A quasi-optimal clustering algorithm for MIMO-NOMA downlink systems

Fabio Saggese, *Student Member, IEEE*, Marco Moretti, *Member, IEEE*, and Andrea Abrardo, *Senior Member, IEEE*

***Abstract*—In this paper we consider a resource allocation problem for multi-user MIMO non orthogonal multiple access (MU-MIMO-NOMA) downlink transmissions. Under the NOMA paradigm, users are organized in clusters of strong/weak pair and our aim is to find an optimal clustering, beamforming and power allocation scheme to minimize the power transmitted subject to a rate constraint for each user. Since the joint optimization problem is intractable, we split it in three sub-problems: clustering, which is formulated as a mixed integer linear programming (MILP) problem, beamforming and power allocation. Simulations results show that the our proposed scheme greatly outperforms both the classical OMA scheme and state-of-the-art NOMA techniques.**

***Keywords*—MIMO, non-orthogonal multiple access, clustering, weighted bipartite matching.**

I. Introduction

Non-orthogonal multiple access (NOMA) is one of the key radio access technologies envisioned to meet the heterogeneous demands on low latency, high reliability, massive connectivity, improved fairness, and high throughput of fifth generation (5G) mobile networks [1]. The key idea of NOMA is to exploit the power domain for allowing multiple users to be served concurrently at the same time on the same frequency channel. NOMA, initially developed for a SISO setting [2], faces several challenges in MIMO systems and only recent works have shown the potential to outperform classical MIMO-OMA systems [3]–[10]. Choosing the users that are matched together on the same channel has a very relevant impact on the performance of any implementation of the NOMA paradigm. In a MIMO setting, where users are separated by a beamforming precoder, *user clustering*, the problem of choosing the users belonging to the same beam, becomes even more important because it has consequences also on the shape of the beams. Because of the complexity of the task, most of the works in literature on clustering are either based on heuristic algorithms [3]–[7], or require a large number of antennas at the mobile terminals (MT) [8], or are theoretic analysis of the capacity of the systems [9].

The contribution of this letter is a novel design algorithm that solves the optimization problem for downlink transmissions in a general MU-MIMO-NOMA scenario by splitting it in three steps: beaforming, power allocation and clustering. For a given number N of users, we employ block-diagonalization (BD) beamforming to separate the users in $N/2$ clusters, each composed by a *strong* and a *weak* user [11]. Following the NOMA paradigm, the strong user in the cluster performs successive interference cancellation (SIC) to remove the interference given by the paired weak user. For each possible strong/weak pair, the optimal beamforming matrices are computed according to a minimum power criterion subject to rate constraints. Unlike the other work in literature, the clustering technique we propose is *optimal* being the solution of a mixed integer linear programming (MILP) problem. Simulations results show that the proposed NOMA approach allows to clearly outperform classical OMA approach based on block diagonalization and state-of-the-art alternatives for MISO communications.

Fabio Saggese (fabio.saggese@phd.unipi.it) and Marco Moretti (marco.moretti@iet.unipi.it) are with Dipartimento di Ingegneria del'Informazione, University of Pisa, Italy. Andrea Abrardo (abrardo@dii.unisi.it) is with Dipartimento di Ingegneria del'Informazione, University of Siena, Italy. Corresponding author: F. Saggese.

II. System model

We focus on a NOMA-MIMO downlink transmission scenario with N users uniformly distributed in a cell of radius R. Each mobile user is equipped with N_r antennas. The number of transmit antennas at the base station (BS) is $N_t \geq NN_r$, so that there are enough spatial degrees of freedom to ideally multiplex all users in the cell without introducing intra-cell interference. In the following, we make the non-restrictive assumption that N is even. To implement the NOMA paradigm, the N users are grouped in $N_C = N/2$ clusters composed by couples of mutually interfering users, generally referred to as *strong* and *weak* users, so labelled on the base of the quality of their propagation channel. Assuming perfect channel state information at the BS, we sort in ascending order the users employing the Frobenius norm of the channel gain matrix as quality indicator. Accordingly, weak and strong users have indexes $w \in \mathcal{W} = \{1, \ldots, N_C\}$ and $s \in \mathcal{S} = \{N_C + 1, \ldots, N\}$, respectively.

