Morales-Cespedes, M., Dobre, O. A. & Garcia-Armada, A. (2020). Semi-Blind Interference Aligned NOMA for Downlink MU-MISO Systems. *IEEE Transactions on Communications*, 68(3), pp. 1852–1865.

DOI: 10.1109/tcomm.2019.2960334

Semi-Blind Interference Aligned NOMA for Downlink MU-MISO Systems

Máximo Morales-Céspedes, *Member, IEEE,* Octavia Dobre, *Fellow, IEEE,* and Ana García-Armada, *Senior Member, IEEE*

***Abstract*—The application of non-orthogonal multiple access (NOMA) to downlink multi-user multiple-input single-output systems involves the design of a beamforming strategy in which the spatial dimension provided by each beam is shared among several users performing NOMA. This approach requires the management of both inter-cluster and intra-cluster interference. Moreover, the beamforming design is subject to instantaneous knowledge of the channel state information at the transmitter (CSIT). We propose a novel transmission scheme that combines blind interference alignment and NOMA. The proposed scheme fully cancels the inter-cluster interference for all users without the need for instantaneous CSIT, which is limited to the knowledge of the large scale effects of the channel in order to implement NOMA within each cluster. Considering user pairing, i.e., each cluster is composed of two users, we derive a method for determining the NOMA power coefficients that maximize the sum-rate, the user fairness or satisfy first the rate of a specific user by simply solving a polynomial function. Furthermore, we propose an alternative methodology based on some approximations in order to provide sub-optimal closed-form expressions of these NOMA power coefficients. Simulation results show that the proposed scheme outperforms conventional MISO-NOMA taking into consideration the costs of providing CSIT.**

***Index Terms*—Non Orthogonal Multiple Access, Blind Interference Alignment, Degrees of Freedom, Multiple-User Multiple-Input Single-Output, Channel State Information.**

I. Introduction

Non-orthogonal multiple access (NOMA) has received significant attention as a means of serving a large number of users with diverse data rate requirements by transmitting to more than one user in the same time, frequency or code [1]. In this sense, it is considered a candidate technique in the evolution of future wireless communications [2]. In this work, we focus on power-domain NOMA. In such a way, the transmitted symbols are superposed at the same time and frequency, each associated to a power coefficient. Thus, by either performing successive interference cancellation (SIC) or treating the interference caused by transmission to other users as noise, both the spectral efficiency and the number of users served satisfactorily can be increased. In this sense, determining the power allocation coefficients for NOMA is subject to several criteria such as maximizing the sum-rate subject to a quality-of-service (QoS) per user or improving the fairness. In [3], a survey of common myths and misunderstands of NOMA is presented.

For the multi-user single-input single-output channel (MU-SISO), it is demonstrated in [4] that NOMA outperforms the achievable rate of orthogonal multiple access (OMA) techniques such as time division multiple access. At this point, it is necessary to remark that NOMA requires to order the users according to their channel gain and determine the power allocated to each user. In [5], [6], it is shown that the performance of NOMA improves as the difference of the channel gain between a pair of users increases. In this context, an algorithm referred to as next-largest-difference user pairing (NLUPA) is proposed to pair the user with worst channel gain with the user of highest channel gain, the second user of lowest channel gain with the second best user and so on. A pairing approach referred to as divide-and-NLUPA is proposed in [7] to guarantee a minimum user-rate in each cluster composed of a pair of users. Since the users are characterized by a single channel gain for the MU-SISO channel, determining the order and the power allocation coefficients results naturally.

Multiple-input multiple-output (MIMO) schemes provide additional Degrees of Freedom (DoF), which can improve the performance of NOMA [8]. Denoting the rate of user k_t, $k_t = \{1, \ldots, K_T\}$ as $R^{[k_t]}$ the sum-DoF is defined as

$$\mathrm{DoF} = \lim_{P \to \infty} \frac{R^{[1]} + \cdots + R^{[K]}}{\log(P)}, \tag{1}$$

which results equivalent to the number of simultaneous data streams transmitted free of interference, i.e., the multiplexing gain. However, NOMA is based on transmitting multiple data streams sharing the same spatial dimension, i.e., a DoF. Therefore, these data streams are subject to certain level of interference. As a consequence, both concepts DoF and number of data streams simultaneously transmitted are related in MIMO systems when applying NOMA. For MIMO systems, the implementation of NOMA is analyzed in [9] considering the effects of user pairing and power allocation. In [10], the capacity of MIMO-NOMA schemes is analyzed assuming a clustering strategy while performing NOMA in each cluster. It is interesting to remark that these works assume that the users are equipped with a number of receiving antennas equal or larger than the transmitting base station (BS). Then, considering downlink transmission, the interference can be canceled at the receiver side. If this condition is not satisfied, beamforming transmission is considered so that each beam corresponds to a

Máximo Morales-Céspedes and Ana García Armada are with the Department of Signal Theory and Communications. Universidad Carlos III de Madrid, Leganés, Spain (e-mail: {maximo, agarcia}@tsc.uc3m.es).

Octavia Dobre is with Faculty of Engineering and Applied Science, Memorial University, Canada (e-mail: odobre@mun.ca).

This work has been partially funded by the research projects TERESA (TEC2017-90093-C3-2-R), Chair of Excellence program at UC3M and Natural Sciences and Engineering Research Council of Canada (NSERC), through its Discovery program.

cluster in which several users perform NOMA [11]. Therefore, channel state information is required at both transmitter (CSIT) and receiver (CSIR) sides in most of MIMO scenarios, which involves a considerable use of the network resources [12]. In this context, in [13], the impact of imperfect CSI for MIMO-NOMA systems is analyzed.

For multi-user multiple-input single-output (MU-MISO) systems, the users are organized in clusters in which a single beamforming vector is assigned to all users of the cluster while managing the intra-cluster interference through NOMA [8], [14]. In [15], the beamforming vectors are determined considering only the user with best channel gain in each cluster. Therefore, all other users, which are not taken into consideration for the calculation of the beamforming, are subject to inter-cluster interference. This approach is proposed in [16] analyzing the use of zero forcing (ZF) and regularized zero forcing (RZF) beamforming [17]. Assuming normalized beamforming vectors, in [16], the authors propose an optimization problem to obtain the NOMA power allocation coefficients that maximize the sum-rate of each cluster. It should be noticed that determining the beamforming vectors for managing the intra and inter-cluster interference is not straightforward and typically leads to complex optimization problems. In [18], the authors propose the concept of quasi-degradation for minimizing the power consumption in MISO-NOMA systems. Interestingly, it is shown that MISO-NOMA achieves capacity when the channels of the users are quasi-degraded. In [19], the authors propose the rate fairness as a criterion to determine the beamforming vectors. Focussing on mmWaves, the power allocation coefficients that maximize the sum-rate subject to a minimal rate constraint per user are derived in [20]. It is shown that this optimization problem is non-convex. Specifically, the authors propose a solution based on decomposing the main problem into two sub-problems; maximizing the sum-rate given by the channel gain associated to each beam and determining the beams that maximize the modulus of the beamforming vectors. Considering a multi-user scenario, in [21], the optimization of the sum-rate assuming several clusters of distinct number of users each is considered for mmWaves. It is worthy to notice that CSIT knowledge is required in these schemes, while on the other hand, the inter-cluster interference is not completely avoided for all users.

Focussing on the MISO systems, an interesting scheme referred to as blind interference alignment (BIA) is proposed to achieve an increase of the DoF as the number of users and/or antennas increases without the need for CSIT [22]. Basically, BIA is based on exploiting a predefined pattern of channel correlations, referred to as *supersymbol*, so that the interference can be measured and subtracted afterwards. In contrast to NOMA, BIA completely removes the interference for all users. In [23], the use of reconfigurable antennas that provide a set of radiation patterns, referred to as preset modes, is proposed for the implementation of BIA. In this sense, the design and manufacturing of reconfigurable antennas is a current topic in the radio-frequency research [24], [25]. The performance of BIA in a cellular network considering clusters of cells is analyzed in [26], [27], [28]. These works show that BIA achieves a sum-rate similar or even higher than ZF beamforming when the costs of providing CSIT are taken into consideration. Moreover, there exist alternative BIA schemes for providing diversity or managing users mobility [29], [30], [31], [32].

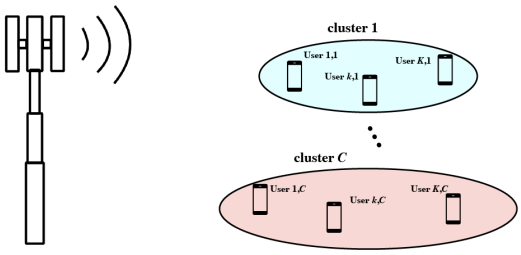
In this work, we propose the use of BIA for managing the inter-cluster interference while performing NOMA within each cluster for MU-MISO systems. In contrast to standard BIA, we do not assume a fully blind system but the large scale effects of the channel are known at the transmitter. It is worthy to remark that the large scale effects vary slowly, and therefore, the costs of obtaining their knowledge can be considered negligible. The main contributions of this work are:

1) A transmission scheme referred to as B-NOMA is devised for MU-MISO systems, which fully cancels the inter-cluster interference for users while performing NOMA in each cluster. Moreover, the proposed scheme only requires knowledge of the large scale effects at the transmitter for determining the order of users and the NOMA power allocation coefficients.
2) Considering that each cluster is composed of a pair of users, we propose a method based on random matrix theory that provides the NOMA power coefficients that aim to maximize the sum-rate or the fairness for the proposed B-NOMA scheme. Besides, we obtain the sub-optimal closed-form expressions of these coefficients considering some approximations.
3) A cognitive approach based on NOMA power allocation is derived for the B-NOMA scheme. The BS first tries to satisfy a quality-of-service (QoS) for a primary user, while the remaining power is assigned to the other user of the pair in a cognitive fashion.

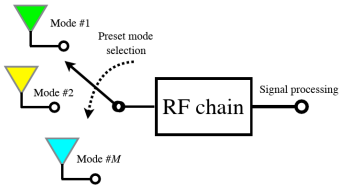
Simulation results show that the proposed B-NOMA scheme outperforms the sum-rate of other schemes such as MISO-OMA and MISO-NOMA based on RZF [16]. Furthermore, in comparison to BIA, the B-NOMA scheme has the advantage of reducing the length of the supersymbol, i.e., it relaxes the channel coherence requirements as well as the noise enhancement due to the subtraction of the interference terms.

The remainder of the paper is organized as follows. The system model is described in Section II. The benchmarking schemes for the proposed B-NOMA, taking into consideration the costs of providing CSI, are presented in Section III. The proposed B-NOMA scheme is described in Section IV. Considering a user pairing strategy, the methods for determining the NOMA power coefficients according to a predefined strategy are presented in Section V. Simulation results are presented in Section VI, and finally, concluding remarks are provided in Section VII.

Notation. The following notation is considered in this work. Bold upper case and lower case letters denote matrices and vectors, respectively, $\mathbf{I}_M$ and $\mathbf{0}_M$ denote the $M \times M$ identity and zero matrices, respectively, while $\mathbf{0}_{M,N}$ corresponds to the $M \times N$ zero matrix, $[\,]^T$ and $[\,]^H$ are the transpose and the hermitic transpose operators, respectively, $\mathbb{E}$ is the statistical expectation, $\mathrm{col}\{\}$ is the column operator that stacks the considered vectors in a column and $[x]^+ = \max(0, x)$.



(a) Considered NOMA scenario. The BS is equipped with M antenna. Transmission to C clusters composed of K users each.



(b) Reconfigurable antenna. It can switch among M preset modes connected to a single signal processing chain. Each mode can be considered as an independent antenna that provides a linearly independent radiation pattern.

