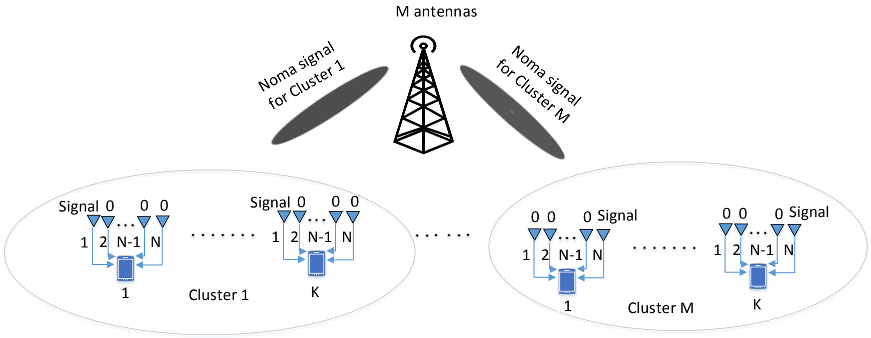

Assessment of Non-orthogonal Multiple Access for 5G systems

Ricardo José Neves Alberto

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Prof. António José Castelo Branco Rodrigues

Prof. Francisco António Taveira Branco Nunes Monteiro

Examination Committee

Chairperson: Prof. José Eduardo Charters Ribeiro da Cunha Sanguino

Supervisor: Prof. Francisco António Taveira Branco Nunes Monteiro

Member of the Committee: Prof. Fernando Duarte Nunes

November 2016

To my family and girlfriend

Acknowledgments

I want to express my appreciation and thanks to my supervisors Professor António Rodrigues and Professor Francisco Monteiro for giving me the opportunity of writing this thesis and all the hours they lost by helping me with it. I also want to express my thanks to Paulo Barata for all the help with the computer room in IST, without whom it would have been impossible to have considered that room my second home during these months. Furthermore, I would like to express my gratitude to my girlfriend Marta Veríssimo for the endless hours of help and conversation, without her this thesis would simply not exist, her contribution was priceless. Last but not least, to my family who supported me in every way, providing a stable environment that allowed me to work profitably every day while at home. I am also thankful to the Instituto de Telecomunicações and FCT for providing me a scholarship under the project LDimMNetcod.

Resumo

O acesso múltiplo não ortogonal (NOMA) tornou-se recentemente um candidato para a próxima geração de sistemas sem fios. Com NOMA, todos os sinais são adicionados no domínio da potência e separados nos receptores conforme os seus níveis de potência. O NOMA tem sido estudado na literatura numa perspectiva de teoria da informação em termos de capacidade e *outage probability.* Esta tese utiliza configurações de sistemas com múltiplos utilizadores MIMO e fornece análises de desempenho através de simulações numéricas para *downlink* NOMA. No primeiro sistema não é assumida pre-codificação na estação de base, para além da alocação de potência. A interferência entre grupos é eliminada através de uma filtragem linear nos terminais como nos sistemas de múltiplos utilizadores MIMO. No segundo sistema é usada pre-codificação na estação base para eliminar a a interferência entre grupos. Em ambos os sistemas os utilizadores são agrupados em feixes MIMO e o NOMA é posteriormente aplicado dentro desses grupos, onde um método de cancelamento de interferência sucessivo é aplicado para separar os diferentes sinais multiplexados em NOMA. Os utilizadores são ordenados de acordo com o seu coeficiente de canal e é observado que dentro de cada grupo se pode encontrar um regime dual dependente da relação sinal ruído: Com uma relação sinal ruído alta os terminais com melhores coeficientes de canal têm melhor desempenho mas com uma relação sinal ruído baixa em certas condições os utilizadores com coeficientes de canal mais baixo têm melhor desempenho. Algumas das limitações práticas do cancelamento sucessivo de interferência são apresentadas e são dados exemplos específicos disso.

É feita uma análise em termos de débito binário, onde um sistema MIMO-NOMA é comparado com um sistema MIMO-OMA. O pressuposto de que NOMA tem melhor desempenho que OMA é confirmado e o regime dual encontrado nas curvas de symbol error rate (SER) é demonstrado nas curvas de débito binário.

Palavras-chave: Acesso múltiplo não ortogonal , Cancelamento sucessivo de interferência, Cancelamento linear de interferência entre grupos.

Abstract

Non-orthogonal multiple access (NOMA) recently became a prominent candidate for next generation wireless systems. With NOMA all signals are added and separated at the receivers, taking in consideration their different power levels. NOMA has been much studied in the literature from an information-theoretic perspective in terms on capacity and outage probability. This thesis looks at two typical system configurations of multiuser MIMO systems and provides performance via traditional numerical simulations for downlink NOMA. In the first system no precoding is assumed at the base station, apart from the power allocation policy. Inter-cluster interference is dealt with by linear filtering at the terminals, as in MIMO multiuser systems. For the second system precoding is used at the base station to eliminate inter-cluster interference, and a massive MIMO array is considered at the base station. In both systems the 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. The intra-cluster users are ordered according to their channel power and it is observed that within each cluster one finds a dual regime depending on the signal-to-noise ratio (SNR): at high SNR the terminals with the best channel gains are the ones performing better, but in the low SNR regime, in certain conditions, the users with the weaker channels are the ones with better performance. Some of the practical limitations of successive interference cancellation (SIC) are highlighted and precise examples of that are given. A rate analysis is made where a MIMO-NOMA system is compared to a MIMO-OMA. The assumption that NOMA outperforms OMA is confirmed and the dual regime found in the SER curves is shown in the rate curves.

Keywords: Non-orthogonal multiple access, successive interference cancellation, linear inter-cluster cancellation.

Contents

List of Figures

Nomenclature

Operators and Sets

$\mathcal{CN}(\mu, \sigma^2)$	Complex normal distribution with mean value μ and variance σ^2.
$D_x^{(2)}[f(x)]$	Second derivative of $f(x)$
$\mathbb{E}\{\cdot\}$	Expectation operator.
$\log(x)$	base-e logarithm of x.
$\log_2(x)$	base-2 logarithm of x.
$\min(a, b)$	Minimum between a and b.
$\mathbb{C}$	Set of complex numbers.
$\mathbb{R}$	Set of real numbers.

Chapter 1

T	Number of OFDMA sub-carriers.

Chapter 2

α_i	Square root of the power allocation coefficient for user i in a NOMA system.
C	Capacity.

C_{sum}	Sum capacity.
C_{sym}	Symmetric capacity.
C_{MIMO}	Capacity of a MIMO system.
C_{SISO}	Capacity of a SISO system.
d_{order}	Diversity order.
γ_i	Power allocation coefficient for user i in a OMA system.
$\mathbf{H}$	Channel matrix.
K_u	Number of users.
λ_i	$i-th$ eigenvalue of matrix $\mathbf{H}$.
r	Multiplexing gain.
M	Number of antennas in the transmitter.
N	Number of antennas in the receiver.
N_{min}	Min(N,M).
$\mathbf{n}$	Unit power additive white gaussian noise vector.
N_0	Noise spectral density.
P_{error}^{i}	Probability that a user i in the SIC decoding chain misinterprets a symbol.
P_i	Transmission power of user i.
P_n	Power of vector n.
P_x	Power of vector x.
R	Rate.
$\mathbf{\Sigma}$	Matrix that contains the singular values of $\mathbf{H}$.

SNR	Signal-to-noise ratio.
$\mathbf{U}$	Matrix that contains the left singular vectors of $\mathbf{H}$.
$\mathbf{V^T}$	Matrix that contains the right singular vectors of $\mathbf{H}$.
x_i	Modulation symbol of user i.
$\mathbf{x}$	Vector that contains the symbols to be transmitted.
$\mathbf{y}$	Received vector.

Chapter 3

$\alpha_{\mathrm{m,k}}$	Power allocation coefficient for user k from cluster m.
d	Euclidean distance.
$\mathbf{H}_{\mathrm{m,k}}$	Channel matrix of user k from cluster m.
$\tilde{\mathbf{H}}_{\mathrm{i,k}}$	$\mathbf{H}_{\mathrm{m,k}}$ matrix without the $h_{\mathrm{m,i}k}$ column.
$\mathbf{h}_{\mathrm{m,i}k}$	m-th column of the $\mathbf{H}_{\mathrm{m,k}}$ matrix.
K	Number of users per cluster.
λ_i	$i-th$ eigenvalue of matrix $\mathbf{H}_{\mathrm{m,k}}$.
M	Number of antennas in the BS and also the number of clusters.
$\mathbf{n}_{\mathrm{m,k}}$	Unit power additive white gaussian noise vector of user k from cluster m.
N	Number of antennas in each user.
$\mathbf{P}$	Power allocation matrix.
$\mathbf{P}_U$	Projection matrix that projects a vector into the space of $\tilde{\mathbf{U}}_{\mathrm{i,k}}$.
$\tilde{\mathbf{s}}$	NOMA vector that is transmitted by the BS.
$s_{m,k}$	Symbol to be transmitted to the user k from cluster m.

$\tilde{\mathbf{s}}_{\backslash 1}$	NOMA vector $\tilde{\mathbf{s}}$ without the contribution of the symbols from the first cluster.
$\tilde{\mathbf{U}}_{\mathrm{i,k}}$	Left singular values of $\tilde{\mathbf{H}}_{\mathrm{i,k}}$ that correspond to zero singular values.
$\mathbf{V}_{\mathrm{m,k}}$	Detection matrix of user k from cluster m.
$\mathbf{y}_{\mathrm{m,k}}$	Received vector in the user k from cluster m.

Chapter 4

β	Splitting of resources in an OMA system.
β^*	Optimal split of resources in an OMA system to maximize the sum-rate.
$C_{m,k}^{MIMO-NOMA}$	Capacity for a MIMO-NOMA system in user k from cluster m.
$C_{m,k}^{MIMO-OMA}$	Capacity for a MIMO-OMA system in user k from cluster m.
γ	Power allocation coefficient for an OMA system.
ρ	Transmit signal to noise ratio.
$R_{m,k}^{MIMO-NOMA}$	Rate for a MIMO-NOMA system in user k from cluster m.
$R_{m,k}^{MIMO-OMA}$	Rate for a MIMO-OMA system in user k from cluster m.
$SINR_{m,k}$	Signal to noise plus interference ratio for user k from cluster m.
σ_x^2	Variance of the signal.
σ_n^2	Variance of the the unit power additive white Gaussian noise.

Chapter 1

Introduction

In this initial chapter, the motivation behind this thesis will be explained and an historical overview of what lead to this topic of research will be presented. The goals of this thesis will be formulated, and the structure of the document will also be provided.

1.1 History: From 1G to 5G

In the 1970s, the foundations of mobile telecommunications were laid with the first generation of mobile networks (1G). It introduced seamless connectivity of voice services in determined zones of the world. By being an analogue technology, it had some limitations, for example, it only supported one user per channel. The channel occupied 25 KHz. In terms of its radio access technology (RAT), frequency division multiple access (FDMA) was used, multiple users were assigned to different frequencies but frequency gaps in-between channels were needed, to minimize adjacent-channel interference (ACI).

The second generation of mobile networks (2G), also known as global system for mobile communications (GSM), used time division multiple access (TDMA), a RAT that allowed eight users to share the same channel, occupying 200 KHz. Therefore, there was no improvement in terms of spectral efficiency but GSM, being a digital technology, was the first one to bring mobile devices to the masses.

The third generation of mobile networks (3G), also known as universal mobile telecommunications system (UMTS) brought code division multiple access (CDMA) as its RAT, allowing users to share the same frequency and communicate at the same time, using different orthogonal codes. This was a breakthrough from a spectral efficiency theoretical point of view since users could share the same time/frequency but it had some limitations, namely, the bandwidth used was very large compared to 2G and there was a limitation in the number of codes that could be assigned to the users.

The fourth generation of mobile networks (4G), also known as long term evolution (LTE), came as a response to the need of faster and better mobile broadband. From a spectral efficiency perspective, the RAT used in downlink LTE, orthogonal frequency division multiple access (OFDMA), is not great because the OFDMA sub-carriers are packed in 20 MHz of spectrum. However, the interference between these sub-carriers is avoided by choosing orthogonal frequencies to each sub-carrier, so that, although the spectrum of the subcarriers overlaps, they do not interfere with each other and that is the main advantage of OFDMA. But, the need for more capacity continues and LTE will not provide enough capacity for the near future.

The fifth generation of mobile networks (5G) is now in research stage, with the first implementations being scheduled for 2020. This thesis will aim to explore in detail the advantages and limitations of non-orthogonal multiple access (NOMA), as a candidate for the 5G RAT.

1.2 Motivation

In the last section, it was said that LTE would not provide enough capacity for the near future. With applications such as internet of things (IoT) and machine-to-machine (M2M) communications, it is expected that by 2020 the world will have 26 billion connected devices [1]. In 2014 there were already three countries with more than two mobile subscriptions per person and one country with three mobile subscriptions per person [2]. All this caused a four thousand-fold increase in mobile data traffic in the past ten years and will result in an eightfold increase of mobile data traffic between 2015 and 2020 [3].

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, T, of OFDMA sub-carriers 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 sub-carriers and grows according to $\log(T)$ [4].

An opportunity arises not only for a new RAT that solves this problem but for a new generation of mobile communications. Both the industry and academics are sensitive to this issue and several projects already exist such as [5], [6] and [7].The main drivers for 5G will be [4]:

- IoT: The challenge will be how to support connection of up to one hundred thousand machine-type communication (MTC) in a cell, not forgetting the low cost and long lifetime premises.

- Gigabit wireless connectivity: The challenge will be how to provide wireless rates that rival the wired ones.

- Tactile internet: The challenge will be lowering the latency to a level that the human user will perceive as zero latency in applications like remote health interventions.

These drivers will define the requirements of 5G, namely, compared to 4G, 5G will need to support [4, 8]:

- 1000 times higher mobile data volume per area,
- 10 to 100 higher number of connected devices,
- 10 to 100 times higher user data-rate, to around 10 Gbps,
- 10 times longer battery life,
- 5 times reduced end to end latency, to around 1ms.

As seen in the points above, 5G will require drastic improvements in many areas. It has been seen previously that LTE will struggle to provide the capacity increase needed, even for 2020 (which is the expected year to have the first 5G network operating in Japan, at the time of the Olympic Games).

Therefore, it becomes extremely important to understand the relation between user throughput and spectral efficiency, in order to understand how it can increase the former via a new RAT that increases the latter. This relation can be expressed as:

$$\begin{aligned}&\text{Throughput per area [bit/s/area]} = \text{spectral bandwidth [Hz]}\times \\ &\text{cell density [cell/area]} \times \text{spectral efficiency[bit/s/Hz/cell]}\end{aligned} \tag{1.1}$$

The throughput of a user increases linearly with the spectral efficiency. The equation (1.1) shows the importance of spectrum, which is a very scarce resource and the importance of cell density, which unfortunately cannot be increased ad infinitum. In 2015, the spectrum auction advanced wireless services (AWS)-3, in the USA, where spectrum between 1700 MHz and 2100 MHz was being auctioned, raised 44.9 billion dollars in bids.

It can be argued that 5G is expected to explore frequency bands that are not currently in use (or at least, not so much), namely, frequency bands above 3 GHz, the so-called millimeter waves communications [9]. However, moving to higher spectrum frequencies also plays with another term of the equation (1.1), the cell density. As it moves higher in the frequency plane, the attenuation that the communications suffer increase, meaning that it needs to cover the same space with more cells if it wants equal levels of signal received. At first glance that would solve the problem completely but, as more BS occupy the space, the interference between them also increases and if not planned carefully, may lead to zero throughput because of the interference [10].

It is then essential to research a new RAT that allows for major improvements in spectral efficiency compared to OFDMA. In fact the most simple version of NOMA has already been included in long term evolution advanced (LTE-A), release 13, to support two users sharing the same frequency and time resources, called multi-user superposition transmission (MUST) [11].

1.3 Goals and Structure of this Document

The organisation of this dissertation in terms of chapters and respective contents will be as follows:

Chapter 2 - NOMA and MIMO Overview: A description of state of the art NOMA ideas will be presented along with some fundamentals of MIMO.

Chapter 3 - NOMA Model: A working NOMA system with MIMO but without inter-cluster-interference will be presented and its performance is evaluated.

Chapter 4 - Rate and capacity analysis: A detailed analysis of rate and capacity will be done using the models from chapter 3. A comparison of NOMA and OMA using these two parameters will be made.

Chapter 5 - Conclusions: The conclusions of the thesis will be presented, along with some topics that could be used for future work.

1.4 Contributions of this Dissertation

The original contribution of this research work is presented in chapter 3. It can be summarised as follows:

NOMA: In chapter 3 a model that uses MIMO-NOMA without precoding at the base-station, with user clustering but without inter-cluster interference (dealt by linear filtering at the receivers) is analysed, with various simulations for SER curves depending on the modulations. In chapter 4 a rate analysis is made using the model from chapter 3, that is compared against a MIMO-OMA system. The assumption that NOMA outperforms OMA is confirmed and the SER curves are shown to be consistent with the rate curves.This part of the work is planned to result in the following submission:

- R. Alberto and F. Monteiro “Performance of Multiuser NOMA with High-Order Modulations and More than Two Users,” to be submitted to a conference, soon.

Chapter 2

An overview of Non-orthogonal multiple access and MIMO

In this chapter a description of state of the art NOMA ideas and peculiarities will be provided along with a overview of the fundamentals of multiple-input multiple-output (MIMO). A summary of the few literature on NOMA will be made, namely, rate expressions for the rates in the single-input single-output (SISO) case will be presented and some known successive interference cancellation (SIC) problems will be detailed and discussed.

2.1 Introduction to MIMO Communications

In wireless communications the traditional structure of a point to point communication is a single antenna at the transmitter and a single antenna at the receiver. This structure is called SISO. This structure was widely used for many years and it is still used in contemporary 2G and 3G systems. However, the requirements in capacity forced this structure to evolve to MIMO, where it has generically M antennas in the transmitter and N antennas in the receiver. This structure was validated for the first time in [12], where the authors proved the superior spectral efficiency (and thus the superior throughput) of this structure compared to SISO.

A MIMO system can be described by the equation:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{2.1}$$

where $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is a vector containing the symbols to be transmitted, $\mathbf{H} \in \mathbb{C}^{N\times M}$ is the channel matrix, which represents the "mixture" of the signals caused by the channel, $\mathbf{n} \sim \mathcal{CN}(0, \sigma_n^2) \in \mathbb{C}^{N\times 1}$ is the unit power additive white Gaussian noise vector and $\mathbf{y} \in \mathbb{C}^{N\times 1}$ is the received vector. A simplified example of an uplink MIMO system can be seen in Figure 2.1 with both M and N bigger then 1:

If $M > 1$ while $N = 1$, the structure is called multiple input single output (MISO). Similarly, if $M = 1$ while $N > 1$, the structure is called single-input multiple-output (SIMO).