The message for the generic user i, classified either as strong or weak, is spatially multiplexed on N_r data streams, so that its vector is $\mathbf{x}_i = [\sqrt{p_{1,i}}s_{1,i}, \sqrt{p_{2,i}}s_{2,i}, \ldots, \sqrt{p_{N_r,i}}s_{N_r,i}]^T$ where $p_{j,i}$ and $s_{j,i}$ are the power and the i.i.d. unitary symbol transmitted on the j-th data stream of user i, respectively, so that it is

$$\mathbb{E}\left\{s_{j,i}s^*_{\ell,m}\right\} = \begin{cases} 1 & \text{if } \ell = j \text{ and } i = m \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

and $\mathbf{P}_i = \mathbb{E}\{\mathbf{x}_i \mathbf{x}_i^H\}$ is the diagonal covariance matrix with the vector $\mathbf{p}_i = \{p_{1,i}, p_{2,i}, \ldots, p_{N_r,i}\}$ on the main diagonal.

Denoting with $\mathbf{W}_i$ the beamforming matrix for the generic user i, the transmitted signal is:

$$\mathbf{y}_i = \mathbf{W}_i \mathbf{x}_i. \tag{2}$$

The signal at the i-th receiver is

$$\mathbf{r}_i = \mathbf{H}_i \left(\sum_{w \in \mathcal{W}} \mathbf{W}_w \mathbf{x}_w + \sum_{s \in \mathcal{S}} \mathbf{W}_s \mathbf{x}_s \right) + \mathbf{n}_i, \tag{3}$$

where $\mathbf{H}_i \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix and $\mathbf{n}_i \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_r})$ is the AWGN vector at the receiver. As for the channel model, we assume uncorrelated antennas, i.e., the entries of $\mathbf{H}_i$ are i.i.d. circularly symmetric, complex Gaussian random variables.

Let us denote by $\mathcal{C} = \{w, s\}$ a generic cluster composed by $w \in \mathcal{W}$ and $s \in \mathcal{S}$ and denote by $k \in \{w, s\}$ a generic element of the cluster. The inter-cluster interference covariance matrix seen by user k can be expressed as:

$$\mathbf{J}_k = \mathbf{H}_k \left(\sum_{\substack{i \in \mathcal{W} \\ i \neq w}} \mathbf{W}_i \mathbf{P}_i \mathbf{W}_i^H + \sum_{\substack{j \in \mathcal{S} \\ j \neq s}} \mathbf{W}_j \mathbf{P}_j \mathbf{W}_j^H \right) \mathbf{H}_k^H. \tag{4}$$

In each cluster the strong NOMA user is able to remove the interference of the weak NOMA user by performing interference cancellation, while the weak user is detected in the presence of the interference caused by the strong user. This implies that the strong users have knowledge of the codebook of its weak associated user, after optimal clustering scheme has been found. Accordingly, the noise covariance matrix, obtained by summing the covariance of thermal noise, intra-cluster and intra-cluster interference is:

$$\mathbf{N}_k = \begin{cases} \sigma_n^2 \mathbf{I}_{N_r} + \mathbf{H}_k \mathbf{W}_s \mathbf{P}_s \mathbf{W}_s^H \mathbf{H}_k^H + \mathbf{J}_k, & \text{if } k = w \\ \sigma_n^2 \mathbf{I}_{N_r} + \mathbf{J}_k, & \text{if } k = s. \end{cases} \tag{5}$$

