Assessment of Non-orthogonal Multiple Access for 5G systems

Ricardo Alberto
ricardo.alberto@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2016

Abstract

Non-orthogonal multiple access (NOMA) recently became a prominent candidate for next generation wireless systems when dense networks are envisaged due to its ability to further enhance the overall spectral efficiency. In NOMA the signals are separated at the receivers taking in consideration their different power levels. NOMA as been much studied in the literature from an information-theoretic perspective in terms on capacity and outage probability. This dissertation looks at two system configurations for downlink multiuser MIMO systems and compares both via traditional numerical simulation. In the first system no beamforming precoding is made at the base-station. Users are clustered into MIMO beams and NOMA is used within each cluster, where successive interference cancellation (SIC) is used at each terminal to separate the NOMA signals. Inter-cluster interference is dealt with by linear filtering at the terminals, as in MIMO multiuser systems. Some of the practical limitations of SIC are highlighted and precise examples of that are given. In the second system, a Karhunen-Loève decomposition is applied assuming that the users are grouped into clusters that share the same correlation matrix that varies slowly along, and for that reason precoding is made possible along with the use of massive MIMO at the base-station. While the first system requires a larger number of antennas at the terminals than the ones of the base-station, in the precoded system, that limitations is waived and massive MIMO is used.
Keywords: Non-orthogonal multiple access, successive interference cancellation, linear inter-cluster

1. Introduction

One of the goals for the the Fifth Generation of mobile communications is to increase by a factor of one thousand the system's capacity in comparison with the current standard LTE-Advanced [2]. LTE's capacity has mostly grown due to carrier aggregation at the expense of higher system's complexity, and this is to be avoided in 5G. On a related front, increasing the number N of OFDM subcarriers is limited by the fact that the amplification of basic small errors (e.g., frequency offsets and imperfect synchronization) is not independent of the number of subcarriers and grows according to $\log(N)$ [13].

The key technologies to be part of physical layer in the upcoming system are massive MIMO [3], in-band full-duplex [8], and non-orthogonal multiple access (NOMA), separating users in the power domain, which is the focus of this dissertation. In fact, the most simple version of NOMA has already been included in LTE-A release 13 to support only two users sharing the same frequency and time resources, named multi-user superposition transmission (MUST) [6].

An information-theoretic analysis of NOMA in wireless communications shows that it's capacity surpasses the ones of conventional orthogonal multiple access schemes [9]. However, the capacity of each user is largely dependant on the its power allocation coefficient, and therefore dependent on the individual channel gains. Moreover, the water-filling although being optimal from the system's total throughput, may lead to highly unfair allocation of the system's capacity to users with bad channels. In [11] the authors tackled this problem imposing restrictions in the power allocation algorithm.

This dissertation considers two system model that have been proposed by Ding et al. in [5], which has also been recently analyzed in [9] and the system proposed by Ding et al. in [7]. In both system models the users are grouped in clusters, with the signals for each cluster being separated at the terminals in the spatial domain using MIMO processing (with the restriction that the number of antennas at the terminals is least equal to the number of antennas at the transmitter), and with intra-cluster multiple access applying NOMA with SIC detection. The results presented in this dissertation are obtained not from an information-theoretic point of view, as it has been traditional in the NOMA literature, but rather via numerical system simulation. To the best of our knowledge, comprehensive simulation results are presented for the first time for

multiuser NOMA systems with more than just two users.

2. System model without precoding

Consider a downlink multi-user MIMO system with M antennas at the BS and $N \geq M$ antennas at each user similar to the one in [5]. To apply the NOMA concept, users will be grouped in M clusters of K users each. The BS transmits the signal x

$$\mathbf{x} = \mathbf{P}\tilde{\mathbf{s}}, \tag{1}$$

where $\mathbf{P}$ is the $M \times M$ precoding matrix, which will be chosen to be the identity matrix, given that in this chapter no precoding will be considered at the BS and, therefore, the users do not have to feedback their channel state information (CSI) to the BS. The symbols to be transmitted from the BS can be represented in a matrix $\mathbf{S} \in \mathbb{C}^{M \times K}$.

$$\mathbf{S} = \begin{bmatrix} s_{1,1}, \cdots, s_{1,K} \\ \vdots \\ s_{M,1}, \cdots, s_{M,K} \end{bmatrix}. \tag{2}$$

The vector $\tilde{\mathbf{s}} \in \mathbb{C}^{M \times 1}$, which is effectively the vector that is transmitted from the BS to the users, is:

$$\tilde{\mathbf{s}} = \begin{bmatrix} \alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K}s_{1,K} \\ \vdots \\ \alpha_{M,1}s_{M,1} + \cdots + \alpha_{M,K}s_{M,K} \end{bmatrix}, \tag{3}$$

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 $\alpha_{m,k}^2 \in [0,1]$ 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) (broadcast channel) where each cluster plays the role of an aggregated "super-user", and later the information to each one of the users within each cluster is distilled from the symbol that was sent to the cluster. The set of power coefficients is selected having in consideration the following power constraint [5]:

$$\sum_{k=1}^{K} \alpha_{m,k}^2 = 1. \tag{4}$$

This condition guarantees that, for example, the power from the superimposed signal, assuming that all symbols are BPSK, is the same as just one BPSK symbol. Note that NOMA will be applied in each cluster, hence, in the worst case, a user 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:

$$\mathbf{y}_{1,k} = \mathbf{H}_{1,k}\tilde{\mathbf{s}} + \mathbf{n}_{1,k}, \tag{5}$$

where $\mathbf{H}_{1,k} \in \mathbb{C}^{N \times M}$ is the Rayleigh fading matrix from the BS to the k-th user in the first cluster and $\mathbf{n}_{1,k} \sim \mathcal{CN}(0, \sigma_n^2) \in \mathbb{C}^{1 \times K}$ is the unit power additive white Gaussian noise vector for the first cluster. Note that this noise is generated by a random variable taken from an independent circularly symmetric complex Gaussian distribution with zero average and variance σ_n^2. The matrix $\mathbf{H}_{1,1} \in \mathbb{C}^{N \times M}$ is the channel matrix for the first user in the first cluster:

$$\mathbf{H}_{1,1} = \begin{bmatrix} \mathbf{h}_{1,1}, \cdots, \mathbf{h}_{1,M} \\ \vdots \\ \mathbf{h}_{N,1}, \cdots, \mathbf{h}_{N,M} \end{bmatrix}. \tag{6}$$