The concept of "capacity of a channel" was defined by Shannon in 1948 as the mutual information maximized over all possible input distributions (in bit/s/Hz) [13], and he also defined the equation for the channel capacity of a SISO structure:

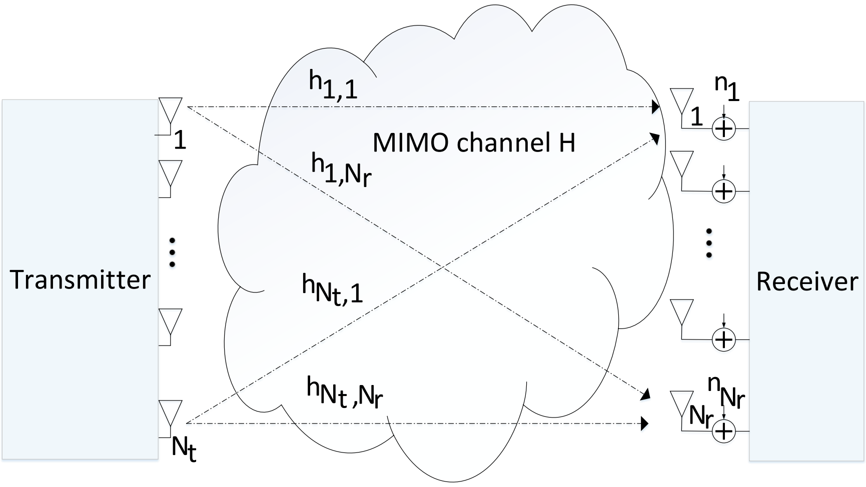

Figure 2.1: MIMO uplink channel with M transmit antennas and N receive antennas.

$$C_{SISO} = \log_2(1 + SNR), \tag{2.2}$$

where $SNR = \frac{P_x}{P_n}$, with P_x being the power of vector $\mathbf{x}$ and P_n being the power of the noise vector $\mathbf{n}$. Later, in 1995, Telatar derived the equation for the channel capacity of MIMO [14, 15]:

$$C_{MIMO} = \log_2[\det(\mathbf{I}_N + \frac{SNR}{M}\mathbf{H}\mathbf{H}^{\mathrm{H}}), \tag{2.3}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix and M is the number of transmitter antennas.

Later on, Goldsmith [16] in 2003 describes in a systematic way the capacities of the various possible scenarios in MIMO (MAC channel, broadcast channel, SIMO, MISO, etc). Later in the dissertation, the capacity of the SISO channel and the MIMO channel will be compared and a detailed analysis of the capacity of both MIMO-NOMA and MIMO-orthogonal multiple access (OMA) systems from chapter 3 will be presented in chapter 4.

Some considerations will be assumed throughout the thesis:

- The transmitted signal is narrowband enough so that the channel can be considered frequency non-selective.

- The channel is modelled as slow-fading, which means that the channel matrix $\mathbf{H}$ is constant for the duration of a coding block (tens to thousands of channel symbols).

- Antenna elements are spaced sufficiently far apart such that the entries of matrix can be modelled as independent and identically distributed (i.i.d.) Gaussian random variables with zero-mean and unit variance ($\sim \mathcal{CN}(0,1)$), the denominated Rayleigh fading.

- Perfect information of the channel matrix $\mathbf{H}$, denoted as channel state information (CSI), is often considered to be known at the receiver (CSI-R). Due to the inherent complexity in estimating CSI at the transmitter side (CSI-T), unless otherwise stated, it is generally not considered to be available.

- Transmitted symbols $\mathbf{x}$ will be often picked from n M-ary quadrature amplitude modulation (QAM). To guarantee symmetry in the constellation, M is chosen as an even power of two, that is $M = 2^{2n}$, with n being an integer. A more basic modulation, binary phase-shift keying (BPSK) will also be used.

An important concept in MIMO is the one of diversity, which can be either transmit diversity or receiver diversity. Receiver (or equivalently transmit) diversity is present when there are more than one antennas at the receiver (or transmitter) side. In this case, and in the presence of fading, independent signals from multiple antennas can be combined in order to mitigate the fading fluctuations. In the limit of having an infinite number of antennas at the receiver side, one would get the additive white gaussian noise (AWGN) channel, since fading would be eliminated. However, in order to get independence between the signals, the antennas need to be sufficiently spaced, normally, a distance of half wavelength is considered sufficient. When plotting a system symbol error rate (SER) or bit error rate (BER) against signal-to-noise ratio (SNR), one can quantify the diversity order as:

$$d_{order} = \lim_{SNR \to \infty} \frac{\log(SER)}{\log(SNR)}, \tag{2.4}$$

which is simply the slope of the SER curve in the high SNR regime.

There is another gain in MIMO, the multiplexing gain. While the diversity gain improves the link by making it more fading resistant (gaining reliability), transmitting identical signals from various antennas, with the multiplexing gain the idea is to improve the throughput of the channel, using the various antennas to send separated and complementary parts of a message. This difference is illustrated in Figure 2.2 for the transmission of a sequence of bits “010”.

By analysing Figure 2.2 it is intuitively clear that there must be a trade-off between these gains, for the example given by the figure, if one wants a multiplexing gain of three, it also gets a diversity gain of zero and vice versa. This intuitive trade-off has been quantified firstly in research papers and more recently put into MIMO textbooks,

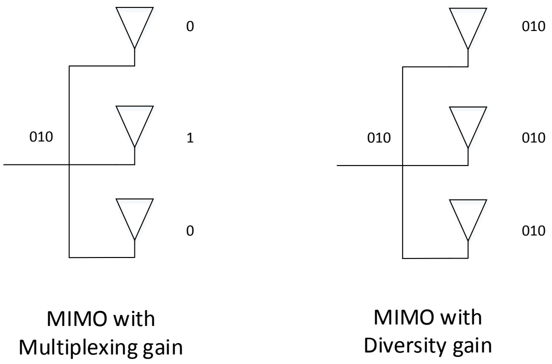

Figure 2.2: A multi-antenna transmitter exemplifying the difference between MIMO diversity and multiplexing.

for example in [17] and, defining r as the multiplexing gain and d as the diversity gain, under the assumption that the fading block length exceeds the total number of antennas at the transmitter and receiver, the optimal d in function of r is:

$$d_{order}(r) = (M - r)(N - r), \quad 0 \leq r \leq \min(M, N). \tag{2.5}$$

As a final remark about the topic, multiplexing is nowadays more important than diversity because there are techniques like orthogonal frequency division multiplexing (OFDM) that introduce diversity by themselves. This features of MIMO will not be profoundly studied in the thesis, since in chapter 3, the number of antennas will be fixed and the same signal will be transmitted by every antenna. The diversity gain will be seen in the SER plots.

Another interesting feature of MIMO is that as long as the CSI (i.e., the knowledge of the channel matrix $\mathbf{H}$) is available at both the receiver and the transmitter, it is possible to decompose the MIMO channel into a set of parallel channels. This can be done using the singular value decomposition (SVD). The SVD of matrix $\mathbf{H}$ is:

$$\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}, \tag{2.6}$$

where $\mathbf{U} \in \mathbb{C}^{N\times N}$ and $\mathbf{V}^{\mathbf{T}} \in \mathbb{C}^{M\times M}$ are unitary matrices ($\mathbf{U}\mathbf{U}^{\mathbf{T}} = \mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{V}\mathbf{V}^{\mathbf{T}} = \mathbf{V}^{\mathbf{T}}\mathbf{V} = \mathbf{I}$), with $\mathbf{U}$ containing the left singular vectors of $\mathbf{H}$ and $\mathbf{V}^{\mathbf{T}}$ containing the right singular vectors. Also, $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times M}$ is in general a rectangular matrix whose diagonal elements are the singular values of $\mathbf{H}$ ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{N_{min}}$) where N_{min} is $\min(N, M)$. The vector received is $\mathbf{y}$, as given in (2.1), and when H is known at the transmit side, the transmitter can choose the precoding matrix to be $\mathbf{V}$, transmitting not $\mathbf{x}$ but $\mathbf{Vx}$ and it can choose the detection matrix to be $\mathbf{U}^{\mathbf{T}}$, multiplying it in the receiver: $\mathbf{U}^{\mathbf{T}}\mathbf{y}$. With these operations, one gets:

$$\begin{aligned} &\mathbf{U}^{\mathbf{T}}\mathbf{y} = \mathbf{U}^{\mathbf{T}}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}\mathbf{V}\mathbf{x} + \mathbf{U}^{\mathbf{T}}\mathbf{n} <=> \\ &<=> \mathbf{U}^{\mathbf{T}}\mathbf{y} = \boldsymbol{\Sigma}\mathbf{x} + \mathbf{U}^{\mathbf{T}}\mathbf{n}. \end{aligned} \tag{2.7}$$

Now, the MIMO system has been reduced to N_{min} parallel SISO systems:

$$\mathbf{U}^{\mathbf{T}}\mathbf{y} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ \dots & & & & \\ 0 & 0 & \dots & \lambda_{N_{min}} & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \mathbf{x} + \mathbf{U}^{\mathbf{T}}\mathbf{n}. \tag{2.8}$$

Note that the multiplication by a unitary matrix does not change the noise distribution. The equivalent parallel SISO channels can be seen in Figure 2.3.

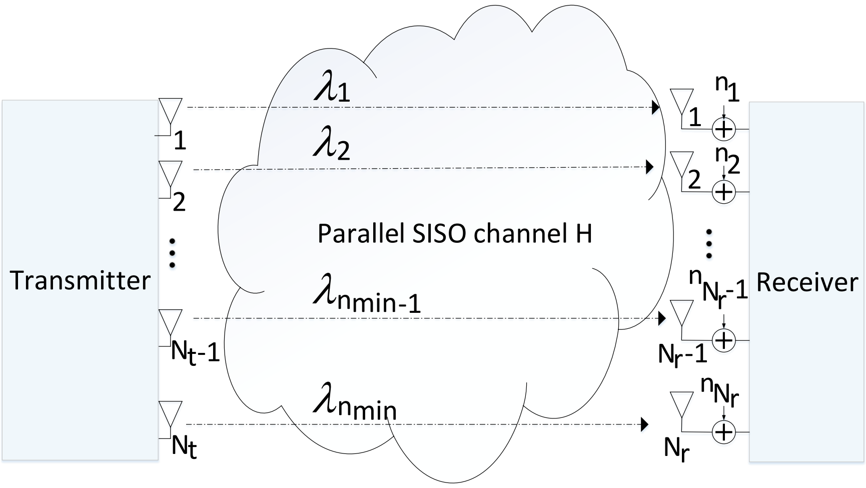

Figure 2.3: MIMO channel decomposition with M=N=N_{min}.

This strategy greatly reduces the system complexity, at the cost of having full CSI. Since it has N_{min} parallel SISO systems, it follows naturally that:

$$C_{MIMO} = N_{min} \times \log_2(1 + SNR) = N_{min} \times C_{SISO}. \tag{2.9}$$

In chapter 3, this technique will not be applied (full CSI will not be assumed) but, by means of a zero forcing (ZF) in the receiver, the MIMO-NOMA system will be converted into a SISO-NOMA system.

2.2 Introduction to non-orthogonal multiple access

The motivation for a new RAT for 5G has been discussed in the previous chapter. In contrast to OFDMA, NOMA will not depend on the frequency domain to divide users, it will superpose multiple users in the power domain, although the signal waveform could be based on the OFDMA one[18]. NOMA has been recognized as a very promising RAT for 5G [19], but to full-fill its promises, it requires a SIC receiver, so each user can demodulate and decode the signals from other users that share the same NOMA channel, apart from his own signal.

This requirement is a drawback in terms of receiver complexity, although similarly complex techniques are already applied, e.g., the turbo decoder[20]. Results show that NOMA using superposition coding (SC) in the transmitter and SIC in the receiver not only outperforms orthogonal multiplexing [21], but is also optimal in the sense of achieving the capacity region of the downlink broadcast channel [22, 23]. Finally, note that although the model that will be studied is for downlink NOMA, it has also been proved that NOMA can also be used in uplink, with SIC applied on the BS side [24].

2.3 Non-orthogonal multiple access concepts

In current OMA schemes, users had to share either time (TDMA) or frequency (FDMA and OFDMA). This share of resources is what limits the OMA schemes, because it limits the bandwidth that such schemes are able to offer, to each individual user. Taking FDMA has an example, the sharing of resources in the frequency can be seen in Figure 2.4.

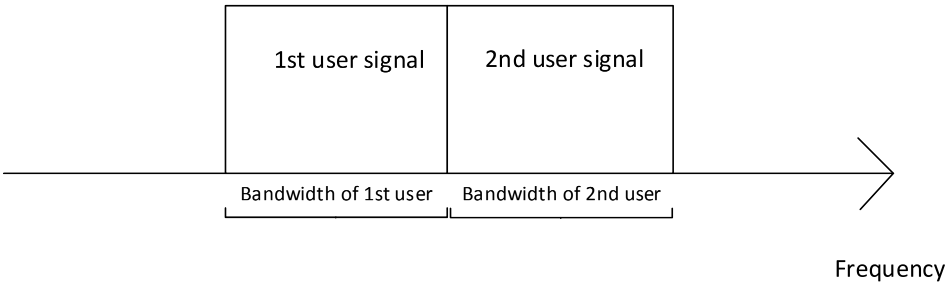

Figure 2.4: Scheme that represents the sharing of frequency between 2 users in a FDMA scheme.

The main advantage of NOMA is that users no longer need to share neither frequency or time, since they are superimposed in the power domain, they can use the same frequency and time, as can be seen in Figure 2.5.

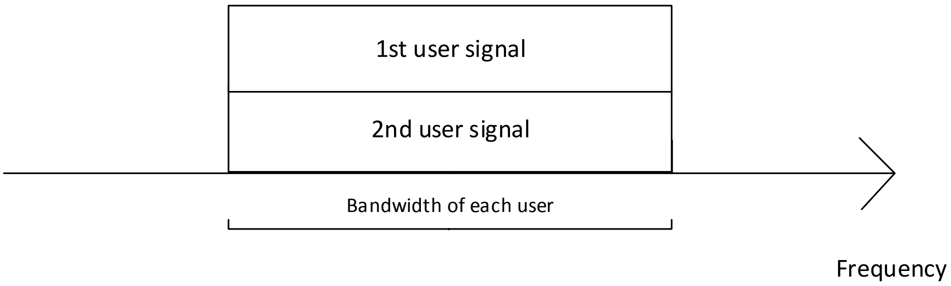

Figure 2.5: Scheme that represents the frequency spectrum being utilized by 2 users in a NOMA scheme.

When the SC and decoding is only done in the power domain, NOMA also has the advantage of not needing spreading codes (compared to CDMA). Taking into account equation (1.1), NOMA with just 2 users is theoretically able to double the throughput compared to current OMA schemes. However, for NOMA to be implemented, the decoder needs to decode both user signals, in this example, and in the general case if NOMA is implemented with K_u users, the receiver needs to decode K_u user signals in the worst case ($K_u - 1$ signals from other users and one from himself). This is where the difficulty of NOMA resides. A simple scheme that shows how decoding can be performed with 2 users can be seen in Figure 2.6:

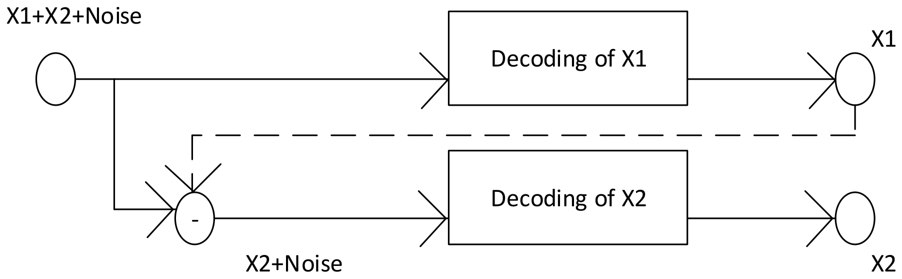

Figure 2.6: Scheme that represents the SIC decoding for two users.

It is also important to discuss how the users know their power allocation coefficients and thus their respective order in the decoding chain. When a user enters a cluster it is assigned an ID. Then, each user has to send pilots so that the base station is able to estimate each channel coefficient associated to each user. After that, the BS broadcasts a packet containing the power coefficients allocated to each user. By looking at the position associated to its ID, each user gets to know its own power coefficient and similarly gets to know the coefficients of all the other users, being able to calculate his position in the decoding chain, comparing his coefficient with the others.

There are at least two approaches on how to implement NOMA [25], in the first approach, different beams are assigned to different users [25]. In this thesis, namely in chapter 3, the second approach, which is to decompose the MIMO-NOMA system into a SISO-NOMA system, will be used, where each user signal is independently channel coded and modulated and then added with other users signals. The idea of this superposition is that the capacity of the system is improved without extending the bandwidth required, due to a better exploitation of the available resources.

There are two proposed ways to generate the superimposed NOMA signal, under investigation at 3rd generation partnership project (3GPP): semi-orthogonal multiple access (SOMA) and rate-adaptive constellation expansion multiple access (RA-CEMA). In chapter 3, SOMA will be used. It was firstly introduced in [26] and [27]. The idea is that each user information is independently modulated in a QAM constellation and then the information is summed with appropriate power allocation coefficients that ensure that the resulting signal forms a higher-order QAM modulation:

$$\mathbf{s} = \sum_{i=1}^{K_u} \alpha_i x_i, \tag{2.10}$$

where α_i is the square root of the power allocation coefficients, with the constraint $\sum_{i=1}^{K_u} \alpha_i^2 = 1$, K_u is the total number of users that share the same NOMA channel and x_i is the modulation symbol of user i. It is an open question if it needs gray mapping or not after the superposition.

In chapter 3, gray mapping will not be considered. SER will be used as a performance metric instead of BER. A visual example of SOMA can be seen in Figure 2.7.

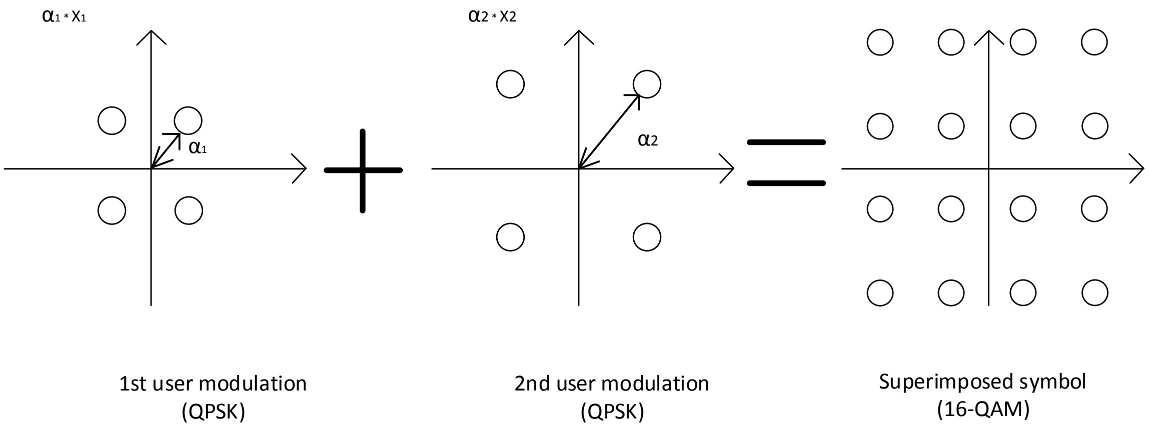
α1 * X1
α1
α2 * X2
α2
1st user modulation (QPSK)
2nd user modulation (QPSK)
Superimposed symbol (16-QAM)

Figure 2.7: Diagram with a SOMA example for two 4-QAM modulations.

Note than when using SOMA, a high difference in the power level of different symbols is going to be required in order to decode them with high probability. Looking carefully at Figure 2.6, it can see that the SIC receiver struggles to decode users whose channel coefficients are alike since the points in the superimposed symbol become very close to each others. SIC operates on the premise that users have very different channel coefficients and therefore it disregards the contribution of user 2 when decoding user 1 and then uses that result to decode the signal of user 2. This approach has obvious problems, hence why most of the research in NOMA uses a 2 user system model. This will be deeply analysed and explained in chapter 3.