Fig. 1. Considered scenario and concept of reconfigurable antenna.

II. System model

We consider a MU-MISO downlink transmission scenario where the BS is equipped with M, $m \in \{1, \dots M\}$, antennas that transmit to K_T, $k_t = \{1, \dots, K_T\}$, users. The set of users is organized in C, $c \in \{1, \dots, C\}$, clusters each composed of $K = \frac{K_T}{C} \in \mathbb{N}^+$, $k \in \{1, \dots, K\}$, users[1] equipped with a reconfigurable antenna each, as shown in Fig. 1(a). Basically, each reconfigurable antenna provides ν, $\nu = \{1, \dots, M\}$, distinct and linearly independent channel responses, referred to as preset modes, connected to a single signal processing chain as is shown in Fig. 1(b). The transmitted signal at time t can be written in a vector form as $\mathbf{x}[t] = \begin{bmatrix} x_1 & \dots & x_M \end{bmatrix}^T \in \mathbb{C}^{M\times 1}$. Thus, the signal received by user k in cluster c at symbol extension[2] t is given by

$$y^{[k,c]}[t] = \mathbf{h}^{[k,c]}\left(\nu^{[k,c]}[t]\right)^T \mathbf{x}[t] + z^{[k,c]}[t], \tag{2}$$

where $\mathbf{h}^{[k,c]}\left(\nu^{[k,c]}[t]\right) \in \mathbb{C}^{M\times 1}$ is the channel vector between the M antennas of the BS and user k of cluster c at preset mode $\nu^{[k,c]}$, and $z^{[k]}[t]$ is complex-valued additive white Gaussian noise with zero mean and variance σ_z^2. The channel response corresponding to the preset mode $\nu^{[k,c]}$ can be written as $\mathbf{h}^{[k,c]}\left(\nu^{[k,c]}[t]\right) = \tilde{\mathbf{h}}^{[k,c]}\left(\nu^{[k,c]}[t]\right)\sqrt{g^{[k,c]}}$, where each entry of $\tilde{\mathbf{h}}^{[k,j]}\left(\nu^{[k,c]}[t]\right)$ follows an independent Gaussian distribution $\mathcal{N}(0,1)$ and $g^{[k,c]}$ represents the large scale channel effects for user k in cluster c.

We assume that CSIT knowledge is limited to the large scale effects, i.e., $g^{[k,c]}$, and there neither exists coordination nor data sharing among the antennas of the BS. Moreover, the switching pattern for selecting the preset mode $\nu^{[k,c]}[t]$ is known beforehand. The transmitted signals are subject to an average power constraint $\mathbb{E}\left[\|\mathbf{x}[t]\|^2\right] \leq P$.

[1]If $\frac{K_T}{C} \in \mathbb{Z}^+$ the first $c = \{1, \dots, \mathrm{mod}(K_T, C)\}$ clusters contain $\lceil \frac{K_T}{C} \rceil$ users while the last $c = \{\mathrm{mod}(K_T, C) + 1, \dots, C\}$ clusters are composed of $\lfloor \frac{K_T}{C} \rfloor$ users.

[2]We focus on the temporal dimension without loss of generality and each symbol extension corresponds to a time slot. Nevertheless, all the results can be easily applied to the frequency domain.

III. OMA, NOMA and BIA schemes for the MISO BC

In this section, we describe two benchmark methodologies based on instantaneous and accurate CSIT. In these cases, we consider conventional antennas so that $\mathbf{h}^{[k,c]}(\nu) = \mathbf{h}^{[k,c]}$. For illustration purpose, we assume $C = M$. After that, we describe the conventional BIA scheme while introducing some useful notation, and consider it also as a benchmarking scheme. Furthermore, we describe the costs of providing CSIT and CSIR for the considered schemes.

A. OMA for the MISO BC

Assuming a system with MK users, the beamforming schemes typically manage the interference through an OMA strategy comprising K time slots. Thus, M data streams can be transmitted simultaneously during each time slot. Without loss of generality, we consider that transmission to the users with the same index k in the C clusters occurs during the time slot k. The symbol intended to user k in cluster c is denoted as $s^{[k,c]}$, $\mathbb{E}\left[s^{[k,c]}\right] = 1$. Under CSIT knowledge, the symbol $s^{[k,c]}$ is transmitted employing a beamforming vector $\mathbf{w}^{[k,c]} \in \mathbb{C}^{M\times 1}$. We assume normalized beamforming vectors, i.e., $\|\mathbf{w}^{[k,c]}\| = 1$. Thus, the signal transmitted during time slot k is

$$\mathbf{x}[k] = \sum_{c=1}^{C} \mathbf{w}^{[k,c]} s^{[k,c]} \sqrt{P_{k,c}}, \tag{3}$$

where $P_{k,c}$ is the power allocated to the symbol $s^{[k,c]}$. Moreover, we define $\bar{s}^{[k,c]} = s^{[k,c]}\sqrt{P_{k,c}}$. Thus, the signal received by user k in cluster c can be written as

$$y^{[k,c]} = \mathbf{h}^{[k,c]^T}\mathbf{w}^{[k,c]}\bar{s}^{[k,c]} + \sum_{\substack{c'=1 \\ c'\neq c}}^{C} \mathbf{h}^{[k,c]^T}\mathbf{w}^{[k,c']}\bar{s}^{[k,c']} + z^{[k,c]}. \tag{4}$$

Therefore, the inter-cluster interference among users can be minimized, or even canceled, by selecting properly the beamforming vectors. Moreover, recall that the intra-cluster interference is avoided through an OMA strategy using K time slots. In this work, we employ ZF and RZF beamforming assuming perfect CSIT [16]. Under the proposed MISO-OMA strategy, the achievable rate of user k in cluster c is given by

$$R_{\mathrm{mo}}^{[k,c]} = \frac{1}{K}\eta_{\mathrm{mo}} \log_2\left(1 + \frac{P_{k,c}|\mathbf{h}^{[k,c]^T}\mathbf{w}^{[k,c]}|^2}{\sum_{c'\neq c}^{C} P_{k,c}|\mathbf{h}^{[k,c]^T}\mathbf{w}^{[k,c']}|^2 + \sigma_z^2}\right), \tag{5}$$

where η_{mo} denotes the costs of providing CSIT and CSIR for MISO-OMA. Notice that $N_{\mathrm{MISO-OMA}} = M$ data streams are transmitted simultaneously for the MISO-OMA approach in each time slot.

Although several beamforming schemes are proposed in the literature, we focus on linear beamforming techniques such as

ZF or RZF. Focussing on RZF, the beamforming vector $\mathbf{w}^{[k,c]}$ is given by

$$\mathbf{w}^{[k,c]} = \frac{\mathbf{H}_k^H \mathbf{r}_k^{\mathrm{RZF}}}{\|\mathbf{H}_k^H \mathbf{r}_k^{\mathrm{RZF}}\|} \in \mathbb{C}^{M\times 1}, \tag{6}$$

where $\mathbf{H}_k = \begin{bmatrix}\mathbf{h}^{[k,1]} & \dots & \mathbf{h}^{[k,C]}\end{bmatrix} \in \mathbb{C}^{M\times M}$ and $\mathbf{r}_k^{\mathrm{RZF}}$ is the k-th column of $\left(\mathbf{H}_k\mathbf{H}_k^H + \delta\mathbf{I}_M\right)^{-1}$, where δ is a non-negative constant value that ensures the existence of the inverse matrix. This beamforming scheme is equivalent to ZF if $\delta = 0$. It is worth to remark that for ZF these data streams are not subject to interference, i.e., the denominator in (5) only contains the contribution of the noise.

B. NOMA for the MISO BC

For the MISO-NOMA approach, the inter-cluster interference is managed through beamforming while applying NOMA within each cluster. Similarly to [16], we consider that the normalized beamforming vector $\mathbf{w}^{[c]} \in \mathbb{C}^{M\times 1}$ corresponding to cluster c is calculated considering exclusively the user with the best channel gain within the cluster. Notice that this approach minimizes the inter-cluster interference among the strongest users of each cluster, which perform SIC. In this case, the transmitted signal is given by

$$\mathbf{x} = \sum_{c=1}^{C} \mathbf{w}^{[c]} \sum_{k=1}^{K} s^{[k,c]} \sqrt{P_c \alpha^{[k,c]}}, \tag{7}$$

where P_c is the power allocated to cluster c and $\alpha^{[k,c]}$ is the NOMA power coefficient of user k in cluster c. Moreover, we define the symbol transmitted to each user after considering the power allocation as $\bar{s}^{[k,c]} = s^{[k,c]}\sqrt{P_c\alpha^{[k,c]}}$. Thus, the signal received by user k in cluster c can be written as

$$\begin{aligned} y^{[k,c]} =& \mathbf{h}^{[k,c]^T}\mathbf{w}^{[c]}\bar{s}^{[k,c]} + \underbrace{\sum_{k'=1,k'\neq k}^{K} \mathbf{h}^{[k,c]^T}\mathbf{w}^{[c]}\bar{s}^{[k',c]}}_{\text{Intra-cluster interference}} \\ &+ \underbrace{\sum_{c'=1,c'\neq c}^{C} \mathbf{h}^{[k,c]^T}\mathbf{w}^{[c']}\sum_{k=1}^{K}\bar{s}^{[k,c']}}_{\text{Inter-cluster interference}} + z^{[k,c]}. \end{aligned} \tag{8}$$

Following the lines of [16], we assume that the resulting channel gains are ordered as:

$$|\mathbf{h}^{[1,c]^T}\mathbf{w}^{[c]}|^2 \leq \dots \leq |\mathbf{h}^{[k,c]^T}\mathbf{w}^{[c]}|^2 \leq \dots \leq |\mathbf{h}^{[K,c]^T}\mathbf{w}^{[c]}|^2 \tag{9}$$

After that, the NOMA power coefficients are determined according to a predefined strategy such as maximizing the sum-rate subject to a QoS per user or improving the fairness, e.g., [16], [19], [20]. For the considered MISO-NOMA, the user k performs SIC in order to remove the interference because of transmission to users $k^\star < k$, $k^\star = \{1,\dots,k-1\}$, while treating as noise the interference given by transmission to the users $k' > k$, $k' = \{k+1,\dots,K\}$. Assuming that the user k of the cluster c decodes successfully the symbols intended to the users $k^\star = \{1,\dots,k-1\}$ and cancels their interference, the achievable rate is given by

$$\begin{aligned} R_{\mathrm{mn}}^{[k,c]} &= \eta_{\mathrm{mn}} \\ &\times \log_2\left(1 + \frac{P_c\alpha^{[k,c]}\,|\mathbf{h}^{[k,c]^T}\mathbf{w}^{[c]}|^2}{\sum_{k'=k}^{K} P_c\alpha^{[k',c]}\,|\mathbf{h}^{[k,c]^T}\mathbf{w}^{[c]}|^2 + I^{[k,c]} + \sigma_z^2}\right), \end{aligned} \tag{10}$$

where η_{mn} denotes the costs of providing CSIT and CSIR for MISO-NOMA and $I^{[k,c]} = \sum_{c'=1,c'\neq c}^{C} P_{c'}|\mathbf{h}^{[k,c]^T}\mathbf{w}^{[c']}|^2$ is the inter-cluster interference that remains after applying the beamforming scheme. It is worth noting that user $k=1$ of each cluster treats the interference because of transmission to all other $K-1$ users of the same cluster as noise, whereas user K removes all the intra-cluster interference through SIC before decoding the intended symbol. The MISO-NOMA scheme transmits $N_{\mathrm{MISO-NOMA}} = MK$ data streams simultaneously per time/frequency resource. However, in contrast to MISO-OMA these data streams are subject to a certain level of intra-cluster interference given by NOMA. Besides, inter-cluster interference can appear depending on the calculation of the beamforming vectors.