In each user, the signal $\mathbf{H}_{1,k}\tilde{\mathbf{s}} + \mathbf{n}_{1,k}$ will be multiplied by the detection vector, leading to:

$$\mathbf{v}_{1,k}^H \mathbf{y}_{1,k} = \mathbf{v}_{1,k}{}^H \mathbf{H}_{1,k}\tilde{\mathbf{s}} + \mathbf{V}_{1,k}^H \mathbf{n}_{1,k}, \tag{7}$$

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

$$\begin{aligned} &\mathbf{v}_{1,k}^H \mathbf{y}_{1,k} = \\ &= \mathbf{v}_{1,k}^H \mathbf{H}_{1,k}(\alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K}s_{1,K}) + \cdots + \\ &\sum_{l=2}^{M} \mathbf{v}_{1,k}^H \mathbf{H}_{1,k}\tilde{\mathbf{s}}_{\setminus 1} + \mathbf{v}_{1,k}^H \mathbf{n}_{1,k}, \end{aligned} \tag{8}$$

where $\tilde{\mathbf{s}}_{\setminus 1} \in \mathbb{C}^{M-1 \times 1}$ denotes the vector $\tilde{\mathbf{s}}$ without the contribution of the signals from the first cluster. The aim is to eliminate the inter cluster interference $\sum_{m=2}^{M} \mathbf{v}_{1,k}^H \mathbf{H}_{1,k}\tilde{\mathbf{s}}_{\setminus 1}$ in the first cluster, such that NOMA detection can be performed on the remaining signal. In short, the problem amounts to having:

$$\mathbf{v}_{m,k}^H \mathbf{H}_{i,k} = 0, \tag{9}$$

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

$$\mathbf{v}_{m,k}^H \underbrace{\left[\mathbf{h}_{1,ik} \cdots \mathbf{h}_{m-1,ik}\, \mathbf{h}_{m+1,ik} \cdots \mathbf{h}_{M,ik}\right]}_{\tilde{\mathbf{H}}_{i,k}} = 0, \tag{10}$$

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 (10) 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 $M = N$:

$$\tilde{\mathbf{H}}_{i,k} = \begin{bmatrix} U_{1,1} & U_{1,2} & \dots & U_{1,N-1} & U_{1,N} \\ \vdots & & & & \vdots \\ U_{N,1} & U_{N,2} & \dots & U_{N,N-1} & \underbrace{U_{N,N}}_{\tilde{\mathbf{U}}_{i,k}} \end{bmatrix} \times \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ \dots & & & & \\ 0 & 0 & \dots & \lambda_{min(M,N)} & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \mathbf{v^T} \tag{11}$$

Note that the matrix with the eigenvalues of $\tilde{\mathbf{H}}_{i,k}$ has a row of zeros at the bottom. This happens because, even if $M = N$, after removing a column from $\mathbf{H}_{m,k}$ to create $\tilde{\mathbf{H}}_{i,k}$, the matrix becomes tall and, after the SVD decomposition, it always leads to an eigenvalue matrix that has at least one row of zeros on the bottom. In general, there will be $(M - N)+1$ rows of zeros in the eigenvalues matrix. Note that the column highlighted in (11) (which in the general case is a matrix), $\tilde{\mathbf{U}}_{i,k} \in \mathbb{C}^{N\times(N-M-1)}$, does not contribute at all to $\tilde{\mathbf{H}}_{i,k}$ since it is multiplied by the row of zeros, thus, it spans in fact a space orthogonal to $\tilde{\mathbf{H}}_{i,k}$. Now, one could use $\mathbf{v}_{m,k} = \frac{\tilde{\mathbf{U}}_{i,i}}{\|\tilde{\mathbf{U}}_{i,k}\|}$ as the detection matrix, but based on the maximum ratio combining (MRC) approach, it can project the $\mathbf{h}_{m,ik}$ column onto the orthogonal space using a projection matrix $\mathbf{P}_U = \tilde{\mathbf{U}}_{i,k}\tilde{\mathbf{U}}^H_{i,k}$, choosing:

$$\mathbf{v}_{m,k} = \tilde{\mathbf{U}}_{i,k}\frac{\tilde{\mathbf{U}}^H_{i,k}\mathbf{h}_{m,ik}}{\|\tilde{\mathbf{U}}^H_{m,i,k}h_{m,ik}\|}. \tag{12}$$

Note that this MRC is done per cluster (maximising the SNR at one single antenna) and the inter-cluster interference is eliminated because (12) fulfils the requirement of (9). This matrix is the reason why $N \geq M$ antennas are needed at each user, otherwise the $\tilde{\mathbf{H}}_{i,k}$ matrix is fat instead of being tall and thus there is no orthogonal space spanned by the columns of $\mathbf{U}_{m,k}$, the left matrix in (11). Without loss of generality, continuing to focus on the first cluster, the channel gains of the users in the first cluster should be be ordered in this manner:

$$\|\mathbf{v}^H_{1,1}\mathbf{H}_{1,1}\|^2 \geq \dots \geq \|\mathbf{v}^H_{1,k}\mathbf{H}_{1,k}\|^2, \tag{13}$$

which is equivalent to sorting the power allocation coefficients as:

$$\alpha_{1,1} \leq \dots \leq \alpha_{1,k}. \tag{14}$$

It should note that the norms in (13) are taken from vectors where all but one elements are zero. Note that this ordering happens within each cluster, and all clusters are statistically identical. The detection process to be applied will be ZF:

$$\begin{aligned} &\tilde{\mathbf{y}}_{1,k} = (\mathbf{v}^H_{1,k}\mathbf{H}_{1,k})^{-1}\mathbf{v}^H_{1,k}\mathbf{H}_{1,k}(\alpha_{1,1}s_{1,1} + \dots + \\ &\alpha_{1,K}s_{1,K}) + (\mathbf{v}^H_{1,k}\mathbf{H}_{1,k})^{-1}\mathbf{v}^H_{1,k}\mathbf{n}_{1,k} = \\ &= (\alpha_{1,1}s_{1,1} + \dots + \alpha_{1,K}) + (\mathbf{H}_{1,k})^{-1}\mathbf{n}_{1,k} \end{aligned} \tag{15}$$

3. The Limitations of successive interference cancellation Detection

In systems where only two users are multiplexed in the power domain, which is the case analysed in almost all the NOMA literature, SIC performs quite well. Unfortunately, as it will be seen in section 4, with more than two users, SIC rapidly starts malfunctioning. This section looks at this phenomenon with some examples.