RA-CEMA is another way to generate the superimposed symbol and was firstly introduced in [28]. The idea is that, similarly to SOMA, a QAM modulation is transmitted, but now the mapping of the coded bits of each user is adaptively controlled depending of their channel conditions, allowing a more detailed rate control. An example of this idea can be seen in Figure 2.8.

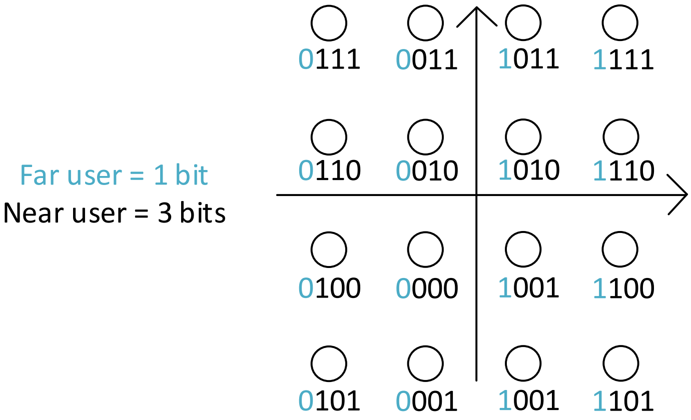

Figure 2.8: Diagram with a RA-CEMA example for 16-QAM modulation and 2 users.

User pairing between NOMA users is also being investigated, namely, pairing users with different quality of service (QoS) requisites or different channel gains [29]. By pairing very different types of users, the idea of user pairing is that the user with the highest QoS requirement should be assigned a higher power allocation coefficient (since it needs a high QoS) and the user with the lowest QoS requirement should be assigned a lower power allocation coefficient (since it needs a lower QoS). An example could be a user downloading a 4K video and another user doing MTC communications. This idea will be discussed in chapter 3, supported by the results obtained.

Another popular approach is to use a cooperative model, where near NOMA users that are close to the source act as relays to help far NOMA users [30], and local short-range communication techniques,such as bluetooth and ultra wide band (UWB) have been suggested for the relaying [31]. In terms of literature, SC with 2 users has been studied both for the uplink and downlink case, mostly in [22]. It states that the main difference between NOMA (SC with SIC at the receiver) and CDMA is that CDMA decodes information of a user treating all the other users as interference, while NOMA uses SIC, that decodes first the users with stronger channel gains (or stronger power allocation coefficients in our model), treating only the users with lower channel gains as interference, which allows a significant improvement.

2.3.1 Capacity region of the two-user uplink additive white Gaussian noise channel with successive interference cancellation

Let us consider the uplink AWGN channel with two users:

$$y = x_1 + x_2 + n, \tag{2.11}$$

where n is independent identically distributed complex Gaussian noise and x_1 and x_2 are, respectively, the symbol of the first and the second user. When point to point cases are analysed, the capacity of a channel is a performance limitation, namely, it can reliably communicate at rates $R < C$ and cannot at rates $R > C$.

In multi-user cases, this concept is extended to a capacity region $\mathcal{C}$, namely, a region where (e.g, for the 2 users case), user 1 and user 2 can reliably communicate at R_1 and R_2. Since they are sharing the bandwidth, when one needs to communicate at a higher rate, the other may need to lower his rate. Some performance measures can be derived from this region $\mathcal{C}$, namely:

- The symmetric capacity:

$$\mathrm{C}_{sym} = \max_{(R,R)\in\mathcal{C}} R \tag{2.12}$$

which is the maximum rate at which both users can simultaneously communicate.

- The sum capacity:

$$\mathrm{C}_{sum} = \max_{(R_1,R_1)\in\mathcal{C}} R_1 + R_2 \tag{2.13}$$

which is the maximum total throughput that can be achieved by the system. In the two user case, the capacity region needs to satisfy three constrains:

$$\begin{aligned} R_1 &< \log_2(1+\frac{P_1}{N_0}),\\ R_2 &< \log_2(1+\frac{P_2}{N_0}),\\ R_1 + R_2 &< \log_2(1+\frac{P_1+P_2}{N_0}). \end{aligned} \tag{2.14}$$

The first two constrains basically say that the rate of a user cannot be higher than the capacity of a point to point link where the other user is absent, which is obvious since the introduction of more users will degrade the individual rates and not improve them. The third constrain says that the total throughput of the system cannot exceed the capacity of a point-to-point AWGN channel with the sum of the received powers of the two users.

This applies since the signals the two users send are independent and therefore the power of the aggregate received signal is the sum of the powers of the individual received signals. Without this constrain the capacity region would be a rectangle, meaning that both users could simultaneously transmit at the point to point capacity, as if the other user did not exist, which would not make sense. The capacity region can be seen in Figure 2.9.

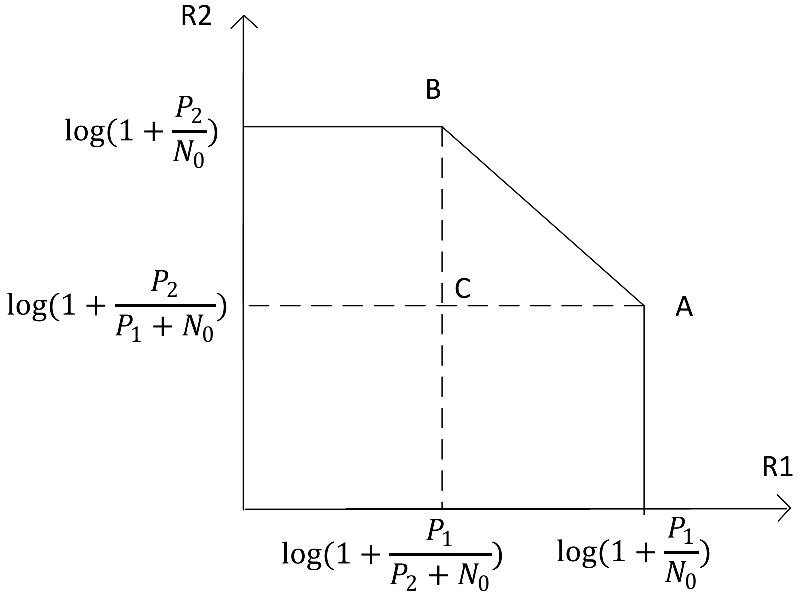

Figure 2.9: Capacity region of the two-user uplink AWGN channel with SIC.[22]

The “magic” of SIC is that user 2 can achieve its point-to-point bound while user 1 has a non zero rate, i.e., user 1 has the rate at point B:

$$R_1^* = \log_2(1+\frac{P_1+P_2}{N_0}) - \log_2(1+\frac{P_2}{N_0}) = \log_2(1+\frac{P_1}{P_2+N_0}). \tag{2.15}$$

This is possible because the receiver firstly decodes the data of user 1, treating the signal from user 2 as interference. Then it reconstructs the user 1’s signal and subtracts it from the aggregated signal. Finally, it decodes the data of user 2. Since now only AWGN noise is left in the system, user 2 can transmit at its point to point bound $\log_2(1+\frac{P_2}{N_0})$. If the decoding order is changed, it ends up at point A. The sum capacity is maximized in the segment AB. However, if the power received by the two users is different, the SIC strategy it should consider a corner point such that the stronger user is decoded first, so that the weak user can get the best rate. This point is said to be max-min fair. These specifics will be applied in chapter 3.

Point C in Figure 2.9 is the point achieved by CDMA, which is sub-optimal since in CDMA every user is treated as interference while with SIC the second user in the decoding chain has no added interference from the first user, allowing SIC to achieve the points in the AB segment.

One other advantage of NOMA is that the near-far problem in CDMA is turned into a near-far advantage in SIC because this detection technique benefits from users having very different channel coefficients. For the general case, with K_u users, the K_u-user capacity region can be described by $2_u^K - 1$ constrains:

$$\sum_{k \in S} R_k < \log_2(1 + \frac{\sum_{k \in S} P_k}{N_0}), \tag{2.16}$$

for all $S \subset (1, ..., K_u)$. Also, the sum capacity can be determined as:

$$\mathrm{C}_{sum} = \log_2(1 + \frac{\sum_{k=1}^{K_u} P_k}{N_0}). \tag{2.17}$$

For the simple case of equally received power, the sum capacity simplifies:

$$\mathrm{C}_{sum} = \log_2(1 + \frac{K_u P}{N_0}), \tag{2.18}$$

and therefore the symmetric capacity with K_u users can be written as

$$\mathrm{C}_{sym} = \frac{1}{K_u}\log_2(1 + \frac{K_u P}{N_0}). \tag{2.19}$$

2.3.2 Capacity region of the two-user asymmetric downlink additive white Gaussian noise channel with successive interference cancellation

Let us now consider the downlink case, with $|h_1| < |h_2|$ and both coefficients are constant (they are time invariant). In this case, since user 2 will have a better channel, it will perform SIC, decoding the data of user 1 and then subtracting that data from the linear superimposed signal ($y_2 = h_1x_1 + h_2x_2 + n_1$). Finally, it decodes its own data. Since user 1 will have the worst channel, it will just treat user 2's signal as interference and decodes its own data from the superimposed signal ($y_2 = h_1x_1 + h_2x_2 + n_2$). Since power is being shared between the users, it has $P = P_1 + P_2$ and the following rate pairs can be achieved:

$$\begin{aligned} R_1 &= \log_2(1 + \frac{P_1 |h_1|^2}{P_2 |h_1|^2 + N_0}), \\ R_2 &= \log_2(1 + \frac{P_2 |h_2|^2}{N_0}). \end{aligned} \tag{2.20}$$

It is now important to compare these rates with the rates that can be achieved by orthogonal schemes in order to understand the advantage of NOMA. If it considers that orthogonal schemes allocate a fraction β of resources to user 1 and $1 - \beta$ to user 2 (it is irrelevant if the resources are time or frequency), and that the power is also split as

$P = P_1 + P_2$, the following rate pairs can be achieved by orthogonal schemes:

$$\begin{aligned} R_1 &= \beta \log_2(1 + \frac{P_1 |h_1|^2}{\beta N_0}) \\ R_2 &= (1-\beta) \log_2(1 + \frac{P_2 |h_2|^2}{(1-\beta) N_0}). \end{aligned} \tag{2.21}$$

For generic SNR, the boundaries of the rate regions can be seen in Figure 2.10.

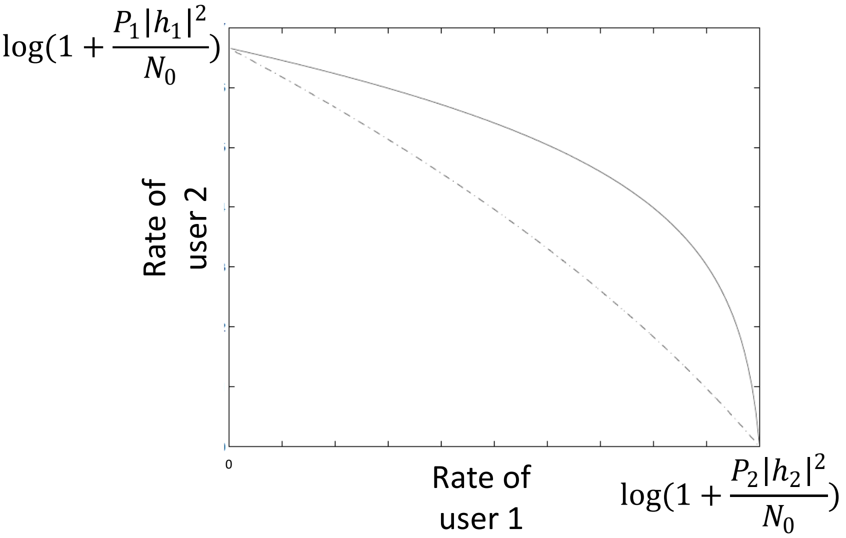

Figure 2.10: Two user downlink asymmetric AWGN rates for orthogonal and non-orthogonal schemes. Superposition coding in solid line and orthogonal schemes in dashed line.[22]

The dashed straight line is easy to understand, as β variates, it goes linearly from point $\log_2(1 + \frac{P_1|h_1|^2}{\beta N_0})$ to the point $\log_2(1 + \frac{P_2|h_2|^2}{(1-\beta) N_0})$. With SC this rate is able to improve due to the same reasons that allow user 2 to achieve its point-to-point bound while user 1 has a non-zero rate in Figure 2.9, namely, user 2 is able to decode its signal without interference from user 1. Figure 2.10 proves that for any rate pair achieved by orthogonal schemes, there exists a SIC scheme that achieves better performance (except for the two corners points where only one user is communicating).

The difference in performance is larger when the two SIC users have very different channel coefficients because orthogonal schemes normally have to allocate a great number of resources to the weak user, causing a degradation in the strong user, while with SIC what is required is that the channel coefficients are very different.

As in the uplink case, one can generalise the expressions for the general case, with K_u users, and the boundaries of the capacity region of the downlink AWGN channel can be written as:

$$R_k = \log_2(1 + \frac{P_k \left|h_k\right|^2}{(\sum_{j=k+1}^{K_u} P_j) \left|h_k\right|^2 + N_0}), \tag{2.22}$$

where $\sum_{k=1}^{K_u} P_k$ is the power splits between the users. It is interesting to analyse that, in the uplink case, the sum capacity is achieved when all the users transmit simultaneously and with equal power, while in the downlink case, looking at (2.22) it can be seen that the sum capacity is now maximised by allowing only the user with higher SNR to transmit. This is important to warn that the sum capacity in the downlink case cannot be the sole criteria by which the power allocation is guided. Although the water-filling idea is optimal from the downlink system's total throughput, it may lead to highly unfair allocation of the system's capacity to users with bad channels. In [32] the authors tackled this issue imposing restrictions in the power allocation algorithm. In this thesis users with poor CSI will get more transmission power, ensuring that they can detect their messages directly by treating the other user's information as noise, as in [33] and maintaining fairness between users.

2.3.3 The Limitations of the successive interference cancellation Receiver

SIC plays an important role in achieving the rate results in Figures 2.9 and 2.10 and will be the decoding method implemented in chapter 3. However, since its inception, SIC is known to have some problems [22]:

- Complexity scaling with the number of users: although in the uplink every access scheme has to decode the signal from every user in the cell (and SIC is no exception), in the downlink, normally each user only has to decode its own information. With SIC, every user has to decode not only its own signal but also the signal from every user with a higher channel gain (or power allocation coefficient, in chapter 3). This means that the decoding complexity grows linearly with the number of users, which may result in a significant delay. In our model this will be mitigated with the use of clusters of users, although no measure of the delay improvement will be done, nor any other delay analysis.

- Error propagation: errors may happen in the system, because of noise. When an error occurs in a decoding chain, lets say for user i, all the users that come later in the decoding chain are likely to be decoded incorrectly as well. If the probability of user i being decoded incorrectly is P^i_{error}, assuming that previous users are decoded correctly, it is known that the error probability of the k-th user in the detection chain is

$$\sum_{i=1}^{k} P^i_{error}, \tag{2.23}$$

with the sum being limited by 1. This means that error propagation will affect the error probability, in the worst case, by a factor of K_u number of users, with K_u being the total number of users in the system. This effect will appear in our model and plays a major role in the behaviour of the SER curves.

- Imperfect channel state estimation. Note that in chapter 3, the techniques used allow the SIC decoding process to be independent from the channel state coefficients. However, SIC can also be performed on a signal $s = \sum_{i=1}^{K_u} h_i x_i$, where h_i is the i-th channel state coefficient. This involves, of course, in the uplink case, a feedback of the CSI to the BS. This feedback is not error free, which will then create residual errors in the decoding chain, degrading the performance. As said earlier, this will not be considered in chapter 3.

- Analog-to-digital quantisation error. Since the received powers of different users can be very distinct, the dynamic range of the analogue-to-digital (A/D) converter needs to be very large. Additionally, the resolution of the quantisers needs to be high enough to represent the weakest signal, otherwise, some quantization errors will appear, degrading the performance of the system. This issue will also not be considered in the analysis contained in chapter 3.

Chapter 3

NOMA-MIMO systems with and without precoding

This chapter presents and analyses a NOMA-MIMO system with users grouped in clusters and with MIMO processing removing the inter-cluster interference. Numerical results are presented for two users per cluster with different modulations and 5 users per cluster with BPSK.

3.1 Introduction

The system model that is considered in this chapter allows to assess the performance of NOMA for typical configurations of multi-user MIMO systems without precoding at the base station, which is possible independently of the NOMA power allocation policy. Users are clustered into MIMO "beams" and NOMA is used within each cluster, where SIC is used at each terminal to separate the NOMA signals. Inter-cluster interference is dealt with linear filtering at the terminals, as in MIMO multi-user systems. The intra-cluster users are ordered according to their channel power. The results 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. Some of the practical limitations of SIC are highlighted and precise examples are given. This model is similar to the one that was proposed by Ding et al. in [34], which has also been recently analysed in [21]. To the best of our knowledge, comprehensive simulation results are presented for the first time for multi-user NOMA systems with more than just two users.

