Multiple-Antenna-Assisted Non-Orthogonal Multiple Access

Yuanwei Liu, Hong Xing, Cunhua Pan, Arumugam Nallanathan, Maged Elkashlan, and Lajos Hanzo

Abstract

Non-orthogonal multiple access (NOMA) is potentially capable of circumventing the limitations of the classic orthogonal multiple access schemes. Hence, it has recently received significant research attention in both industry and academia. This article is focused on exploiting multiple antenna techniques in NOMA networks, with an emphasis on investigating the rate region of MIMO-NOMA, while reviewing two popular multiple antennas aided NOMA structures, as well as underlining resource management problems of both single-carrier and multi-carrier MIMO-NOMA networks. This article also points out several effective methods of tackling the practical implementation constraints of multiple-antenna NOMA networks. Finally, some promising open research directions are provided in the context of multiple-antenna-aided NOMA.

Introduction

Given the popularity of bandwidth-thirsty multimedia applications, such as online gaming and virtual reality, the bandwidth demand for high-rate services has been higher. Moreover, the proliferation of Internet of Things (IoT) devices imposes additional challenges on the next generation networks. Non-orthogonal multiple access (NOMA), is potentially capable of improving the spectral efficiency while supporting the connectivity of a myriad of devices [1]. Hence, NOMA has been considered as a promising candidate for the fifth generation (5G) networks [2]. The key concept of NOMA relies on allowing multiple users to occupy the same resource block, while identifying users based on their different power levels. More particularly, NOMA applies superposition coding (SC) at the transmitters for multiplexing users within the power domain and invokes successive interference cancellation (SIC) at receivers for detection.

We focus our attention on investigating the family of multiple-antenna-aided NOMA systems, since it is important to explore the spatial degrees of freedom for improving spectral efficiency. A remarkable advantage of a multiple-antenna-aided NOMA design is that it is capable of providing array gains by invoking directional beamforming or by increasing the system's throughput by applying spatial multiplexing. Moreover, another astute application of multiple antennas in NOMA systems is based on creating unique, user-specific channels by adopting appropriate transmit precoding matrix designs. These extra manipulations are capable of eliminating the channel difference constraints of NOMA, which leads to a generalized NOMA design for satisfying the heterogeneous quality of service (QoS) requirements of users. Although, as mentioned above, there are potential benefits, when applying multiple antennas in NOMA, numerous open research challenges arise, which motivate us to develop this article.

The main issues in the context of multiple-antenna NOMA addressed by this article are summarized as follows:

1. The rate region gains of multiple-input multiple-output (MIMO)-NOMA compared to MIMO-OMA are presented from a theoretical perspective.
2. A pair of representative multiple-antenna-aided designs are illustrated. Based on those designs, the resource allocation problems of multiple-antenna-aided NOMA are discussed in terms of both single-carrier and multi-carrier forms.
3. The associated practical implementation issues are identified and a range of promising solutions were discussed.

Multiple-Antenna NOMA: Information-Theoretic Perspective

In this article, we confine our discussions of the multiple-antenna NOMA downlink to specific types of multi-user broadcast channels (BCs). In this section, we discuss the fundamental limits of the multi-user BC, revealing that NOMA is capable of improving the downlink spectral efficiency, because information-theoretically it is optimal in terms of its achievable rate region in several important special cases.

MISO-NOMA

It is widely recognized that the capacity region of a *degraded* BC is achievable by using superposition coding at transmitters (Txs) and SIC at receivers (Rxs). Specifically, assuming a two-user single-input single-output (SISO) case with the users ordered naturally by their channel gains (e.g., $|h_n|^2 > |h_m|^2$), the condition of $R_{n\to m} > R_{m\to m}$)[1] that guarantees successful SIC is *automatically* satisfied.

On the other hand, assuming a base station (BS) equipped with more than one transmit anten-

This work was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/N029666/1. L. Hanzo would like to acknowledge the ERC's financial support of his Advance Fellow Grant.

[1] $R_{i\to j}$ denotes the rate at which user *i* decodes user *j*'s message throughout the article.

Yuanwei Liu, Hong Xing, Cunhua Pan, Arumugam Nallanathan, and Maged Elkashlan are with Queen Mary University of London; Lajos Hanzo is with the University of Southampton.

Digital Object Identifier: 10.1109/MWC.2018.1700080

1536-1284/18/$25.00 © 2018 IEEE

	Capacity region	DPC rate region	NOMA rate region
SISO	=	=	=
MISO/ MIMO	$\begin{cases} = \text{DPC; } \textit{special cases} \\ \textbf{unknown}\text{; } \geq \text{DPC [3]. in general} \end{cases}$	$\begin{cases} = \text{NOMA; } \textit{special cases}\text{, e.g., (}\textit{quasi-}\text{) deg}\textit{raded}\text{ [4]} \\ \geq \text{NOMA: in general, given the same encoding/decoding order} \end{cases}$	Assuming fixed decoding order $m \rightarrow n$: successful SIC $\Leftrightarrow R_n \rightarrow m > R_m \rightarrow m$

TABLE 1. Relationships among the rate regions achieved by NOMA and others.