We are now in the position of deriving the achievable rate of user k as:

$$R_k = \log_2 \det\left(\mathbf{I}_{N_r} + \mathbf{H}_k \mathbf{W}_k \mathbf{P}_k \mathbf{W}_k^H \mathbf{H}_k^H \mathbf{N}_k^{-1} \right). \tag{6}$$

NOMA interference cancelation is possible under the condition that the strong user s can correctly decode (and cancel!) the signal of the weak user w. This condition can be translated into the relation

$$R_w \leq R_{w,s} \tag{7}$$

where $R_{w,s}$ is the so called *weak-strong* rate, i.e., the rate of the weak user at the receiver of the strong user

$$R_{w,s} = \log_2 \det\left(\mathbf{I}_{N_r} + \mathbf{H}_s \mathbf{W}_w \mathbf{P}_w \mathbf{W}_w^H \mathbf{H}_s^H \mathbf{N}_{w,s}^{-1} \right) \tag{8}$$

with $\mathbf{N}_{w,s} = \sigma_n^2 \mathbf{I}_{N_r} + \mathbf{H}_s \mathbf{W}_s \mathbf{P}_s \mathbf{W}_s^H \mathbf{H}_s^H + \mathbf{J}_s$.

III. Problem formulation

In this paper we address the problem of minimizing the transmit power in the presence of per-user rate constraints η_i, so that each user gets a portion of the available spectrum. Unlike rate maximization, the problem we study is oriented to a fair resource allocation between users.

Implementing the NOMA paradigm [12], we need to define an optimal strategy for grouping the N users into *clusters* of strong/weak users. The partition of all users into disjoint clusters of two users is called a *clustering*. In order to properly formulate the clustering problem, we introduce the allocation variable $\rho_{m,n} \in \{0, 1\}$, $m \in \mathcal{W}$, $n \in \mathcal{S}$, which is $\rho_{m,n} = 1$, if users $m \in \mathcal{W}$ and $n \in \mathcal{S}$ are clustered together, and $\rho_{m,n} = 0$ otherwise. Note that the number of allocation variables is N_c^2, equal to the number of possible clusters. Since each user can belong at most to a cluster and each cluster must contain a strong and a weak user only, any *feasible* clustering must satisfy the two following constraints

$$\begin{aligned} &\sum_{w \in \mathcal{W}} \rho_{w,s} = 1, \quad \forall s \in \mathcal{S}, \\ &\sum_{s \in \mathcal{S}} \rho_{w,s} = 1, \quad \forall w \in \mathcal{W}, \end{aligned} \tag{9}$$

Accordingly, the number of feasible clustering, i.e, the number of combination of N_c clusters fulfilling (9) is $N_c!$. Let us now denote by $\mathbf{P}$, $\mathbf{W}$ and $\boldsymbol{\rho}$ the vectors and matrices stacking all the optimization variables, i.e., the transmitting powers, the beamforming matrices and the cluster allocation variables, respectively. We can formulate the power minimization problem as

$$\min_{\mathbf{P}, \mathbf{W}, \boldsymbol{\rho}} \sum_{w \in \mathcal{W}} \sum_{s \in \mathcal{S}} \rho_{w,s} \left(\mathrm{tr}\{\mathbf{W}_w \mathbf{P}_w \mathbf{W}_w^H\} + \mathrm{tr}\{\mathbf{W}_s \mathbf{P}_s \mathbf{W}_s^H\} \right) \tag{10}$$

subject to

$$\begin{aligned} &R_i \geq \eta_i, \quad \forall i \in \{\mathcal{W}, \mathcal{S}\} \\ &\mathbf{P}_i \succeq 0, \quad \forall i \in \{\mathcal{W}, \mathcal{S}\} \\ &\rho_{w,s} \in \{0, 1\}, \quad \forall w \in \mathcal{W}, \forall s \in \mathcal{S} \\ &\sum_{w \in \mathcal{W}} \rho_{w,s} = 1, \quad \forall s \in \mathcal{S}, \\ &\sum_{s \in \mathcal{S}} \rho_{w,s} = 1, \quad \forall w \in \mathcal{W}, \\ &R_w \leq R_{w,s}, \quad \forall w \in \mathcal{W}, \forall s \in \mathcal{S}, \, \rho_{w,s} = 1, \end{aligned}$$

Problem (10) is a mixed-integer non-convex problem, and its global optimum solution is unknown. In order to obtain a solution of this problem, we follow a sub-optimal approach and we split (10) into three different steps.

