Multiuser Random Beamforming in Millimetre-Waves Channels

Corso di Laurea Magistrale in
ICT for Internet and Multimedia
(LM-27 Ingegneria delle Telecomunicazioni)

Laureando:
Francesco B. L. SANTAGIUSTINA

Relatore:
Prof. Stefano TOMASIN

Correlatore:
Prof. Jaume RIBA SAGARRA (Università Politecnica della Catalogna)

Lunedì 30 Settembre 2019 Anno Accademico 2018-2019

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

UNIVERSITÀ DEGLI STUDI DI PADOVA

Multiuser Random Beamforming in Millimetre-Waves Channels

A Master's Thesis submitted in partial fulfilment of the requirements for the master degree in ICT for Internet and Multimedia

Candidate:
Francesco B. L. SANTAGIUSTINA

Advisor:
Prof. Jaume RIBA SAGARRA

Co-advisor:
Prof. Stefano TOMASIN

Academic Year 2018/2019

Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona
UNIVERSITAT POLITÈCNICA DE CATALUNYA

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

To my old and new friends.

"There are friendships imprinted in our hearts that will never be diminished by time and distance."

Dodinsky

Acknowledgements

I would like to start by thanking my advisor, Professor Jaume Riba Sagarra who accorded me a great deal of trust by assigning me the topic before having had me as a student, and then helped me all the way in the redaction of this document with considerable dedication.
My gratitude also goes to my Paduan advisor Professor Stefano Tomasin, who followed at distance my Erasmus academic path, helped me to face the big amount of bureaucratic issues related to the exchange, and gave me useful advices on how to complete my thesis and steer my future career.
Then I would like to thank my parents, my siblings (in-law included) and my grandparents, who always supported me in every aspect of my life. They understood and encouraged my decision of spending this last year abroad even if, or maybe because, this would entail seeing me only on rare occasions.
And last but not least, all the friends who filled my spare time in this semester of thesis research and redaction of good moments together.

Abstract

Cellular communications exploiting the mmWaves frequency range are coming within our technological reach. However the specificities of propagation at these frequencies call for new transmission schemes. Concerning the downlink there are signs that opportunistic beamforming may be an effective solution. This thesis aims to show that in mmWaves channels, schemes based on randomly-directional beamforming allow to harness both the spatial multiplexing and multiuser diversity characterizing the broadcast channel by using only limited feedback and a simple transmitter architecture. It is well-known that performances of random beamforming schemes become optimal when the number of users tends to infinity. Hence, the number of necessary users with respect to the number of transmitting antennas is investigated and the necessity of a linear relation between the two is confirmed. Opportunistic beamforming is furthermore analysed under the aspect of fairness. The possibility to combine it with proportional-fair scheduling with only a small sum-rate loss is shown. Finally, the allocation of multiple users per beam is considered and the advantage of NOMA over OMA under the point of view of fairness is displayed.