One starts by taking the example of a case with three users transmitting the bits $s_{m,1} = +1, s_{m,2} = +1, s_{m,3} = -1$, with $\alpha_{m,1} = \sqrt{\frac{1}{6}}, \alpha_{m,2} = \sqrt{\frac{1}{3}}$ and $\alpha_{m,3} = \sqrt{\frac{1}{2}}$ and no noise is added. In the first iteration the receiver decides for a positive $s_{m,1}$, given that $\sqrt{\frac{1}{6}} + \sqrt{\frac{1}{3}} - \sqrt{\frac{1}{2}} > 0$, even though the bit with the largest power is -1. In such situation the second bit to be decoded is guaranteed to also be wrongly detected due to error propagation, i.e., subtracting $\sqrt{\frac{1}{2}}$ from $\tilde{s}_1$ leads to a negative value: $\sqrt{\frac{1}{6}} + \sqrt{\frac{1}{3}} - \sqrt{\frac{1}{2}} - \sqrt{\frac{1}{2}} < 0$, which causes the second bit to be decoded as a -1, when a $+1$ had been transmitted. This type of events leads to a disastrous performance of the SIC receiver with more than two users.

When using higher modulation schemes such as 16-QAM or 4-pulse amplitude modulation (PAM), this type of error propagation, even in the absence of noise, happens even more frequently. Consider one further example, now with two users using 16 QAM: take $s_{m,1} = 3 - j$ and $s_{m,2} = -1 - j$ and $\alpha_{m,1} = \sqrt{\frac{1}{4}}$ and $\alpha_{m,2} = \sqrt{\frac{3}{4}}$, also without noise. Disregarding the complex part, the real part of $\tilde{s}_1$ will be positive, as $3 \times \sqrt{\frac{1}{4}} > \sqrt{\frac{3}{4}}$. This means that the real part of the first symbol decoded is positive, in this case it will be $+1$, since $3 \times \sqrt{\frac{1}{4}} - \sqrt{\frac{3}{4}} \approx 0.63$, however, the real part of the symbol transmitted is -1, so the symbol will be misinterpreted.

It is important to note that this problem becomes significantly worse when QAM is used instead of BPSK because with QAM symbols it is not sufficient that the highest alpha is greater than the sum of all the remaining (less powerful) alphas, as in the multi-user problem that was explained using BPSK. For these higher-order modulations, a distribution for the power allocation coefficients that leads to correctly decodable symbol sets allowing

more than two users is not known. It is important to understand the relation that the power allocation coefficients need to satisfy in order to this particular simulation to be decodable, namely, how much smaller the first alpha needs to be in respect to the second and so forth.

A simulation for two users using two different QAM constellations will be later shown. In these cases, the decision region between two points has Euclidean distance $d = 2$ in standard QAM modulations. As it is well-known, maximum likelihood (ML) decisions will lead to errors when deciding for points where the real or quadrature components deviated by more than $\frac{d}{2}$ from the correct constellation point. In NOMA systems this distance will be reduced by the factor α_1. Consider for example the outer symbol $3+3i$ of a standard 16-QAM constellation, whose real and imaginary components of $\tilde{s}_1$ after applying the power coefficients become $3\ \alpha_1$. Hence, one needs to have $\alpha_1 < \frac{1}{3}\alpha_2$. Later, when simulating a multi-user scenario with five users, BPSK will be the only modulation used by each user, precisely due to these limitations for higher modulation schemes.

For SIC detection to be possible with BPSK, the following constrain needs to be imposed:

$$\alpha_{m,k} > \sum_{k=1}^{K-1} \alpha_{m,k}, \tag{16}$$

for users $1 \leq k \leq K$ in the i-th cluster. It should be noted though that this relation disregards fairness. To minimise this problem, it will apply a simple rule where:

$$\alpha_{m,k-1} = 0.5 \times \alpha_{m,k}, \tag{17}$$

and since $\sum_{k=1}^{N}(1/2)^k$ (which is the geometric progression of ratio 1/2 deprived from its first term) tends to 1 as N tends to infinity, the restriction (16) will be fulfilled and the lower users in the decoding order will be allocated the maximum possible power. 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 [10] (and thus not taking fairness into equation).

4. Numerical Results

This section presents the simulation results for a number of multi-user NOMA scenarios. The two-user case is assessed with BPSK and with different QAM modulations and the case of five users using BPSK is assessed. The results are depicted in Figures 1, 2 and 3.

One interesting result that emerges is that the performance results show in some cases two distinct regimes, depending on the SNR. One could naively think that the user with an highest power allocation coefficient would experience a lower symbol error rate (SER) than the other user. In fact, this

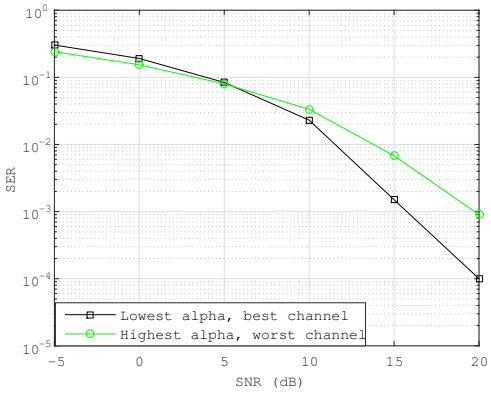

Figure 1: SER curves for two users per cluster with BPSK modulation. M=2, N=3 and K=2.