3.2 System Model

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 [34]. To apply the NOMA concept, users will be grouped in M clusters of K users each. The BS transmits the signal x

$$\mathbf{x} = \mathbf{P}\tilde{\mathbf{s}}, \tag{3.1}$$

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 $\mathbf{S} \in \mathbb{C}^{M \times K}$.

$$\mathbf{S} = \begin{bmatrix} s_{\mathrm{1,1}}, \cdots, s_{\mathrm{1,K}} \\ \vdots \\ s_{\mathrm{M,1}}, \cdots, s_{\mathrm{M,K}} \end{bmatrix}. \tag{3.2}$$

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:

$$\tilde{\mathbf{s}} = \begin{bmatrix} \alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,K}s_{1,K} \\ \vdots \\ \alpha_{M,1}s_{M,1} + \cdots + \alpha_{M,K}s_{M,K} \end{bmatrix}, \tag{3.3}$$

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 $\alpha^2_{m,k} \in [0, 1]$ 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) (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 [34]:

$$\sum_{k=1}^{K} \alpha^2_{m,k} = 1. \tag{3.4}$$

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:

$$\mathbf{y}_{1,\mathrm{k}} = \mathbf{H}_{1,\mathrm{k}}\tilde{\mathbf{s}} + \mathbf{n}_{1,\mathrm{k}}, \tag{3.5}$$

where $\mathbf{H}_{1,\mathrm{k}} \in \mathbb{C}^{N \times M}$ is the Rayleigh fading matrix from the BS to the k-th user in the first cluster and $\mathbf{n}_{1,\mathrm{k}} \sim \mathcal{CN}(0, \sigma_n^2) \in \mathbb{C}^{1 \times K}$ 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:

$$\mathbf{H}_{1,1} = \begin{bmatrix} \mathbf{h}_{_{1,1}}, \cdots, \mathbf{h}_{_{1,\mathrm{M}}} \\ \vdots \\ \mathbf{h}_{_{\mathrm{N},1}}, \cdots, \mathbf{h}_{_{\mathrm{N},\mathrm{M}}} \end{bmatrix}. \tag{3.6}$$

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

$$\mathbf{v}_{1,\mathrm{k}}^{\mathrm{H}}\mathbf{y}_{1,\mathrm{k}} = \mathbf{v}_{1,\mathrm{k}}{}^{\mathrm{H}}\mathbf{H}_{1,\mathrm{k}}\tilde{\mathbf{s}} + \mathbf{V}_{1,\mathrm{k}}^{\mathrm{H}}\mathbf{n}_{1,\mathrm{k}}, \tag{3.7}$$

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

$$\begin{aligned} &\mathbf{v}_{1,\mathrm{k}}^{H}\mathbf{y}_{1,\mathrm{k}} = \\ &= \mathbf{v}_{1,\mathrm{k}}^{\mathrm{H}}\mathbf{H}_{1,\mathrm{k}}(\alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,\mathrm{K}}s_{1,\mathrm{K}}) + \sum_{m=2}^{M}\mathbf{v}_{1,\mathrm{k}}^{H}\mathbf{H}_{1,\mathrm{k}}\tilde{\mathbf{s}}_{l} + \mathbf{v}_{1,\mathrm{k}}^{H}\mathbf{n}_{1,\mathrm{k}}, \end{aligned} \tag{3.8}$$

where $\tilde{\mathbf{s}}_m \in \mathbb{C}^{1\times 1}$ denotes the contribution of cluster m to the $\tilde{\mathbf{s}}$ vector. The aim is to eliminate the inter cluster interference $\sum_{m=2}^{M}\mathbf{v}_{1,\mathrm{k}}^{H}\mathbf{H}_{1,\mathrm{k}}\tilde{\mathbf{s}}_m$ in the first cluster, such that NOMA detection can be performed on the remaining signal. In short, the problem amounts to having:

$$\mathbf{v}_{\mathrm{m,k}}^{H}\mathbf{H}_{\mathrm{i,k}} = 0, \tag{3.9}$$

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

$$\mathbf{v}_{\mathrm{m,k}}^{H}\underbrace{\left[\mathbf{h}_{1,\mathrm{i}k}\cdots\mathbf{h}_{\mathrm{m}-1,\mathrm{i}k}\,\mathbf{h}_{\mathrm{m}+1,\mathrm{i}k}\cdots\mathbf{h}_{\mathrm{M},\mathrm{i}k}\right]}_{\tilde{\mathbf{H}}_{\mathrm{i,k}}} = 0, \tag{3.10}$$

where $\mathbf{h}_{m,ik} \in \mathbb{C}^{N\times 1}$ is the m-th column of the $\mathrm{H}_{i,k}$ matrix. It is clear from equation (3.10) that $\mathbf{v}_{\mathrm{m,k}}^{H} \in \mathbb{C}^{N\times 1}$ must belong to a space that is orthogonal to $\tilde{\mathbf{H}}_{\mathrm{i,k}}$. Let us expand the matrix $\tilde{\mathbf{H}}_{\mathrm{m,k}}$ into its SVD decomposition for the case $M = N$:

$$\tilde{\mathbf{H}}_{\mathrm{i,k}} = \begin{bmatrix} U_{1,1} & U_{1,2} & \dots & U_{1,N-1} & U_{1,N} \\ \vdots & & & & \vdots \\ U_{N,1} & U_{N,2} & \dots & U_{N,N-1} & \underbrace{U_{N,N}}_{\tilde{\mathbf{U}}_{\mathrm{i,k}}} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ \dots & & & & \\ 0 & 0 & \dots & \lambda_{min(M,N)} & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix} \mathbf{V}^{\mathbf{T}} \tag{3.11}$$

Note that the matrix with the eigenvalues of $\tilde{\mathbf{H}}_{\mathrm{i,k}}$ has a row of zeros at the bottom. This happens because, even if $M = N$, after removing a column from $\mathbf{H}_{\mathrm{m,k}}$ to create $\tilde{\mathbf{H}}_{\mathrm{i,k}}$, the matrix becomes tall and, after the SVD decomposition, it always leads to an eigenvalues matrix that has at least one row of zeros at the bottom. In general, there will be $(M - N) + 1$ rows of zeros in the eigenvalues matrix. Note that the column highlighted in (3.11) (which in the general case is a matrix), $\tilde{\mathbf{U}}_{\mathrm{i,k}} \in \mathbb{C}^{N \times (N-M-1)}$, does not contribute at all to $\tilde{\mathbf{H}}_{\mathrm{i,k}}$ since it is multiplied by the row of zeros, thus, it spans in fact a space orthogonal to $\tilde{\mathbf{H}}_{\mathrm{i,k}}$. Now, one could use $\mathbf{v}_{\mathrm{m,k}} = \frac{\tilde{\mathbf{U}}_{\mathrm{i,i}}}{\|\tilde{\mathbf{U}}_{\mathrm{i,k}}\|}$ as the detection matrix, but based on the maximum ratio combining (MRC) approach, it can project the $\mathbf{h}_{\mathrm{m,i}k}$ column onto the orthogonal space using a projection matrix $\mathbf{P}_U = \tilde{\mathbf{U}}_{\mathrm{i,k}} \tilde{\mathbf{U}}_{\mathrm{i,k}}^H$, choosing:

$$\mathbf{v}_{\mathrm{m,k}} = \tilde{\mathbf{U}}_{\mathrm{i,k}} \frac{\tilde{\mathbf{U}}_{\mathrm{i,k}}^H \mathbf{h}_{\mathrm{m,i}k}}{\|\tilde{\mathbf{U}}_{\mathrm{m},i,\mathrm{k}}^H h_{\mathrm{m,i}k}\|}. \tag{3.12}$$

Note that this MRC is done per cluster (maximising the SNR at one single antenna) and the inter-cluster interference is eliminated because (3.12) fulfils the requirement of (3.9).

This matrix is the reason why $N \geq M$ antennas are needed at each user, otherwise the $\tilde{\mathbf{H}}_{\mathrm{i,k}}$ matrix is fat instead of being tall and thus there is no orthogonal space spanned by the columns of $\mathbf{U}_{\mathrm{m,k}}$, the left matrix in (3.11). 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:

$$\|\mathbf{v}_{1,1}^{H}\mathbf{H}_{1,1}\|^2 \geq \cdots \geq \|\mathbf{v}_{1,\mathrm{k}}^{H}\mathbf{H}_{1,\mathrm{k}}\|^2, \tag{3.13}$$

which is equivalent to sorting the power allocation coefficients as:

$$\alpha_{1,1} \leq \cdots \leq \alpha_{1,\mathrm{k}}. \tag{3.14}$$

It should note that the norms in (3.13) 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:

$$\begin{aligned} \tilde{\mathbf{y}}_{1,\mathrm{k}} &= (\mathbf{v}_{1,\mathrm{k}}^{\mathrm{H}}\mathbf{H}_{1,\mathrm{k}})^{-1}\mathbf{v}_{1,\mathrm{k}}^{\mathrm{H}}\mathbf{H}_{1,\mathrm{k}}(\alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,\mathrm{K}}s_{1,\mathrm{K}}) + (\mathbf{v}_{1,\mathrm{k}}^{\mathrm{H}}\mathbf{H}_{1,\mathrm{k}})^{-1}\mathbf{v}_{1,\mathrm{k}}^{H}\mathbf{n}_{1,\mathrm{k}} = \\ &= (\alpha_{1,1}s_{1,1} + \cdots + \alpha_{1,\mathrm{K}}) + (\mathbf{v}_{1,\mathrm{k}}^{\mathrm{H}}\mathbf{H}_{1,\mathrm{k}})^{-1}\mathbf{v}_{1,\mathrm{k}}^{H}\mathbf{n}_{1,\mathrm{k}} \end{aligned} \tag{3.15}$$

The system model is depicted in Figure 3.1.

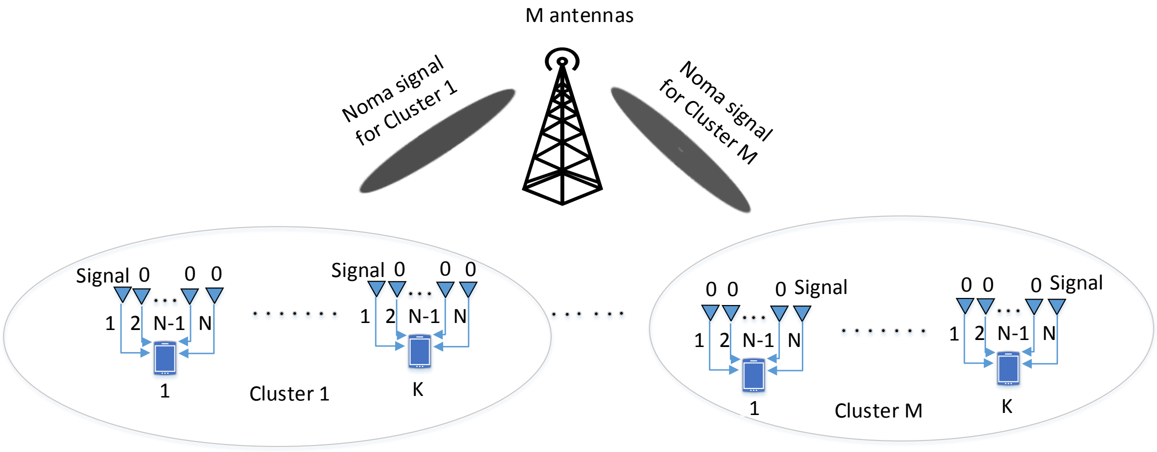

Figure 3.1: Proposed MIMO-NOMA system model.

3.3 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 3.4, 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 $\alpha_{\mathrm{m},1} = \sqrt{\frac{1}{6}}, \alpha_{\mathrm{m},2} = \sqrt{\frac{1}{3}}$ and $\alpha_{\mathrm{m},3} = \sqrt{\frac{1}{2}}$ 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 $\alpha_{\mathrm{m},1} = \sqrt{\frac{1}{4}}$ and $\alpha_{\mathrm{m},2} = \sqrt{\frac{3}{4}}$, 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.

The concept of superposition of two constellations, each one scaled by a power coefficient, was illustrated in Figure 2.7.

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 α_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:

$$\alpha_{m,k} > \sum_{k=1}^{K-1} \alpha_{m,k}, \tag{3.16}$$

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:

$$\alpha_{m,k-1} = 0.5 \times \alpha_{m,k}, \tag{3.17}$$

and since $\sum_{k=1}^{\mathrm{N}} (1/2)^k$ (which is the geometric progression of ratio 1/2 deprived from its first term) tends to 1 as N tends to infinity, the restriction (3.16) 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 [35] (and thus not taking fairness into equation).

3.4 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 3.2, 3.3, 3.4 and 3.5.

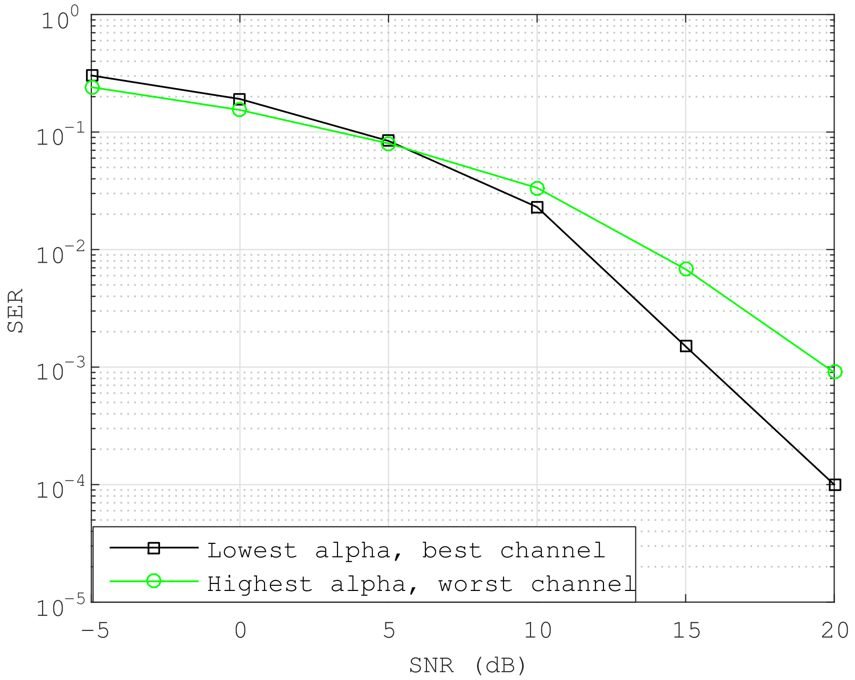

Figure 3.2: SER curves for two users per cluster with BPSK modulation. M=2, N=3 and K=2.

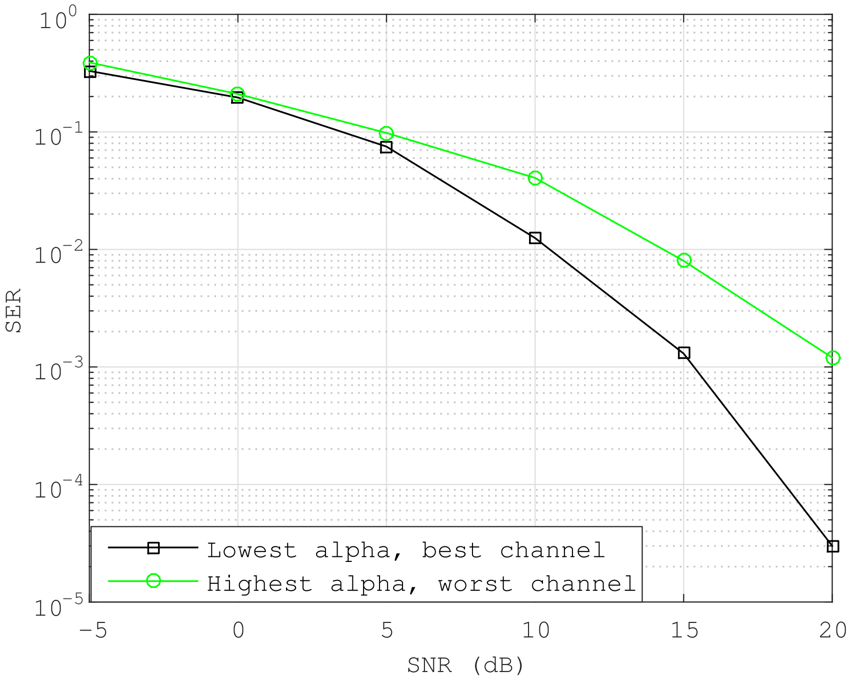

Figure 3.3: SER curves for two users per cluster with BPSK modulation in the first user and 4-QAM modulation in the second user. M=2, N=3 and K=2.

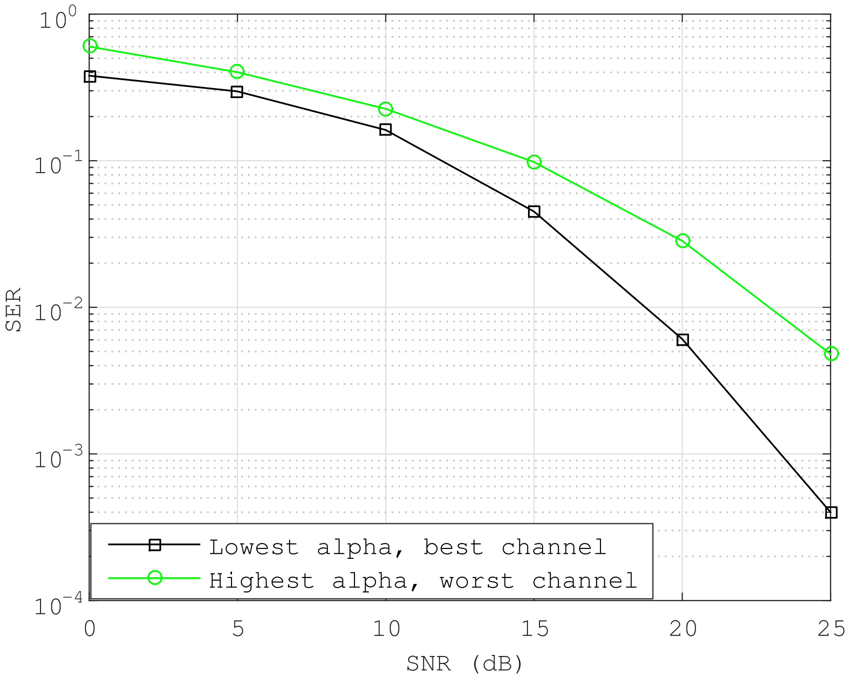

Figure 3.4: SER curves for two users per cluster with BPSK modulation in the first user and 16-QAM modulation in the second user. M=2, N=3 and K=2.

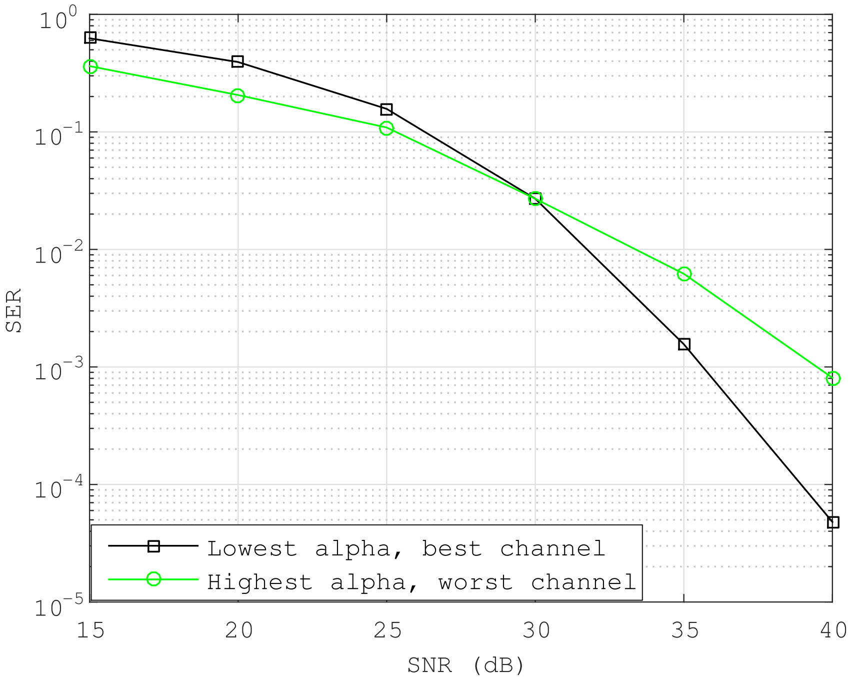

Figure 3.5: SER curves for two users with 16-QAM modulation in the first user and 64-QAM modulation in the second user. M=2, N=3 and K=2.

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 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. From equation These limitations due to this interference and noise explain 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 3.2 and 3.5, 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 3.3 and 3.4), 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 3.2 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 3.2, 3.3 and 3.4, $\alpha_1 = \sqrt{\frac{1}{4}}$ and $\alpha_2 = \sqrt{\frac{3}{4}}$ were used in order to compare with the results in Figure 1 in [34]. (Later, in Figures 3.8 and 3.9, the same power coefficients were used to evaluate the gain of using relaying, but only the users with lower power allocation coefficients were compared for Figure 3.8 and with higher power allocation coefficients for Figure 3.9). In Figure 3.5, it has $\alpha_1 = \sqrt{\frac{1}{17}}$ and $\alpha_2 = \sqrt{\frac{16}{17}}$ which were used to comply with restrictions (3.4) and (3.16). Comparing the first simulation with Figure 1 in [34], 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 3.6. The users that are close to the base station (and thus having a better channel

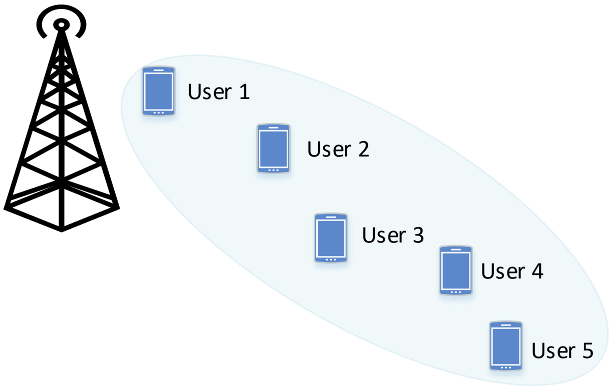

Figure 3.6: A five user MIMO-NOMA system with users sorted according to their channel gain.