Contents

List of Figures

List of Tables

List of Acronyms

3G 3^{rd} Generation of the global cellular network. 33

3GPP 3rd Generation Partnership Project. 14

5G 5^{th} Generation of the global cellular network. 1

AoD Angle of Departure. 17

AWGN Additive White Gaussian Noise. 15

BC Broadcast Channel. 3

BS Base Station. 10

CDF Cumulative Distribution Function. 67

CQI Channel Quality Indicator. 45

CSI Channel State Information. 1

CSIT Channel State Information at the Transmitterl. 14

DPC Dirty-Paper Coding. 24

EB-OBF Equispaced Beams Opportunistic Beamforming. 81

FDMA Frequency Division Multiple Access. 3

FSPL Free-Space Path Loss. 10

GBC Gaussian Broadcast Channel. 23

IDFT Inverse Discrete Fourier Transform. 29

ITU International Telecommunication Union. 9

LOS Line-Of-Sight. 10

LTE Long Term Evolution. 33

MAC Multiple Access Channel. 25

MIMO Multiple-Input Multiple-Output. 1

MISO Multiple-Input Single-Output. 16

MMSE Minimum Mean Squared Error. 39

mmWaves millimetre waves. 1

MRT Maximal Ratio Transmission. 40

NLOS Non Line-Of-Sight. 13

NOMA Non-Orthogonal Multiple Access. 3

OBF OBF Opportunistic Beam-Forming. 2

OFDM Orthogonal Frequency Division Multiplexing. 61

OMA Orthogonal Multiple Acces. 3

PDF Probabilistic Density Function. 68

PF Proportional Fair(ness). 60

QAM Quadrature Amplitude Modulation. 29

RF Radio-Frequency. 47

RMS Root Mean Square. 13

SB-OBF Single Beam Opportunistic Beamforming. 78

SIC Successive Interference Cancellation. 3

SINR Signal to Interference plus Noise Ratio. 2

SISO Single-Input Single-Output. 22

SNR Signal to Noise Ratio. 29

SVD Singular Value Decomposition. 29

TDMA Time Division Multiple Access. 3

UE User Equipment. 10

ULA Uniform Linear Array. 16

UR-MP Uniform Random Multi-Path. 18

UR-SP Uniform Random Single Path. 18

ZF Zero Forcing. 40

Notation

$\lvert\mathbf{A}\rvert$	Determinant of $\mathbf{A}$
$\lVert\mathbf{a}\rVert$	Norm of $\mathbf{a}$
$\log(\cdot)$	Base-2 logarithm
$\mathbf{A}$	Matrix
$\mathbf{a}$	Vector
$\mathbf{A}^*$	Complex conjugate of $\mathbf{A}$
$\mathbf{A}^\dagger$	Conjugate transpose of $\mathbf{A}$
$\mathbf{A}^T$	Transpose of $\mathbf{A}$
$\mathrm{col}_k(\mathbf{A})$	k^{th} column of $\mathbf{A}$
$\mathrm{Convex}\{\cdot\}$	Convex envelope operator
$\mathrm{E}[\cdot]$	Expected value
$\mathrm{Tr}(\mathbf{A})$	Trace of $\mathbf{A}$
a	Scalar

Introduction

In the two past decades internet access has been shifting from fixed only connections to mobile ones, and 2017 was the first year where more web-pages were served to mobiles users than to fixed ones. This global phenomenon is especially fostered by newly industrialized countries users which are mainly accessing internet through mobile connections (98 percent of Internet users in China are mobile). Furthermore mobile users data rates requirements have boomed with the consumption (and production) of video content on smartphones. This trend is continuing, plus the connections density is increasing as the spectrum must also be shared with machines forming the new Internet of Things. In this context increasing spectral efficiency and frequency reuse is becoming harder and harder. That's why among its new features, the fifth generation (5G) of the global cellular network introduces a radical change in the range of frequencies over which the communications may occur. Indeed it includes parts of the millimetre-waves (mmWaves) spectrum, also referred to as Extremely High Frequency spectrum, which ranges from 30 GHz to 300 GHz, while previous generations of cellular networks used only frequencies under 4 GHz. Unfortunately this huge amount of bandwidth wasn't allocated before to mobile communications for a sensible reason: its propagation properties are pretty poor. Indeed mmWaves suffer of a much higher propagation power loss, specifically when passing through solid and liquid obstacles, which may completely block them. On the other hand, given the shortness of the wavelength it is possible to fit in a regular base-station antennas arrays composed of a great many elements and thus exploit the advantages of Multiple-Inputs Multiple-Outputs (MIMO) systems.

Unfortunately usual precoding techniques require, not only a complex hardware structure which may be unrealizable in mmWaves, but also the knowledge of the state of the channel at the transmitter, and as the number of users and antennas becomes very large the Channel State Information (CSI) that should be fed-back from users to the base station may become too large. One way to overcome this problem is to simply focus the transmitted power towards users through beamforming. The elements of an antennas array can be coordinated to provide a high directivity to the signal. Furthermore, this directivity is con-

served by the mmWaves channel which has a quasi-optical nature: a beam may be blocked or reflected but will generally suffer little scattering and diffusion. However the mmWave channel presents an high variability, so to ulteriorly decrease the training overhead beamforming techniques where the beams are transmitted in random directions and then assigned to the user presenting the highest signal power have been developed under the name of random and Opportunistic Beam-Forming (OBF). This kind of schemes try to exploit the concept of multiuser diversity, which consists of exploiting the variability of the channel to transmit to each user when it is close to its peak channel quality level. In this optic, users only need to feed-back an aggregated channel quality indicator or their Signal to Noise plus Interference Noise Ratio (SINR) to the base station, consequently the feedback overhead is greatly reduced.