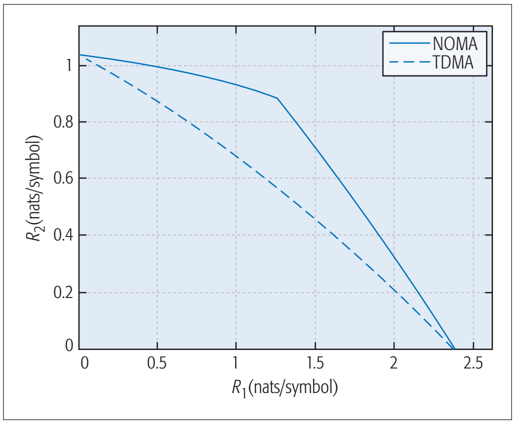


FIGURE 1. The boundaries of the two-user ADBC rate regions of both NOMA and TDMA, in conjunction with $\mathbf{N}_1$ = [.5,.18;.18,.7], $\mathbf{N}_2$ = [.7, .08;.08, 10.7], and $\mathbf{S}$ = [1, .6;.6,2]. Here, $\mathbf{N}_i$, i = 1, 2, denotes the additive noise covariance matrix at the ith Rx, and $\mathbf{S}$ is the given matrix covariance matrix.

na serving two users each associated with $(_i$ (i = m, n) receive antennas, even when we have r_m = r_n = 1, there are no known results on its capacity region of this general Gaussian multiple-input single-output (MISO) BC. Hence, only the dirty paper coding (DPC) rate region is recognized as being achievable, which is in general larger than NOMA's rate region assuming the same fixed DPC encoding and SIC decoding order. The reason for the reduction of the NOMA rate compared to DPC is that for two-user MISO downlink transmission assuming a fixed order of $m \rightarrow n$, DPC ensures that the achievable rate of user n is $R_n = \log_2(1 + |\mathbf{h}_n^{\mathbf{H}}\mathbf{w}_n|^2)$, where $\mathbf{h}_n$ and $\mathbf{w}_n$ are the channel vectors and beamformer vectors of user n, respectively. This is because the interference caused by user m has been assumed to be non-causally known and thus has been pre-cancelled by the Tx. By contrast, NOMA, whose rate region is achieved by SIC, entails the extra constraint of $R_{n \rightarrow m} > R_{m \rightarrow m}$ so that once the message of user m is successfully decoded, it can also be successfully remodulated and cancelled by user n. As a result, the rate region achieved by NOMA is contained with that achieved by DPC. Analogous to the SISO case, one may assume another natural ordering of the users, according to $||\mathbf{h_n}|| > ||\mathbf{h_m}||$), which, however, does not necessarily yield $R_{n \rightarrow m} > R_{m \rightarrow m}$, as can be readily verified by simple calculation.

It is also intriguing to find in the literature that in some *special cases,* MISO NOMA is capable of achieving the same performance as DPC. For instance, the sufficient and necessary conditions for a *quasi-degradation* were recently developed in [4] in order to bridge the gap between NOMA and DPC.

MIMO-NOMA

Similar to the MISO case, the capacity region of a general MIMO downlink transmission is still unknown, while the DPC rate region coincides with the capacity region of the MIMO BC in several special cases, such as that of the *aligned* and *degraded* MIMO BC (ADBC), and that of the *aligned MIMO BC* (not necessarily degraded) [3]. More particularly, it was shown in [3] that the NOMA rate region under a covariance matrix input constraint of $\mathbf{S}$ ($\mathbf{S} \preceq \mathbf{0}$) is readily achievable for the ADBC. Furthermore, it was also shown that in this case we have capacity region = DPC rate region = NOMA rate region. The relations among the rate regions achieved by the different transmission schemes considered are summarized in Table 1.

We continue by providing a numerical example for the ADBC downlink. The MIMO BC is said to be aligned if the number of transmitt antennas is equal to the number of receive antennas at each of the Rxs (i.e., we have $t = r_m = r_n$), and the channel gain matrices are all identity matrices. In contrast, the MIMO BC is said to be degraded if the covariance matrices of the additive Gaussian noise at each of the Rxs are ordered as **0** prec bf $\mathbf{N_n} \in \mathbf{N_m}$).[2] We compare in Fig. 1 the downlink NOMA rate region (i.e., the capacity region) of a two-user (2 × 2) ADBC against that achieved by an OMA scheme, namely classic time-division multiple access (TDMA).

Multiple-Antenna NOMA: Beamformer-Based Structure

While the previous section has laid the foundation for multi-antenna NOMA from a theoretical perspective, in the following sections we discuss the promising multiple-antenna-aided NOMA solutions. Broadly speaking, current applications can be primarily classified into a pair of categories: the beamformer-based structure and cluster-based structure. The key difference between these two structures is whether one beamformer serves multiple users or one user. In this section, we focus on investigating each beamformer design for each single user, by invoking both the centralized and coordinated beamforming approaches.

