fellow, IEEE) was born in Berlin, Germany, in 1975. He received the Dipl.-Ing. degree in computer engineering and the Ph.D. degree in electrical engineering from the Berlin Institute of Technology, Berlin, Germany, in 2000 and 2004, respectively. From 2006 to 2008, he was a Postdoctoral Research Fellow and an Assistant Professor with the Signal Processing Department, KTH Royal Institute of Technology, Stockholm, Sweden. Since 2020, he has been the Managing Director of the Institute of Communications Technology, the Head of the Chair for Communications Systems, and a Full Professor with Technische Universitaet Braunschweig, Brunswick, Germany. From 2008 to 2019, he was the Head of the Chair of Communications Theory and a Full Professor with the Dresden University of Technology, Dresden, Germany. He has authored and coauthored more than 135 journal papers, 15 book chapters, three monographs, and some 285 conference papers in his research fields, which include communications, signal processing, and applied information theory. He was a recipient of the IEEE Signal Processing Society Best Paper Award. He and his colleagues were also a recipient of the Best Paper and Best Student Paper Awards at the IEEE CAMSAP 2011, IEEE WCSP 2021, IEEE SPAWC 2021, IEEE ICUFN 2018, PETS 2019, and ISWCS 2019. Since 2017, he has been the Editor-in-Chief of the *EURASIP Journal on Wireless Communications and Networking*. He was on the editorial boards for the IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

DAVID J. LOVE (Fellow, IEEE) received the B.S. (with highest Hons.), M.S.E., and Ph.D. degrees in electrical engineering from the University of Texas at Austin, Austin, TX, USA, in 2000, 2002, and 2004, respectively. Since 2004, he has been with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA, where he is currently a Nick Trbovich Professor of Electrical and Computer Engineering. He holds 32 issued U.S. patent filings. His research interests include the design and analysis of broadband wireless communication systems, 5G wireless systems, multiple-input multiple-output (MIMO) communications, millimeter wave wireless, software defined radios and wireless networks, coding theory, and MIMO array processing. Along with Coauthors, he won best paper awards the 2016 Stephen O. Rice Prize and the 2020 Fred Ellersick Prize from the IEEE Communications Society; the 2015 Best Paper Award from the IEEE Signal Processing Society; and the 2009 Jack Neubauer Memorial Award from the IEEE Vehicular Technology Society. He was named a Thomson Reuters Highly Cited Researcher in 2014 and 2015. He is currently a Senior Editor of the *IEEE Signal Processing Magazine*. He was an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and a Guest Editor of special issues of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and the *EURASIP Journal on Wireless Communications and Networking*. He is a Fellow of the Royal Statistical Society and inducted into Tau Beta Pi and Eta Kappa Nu. He is a member of the Executive Committee for the National Spectrum Consortium.

JINHONG YUAN (Fellow, IEEE) is a Professor and the Head of Telecommunication Group, School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia. He has published two books, five book chapters, over 300 papers in telecommunications journals and conference proceedings, and 50 industrial reports. He is a Co-Inventor of one patent on MIMO systems and two patents on low-density-parity-check codes. He has coauthored four Best Paper Awards and one Best Poster Award, including the Best Paper Award from the IEEE International Conference on Communications, Kansas City, USA, in 2018, the Best Paper Award from IEEE Wireless Communications and Networking Conference, Cancun, Mexico, in 2011, and the Best Paper Award from the IEEE International Symposium on Wireless Communications Systems, Trondheim, Norway, in 2007. His current research interests include error control coding and information theory, communication theory, and wireless communications. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and IEEE TRANSACTIONS ON COMMUNICATIONS. He served as the IEEE NSW Chapter Chair of Joint Communications/Signal Processions/Ocean Engineering Chapter from 2011 to 2014, and served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS from 2012 to 2017.

LAJOS HANZO (Fellow, IEEE) received the master's and Doctorate degrees from the Technical University of Budapest in 1976 and 1983, respectively, the Doctor of Sciences degree from the University of Southampton, U.K., in 2004, and the first Honorary Doctorate degree from TU of Budapest in 2009, and the second Honorary Doctorate degree from the University of Edinburgh in 2015. He is a Former Editor-in-Chief of the IEEE Press. He has served several terms as Governor of both IEEE ComSoc and of VTS. He has published more then 1900 contributions at IEEE Xplore, 19 Wiley-IEEE Press books, and has helped the fast-track career of 123 Ph.D. students. He holds the Chair of Telecommunications and directs the research of Next-Generation Wireless with the University of Southampton. He is a Fellow of the Royal Academy of Engineering, the IET, and of EURASIP. He is a Foreign Member of the Hungarian Academy of Sciences.