1) For each of the possible N_c^2 clusters of users, we adopt block diagonalization beamforming so that the spatial precoder $\mathbf{D}$ is employed to nullify the interference between clusters.
2) Having set to zero the inter-cluster interference, we can now compute for each cluster the *optimal* beamformer $\mathbf{B}$ and optimally distribute the power on the various streams so that the rate constraints are met.

3) We select the feasible clustering strategy that yields the minimum transmit power.

IV. Beamforming design

Implementing the beamforming strategy described in the previous section, determines that the spatial precoder for user i is computed as the cascade of two spatial filters, i.e. $\mathbf{W}_i = \mathbf{D}_i\mathbf{B}_i$.

A. Block Diagonalization

The inter cluster interference can be set to zero by employing block diagonalization (BD) beamforming [13] to separate the signals of the various clusters. Let us consider a generic cluster $\mathcal{C} = \{w, s\}$ and introduce $\mathbf{H}_{-(w,s)} \in \mathbb{C}^{d_{NOMA}\times N_t}$ as the matrix obtained by stacking the channel matrices of all $N-2$ users in the system different from w and s, i.e.,

$$\begin{aligned}\mathbf{H}_{-(w,s)} = [&\mathbf{H}_1^H, \ldots, \mathbf{H}_{w-1}^H, \mathbf{H}_{w+1}^H, \ldots, \mathbf{H}_{N_C}^H \\ &\mathbf{H}_{N_C+1}^H, \ldots, \mathbf{H}_{s-1}^H, \mathbf{H}_{s+1}^H, \ldots, \mathbf{H}_N^H]^H,\end{aligned} \tag{11}$$

where $d_{NOMA} = (N-2)N_r$. The singular value decomposition (SVD) of $\mathbf{H}_{-(w,s)}$ is

$$\mathbf{H}_{-(w,s)} = \mathbf{U}_{-(w,s)}\mathbf{\Sigma}_{-(w,s)}[\mathbf{V}_{-(w,s)}^{(1)}, \mathbf{V}_{-(w,s)}^{(0)}]^H. \tag{12}$$

where the columns of $\mathbf{V}_{-(w,s)}^{(0)} \in \mathbb{C}^{N_t\times(N_t-d_{NOMA})}$ span the null space vector of $\mathbf{H}_{-(w,s)}$, so that by employing the spatial filter

$$\mathbf{D}_{w,s} = \mathbf{V}_{-(w,s)}^{(0)}, \tag{13}$$

the inter-cluster interference generated by w and s is completely eliminated.

B. Strong User's Beamforming

Let us focus on the strong user s and assume that condition (7) is fulfilled, i.e., the interference of weak user can be canceled. The cases where (7) is not satisfied is treated separately in Section IV-E. Because of BD precoding, the inter cluster interference is zero and the received signal is

$$\begin{aligned}\mathbf{r}_s &= \mathbf{H}_s\mathbf{D}_{w,s}\mathbf{B}_s\mathbf{x}_s + \underbrace{\mathbf{H}_s\mathbf{D}_{w,s}\mathbf{B}_w\mathbf{x}_w}_{\text{canceled by SIC}} + \mathbf{n}_s \\ &= \bar{\mathbf{H}}_s\mathbf{B}_s\mathbf{x}_s + \mathbf{n}_s,\end{aligned} \tag{14}$$

where $\bar{\mathbf{H}}_s = \mathbf{H}_s\mathbf{D}_{w,s}$. The precoding and combining matrices are obtained by the SVD of the $\bar{\mathbf{H}}_s = \bar{\mathbf{U}}_s\bar{\mathbf{\Sigma}}_s\bar{\mathbf{V}}_s^H$, by setting

$$\begin{aligned}\mathbf{B}_s &= \bar{\mathbf{V}}_s \in \mathbb{C}^{(N_t-d_{NOMA})\times N_r}, \\ \mathbf{C}_s &= \bar{\mathbf{U}}_s \in \mathbb{C}^{N_r\times N_r}.\end{aligned} \tag{15}$$