Centralized Beamforming

The centralized beamforming terminology is used in this article to indicate that the transmit precoding (TPC) is carried out by a single BS, and there is no coordination at the BS. In centralized beamforming, the BSs are equipped with M antennas. As shown in Fig. 2a, let us consider a simple two-user downlink MISO scenario as our

[2] $\mathbf{A} \preceq \mathbf{B}$ denotes ($\mathbf{A} - \mathbf{B}$) is a negative semi-definite matrix.

example, where the BS transmits a superposition of individual messages of user m and user n with the aid of two beamformers, $\mathbf{w}_m$ and $\mathbf{w}_n$, specifically constructed for each user. We denote the channel vectors of user m and user n by $\mathbf{h}_m$ and $\mathbf{h}_n$, respectively. User n decodes the message of user m at the rate of

$$R_{n\to m} = \log_2\left(1+\frac{\left|\mathbf{h}_n^H\mathbf{w}_m\right|^2}{\left|\mathbf{h}_n^H\mathbf{w}_n\right|^2+\sigma^2}\right),$$

where σ^2 is the noise variance. If this operation is successful, user n invokes the classic SIC and then decodes its own message at the rate of

$$R_{n\to n} = \log_2\left(1+\frac{\left|\mathbf{h}_n^H\mathbf{w}_n\right|^2}{\sigma^2}\right).$$

As for user m, it will directly decode its own message by treating the message of user n as interference, which is given by

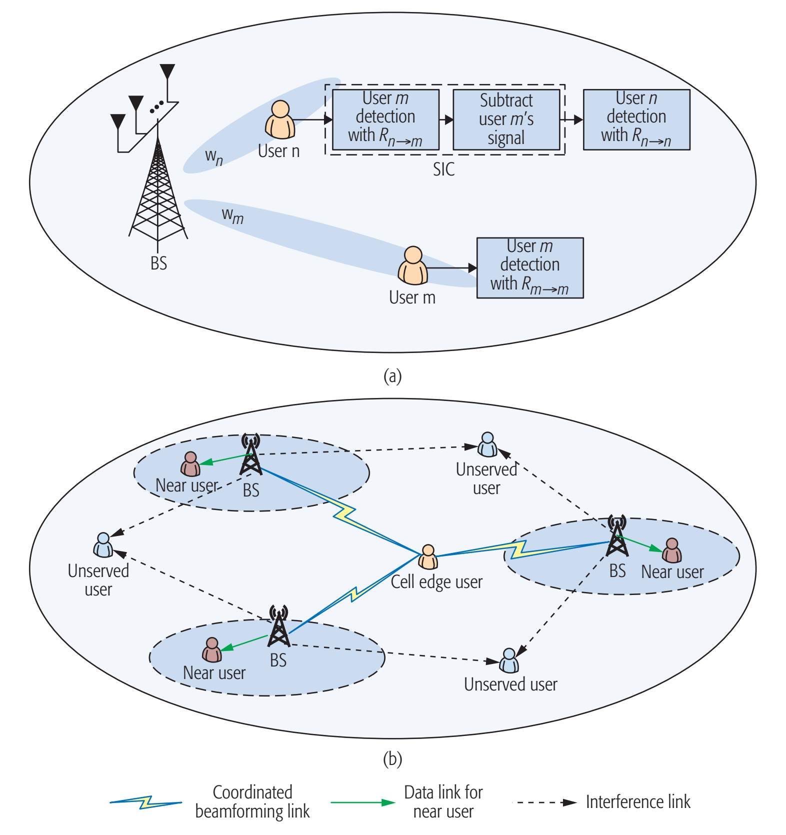
$$R_{m\to m} = \log_2\left(1+\frac{\left|\mathbf{h}_m^H\mathbf{w}_m\right|^2}{\left|\mathbf{h}_m^H\mathbf{w}_n\right|^2+\sigma^2}\right).$$

As mentioned earlier, it is worth pointing out that successful SIC can only be guaranteed if the condition of $R_{n\to m} > R_{m\to m}$ is satisfied. Nevertheless, this constraint will make the resultant optimization problem very challenging to solve, since some existing research contributions of MISO NOMA rely on alternative approaches to circumvent this issue, such as assuming a significantly different path loss for the users or assigning a predefined decoding order for each user. Furthermore, finding the optimal ordering for MISO NOMA is still an open problem; hence, further research efforts are expected to find the optimal performance bound.

