Downlink MIMO-NOMA With and Without CSI: A Short Survey and Comparison

Ricardo Alberto
Instituto de Telecomunicações, and
Instituto Superior Técnico, Universidade de Lisboa,
Lisboa, Portugal

Francisco A. Monteiro
Instituto de Telecomunicações, and
Instituto Universitário de Lisboa (ISCTE-IUL),
Lisboa, Portugal
frmo@lx.it.pt

***Abstract*—Non-orthogonal multiple access (NOMA) concatenated with multiple-input multiple-output (MIMO) or with massive MIMO, has been under scrutiny for both broadband and machine-type communications (MTC), even though it has not been adopted in the latest 5G standard (3GPP Release 16), being left for beyond 5G. This paper dwells on the problems causing such cautiousness, and surveys different NOMA proposals for the downlink in cell-centered systems. Because acquiring channel state information at the transmitter (CSIT) may be hard, open-loop operation is an option. However, when users clustering is possible, due to some common statistical CSI, closed-loop operation should be exploited. The paper numerically compares these two operating modes. The users are clustered in beams and then successive interference cancellation (SIC) separates the power-domain NOMA (PD-NOMA) signals at the terminals. In the precoded closed-loop system, the Karhunen-Loève channel decomposition is used assuming that users within a cluster share the same slowly changing spatial correlation matrix. For a comparable number of antennas the two options perform similarly, however, while in the open-loop downlink the number of antennas at the BS is limited in practice, this restriction is waived in the precoded systems, with massive MIMO allowing for a larger number of clusters.**

***Index Terms*—Non-orthogonal multiple access, successive interference cancellation, downlink, linear inter-cluster cancellation.**

I. Introduction

Non-orthogonal multiple access (NOMA) has been much studied for next generation wireless systems, including 5G, when dense networks are envisaged due to its ability to further enhance the overall spectral efficiency [1, 2, 3]. In contrast to orthogonal multiple access (OMA), in NOMA all users are superimposed in the time, frequency, or the code domains and then separated by means of successive interference cancellation (SIC) or parallel interference cancellation (PIC), and achieve all points in the capacity region of the multiple access channel (MAC) region [4]. In power-domain NOMA (PD-NOMA) the signals are separated at the receivers taking in consideration their different power levels, although other types of NOMA exist; a range of other categories of NOMA are described in [5]: scrambling-based NOMA, spreading-based NOMA, coding-based NOMA, and interleaving-based NOMA. In addition to those, NOMA based on the partitioning of lattices [6] have also proved well even when the channel gains of the users in the same cluster are similar [7, 8, 9, 10], which is a requirement for a good performance in PD-NOMA. Both multiple-input multiple-output (MIMO) schemes compared in this paper use PD-NOMA, overwhelmingly the most studied type of NOMA. Clustered MIMO-NOMA has also been studied for millimeter waves ranging from 28 GHz up to 73 GHz [11], exhibiting a superior sum-rate than OMA, as initially suggested by [3]. In [12] multiple analog beams are formed to further create more NOMA groups and therefore increase performance for any angular distribution of the users' positions.

NOMA was included in long-term evolution (LTE) Release 14, under the name multi-user superposition transmission (MUST) [13], multiplexing two users, and despite the information-theoretic foundations of NOMA [14], its practical application has been postponed by 3GPP for the 5G standard, after having been analyzed during the 3GPP technical tasks. The authors in [15] show some reasons underpinning that decision. They show how NOMA only outperforms multi-user MIMO (MU-MIMO) when the system loading (defined in respect to the length of the quasi-orthogonal sequences used in a spreading-based or coding-based NOMA system) gets larger. However, at lower overloading factors, the use of several slots (or resources in general) by a spreading-based or coding-based NOMA makes it less spectrally efficient than MU-MIMO. NOMA only outperforms MU-MIMO at high signal-to-noise ratio (SNR), due to the inherent interference-limitation in MU-MIMO, but even so with almost negligible gain (around 1 dB).

In the case of PD-NOMA, the number of users supported in the power-domain is always very low, typically two or three users, due to SIC error propagation. The issue of fair power allocation is prone to a discussion over the definition of fairness in the context of NOMA (see [16] and references therein). In [17] it was proposed a power allocation policy based on stochastic geometry to take into account the distribution of the users' location. While most papers analyze NOMA in a single-cell scenario, the problem in real multi-cell scenarios raised the problem of how to associate users to a cell while maximizing the system's sum-rate. This problem is dealt with in [18], applying matching-theoretic algorithms. The multi-cell scenario can be enhanced by considering different types of service requirements in different cell and a power allocation algorithm that takes in consideration different types of data traffic is proposed in [19]. The benefits of PD-NOMA over OMA highly depend on the differences between the channel gains to each user and a well-designed system should have the option of switching to a OMA at times. The authors in [20] recently bridged this gap by proposing a utility cost that takes in consideration the costs and gains associated to each MAC mode such that the system can opt between them. When device-to-device (D2D) communication exists in a cellular communication environment, NOMA can be used to multiplex the communication from one transmitting