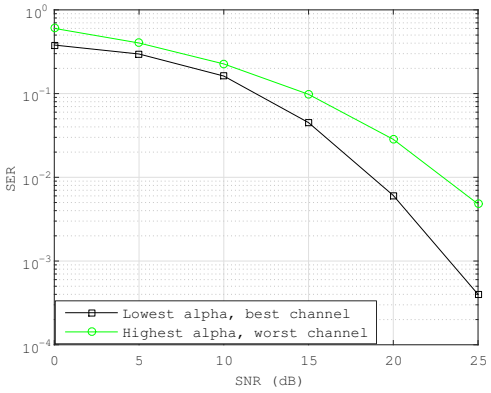

Figure 2: SER curves for two users per cluster with BPSK modulation in the first user and 16-QAM modulation in the second user. M=2, N=3 and K=2.

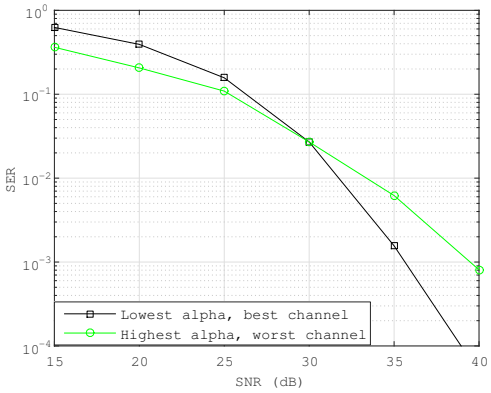

Figure 3: SER curves for two users with 16-QAM modulation in the first user and 64-QAM modulation in the second user. M=2, N=3 and K=2.

is only true in the low SNR regime and not even in all cases. Consider that user 1 is the one with the lowest power allocation coefficient and user 2 the highest one. In the high SNR regime, both receivers experience low noise, nevertheless, user 1 has

a channel with a larger gain than the one that user 2 experiences. Moreover, user 2 also has to deal with an increased noise level due to the superimposed interference signal intended to user 1, even when the noise tends to zero. The limitation due to this interference explains why user 1, with a better channel and a lower power allocation, holds a better SER at high SNR.

At low SNR, because user 1 has to firstly decode the symbol intended to user 2, the errors will propagate to the detection of its own symbol, degrading its SER while user 2 does not suffer any degradation. This explains why in Figures 1 and 3, user 2 surpasses the SER of user 1 in the low SNR regime.

As expected, when using higher modulation schemes, the performance degrades given that, when maintaining a normalised unit power, they hold a shorter Euclidean distance between symbols. It can also be noted than when the modulation of user 1 is a simple BPSK and user 2 uses a QAM modulation (see Figure 2), the dual regime does not appear since at low SNR the errors that could propagate and influence the detection of user 1 signal are not meaningful, because when detecting BPSK there are only two detection regions: above or below 0. This does not occur in Figure 1 since both users are using BPSK.

In general, the robustness of the systems is chiefly defined by the relations between the power coefficients. In Figures 1 and 2, $\alpha_1 = \sqrt{\frac{1}{4}}$ and $\alpha_2 = \sqrt{\frac{3}{4}}$ were used in order to compare with the results in Figure 1 in [5]. In Figure 3, it has $\alpha_1 = \sqrt{\frac{1}{17}}$ and $\alpha_2 = \sqrt{\frac{16}{17}}$ which were used to comply with restrictions (4) and (16). Comparing the first simulation with Figure 1 in [5], it sees that the SER results are effectively bounded by the outage probability, as expected. For the five user simulation, the users are ordered as in figure 4. The users that are close to the base station (and thus having a better channel coefficient) being numbered from 1 to 5 and having a lower power allocation coefficient to maintain fairness. In Figure 5, one can observe that the users with higher power allocation coefficients have a better (lower) SER at low SNR and then worse performance at high SNR. The $\alpha_{m,k}$ were obtained by using the set $\{1, 2, 4, 8, 16\}$, normalized by its sum $\sqrt{341}$, in order to comply with equation (4). With six users and similar power allocations coefficients, $\alpha_{m,1}$ becomes too small, and user 1 is much affected in a noise detection, with a $SER > 0.5$ for $SNR = 10$ dB, which was chosen as the limit for which the simulations are run.

Figure 4: A five user MIMO-NOMA system with users sorted according to their channel gain.

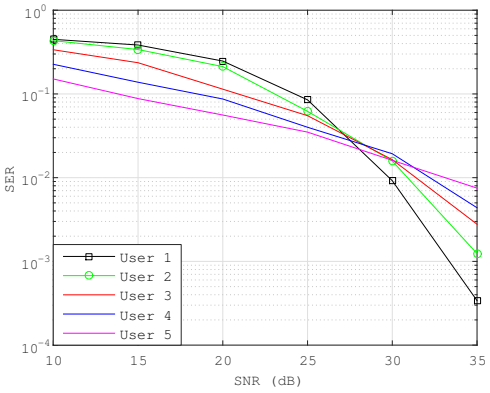

Figure 5: SER curves for five users per cluster with BPSK modulation. M=2, N=2 and K=5. User 5 has the highest power allocation coefficient and the lowest channel gain. User 1 has the lowest power allocation coefficient and the largest channel gain. $(\alpha_{m,1} = 0.0542, \alpha_{m,2} = 0.1083, \alpha_{m,3} = 0.2166, \alpha_{m,4} = 0.4332, \alpha_{m,5} = 0.8664)$.