The combining operation at the strong receiver yields:

$$\mathbf{z}_s = \mathbf{C}_s\mathbf{r}_s = \bar{\mathbf{\Sigma}}_s\mathbf{x}_s + \bar{\mathbf{n}}_s, \tag{16}$$

where the noise vector $\bar{\mathbf{n}}_s = \mathbf{C}_s\mathbf{n}_s$ has the same statistics of $\mathbf{n}_s$. The achievable rate of strong user s can be now expressed as

$$R_s = \sum_{j=1}^{N_r}\log_2\left(1 + p_{j,s}\Lambda_{j,s}\right) \tag{17}$$

where $\Lambda_{j,s} = \sigma_{j,s}^2/\sigma_n^2$, where $\sigma_{j,s}$ are the diagonal elements of $\bar{\mathbf{\Sigma}}_s$.

C. Weak User's Beamforming

Given the above, the signal at weak user receiver is

$$\mathbf{r}_w = \bar{\mathbf{H}}_w\mathbf{B}_w\mathbf{x}_w + \bar{\mathbf{H}}_w\mathbf{B}_s\mathbf{x}_s + \mathbf{n}_w, \tag{18}$$

where $\bar{\mathbf{H}}_w = \mathbf{H}_w\mathbf{D}_{w,s}$. The optimal beamforming matrices for the weak users is obtained by exploiting the knowledge of the covariance matrix $\mathbf{N}_w$ of the interference plus noise vector [14]. At the weak receiver the correlation matrix $\mathbf{N}_w$ in (5) becomes

$$\mathbf{N}_w = \bar{\mathbf{H}}_w\mathbf{B}_s\mathbf{P}_s\mathbf{B}_s^H\bar{\mathbf{H}}_w^H + \sigma_n^2\mathbf{I}_{N_r}. \tag{19}$$

The noise at the weaker receiver can be whitened by the following operation:

$$\mathbf{N}_w^{-\frac{1}{2}}\mathbf{r}_w = \mathbf{N}_w^{-\frac{1}{2}}\bar{\mathbf{H}}_w\mathbf{B}_w\mathbf{x}_w + \tilde{\mathbf{n}}_w, \tag{20}$$

where $\tilde{\mathbf{n}}_w = \mathbf{N}_w^{-\frac{1}{2}}(\bar{\mathbf{H}}_w\mathbf{B}_s\mathbf{x}_s + \mathbf{n}_w) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_r})$. The optimal transmit and receive filter are then obtained by computing the SVD of the filtered equivalent weak channel

$$\mathbf{N}_w^{-\frac{1}{2}}\bar{\mathbf{H}}_w = \bar{\mathbf{U}}_w\bar{\mathbf{\Sigma}}_w\bar{\mathbf{V}}_w^H. \tag{21}$$

and setting

$$\begin{aligned}\mathbf{B}_w &= \bar{\mathbf{V}}_w \in \mathbb{C}^{(N_t-d_{NOMA})\times N_r}, \\ \mathbf{C}_w &= \mathbf{N}_w^{-\frac{1}{2}}\bar{\mathbf{U}}_w \in \mathbb{C}^{N_r\times N_r}.\end{aligned} \tag{22}$$