Coordinated Beamforming

In addition to the centralized beamforming approach, coordinated beamforming is another effective technique of invoking multiple-antenna technologies. The concept of coordinated beamforming relies on the cooperation of a number of single-antenna devices for the sake of forming virtual antenna arrays, which, however, relies on setting aside much of the achievable capacity for inter-node information exchange. Figure 2 illustrates a possible implementation of coordinated beamforming in NOMA. More particularly, several BSs are engaged in coordinated beamforming to serve a cell edge user, where each BS serves a single user within its own cell by applying SIC for cancelling the intra-cell interference received from the cell edge user. By doing so, the performance of the cell edge user is enhanced, which results in better fairness for the entire network.

The initial idea of coordinated beamforming of NOMA, which is similar to the coordinated multipoint (CoMP) transmission scheme, was proposed in [5], where two coordinated BSs invoke an Alamouti-code-based coordinated SC to simultaneously serve a pair of users in each others' vicinity as well as a cell edge user. As an evolution of the single-antenna-based systems of [5], the authors also considered multiple-antenna-assisted BSs and users in [6], which was essentially a coordinated MIMO-NOMA system setup. In particular, a joint centralized and coordinated beamforming design was developed for suppressing the inter-cell interference as well as for enhancing the throughput of the cell edge users.



BS
w_n
User n
User m detection with $R_{n\to m}$
Subtract user m's signal
User n detection with $R_{n\to n}$
SIC
w_m
User m
User m detection with $R_{m\to m}$
(a)
Near user
BS
Unserved user
Unserved user
Cell edge user
BS
Near user
BS
Near user
Unserved user
(b)
Coordinated beamforming link
Data link for near user
Interference link

FIGURE 2. Beamformer-based structure of multiple-antenna-aided NOMA: a) illustration of centralized beamforming of multiple-antenna-aided NOMA; b) illustration of coordinated beamforming of multiple-antenna-aided NOMA.

Multiple Antenna NOMA: Cluster-Based Structure

Another popular multiple-antenna NOMA design relies on partitioning the users into several different clusters, where the users in a specific cluster share the same beamformers. Then, by applying appropriate TPC and detector designs, the inter-cluster interference can be suppressed. In this section, we introduce two typical cluster-based designs, depending on whether the inter-cluster interference can be completely cancelled.

Inter-Cluster Interference-Free Design

In an effort to tackle the channel ordering, an appealling low-complexity technique is that of decomposing the MIMO-NOMA channels to multiple SISO-NOMA channels [7, 8]. As shown in Fig. 3a, a BS equipped with M antennas communicates with $K = \sum_{m=1}^{M} L_m$ users, who are randomly partitioned into M clusters and equipped with n antennas each. The spirit of this decomposition-based design is to adopt zero-forcing detection for each user, which results in a

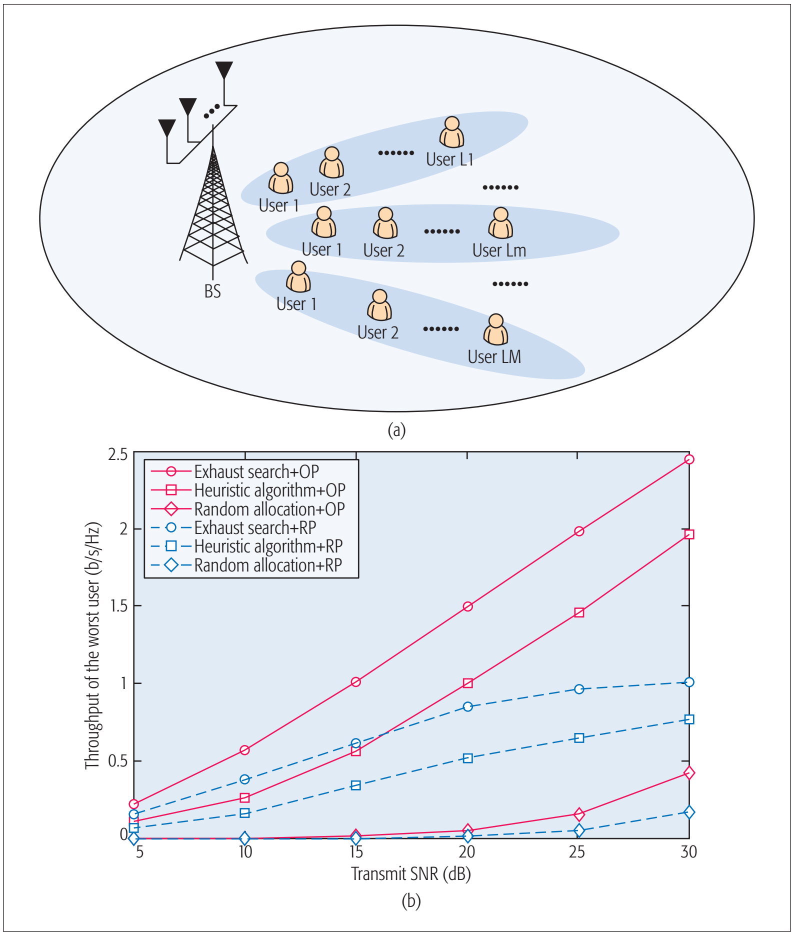


FIGURE 3. Cluster-based structure of multiple-antenna-aided NOMA: a) illustration of cluster-based structure for multiple-antenna NOMA; b) resource allocation for guaranteeing the fairness of cluster-based MIMO-NOMA networks [11]. OP refers to optimal power allocation, which is obtained by bi-section search. RP refers to random power allocation.

low-complexity SISO-NOMA model. Hence, the conventional NOMA technique can be applied. A range of TPC designs can also be correspondingly invoked at the BS [8].

