Multiple Antenna Assisted Non-Orthogonal Multiple Access

Yuanwei Liu, *Member, IEEE,* Hong Xing, *Member, IEEE,* Cunhua Pan, *Member, IEEE,* Arumugam Nallanathan, *Fellow, IEEE,* Maged Elkashlan, *Member, IEEE,* and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Non-orthogonal multiple access (NOMA) is potentially capable of circumventing the limitations of the classic the orthogonal multiple access schemes, hence it has recently received significant research attention both in industry and academia. This article is focused on exploiting multiple antenna techniques in NOMA networks, with an emphasis on investigating the rate region of multiple-input multiple-output (MIMO)-NOMA, whist 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 context of multiple antenna aided NOMA.**

I. 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 whilst supporting the connectivity of a myriad of devices. Hence NOMA has been considered as a promising candidate for the fifth generation (5G) networks [1]. The key concept of NOMA relies on allowing multiple users to occupy the same resource block, whilst identifying users based on their different power levels [2]. 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.

Given the extensive research on single-antenna NOMA, in this article, we focus our attention on investigating the family of multiple-antenna aided NOMA systems, which is motivated by exploring the spatial degrees of freedom for improving the spectral efficiency. For instance, the multiple-antenna aided NOMA design is capable of providing array gains by invoking directional beamforming or increasing the system's throughput by applying spatial multiplexing. Moreover, another astute application of multiple antennas in NOMA is to create unique, user-specific effective channels by adopting appropriate 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. When aiming for exploiting the potential benefits of applying multiple antennas in NOMA, numerous open research challenges arise, which motivates us to develop this treatise.

The rest of this article is structured as follows: Section II presents the rate region gains of multiple-input multiple-output (MIMO)-NOMA compared to MIMO-OMA. A pair of representative multiple-antenna aided designs are illustrated in Section III and Section IV, respectively. In Section V, resource allocation problems of multiple-antenna NOMA are discussed. The associated practical implementation issues were identified in Section VI, which are followed by our conclusions in Section VII, including promising research directions.

II. NOMA:INFORMATION THEORETIC PERSPECTIVE

In this treatise, we confine our discussions of the multiple-antenna NOMA downlink to the specific type of multi-user broadcast channels (BCs) over which a superposition of Gaussian signals emanate from the Transmitter (Tx), and each user's receiver (Rx), relies on SIC. In this section, we discuss the fundamental limits of the multi-user BC unveiling 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.

A. MISO-NOMA

It is widely recognized that the capacity region of a *degraded* BC is achievable by using superposition coding at the Tx and SIC at the Rxs. Specifically, assuming a two-user SISO case with the users ordered naturally by their channel gains, e.g., $|h_n|^2 > |h_m|^2$, the conditions 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 equipped with more than one transmit antenna serving two users each associated with r_i $(i = m, n)$ receive antennas, even when we have $r_m = r_n = 1$, there is no known results on its capacity region

Y. Liu, H. Xing and A. Nallanathan are with King's College London, London, UK (email: {yuanwei.liu, hong.1.xing, arumugam.nallanathan}@kcl.ac.uk).

C. Pan and M. Elkashlan is with Queen Mary University of London, London, UK (email: {c.pan, maged.elkashlan}@qmul.ac.uk).

L. Hanzo is with University of Southampton, Southampton, UK (email:lh@ecs.soton.ac.uk).

IEEE Wireless Communications Magazine, 2018, accepted on 1/01/2018

The financial support of the European Research Council's Advanced Fellow Grant is gratefully acknowledged

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

of this general Gaussian multiple-input single-output (MISO) BC. Hence it is only that 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 that accounts 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 $R_n = \log_2(1 + |\mathbf{h_n^H w_n}|^{\mathbf{2}})$. 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 it 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* was recently developed in [4] in order to bridge the gap between NOMA and DPC.

B. 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 $\mathbf{S}$ ($\mathbf{S} \succeq \mathbf{0}$) 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 I.