C. BIA for the MISO BC

The BIA schemes consider transmission during several symbol extensions, i.e., channel uses, in which the reconfigurable antenna of each user switches among a set of preset modes that provide distinct channel responses each. The set of symbol extensions that satisfy the BIA criterion is referred to as *supersymbol* from now on. Moreover, in this work we consider the temporal dimension, and therefore, each symbol extension corresponds to a time slot.

Let us consider first a toy example where a BS equipped with $M = 2$ antennas transmits to $K_t = 2$ users. Following [23], the reconfigurable antenna of each user follows a pattern for selecting the preset mode during 3 time slots as is shown in Fig. 2. Thus, the superysmbol comprises 3 time slots and it is composed of these two patterns. The pattern of preset modes that each user follows during the supersymbol is denoted as $\zeta^{[k]}[t]$. Moreover, the cluster index is removed for ease of notation. For the proposed supersymbol, the transmitted signal is

	1	2	3
User 1	$\mathbf{h}^{[1]}(1)$	$\mathbf{h}^{[1]}(2)$	$\mathbf{h}^{[1]}(1)$
User 2	$\mathbf{h}^{[2]}(1)$	$\mathbf{h}^{[2]}(1)$	$\mathbf{h}^{[2]}(2)$

Fig. 2. Supersymbol of the BIA scheme for $M = 2$ and $K = 2$ users. The channel of users 1 and 2 varies according to the pattern of preset modes $\zeta^{[1]}[t] = \{1,2,1\}$ and $\zeta^{[2]}[t] = \{1,1,2\}$, respectively. Each color represents a preset mode.

$$\mathbf{x}_{\mathrm{BIA}} = \begin{bmatrix}\mathbf{x}[1]\\ \mathbf{x}[2]\\ \mathbf{x}[3]\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{I}_2\\ \mathbf{I}_2\\ \mathbf{0}_2\end{bmatrix}}_{\mathbf{W}^{[1]}}\mathbf{s}^{[1]} + \underbrace{\begin{bmatrix}\mathbf{I}_2\\ \mathbf{0}_2\\ \mathbf{I}_2\end{bmatrix}}_{\mathbf{W}^{[2]}}\mathbf{s}^{[2]}, \tag{11}$$

where $\mathbf{s}^{[k]} = [s_1^{[k]}, s_2^{[k]}]^T$ is the symbol intended to user k, which carries 2 DoF, i.e., 2 data streams, and $s_m^{[k]}$ is the symbol

transmitted from antenna m of the BS. In (11), $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$ are the precoding matrices for users 1 and 2, respectively. It is worth noting that these matrices are composed of 0 and 1 values and CSIT is not required to determine their structure. Thus, for user 1, the signal received during the three time slots of the supersymbol is

$$\begin{bmatrix} y^{[1]}[1] \\ y^{[1]}[2] \\ y^{[1]}[3] \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{h}^{[1]}(1)^T \\ \mathbf{h}^{[1]}(2)^T \\ \mathbf{0}_{2,1}^T \end{bmatrix}}_{\text{rank=2}} \mathbf{s}^{[1]} + \underbrace{\begin{bmatrix} \mathbf{h}^{[1]}(1)^T \\ \mathbf{0}_{2,1}^T \\ \mathbf{h}^{[1]}(1)^T \end{bmatrix}}_{\text{rank=1}} \mathbf{s}^{[2]} + \begin{bmatrix} z^{[1]}[1] \\ z^{[1]}[2] \\ z^{[1]}[3] \end{bmatrix}. \tag{12}$$

Notice that considering the pattern of preset modes for user 1, $\zeta^{[1]}[t]$, the symbol $\mathbf{s}^{[1]}$ precoded by the matrix $\mathbf{W}^{[1]}$ is contained in a rank-2 matrix. Further, the interference due to the transmission of $\mathbf{s}^{[2]}$, which is precoded according to $\mathbf{W}^{[2]}$ is aligned in a rank-1 matrix. As a consequence, user 1 can cancel the interference by measuring it at the third time slot. After subtracting the interference, the received signal is

$$\begin{bmatrix} y^{[1]}[1] - y^{[1]}[3] \\ y^{[1]}[2] \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{h}^{[1]}(1)^T \\ \mathbf{h}^{[1]}(2)^T \end{bmatrix}}_{\mathbf{H}^{[1]}} \mathbf{s}^{[1]} + \underbrace{\begin{bmatrix} z^{[1]}[1] - z^{[1]}[3] \\ z^{[1]}[2] \end{bmatrix}}_{\text{Noise enhancement}}. \tag{13}$$

Thus, since $\mathbf{H}^{[1]}$ is a full rank matrix, user 1 achieves 2 DoF by solving the equation system (13). Similarly, user 2 can obtain 2 DoF. Therefore, 4 DoF are achievable during 3 time slots for the considered setting. That is, the BIA scheme achieves $\frac{4}{3}$ DoF per time slot. In turn, note that orthogonal resource allocation obtains 1 DoF per time slot.

For the general case, BIA is based on creating a supersymbol in which the channel state of user k varies among M preset modes, while all other users maintain a constant channel mode, i.e., do not vary their preset mode. The set of symbol extensions, i.e., time slots in the temporal dimension, that satisfies this criterion is referred to as an *alignment block* of user k. Thus, the BIA scheme is given by the functions that determine the pattern of preset modes $\zeta^{[k]}[t]$ and the precoding matrices $\mathbf{W}^{[k]}$ [23].

The supersymbol and the transmission structure are obtained creating alignment blocks. Notice that distinct symbols with M DoF each are transmitted in each alignment block. The first $M-1$ time slots of each alignment block belong to Block 1 of the supersymbol, where simultaneous transmission occurs. On the other hand, the last time slot is assigned to Block 2, in which each symbol is transmitted in orthogonal fashion. For instance, considering the supersymbol of the previous example (see Fig. (2)), the time slots $\{1,2\}$ and $\{1,3\}$ form an alignment block for users 1 and 2, respectively. The interaction between an alignment block of user k and user $k+1$ is as shown in Fig. 3. The interference because of transmission to user k can be measured by user $k+1$ and subtracted from the time slots polluted by this interference. It can be easily checked that the same procedure can be applied to any alignment block of user $k+1$.

Following this methodology, Block 1 and Block 2 comprise $\Lambda_{\mathrm{B1}} = (M-1)^{K_T}$ and $\Lambda_{\mathrm{B2}} = K_T(M-1)^{K_T-1}$ time slots, respectively. Therefore, the supersymbol comprises

$$\Lambda(M, K_T) = (M-1)^{K_T} + K_T(M-1)^{K_T-1}, \tag{14}$$

time slots. The BIA scheme provides $(M-1)^{K_T-1}$, $\ell \in \{1, \ldots, (M-1)^{K_T-1}\}$, alignment blocks per user, in which a distinct symbol carrying M DoF is transmitted. Therefore, the achievable sum-DoF per time slot is

$$\mathrm{DoF}_{\mathrm{BIA}} = N_{\mathrm{BIA}} = \frac{MK_T}{M+K_T-1}. \tag{15}$$

In [33], it is demonstrated that (15) corresponds to the optimal sum-DoF without CSIT. Notice that this sum-DoF also corresponds to the number of data streams that are simultaneously transmitted per time slot, which is denoted as N_{BIA}. In terms of resource management, BIA orthogonalizes the transmission to K_T users without the need for CSIT while providing multiplexing gain. Specifically, each user decodes M DoF per alignment block where the ratio between the number of alignment blocks and the supersymbol length is $\frac{(M-1)^{K_T-1}}{(M-1)^{K_T}+K_T(M-1)^{K_T-1}} = \frac{1}{M+K_T-1}$. Notice that this ratio corresponds to the pre-log factor of a MIMO channel given by the channel matrix defined as

$$\tilde{\mathbf{H}}^{[k_t]} = \left[\tilde{\mathbf{h}}^{[k_t]}(1)^T, \quad \ldots, \quad \tilde{\mathbf{h}}^{[k_t]}(M)^T\right]^T \in \mathbb{C}^{M\times M}. \tag{16}$$

Focussing on the achievable rate BIA is based on subtracting the interference, which has been measured previously in the proper time slots, generating a noise enhancement (see (13)). Therefore, a SNR degradation occurs as the number of users increases. Assuming constant power allocation during the entire supersymbol [26], the achievable user rate of user k_t is

$$\begin{aligned} R_{\mathrm{BIA}}^{[k_t]} =& \frac{\eta_{\mathrm{bia}}}{M+K_T-1} \\ & \times \mathbb{E}\left[\log\det\left(\mathbf{I} + \frac{\rho^{[k_t]}}{M}\tilde{\mathbf{H}}^{[k_t]}\tilde{\mathbf{H}}^{[k_t]^H}\mathbf{R}_z^{-1}\right)\right], \end{aligned} \tag{17}$$

where η_{bia} represents the costs of providing CSIR for BIA, $\rho^{[k_t]} = \frac{Pg^{[k_t]}}{M\sigma_z^2}$ is the SNR associated to user k_t, and

$$\mathbf{R}_z = \begin{bmatrix} (2K_T-1)\mathbf{I}_{M-1} & 0 \\ 0 & 1 \end{bmatrix}, \tag{18}$$

is the resulting covariance matrix of the noise after interference subtraction for constant power allocation. Notice that the noise enhancement is proportional to the total number of users.

Remark 1. *The implementation of BIA is subject to transmission within a coherence time block larger than the supersymbol length* (14)*. There exist several alternative BIA schemes for managing situations with short coherence time efficiently such as [28], [31], [32].*

D. The costs of providing CSIT and CSIR

Assuming frequency division duplex in a MU-MISO system, providing CSIT requires first to transmit downlink estimation pilots (EP) in orthogonal fashion so that the contribution to the CSIR of each antenna of the BS can be estimated [12]. After that, each user quantizes the estimated CSIR and feeds

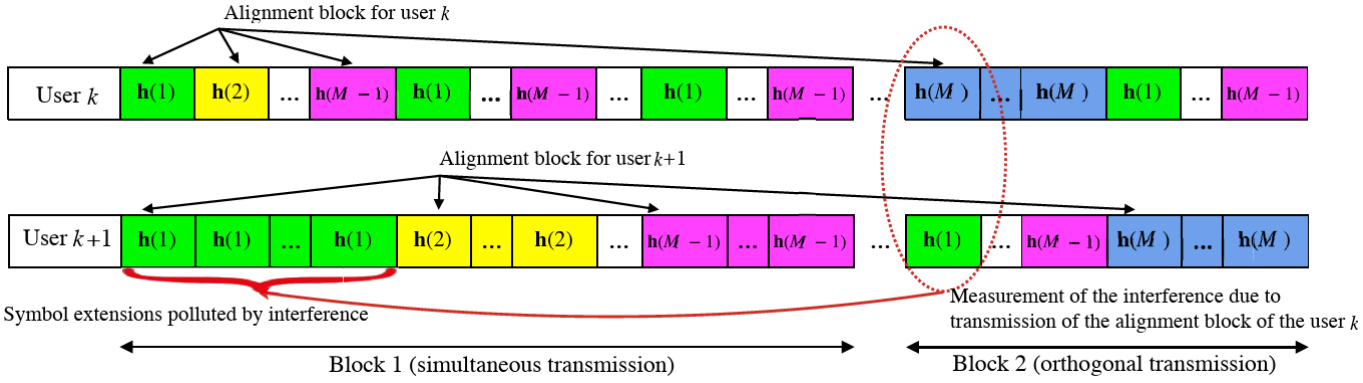


Fig. 3. Interaction between alignment blocks of different users for BIA. Each color represents a preset mode.