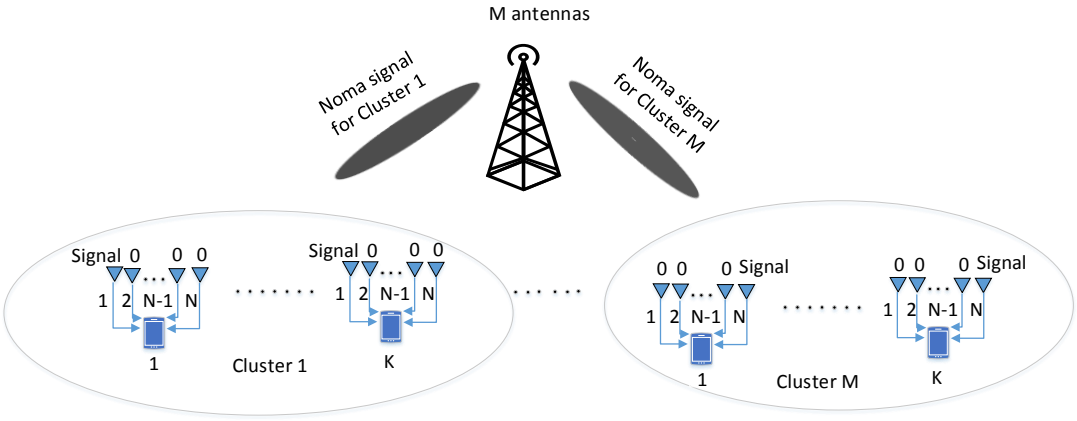

Figure 1: Open-loop MIMO-NOMA with clustered terminals with null-space-based detectors.

terminal to two receiving devices [21]. While the number of users possible to multiplex in PD-NOMA is extremely limited (2 or 3 users only), the concatenation of PD-NOMA with orthogonal frequency division multiplexing (OFDM) largely increases the number of multiplexed users in the overall system. The problem then becomes the one of user grouping and power allocation, for which simple greedy algorithms can perform quite decently [22].

When MIMO-NOMA is considered, the natural approach is to spatially cluster users which, from a MU-MIMO precoding point of view, behave as one virtual-user [23, 24]. Subsequently, the messages to each user in a cluster are separated by SIC. This requires channel state information at the transmitter (CSIT), which can be challenging to obtain, and [25] shows how to refine the quality of the CSIT. In the scenario of machine-type communications (MTC) [26], where terminals are particularly simple and energy-constrained, CSIT is even harder to attain. Moreover, user grouping is also hard given the combinatorial nature of the problem. An optimal solution to the user-selection problem to form NOMA groups (which are then differentiated in some orthogonal domain) is given in [27], but only for single-antenna BS and single-antenna users, and only for groups with two users. Obtaining CSIT with a massive MIMO base station (BS) is even more challenging, due to the sheer number of channels; a technique to mitigate intra-cluster pilot contamination has been proposed in [28].

This paper looks at the two most important MIMO-NOMA downlink setups using clusters of users: the first system operating in open-loop (analyzed in Section II) and the second in closed-loop using precoding at the BS (analyzed in Section III). Both schemes were respectively proposed by Ding et al. in [29] and [30]. The former system has also been analyzed in [31] in terms of its information-theoretic achievable rates. The uncoded system requires the number of antennas at the terminals to be equal or greater than the number of antennas at the BS in order to take advantage of the null space that the extra dimensions permit, which constitutes a strong limitation on the number of clusters. Both schemes are assessed and compared in this paper not from an information-theoretic point of view, as typical in the NOMA literature [31, 32], but rather via numerical simulation.

II. Open-loop MIMO-NOMA

A. System Model

Consider a downlink multi-user open-loop MIMO transmission with M antennas at the BS and $N \geq M$ antennas at each user, similar to the one in [29, 31], where users are grouped in M clusters of K users, multiplexed with PD-NOMA (see Fig. 1). The BS transmits $\mathbf{x} = \mathbf{P}\tilde{\mathbf{s}}$, where $\mathbf{P}$ is the $M \times M$ precoding matrix, which in the open-loop system corresponds to an identity matrix, given that there is precoding at the BS, and therefore no CSIT is needed at the BS. The transmitted vector $\tilde{\mathbf{s}} \in \mathbb{C}^{M\times 1}$ is constructed as:

$$\tilde{\mathbf{s}} = \begin{bmatrix} \alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K}s_{1,K} \\ \vdots \\ \alpha_{M,1}s_{M,1} + \cdots + \alpha_{M,K}s_{M,K} \end{bmatrix}, \tag{1}$$

where $s_{m,k} \in \mathbb{C}$ is the BPSK or QAM symbol to be transmitted to the k-th user in the m-th cluster and the coefficient $\alpha^2_{m,k} \in [0, 1]$ defines the power allocation for the k-th user in the m-th cluster. This system can be seen as a multi-user MIMO (MU-MIMO) (also known as the broadcast channel [6]), where each cluster plays the role of an aggregated virtual-user, and later the information to each user within each cluster is distilled from the NOMA symbol detected by the cluster. The set of power coefficients is selected such that $\sum_{k=1}^{K} \alpha^2_{m,k} = 1$ [29]. In the worst case, a user within a cluster will have to decode $K - 1$ signals from other users with higher power allocation coefficients than its own. The signal received at the k-th user in the first cluster is:

$$\mathbf{y}_{1,k} = \mathbf{H}_{1,k}\tilde{\mathbf{s}} + \mathbf{n}_{1,k}, \tag{2}$$

where $\mathbf{H}_{1,k} \in \mathbb{C}^{N\times M}$ is the Rayleigh flat-fading matrix from the BS to the k-th user in the first cluster and $\mathbf{n}_{1,k}$ is the unit power additive white Gaussian noise vector for k-th user in the first cluster. The noise is taken from an independent circularly symmetric complex Gaussian distribution. i.e., $\mathbf{n}_{1,k} \sim \mathcal{CN}(0, \sigma_n^2) \in \mathbb{C}^{1\times K}$. The channel matrix for the first user in the first cluster, is denoted as $\mathbf{H}_{1,1} \in \mathbb{C}^{N\times M}$. Linear detection at each terminal is made by multiplying the incoming signal (2)

by the detection vector, leading to:

$$\mathbf{v}_{1,k}^H \mathbf{y}_{1,k} = \mathbf{v}_{1,k}{}^H \mathbf{H}_{1,k} \mathbf{w}_m \tilde{\mathbf{s}} + \mathbf{v}_{1,k}^H \mathbf{n}_{1,k}, \tag{3}$$

where $\mathbf{v}_{1,k}{}^H$ denotes the Hermitian transpose of $\mathbf{v}_{1,k}$, and $\mathbf{w}_l$ is an indicator vector. This relation can be expanded, knowing that at the first cluster one is interested only in the sum $\alpha_{1,1}\mathbf{s}_{1,1} + \cdots + \alpha_{1,K}\mathbf{s}_{1,K}$:

$$\begin{aligned} \mathbf{v}_{1,k}^H \mathbf{y}_{1,k} = \mathbf{v}_{1,k}^H \mathbf{H}_{1,k} \mathbf{w}_1 (\alpha_{1,1} s_{1,1} + \cdots + \alpha_{1,K} s_{1,K}) + \\ \sum_{m=2}^{M} \mathbf{v}_{1,k}^H \mathbf{H}_{1,k} \mathbf{w}_m \tilde{\mathbf{s}}_m + \mathbf{v}_{1,k}^H \mathbf{n}_{1,k}, \end{aligned} \tag{4}$$

where $\tilde{\mathbf{s}}_m \in \mathbb{C}$ is the contribution of cluster m to the $\tilde{\mathbf{s}}$ vector. The aim is to eliminate the inter-cluster interference $\sum_{m=2}^{M} \mathbf{v}_{1,k}^H \mathbf{H}_{1,k} \tilde{\mathbf{s}}_m$ in the first cluster, which amounts to imposing:

$$\mathbf{v}_{m,k}^H \mathbf{H}_{i,k} \mathbf{w}_m = 0, \tag{5}$$

for any $i \neq m$. The matrix $\tilde{\mathbf{H}}_{i,k} \in \mathbb{C}^{N \times M-1}$ is constructed by removing the m-th column of the matrix $\mathbf{H}_{m,k}$. The problem can now be rewritten as:

$$\mathbf{v}_{m,k}^H \underbrace{\left[\mathbf{h}_{1,ik} \cdots \mathbf{h}_{m-1,ik}\, \mathbf{h}_{m+1,ik} \cdots \mathbf{h}_{M,ik}\right]}_{\tilde{\mathbf{H}}_{i,k}} = 0, \tag{6}$$

where $\mathbf{h}_{m,ik} \in \mathbb{C}^{N \times 1}$ is the m-th column of the $\mathbf{H}_{i,k}$ matrix. It is clear from equation (6) that $\mathbf{v}_{m,k}^H \in \mathbb{C}^{N \times 1}$ must belong to a space that is orthogonal to $\tilde{\mathbf{H}}_{i,k}$. Let us expand the matrix $\tilde{\mathbf{H}}_{m,k}$ into its SVD decomposition for the case $M = N$:

$$\tilde{\mathbf{H}}_{i,k} = \mathbf{U}_{i,k} \lambda \mathbf{V}^{\mathbf{T}} \tag{7}$$

where $\mathbf{U}_{i,k}$ is a unitary matrix:

$$\begin{bmatrix} U_{1,1} & U_{1,2} & \dots & U_{1,N-1} & \underbrace{\begin{matrix} U_{1,N} \\ \vdots \\ U_{N,N} \end{matrix}}_{\tilde{\mathbf{U}}_{i,k}} \\ \vdots & & & & \\ U_{N,1} & U_{N,2} & \dots & U_{N,N-1} & \end{bmatrix} \tag{8}$$

and λ has the diagonal form:

$$\begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & \lambda_{min(M,N)} & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}. \tag{9}$$