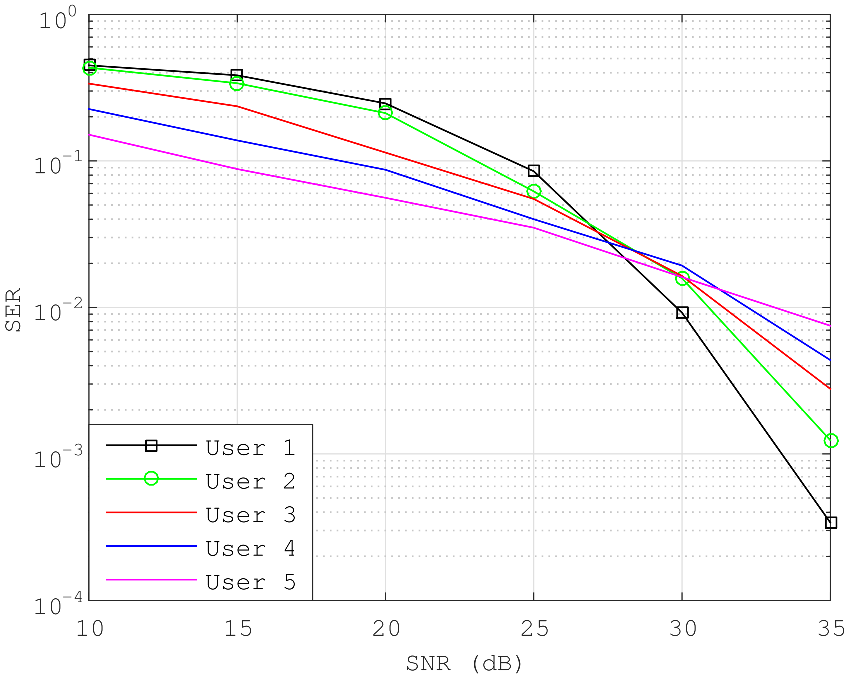

Figure 3.7: SER curves for five users per cluster with BPSK modulation. M=2, N=2 and K=5. User 5 has the highest power allocation coefficient and the lowest channel gain. User 1 has the lowest power allocation coefficient and the largest channel gain. ($\alpha_{\mathrm{m},1} = 0.0542, \alpha_{\mathrm{m},2} = 0.1083, \alpha_{\mathrm{m},3} = 0.2166, \alpha_{\mathrm{m},4} = 0.4332, \alpha_{\mathrm{m},5} = 0.8664$).

coefficient) being numbered from 1 to 5 and having a lower power allocation coefficient to maintain fairness. In Figure 3.7, 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 $\alpha_{\mathrm{m,k}}$ were obtained by using the set $\{1, 2, 4, 8, 16\}$, normalized by its sum $\sqrt{341}$, in order to comply with equation (3.4). With six users and similar power allocations coefficients, $\alpha_{\mathrm{m,1}}$ 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.

NOMA lends itself to using relaying between the chain of users, with the ones with a better channel able to relay to the ones with less favourable channels. The idea of user pairing with relaying was firstly introduced in [36], where users with better channel coefficients (that have to decode both their own signal and the other user signal) act as a relay for users with lower channel coefficients.

The results in Figure 3.8 firstly analyse the existence of such relaying mechanism but now with the users with higher power allocation coefficients relaying to users with lower power allocation coefficients. The relaying process is assumed to be error-free. Looking at Figure 3.2, one would think that this relaying would be at least beneficial in the low SNR, however, this relay actually increases the SER of lower power allocation coefficient users at every SNR. This is a proof that as long as the signal is decodable ($\alpha_2 > \alpha_1$, for the 2 user BPSK case), it is always beneficial to decode the signal with a better channel. The results in Figure 3.9 (where the SER of users with lower channel coefficients are shown) assess the relay mechanism suggested in [36], with the users with better channel coefficients relaying to users with lower channel coefficients. There exists a very clear performance improvement when doing that, with a ≈ 7 dB improvement for $SER \approx 10^{-3}$.

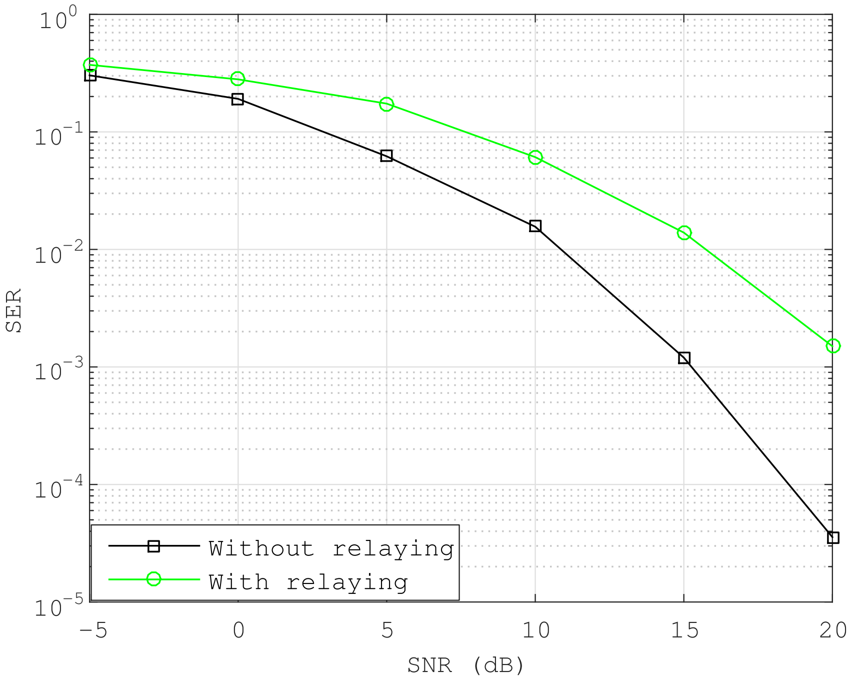

Figure 3.8: SER curves for the users with lower power allocation coefficients, in a two users per cluster NOMA system, with BPSK modulation, for the cases with and without relaying, where the relaying takes place from users with higher power allocation coefficients to the ones with lower power allocation coefficients. M=2, N=3 and K=2.

3.5 Massive MIMO Model

The previous model could not use massive MIMO at the BS due to the restrictions imposed by (3.12), 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 [37]. 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

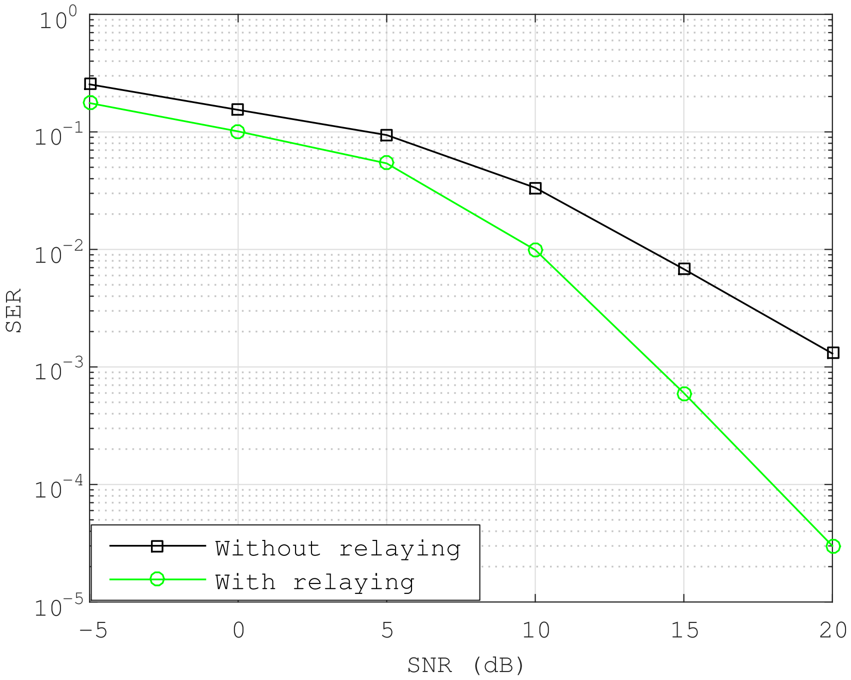

Figure 3.9: SER curves for the users with higher power allocation coefficients, in a two users per cluster NOMA system, with BPSK modulation, for the cases with and without relaying, where the relaying takes place from users with lower power allocation coefficients to the ones with higher power allocation coefficients. M=2, N=3 and K=2.

sharing the same spatial correlation matrix, denoted by $\mathbf{R}_l$. Using the Karhunen-Loève decomposition [38, 39], the k-th user in the l-th cluster can have its channel matrix decomposed as:

$$\mathbf{H}_{\mathrm{l,k}} = \mathbf{G}_{\mathrm{l,k}} \Lambda_{\mathrm{l}}^{\frac{1}{2}} \mathbf{U}_{\mathrm{l}}, \tag{3.18}$$

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

$$\mathbf{R}_l = \mathbf{U}_{\mathrm{l}}^{H} \Lambda_{\mathrm{l}} \mathbf{U}_{\mathrm{l}} = \mathbb{E}\{\mathbf{H}_{\mathrm{l,k}}^{H} \mathbf{H}_{\mathrm{l,k}}\}, \tag{3.19}$$

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:

$$\Lambda_\mathrm{l} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & 0 \\ \dots & & & & & \\ 0 & 0 & \dots & \lambda_{M-r_k,M-r_k} & 0 & 0 \\ 0 & 0 & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & 0 & \lambda_{M,M} \end{bmatrix}, \tag{3.20}$$

and thus can be reduced to a $r_l \times r_l$ matrix, making $\mathbf{G}_{\mathrm{l,k}}$ a $M \times r_l$ matrix and $\mathbf{U}_\mathrm{l}$ a $r_l \times M$ matrix. The Karhunen-Loève decomposition is useful because, while the CSI-T concerning the fast fading matrix $\mathbf{G}_{\mathrm{l,k}}$ 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

$$\mathbf{S} = \sum_{l=1}^{L} \mathbf{P}_l \sum_{k=1}^{K} w_l \alpha_{l,k} s_{l,k}, \tag{3.21}$$

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:

$$\mathbf{y}_{\mathrm{l,k}} = \mathbf{G}_{\mathrm{l,k}}\Lambda_{\mathrm{l}}^{\frac{1}{2}}\mathbf{U}_{\mathrm{l}}\sum_{l=1}^{L}\mathbf{P}_l\sum_{k=1}^{K}w_l\alpha_{l,k}s_{l,k} + n_{l,k}, \tag{3.22}$$

where $n_{l,k}$ is the noise value for the k-th user in the l-th cluster. By looking at equation (3.22), the precoding matrix $\mathbf{P}_l$ will need to satisfy the following constrain to eliminate inter cluster interference:

$$[\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H\mathbf{U}_{l+1}^H \cdots \mathbf{U}_{\mathrm{L}}^H]^H\mathbf{P}_l = 0. \tag{3.23}$$

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

$$\mathbf{P}_l = \mathrm{Null}([\mathbf{U}_1^H \cdots \mathbf{U}_{l-1}^H\mathbf{U}_{l+1}^H \cdots \mathbf{U}_{\mathrm{L}}^H]^H). \tag{3.24}$$

Using $\mathbf{P}_l$ in (3.24), (3.22) can be simplified to:

$$\mathbf{y}_{\mathrm{l,k}} = \mathbf{G}_{\mathrm{l,k}}\Lambda_{\mathrm{l}}^{\frac{1}{2}}\mathbf{U}_{\mathrm{l}}\mathbf{P}_l\sum_{k=1}^{K}w_l\alpha_{l,k}s_{l,k} + n_{l,k}. \tag{3.25}$$

Let us specify (3.25) for the case of $l = 1$ and $k = 2$ and analyse the signal received by the first user:

$$\mathbf{y}_{1,1} = \mathbf{G}_{1,1}\Lambda_1^{\frac{1}{2}}\mathbf{U}_1\mathbf{P}_1\mathbf{w_1}(\alpha_{1,1}s_{1,1} + \alpha_{1,2}s_{1,2}) + n_{1,1}. \tag{3.26}$$

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

$$[\alpha_{1,1}s_{1,1} + \alpha_{1,2}s_{1,2} \quad 0 \cdots 0]^T, \tag{3.27}$$

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. This can be written as:

$$\begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,\tilde{M}-1} & c_{1,\tilde{M}} \\ \vdots & & & & \vdots \\ c_{N,1} & c_{N,2} & \cdots & c_{N,\tilde{M}-1} & c_{N,\tilde{M}} \end{bmatrix} \begin{bmatrix} \alpha_{1,1}s_{1,1} + \alpha_{1,2}s_{1,2} \\ \vdots \\ 0 \end{bmatrix} + n_{1,1}, \tag{3.28}$$

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:

$$\begin{aligned} &\tilde{\mathbf{y}}_{1,1} = (\mathbf{G}_{1,1}\Lambda_1^{\frac{1}{2}}\mathbf{U}_1\mathbf{P}_1\mathbf{w_1})^{-1}[\mathbf{G}_{1,1}\Lambda_1^{\frac{1}{2}}\mathbf{U}_1\mathbf{P}_1\mathbf{w_1}(\alpha_{1,1}s_{1,1} + \alpha_{1,2}s_{1,2}) + n_{1,1}] = \\ &(\alpha_{1,1}s_{1,1} + \alpha_{1,2}s_{1,2}) + (\mathbf{G}_{1,1}\Lambda_1^{\frac{1}{2}}\mathbf{U}_1\mathbf{P}_1\mathbf{w_1})^{-1}n_{1,1}. \end{aligned} \tag{3.29}$$

Comparing figures 3.10 and 3.11 with figures 3.2 and 3.4, one can see that they are identical. To understand why this happens one needs to look at equations (3.29) and (3.15). Noting that $\mathbf{G}_{1,1}\Lambda_1^{\frac{1}{2}}\mathbf{U}_1 = \mathbf{H}_{1,1}$ (Karhunen-Loève decomposition) and that $\|\mathbf{v}_{\mathrm{m,k}}\| = \|\mathbf{P}_1\| = 1$, one sees that both equations are equivalent in terms of the ratio between the signal power and noise power.

From one model to the other the major difference is that with the massive MIMO model there is no limitation to the number of transmit antennas. Remember that the model without precoding and with conventional MIMO had the number of cluster limited by the number of receive antennas (so at maximum we could have 8 clusters if we consider the limit of $N = 8$). In the massive MIMO model the maximum number of clusters is limited by $\tilde{M}_l = (M - r_l(L-1))$, because of the precoding, which with an increasing number of antennas at the BS, can lead and an arbitrarily large number of clusters.

With the channel coefficients coming from a Gaussian distribution we have $r_l = N$, so if we consider the case of $N = 8$ and $M = 128$ ($N = 8$ is the maximum number of antennas considered in LTE-A and $M = 128$ is a number being studied for the massive MIMO arrays, although this number may reach several hundreds), the maximum number of clusters would be $L = 15$ (we have the condition $\tilde{M} < M$), with $\tilde{M} = 16$, which means that the massive MIMO model can duplicate the number of clusters. With a more conservative number of user antennas, $N = 1$, one could get $\tilde{M} = 26$ and $L = 25$.

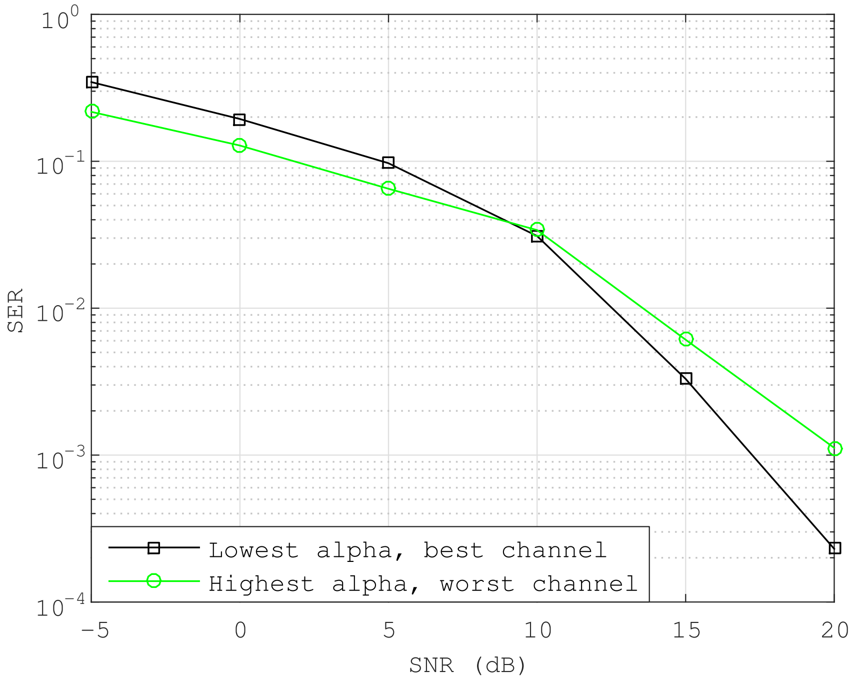

Figure 3.10: SER curves for two users with BPSK modulation with the massive MIMO system. M=50, N=3 and K=2.

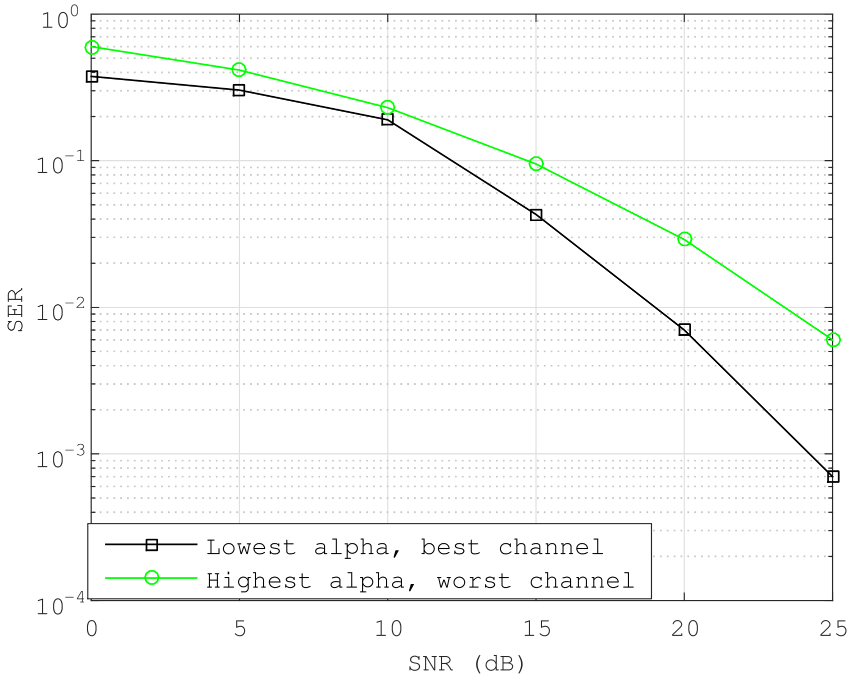

Figure 3.11: SER curves for two users with BPSK modulation in the first user and 16-QAM modulation in the second user with the massive MIMO system. M=50, N=3 and K=2.