TABLE I
COSTS OF PROVIDING CSIR AND CSIT

Type	MISO-OMA	MISO-NOMA	BIA
EP	$M\theta_{\text{csi}}$	$M\theta_{\text{csi}}$	0
FB	$M\theta_{\text{fb}}$	$C\theta_{\text{fb}}$	0
CD	$KC\theta_{\text{cd}}$	$KC\theta_{\text{cd}}$	$M\theta_{\text{cd}}$

back the channel through the uplink. Once the beamforming vectors are obtained, additional pilots are required for coherent detection (CD) taking into consideration the transmitted signal (see (3) or (7)).

The fraction of downlink transmission resources for pilot estimation and CD is denoted as θ_{ep} and θ_{cd}, respectively. Similarly, the fraction of transmission resources that are assigned to the uplink for feedback (FB) is denoted as θ_{fb}. The feedback overhead is proportional to the number of users that send the estimated CSIR to the BS. Recall that for the proposed MISO-NOMA approach the beamforming only considers the user with highest channel gain of each cluster, and therefore, only these users feed back the estimated channel. Considering users equipped with a single conventional antenna, the overhead due to CD is proportional to the number of users, i.e., $K\theta_{\text{cd}}$. On the other hand, reconfigurable antennas allow us to implement blind transmission techniques in the absence of CSIT, i.e., $\theta_{\text{csi}} = \theta_{\text{fb}} = 0$. However, the costs of CD are proportional to the number of preset modes, i.e., $M\theta_{\text{cd}}$. Table I summarizes the costs of providing CSIR and CSIT [12]. For instance, the efficiency of the MISO-NOMA is $\eta_{\text{mn}} = 1-(M\theta_{\text{csi}}+C\theta_{\text{fb}}+KC\theta_{\text{cd}})$. Furthermore, the costs of providing knowledge of the channel large scale effects are considered negligible.

IV. B-NOMA BASED ON RECONFIGURABLE ANTENNAS

We propose a transmission scheme that combines the BIA and NOMA schemes, which is henceforth referred to as B-NOMA. The key idea is to employ BIA to align the inter-cluster interference, while applying NOMA within each cluster. For the sake of an easy explanation, we first focus on a specific case and after that we describe the proposed B-NOMA scheme for the general case.

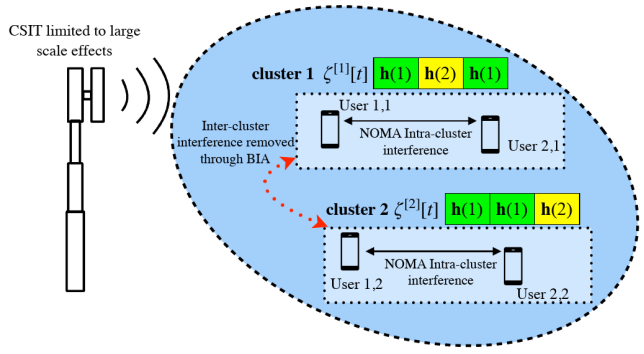


Fig. 4. B-NOMA transmission for the considered toy example. $M = C = 2$, $K = 2$.

A. Toy example

Let us consider a scenario where a BS equipped with $M = 2$ antennas transmits to $C = 2$ clusters composed of $K = 2$ users each. The users of each cluster c, $c = \{1, 2\}$, are paired so that $g^{[1,c]} \leq g^{[2,c]}$. The inter-cluster interference is managed through BIA. Thus, the users of the same cluster reuse the same switching pattern of preset modes is as shown in Fig. 4. That is, both users of cluster $c = 1$ follow the switching pattern $\zeta^{[1]}[t]$ corresponding to user 1 of the BIA scheme for $M = 2$ and $K_t = 2$, while the users of cluster 2 employ the same switching pattern as for user 2 in the BIA scheme, i.e., $\zeta^{[2]}[t]$ (see Fig. 2). Moreover, NOMA transmission is performed within each cluster. Therefore, the transmitted signal is given by

$$\mathbf{x}_{\text{B-NOMA}} = \begin{bmatrix}\mathbf{x}[1]\\ \mathbf{x}[2]\\ \mathbf{x}[3]\end{bmatrix} = \mathbf{W}^{[1]}\left(\alpha^{[1,1]}\mathbf{s}^{[1,1]} + \alpha^{[2,1]}\mathbf{s}^{[2,1]}\right)\sqrt{P_c} + \mathbf{W}^{[2]}\left(\alpha^{[1,2]}\mathbf{s}^{[1,2]} + \alpha^{[2,2]}\mathbf{s}^{[2,2]}\right)\sqrt{P_c}, \tag{19}$$

where $\mathbf{s}^{[k,c]} \in \mathbb{C}^{2\times 1}$ is the symbol intended to user k of cluster c and $\alpha^{[k,c]}$ is the NOMA power coefficient of the corresponding symbol. Moreover, in (19), the precoding matrices $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$ are given by the considered BIA scheme (see (11)).

Focussing on the first cluster without loss of generality, the

signal received by the first user can be written as

$$\mathbf{y}^{[1,1]} = \begin{bmatrix} \mathbf{h}^{[1,1]}(1)^T \\ \mathbf{h}^{[1,1]}(2)^T \\ \mathbf{0}_{2,1}^T \end{bmatrix} \left(\mathbf{s}^{[1,1]}\sqrt{\alpha^{[1,1]}} + \mathbf{s}^{[2,1]}\sqrt{\alpha^{[2,1]}} \right) \sqrt{P_1} + \begin{bmatrix} \mathbf{h}^{[1,1]}(1)^T \\ \mathbf{0}_{2,1}^T \\ \mathbf{h}^{[1,1]}(1)^T \end{bmatrix} \underbrace{\left(\mathbf{s}^{[1,2]}\sqrt{\alpha^{[1,2]}} + \mathbf{s}^{[2,2]}\sqrt{\alpha^{[2,2]}} \right) \sqrt{P_2}}_{\text{Inter-cluster interference}}, \tag{20}$$

where $\mathbf{y}^{[1,1]} = \text{col}\{y^{[1,1]}[t]\}_{t=1}^{3}$.

Similarly to (12), it can be seen that the inter-cluser interference is aligned. Therefore, the interference because of transmission to both users of cluster $c = 2$ can be measured at the third time slot and subtracted afterwards. The signal received by user 1 in cluster 1 after the subtraction of the inter-cluster interference, which is denoted as $\tilde{\mathbf{y}}^{[1,1]}$, is

$$\tilde{\mathbf{y}}^{[1,1]} = \begin{bmatrix} \mathbf{h}^{[1,1]}(1)^T \\ \mathbf{h}^{[1,1]}(2)^T \end{bmatrix} \left(\mathbf{s}^{[1,1]}\sqrt{\alpha^{[1,1]}} + \mathbf{s}^{[2,1]}\sqrt{\alpha^{[2,1]}} \right) \sqrt{P_1} + \begin{bmatrix} z^{[1,1]}[1] - z^{[1,1]}[3] \\ z^{[1,1]}[2] \end{bmatrix}. \tag{21}$$

Thus, the sum of the symbols $\left(\alpha^{[1,1]}\mathbf{s}^{[1,1]} + \alpha^{[2,1]}\mathbf{s}^{[2,1]}\right)$ is decodable by solving (21). After that, NOMA is applied to decode the intended symbol $\mathbf{s}^{[1,1]}$ treating the interference caused by the transmission of $\mathbf{s}^{[2,1]}$ as noise. The same procedure can be carried out for the second user, which first performs SIC to remove the interference caused by the transmission of $\mathbf{s}^{[1,1]}$ and, after that, decodes the intended symbol $\mathbf{s}^{[2,1]}$. It can be easily checked that this methodology can be also applied to the second cluster.

Note that the proposed B-NOMA scheme reduces the noise enhancement in comparison to BIA for transmitting the whole set of users, i.e., $K_T = 4$, while cancelling the inter-cluster interference. Moreover, the length of the supersymbol depends on the number of clusters instead of the number of users, and therefore, the channel coherence requirement is relaxed. Furthermore, the combination of BIA and NOMA increases the number of data streams simultaneously transmitted by sharing the DoF, i.e., the spatial dimensions, provided by BIA through NOMA. In this particular case, the number of data streams transmitted simultaneously is multiplied by two in comparison with BIA. Therefore, $\frac{8}{3}$ data streams are transmitted per time slot, which are subject to certain level of intra-cell interference because of the use of NOMA. Besides, in comparison with MISO-NOMA, the inter-cluster interference is completely canceled.

B. General case

For the general case, each cluster c, $c \in \{1, \ldots, C\}$, contains K, $k \in \{1, \ldots, K\}$ users. As described above for a specific case, the inter-cluster interference can be canceled by applying BIA. Thus, B-NOMA considers a supersymbol structure for a BS equipped with M antennas and C receivers. Each specific switching pattern of preset modes $\zeta^{[c]}[t]$ is reused by the K users of the same cluster c as shown in Fig. 4. Once the inter-cluster interference is managed using BIA, the multi-user transmission within each cluster performs NOMA. Note that the same BIA precoding matrix $\mathbf{W}^{[c]}$ is considered for the users of the same cluster. Hence, the transmitted signal for B-NOMA is

$$\mathbf{X} = \sum_{c=1}^{C} \sqrt{P_c}\mathbf{W}^{[c]} \underbrace{\sum_{k=1}^{K} \sqrt{\alpha^{[k,c]}}\mathbf{s}_T^{[k,c]}}_{\mathbf{s}_{T,\text{NOMA}}^{[c]}}, \tag{22}$$

where $\mathbf{X} = \text{col}\{\mathbf{x}[t]\}_{t=1}^{\Lambda(M,C)}$, $\mathbf{s}_{T,\text{NOMA}}^{[c]}$ is the vector that contains the symbols intended to the K users of cluster c superposed in the power domain, $\mathbf{s}_T^{[k,j]} = \text{col}\left\{\mathbf{s}_\ell^{[k,c]}\right\}_{\ell=1}^{(M-1)^{K-1}}$ is the vector that contains the symbols allocated to each alignment block of the BIA scheme and $\mathbf{s}_\ell^{[k,c]} \in \mathbb{C}^{M\times 1}$ is the symbol of the ℓ-th alignment block intended to user k in cluster c. Notice that the extension to an arbitrary number of users per cluster results straightforward by simply considering $\mathbf{s}_{T,\text{NOMA}}^{[c]} = \sum_{k=1}^{K_c} \sqrt{\alpha^{[k,c]}}\mathbf{s}_T^{[k,c]}$, where K_c denotes the number of users in the cluster c.

Notice that determining the precoding matrices does not require any CSIT since they are based on the structure of the supersymbol. Basically, the precoding matrices manage the transmission of each alignment block at the corresponding time slots. Thus, each alignment block corresponds to a column in the precoding matrix which is obtained by stacking $M \times M$ identity matrices in the rows corresponding to the time slots where that alignment block is transmitted. For instance, in the toy example considered in section III-C, it can be seen that users 1 and 2 obtains a single alignment block at time slots $\{1, 2\}$ and $\{1, 3\}$, respectively. Thus, the precoding matrices, which are shown in (11), are composed by a single column with a 2×2 identity matrix in the rows corresponding to the aforementioned time slots. In this sense, the procedure to determine the precoding matrices and their general expression are described in detail in [23]. For B-NOMA the same precoding matrix $\mathbf{W}^{[c]}$ is employed for all the user belonging to the same cluster. The proposed B-NOMA applies BIA for managing the inter-cluster interference assuming that the reconfigurable antennas of all the users belonging to the cluster c follow the same switching pattern $\zeta^{[c]}[t]$. Therefore, the supersymbol for B-NOMA comprises $\Lambda(M, C)$ time slots (see (14)). It is worth to remark that B-NOMA reduces the length of the supersymbol in comparison with BIA since multiple users reuse the same switching pattern. Furthermore, the implementation of BIA techniques for managing the supersymbol length referred in Remark 1 can be easily applied for the proposed B-NOMA scheme.