The interference between signals intended to different users is greatly decreased by using different beams focused in different directions, as they become asymptotically orthogonal when the number of transmitting antennas increases. In this way we can simultaneously transmit to many users located in different directions without needing to use separate frequency bands or codes reducing the available degrees of freedom. This concept is called spatial multiplexing and it is known that the maximal number of achievable parallel transmissions is, in the general MIMO case, limited by the number of transmitting antennas and receiving antennas. As the achievable angular beam width is inversely proportional to the number of antennas we may expect to be able to achieve a similar result with random beamforming using equispaced beams, but only at the condition of being able to find users aligned with each beam. The relationship between the number of transmission antennas and the necessary number of users have been object of studies and it has been shown that in mmWaves the necessary number of users to get both a full power and multiplexing gain asymptotically grows only linearly with the number of transmitting antennas, while for lower frequency channels modelled by Rayleigh fading this growth was exponential in the number of antennas. Random beamforming hence appears much more suitable for mmWaves frequencies.

Another issue is represented by the fact that focusing resources on the best users lead to fairness issues. Indeed if our criterion for transmission is simply the quality of the channel with respect to the random beam we can be led to neglect users which are far away from the base-station. To assure fairness we will probably be forced to sacrifice some multiuser diversity but the trade-off should be dealt with smartly. This kind of problems is generally dealt with through scheduling policies, but in the context of random beamforming the beam directions determine a subset of well-aligned users which are suitable for transmission so the scheduling policies must be rethought on a sectors basis, while trying to maintain a global optimality. Furthermore an advantage of Non-Orthogonal Multiple Access Tech-

niques (NOMA) (which consists in the superposition of the signals intended to different users followed by interference cancellation) with respect to Orthogonal Multiple Access Techniques OMA, like Time Division Multiple Acces (TDMA) or Frequency Division Multiple Access (FDMA), has been proved in terms of achievable sets of rates in the multiuser scenario. So it would be interesting to see how this can be applied to fairness enforcing for random beamforming in mmWaves channels. Interference cancellation of NOMA requires a computational complexity at the receiver scaling with the number of involved users, indeed each users should decode all the data intended to users with a worst channel than its own to remove their interference. In order to limit this complexity, the separation of users in sectors associated to each beam becomes advantageous, as we can perform Successive Interference Cancellation (SIC) just on a few users at the time.

Thesis outline and organization

The general aim of this thesis is to show that, in mmWaves channels, schemes based on random-directional beamforming allow to harness both spatial multiplexing and multiuser diversity of the Broadcast Channel (BC) using only limited feedback from users and a simple transmitter architecture.

The thesis is organized in two parts. The first part is named "Theoretical framework and analysis" and contains the chapters 1 to 4. This title shouldn't mislead the reader: this part deals with very practical problems, but with an analytical and theoretical approach. The numerous aspects which contributes to the suitability of random beamforming in mmWaves channels (like channel characteristic and hardware limitations) are presented. The general performances we can expect from random beamforming are put in the larger contest of linear precoding and the general information-theoretic broadcast channel. The opportunistic beamforming scheme itself is characterized and the criticalities emerging from a theoretical analysis are pointed out. The second part is named "Computational analysis and numerical results" and it contains Chapter 5 and 6. These two chapters aim to elucidate the aforementioned issues of the necessary number of users and fairness by computational methods: including simulations and numerical analysis. Different fairness policies are extended to opportunistic beamforming and their performances are compared.

- Chapter 1: mmWaves and simplified channel models
  The electromagnetic waves propagation characteristics in the mmWave frequency range are reported and compared the those in the traditional frequencies. The channel models successively used are defined and justified.
- Chapter 2: Optimal performances of the MIMO Broadcast Channel

The broadcast channel is analysed from an information theoretical point of view, leading to the sum-capacity formulas and it's asymptotical analysis. The concepts of spatial multiplexing and multiuser diversity, which explain the asymptotical performances of opportunistic beamforming, are explained and their upper bound is defined.