A remarkable advantage of this decomposition-aided design is that by transforming to SISO-NOMA, the sophisticated channel ordering may be circumvented, which reduces the system complexity. Moreover, this design is capable of completely cancelling the inter-cluster interference. Nevertheless, the zero-forcing detection adopted relies on a specific relationship between the number of transmitter antennas and receiver antennas. Specifically, either the condition of $N \geq M$ of [7] or that of $N \geq M/2$ of [8] has to be satisfied. It is worth pointing out that fairness is of great significance in the cluster-based NOMA structure. Figure 3b illustrates the resource allocation conceived for maintaining max-min fairness in clustered MIMO-NOMA scenarios, applying a heuristic algorithm for user scheduling and bi-section-based search for optimal power allocation. Further comparisons relying on other optimization techniques are detailed in Table 2.

Inter-Cluster Interference-Tolerant Design

In contrast to the inter-cluster interference-free design mentioned in the last subsection, we introduce another cluster-based MIMO-NOMA design, which was proposed in [9]. This design allows the existence of inter-cluster interference and applies the so-called decoding scaling weight for increasing the strength of the desired signals. We consider Fig. 3a as our example, and in contrast to the random clustering in [7, 8], the user clustering in [9] follows specific techniques for making the channel gains of users more distinctive. The users within a specific cluster are sorted according to their equivalent normalized channel gain. As a result, the user experiencing the highest channel gain is the cluster head, capable of completely cancelling all intra-cluster interference by invoking SIC. The key idea behind this detection approach is that the BS has to send the decoding scaling weight to the users prior to the data transmission process. The inter-cluster interference can be efficiently suppressed by exploiting the fact that all cluster users except for the cluster head will estimate their own equivalent cluster channels.

The key advantage of this inter-cluster interference-tolerant design is that it does not imposed any constraints at the BS and users on their numbers of antennas. Hence, such designs can be directly extended to the scenarios, where the BS is equipped with a large antenna array, as in massive-MIMO-NOMA or NOMA millimeter-wave communication scenarios [10]. Actually, a large antenna array will facilitate highly directional transmission, which in turn results in correlations among the users. This channel characteristic is ideal for the application of NOMA principles in the context of large antenna array systems.

Resource Allocation for Multiple-Antenna NOMA

Due to the complex nature of interference in multiple-antenna-aided NOMA networks, especially in the cluster-based design, one of the main challenges is to enhance both the spectral and energy efficiency by exploiting the sophisticated reuse of resources, with the aid of appropriately designed beamformers, by opportunistic scheduling of users into different clusters by applying efficient algorithms, and by intelligent power allocation. In this section, we first introduce the resource allocation problems of multiple-antenna-aided NOMA scenarios from the perspective of both single-carrier and multi-carrier solutions, and then summarize some potential mathematical modeling techniques and tools conceived for tackling these problems.

Single-Carrier Resource Allocation

We commence our discussion on resource allocation in terms of MIMO-NOMA in a single resource block, relying on a single carrier and the same time slot or spreading code. We still use the cluster-based MIMO-NOMA structure of Fig. 3a as an example, where several resource allocation problems have to be carefully tackled:

- The number of clusters
- The number of NOMA users to allocate to each cluster
- Which users should be assigned to which clusters
- Power sharing among the different clusters as well as among the users within the same cluster.

Several MIMO-NOMA resource allocation contributions have considered a simplified model with a particular focus on tackling the third and fourth issues, such as fixing both the number as well as the size of clusters [9, 11]. More specifically, dynamic user scheduling and power allocation problems were optimized in [9, 11], aiming to maximize the system's throughput and addressing the max-min fairness of clustered MIMO-NOMA scenarios, respectively, when utilizing a particular TPC and decoder. Moreover, optimally solving all the aforementioned problems is a rather challenging problem, which constitutes a promising research direction.

Multi-Carrier Resource Allocation

Multi-carrier MIMO-NOMA constitutes a natural extension of single-carrier MIMO-NOMA, which relies on the multiplexing also in the frequency domain in addition to the power and spatial domains. Multi-carrier resource allocation for MIMO-NOMA can be viewed as a hybrid NOMA resource allocation problem, which aims to simultaneously manage multi-dimensional resources. As the number of MIMO-NOMA users to be served in a single resource block increases, the intra/inter beams interference suppression techniques have to become more sophisticated, which may also require a large number of antennas at the BS. Driven by this, we can first partition the users into different subcarrier bands, which are orthogonal to each other. In the same subband, MIMO-NOMA designs are invoked. By adopting such designs, the system's implementational complexity can be significantly reduced.