According to the B-NOMA scheme, the proposed supersymbol allocates $(M-1)^{C-1}$ alignment blocks carrying M DoF to each of the C clusters during $\Lambda(M, C)$ time slots. Therefore, each cluster achieves $\frac{M}{M+C-1}$ DoF per time slot, which can be interpreted as the spatial dimensions that are shared through NOMA by the K users of each cluster. Therefore, considering C clusters containing K users each, the number of data

streams that are simultaneously transmitted is

$$N_{\mathrm{B-NOMA}} = \frac{CM}{C+M-1} \times K. \tag{23}$$

Notice that for MISO-NOMA the users of the same cluster share a single-dimension, i.e. a DoF, given by their resulting channel gain, which is determined by the precoding matrices, i.e., $|\mathbf{h}^{[k,c]^T}\mathbf{w}^{[c]}|^2$ in (10). Besides, the user can be subject to inter-cluster cluster for MISO-NOMA as can be seen in (10). In contrast, the proposed B-NOMA cancels the inter-cluster interference for all users, while sharing M dimensions during the M time slots that compose each alignment block. Taking into consideration the ratio of alignment blocks per supersymbol for each user denoted as $\kappa = \frac{1}{K+C-1}$, the users of each cluster share $\frac{M}{M+C-1}$ dimensions per time slot.

Then, transmission to the users within each cluster occurs during an alignment block. In this sense, each alignment block follows the same structure as shown in Fig. 3. That is, the first $M-1$ slots of the alignment block, which belong to Block 1, are polluted by interference due to transmission to all other clusters, while the last slot does not contain any term of interference since it belongs to Block 2. By applying the BIA scheme, the inter-cluster interference is measured and subtracted afterwards. Thus, the received signal at user k of cluster c after removing the inter-cluster interference can be written as

$$\begin{bmatrix} y^{[k,c]}[1] \\ \vdots \\ y^{[k,c]}[M-1] \\ y^{[k,c]}[M] \end{bmatrix} = \begin{bmatrix} \mathbf{h}^{[k,c]}(1)^T \\ \vdots \\ \mathbf{h}^{[k,c]}(M-1)^T \\ \mathbf{h}^{[k,c]}(M)^T \end{bmatrix} \sum_{k=1}^{K} \mathbf{s}_{\ell}^{[k,c]} \sqrt{P_c \alpha^{[k,c]}} + \begin{bmatrix} z^{[k,c]}[1] - \sum_{c' \neq c}^{C} z^{[k,c]}[\tau_{c'}] \\ \vdots \\ z^{[k,c]}[M-1] - \sum_{c' \neq c}^{C} z^{[k,c]}[\tau_{c'}] \\ z^{[k,c]}[M] \end{bmatrix}, \tag{24}$$

where for the sake of simplicity, the temporal index refers to the position in the alignment block and $\tau_{c'}$ corresponds to the time slot in which the interference caused by the transmission in cluster c' is measured. Thus, the achievable rate of user k in cluster c can be written as

$$R_{\mathrm{B-NOMA}}^{[k,c]} = \eta_{\mathrm{bia}} \kappa \times \mathbb{E}\left[\log\det\left(\mathbf{I} + \gamma^{[k,c]} \tilde{\mathbf{H}}^{[k,c]} \tilde{\mathbf{H}}^{[k,c]^H} \mathbf{R}_z^{-1}\right)\right], \tag{25}$$

where $\gamma^{[k,c]}$ is the resulting SINR of the data stream associated to user k in cluster c taking into consideration the NOMA transmission, $\tilde{\mathbf{H}}^{[k,c]}$ is the channel matrix considering the M preset modes of the reconfigurable antenna, which follows the same structure as (16), and

$$\mathbf{R}_z = \begin{bmatrix} (2C-1)\mathbf{I}_{M-1} & 0 \\ 0 & 1 \end{bmatrix} \tag{26}$$

is the noise enhancement matrix. Notice that the noise increase is proportional to the number of clusters for B-NOMA. Moreover, the costs of providing CSIR are the same as for BIA.

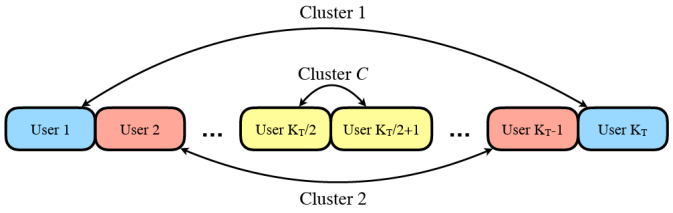


Fig. 5. NLUPA user pairing. $g^{[1]} \leq g^{[2]} \leq \cdots \leq g^{[K_T]}$.

For transmission based on BIA, constant power allocation during the entire supersymbol is assumed as proposed in [26]. That is, the resulting transmission scheme is normalized so that the same power is assigned to each time slot whether it belongs to Block 1 or Block 2. Moreover, the power allocated to each cluster is $P_c = \frac{P}{C}$. In this sense, we can define the SNR of user k in cluster c without considering the NOMA power allocation coefficients as $\rho_{k,c} = \frac{P_c g^{[k,c]}}{\sigma_z^2}$. Following the lines of [16], the users of each cluster are organized such that $g^{[1,c]} \leq \cdots \leq g^{[k,c]} \leq \cdots \leq g^{[K,c]}$. Moreover, the power P_c is distributed among the K users of cluster c according to the NOMA power coefficients satisfying $\sum_{k=1}^{K} \alpha^{[k,c]} = 1$. These coefficients are determined according to some power allocation strategy afterwards. Thus, the first user of each cluster treats the interference caused by the transmission to other users of the same cluster as noise. Therefore, the resulting SINR for the first user of cluster c is given by

$$\gamma_{1,c} = \frac{P_c g^{[1,c]} \alpha^{[1,c]}}{P_c g^{[1,c]} \sum_{k'=2}^{K} \alpha^{[k',c]} + \sigma_z^2} = \frac{\alpha^{[1,c]}}{\sum_{k'=2}^{K} \alpha^{[k',c]} + \frac{1}{\rho_{1,c}}}. \tag{27}$$

On the other hand, user K of the cluster performs SIC to remove the interference caused by transmission to the other $K-1$ users before decoding the desired symbol. Thus, the resulting SINR for user K is

$$\gamma_{K,c} = \frac{P_c g^{[K,c]} \alpha^{[K,c]}}{\sigma_z^2} = \rho_{K,c} \alpha^{[K,c]}. \tag{28}$$

Although user pairing is considered in the following sections, it results useful to determine the SINR for the general case in order to provide a better understanding of the B-NOMA scheme. For user k, the interference caused by transmission to the first $k-1$ users is subtracted based on SIC, while the interference because of transmission to the following users $k' > k$ is treated as noise. Thus, the resulting SINR of user k in cluster c is

$$\gamma_{k,c} = \frac{P_c g^{[k,c]} \alpha^{[k,c]}}{P_c g^{[k,c]} \sum_{k'>k}^{K} \alpha^{[k',c]} + \sigma_z^2} = \frac{\alpha^{[k,c]}}{\sum_{k'>k}^{K} \alpha^{[k',c]} + \frac{1}{\rho_{k,c}}}. \tag{29}$$

V. Impact of User Pairing in B-NOMA

The B-NOMA scheme derived in the previous section considers clusters composed of K users. Following the lines of other works, e.g., [5], [6], in this section we focus on the case in which each cluster is composed of $K = 2$ users. Notice, that the performance of the proposed scheme

is strongly determined by the differences in the channel gain among the users of the same cluster (see (27) and (28)). In this sense, the NOMA power allocation coefficients that satisfy a specific criterion depend on the channel gain of the users that compose each cluster. Specifically, applying the NLUPA grouping strategy over the K_T users, which are previously organized in ascending order of their channel gain, the resulting clusters are obtained as described in Fig. 5. Without loss of generality, we focus on a generic cluster denoting the users with lower and higher channel gains as j and k, respectively. Moreover, we remove the cluster index for the sake of simplicity. Hence, the resulting SINR of users j and k are

$$\gamma_j = \frac{1-\alpha}{\alpha + \frac{1}{\rho_j}} \tag{30}$$

and

$$\gamma_k = \rho_k \alpha, \tag{31}$$

respectively.

A. Maximize sum-rate solution

Traditional NOMA schemes, e.g., MISO-NOMA, are based on sharing a single spatial dimension among several users by assigning a power coefficient to each user. In this sense, each user receives a data stream through a channel gain, which is determined by the precoding vector that defines each cluster in each time slot. In this case, maximizing the sum-rate of each cluster leads to allocate all the power to the user with the highest channel gain. In contrast to this approach, the rate of each user for B-NOMA is determined by the channel response of the M preset modes of the reconfigurable antennas. These channel responses form the channel matrix $\tilde{\mathbf{H}}^{[k,c]} \in \mathbb{C}^{M\times M}$ (see (16) and (25). Moreover, the entries of $\tilde{\mathbf{H}}^{[k,c]}$ are Gaussian, and therefore, $\tilde{\mathbf{H}}^{[k,c]}\tilde{\mathbf{H}}^{[k,c]^T}$ is a random matrix following a Wishart distribution [34]. As a consequence, the power allocation that maximizes the sum-rate of a pair of users for B-NOMA can be between 0 an 1.

In the following, we focus on deriving the NOMA power coefficients that maximize the sum-rate of the considered pair of users for the proposed B-NOMA scheme. As can be seen in Theorem 1, this solution is given by a polynomial function, which can be easily solved using the polynomial long division algorithm [35]. Furthermore, with the aim of providing a closed-form expression of the NOMA power coefficients, we consider only the strongest contribution of this polynomial function in Lemma 1.

Theorem 1. *For the proposed B-NOMA scheme, the NOMA power coefficients that maximize the rate of the user pair is given by maximizing the following polynomial function*

$$\begin{aligned} \max_{\alpha} \; & f(\alpha) = \sum_{m=0}^{M}\sum_{n=0}^{M} \beta(m)\beta(n)\tilde{\gamma}_k^m(\alpha)\tilde{\gamma}_j^n \\ \text{s.t.} \quad & \alpha \in [0,1], \end{aligned} \tag{32}$$

where $\beta(i) = \binom{M}{i}\frac{M!}{(M-i)!}$, $\tilde{\gamma_j} = \frac{\gamma_j}{2^C-1}$, $\tilde{\gamma_k} = \frac{\gamma_k}{2^C-1}$ *and* γ_j *and* γ_k *are defined in* (30) *and* (31)*, respectively.*

Proof. See Appendix A. □

The function $\beta(m)$ is composed of factorial and combinatorial numbers, and therefore, it is easy to find a maximum value $\beta_{\max} = \max_m \beta(m)$ given by $m_{\max}$ that provides $\beta_{\max} >> \beta(m)$, $\forall m \neq m_{\max}$. For some specific cases, the function $\beta(m)$ provides the same value for two adjacent values $m_{\max}$ and $m_{\max}+1$, while all other values of m provide a much lower evaluation of $\beta(m)$. In this cases, we select the maximum m that satisfies this condition, i.e., $m_{\max}+1$. In the following, we propose a Lemma based on the Theorem 1 that determines the NOMA power coefficients considering exclusively $\beta_{\max}$.

Lemma 1. *Considering the strongest contribution of the polynomial function* (32)*, i.e.,* $\beta_{\max} = \max_m \beta(m)$*, the NOMA power coefficients are given by solving the equation* $A\alpha^2 + B\alpha + C = 0$*, where* $A = -1$*,* $B = -\frac{2}{\rho_j}$ *and* $C = \frac{1}{\rho_j}$*. The closed-form expression for the NOMA power coefficient is given by taking the positive solution of* (34)*. In this case, it can be written as* $\alpha = \min\left(\left[\frac{1\pm 2\sqrt{1+\rho_j}}{\rho_j}\right]^+, 1\right)$.