Chapter 4

Rate and capacity analysis

In this chapter a series of simulations for the achievable rates of the MIMO-NOMA model without precoding of the last chapter and also a MIMO-OMA counterpart model will be performed, both for the two and for the five user cases. A comparison with the results for the rates presented in chapter 2 will be made and the performance of the new simulations will be evaluated.

The conclusion that NOMA outperforms than current orthogonal schemes in terms of rate has already been given in 2.3.2 and 2.3.1, based on the results from [22]. 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 [34], the SINR for the k-th user in the m-th cluster can be written as:

$$SINR_{m,k} = \frac{\rho\|\mathbf{v}_{\mathrm{m,k}}^{H}\mathbf{H}_{\mathrm{m,k}}\|^{2}\alpha_{\mathrm{m,k}}^{2}}{\rho\sum_{l=1,l\neq k}^{K}\|\mathbf{v}_{\mathrm{m,k}}^{H}\mathbf{H}_{\mathrm{m,k}}\|^{2}\alpha_{\mathrm{m,l}}^{2} + \|\mathbf{v}_{\mathrm{m,k}}\|^{2}}, \tag{4.1}$$

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 $\|\mathbf{v}_{\mathrm{m,k}}\|^{2} = 1$ and $\sum_{l=1,l\neq k}^{K}\|\mathbf{v}_{\mathrm{m,k}}^{H}\mathbf{H}_{\mathrm{m,k}}\|^{2}\alpha_{\mathrm{m,l}}^{2} = 0$, for $k = 1$, (4.1) can be written for the case of user 1 as:

$$SINR_{m,1} = \rho\|\mathbf{v}_{\mathrm{m,1}}^{H}\mathbf{H}_{\mathrm{m,1}}\|^{2}\alpha_{\mathrm{m,1}}^{2}, \tag{4.2}$$

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

$$SINR_{m,2} = \frac{\rho\|\mathbf{v}_{\mathrm{m,2}}^{H}\mathbf{H}_{\mathrm{m,2}}\|^{2}\alpha_{\mathrm{m,2}}^{2}}{\rho\|\mathbf{v}_{\mathrm{m,2}}^{H}\mathbf{H}_{\mathrm{m,2}}\|^{2}\alpha_{\mathrm{m,1}}^{2} + 1}. \tag{4.3}$$

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:

$$R_{m,1}^{MIMO-NOMA} \leq \log_2(1 + \rho\|\mathbf{v}_{\mathrm{m,1}}^{H}\mathbf{H}_{\mathrm{m,1}}\|^{2}\alpha_{\mathrm{m,1}}^{2}), \tag{4.4}$$

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

$$R_{m,2}^{MIMO-NOMA} \leq \log_2(1 + \frac{\rho\|\mathbf{v}_{\mathrm{m,2}}^{H}\mathbf{H}_{\mathrm{m,2}}\|^{2}\alpha_{\mathrm{m,2}}^{2}}{\rho\|\mathbf{v}_{\mathrm{m,2}}^{H}\mathbf{H}_{\mathrm{m,2}}\|^{2}\alpha_{\mathrm{m,1}}^{2} + 1}). \tag{4.5}$$

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:

$$R_{m,1}^{MIMO-OMA} \leq (1-\beta)\log_2(1+\frac{(1-\gamma)\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{1-\beta}), \tag{4.6}$$

and the rate of the second user is bounded by

$$R_{m,2}^{MIMO-OMA} \leq \beta\log_2(1+\frac{\gamma\rho\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta}). \tag{4.7}$$

Now, to get the total sum-rate of the OMA systems, one has to use the Jensen's inequality. The Jensen's inequality gives a lower bound on expectations of convex functions and an upper bound on the expectations of concave functions. A function $f(x)$ is convex if, for any $0 < \lambda < 1$:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \tag{4.8}$$

or, more simply:

$$D_x^{(2)}[f(x)] \geq 0. \tag{4.9}$$

By opposition, $f(x)$ is concave if $-f(x)$ is convex. Now, let us suppose that X is a random variable with expectation μ and function $f(x)$ is convex and finite. In short the Jensen's inequality, for the case of convex functions, states that

$$\mathbb{E}_x[f(X)] \geq f(\mathbb{E}_x[X]), \tag{4.10}$$

with equality if and only if, $f(x)$ is linear. Analogously, for a concave function, $\mathbb{E}_x[f(X)] \leq f(\mathbb{E}_x[X])$. Now, since $D_x^{(2)}[log(x)] = -\frac{1}{x^2} < 0$, it can finally state that:

$$\mathbb{E}_x[\log(X)] \leq f(\mathbb{E}_x[X]). \tag{4.11}$$

In other words it means that:

$$\log(\lambda_1 x_1 + \cdots + \lambda_n x_n) \geq \lambda_1 \log(x_1) + \cdots + \lambda_n \log(x_n). \tag{4.12}$$

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

$$\begin{aligned} &R_{m,1}^{MIMO-OMA} + R_{m,2}^{MIMO-OMA} \leq \\ &\leq (1-\beta)\log_2(1 + \frac{(1-\gamma)\rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2}{1-\beta}) + \beta \log_2(1 + \frac{\gamma\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2}{\beta}) \leq \\ &\leq \log_2((1-\beta) + \beta + (1-\beta)\frac{(1-\gamma)\rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2}{1-\beta} + \beta\frac{\gamma\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2}{\beta}) = \\ &= \log_2(1 + \rho(1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 + \rho\gamma\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2), \end{aligned} \tag{4.13}$$

where the equality in the second inequality only holds if

$$\frac{\gamma\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2}{\beta} = \frac{(1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2}{1-\beta}. \tag{4.14}$$

This can be seen by defining $\frac{\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta} = \frac{(1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{1-\beta} = A.$, which allows to write:

$$\begin{aligned} & R_{m,1}^{MIMO-OMA} + R_{m,2}^{MIMO-OMA} \leq \\ & \leq (1-\beta)\log_2(1+\frac{(1-\gamma)\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{1-\beta}) + \beta\log_2(1+\frac{\gamma\rho\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta}) = \\ & = (1-\beta)\log_2(1+\rho A) + \beta\log_2(1+\rho A) = \log_2(1+\rho A). \end{aligned} \tag{4.15}$$

Taking the expression on the right-hand side of the second inequality in equation (4.13):

$$\begin{aligned} & \log_2(1+(1-\beta)\frac{(1-\gamma)\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{1-\beta} + \beta\frac{\gamma\rho\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta}) = \\ & = \log_2(1+(1-\beta)\rho A + \beta\rho A) = \log_2(1+\rho A) \end{aligned} \tag{4.16}$$

From (4.14) it can now be derived the optimal split of the resources to achieve the maximum sum-rate of MIMO-OMA:

$$\beta^* = \frac{\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2 + (1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}. \tag{4.17}$$

Under the optimal split of resources from (4.17), the channel capacities achieved by MIMO-OMA can be written as:

$$\begin{aligned} C_{\mathrm{m},1}^{MIMO-OMA*} = & \frac{(1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2 + (1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2} \times \\ & \times \log_2(1+\rho(1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2 + \rho\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2), \end{aligned} \tag{4.18}$$

$$C_{\mathrm{m},2}^{MIMO-OMA*} = \frac{\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2 + (1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}\times \\ \times \log_2(1+\rho(1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2 + \rho\gamma\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2). \tag{4.19}$$

The rates for MIMO-NOMA and MIMO-OMA as a function of the power allocation coefficient $\alpha_{\mathrm{m},2}^2 = \gamma$ and $\beta = \beta^*$ can be seen in Figures 4.1 and 4.2. Each point was simulated 10^4 times, because it was seen that with more than 10^4 simulations the points were very close to the point simulated with 10^4 simulations.

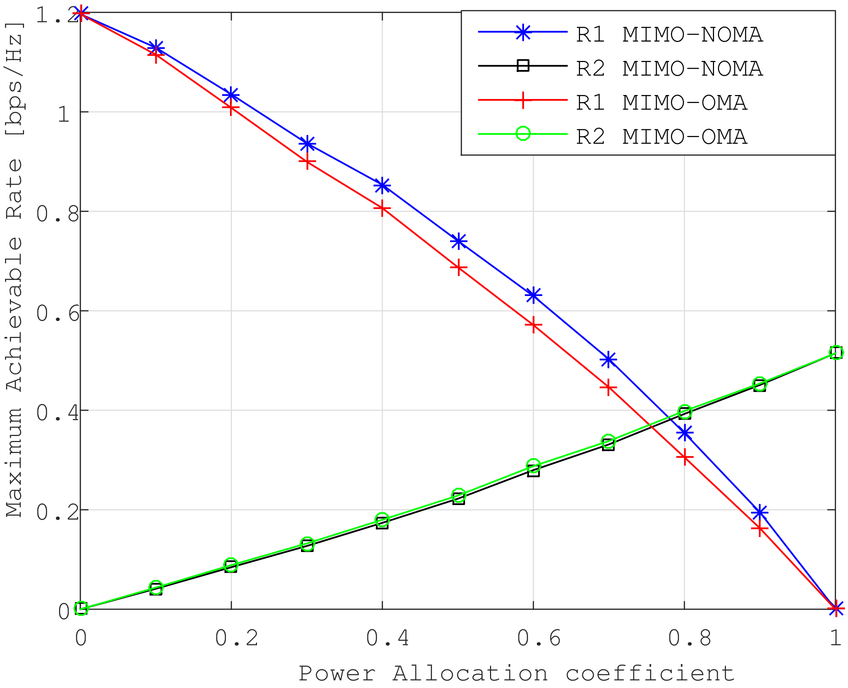

Figure 4.1: Maximum rates achieved by a MIMO-NOMA and a MIMO-OMA schemes with two users each. The SNR is $\rho = 0dB$. Power allocation coefficient $\alpha_{\mathrm{m},2}^2 = \gamma$.

This fact can be seen in Figure 4.3 were it is shown the variation of the first point of $R_{m,1}^{MIMO-OMA}$ in Figure 4.1, with the number of simulations. Also, the results from 2 sets of 10000 simulations each, varied the rates in less than 0.01%. Important conclusions can be made about the Figures 4.1 and 4.2, namely, the only situation where MIMO-NOMA

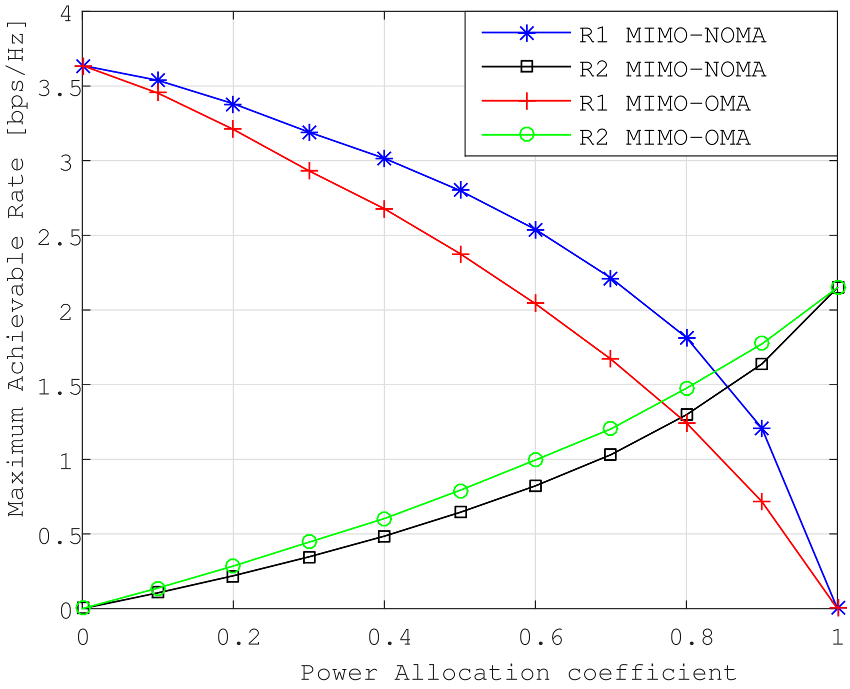

Figure 4.2: Maximum rates achieved by a MIMO-NOMA and a MIMO-OMA scheme. The SNR is $\rho = 10dB$. Power allocation coefficient $\alpha^2_{\mathrm{m},2} = \gamma$.

rates are equal to MIMO-OMA rates is when one of the users is not communicating ($\alpha^2_{\mathrm{m},2} = \gamma = 0$ or $\alpha^2_{\mathrm{m},2} = \gamma = 1$). Also, $R^{MIMO-NOMA}_{m,1} > R^{MIMO-OMA}_{m,1}$ for every $0 < \alpha^2_{\mathrm{m},2} = \gamma < 1$, 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.

Although it may be hard to see in Figure 4.1 (although it is clear in Figure 4.2), it seems odd that $R^{MIMO-OMA}_{m,2} > R^{MIMO-NOMA}_{m,2}$ for every $0 < \alpha^2_{\mathrm{m},2} = \beta = 1 < 1$, but this can be explained, by the fact that the second NOMA user $R^{MIMO-NOMA}_{m,2}$ 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.

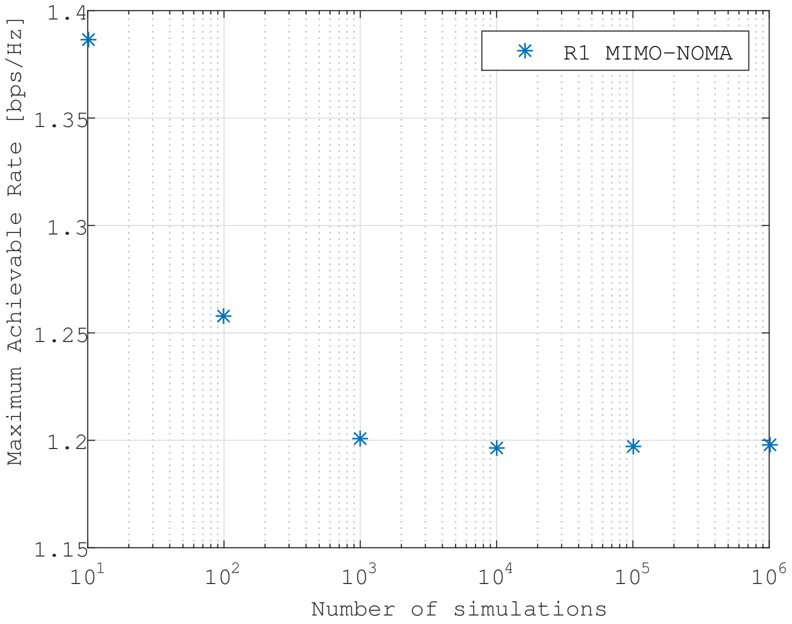

Figure 4.3: Variation of the R_1 MIMO-NOMA rate from Figure 4.1 with the number of simulations. $\alpha_{\mathrm{m},2}^2 = \gamma = 0$.

It is also noteworthy that the rates seem to grow with ρ. Looking at equations (4.4), (4.5), (4.6) and (4.7), it can be seen that ρ increases the term inside the logarithm, except in the case of (4.5), where that relation is less obvious. However, since $\alpha_{\mathrm{m},2}^2 > \alpha_{\mathrm{m},1}^2$, 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 [22], one should represent the boundaries of rate pairs achieved by MIMO-NOMA and MIMO-OMA, as it is presented in Figures 4.4 and 4.5.

Comparing the results to Figure 2.10, 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.

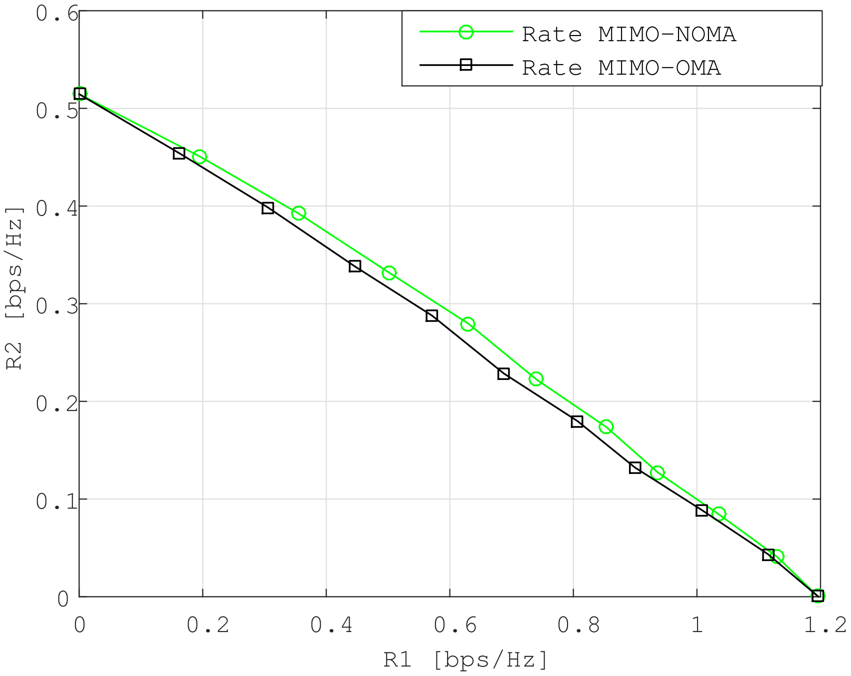

Figure 4.4: Boundary of rate pairs achieved by MIMO-NOMA and MIMO-OMA systems, each with two users. $\rho = 0dB$.

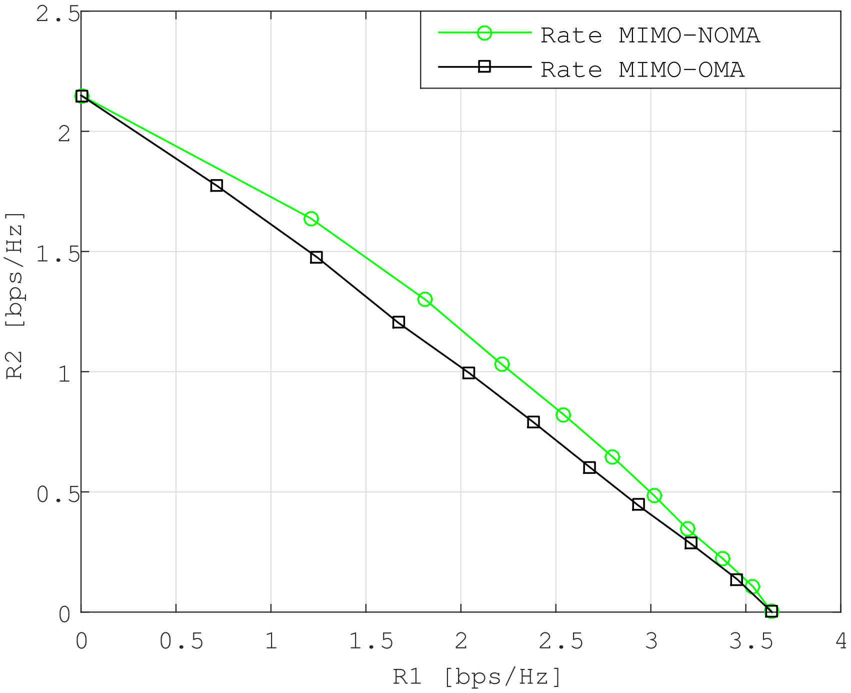

Figure 4.5: Boundary of rate pairs achieved by MIMO-NOMA and MIMO-OMA systems, each with two users. $\rho = 10dB$.