Note that (9) has a zero row at the bottom (even if $M = N$) because after removing a column from $\mathbf{H}_{m,k}$ to create $\tilde{\mathbf{H}}_{i,k}$, the matrix becomes tall and thus rank-deficient. In general, there will be $(M - N) + 1$ rows of zeros in the matrix of singular values. One can see that the column highlighted in (8) (which is a matrix in the general case), $\tilde{\mathbf{U}}_{i,k} \in \mathbb{C}^{N \times (N-M+1)}$, does not contribute to $\tilde{\mathbf{H}}_{i,k}$ since it is multiplied by the row of zeros (or a zero fat matrix in general), thus spanning a space orthogonal to $\tilde{\mathbf{H}}_{i,k}$. Next, one projects the $\mathbf{h}_{m,ik}$ column onto the orthogonal space using the projection matrix $\mathbf{P}_U = \tilde{\mathbf{U}}_{i,k} \tilde{\mathbf{U}}_{i,k}^H$, choosing:

$$\mathbf{v}_{m,k} = \tilde{\mathbf{U}}_{i,k} \frac{\tilde{\mathbf{U}}_{i,k}^H \mathbf{h}_{m,ik}}{\|\tilde{\mathbf{U}}_{m,i,k}^H h_{m,ik}\|}, \tag{10}$$

which is equally applied by all the users in the m-th cluster, eliminating the inter-cluster interference because (10) fulfils the requirement established in (5). Consequently, $N \geq M$ antennas are needed at each user, otherwise the $\tilde{\mathbf{H}}_{i,k}$ matrix becomes fat rather than tall and thus there is no orthogonal space spanned by the columns of $\mathbf{U}_{i,k}$ in (8). Without loss of generality, focusing on the first cluster, the channel gains of the different users in the first cluster should be ordered such that $\|\mathbf{v}_{1,1}^H \mathbf{H}_{1,1}\|^2 \geq \cdots \geq \|\mathbf{v}_{1,k}^H \mathbf{H}_{1,k}\|^2$, which is equivalent to choosing $\alpha_{1,1}^2 \leq \cdots \leq \alpha_{1,k}^2$. Note that this ordering happens within each cluster, and all clusters are statistically identical. Zero-forcing (ZF) detection is then applied at each terminal in the cluster:

$$\begin{aligned} \tilde{\mathbf{y}}_{1,k} &= (\mathbf{v}_{1,k}^H \mathbf{H}_{1,k})^{-1} \mathbf{v}_{1,k}^H \mathbf{H}_{1,k} \mathbf{w}_1 \alpha_{1,1} s_{1,1} + \cdots + \alpha_{1,K} s_{1,K}) + \\ &+ (\mathbf{v}_{1,k}^H \mathbf{H}_{1,k})^{-1} \mathbf{v}_{1,k}^H \mathbf{n}_{1,k} \\ &= (\alpha_{1,1} s_{1,1} + \cdots + \alpha_{1,K}) + (\mathbf{v}_{1,k}^H \mathbf{H}_{1,k})^{-1} \mathbf{v}_{1,k}^H \mathbf{n}_{1,k}, \end{aligned} \tag{11}$$

leading to a sum of the intended NOMA signal for that cluster perturbed by a noise term.

For SIC detection to be possible with BPSK, the following constraint is imposed:

$$\alpha_{m,k} > \sum_{i=1}^{k-1} \alpha_{m,i}, \tag{12}$$

for users $1 \leq k \leq K$ in the m-th cluster, even though it disregards fairness. One follows the rule $\alpha_{m,k-1}^2 = 0.5 \times \alpha_{m,k}^2$. The rule follows the geometric progression of ratio 1/2 deprived from its first term with $k = 0$, the value of $\sum_{k=1}^{N} (1/2)^k$tends to 1 as N tends to infinity, and the restriction (12) is naturally fulfilled. A similar strategy was proposed in the context of visible light communications (VLC) using decaying factors 0.3 and 0.4 instead of 0.5 [33].

B. Performance

A two-user case PD-NOMA with null-space based MIMO is assessed with BPSK and different QAM modulation schemes in Figures 2, 3 and 4, with $M = 2$, $N = 3$, and $K = 2$ in all cases. Subsequently, a five-users case with BPSK is also assessed.

The well-known two regimes of PD-NOMA emerge, depending on the SNR. Consider user 1 the one with the lowest power

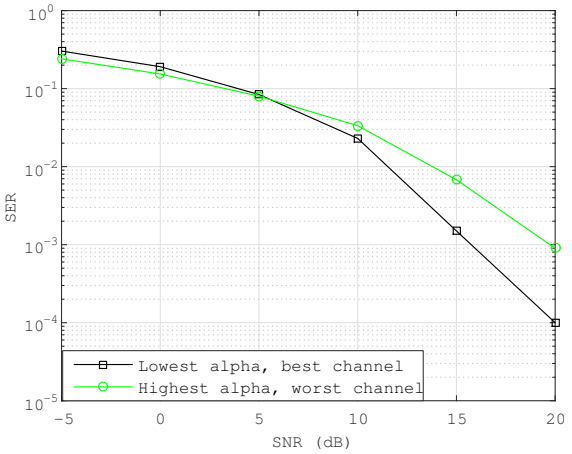

Figure 2: Open-loop with two users, both using BPSK.