Proof. Considering exclusively the contribution of the value $\beta_{\max}$, the function $f(\alpha)$ can be re-written as

$$\tilde{f}(\alpha) = \beta_{\max}^2\gamma_j\gamma_k = \beta_{\max}^2\left(\frac{\rho_k\alpha(1-\alpha)}{\alpha+\frac{1}{\rho_j}}\right). \tag{33}$$

Thus, maximizing $\tilde{f}(\alpha)$ requires to obtain the first derivative, which equalized to zero leads to

$$\frac{\partial\tilde{f}(\alpha)}{\partial\alpha} = 0 \Rightarrow A\alpha^2 + B\alpha + C = 0, \tag{34}$$

where $A=-1$, $B=-\frac{2}{\rho_j}$ and $C=\frac{1}{\rho_j}$. □

B. Fair solution

Typically, maximizing the sum-rate leads to unfair rates between both users. In the following, we derive the NOMA coefficients that provide the same rate for both users. Similar to the solution that maximizes the sum-rate, we propose an approximation to obtain the closed-form expression of the NOMA coefficients.

Theorem 2. *For the B-NOMA scheme, the NOMA power coefficients that obtain a fair solution for each user pair, i.e.,* $R^{[j]} = R^{[k]}$*, is given by solving the following polynomial equation*

$$\begin{aligned} \min_{\alpha} \; & \sum_{m=0}^{M} \beta(m)\left(\tilde{\gamma}_k^m - \tilde{\gamma}_j^m\right) \\ \text{s.t.} \quad & \alpha \in [0,1]. \end{aligned} \tag{35}$$

Proof. A fair solution minimizes the difference between the rates achieved by users j and k. Thus, taking into consideration the expression of the user rate derived in Appendix A (see (40)), the fairness solution corresponds to solving the above problem. □

Lemma 2. *Considering the strongest contribution of the polynomial function that obtains the fairness solution, i.e.,* $\beta_{\max} = \max_m \beta(m)$*, the NOMA power allocation coefficient is given by solving the equation* $A\alpha^2 + B\alpha + C = 0$*, where* $A = \rho_k$*,* $B = \left(\frac{\rho_k}{\rho_j} + 1\right)$ *and* $C = -1$*. The closed-form expression of the NOMA power coefficient is given by taking the positive solution of this equation. This value is omitted for space limitations.*

Proof. If only one term of the objetive function of (35) is considered, the fair solution corresponds to solving the equation $\tilde{\gamma}_k - \tilde{\gamma}_j = 0$. Given the definition of γ_k and γ_j, it can be easily checked that the solution corresponds to the proposed equation. □

C. Cognitive B-NOMA

Until now we have considered a NOMA power allocation strategy based on maximizing the sum-rate or providing a fairness among users. However, NOMA schemes are usually proposed for serving a large number of devices with diverse data rate. Thus, we consider user k as a primary user, which defines a required QoS given by a minimum rate denoted as $R_\star^{[k]}$. Therefore, the BS first tries to satisfy this target rate, which is given by a NOMA power allocation coefficient α. Once the QoS of user k is guaranteed, user j can obtain a non-zero rate given by the remaining available power, i.e., $1-\alpha$, in a cognitive fashion.

Theorem 3. *For a target rate* $R_\star^{[k]}$*, the NOMA power coefficient for the proposed B-NOMA scheme that satisfies the QoS of the user* k *is*

$$\alpha = \min\left(\frac{2^{\left(\frac{R_\star^{[k]}}{\kappa} - \sum_{m=0}^{M-1}\Psi(M-m)+(M-1)\log(2C-1)\right)\frac{1}{M}}}{\rho_k}, 1\right), \tag{36}$$

where $\kappa = \frac{\eta_{\text{bia}}}{M+C-1}$ *and* Ψ *is the digamma function [34].*

Proof. See Appendix B. □

VI. Simulation Results

We now present the simulation results with the aim of characterizing the performance of the proposed B-NOMA scheme in comparison to the benchmark schemes introduced in Section III. Moreover, we consider the RZF beamforming scheme described in [16], and a user pairing strategy in which the BS transmits to $2M$ users organized in $C = M$ clusters. For determining the costs of providing CSIT, it is assumed that $\theta_{\text{fb}} = \theta_{\text{cd}} = 1\%$ [12]. Moreover, for the EP we consider a 5G-based orthogonal frequency division multiplexing system where each resource block is composed of 12 subcarriers and 14 symbols in which 8 pilots are transmitted as reference signals then, $\theta_{\text{ep}} = \frac{8}{12\times 14} \approx 4.8\%$ [36].

In Fig. 6, we plot the rate region for $M = 2$ and ρ_k fixed to 12 dB while $\rho_j = \{2, 7\}$ dB, i.e., a difference of 10 dB and 5 dB between both users, respectively. It can be seen that the proposed B-NOMA outperforms the considered benchmark schemes. For both MISO-OMA and BIA schemes, the user

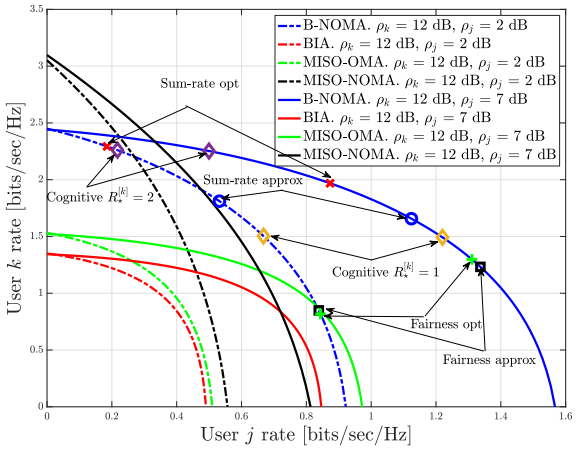


Fig. 6. User-rate region. $M = 2$, $\rho_k = 12$ dB and $\rho_j = \{2, 7\}$ dB.

rate is considerably penalyzed by the utilization of orthogonal resource allocation and the number of alignment blocks per supersymbol length of each user, respectively. In comparison to MISO-NOMA, it should be noted that user k can achieve a greater rate than for B-NOMA at the cost of a lower rate for user j. Moreover, we plot the points corresponding to the optimal sum-rate and fairness for B-NOMA with the aim of comparing these points with the values obtained by the closed-form expressions of the NOMA power coefficient provided in Lemma 1 and Lemma 2. Interestingly, the approximation for maximizing the sum-rate for B-NOMA provides a more fair solution, i.e., the rate of user j is closer to the rate of user k than for the optimal solution. Considering the optimal power allocation for maximizing the sum-rate, B-NOMA obtains a sum-rate equal to 2.47 bits/sec/Hz and 2.84 bits/sec/Hz for ρ_j equal to 2 dB and 7 dB, respectively, while these values are 2.34 bits/sec/Hz and 2.78 bits/sec/Hz for the proposed approximation. Similarly, the fairness approximation obtains a small difference between the rate achieved by both users. Furthermore, it can be seen that the cognitive approach for $R_\star^{[k]} = \{1, 2\}$ bits/sec/Hz is more effective for higher target data rates.

The rate region for $M = 2$, $\rho_k = 20$ dB and $\rho_j = \{10, 15\}$ dB is shown in Fig. 7. For the B-NOMA scheme, the maximum rate of user j and user k increases about 66% and 90% when compared with the case described in Fig. 6. This behaviour also occurs for the MISO-OMA and BIA schemes. However, the rate of user j for MISO-NOMA barely increases, since enhancing the transmitted power also increases the inter-cluster interference. For the proposed sum-rate approximation, B-NOMA achieves a sum-rate of 4.59 bits/sec/Hz and 5.28 bits/sec/Hz in the considered cluster for ρ_j equal to 15 dB and 20 dB, respectively. On the other hand, the fair solution provides about 1.96 bits/sec/Hz and 2.6 bits/sec/Hz for both users in these cases. It can be seen that both approximations are close to the optimal solution. Interestingly, it is noticed that the cognitive approach is effective for $R_\star^{[k]} = 3$ bits/sec/Hz, while it is not so accurate for lower rates.

Increasing the number of antennas of the BS to $M = 4$,

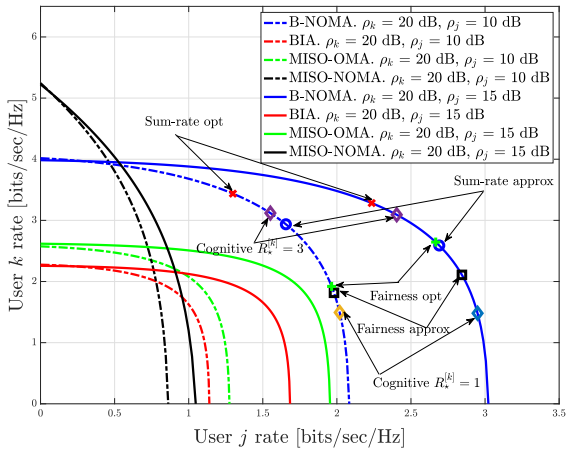


Fig. 7. User-rate region. $M = 2$, $\rho_k = 20$ dB and $\rho_j = \{10, 15\}$ dB.

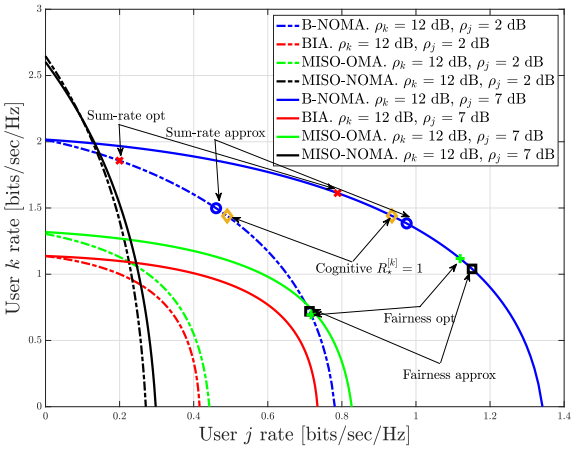


Fig. 8. User-rate region. $M = 4$, $\rho_k = 12$ dB and $\rho_j = \{2, 7\}$ dB.

we plot the rate region for $\rho_k = 12$ dB and $\rho_j = \{2, 7\}$ dB in Fig. 8. Since the considered scenario comprises 8 users, the user-rate achieved by B-NOMA and BIA slightly decreases due to the noise enhancement of subtracting more terms of interference in comparison to the case $M = 2$ (see Fig. 6). For MISO-OMA, the rate achieved by both users also suffers a small decrease since the probability of serving users whose channel response is more correlated increases, which penalyzes the performance of RZF beamforming. However, for MISO-NOMA the user rate of user k is barely penalized while the rate of user j decreases considerably. Besides, note that increasing the SNR of user j barely increases the rate achieved by MISO-NOMA since this user is subject to inter-cluster interference.

In Fig. 9 we plot the rate region for $M = 4$, $\rho_k = 20$ dB and $\rho_j = \{10, 15\}$ dB. In this case, the B-NOMA, BIA and MISO-OMA schemes follow a similar behaviour as described above. However, note that user j achieves a poor rate for MISO-NOMA independently of the SNR of the user. Therefore, MISO-NOMA provides an unfair rate distribution since the user with the best channel gain of each cluster attains a great rate while all other users, which are not considered for determining the beamforming vectors, achieve a poor rate.