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

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

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; By 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

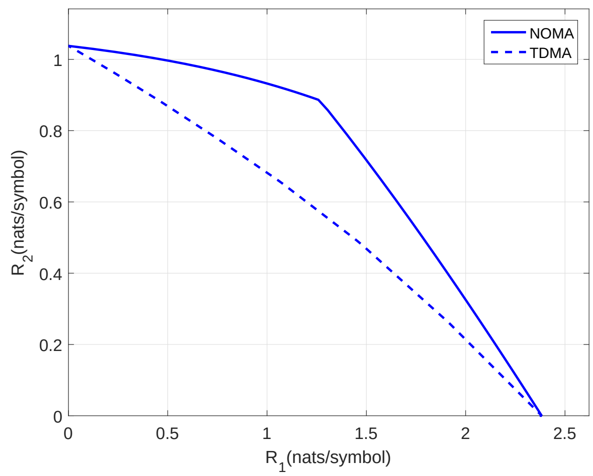


Fig. 1: The boundaries of the two-user ADBC rate regions of both NOMA and of TDMA, in conjunction with $\mathbf{N_1} = [.5, .18; .18, .7]$, $\mathbf{N_2} = [.7.08; .0810.7]$ and $\mathbf{S} = [1.6; .6, 2]$.

$\mathbf{0} \prec \mathbf{N_n} \preceq \mathbf{N_m}$[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 the classic time-division multiple access (TDMA).

III. Multiple-Antenna NOMA: Beamformer Based Structure

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

A. Centralized Beamforming

To elaborate on the centralized beamforming, the base stations (BSs) are equipped with M antennas. As shown in Fig. 2(a), let us consider a simple two-user dnowlink MISO scenario as our 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 [5]. 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 $R_{n\rightarrow m} = \log_2(1 + \frac{|\mathbf{h}_n^H \mathbf{w}_m|^2}{|\mathbf{h}_n^H \mathbf{w}_n|^2 + \sigma^2})$, 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 $R_{n\rightarrow n} = \log_2(1 + \frac{|\mathbf{h}_n^H \mathbf{w}_n|^2}{\sigma^2})$. As for

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

User m, it will directly decode its own message by treating the message of User n as interference, which is given by $R_{m\to m} = \log_2(1 + \frac{|\mathbf{h}_m^H \mathbf{w}_m|^2}{|\mathbf{h}_m^H \mathbf{w}_n|^2 + \sigma^2})$.

As mentioned in Section II, 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 [5]. 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.

B. 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. Fig. 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. 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 multi-point (CoMP) transmission scheme, was proposed in [6], 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 [6], the authors also considered multiple-antenna assisted BSs and users in [7], which is 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.

IV. 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 with applying appropriate transmit precoding (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.

A. 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 [8, 9]. As shown in Fig. 3, 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 low complexity single-input single-output (SISO)-NOMA model. Hence the conventional NOMA technique can be applied. A range of TPC designs can also be correspondingly invoked at the BS [9].

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 canceling 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 [8] or that of $N \geq M/2$ of [9] has to be satisfied.

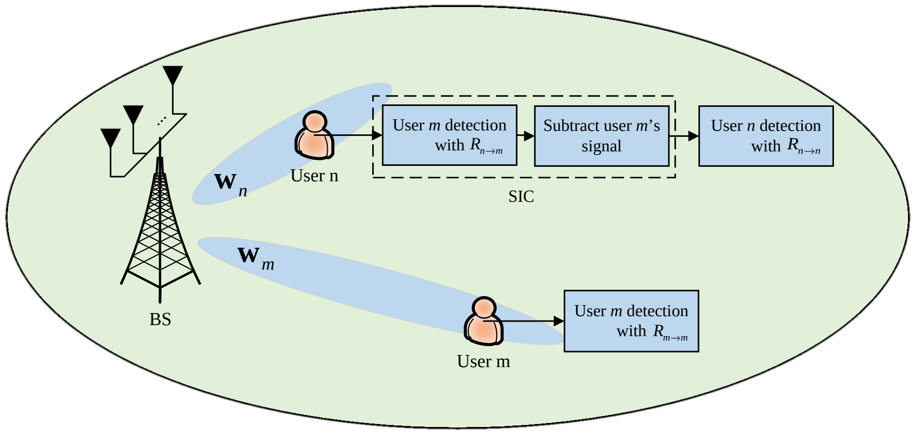
B. Inter-Cluster Interference Allowance Design

In contrast to the inter-cluster interference free design mentioned in last subsection, we introduce another cluster based MIMO-NOMA design which was proposed in [10]. 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. 3 as our example, and in contrast to the random clustering in [8, 9], the user clustering in [10] 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, which is the cluster-head, is capable of completely canceling 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.