5. Massive MIMO Model

The previous model could not use massive MIMO at the BS due to the restrictions imposed by (12), where the detection vector required $N \geq M$. However, it is intuitive that when doing the ZF of the inter-cluster interference at the receivers, that condition does not need to hold true. That is the theory Ding et al. have presented in [7]. Consider a scenario similar to the previous one, where one base station with M antennas is communicating with multiple users, each of which with N antennas, but now it will be considered that $M >> N$ with a massive MIMO BS. The users are separated into L clusters with $L \neq M$, and in each cluster there are K different users, with different channel matrices, but all sharing the same spatial correlation matrix, denoted by $\mathbf{R}_l$. Using the Karhunen-Loève decomposition [1, 4], the k-th user in the l-th cluster can have its channel matrix decomposed as:

$$\mathbf{H}_{l,k} = \mathbf{G}_{l,k} \Lambda_l^{\frac{1}{2}} \mathbf{U}_l, \tag{18}$$

where $\mathbf{G}_{l,k} \in \mathbb{C}^{M \times M}$ 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

$$\mathbf{R}_l = \mathbf{U}_l^H \Lambda_l \mathbf{U}_l = \mathbb{E}\{\mathbf{H}_{l,k}^H \mathbf{H}_{l,k}\}, \tag{19}$$

given that a correlation matrix is always symmetric. However, $\mathbf{R}_l$ is only going to have r_l non-zero eigenvalues, where r_l is the rank of $\mathbf{R}_l$. The Λ_l matrix

has the form:

$$\Lambda_l = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & 0 \\ \dots & & & & & \\ 0 & 0 & \dots & \lambda_{M-r_k,M-r_k} & 0 & 0 \\ 0 & 0 & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & 0 & \lambda_{M,M} \end{bmatrix}, \tag{20}$$

and thus can be reduced to a $r_l \times r_l$ matrix, making $\mathbf{G}_{l,k}$ a $M \times r_l$ matrix and $\mathbf{U}_l$ a $r_l \times M$ matrix. The Karhunen-Loève decomposition is useful because, while the CSI-T concerning the fast fading matrix $\mathbf{G}_{l,k}$ is hard to get at the BS, the $\mathbf{R}_l$ matrix represents the channel correlation and thus varies slowly, so it is reasonable to assume that the BS has easier access to $\mathbf{R}_l$. The BS will send the $M \times 1$ NOMA superimposed symbol

$$\mathbf{S} = \sum_{l=1}^{L} \mathbf{P}_l \sum_{k=1}^{K} w_l \alpha_{l,k} s_{l,k}, \tag{21}$$

where $s_{l,k}$ is the modulated symbol to be transmitted to the k-th user in the l-th cluster, $\alpha_{l,k}$ is the power allocation coefficient for the k-th user in the l-th cluster that fulfils the previous model condition of $\sum_{k=1}^{K} \alpha_{l,k}^2 = 1$, $w_l = [0 \cdots 0 \ 1 \ 0 \cdots 0]^T$ is the $\tilde{M}_l \times 1$ precoding vector that has a 1 in the l-th position. The number of effective antennas $\tilde{M}_l = (M - r_l(L-1))$ and, $\mathbf{P}_l$ is the $M \times \tilde{M}_l$ precoding matrix of the l-th cluster that is used to eliminate inter-cluster interference. The k-th user in the l-th cluster will observe the following:

$$\mathbf{y}_{\mathrm{l,k}} = \mathbf{G}_{l,k} \Lambda_l^{\frac{1}{2}} \mathbf{U}_l \sum_{l=1}^{L} \mathbf{P}_l \sum_{k=1}^{K} w_l \alpha_{l,k} s_{l,k} + n_{l,k}, \tag{22}$$

where $n_{l,k}$ is the noise value for the k-th user in the l-th cluster. By looking at equation (22), the precoding matrix $\mathbf{P}_l$ will need to satisfy the following constrain to eliminate inter cluster interference:

$$[\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H \mathbf{U}_{l+1}^H \cdots \mathbf{U}_L^H]^H \mathbf{P}_l = 0. \tag{23}$$

Since $[\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H \mathbf{U}_{l+1}^H \cdots \mathbf{U}_L^H]^H$ is always going to be a fat matrix (and thus has always a defined nullspace), $\mathbf{P}_l$ will simply be chosen as:

$$\mathbf{P}_l = \mathrm{Null}([\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H \mathbf{U}_{l+1}^H \cdots \mathbf{U}_L^H]^H). \tag{24}$$

Using $\mathbf{P}_l$ in (24), (22) can be simplified to:

$$\mathbf{y}_{\mathrm{l,k}} = \mathbf{G}_{l,k} \Lambda_l^{\frac{1}{2}} \mathbf{U}_l \mathbf{P}_l \sum_{k=1}^{K} w_l \alpha_{l,k} s_{l,k} + n_{l,k}. \tag{25}$$

Let us specify (25) for the case of $l = 1$ and $k = 2$ and analyse the signal received by the first user:

$$\mathbf{y}_{1,1} = \mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1} (\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2}) + n_{1,1}. \tag{26}$$

Note that the information from all the users information is being carried by a $\tilde{M}_l \times 1$ vector that has the form:

$$[\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2} \quad 0 \cdots 0]^T, \tag{27}$$

and 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 $N \times \tilde{M}$. Let us call $c_{n,\tilde{m}}$ to the elements of this matrix. Disregarding noise, this can be written as:

$$\begin{bmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,\tilde{M}-1} & c_{1,\tilde{M}} \\ \vdots & & & & \vdots \\ c_{N,1} & c_{N,2} & \dots & c_{N,\tilde{M}-1} & c_{N,\tilde{M}} \end{bmatrix} \times \begin{bmatrix} \alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2} \\ \vdots \\ 0 \end{bmatrix}, \tag{28}$$

so, only the first column of $\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1$ is going to influence the received $N \times 1$ vector $\mathbf{y}_{1,1}$. This leads to an MRC detection of a column vector, i.e., the detection is performed by an "inverse vector" in the following manner:

$$\begin{aligned} &\tilde{\mathbf{y}}_{1,1} = (\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1})^{-1} \times \\ &[\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1} (\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2}) + n_{1,1}] = \\ &(\alpha_{1,1} s_{1,1} + \alpha_{1,2} s_{1,2}) + (\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 \mathbf{P}_1 \mathbf{w_1})^{-1} n_{1,1}. \end{aligned} \tag{29}$$

Comparing figures **??** and **??** with figures 1 and 2, one can see that they are identical. To understand why this happens one needs to look at equations (29) and (15). Noting that $\mathbf{G}_{1,1} \Lambda_1^{\frac{1}{2}} \mathbf{U}_1 = \mathbf{H}_{1,1}$ (Karhunen-Loève decomposition) and that $\|\mathbf{v}_{m,k}\| = \|\mathbf{P}_1\| = 1$, one sees that both equations are equivalent in terms of the ratio between the signal power and noise power.