Some initial research results are already available on multi-carrier SISO-NOMA [12, 13], but multi-carrier MIMO-NOMA designs are still in their infancy. Note that the resource allocation of multi-carrier MIMO-NOMA requires more sophisticated design, which includes two rounds of user scheduling as well as power allocation for both the subcarriers and MIMO clusters. It is worth pointing out that other practical forms of multi-carrier NOMA, such as sparse code multiple access (SCMA) and pattern-division multiple access (PDMA) may also be combined with MIMO techniques in the context of generalized multi-carrier MIMO NOMA resource allocation designs.

Approaches for Resource Allocation

Given the distinct characteristics of MIMO-NOMA channels and the interference constraints, the optimization problems formulated for resource allocation usually constitute a mixed-integer non-convex problem. There are two popular methods for tackling this kind of problem. The *first* one is based on the joint optimization of both user scheduling and power allocation [13]. Representative approaches invoked for solving the joint optimization problems can be monotonic optimization and Branch-and-Bound, which may result in globally optimal solutions. The *second* one is to decouple the user scheduling and power allocation into two sub-problems to be optimized [9, 11, 12]. More particularly, matching theory can be an effective technique of moderate complexity for scheduling users [12]. Regarding power allocation, geometric programming and non-cooperation games are promising tools for allocating the power in an intelligent way. Table 2 summarizes some representative optimization strategies, which can be potentially applied for tackling the resource allocation problem in MIMO-NOMA scenarios.

Categories	Optimization variables	Optimization approaches	Characteristics	Ref.
Jointly	User scheduling & power allocation	• Monotonic optimization	Optimality achievable	[13]
		• Branch-and-Bound	Optimality achievable	–
Decoupled	User scheduling	• Matching theory	Moderate complexity	[12]
		• Heuristic algorithms	High flexibility	[9, 11]
	Power allocation	• Geometric programming	Low complexity	[12]
		• Lagrangian algorithms	Closed-form solutions	[9]
		• Bi-section search	Low complexity	[11]

TABLE 2. Summary of representative optimization approaches for resource allocation in MIMO-NOMA.

Tackling the Practical Implementation Constraints of Multiple-Antenna NOMA

Although the application of multiple-antenna techniques in NOMA potentially enhances the performance of networks upon scaling up the number of antennas, it also imposes several implementational constraints in practical scenarios, including potential overhead, sophisticated TPC and detector design, energy and security related issues, and so on. Motivated by this, in this section, we provide several effective solutions for tackling these implementational constraints.

Reducing Complexity with Imperfect CSI

The sophisticated TPC and detector designs of MIMO-NOMA may impose high feedback overhead compliant with tight specifications, which raise high requirements for channel estimations. Although many existing research contributions are based on the idealized simplifying assumption of having perfect channel state information (CSI), in practical systems, this cannot be achieved. Hence, low-complexity imperfect-CSI-based designs are desired for MIMO-NOMA networks. There are two popular approaches:

1. The first approach is to rely on partial CSI, such as the path loss, which does not fluctuate rapidly.
2. The second approach is to use limited feedback for reducing the system's overhead, which is particularly significant for MIMO-NOMA networks.

By doing so, each user feeds back a limited number of bits to the BS, which may be the TPC codebook index.

Tackling Energy Issues with Multi-Antenna-Aided Wireless Power Transfer

Given the fact that NOMA is capable of supporting massive conductivity, it is eminently suitable for the IoT. However, the energy is severely limited in IoT scenarios, especially in wireless sensor networks where the user equipments (UEs) are usually energy constrained. This motivates the application of a new member of the energy harvesting family, simultaneous wireless information and power transfer (SWIPT), which

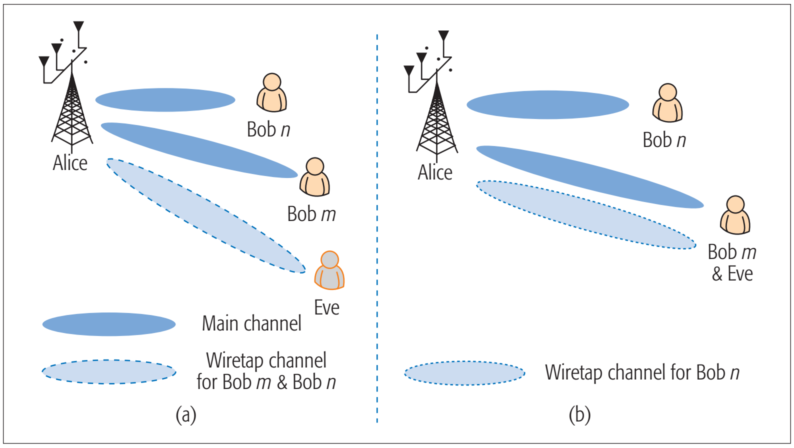


FIGURE 4. Illustration of PLS of multiple-antenna NOMA.

was initially proposed for cooperative NOMA scenarios in [14]. Multi-antenna techniques are capable of supporting the application of SWIPT in NOMA networks, which opens up several exciting opportunities. However, these applications also pose new challenges in terms of jointly designing the energy and information beams, which requires further research efforts in this area.