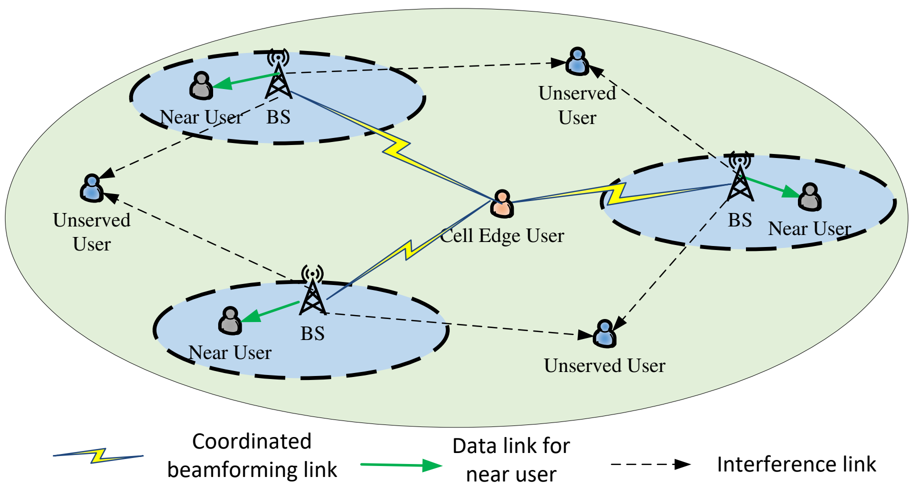
The key advantage of this inter-cluster interference tolerant design is that it does not imposed any constraints at the BS and users conceiving their number of antennas. Hence such designs can be easily extended to the scenarios where the BS is equipped with a large antenna array, in massive-MIMO-NOMA or NOMA millimeter wave communication scenarios. However, this design assigns a priority to the cluster-head and requires large channel differences for the users within the same cluster, which is different from the horizonal design principles adopted in [8, 9].

V. Resource Allocation for Multiple-Antenna NOMA

Due to the complex nature of interference in multiple-antenna aided NOMA networks, especially in the cluster based



(a) Illustration of centralized beamforming of multiple-antenna aided NOMA.



(b) Illustration of coordinated beamforming of multiple-antenna aided NOMA.

Fig. 2: Beamformer based structure of multiple-antenna aided NOMA

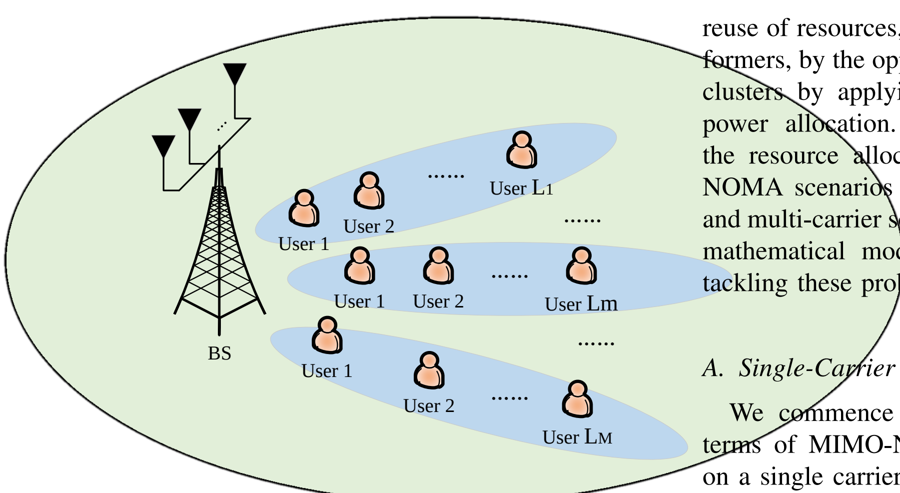


Fig. 3: Illustration of cluster based structure for multiple-antenna NOMA.

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 design beamformers, by the opportunistic scheduling of users into different clusters by applying efficient algorithms and by intelligent power allocation. In this section, we will 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 technique and tools conceived for tackling these problems.

A. 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. 3, as an example, where several resource allocation problems have to be carefully tackled: i) the number of clusters; ii) the number of NOMA users to allocate to each cluster; iii) which users should be assigned to which clusters; and iv) power sharing among the different clusters as well as among the users within the same cluster. Note that optimally solving all the aforementioned issues simultaneously is a rather challenging problem. Indeed, several resource allocation contributions of MIMO-NOMA scenarios considered a simplified model with a particular focus on tackling two or three issues, such as fixing both the number as well as the size of clusters [10, 11]. More specifically, dynamic user scheduling and power allocation problems were optimized in [10] and [11], aiming for maximizing the system's throughput and addressing the max-min fairness of clustered MIMO-NOMA scenarios, respectively, when utilizing particular TPC and decoder.

B. *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 for simultaneously managing 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 sub-carrier bands, which are orthogonal to each other. In the same subband, MIMO-NOMA design 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 power allocation for both the sub-carriers 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.