GEOFFREY YE LI (Fellow, IEEE) has been a Chair Professor with Imperial College London, U.K., since 2020. Before moving to Imperial, he was with the Georgia Institute of Technology, Georgia, USA, as a Professor for twenty years and with AT&T Labs—Research, New Jersey, USA, as a Principal Technical Staff Member for five years. His general research interests include statistical signal processing and machine learning for wireless communications. In the related areas, he has published over 500 journal and conference papers in addition to over 40 granted patents. He was awarded the IEEE Fellow for his contributions to signal processing for wireless communications in 2005. He won several prestigious awards the Donald G. Fink Overview Paper Award in 2017 from IEEE Signal Processing Society; the James Evans Avant Garde Award in 2013 and the Jack Neubauer Memorial Award in 2014 from IEEE Vehicular Technology Society; and the Stephen O. Rice Prize Paper Award in 2013, the Award for Advances in Communication in 2017, and the Edwin Howard Armstrong Achievement Award in 2019 the from IEEE Communications Society. He also received the 2015 Distinguished ECE Faculty Achievement Award from Georgia Tech. His publications have been cited over 46 000 times and he has been recognized as a Highly Cited Researcher, by Thomson Reuters, almost every year. He has organized and chaired many international conferences, including a Technical Program Vice-Chair of the IEEE ICC'03, and a General Co-Chair of the IEEE GlobalSIP'14, the IEEE VTC'19 (Fall), and the IEEE SPAWC'20. He has been involved in editorial activities for over 20 technical journals, including the Founding Editor-in-Chief of IEEE JSAC Special Series on ML in Communications and Networking.

ERIK G. LARSSON (Fellow, IEEE) is a Professor with Linköping University, Sweden. He coauthored the textbook *Fundamentals of Massive MIMO* (Cambridge University Press, 2016). His main professional interests are within signal processing, communication theory, applied information theory, wireless systems, and 5G. He received among others, the IEEE ComSoc Stephen O. Rice Prize in Communications Theory in 2015, the IEEE ComSoc Leonard G. Abraham Prize in 2017, the IEEE ComSoc Best Tutorial Paper Award in 2018, and the IEEE ComSoc Fred W. Ellersick Prize in 2019. He was the Chair of the IEEE Signal Processing Society Signal Processing for Communications and Networking Technical Committee in 2015 and 2016, and Steering Committee of IEEE WIRELESS COMMUNICATIONS LETTERS in 2014 and 2015, and he organized the Asilomar Conference on Signals, Systems, and Computers as a General Chair in 2015, and a Technical Chair in 2012.

GIUSEPPE CAIRE (Fellow, IEEE) was born in Torino, in 1965. He received the B.Sc. degree in electrical engineering from the Politecnico di Torino, in 1990, the M.Sc. degree in electrical engineering from Princeton University, in 1992, and the Ph.D. degree from the Politecnico di Torino, in 1994. He has been a Postdoctoral Research Fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands), from 1994 to 1995, an Assistant Professor of Telecommunications with the Politecnico di Torino, an Associate Professor with the University of Parma, Italy, a Professor with the Department of Mobile Communications, Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA. He is currently an Alexander von Humboldt Professor with the Faculty of Electrical Engineering and Computer Science, Technical University of Berlin, Germany. His main research interests include the field of communications theory, information theory, channel, and source coding with particular focus on wireless communications. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society and Information Theory Society Joint Paper Award in 2004 and in 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, the Vodafone Innovation Prize in 2015, the ERC Advanced Grant in 2018, the Leonard G. Abraham Prize for Best IEEE JSAC Paper in 2019, and the IEEE Communications Society Edwin Howard Armstrong Achievement Award in 2020. He was a recipient of the 2021 Leibinz Prize of the German National Science Foundation (DFG). He has served in the board of governors of the IEEE Information Theory Society from 2004 to 2007, and as an Officer from 2008 to 2013. He was the President of the IEEE Information Theory Society in 2011.