Low-Complexity Design with Antenna Selection

The classic antenna selection (AS) technique can be invoked as an effective solution for MIMO-NOMA networks due to the fact that it brings about two distinct benefits. The first one is that AS reduces the hardware costs imposed by the expensive RF chains without losing the spatial diversity, which is the advantage of AS for conventional MIMO. The second one, which is also quite significant, is that AS transforms MIMO-NOMA to SISO-NOMA in a straightforward manner; hence, the sophisticated channel ordering operation of MIMO-NOMA can be avoided, leading to an appealing performance-complexity trade-off. It is worth pointing out that the design of effective algorithms for selecting antennas at both transmitters and receivers to strike an attractive performance-complexity trade-off is a promising research area.

Security Provision Relying on Multiple Antennas

Invoking physical layer security (PLS) is beneficial in NOMA networks in order to counteract the broadcast nature of wireless transmissions as well as the security threat of SIC. When dealing with security issues, exploiting multiple-antenna-aided NOMA is capable of enhancing the PLS by increasing the desired destination user's capacity. Alternatively, artificial noise (AN) may be invoked on the eavesdroppers (Eves) without degrading the reception of the desired user (Bob). This AN aided approach was applied in a MISO-NOMA scenario, where eavesdroppers are external [15], as shown in Fig. 4a. By contrast, for the scenarios when the internal NOMA users are the Eves, as seen in Fig. 4b, Bob *M* having a poor channel tries to detect the message of Bob *n* having a good channel, multiple antennas can be used to artificially create the required effective channel differences between two users, which is beneficial for preventing eavesdropping. It is worth pointing out that how to prevent Bob *n* from detecting the signal of Bob *m* is still an open problem.

Conclusions and Promising Research Directions

In this article, the application of multiple-antenna techniques to NOMA has been exploited. The capacity gain of MIMO-NOMA over MIMO-OMA has been first demonstrated from an information-theoretic perspective. Then two dominant multiple-antenna NOMA structures, namely the beamformer-based and cluster-based designs, have been highlighted. The resource allocation problems of MIMO-NOMA networks have also been identified, followed by discussing several implementation issues as well as the corresponding potential solutions conceived for multiple-antenna-aided NOMA. There are still numerous open research problems in this area, which are listed as follows.

Spatial Effect Investigation: Stochastic geometry has been recognized as a powerful mathematical tool used for capturing the topological randomness of large-scale networks. Some initial stochastic-geometry-based investigations have been conducted in the context of NOMA relying on the binomial point process (BPP) and Poisson point process (PPP) [8, 14, 15], leading to the more practical but challenging Poisson cluster process (PCP). By utilizing multiple-antenna arrays at the BS, more effective directional beamforming design can be adopted, while considering the spatial position of BSs and NOMA users on the attractive system performance, which is a promising research direction.

Optimal Decoding Order Design: The SIC decoding order has a significant impact on the performance of NOMA networks. Compared to the SISO-NOMA systems, the optimal ordering design problem of MIMO-NOMA is more challenging, since it depends on the TPC and the detector design. Most of the existing research contributions on MIMO-NOMA were based on a particular ordering, which does not result in approaching the optimal performance bound. Such optimal designs still constitute an open area.

Modulation Design for MIMO-NOMA: Given the maturity of orthogonal frequency-division multiplexing (OFDM), the MIMO-NOMA designs are expected to be incorporated into OFDM. Various other novel modulation schemes have also been proposed for 5G networks, which have to be investigated.

Emerging MIMO-NOMA for 5G: Massive MIMO and millimeter-wave, as two important technologies advocated for the forthcoming 5G networks, are capable of enhancing the attainable system performance as a benefit of their large antenna array gain and large bandwidth, respectively. NOMA is expected to coexist with these two technologies for further improving the spectral efficiency as well as for supporting massive connectivity. However, the distinct characteristics of a large number of antennas at the BS inevitably necessitates the redesign of TPC and detection techniques, which require more research contributions in this field.

References

[1] Y. Liu *et al.*, "Nonorthogonal Multiple Access for 5G and Beyond," *Proc. IEEE*, vol. 105, no. 12, Dec. 2017, pp. 2347–81.

[2] L. Dai *et al.*, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 74–81.

[3] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel," *IEEE Trans. Info. Theory*, vol. 52, no. 9, Sept. 2006, pp. 3936–64.

[4] Z. Chen *et al.*, "On the Application of Quasi-Degradation to MISO-NOMA Downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, Dec. 2016, pp. 6174–89.