6. Rate and capacity analysis

The conclusion that NOMA outperforms than current orthogonal schemes in terms of rate has already been given in [12]. However, these results are only obtained for single antenna systems and, in this chapter, taking the system with no precoding and with two users per cluster analysed in chapter 2, the results for MIMO-NOMA will be obtained in order to see if the conclusions can be extended to the MIMO-NOMA case.

As in equation (14) from [5], the SINR for the k-th user in the m-th cluster can be written as:

$$SINR_{m,k} = \frac{\rho \|\mathbf{v}_{m,k}^H \mathbf{H}_{m,k}\|^2 \alpha_{m,k}^2}{\rho \sum_{l=1, l \neq k}^{K} \|\mathbf{v}_{m,k}^H \mathbf{H}_{m,k}\|^2 \alpha_{m,l}^2 + \|\mathbf{v}_{m,k}\|^2}, \tag{30}$$

where $\rho = \frac{\sigma_x^2}{\sigma_n^2}$ is the SNR, with σ_x^2 being the variance of the signal and σ_n^2 being the variance of the

unit power additive white Gaussian noise. Noting that $\|\mathbf{v}_{m,k}\|^2 = 1$ and $\sum_{l=1,l\neq k}^{K} \|\mathbf{v}_{m,k}^H \mathbf{H}_{m,k}\|^2 \alpha_{m,l}^2 = 0$, for $k = 1$, (30) can be written for the case of user 1 as:

$$SINR_{m,1} = \rho \|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2, \tag{31}$$

and, taking into account that $\sum_{l=1,l\neq k}^{K} \|\mathbf{v}_{m,k}^H \mathbf{H}_{m,k}\|^2 \alpha_{m,l}^2 = \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2$ for $k = 2$ it can be further written as:

$$SINR_{m,2} = \frac{\rho \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2}{\rho \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2 + 1}. \tag{32}$$

With these expressions, one can now formulate the equations for the rates of the two NOMA users. The rate for the first user of the m-th cluster, after it decodes and removes the signal from the second user, is bounded by:

$$R_{m,1}^{MIMO-NOMA} \leq \log_2(1 + \rho \|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2), \tag{33}$$

and the achievable rate for the second user in the m-th cluster is bounded by:

$$R_{m,2}^{MIMO-NOMA} \leq \log_2(1 + \frac{\rho \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2}{\rho \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2 + 1}). \tag{34}$$

Now, in order to compare MIMO-NOMA to MIMO-OMA, one now also needs to have expressions for the OMA rates as well. In orthogonal schemes there is a splitting of resources (time or frequency) between users. Let us define β as the fraction of resources allocated to the second user in the m-th cluster, and hence $1 - \beta$ is the fraction allocated for the first user. Further, consider that $\frac{\gamma\rho}{\beta}$, with the power allocation coefficient $0 \leq \gamma \leq 1$, is the SNR of the second user and hence, $\frac{(1-\gamma)\rho}{1-\beta}$ is the SNR for the second user. Now, the rate of the first user is bounded by:

$$R_{m,1}^{MIMO-OMA} \leq (1-\beta) \log_2(1 + \frac{(1-\gamma)\rho \|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2}{1-\beta}), \tag{35}$$

and the rate of the second user is bounded by

$$R_{m,2}^{MIMO-OMA} \leq \beta \log_2(1 + \frac{\gamma\rho \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2}{\beta}). \tag{36}$$

By applying Jensen's inequality, the following upper bound can be established:

$$\begin{aligned} &R_{m,1}^{MIMO-OMA} + R_{m,2}^{MIMO-OMA} \leq \\ &\leq (1-\beta) \log_2(1 + \frac{(1-\gamma)\rho \|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2}{1-\beta}) + \\ &+ \beta \log_2(1 + \frac{\gamma\rho \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2}{\beta}) \leq \\ &\leq \log_2((1-\beta) + \beta + (1-\beta)\frac{(1-\gamma)\rho \|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2}{1-\beta} + \\ &+ \beta \frac{\gamma\rho \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2}{\beta}) = \\ &= \log_2(1 + \rho(1-\gamma)\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 + \rho\gamma\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2), \end{aligned} \tag{37}$$

where the equality in the second inequality only holds if

$$\frac{\gamma \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2}{\beta} = \frac{(1-\gamma)\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2}{1-\beta}. \tag{38}$$

From (38) it can now be derived the optimal split of the resources to achieve the maximum sum-rate of MIMO-OMA:

$$\beta^* = \frac{\gamma \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2}{\gamma \|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 + (1-\gamma)\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2}. \tag{39}$$

The rates for MIMO-NOMA and MIMO-OMA as a function of the power allocation coefficient $\alpha_{m,2}^2 = \gamma$ and $\beta = \beta^*$ can be seen in Figure 6.

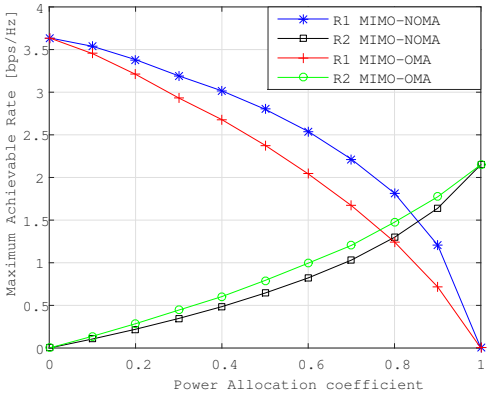

Figure 6: Maximum rates achieved by a MIMO-NOMA and a MIMO-OMA scheme. The SNR is $\rho = 10dB$. Power allocation coefficient $\alpha_{m,2}^2 = \gamma$.