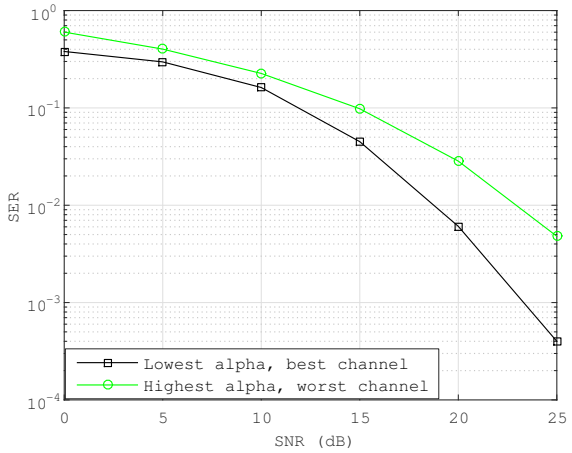

Figure 3: Open-loop with two users. User 1: BPSK; user 2: 16-QAM .

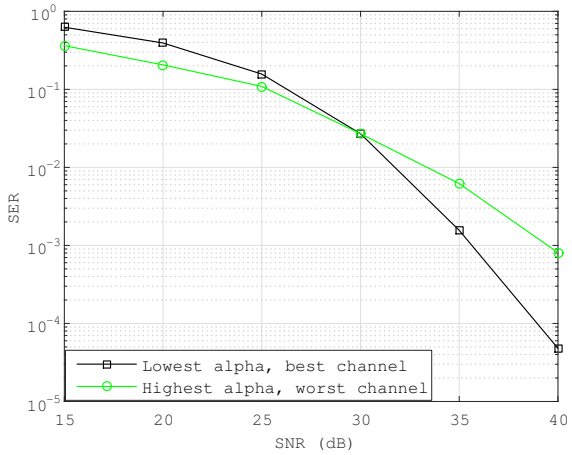

Figure 4: Open-loop with two users. User 1: 16-QAM; user 2: 64-QAM .

allocation (i.e., the one with larger channel gain). At low SNR, user 1 can incorrectly detect the signal with the larger power coefficient and propagate the error, incorrectly decoding its own signal. At high SNR, user 2 still has to cope with the extra degradation imposed by the interference from the signal to user 1, yielding a poorer performance. In Figures 2 and 4 user 2 clearly outperforms user 1 in the low SNR regime. As expected, when using higher modulation schemes at user 2, that user's performance is degraded. Interestingly, when users 1 and 2 respectively apply BPSK and 16-QAM, the two regimes do not appear in Fig. 3 because at low SNR the errors arising at initial detection stage in user 1 are not significant to corrupt the BPSK detection of user 1.

The robustness of the system is chiefly defined by the relations between the power coefficients. In Figures 2 and 3, $\alpha_1 = \sqrt{1/4}$ and $\alpha_2 = \sqrt{3/4}$ were used to compare with the results in Fig. 1 in [29]. In Fig. 4, one has $\alpha_1 = \sqrt{1/17}$ and $\alpha_2 = \sqrt{16/17}$. Comparing Fig. 2 with Fig. 1 in [29], one observes that the SER is bounded by the outage probability.

For the five user case, with $M = 2$ and $N = 2$, the users are ordered such that user 1 has the best channel and user 5 the worst. In Fig. 5, one can find that users with higher power

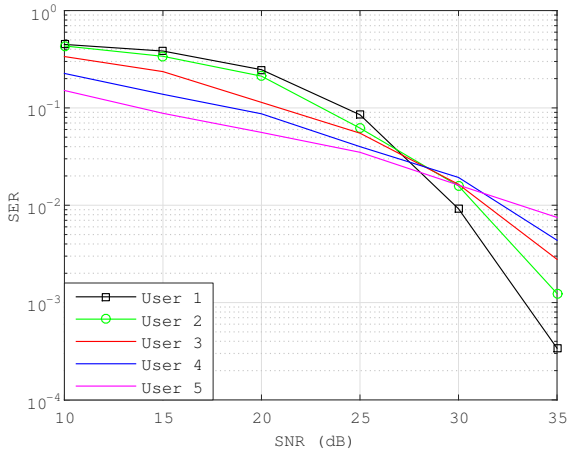

Figure 5: Open-loop with five users, all using BPSK: $\alpha_{m,1} = 0.0542$, $\alpha_{m,2} = 0.1083$, $\alpha_{m,3} = 0.2166$, $\alpha_{m,4} = 0.4332$, $\alpha_{m,5} = 0.8664$).

allocation coefficients have a better (lower) SER at low SNR and then worse performance at high SNR, exhibiting the same dual-regime. The $\alpha_{m,k}$ were defined by the set $\{1, 2, 4, 8, 16\}$, normalized by $\sqrt{\sum \alpha_i^2} = \sqrt{341}$), such that $\sum_{k=1}^{K} \alpha_{m,k}^2 = 1$. With six users and the same power allocation rule $\alpha_{m,1}$ becomes too small, and user 1 gets a SER > 0.5 for SNR $= 10$ dB, showcasing the limitations of PD-NOMA.