C. *Effective 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 exist two popular methods for tackling this kind of problem. The *first* one is based on the joint optimization of both user scheduling and of 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 [10–12]. 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 II 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	Characteris
Jointly	User Scheduling & Power Allocation	• Monotonic Optimization	Optimality a
		• Branch-and-Bound	Optimality a
Decoupled	User Scheduling	• Matching Theory	Moderate co
		• Heuristic Algorithms	High flexibil
	Power Allocation	• Geometric Programming	Low comple
		• Lagrangian Algorithms	Closed-form
		• Bi-section search	Low comple

TABLE II: Summary of representative optimization approaches for resource allocation in MIMO-NOMA

VI. 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, such as a potentially overhead, sophisticated TPC and detector design, energy and security related issues, etc. Motivated by this, in this section, we will provide several effective solutions for tackling these implementatonal constraints.

A. *Reducing Complexity with Imperfect CSI*

The sophisticated TPC and detector designs of MIMO-NOMA may impose a high feedback overhead complied with tight specifications, which raise high requirements for channel estimations. Although many existing research contributions are based on idealized simplifying assumption of having perfect CSI, in practical systems, this cannot be achieved. Hence low-complexity imperfect CSI based designs are desired for MIMO-NOMA networks. There exists two popular approaches: i) The first approach is to rely on partial CSI, such as the path loss, which does not fluctuate rapidly. ii) 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.

B. *Tackling Energy Issues with Wireless Power Transfer*

Given the fact that NOMA is capable of supporting massive conductivity, is eminently suitable for the Internet of Things (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—simultaneously wireless information and power transfer (SWIPT), which 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 for jointly designing the energy and information beams.

C. 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 two distinct benefits. On the one hand, AS reduces the hardware costs imposed by the expensive radio frequency (RF) chains without losing the spatial diversity. On the other hand, 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 tradeoff.

D. Security 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 user's capacity. Alternatively, artificial noise (AN) may be invoked on the eavesdroppers (Eves) without degrading the desired user's reception. This AN aided approach was applied in a MISO-NOMA scenario, where eavesdroppers are external [15], as shown in Fig. 4(a). By contrast, for the scenarios when the internal NOMA users are the Eves, as seen in Fig. 4(b), 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 to pointing out that how to prevent Bob n from detecting the signal of Bob m is still an open problem.

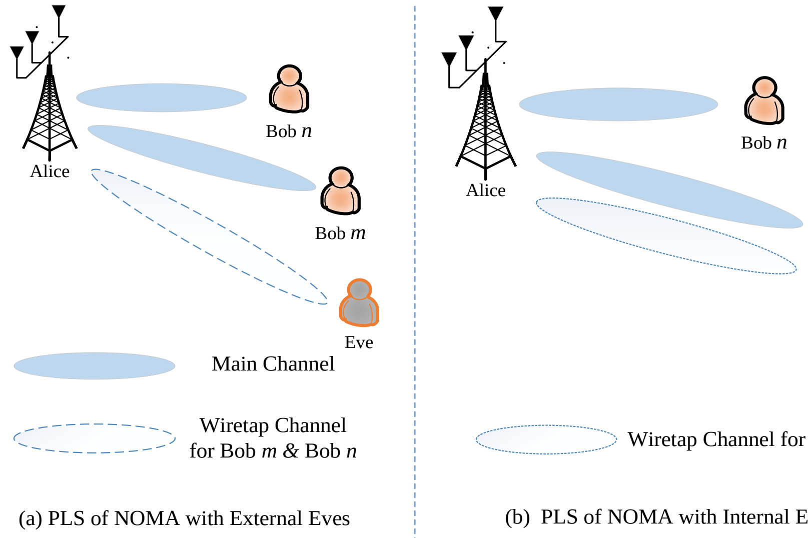


Fig. 4: Illustration of PLS of Multiple Antenna NOMA.

VII. 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 the 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 has been conducted in context of NOMA relying on the Binomial point process (BPP) and Poisson point process (PPP), 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 will be 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 on 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 co-exist 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 antennas at the BS inevitably necessitates the redesign of TPC and detection techniques, which require more research contributions in this field.

References

[1] 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.

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

[3] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "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.

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

[5] 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.

[6] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 313–316, Feb. 2014.

[7] 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.

[8] 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. 2015.

[9] 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.

[10] 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, pp. 565-577, 2017.

[11] 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.

[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, pp. 7686–7698, Nov. 2016.

[13] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Wireless Commun.*, vol. 65, no. 3, pp. 1077-1091, Mar. 2017.

[14] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, Apr. 2016.

[15] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656-1672, Mar. 2017.