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:

$$\begin{aligned} & C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} = \\ & = \log_2(1 + \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2) + \log_2(1 + \frac{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2}{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},1}^2 + 1}), \end{aligned} \tag{4.20}$$

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

$$\begin{aligned} & C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} = \log_2(1 + \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 + \\ & + \frac{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2}{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},1}^2 + 1} + \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 \frac{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2}{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},1}^2 + 1}) = \\ & = \log_2(1 + \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 + \\ & + \frac{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2 + \rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2 \times \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2}{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},1}^2 + 1}) = \\ & = \log_2(1 + \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 + \rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2 \frac{\rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 + 1}{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},1}^2 + 1})). \end{aligned} \tag{4.21}$$

Now, from (3.13), is known that $\frac{\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2}{\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2} > 1$, and therefore:

$$\begin{aligned} & C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} = \\ & = \log_2(1 + \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 + \rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2 \frac{\rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 + 1}{\rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},1}^2 + 1}) \geq \\ & \geq \log_2(1 + \rho\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 \alpha_{\mathrm{m},1}^2 + \rho\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2 \alpha_{\mathrm{m},2}^2). \end{aligned} \tag{4.22}$$

Recalling from equation (4.13):

$$\begin{aligned} & C_{m,1}^{MIMO-OMA} + C_{m,2}^{MIMO-OMA} = \\ & = \log_2(1 + \rho(1-\gamma)\|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 + \rho\gamma\|\mathbf{v}_{\mathrm{m},2}^H \mathbf{H}_{\mathrm{m},2}\|^2), \end{aligned} \tag{4.23}$$

substituting in equation (4.22) if it agrees that the power allocation coefficients for NOMA are the same as for OMA ($\alpha_{\mathrm{m},2}^2 = \gamma$):

$$C_{m,1}^{MIMO-NOMA} + C_{m,2}^{MIMO-NOMA} \geq C_{m,1}^{MIMO-OMA} + C_{m,2}^{MIMO-OMA}, \tag{4.24}$$

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 4.6, 4.7 and 4.8.

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 $\alpha_{\mathrm{m},2}^2 = \gamma$. This is in accord with the results in section 3.4, namely, the fact that allocating too much power to the user with the best channel (in the case of Figure 4.8) 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. For comparison reasons, remember that the maximum difference in capacity between MIMO-NOMA and MIMO-OMA is obtained in Figure 4.8 for $\rho = 30$ dB

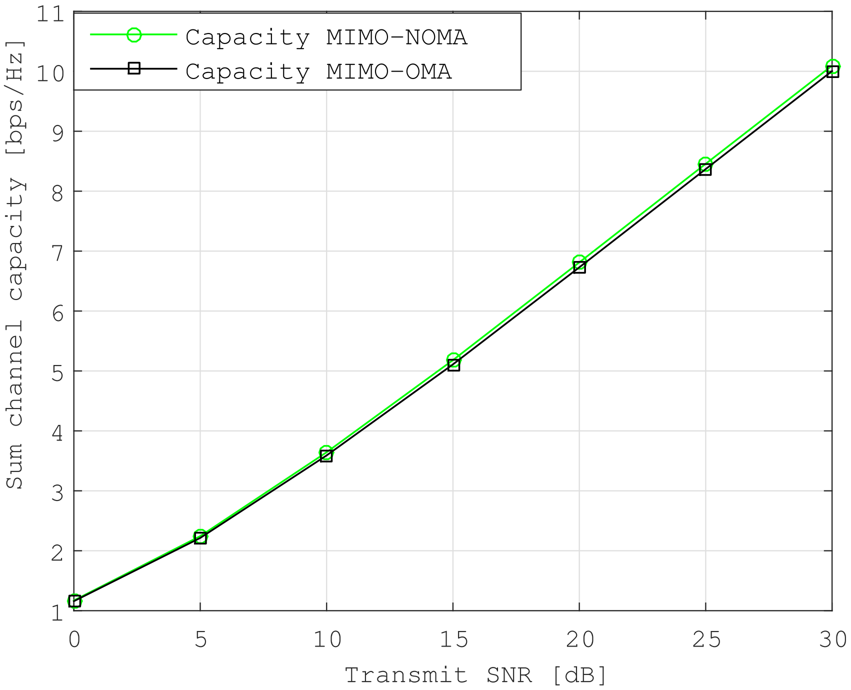

Figure 4.6: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with two users each, with $\alpha_{\mathrm{m},2}^2 = \gamma = 0.1$ and $\alpha_{\mathrm{m},1}^2 = 1 - \gamma = 0.9$.

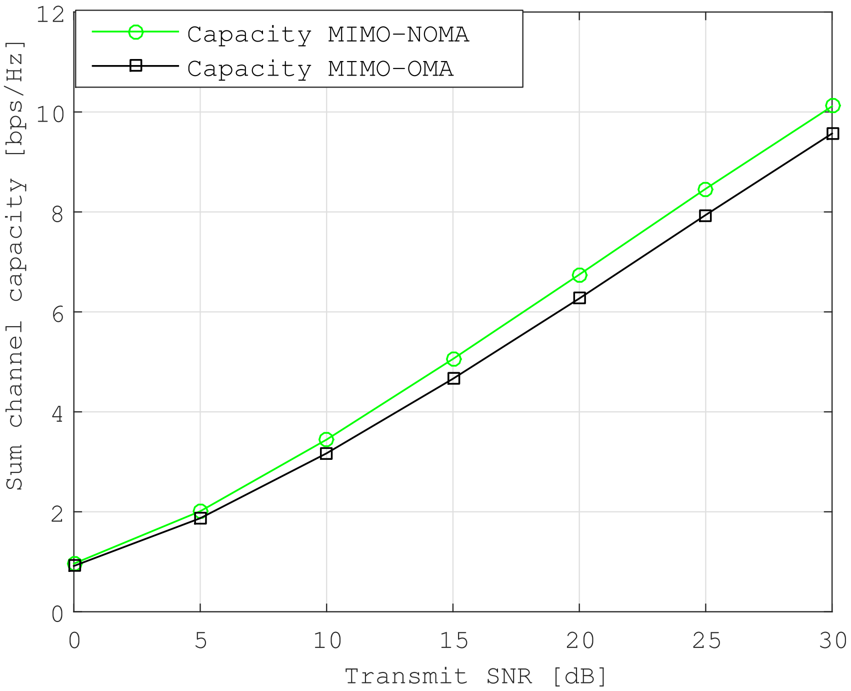

Figure 4.7: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with two users each, with $\alpha_{\mathrm{m},2}^2 = \gamma = 0.5$ and $\alpha_{\mathrm{m},1}^2 = 1 - \gamma = 0.5$.

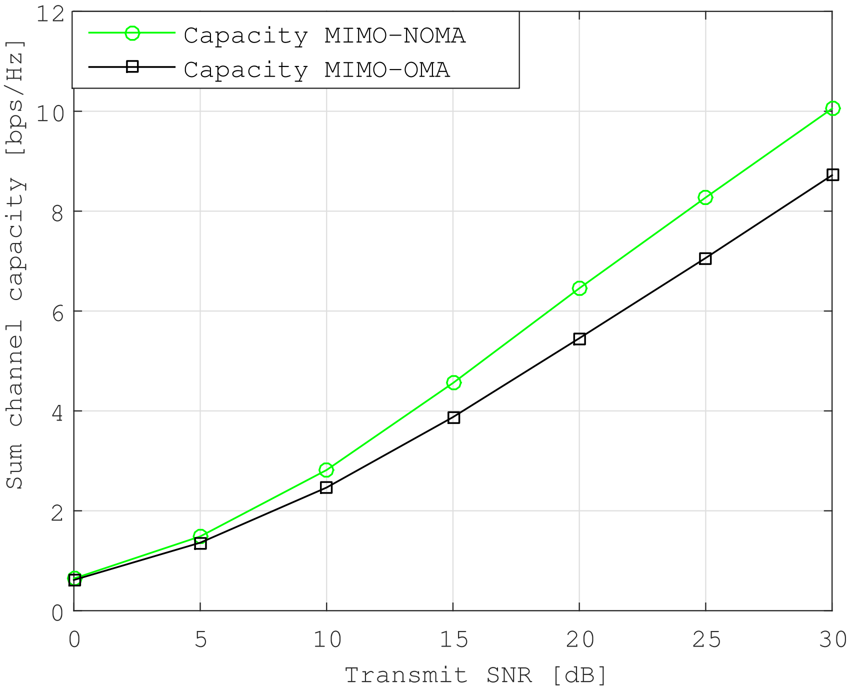

Figure 4.8: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with two users each, with $\alpha^2_{\mathrm{m},2} = \gamma = 0.9$ and $\alpha^2_{\mathrm{m},1} = 1 - \gamma = 0.1$.

and, for the reference value of 8 dB of SNR, is 3.58 dB. Also, in Figure 4.7 the maximum difference in capacity between MIMO-NOMA and MIMO-OMA, for $\rho = 30$ dB and for the reference value of 8 dB of SNR, amounts to 1.53 dB. Up until now all the results are valid for models with two users and they follow closely the work that has been done in [21]. However, in section 3.4, this system performed well ($SER < 0.5$ for $\rho = 10$ dB) in terms of SER up to five users. With this in mind, the results previously formulated for the two user case can be extended to the five user case and the performance of MIMO-NOMA compared to MIMO-OMA can be studied. It follows naturally from the previous analysis that:

$$R^{MIMO-NOMA}_{m,1} \leq \log_2(1 + \rho\|\mathbf{v}^H_{\mathrm{m},1}\mathbf{H}_{\mathrm{m},1}\|^2\alpha^2_{\mathrm{m},1}), \tag{4.25}$$

$$R^{MIMO-NOMA}_{m,2} \leq \log_2(1 + \frac{\rho\|\mathbf{v}^H_{\mathrm{m},2}\mathbf{H}_{\mathrm{m},2}\|^2\alpha^2_{\mathrm{m},2}}{\rho\|\mathbf{v}^H_{\mathrm{m},2}\mathbf{H}_{\mathrm{m},2}\|^2\alpha^2_{\mathrm{m},1} + 1}), \tag{4.26}$$

$$R_{m,3}^{MIMO-NOMA} \leq \log_2(1+\frac{\rho\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2\alpha_{\mathrm{m},3}^2}{\rho\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2\alpha_{\mathrm{m},1}^2+\rho\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2\alpha_{\mathrm{m},2}^2+1}), \tag{4.27}$$

$$\begin{aligned} R_{m,4}^{MIMO-NOMA} &\leq \\ &\leq \log_2(1+\frac{\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2\alpha_{\mathrm{m},4}^2}{\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2\alpha_{\mathrm{m},1}^2+\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2\alpha_{\mathrm{m},2}^2+\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2\alpha_{\mathrm{m},3}^2+1}), \end{aligned} \tag{4.28}$$

$$\begin{aligned} R_{m,5}^{MIMO-NOMA} &\leq \log_2(1+ \\ &+\frac{\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},5}^2}{\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},1}^2+\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},2}^2+\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},3}^2+\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},4}^2+1}), \end{aligned} \tag{4.29}$$

and:

$$R_{m,1}^{MIMO-OMA} \leq \beta_1\log_2(1+\frac{\gamma_1\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{\beta_1}), \tag{4.30}$$

$$R_{m,2}^{MIMO-OMA} \leq \beta_2\log_2(1+\frac{\gamma_2\rho\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta_2}), \tag{4.31}$$

$$R_{m,3}^{MIMO-OMA} \leq \beta_3\log_2(1+\frac{\gamma_3\rho\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2}{\beta_3}), \tag{4.32}$$

$$R_{m,4}^{MIMO-OMA} \leq \beta_4\log_2(1+\frac{\gamma_4\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2}{\beta_4}), \tag{4.33}$$

$$R_{m,5}^{MIMO-OMA} \leq \beta_5\log_2(1+\frac{\gamma_5\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2}{\beta_5}). \tag{4.34}$$

However, it is not obvious what should be the relation between $\beta_1, \beta_2, ..., \beta_5$ to achieve the maximum sum-rate of MIMO-OMA. As in the two user case, the Jensen's inequality can be used:

$$\begin{aligned}
&R_{m,1}^{MIMO-OMA} + R_{m,2}^{MIMO-OMA} + R_{m,3}^{MIMO-OMA} + R_{m,4}^{MIMO-OMA} + R_{m,5}^{MIMO-OMA} \leq \\
&\leq \beta_1 \log_2(1 + \frac{\gamma_1\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{\beta_1}) + \beta_2 \log_2(1 + \frac{\gamma_2\rho\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta_2}) + \\
&+ \beta_3 \log_2(1 + \frac{\gamma_3\rho\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2}{\beta_3}) + \beta_4 \log_2(1 + \frac{\gamma_4\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2}{\beta_4}) + \\
&+ \beta_5 \log_2(1 + \frac{\gamma_5\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2}{\beta_5}) \leq \log_2(\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta 5 + \\
&+ \beta_1\frac{\gamma_1\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{\beta_1} + \beta_2\frac{\gamma_2\rho\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta_2} + \beta_3\frac{\gamma_3\rho\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2}{\beta_3} + \\
&+ \beta_4\frac{\gamma_4\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2}{\beta_4} \beta_5\frac{\gamma_5\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2}{\beta_5}) = \log_2(1 + \rho\gamma_1\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2 + \\
&+ \rho\gamma_2\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2 + \rho\gamma_3\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2 + \rho\gamma_4\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2 + \rho\gamma_5\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2).
\end{aligned} \tag{4.35}$$

Now, by hypothesis, as in the two user case, it will be checked if the equality in the second inequality in equation (4.35) holds if $\frac{\gamma_1\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{\beta_1} = \frac{\gamma_2\rho\|\mathbf{v}_{\mathrm{m},2}^H\mathbf{H}_{\mathrm{m},2}\|^2}{\beta_2} = \frac{\gamma_3\rho\|\mathbf{v}_{\mathrm{m},3}^H\mathbf{H}_{\mathrm{m},3}\|^2}{\beta_3} = \frac{\gamma_4\rho\|\mathbf{v}_{\mathrm{m},4}^H\mathbf{H}_{\mathrm{m},4}\|^2}{\beta_4} = \frac{\gamma_5\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2}{\beta_5}$. Again, for simplicity, let us call $\frac{\gamma_1\rho\|\mathbf{v}_{\mathrm{m},1}^H\mathbf{H}_{\mathrm{m},1}\|^2}{\beta_1} = ... = \frac{\gamma_5\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2}{\beta_5} = A$. So, the left side of the second inequality in equation 4.35 is:

$$\begin{aligned}
&\beta_1 \log_2(1 + A) + \beta_2 \log_2(1 + A) + \beta_3 \log_2(1 + A) + \beta_4 \log_2(1 + A) + \\
&+ \beta_5 \log_2(1 + A) = (\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 +) \log_2(1 + A) = \log_2(1 + A).
\end{aligned} \tag{4.36}$$

And in the right side:

$$\begin{aligned}
&\log_2(1 + \beta_1 A + \beta_2 A + \beta_3 A + \beta_4 A + \beta_5 A) = \\
&= \log_2(1 + (\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5)A) = \log_2(1 + A).
\end{aligned} \tag{4.37}$$

It is then proved that this relation for the splitting of power resources given by $\beta_1 ... \beta_5$

and the power allocation coefficients for OMA $\gamma_1 ... \gamma_5$ maximizes the sum rate of MIMO-OMA for five users. One only needs now to rewrite the expression in terms of the coefficients of the resource splitting β. For simplicity, let one define $\gamma_1 \rho \|\mathbf{v}_{\mathrm{m},1}^H \mathbf{H}_{\mathrm{m},1}\|^2 = B_1, ..., \gamma_5 \rho \|\mathbf{v}_{\mathrm{m},5}^H \mathbf{H}_{\mathrm{m},5}\|^2 = B_5$. Consequently, it comes that:

$$\frac{B_1}{\beta_1} = \frac{B_5}{\beta_5} \implies \frac{B_1}{\beta_1} = \frac{B_5}{1 - \beta_1 - \beta_2 - \beta_3 - \beta_4} \tag{4.38}$$

$$\frac{B_1}{\beta_1} = \frac{B_2}{\beta_2} \implies \beta_2 = \frac{B_2}{B_1}\beta_1 \tag{4.39}$$

$$\frac{B_1}{\beta_1} = \frac{B_3}{\beta_3} \implies \beta_3 = \frac{B_3}{B_1}\beta_1 \tag{4.40}$$

$$\frac{B_1}{\beta_1} = \frac{B_4}{\beta_4} \implies \beta_4 = \frac{B_4}{B_1}\beta_1. \tag{4.41}$$

By substitution:

$$\frac{B_1}{\beta_1} = \frac{B_5}{1 - \beta_1 - \frac{B_2}{B_1}\beta_1 - \frac{B_3}{B_1}\beta_1 - \frac{B_4}{B_1}\beta_1} <=> \beta_1 = \frac{B_1}{B_1 + B_2 + B_3 + B_4 + B_5} \tag{4.42}$$

$$\beta_2 = \frac{B_2}{B_1 + B_2 + B_3 + B_4 + B_5} \tag{4.43}$$

$$\beta_3 = \frac{B_3}{B_1 + B_2 + B_3 + B_4 + B_5} \tag{4.44}$$

$$\beta_4 = \frac{B_4}{B_1 + B_2 + B_3 + B_4 + B_5} \tag{4.45}$$

$$\beta_5 = 1 - \beta_1 - \beta_2 - \beta_3 - \beta_4. \tag{4.46}$$

These resource allocation coefficients were used in the simulations. Due to the mathematical difficulty of having five rates instead of two, the demonstration of equation (4.24) for 5 users is omitted.

A downright equivalent comparison in terms of power allocation coefficient between the five user case and the two user case cannot be made, simply because allocating power to five users is different from allocating power to just two users. With those limitations in mind, the simulations made for five users tried to mirror the ones made for two users, namely, when $\alpha^2_{\mathrm{m},1} = 1-\gamma = 0.1$ and $\alpha^2_{\mathrm{m},2} = \gamma = 0.9$ for the two user case, $\alpha^2_{\mathrm{m},1} = \gamma_1 = 0.0542^2, \alpha^2_{\mathrm{m},2} = \gamma_2 = 0.1083^2, \alpha^2_{\mathrm{m},3} = \gamma_3 = 0.2166^2, \alpha^2_{\mathrm{m},4} = \gamma_4 = 0.4332^2$ and $\alpha^2_{\mathrm{m},5} = \gamma_5 = 0.8664^2$, for the five user case. Remember that these α coefficients are the same that were used in the simulation of Figure 3.7 and were chosen because they are known to have achieved good results, as seen in chapter 3. Analogously, when $\alpha^2_{\mathrm{m},1} = 1-\gamma = 0.9$ and $\alpha^2_{\mathrm{m},2} = \gamma = 0.1$ were used for the two user case, $\alpha^2_{\mathrm{m},1} = \gamma_1 = 0.8664^2, \alpha^2_{\mathrm{m},2} = \gamma_2 = 0.4332^2, \alpha^2_{\mathrm{m},3} = \gamma_3 = 0.2166^2, \alpha^2_{\mathrm{m},4} = \gamma_4 = 0.1083^2$ and $\alpha^2_{\mathrm{m},5} = \gamma_5 = 0.0542^2$ were used for the five user case. Naturally, when $\alpha^2_{\mathrm{m},1} = 1-\gamma = 0.5$ and $\alpha^2_{\mathrm{m},2} = \gamma = 0.5$ were used for the two user case, $\alpha^2_{\mathrm{m},1} = \gamma_1 = 0.2, \alpha^2_{\mathrm{m},2} = \gamma_2 = 0.2, \alpha^2_{\mathrm{m},3} = \gamma_3 = 0.2, \alpha^2_{\mathrm{m},4} = \gamma_4 = 0.2$ and $\alpha^2_{\mathrm{m},5} = \gamma_5 = 0.2$ were used for the five user case.