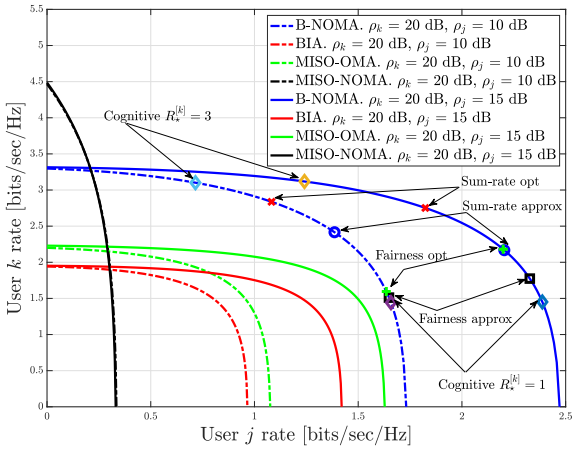


Fig. 9. User-rate region. $M = 4$, $\rho_k = 20$ dB and $\rho_j = \{15, 20\}$ dB.

The achievable sum-rate of the proposed B-NOMA scheme for distinct values of SNR is shown in Fig. 10, in comparison with the benchmarking schemes. The NOMA power allocation coefficients that maximize the sum-rate are considered for B-NOMA. Moreover, we consider a power coefficient $\alpha = 1$ for maximizing the sum-rate of MISO-NOMA. Notice that only the users with highest channel gain of each cluster are served under this condition. In order to provide transmission to both users in each cluster, we also use a NOMA power coefficient $\alpha = \frac{1}{5}$ used in [5], [6] that is sub-optimal in terms of sum-rate. First, notice that B-NOMA and MISO-NOMA for $\alpha = 1$ provide a similar sum-rate, where B-NOMA achieves a slightly greater sum-rate for high SNR values. This behaviour is consistent with the results obtained in [26] taken into consideration the costs of providing CSIT. However, the users with lower channel gain of each cluster attain zero-rates for MISO-NOMA with $\alpha = 1$, which contradict the concept of NOMA, while both users obtain non-zero rate for B-NOMA (see Fig. 6 and Fig. 7). It can be also seen that the sum-rate decreases considerably when both users of each cluster are served through MISO-NOMA with $\frac{1}{5}$. Besides, BIA and MISO-OMA achieve lower sum-rate than B-NOMA as expected from the analysis of the user-rate region described above.

The achievable sum-rate of B-NOMA versus SNR for $M = 4$ is shown in Fig. 11. It can be seen that the sum-rate follows a similar behaviour as for $M = 2$ (see Fig. 10). The B-NOMA and MISO-NOMA schemes achieve a similar sum-rate when maximizing the sum-rate. However, the sum-rate of MISO-NOMA decreases if transmission to both users of each cluster is required, e.g., assuming a NOMA power coefficient $\alpha = \frac{1}{5}$. Furthermore, MISO-OMA obtains a similar sum-rate as BIA, which is considerably lower than the performance of B-NOMA. We can conclude that the proposed B-NOMA is an efficient scheme for managing the simultaneous transmission to several clusters in MISO configurations in comparison with the proposed benchmark schemes.

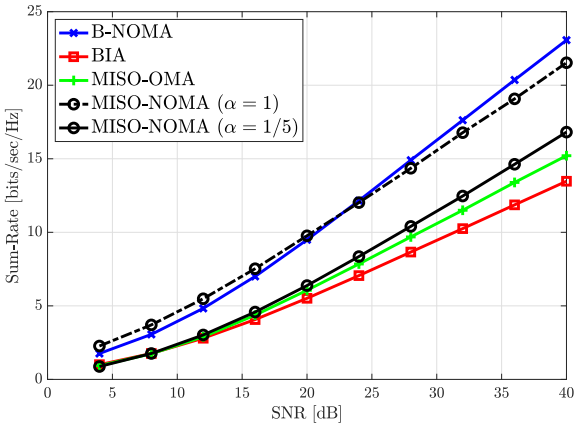


Fig. 10. Sum-rate for B-NOMA versus SNR in comparison to BIA, MISO-OMA and MISO-NOMA. $M = 2$, $\rho_{j,1} = \rho_{k,1} - 10$ dB, $\rho_{k,2} = \rho_{k,1} - 3$ dB and $\rho_{j,2} = \rho_{k,1} - 6$ dB.

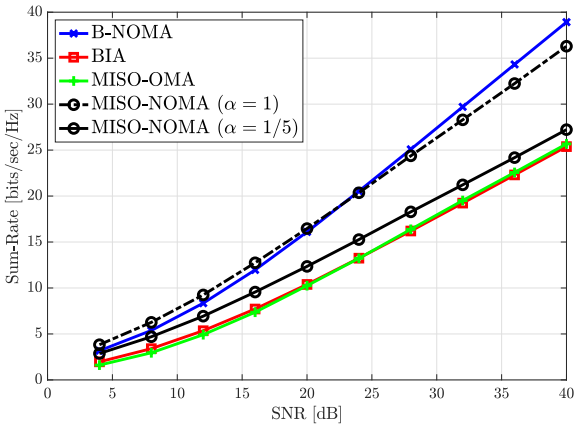


Fig. 11. Sum-rate for B-NOMA versus SNR in comparison to BIA, MISO-OMA and MISO-NOMA. $M = 4$, $\rho_{j,1} = \rho_{k,1} - 12$ dB, $\rho_{k,2} = \rho_{k,1} - 6$ dB, $\rho_{j,3} = \rho_{k,3} - 6$ dB, and $\rho_{j,4} = \rho_{k,4} - 3$ dB.

VII. Conclusions

In this paper, we proposed the use of BIA jointly with NOMA for MU-MISO systems, which is referred to as B-NOMA. In contrast to the NOMA schemes combined with beamforming strategies based on CSIT, the proposed approach completely avoids the intra and inter-cluster interference while increasing the number of data streams that are transmitted simultaneously. It is seen that B-NOMA outperforms the rate achieved by BIA and MISO-OMA schemes. The MISO-NOMA approach provides better rate for the user of each cluster with highest channel gain, when this user is exclusively considered for determining the beamforming vector of the cluster. Under this approach, the user with lower channel gain of each cluster achieves a poor rate since it is subject to inter-cluster interference. This fact makes the increase of data streams simultaneously transmitted by MISO-NOMA irrelevant, since the users that are not considered for determining the beamforming vectors are subject to inter-cluster interference, which hampers the decoding of their corresponding data streams. In this sense, the proposed B-NOMA increases the number of data streams in comparison with BIA and MISO-OMA while cancelling all sources of interference, and therefore, improving the decoding of these data streams.

Future directions

Future directions consider joint optimization of the BIA-based precoding matrices and the NOMA power coefficients based on knowledge of the large scale effects. Furthermore, considering transmission to a massive number of devices with diverse data rate and latency requirements, determining the power allocation coefficients for clusters composed of more than two users and managing the length of the supersymbol are identified as future directions of the presented work.

Appendix A

First, let us consider the following approximation for the user rate of the proposed B-NOMA scheme,

$$R^{[k]}_{\text{B-NOMA}} \approx \tilde{R}^{[k]}_{\text{B-NOMA}} = \kappa \mathbb{E}\left[\log\det\left(\mathbf{I}_M + \frac{\gamma_k}{2C-1}\mathbf{H}^{[k]}\mathbf{H}^{[k]\,H}\right)\right] + \epsilon, \tag{37}$$

where $\tilde{R}^{[k]}_{\text{B-NOMA}}$ corresponds to the case in which it is assumed that the last time slot of each alignment block is also subject to interference subtraction (see (24)), i.e., $\mathbf{R}_z = (2C-1)\mathbf{I}_M$. Assuming high SNR, i.e., $\log(1+\text{SNR}) \approx \log(\text{SNR})$, it can be easily checked that the error of the approximation, $\epsilon = R^{[k]}_{\text{B-NOMA}} - \tilde{R}^{[k]}_{\text{B-NOMA}}$, is given by

$$\begin{aligned}\epsilon &= \kappa\left(\mathbb{E}\left[\log\det\left(\gamma_k\mathbf{A}^{[k]}\mathbf{R}_z^{-1}\right) - \log\det\left(\frac{\gamma_k}{2C-1}\mathbf{A}^{[k]}\right)\right]\right) \\ &= \frac{1}{M+C-1}\log\left(2C-1\right),\end{aligned} \tag{38}$$

where $\mathbf{A}^{[k]} = \mathbf{H}^{[k]}\mathbf{H}^{[k]\,H}$. Notice that the error decreases as the number of antennas and/or clusters increases. Moreover, once these parameters are defined, ϵ can be considered a constant for optimization purposes.

Recall that the channel matrix $\mathbf{H}^{[k]} \sim \mathcal{N}(0, \mathbf{I}_M)$. Therefore, $\mathbf{A}^{[k]}$ is a Wishart matrix $\mathbf{A}^{[k]} \sim \mathcal{W}(M, \mathbf{I}_M)$. Applying Theorem 2.13 in [34],

$$\mathbb{E}\left[\det\left(\mathbf{I}_M + \tilde{\gamma}_k\mathbf{A}^{[k]}\right)\right] = \sum_{m=0}^{M}\binom{M}{m}\frac{M!}{(M-m)!}\tilde{\gamma}_k^m, \tag{39}$$

where $\tilde{\gamma}_k = \frac{\gamma_k}{2C-1}$. Therefore, using the second-order Taylor expansion of $\mathbb{E}[x]$ and evaluating its expectation under the Gaussian approximation [37], $\mathbb{E}[\log(x)] \approx \log(\mathbb{E}[x]) - \frac{\mathbb{V}[x]}{2\mathbb{E}[x]^2}$, we can re-write (39) as

$$R^{[k]} \approx \kappa\log\left(\sum_{m=0}^{M}\binom{M}{m}\frac{M!}{(M-m)!}\tilde{\gamma}_k^m\right) - S_x, \tag{40}$$

where S_x it is the statistical correction because of the considered approximation. Moreover, S_x can be considered

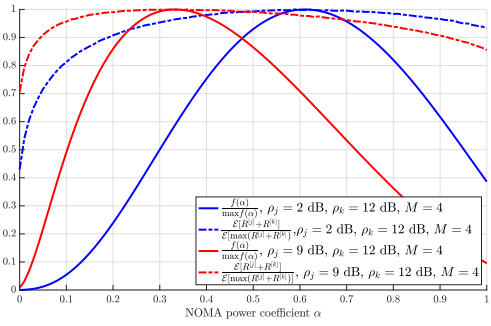


Fig. 12. Evolution of the normalized function $f(\alpha)$ and the expectation of the sum-rate with the NOMA power coefficient α.

negligible when $\mathbb{E}[x] >> 0$. Thus, maximizing the sum-rate $R^{[j]} + R^{[k]}$ in the considered pair can be written as

$$\begin{aligned} &\max_{\alpha} \sum_{m=0}^{M} \beta(m)\tilde{\gamma_k}^m \sum_{m=0}^{M} \beta(m)\tilde{\gamma_j}^m = f(\alpha) \\ &\text{s.t.} \qquad \alpha \in [0,1], \end{aligned} \tag{41}$$

where the objective function is determined by using the logarithm properties and the fact that it is a monotonically increasing function. Thus, $f(\alpha)$ can be written as

$$f(\alpha) = \sum_{m=0}^{M}\sum_{n=0}^{M} \beta(m)\beta(n)\tilde{\gamma}_k^m(\alpha)\tilde{\gamma}_j^n(\alpha). \tag{42}$$

Notice that the weights $\beta(m)$ and $\beta(n)$ are symmetric and positive for each contribution of $\tilde{\gamma}_k^m(\alpha)$ and $\tilde{\gamma}_j^n(\alpha)$. Moreover, the product $\tilde{\gamma}_k^m(\alpha)\tilde{\gamma}_j^n(\alpha)$ is strictly positive and equal to zero for $\alpha = 0$ and $\alpha = 1$ (see (30) and (31)). Thus, according to the Jensen's inequality (it is straightforward to obtain that the opposite of the inequality is true for a concave transformation), $f(\alpha)$ is a concave polynomial function in $\alpha \in [0,1]$. For illustrative purposes, we plot the function normalized function $f(\alpha)$ for $M = 4$ and distinct values of ρ_k and ρ_j in Fig. 12 in comparison with the expectation of the normalized sum-rate. Therefore, the value of α that maximizes the sum-rate of the user pair j and k can be obtained by evaluating $f(\alpha), \alpha \in [0,1]$.