- Chapter 3: Dealing with hardware and feedback limitations
  The problem of limited feedback in linear precoding is exposed, and random beamforming proposed as a solution. The main beamforming transceiver architectures are reviewed, and practical problems calling for a simple directional beamforming protocol are presented.

- Chapter 4: Some open issues in opportunistic beamforming
  The problem of the necessary number of users for optimal performances of opportunistic beamforming is stated and some asymptotical results are exposed. The optimal fairness concept is defined and associated to scheduling policies, as well as NOMA and OMA.

- Chapter 5: Methodology and approach
  The objectives of the numerical analysis are presented, and the simulations methodology is described. A semi-analytical approach to quickly compute the expectations of some r.vs. of interest are presented.

- Chapter 6: Simulations and results:
  Simulations concerning the evolution of the performances of opportunistic beamforming with the number of users are performed and their results are exposed. The proportional fairness policy is examined for random beamforming in heterogeneous networks, the proportional fairness scheduling is applied in conjunction with NOMA in order to see if an advantage over OMA is obtained.

The following conceptual diagram put in relation the main topics and ideas underlying the thesis structure.

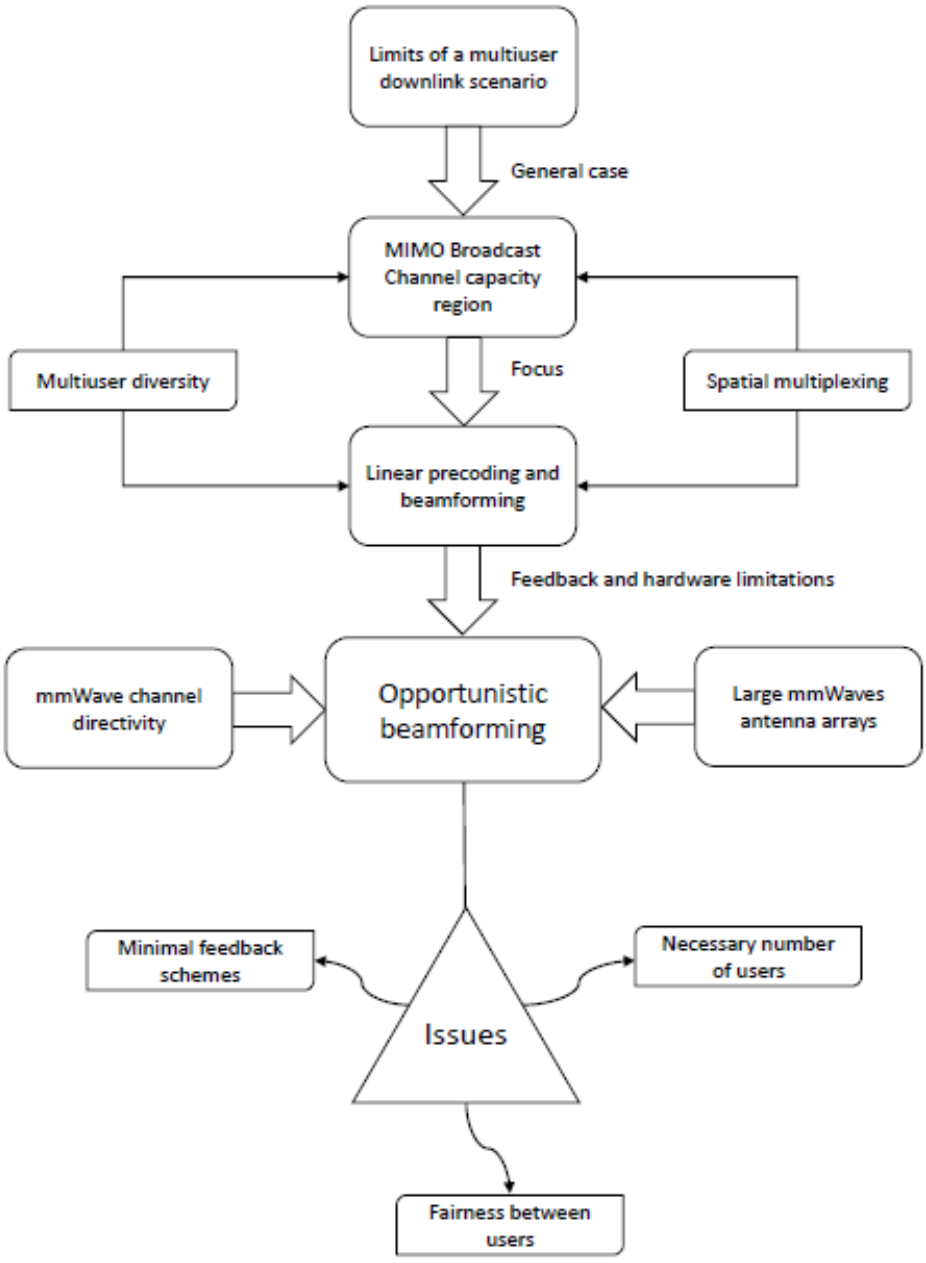


Figure 1: Conceptual scheme of the thesis content

Part I

Theoretical framework and analysis

Chapter 1

mmWaves and simplified channel models

Electromagnetic waves in the frequency range between 30 GHz and 300 GHz are referred to as millimetre waves (often abbreviated as mm-Waves or mmWaves and officially called "extremely high frequencies" by the ITU) because their wavelength span from about one to ten millimetres. This large amount of under-utilised bandwidth (about 250 GHz of usable bands) [12] represents one of the main resources which will allow the next generations of cellular systems, starting from 5G, to meet the massively increasing data rate demand from users and connected devices [13]. In the past these frequencies were mainly used in the backhaul for fixed or satellite point-to-point radio-links because their characteristics seemed to be unfit for mobile communications which were instead deployed in microwave frequencies usually between 700 MHz and 2.6 GHz.