The combining operation at the receiver yields:

$$\mathbf{z}_w = \mathbf{C}_w^H\mathbf{r}_w = \bar{\mathbf{\Sigma}}_w\mathbf{x}_s + \bar{\mathbf{n}}_w, \tag{23}$$

where $\bar{\mathbf{n}}_w = \bar{\mathbf{U}}_w^H\tilde{\mathbf{n}}_w \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_r})$. Hence, the achievable rate of weak user w is

$$R_w = \sum_{j=1}^{N_r}\log_2\left(1 + p_{j,w}\Lambda_{j,w}\right), \tag{24}$$

where $\Lambda_{j,w} = \sigma_{j,w}^2/\sigma_n^2$ and $\sigma_{j,w}$ are the diagonal elements of $\bar{\mathbf{\Sigma}}_w$.

D. Power allocation algorithm

From (17) and (24), one can formulate the minimum power problem subject to rate constraints for a generic user $k \in \{w, s\}$ belonging to the cluster as

$$\begin{aligned}&\min \mathrm{tr}\{\mathbf{P}_k\} \\ &\text{subject to } \sum_{j=1}^{N_r}\log_2\big(1 + p_{j,k}\Lambda_{j,k}\big) \geq \eta_k, \\ &\qquad p_{j,k} \geq 0, \quad j = 1, \ldots, N_r.\end{aligned} \tag{25}$$

Problem (25) is convex and differentiable and it can be efficiently solved in the Lagrangian dual domain. The minimum is obtained by the well known water-filling solution [15]

$$p_{j,k}^{*}=\left(\frac{\mu_k}{\ln 2}-\frac{1}{\Lambda_{j,k}}\right)^{+},\quad \forall j,k. \tag{26}$$

where μ_k are chosen so that the rate constraints in (25) for user k is met.

E. NOMA constraint

In the case constraint (7) is not met in a specific cluster, we adopt for that cluster a conventional OMA scheme, where the two users in the cluster are orthogonalized by means of BD. In this case, to implement BD the number of columns of the effective channel $\bar{\mathbf{H}}$ matrix is reduced by a factor N_r and the users will experience a smaller degree of spatial diversity. This situation typically occurs for those clusters where the strong and the weak users are characterized by similar channel quality.

V. Cluster selection

In order to formulate the optimal clustering scheme for the problem at hand, we denote by $P_{w,s}$ the total transmitting power of cluster $\{w,s\}$, i.e.:

$$P_{w,s}=\sum_j p_{j,w}+\sum_j p_{j,s}, \tag{27}$$

and the optimal clustering problem can be formulated as

$$\begin{aligned}\min_{\boldsymbol{\rho}}\ & P_{tot}(\boldsymbol{\rho})=\sum_{w\in\mathcal{W}}\sum_{s\in\mathcal{S}}\rho_{w,s}P_{w,s},\\ \text{subject to }& \rho_{w,s}\in\{0,1\},\quad \forall w\in\mathcal{W},\forall s\in\mathcal{S},\\ & \sum_{w\in\mathcal{W}}\rho_{w,s}=1,\quad \forall s\in\mathcal{S},\\ & \sum_{s\in\mathcal{S}}\rho_{w,s}=1,\quad \forall w\in\mathcal{W}.\end{aligned} \tag{28}$$

Problem (28) is a particular type of MILP problem called *weighted bipartite matching* (WBM). WBM problems have the important property that the coefficient matrix is totally *unimodular* [16], which guarantees that the *optimal* solution of the linear problem is *integer*, regardless of the solver employed. In particular WBM problems can be solved with very fast algorithms [17].

The quasi-optimal clustering (QOC) algorithm is summarized in Algorithm 1.

VI. Numerical Results

We consider a single cell scenario with radius $R=100$ m. The exponential path loss is $\gamma=4$. The central carrier frequency is $f_0=2$ GHz. The power of the additive Gaussian noise is $\sigma_n^2=-125$ dBm. In all figures, we plot the performance of the proposed NOMA scheme together with the benchmark OMA scheme based on BD proposed in [13], referred to as OMA in the following.