Appendix B

Assuming an SNR regime high enough to consider the approximation $\log(1 + \text{SNR}) \approx \log(\text{SNR})$, the rate of user k can be written as

$$R^{[k]} = \kappa \mathbb{E}\left[\log\det\left(\gamma_k \tilde{\mathbf{H}}^{[k]}\tilde{\mathbf{H}}^{[k]^H}\mathbf{R}_z^{-1}\right)\right], \tag{43}$$

where recall that $\kappa = \frac{\eta_{\text{bia}}}{M+C-1}$. Thus, by employing the properties of the determinant

$$\begin{aligned} \frac{R^{[k]}}{\kappa} &= \log\left(\gamma_k^M \det\left(\mathbf{R}_z^{-1}\right)\right) + \mathbb{E}\left[\log\det\left(\tilde{\mathbf{H}}^{[k]}\tilde{\mathbf{H}}^{[k]^H}\right)\right] \\ &\overset{(a)}{=} M\log(\gamma_k) - (M-1)\log(2C-1) + \sum_{m=0}^{M-1}\Psi(M-m), \end{aligned} \tag{44}$$

where step (a) uses de fact that $\det(\mathbf{R}_z^{-1}) = (2C-1)^{-(M-1)}$ as well as Theorem 2.11 in [34] since $\mathbf{A}^{[k]} = \tilde{\mathbf{H}}^{[k]}\tilde{\mathbf{H}}^{[k]^H}$ is a Wishart matrix $\mathbf{A}^{[k]} \sim \mathcal{W}(M, \mathbf{I}_M)$. That is, $\mathbb{E}\left[\log\det\left(\tilde{\mathbf{H}}^{[k]}\tilde{\mathbf{H}}^{[k]^H}\right)\right] = \sum_{m=0}^{M-1}\Psi(M-m)$. Therefore, since $\gamma_k = \rho_k\alpha$, and after some algebraic manipulation,

$$\alpha = \frac{2^{\left(\frac{R_\star^{[k]}}{\kappa} - \sum_{m=0}^{M-1}\Psi(M-m) + (M-1)\log(2C-1)\right)\frac{1}{M}}}{\rho_k}. \tag{45}$$

References

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

[2] K. Yang, N. Yang, N. Ye, M. Jia, Z. Gao, and R. Fan, "Non-orthogonal multiple access: Achieving sustainable future radio access," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 116–121, February 2019.

[3] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *CoRR*, vol. abs/1809.07224, 2018. [Online]. Available: http://arxiv.org/abs/1809.07224

[4] 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 Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–5.

[5] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[6] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5319–5332, Dec 2017.

[7] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

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

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

[10] 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 Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[11] ——, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 534–537, Aug 2017.

[12] S. A. Ramprashad, G. Caire, and H. C. Papadopoulos, "Cellular and network MIMO architectures: MU-MIMO spectral efficiency and costs of channel state information," in *Proc. Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, 2009, pp. 1811–1818.

[13] J. Cui, Z. Ding, and P. Fan, "Outage probability constrained MIMO-NOMA designs under imperfect CSI," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8239–8255, Dec. 2018.

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

[15] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Military Communications Conference*, Nov. 2013, pp. 1278–1283.

[16] S. Park, A. Q. Truong, and T. H. Nguyen, "Power control for sum spectral efficiency optimization in MIMO-NOMA systems with linear beamforming," *IEEE Access*, vol. 7, pp. 10 593–10 605, 2019.

[17] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

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

[19] H. Al-Obiedollah, K. Cumanan, J. Thiyagalingam, A. G. Burr, Z. Ding, and O. A. Dobre, "Sum rate fairness trade-off-based resource allocation technique for MISO NOMA systems," *Available Online: https://arxiv.org/abs/1902.05735*, 2019.

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

[21] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2019.

[22] S. A. Jafar, "Blind interference alignment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 3, pp. 216–227, June 2012.

[23] T. Gou, C. Wang, and S. A. Jafar, "Aiming perfectly in the dark-blind interference alignment through staggered antenna switching," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2734–2744, June 2011.

[24] R. Qian and M. Sellathurai, "Performance of the blind interference alignment using ESPAR antennas," in *Proc. IEEE International Conference on Communications (ICC)*, June 2013, pp. 4885–4889.

[25] ——, "On the implementation of blind interference alignment with single-radio parasitic antennas," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10 180–10 184, Dec. 2016.

[26] C. Wang, H. C. Papadopoulos, S. A. Ramprashad, and G. Caire, "Design and operation of blind interference alignment in cellular and cluster-based systems," in *Proc. Information Theory and Applications Workshop*, Feb. 2011, pp. 1–10.

[27] M. Morales-Céspedes, J. Plata-Chaves, D. Toumpakaris, S. A. Jafar, and A. G. Armada, "Blind interference alignment for cellular networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 1, pp. 41–56, Jan 2015.

[28] M. Morales-Céspedes, J. Plata-Chaves, A. G. Armada, and L. Vandendorpe, "A blind interference alignment scheme for practical channels," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[29] S. H. Chae and S. Chung, "Blind interference alignment for a class of K-user line-of-sight interference channels," *IEEE Transactions on Communications*, vol. 60, no. 5, pp. 1177–1181, May 2012.

[30] Y. Lu, W. Zhang, and K. B. Letaief, "Blind interference alignment with diversity in K-user interference channels," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2850–2859, Aug 2014.

[31] H. Yang, W. Shin, and J. Lee, "Hierarchical blind interference alignment over interference networks with finite coherence time," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1289–1304, March 2016.

[32] H. Cha, S. Jeon, and D. K. Kim, "Blind interference alignment for the K -user MISO BC under limited symbol extension," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2861–2875, June 2018.

[33] T. Gou, S. A. Jafar, and C. Wang, "On the degrees of freedom of finite state compound wireless networks," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3286–3308, June 2011.

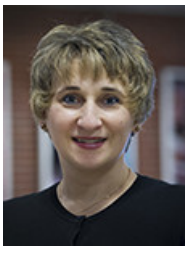
[34] A. M. Tulino and S. Verdu, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.

[35] C. Strickland-Constable, "89.65 a simple method for finding tangents to polynomial graphs," *The Mathematical Gazette*, vol. 89, no. 516, pp. 466–467, 2005.

[36] "TS 38.211 evolved universal terrestrial radio access (e-utra); physical channels and modulation." 3GPP, Tech. Rep.

[37] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 1353–1360.

Máximo Morales-Céspedes (S'10-M'15) was born in Valdepeñas, C. Real, Spain, in 1986. He received the B.Sc., and M.Sc., and Ph.D. degrees from the Universidad Carlos III de Madrid, Spain, in 2010, 2012, 2015, respectively, all in electrical engineering, with a specialization in Multimedia and Communications. In 2012 he was finalist of the IEEE Region 8 Student Paper Contest. From 2015 to 2017 he has been working as a postdoctoral fellow with the Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM) at Universite Catholique de Louvain. Currently, he is with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Spain. He serves as Editor in IEEE COMMUNICATION LETTERS and has been TPC member for numerous IEEE conferences (VTC, Globecom, ICC and WCNC). His research interests are interference management, hardware implementations, MIMO techniques and signal processing applied to wireless communications.



Octavia A. Dobre (M'05-SM'07-F'20) Octavia A. Dobre received the Dipl. Ing. and Ph.D. degrees from Politehnica University of Bucharest (formerly Polytechnic Institute of Bucharest), Romania, in 1991 and 2000, respectively. Between 2002 and 2005, she was with New Jersey Institute of Technology, USA and Politehnica University of Bucharest. In 2005, she joined Memorial University, Canada, where she is currently Professor and Research Chair. She was a Visiting Professor with Massachusetts Institute of Technology, USA and Université de Bretagne Occidentale, France. Her research interests include enabling technologies for beyond 5G, blind signal identification and parameter estimation techniques, as well as optical and underwater communications. She authored and co-authored over 250 refereed papers in these areas. Dr. Dobre serves as the Editor-in-Chief (EiC) of the IEEE Open Journal of the Communications Society, as well as an Editor of the IEEE Communications Surveys and Tutorials, IEEE Vehicular Communications Magazine, and IEEE Systems. She was the EiC of the IEEE Communications Letters, as well as Senior Editor, Editor, and Guest Editor for various prestigious journals and magazines. Dr. Dobre was the General Chair, Technical Program Co-Chair, Tutorial Co-Chair, and Technical Co-Chair of symposia at numerous conferences. Dr. Dobre was a Royal Society Scholar and a Fulbright Scholar. She obtained Best Paper Awards at various conferences, including IEEE ICC, IEEE Globecom and IEEE WCNC. Dr. Dobre is a Distinguished Lecturer of the IEEE Communications Society and a Fellow of the Engineering Institute of Canada.

Ana Garcia Armada (S'96-A'98-M'15-SM'08) received the Ph.D. degree in electrical engineering from the Polytechnical University of Madrid in February 1998. She is currently a Professor at University Carlos III of Madrid, Spain, where she has occupied a variety of management positions (Head of Signal Theory and Communications Department, Vice-dean of Electrical Engineering, Deputy Vice-Chancellor of International Relations, among others). She is leading the Communications Research Group at this university. She has been visiting scholar at Stanford University, Bell Labs and University of Southampton. She has participated (and coordinated most of them) in more than 30 national and 10 international research projects as well as 20 contracts with the industry, all of them related to wireless communications. She is the co-author of eight book chapters on wireless communications and signal processing. She has published around 150 papers in international journals and conference proceedings and she holds four patents. She has contributed to international standards organizations, such as ITU and ETSI, is member of the expert group of the European 5G PPP and member of the advisory committee 5JAC of the ESA as expert appointed by Spain on 5G. She has served on the editorial boards of Physical Communication (2008-2017), IET Communications (2014-2017). She serves on the editorial board of IEEE Communications Letters since 2016 (Editor until Feb 2019, Senior Editor from Mar 2019, Exemplary Editor Award 2017 and 2018) and IEEE Transactions on Communications since 2019. She has served on the TPC of more than 40 conferences and she has been/is part of the organizing committee of IEEE Globecom 2019 and 2021 (General Chair), IEEE Vehicular Technology Conference (VTC) Fall 2018, Spring 2018 and 2019 and IEEE 5G Summit 2017, among others. She was the Newsletter Editor of the IEEE ComSoc Signal Processing and Consumer Electronics Committee (2017-2018) and is now the Secretary of this committee (since 2019). She has been the Secretary of the IEEE ComSoc Women in Communications Engineering Standing Committee (2016-2017) and the Chair of this committee (2018-2019). She has received the Young Researchers Excellence Award, the Award to Outstanding achievement in research, teaching and management and the Award to Best Practices in Teaching, all from University Carlos III of Madrid, She was awarded the third place Bell Labs Prize 2014 for shaping the future of information and communications technology. She received the Outstanding service award from the IEEE ComSoc Signal Processing and Communications Electronics (SPCE) technical committee in 2019. Her main interests are multi-carrier and multi-antenna techniques and signal processing applied to wireless communications.