III. Closed-loop Massive MIMO-NOMA

A. System Model

Consider a scenario similar to the previous one, but now with a massive-MIMO BS with M antennas transmitting to users equipped with N antennas. The users are also grouped into L clusters, each of which with K users, all with different channel matrices, however, they all share the same spatial correlation matrix $\mathbf{R}_l$. In such cases one can apply the Karhunen-Loève channel decomposition [34, 35], according to which the k-th user in the l-th cluster can have its channel matrix written as:

$$\mathbf{H}_{l,k} = \mathbf{G}_{l,k} \Lambda_l^{\frac{1}{2}} \mathbf{U}_l, \tag{13}$$

where $\mathbf{G}_{l,k} \in \mathbb{C}^{N \times N}$ denotes a fast fading complex Gaussian matrix, $\Lambda_l \in \mathbb{C}^{M \times M}$ is a diagonal matrix that contains the eigenvalues of $\mathbf{R}_k$ and $\mathbf{U}_l \in \mathbb{C}^{M \times M}$ is a matrix that contains the eigenvectors of $\mathbf{R}_l$, meaning that

$$\mathbf{R}_l = \mathbf{U}_l^H \Lambda_l \mathbf{U}_l = \mathbb{E}\{\mathbf{H}_{l,k}^H \mathbf{H}_{l,k}\}, \tag{14}$$

since that a correlation matrix is always symmetric. However, $\mathbf{R}_l$ only has r_l non-zero eigenvalues, with r_l being the rank of $\mathbf{R}_l$. Therefore, Λ_l is of the form:

$$\Lambda_l = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & 0 \\ \dots & & & & & \\ 0 & 0 & \dots & \lambda_{M-r_k, M-r_k} & 0 & 0 \\ 0 & 0 & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & 0 & \lambda_{M,M} \end{bmatrix}, \tag{15}$$

and thus can be reduced to a $r_l \times r_l$ matrix, turning $\mathbf{G}_{l,k}$ a $N \times r_l$ matrix and $\mathbf{U}_l$ a $r_l \times M$ matrix. Obtaining CSIT for the fast fading matrix $\mathbf{G}_{l,k}$ may often be difficult. Because $\mathbf{R}_l$

is a slowly-changing channel correlation matrix, its estimation at the BS is easier to obtain. The BS sends a precoded $M \times 1$ NOMA vector with superimposed symbols

$$\mathbf{S} = \sum_{l=1}^{L} \mathbf{P}_l \sum_{k=1}^{K} \mathbf{w}_l \alpha_{l,k} s_{l,k}, \tag{16}$$

where $s_{l,k}$ is the symbol for the k-th user in the l-th cluster, $\alpha_{l,k}$ is the power coefficient for the k-th user in the l-th cluster. The number of effective BS antennas for each cluster is $\tilde{M}_l = (M - r_l(L-1))$ and, $\mathbf{P}_l$ is the $M \times \tilde{M}_l$ precoding matrix of the l-th cluster. $w_l = [0 \cdots 0\ 1\ 0 \cdots 0]^T$ is the $\tilde{M}_l \times 1$ precoding vector that has a 1 in the l-th position. The k-th user in the l-th cluster therefore receives

$$\mathbf{y}_{l,k} = \mathbf{G}_{l,k} \Lambda_l^{\frac{1}{2}} \mathbf{U}_l \sum_{l=1}^{L} \mathbf{P}_l \sum_{k=1}^{K} \mathbf{w}_l \alpha_{l,k} s_{l,k} + \mathbf{n}_{l,k}, \tag{17}$$

where $n_{l,k}$ is the noise at the k-th user in the l-th cluster. Looking at (17), $\mathbf{P}_l$ needs to satisfy the following constraint to eliminate inter-cluster interference:

$$[\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H \mathbf{U}_{l+1}^H \cdots \mathbf{U}_L^H]^H \mathbf{P}_l = 0. \tag{18}$$

Since $[\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H \mathbf{U}_{l+1}^H \cdots \mathbf{U}_L^H]^H$ is always a fat matrix (and thus it always has some non-zero nullspace), then

$$\mathbf{P}_l = \text{Null}([\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H \mathbf{U}_{l+1}^H \cdots \mathbf{U}_L^H]^H). \tag{19}$$

Using a $\mathbf{P}_l$ given by (19), the inter-cluster interference is removed and (17) becomes

$$\mathbf{y}_{l,k} = \mathbf{G}_{l,k} \Lambda_l^{\frac{1}{2}} \mathbf{U}_l \mathbf{P}_l \sum_{k=1}^{K} \mathbf{w}_l \alpha_{l,k} s_{l,k} + \mathbf{n}_{l,k}. \tag{20}$$

Without loss of generality, looking at user $k = 1$ in the first cluster ($l = 1$) of a system with $K = 2$ users, (20) leads to:

$$\mathbf{y}_{1,1} = \mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1} (\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2}) + \mathbf{n}_{1,1}. \tag{21}$$

The information to all users is carried by the $\tilde{M}_l \times 1$ vector:

$$[\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2} \quad 0 \cdots 0]^T, \tag{22}$$

which imposes a limit of $\tilde{M}_l$ to the number of clusters. This vector is then multiplied by the matrix $\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1$ whose dimensions are $N \times \tilde{M}$, and whose elements will be denoted as $c_{n,\tilde{m}}$. Disregarding noise, this can be written as:

$$\begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,\tilde{M}-1} & c_{1,\tilde{M}} \\ \vdots & & & & \vdots \\ c_{N,1} & c_{N,2} & \cdots & c_{N,\tilde{M}-1} & c_{N,\tilde{M}} \end{bmatrix} \begin{bmatrix} \alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2} \\ \vdots \\ 0 \end{bmatrix} \tag{23}$$

where only the first column of $\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1$ influences the received $N \times 1$ vector $\mathbf{y}_{1,1}$. Applying ZF detection leads to

$$\begin{aligned} &\tilde{\mathbf{y}}_{1,1} = (\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1})^{-1} \times \\ &[\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1} (\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2}) [1\ 0]^T + \mathbf{n}_{1,1}] \\ &= (\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2}) [1\ 0]^T + (\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1})^{-1} \\ &+ \mathbf{n}_{1,1}, \end{aligned} \tag{24}$$