The results of the simulations can be seen in Figures 4.9, 4.10 and 4.11. Comparing with Figures 4.6, 4.7 and 4.8, it can be noticed that the general trend of having a higher difference in capacity between MIMO-NOMA and MIMO-OMA as the transmit SNR ρ grows is valid for both sets of Figures.

The most significant fact from the Figures is that the difference of capacity between

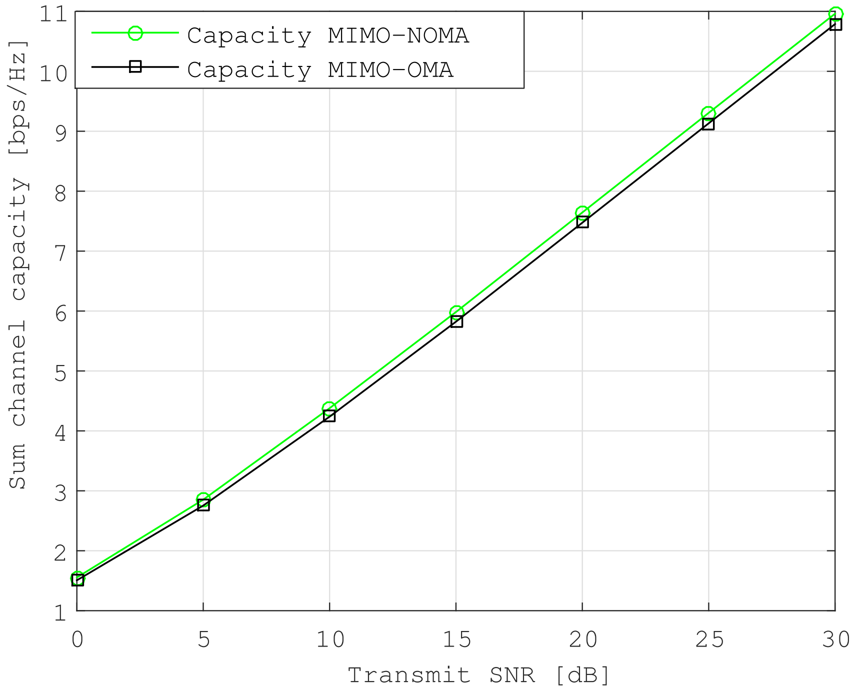

Figure 4.9: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with five users each, with $\alpha_{\mathrm{m},1}^2 = \gamma_1 = 0.8664^2, \alpha_{\mathrm{m},2}^2 = \gamma_2 = 0.4332^2, \alpha_{\mathrm{m},3}^2 = \gamma_3 = 0.2166^2, \alpha_{\mathrm{m},4}^2 = \gamma_4 = 0.1083^2$ and $\alpha_{\mathrm{m},5}^2 = \gamma_5 = 0.0542^2$.

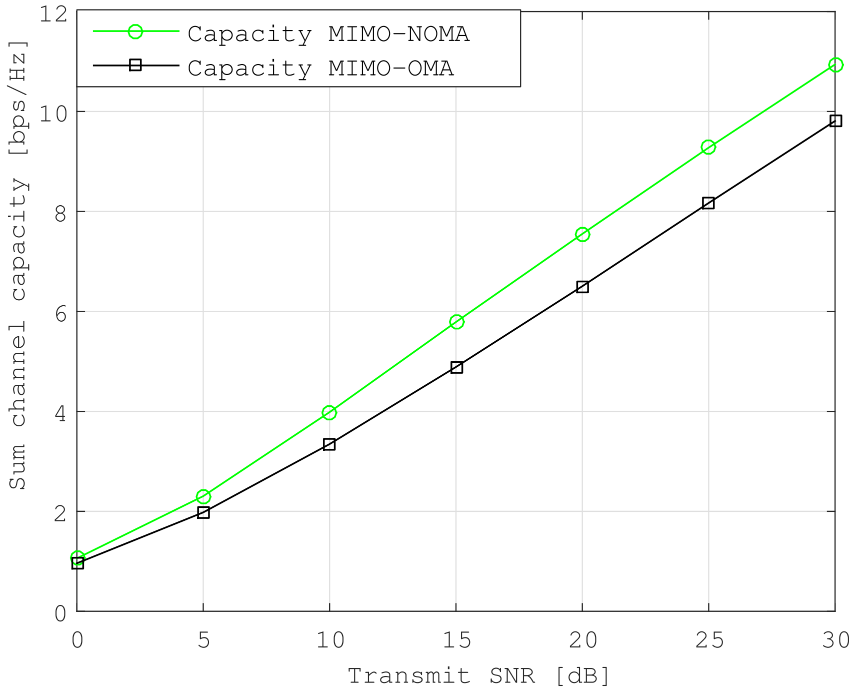

Figure 4.10: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with five users each, with $\alpha_{\mathrm{m},1}^2 = \gamma_1 = 0.2, \alpha_{\mathrm{m},2}^2 = \gamma_2 = 0.2, \alpha_{\mathrm{m},3}^2 = \gamma_3 = 0.2, \alpha_{\mathrm{m},4}^2 = \gamma_4 = 0.2$ and $\alpha_{\mathrm{m},5}^2 = \gamma_5 = 0.2$.

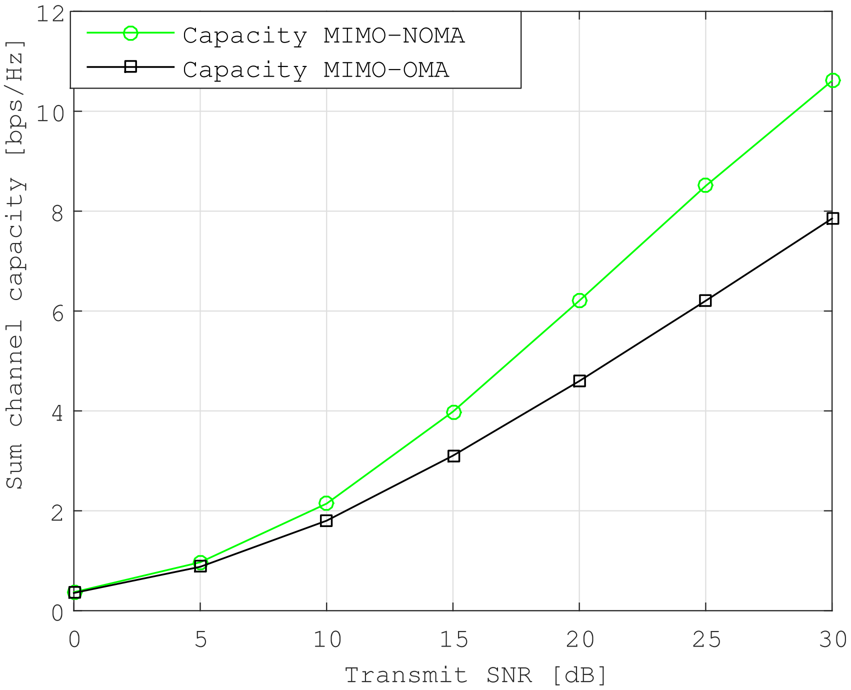

Figure 4.11: Sum channel capacity for MIMO-NOMA and MIMO-NOMA with five users each, with $\alpha^2_{\mathrm{m},1} = \gamma_1 = 0.0542^2, \alpha^2_{\mathrm{m},2} = \gamma_2 = 0.1083^2, \alpha^2_{\mathrm{m},3} = \gamma_3 = 0.2166^2, \alpha^2_{\mathrm{m},4} = \gamma_4 = 0.4332^2$ and $\alpha^2_{\mathrm{m},5} = \gamma_5 = 0.8664^2$.

MIMO-NOMA and MIMO-OMA grows with the number of users, a fact that, while intuitive, was yet to be proven, to the best of the author's knowledge. It should be remembered that for equidistributed power between the users, the two user case exhibited a maximum difference of 1.53 dB in capacity. As one can see in Figure 4.10, the maximum difference in capacity is, for the reference value of 8 dB of SNR, is now of 3.18 dB. Hence, an improvement of 1.65 dB has been reached.

Although the equidistributed power is the most fair case for comparison, the advantage of NOMA is more discernible when more power is allocated to the users with worst channel coefficients. With that in mind and looking at Figure 4.11 and noting that for the two user case with $\alpha^2_{\mathrm{m},1} = 1 - \gamma = 0.1$ and $\alpha^2_{\mathrm{m},2} = \gamma = 0.9$ the maximum difference of capacity, for the reference value of 8 dB of SNR, amounts to 3.58 dB, while in the other hand for the reference value of 8 dB of SNR, it shows a 6.54 dB difference in the five user case. This corresponds to an improvement of 2.96 dB. This is a very important result, taking in consideration that by looking at Figures 2.4, 2.5 and equation (1.1), where one may have thought that the improvement in terms of capacity from OMA to NOMA would be linear. However, the results prove that the improvement is sub-linear (if it was linear from 2 to 4 users one would get +3 dB of difference, instead of getting at maximum 2.96 dB from 2 to 5 users), which makes sense given that the performance of the higher users in the decoding chain is degraded by interference from other users, as in equation (4.29), for instance. Regarding the absolute value of the NOMA capacity, in Figure 4.9 the value is 10.96 [bps/Hz], in Figure 4.10 the value is 10.94 [bps/Hz] and in Figure 4.11 the value is 10.60 [bps/Hz]. These results are coherent with the analysis made in chapter 2, where it was said that the water-filling idea (allocate more power to the users with better channels) is optimal in terms of system's total throughput, but leads to an imbalanced allocation of the system's capacity to users with bad channels and hence poor fairness among users.

To reasure the results in Figure 3.7, simulations for the individual NOMA rates are presented in Figure 4.12. The dual SNR regime that appeared in Figure 3.7 can also be seen in Figure 4.12. With low SNR, the users with higher power allocations coefficients have a better performance, meaning, lower SER and higher rates. When the SNR is high, the users with lower power allocations coefficients have a lower SER and higher rates. An explanation for this was given in chapter 3, but it is now possible to rigorously explain these facts with the derivation of the rate expressions such as (4.29).

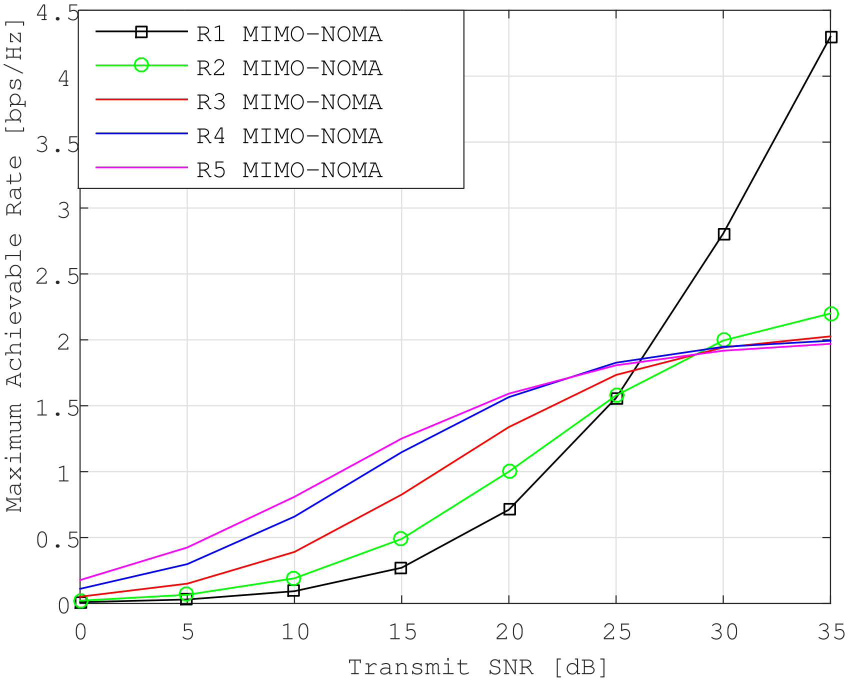

Figure 4.12: Maximum rates achieved by a MIMO-NOMA scheme with five users, with $\alpha_{\mathrm{m},1} = 0.0542, \alpha_{\mathrm{m},2} = 0.1083, \alpha_{\mathrm{m},3} = 0.2166, \alpha_{\mathrm{m},4} = 0.4332$ and $\alpha_{\mathrm{m},5} = 0.8664$.

In the low SNR regime, the term $\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},5}^2$ is larger than the other terms $\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},1}^2$, $\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},2}^2$, $\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},3}^2$ and $\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},4}^2$, noting equation (3.16). But as the transmit SNR ρ increases, the denominator $\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},1}^2 + \rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},2}^2 + \rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},3}^2 + \rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},4}^2 + 1$ will grow faster than $\rho\|\mathbf{v}_{\mathrm{m},5}^H\mathbf{H}_{\mathrm{m},5}\|^2\alpha_{\mathrm{m},5}^2$, until a point that the other rates, which have less interference factors in the denominator, starts to be actually larger than the rate of the fifth user. In both figures this point is around $\rho = 25$ dB.

Chapter 5

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 [34] have proven to hold. It has been explained why uncoded NOMA struggles to serve more than two users and an extension of this model to up to five users was managed, when the users are limited to BPSK. 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.

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.

A setup with massive MIMO and precoding at the base station was also implemented. The obtained performance in terms of SER was worst than the one with conventional MIMO in comparable setups, 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).

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 [32], 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.

Bibliography

(the hyperlinks provided below have proved to be working on 09/10/2016)

[1] Ericsson, "Mobility report," Ericsson, Stockholm, Sweden, White Paper, June 2015. [Online]. Available: http://www.ericsson.com/res/docs/2015/ericsson-mobility-report-june-2015.pdf

[2] T. world bank, "Mobile cellular subscriptions (per 100 people)," 2014. [Online]. Available: http://data.worldbank.org/indicator/IT.CEL.SETS.P2

[3] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020," Cisco, White Paper, February 2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[4] 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, vol. 52, no. 2, pp. 97–105, February 2014.

[5] "The 5G Infrastructure Public Private Partnership". Web platform: http://5g-ppp.eu [June 2015].

[6] "5th Generation Non-Orthogonal Waveforms for Asynchronous Signalling". Web platform: http://www.5gnow.eu [June 2015].

[7] "Mobile and Wireless Communications Enablers for Twenty-Twenty (2020) Information Society". Web platform: https://www.metis2020.com [June 2015].

[8] A. Osseiran, V. Braun, T. Hidekazu, P. Marsch, H. Schotten, H. Tullberg, M. A. Uusitalo, and M. Schellman, "The foundation of the mobile and wireless communications system for 2020 and beyond: Challenges, enablers and technology solutions," in Proc. of IEEE Vehicular Technology Conference (VTC Spring), Dresden, Germany, June 2013.

[9] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" IEEE Access, vol. 1, pp. 335–349, 2013.

[10] X. Zhang and J. G. Andrews, "Downlink cellular network analysis with multi-slope path loss models," IEEE Transactions on Communications, vol. 63, no. 5, pp. 1881–1894, May 2015.

[11] 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.

[12] G. D. Golden, C. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using v-blast space-time communication architecture," Electronics Letters, vol. 35, no. 1, pp. 14–16, Jan 1999.

[13] C. E. Shannon, "A mathematical theory of communication," ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, pp. 3–55, 2001.

[14] I. E. Telatar, "Capacity of multi-antenna gaussian channels," Tech. Rep. Bell Labs, Lucent Technologies., 1995, [Online]. Available: http://mars.bell-labs.com/papers/proof/proof.pdf.

[15] E. Telatar, "Capacity of multi-antenna gaussian channels," European transactions on telecommunications, vol. 10, no. 6, pp. 585–595, 1999.

[16] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” IEEE Journal on Selected Areas in Communications, vol. 21, no. 5, pp. 684–702, June 2003.

[17] A. Goldsmith, Wireless communications. Cambridge university press, 2005.

[18] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, “Non-orthogonal multiple access (NOMA) for cellular future radio access,” in Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th, June 2013, pp. 1–5.

[19] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, “System-level performance evaluation of downlink non-orthogonal multiple access (NOMA),” in Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on, Sept 2013, pp. 611–615.

[20] S. L. Goff, A. Glavieux, and C. Berrou, “Turbo-codes and high spectral efficiency modulation,” in Communications, 1994. ICC ’94, SUPERCOMM/ICC ’94, Conference Record, ’Serving Humanity Through Communications.’ IEEE International Conference on, May 1994, pp. 645–649 vol.2.

[21] 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, May 2016.

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

[23] J. A. Thomas and T. Cover, Elements of information theory. Wiley New York, 2006, vol. 2.

[24] H. Haci, “Non-orthogonal multiple access (NOMA) with asynchronous interference cancellation,” Ph.D. dissertation, University of Kent, 2015.

[25] Z. Ding, “Non-orthogonal multiple access (NOMA): Evolution towards 5G cellular networks,” April 2016, [Online]. Available: http://www.lancaster.ac.uk/staff/dingz/NOMA.pdf.

[26] Huawei, “Candidate schemes for superposition transmission,” April 2015. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R1-81--31256.htm

[27] Qualcomm, “Multiuser superposition schemes,” April 2015. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R1-81--31256.htm

[28] A. G. Perotti and B. M. Popovié, “Non-orthogonal multiple access for degraded broadcast channels: RA-CEMA,” in Wireless Communications and Networking Conference (WCNC), 2015 IEEE, March 2015, pp. 735–740.

[29] Z. Ding, P. Fan, and V. Poor, “Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions,” IEEE Transactions on Vehicular Technology, vol. PP, no. 99, pp. 1–1, 2015.

[30] Y. Liu, Z. Ding, M. Eïkashlan, and H. V. Poor, “Cooperative non-orthogonal multiple access in 5G systems with SWIPT,” in Signal Processing Conference (EUSIPCO), 2015 23rd European, Aug 2015, pp. 1999–2003.

[31] Z. Ding, M. Peng, and H. V. Poor, “Cooperative non-orthogonal multiple access in 5g systems,” IEEE Communications Letters, vol. 19, no. 8, pp. 1462–1465, Aug 2015.

[32] S. Timotheou and I. Krikidis, “Fairness for non-orthogonal multiple access in 5G systems,” Signal Processing Letters, IEEE, vol. 22, no. 10, pp. 1647–1651, Oct 2015.

[33] 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,” CoRR, vol. abs/1511.08610, 2015. [Online]. Available: http://arxiv.org/abs/1511.08610

[34] Z. Ding, F. Adachi, and H. Poor, “The application of MIMO to non-orthogonal multiple access,” IEEE Trans. on Wireless Communications, vol. 15, no. 1, pp. 537–552, Jan 2016.

[35] 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, Jan 2016.

[36] Z. Ding, M. Peng, and H. V. Poor, “Cooperative non-orthogonal multiple access in 5G systems,” IEEE Communications Letters, vol. 19, no. 8, pp. 1462–1465, Aug 2015.

[37] Z. Ding and H. V. Poor, “Design of massive-MIMO-NOMA with limited feedback,” IEEE Sig. Proc. Letters, vol. 23, no. 5, pp. 629–633, May 2016.

[38] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, “Joint spatial division and multiplexing—the large-scale array regime,” IEEE Trans. on Info. Theory, vol. 59, no. 10, pp. 6441–6463, Jun. 2013.

[39] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, “A hierarchical rate splitting strategy for FDD massive MIMO under imperfect CSIT,” in Proc. of 2015 IEEE 20th Inter. Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD), Guildford, UK, Sep 2015, pp. 80–84.