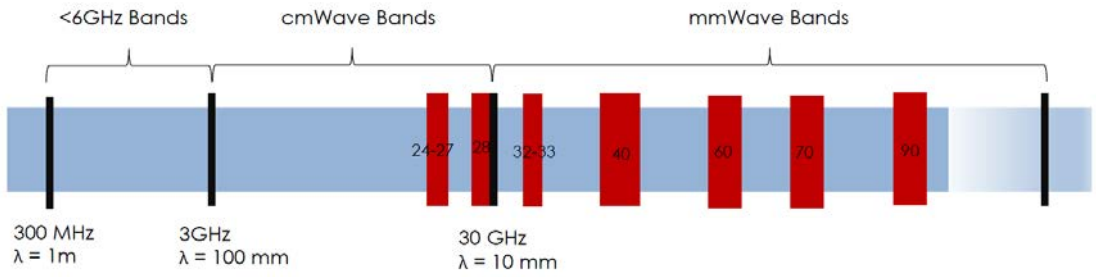


Figure 1.1: Electromagnetic spectrum chart showing the planned allocation of mmWaves bands in 5G (credits: Qualcomm)

The way electromagnetic waves propagate and interact with the environment is strongly frequency dependent. Indeed, mmWaves suffer from high atmospheric attenuation (due to oxygen absorption or precipitation) and have a low penetra-

tion coefficient which may result on the complete shadowing presence of obstacles (buildings but also foliage) [3]. Furthermore, being the propagation of mmWaves very directional, similarly to visible light, it was thought that only line-of-sight (LOS) communication would have been possible. However, the smallness of the wavelength allows to concentrate a large number of antennas in a small surface base station (BS) or user equipment (UE) transceiver which allows to implement massive MIMO or beamforming techniques which greatly improve capacity by increasing directivity gain and leveraging path diversity and reflections. Robust links with moving UE can furthermore be achieved by adaptive beam steering and handover from one BS to another. Also the manufacturing of components for mmWaves transmission, reception and processing developed in the past few years and nowadays network and device components are already being commercialised [14]. In this chapter we are thus going to investigate which are the main characteristics of the mmWaves channel and describe some resulting models which have been proposed in the literature and are fitted for our successive analysis of random beamforming schemes, which implies a trade-off between accuracy and mathematical tractability.

1.1 Propagation characteristics of mmWaves

The aim of this section is to give a quick overview of the main differences between mmWaves bands and traditional cellular bands, focusing on those characteristics that will motivate the adoption of our channel models and the consecutive analysis of random beamforming schemes: increased path loss, high probability of blockage and sparse multipath.