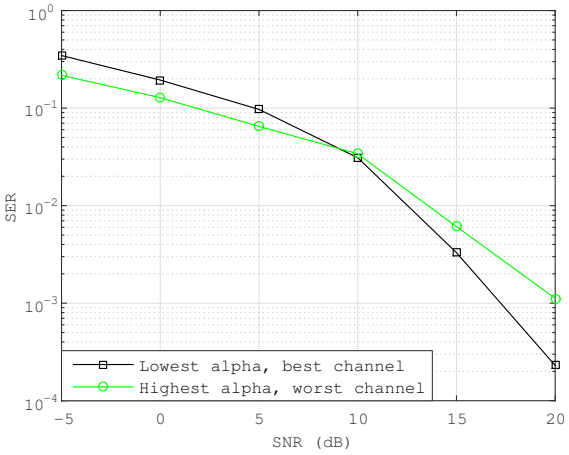

Figure 6: Closed-loop with two users, both using BPSK. M=50, N=3.

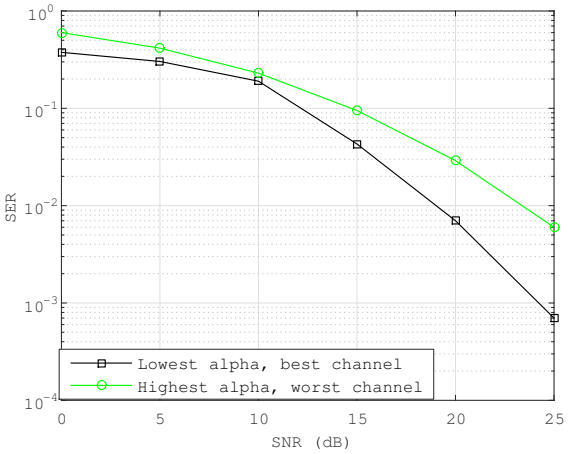

Figure 7: Closed-loop with two users. User 1: BPSK; user 2: 16-QAM. M=50, N=3.

which again, as in the previous closed-loop model, is the intended NOMA mixture for the cluster, added to a noise term.

B. Performance

A system with a massive array with $M = 50$ antennas at the BS and terminals with $N = 3$, which is the same number of antennas considered at the terminals in the open-loop setup. Comparing Figures 6 and 7 with Figures 2 and 3, also with two PD-NOMA users and the same modulations, one can see that the performances very similar. This is because the channel models of (11) and (16) are in fact equivalent in terms of end-to-end SNR per user. To understand why this happens one needs to revisit equations (11) and (24) and note that $\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 = \mathbf{H}_{1,1}$ (Karhunen-Loève decomposition) and that $\|\mathbf{v}_{m,k}\| = \|\mathbf{P}_1\| = 1$. Hence, both equations are in fact equivalent in terms of the ratio between the signal power and the noise power in each of these ZF schemes, when averaging over several channel realizations.

IV. Comparison and Conclusions

While both systems are equivalent in terms of performance, the open-loop cannot uphold a massive array at the BS because

it is limited by the number of receive antennas that the terminals can fit, while in the closed-loop model an increasing number of M antennas at the BS can lead to an arbitrarily large number of clusters. However, one should notice the trade-off that higher-rank correlation matrices impose, forcing to lower the number of clusters or the number of effective transmit antennas per cluster. It is worth mentioning that a system's designer should not only optimize the power coefficients but also consider different modulations for the users. The correlation matrix can have a rank as large as $r_l = N$, so considering for example $N = 8$ antennas at the receivers and $M = 128$ at the BS, it is possible to support $L = 15$ clusters, with $\tilde{M}_l = 128 - 8 \times (15 - 1) = 16$ effective transmit antennas per cluster. In this example, the closed-loop system can almost duplicate the number of NOMA clusters possible in open-loop, which would be $L = N = 8$. Notably, with single antenna terminals ($N = 1$), keeping the $M = 128$ and the same $\tilde{M}_l = 16$ antennas per cluster, one could support $L = 113$ clusters.

ACKNOWLEDGMENTS

This work was funded by FCT (Foundation for Science and Technology) and Instituto de Telecomunicações through national funds, and when applicable co-funded EU funds, under the project UIDB/EEA/50008/2020.

REFERENCES

[1] A. Anwar, B.-C. Seet, M. A. Hasan, and X. J. Li, "A survey on application of non-orthogonal multiple access to different wireless networks," *Electronics*, vol. 8, p. 1355, Nov. 2019.

[2] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[3] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 2181–2195, Oct. 2017.