Important conclusions can be made about the figure 6, namely, the only situation where MIMO-NOMA rates are equal to MIMO-OMA rates is when one of the users is not communicating ($\alpha_{m,2}^2 = \gamma = 0$ or $\alpha_{m,2}^2 = \gamma = 1$). Also, $R_{m,1}^{MIMO-NOMA} > R_{m,1}^{MIMO-OMA}$ for every $0 < \alpha_{m,2}^2 = \gamma < 1$, which makes sense since the first OMA user has to divide the frequency or time resources with the second OMA user while the first NOMA user does not have this restriction.

It seems odd that $R_{m,2}^{MIMO-OMA} > R_{m,2}^{MIMO-NOMA}$ for every $0 < \alpha_{m,2}^2 = \beta = 1 < 1$, but this can be explained, by the fact that the second NOMA user $R_{m,2}^{MIMO-NOMA}$ is interference limited, because the second user decodes its own signal with interference from the first user, while the second OMA user does not suffer any impairment by the presence of the first user.

It is also noteworthy that the rates seem to grow with ρ. Looking at equations (33), (34), (35) and (36), it can be seen that ρ increases the term inside the logarithm, except in the case of (34), where that relation is less obvious. However, since $\alpha_{m,2}^2 > \alpha_{m,1}^2$, the increase of ρ also results in the increasing of $R_{m,2}^{MIMO-NOMA}$ in that case. In order to compare this results with those of Tse [12], one should represent the boundaries of rate pairs achieved by MIMO-NOMA and MIMO-OMA, as it is presented in Figure 7.

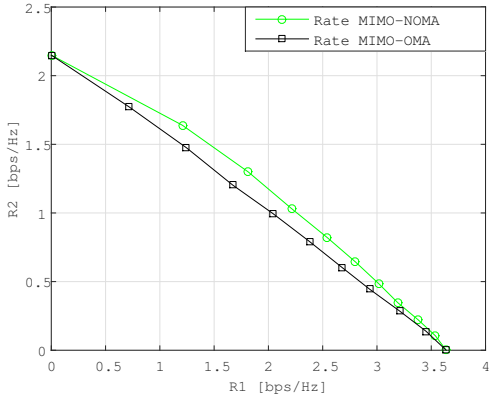

Figure 7: Boundary of rate pairs achieved by MIMO-NOMA and MIMO-OMA systems, each with two users. $\rho = 10dB$.

Comparing the results with what was obtained by Tse in [12], it can be seen that the results from SISO-NOMA and SISO-OMA also apply to MIMO-NOMA and MIMO-OMA. Again, NOMA and OMA have the same performance when only one user is being communicated too, as usual. Otherwise, the sum-rate of MIMO-NOMA is always superior.

Now, we want to compare the sum channel capacity of MIMO-NOMA versus MIMO-OMA. As seen previously, the sum channel capacity of MIMO-NOMA can be written as:

$$\begin{aligned} &C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} = \\ &= \log_2(1 + \rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2) + \\ &+ \log_2(1 + \frac{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2}{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2 + 1}), \end{aligned} \tag{40}$$

knowing that $\log_c(a) + \log_c(b) = \log_c(a \times b)$:

$$\begin{aligned} &C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} = \\ &= \log_2(1 + \rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2 + \\ &+ \frac{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2}{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2 + 1} + \\ &+ \rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2 \frac{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2}{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2 + 1}) = \\ &= \log_2(1 + \rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2 + \\ &= \rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2 \frac{\rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2 + 1}{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2 + 1})). \end{aligned} \tag{41}$$

Now, from (13), is known that $\frac{\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2}{\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2} > 1$, and therefore:

$$\begin{aligned} &C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} = \\ &= \log_2(1 + \rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2 + \\ &+ \rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2 \frac{\rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2 + 1}{\rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,1}^2 + 1}) \geq \\ &\geq \log_2(1 + \rho\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 \alpha_{m,1}^2 + \rho\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2 \alpha_{m,2}^2). \end{aligned} \tag{42}$$

Recalling from equation (37):

$$\begin{aligned} &C_{m,1}^{MIMO-OMA} + C_{m,2}^{MIMO-OMA} = \\ &\log_2(1 + \rho(1-\gamma)\|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\|^2 + \rho\gamma\|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\|^2), \end{aligned} \tag{43}$$

substituting in equation (42) if it agrees that the power allocation coefficients for NOMA are the same as for OMA ($\alpha_{m,2}^2 = \gamma$):

$$\begin{aligned} C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} \geq \\ C_{m,1}^{MIMO-OMA} + C_{m,2}^{MIMO-OMA}, \end{aligned} \tag{44}$$

which proves that there is a power split for which MIMO-NOMA can achieve a larger sum channel capacity than MIMO-OMA (with equality when only one user is being communicated to). These results were also confirmed by simulations, as seen in Figures 8, 9 and 10.

In those Figures it is evident that the difference between $C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA}$ and $C_{m,1}^{MIMO-OMA} + C_{m,2}^{MIMO-OMA}$ grows with $\alpha_{m,2}^2 = \gamma$. This is in accord with the results in section 4, namely, the fact that allocating too much power to the user with the best channel (in the case of Figure 10) tends to significantly lower the performance of the user with the worst channel, leading to a maximum sum-rate of the channel when using NOMA that is very similar to the OMA one. As the power allocation coefficient for the second user grows, the difference between the performance of NOMA compared to OMA also increases.

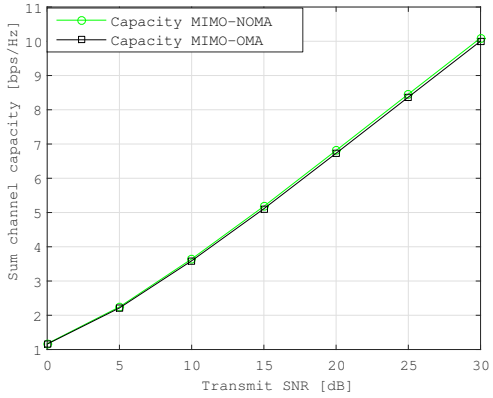

Figure 8: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with two users each, with $\alpha^2_{m,2} = \gamma = 0.1$ and $\alpha^2_{m,1} = 1 - \gamma = 0.9$.