Algorithm 1: QOC

```
Initialize: sort in increasing order the user channel gains;
Set W = {1,...,N_C}, S = {N_C+1,...,N};
for w ∈ W do
    for s ∈ S do
        Compute D_{w,s} according to section IV-A;
        Compute B_s and C_s according to section IV-B;
        Evaluate power allocation for P_s solving
          problem (25);
        Compute B_w and C_w according to section IV-C;
        Evaluate power allocation for P_w solving
          problem (25);
        if R_w > R_{w,s} then
            Perform BD of s and w;
        P_{w,s} = tr{P_w} + tr{P_s}
Solve problem (28) for optimal clustering;
```

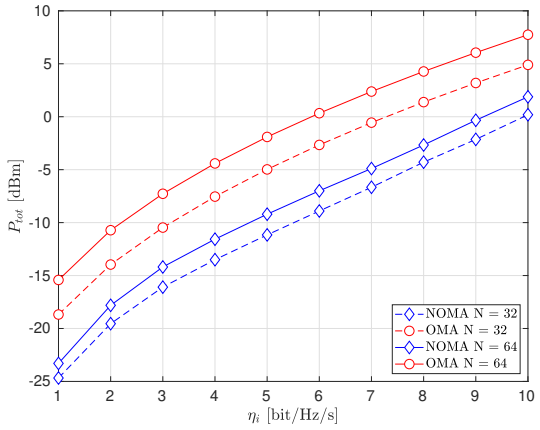

Fig. 1: P_{tot} as a function of η_i for the NOMA proposed algorithm and for the OMA scheme.

Figs. 1 and 2 show a comparison between NOMA and OMA in terms of the overall power spent P_{tot}. For each user, we have considered a number of receiving antenna $N_r=2$, while the number of antennas at the BS is set to $N_t=N_rN=2N$.

Fig. 1, presents the total required power as a function of the target rates for $N=32$ and $N=64$, where the same target rate is assumed for all users. NOMA clearly outperforms OMA, thus demonstrating the effectiveness of the clustering algorithm in exploiting the NOMA potential. As expected, in all cases we observe an exponential increase (linear in log-scale) of the required power with the increase of the target rate.

Fig. 2 shows the total required power as a function of the number of users in the cell, for a fixed target throughput $\eta_i=6$. The figure shows the results for NOMA with $N_r=2$ and OMA with $N_r=2$ and $N_r=4$. For any number of users N, NOMA outperforms the OMA scheme implemented with the same number of receive antennas and, as the number of users increases, the performance of our proposed scheme tends to the performance of OMA with a double number of receive antennas, i.e., $N_r=4$. Indeed, in the presence of a high number of users, the clustering algorithm exploits the multi-user diversity of

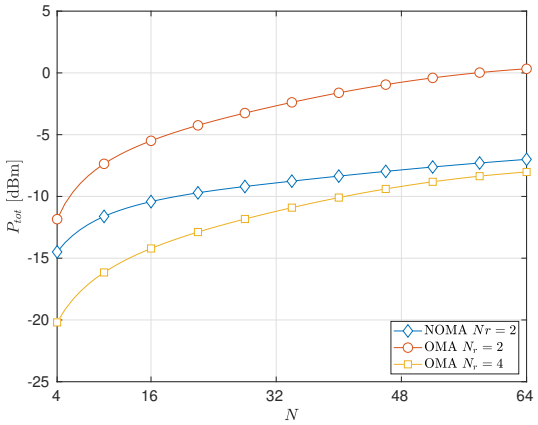

Fig. 2: P_{tot} as a function of the number of users, for $\eta_i = 6$.

the system and is more likely to find a couple of users which can efficiently share the same channel with a low power consumption.