[4] L. Liu, Y. Chi, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving MIMO-NOMA: Iterative LMMSE detection," *IEEE Transactions on Signal Processing*, vol. 67, pp. 1758–1773, Apr. 2019.

[5] Z. Wu, K. Lu, C. Jiang, and X. Shao, "Comprehensive study and comparison on 5G NOMA schemes," *IEEE Access*, vol. 6, pp. 18511–18519, 2018. Conference Name: IEEE Access.

[6] F. A. Monteiro, *Lattices in MIMO Spatial Multiplexing: Detection and Geometry*. PhD thesis, University of Cambridge, UK, 2012.

[7] M. Qiu, Y. Huang, J. Yuan, and C. Wang, "Lattice-partition-based downlink non-orthogonal multiple access without SIC for slow fading channels," *IEEE Transactions on Communications*, vol. 67, pp. 1166–1181, Feb. 2019.

[8] M. Qiu, Y. Huang, S. Shieh, and J. Yuan, "A lattice-partition framework of downlink non-orthogonal multiple access without SIC," *IEEE Transactions on Communications*, vol. 66, pp. 2532–2546, June 2018.

[9] G. Geraci, D. Fang, and H. Claussen, "A new method of MIMO-based non-orthogonal multiuser downlink transmission," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, June 2017.

[10] D. Fang, Y.-C. Huang, Z. Ding, G. Geraci, S.-L. Shieh, and H. Claussen, "Lattice partition multiple access: A new method of downlink non-orthogonal multiuser transmissions," in *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec. 2016.

[11] W. Yi, Y. Liu, A. Nallanathan, and M. Elkashlan, "Clustered millimeter wave networks with non-orthogonal multiple access," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[12] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmwave systems," *IEEE Transactions on Communications*, vol. 67, pp. 1705–1719, Feb. 2019.

[13] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[14] M. Vaezi and H. V. Poor, "NOMA: An information-theoretic perspective," in *Multiple Access Techniques for 5G Wireless Networks and Beyond* (M. Vaezi, Z. Ding, and H. V. Poor, eds.), ch. 5, p. 167–193, Springer, 2019.

[15] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 179–189, 2020.

[16] G. Gui, H. Sari, and E. Biglieri, "A new definition of fairness for non-orthogonal multiple access," *IEEE Communications Letters*, vol. 23, pp. 1267–1271, July 2019.

[17] 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, 2016.

[18] M. W. Baidas, Z. Bahbahani, and E. Alsusa, "User association and channel assignment in downlink multi-cell NOMA networks: A matching-theoretic approach," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, p. 220, Sept. 2019.

[19] F. Mokhtari, M. R. Mili, F. Eslami, F. Ashtiani, B. Makki, M. Mirmohseni, M. Nasiri-Kenari, and T. Svensson, "Download elastic traffic rate optimization via NOMA protocols," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 713–727, Jan. 2019.

[20] M. Baghani, S. Parsaeefard, M. Derakhshani, and W. Saad, "Dynamic non-orthogonal multiple access (NOMA) and orthogonal multiple access (OMA) in 5G wireless networks," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[21] S. Alemaishat, O. A. Saraereh, I. Khan, and B. J. Choi, "An efficient resource allocation algorithm for D2D communications based on NOMA," *IEEE Access*, vol. 7, pp. 120238–120247, 2019.

[22] O. A. Saraereh, A. Alsaraira, I. Khan, and P. Uthansakul, "An efficient resource allocation algorithm for OFDM-based NOMA in 5G systems," *Electronics*, vol. 8, p. 1399, Nov 2019.

[23] 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.

[24] F. A. Monteiro and I. J. Wassell, "Recovery of a lattice generator matrix from its Gram matrix for feedback and precoding in MIMO," in *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 1–6, 2010.

[25] Y. Lan, A. Benjebbour, X. Chen, A. Li, and H. Jiang, "Enhanced channel feedback schemes for downlink NOMA combined with closed-loop SU-MIMO," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–6, May 2016. ISSN: null.

[26] A. Shahini and N. Ansari, "NOMA aided narrowband IoT for machine type communications with user clustering," *IEEE Internet of Things Journal*, vol. 6, pp. 7183–7191, Aug 2019.

[27] J.-M. Kang and I.-M. Kim, "Optimal user grouping for downlink NOMA," *IEEE Wireless Communications Letters*, vol. 7, pp. 724–727, Oct. 2018.

[28] D. Kudathanthirige and G. Amarasuriya, "Massive MIMO NOMA downlink," in *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, Dec. 2018.

[29] 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, 2016.

[30] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 629–633, 2016.

[31] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, pp. 2123–2129, 2016.

[32] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Transactions on Signal Processing*, vol. 65, pp. 5191–5202, Oct. 2017.

[33] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat, "Non-orthogonal multiple access for visible light communications," *IEEE Photonics Technology Letters*, vol. 28, no. 1, pp. 51–54, 2016.

[34] A. Adhikary, J. Nam, J. Ahn, and G. Caire, "Joint spatial division and multiplexing—the large-scale array regime," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.

[35] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A hierarchical rate splitting strategy for FDD massive MIMO under imperfect CSIT," in *2015 IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)*, pp. 80–84, 2015.