1.1.1 Signal attenuation and blockage

The maximum coverage distance and interference levels in our system will depend on the path loss caused by free space and atmospheric attenuation. First and foremost, one could remember that the free-space path loss (FSPL) depends quadratically on the frequency:

$$\text{FSPL} = \left(\frac{4\pi d}{\lambda}\right)^2 = \left(\frac{4\pi df}{c}\right)^2 \tag{1.1}$$

when considering antennas with frequency-independent gains. For instance, we will have a 40dB difference in FSPL between a transmission at 1 GHz and one at 100 GHz. Contrariwise if at both ends of the link we consider transceivers with a fixed physical area and the corresponding antennas gains, we obtain that the path loss decreases quadratically with frequency [15], [16]! Another major concern is the

attenuation caused by gases composing the atmosphere and eventual precipitation. The atmospheric attenuation generally increases with frequency and we can notice in Figure 2 the presence of several stronger absorption windows corresponding to dioxygen and water vapor molecules resonance frequencies that we will preferably reserve for local or personal area communications with coverage distances of a few meters. But if we consider a transmission distance of one kilometre, we realise that between 30 and 100 GHz, taking out the 50 to 70 GHz oxygen absorption window we have 50 GHz of bandwidth left where the atmospheric attenuation is under 1 dB! Since one of the principles of modern cellular systems is frequency reuse through the deployment of a dense network of base-stations, with inter-site distances of a few hundred meters in urban areas, the atmospheric attenuation would be even lower. Using more sophisticated channel models for the urban micro-cell scenario like the 3GPP/ITU ABG or NYUSIM from New York University, it has been shown that beyond the first meter of distance there is virtually no difference in the path loss exponent between mmWaves and standard microwaves [15].

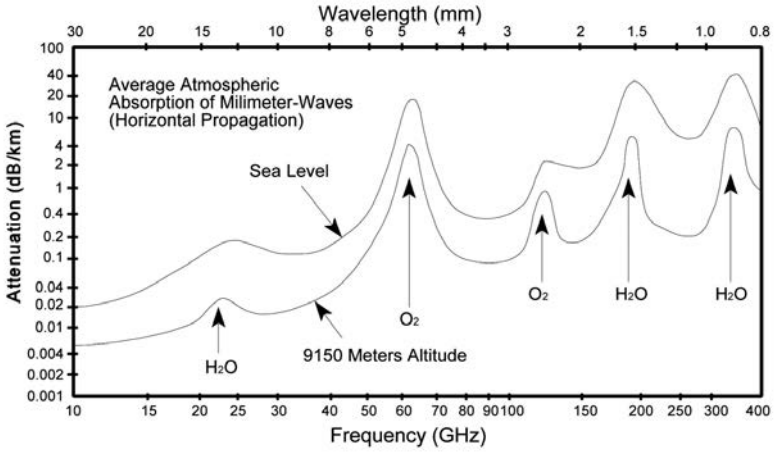


Figure 1.2: Atmospheric absorption of mmWaves [1]

Another factor to be taken into account is rain and the other climatic effects. If the effects of light rain are negligible for small cell distances, heavy rain or fog may severely impair performances of mmWaves systems. The size of raindrops is roughly the same as the wavelengths of mmWaves and therefore they can scatter, but also absorb, diffract or depolarize the signal. The attenuation will depend on the intensity of rain, usually expressed in millimetres per hour, but also on the distribution of droplets size and waves' polarization (see [17] and [18] for more details). We will here focus only on frequency.