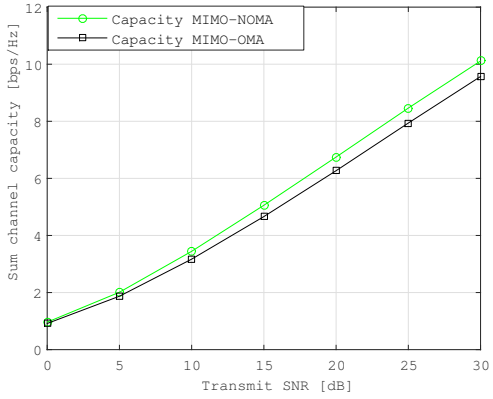

Figure 9: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with two users each, with $\alpha^2_{m,2} = \gamma = 0.5$ and $\alpha^2_{m,1} = 1 - \gamma = 0.5$.

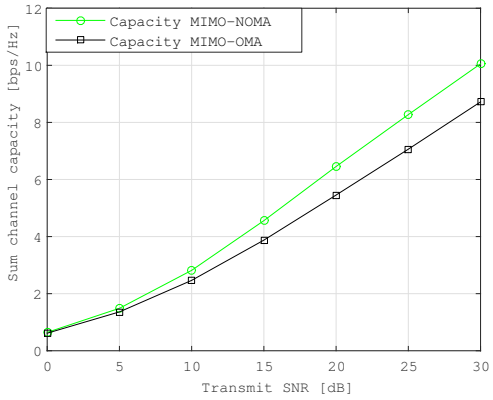

Figure 10: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with two users each, with $\alpha^2_{m,2} = \gamma = 0.9$ and $\alpha^2_{m,1} = 1 - \gamma = 0.1$.

7. Conclusions

This thesis looked at some practical aspects of the implementation of uncoded MIMO-NOMA related to the distribution of the power allocation coefficients and the limitations of SIC detection with alphabets larger than the binary one and the limitations in the number of users that can be supported. The analytical results in [5] have proven to hold. It has been explained why uncoded NOMA struggles to serve more than two users and an extension of this model range to up to five users was managed, when the users make use of BPSK only. The results show that using SIC is actually feasible up to five multiplexed users for the detection of NOMA with BPSK, while maintaining the target of $SER > 0.5$ for $SNR = 10$ dB. A massive MIMO model was also tested, in terms of SER the results were worse than the first model but it allowed a higher number of clusters. While the limitations in terms of users and modulations may be below our expectations for the 5G RAT, it can still be useful for certain type of applications (M2M communications, for example). Results for the intra-cluster relaying concept were also obtained, confirming the benefit of relaying information from the users with better channel coefficients to users with lower channel coefficients.

This thesis also looked at the rates of both MIMO-NOMA and MIMO-OMA systems, confirming that the rate curves for SISO-NOMA and SISO-OMA in the literature are consistent with the MIMO-NOMA and MIMO-OMA rates obtained in this dissertation. The dual SNR regime found in the SER curves was also found in the rate curves. The improvement of using MIMO-NOMA instead of MIMO-OMA was less then linear, in terms of rate, because of the intra-cluster interference (or inter user interference).

In terms of future work, the objectives of the thesis were fulfil but related problems around NOMA detection are still open to research:

- In this thesis, the focus was to explore the limitations of NOMA rather than to optimize the system parameters. However, the optimization of power allocation coefficients is a critical point, since it affects fairness between users and the total system throughput. There is some work done in this regard [11], and this thesis provided an easy formula to maximize fairness for the multi-user BPSK case but the topic is far from closed.
- The order of the users is known at the BS and at each user, however, since this transmission of information is not error free, the effects of errors need to be studied. If there is an error in the pilots that are sent to the BS, the users can be not properly ordered and there will be problems regarding fairness, since users with better channels can be assigned with higher power allocation coefficients and

thus accidentally end up in a water-filling situation.

• Throughout this thesis, any user in a NOMA system would be able to access the symbols being transmitted to any user in its cluster. Users that were positioned later in the decoding chain would even decode the other user's symbols, in order to nullify that user interference on its own signal. Hence, the security topic has been disregarded in this thesis. In a future work, it is imperative to study some mechanisms that prevent this easy access to another user's information.

References

[1] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire. Joint spatial division and multiplexing—the large-scale array regime. *IEEE Transactions on Information Theory*, 59(10):6441–6463, 2013.

[2] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang. What will 5g be? *Selected Areas in Communications, IEEE Journal on*, 32(6):1065–1082, June 2014.

[3] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski. Five disruptive technology directions for 5G. *IEEE Communications Magazine*, 52(2):74–80, February 2014.

[4] 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)*, pages 80–84, Guildford, UK, September 2015. IEEE.

[5] Z. Ding, F. Adachi, and H. Poor. The application of MIMO to non-orthogonal multiple access. *Wireless Communications, IEEE Transactions on*, 15(1):537–552, Jan 2016.

[6] 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. *Submitted to IEEE Communications Magazine*, 2015. [Online]. Available: http://arxiv.org/abs/1511.08610.

[7] Z. Ding and H. V. Poor. Design of massive-MIMO-NOMA with limited feedback. *CoRR*, abs/1511.05583, 2015.

[8] D. Kim, H. Lee, and D. Hong. A survey of in-band full-duplex transmission: From the perspective of phy and mac layers. *IEEE Communications Surveys Tutorials*, 17(4):2017–2046, Fourthquarter 2015.

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

[10] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat. Non-orthogonal multiple access for visible light communications. *IEEE Photonics Technology Letters*, 28(1):51–54, Jan 2016.

[11] S. Timotheou and I. Krikidis. Fairness for non-orthogonal multiple access in 5G systems. *Signal Processing Letters, IEEE*, 22(10):1647–1651, Oct 2015.

[12] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, New York, NY, USA, 2005.

[13] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, and F. Wiedmann. 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. *Communications Magazine, IEEE*, 52(2):97–105, February 2014.