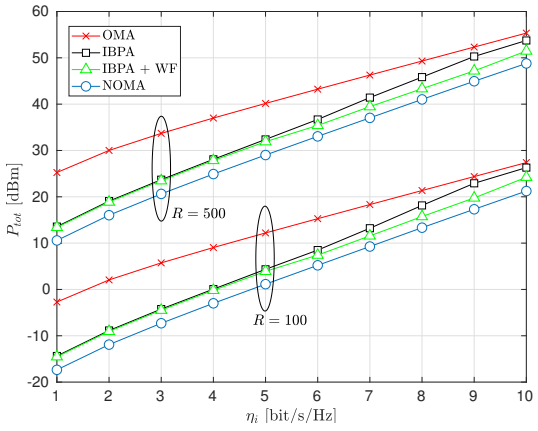

Fig. 3: P_{tot} as a function of η_i, for the NOMA proposed algorithm, IBPA + WF, IBPA [6] and OMA, for $N = 32$ and for cell radius $R = 100$ m and $R = 500$ m.

Finally, Fig. 3 provides the comparison of the proposed scheme with two other schemes: the current state-of-the-art algorithm [6] for NOMA beamforming and clustering, labelled as *inversion based paired algorithm*, 'IBPA', and the scheme obtained by combining IBPA clustering with the beamforming strategy presented here, labelled as 'IBPA + WF'. The results for this second algorithm show the superiority of our clustering algorithm with respect to the heuristic presented in [6]. The results are plotted for $R = 100$ m and $R = 500$ m cell radius with no appreciable difference in the algorithms behaviour. In order to compare the three algorithms, we consider a MISO transmission paradigm, i.e. $N_r = 1$, and we collect the results of the total power spent P_{tot} for the three algorithms in the same simulated scenarios, for $N = 32$. The different number of receive antenna is the reason for the different results in Fig. 1 and Fig. 3. Due to the optimal clustering selection, the proposed algorithm largely outperforms both reference algorithms.

VII. Conclusions

In this paper, the joint clustering, beamforming and power allocation problem for the downlink of multi-user MIMO-NOMA cellular systems has been investigated. In particular, we have proposed a scheme where *strong* and *weak* users are paired together to form NOMA clusters and the various clusters are separated by BD beamforming. The optimal clustering allocation is selected as the solution of a MILP problem. Simulations results show that the proposed MIMO-NOMA scheme outperforms both classical MIMO-OMA and state-of-the-art MIMO-NOMA algorithms.

References

[1] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. J. Bhargava, "A survey on Non-Orthogonal Multiple Access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, Oct. 2017.

[2] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-Orthogonal Multiple Access (noma) for cellular future radio access," in *IEEE VTC Spring*, June 2013.

[3] 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 *IEEE MILCOM*, Nov. 2013.

[4] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, Mar. 2015.

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

[6] Z. Chen, Z. Ding, and X. Dai, "Beamforming for combating inter-cluster and intra-cluster interference in hybrid NOMA systems," *IEEE Access*, vol. 4, 2016.

[7] Z. Chen and X. Dai, "MED precoding for multiuser MIMO-NOMA downlink transmission," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5501–5505, Jun. 2017.

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

[9] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct 2017.

[10] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, 2018.

[11] M. Moretti and A. Perez-Neira, "Efficient margin adaptive scheduling for MIMO-OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 278 –287, Jan. 2013.

[12] 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 Commun.*, vol. 25, no. 2, Apr. 2018.

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

[14] F. R. Farrokhi, G. J. Foschini, A. Lozano, and R. A. Valenzuela, "Link-optimal space-time processing with multiple transmit and receive antennas," *IEEE Commun. Lett.*, vol. 5, Mar. 2001.

[15] D. P. Palomar and J. R. Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 686–695, Feb 2005.

[16] M. Moretti, A. Abrardo, and M. Belleschi, "On the convergence and optimality of reweighted message passing for channel assignment problems," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1428–1432, Nov 2014.

[17] B. Huang and T. Jebara, "Loopy belief propagation for bipartite maximum weight b-matching," in *Artificial Intelligence and Statistics*, 2007, pp. 195–202.