[5] J. Choi, "Non-Orthogonal Multiple Access in Downlink Coordinated Two-Point Systems," *IEEE Commun. Lett.*, vol. 18, no. 2, Feb. 2014, pp. 313–16.

[6] W. Shin *et al.*, "Coordinated Beamforming for Multi-Cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, Jan. 2017, pp. 84–87.

[7] 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, Jan. 2015, pp. 537–52.

[8] 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, June 2016, pp. 4438–54.

[9] M. S. Ali, E. Hossain, and D. I. Kim, "Non-Orthogonal Multiple Access (NOMA) for Downlink Multiuser MIMO Systems: User Clustering, Beamforming, and Power Allocation," *IEEE Access*, vol. 5, 2017, pp. 565–77.

[10] B. Wang *et al.*, "Spectrum and Energy Efficient Beamspace MIMO-NOMA for Millimeter-Wave Communications Using Lens Antenna Array," *IEEE JSAC*, vol. 35, no. 10, Oct. 2017, pp. 2370–82.

[11] Y. Liu *et al.*, "Fairness of User Clustering in MIMO Non-Orthogonal Multiple Access Systems," *IEEE Commun. Lett.*, vol. 20, no. 7, July 2016, pp. 1465–68.

[12] B. Di, L. Song, and Y. Li, "Sub-Channel Assignment, Power Allocation, and User Scheduling for Non-Orthogonal Multiple Access Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, Nov. 2016, pp. 7686–98.

[13] Y. Sun *et al.*, "Optimal Joint Power and Subcarrier Allocation for Full-Duplex Multicarrier Non-Orthogonal Multiple Access Systems," *IEEE Trans. Wireless Commun.*, vol. 65, no. 3, Mar. 2017, pp. 1077–91.

[14] Y. Liu *et al.*, "Cooperative Non-Orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer," *IEEE JSAC*, vol. 34, no. 4, Apr. 2016.

[15] Y. Liu *et al.*, "Enhancing the Physical Layer Security of Non-Orthogonal Multiple Access in Large-Scale Networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, Mar. 2017., pp. 1656–72

Biographies

Yuanwei Liu (yuanwei.liu@qmul.ac.uk) is a lecturer (assistant professor) with Queen Mary University of London, where he received his Ph.D. degree in 2016. His research interests include 5G wireless networks, the Internet of Things, and stochastic geometry. He received Exemplary Reviewer Certificates from *IEEE Wireless Communications Letters* and *IEEE Transactions on Communications*. He serves as an Editor of *IEEE Communications Letters* and *IEEE Access*.

Hong Xing (h.xing@qmul.ac.uk) received her B.Eng. degree in electronic sciences and technologies from Zhejiang University, China, in 2011, and her Ph.D. degree in wireless communications from King's College London, United Kingdom, in 2015. She was a research associate with the Department of Informatics, King's College London, from 2016 to 2017. She is currently a research associate with the School of EECS, Queen Mary University of London.

Cunhua Pan (c.pan@qmul.ac.uk) received his B.S. and Ph.D. degrees from Southeast University, Nanjing, China, in 2010 and 2015, respectively. From 2015 to 2016, he worked as a research associate at the University of Kent, United Kingdom. He is currently a research fellow with Queen Mary University of London. His research interests include ultra-dense C-RAN, UAV, IoT, NOMA, and mobile edge computing. He serves as the Student Travel Grant Chair for ICC 2019.

Arumugam Nallanathan (a.nallanathan@qmul.ac.uk) has been a professor of wireless communications at Queen Mary University of London since September 2017. He was with King's College London from 2007 to 2017, and the National University of Singapore from 2000 to 2007. His research interests include 5G wireless networks, the Internet of Things, and molecular communications. He received the Best Paper Awards at IEEE ICC 2016 and IEEE GLOBECOM 2017. He is an IEEE Distinguished Lecturer. He is a Web of Science Highly Cited Researcher in 2016.

Maged Elkashlan (maged.elkashlan@qmul.ac.uk) received his Ph.D. degree in electrical engineering from the University of British Columbia in 2006. From 2007 to 2011, he was with the Commonwealth Scientific and Industrial Research Organization (CSIRO) Australia. In 2011, he joined the School of Electronic Engineering and Computer Science at Queen Mary University of London. He serves as an Editor of *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *IEEE Transactions on Molecular, Biological and Multi-scale Communications*.

Lajos Hanzo [F'04] (lh@ecs.soton.ac.uk), FREng, FIET, Fellow of EURASIP, D.Sc., received his degree in electronics in 1976 and his doctorate in 1983. He holds honorary doctorates from the Technical University of Budapest (2009) and the University of Edinburgh (2015). He is a member of the Hungarian Academy of Sciences and a former Editor-in-Chief of IEEE Press. He is a Governor of both IEEE ComSoc and VTS. He has published 1700+ contributions at IEEE Xplore and 18 Wiley-IEEE Press books. He has successfully supervised 112 Ph.D. students.