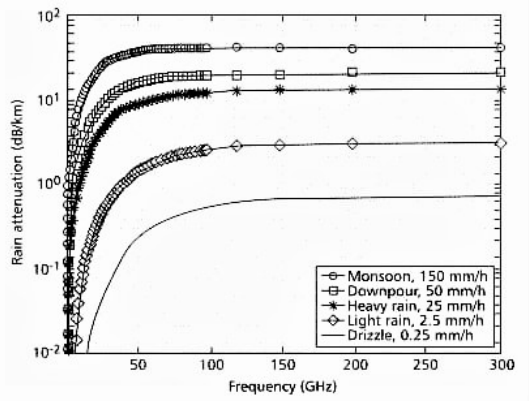


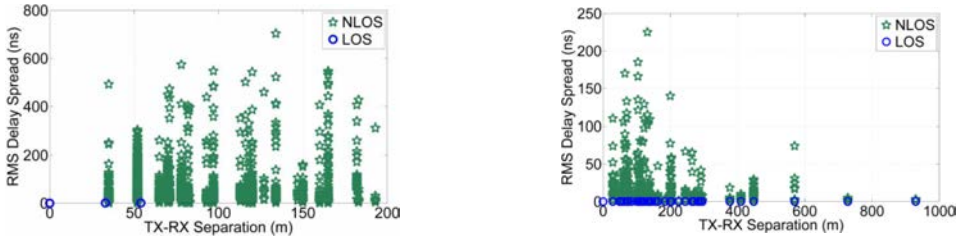
Figure 1.3: Rain attenuation in mmWaves frequencies [2]

As shown in Figure 1.3 the rain attenuation increases of many orders of magnitude in the first ten GHz but then becomes almost constant after 100 GHz. Taking a frequency of 75 GHz one can observe that the heavy rain attenuation is roughly 10dB/km which is quite large. Although this kind of attenuation may be compensated by increased gains, in the regions where intense precipitation is expected it is a good practice to set-up a system where communications in regular cellular bands act as back-up when the severe propagation impairments make mmWaves communication unavailable or inconvenient. Another fundamental challenge for mmWaves signals is their sensitivity to blockage. Considering the mobility of users, the appearance of buildings, vehicles but also of foliage or of the user's own body in the transmission path between the BS and the UE is very likely and may cause a sudden outage if not handled properly. It has been shown that transmission power through lossy media decreases almost uniformly with frequency [15]. It is thus very hard to guarantee connection to indoor users with an outdoor base station, and thinking to an urban environment, coverage can be guaranteed only with a dynamic optimization of beamforming/precoding to exploit the better paths and the handover from one BS to the other when it gets a better channel. The required density of BS will actually be mainly determined by the blockage events rather than attenuation or frequency reuse requirements [19]. Simulations for 5G New Radio have shown that in a metropolitan setting the inter-site distance should be of about two hundred meters to guarantee coverage with present technology. From a Qualcomm simulation for San Francisco it resulted that by reusing existing LTE sites to deploy mmWaves BS an outdoor downlink coverage of 64.8% could be achieved, and that by reaching 73 sites/km^2 a coverage of over 95% could be achieved [20].

1.1.2 Multipath and consequent delay and Doppler spread

Given the high attenuation or blockage that could be suffered by the LOS path and the advantages resulting from spatial diversity, advanced beamforming and beam tracking techniques leveraging path diversity and reflections are one of the key enabling factors for successful mmWaves communications. The quasi-optical nature and high directivity of mmWaves typically produce very few multipaths,

but they can have a sufficiently high intensity to allow for non line-of-sight (NLOS) communication. As the wavelength is so short many surfaces which would appear nearly flat at higher scales become rough (i.e. have surface height deviations comparable to the wavelength) and produce diffuse scattering. This diffusion being essentially isotropic in the extreme case, it entails a large power dissipation [15]. For this reason, in urban areas the material composing building façades becomes a critical factor (consider the difference between smooth glass and concrete or stucco for instance). Diffraction on the other hand strongly decreases at high frequencies, so obstacles like buildings, vehicles or people will produce sharp "shadows", and it will be difficult to be able to exploit a refracted signal to reach a UE around a building corner for instance [21]. We will mainly rely on reflections, obeying the Fresnel equations like visible light. NLOS paths will require channel equalization and will have a higher latency than LOS channels. Each path n having a different length, it will introduce a different delay τ_n, which will result at the receiver in a phase shift of $2\pi\tau_n f$ for that path's spectral components. The impact of the higher frequency of mmWaves on the delay spread is not clear but the root-mean-square (RMS) delay spread is generally below a few hundreds of nanoseconds. The delay spread is actually much more influenced by the type of environment as shown in 1.4 where the New York City environment is denser and more reflective than the one of Austin which implies LOS links only for small ranges and shorter delay spreads.



(a) New York (b) Austin

Figure 1.4: Measures of RMS delay spread as a function of TX-RX separation for all possible pointing angles at 28 GHz in New York City and 38 GHz in Austin, Texas for LOS and NLOS links [3]

In a scenario with mobile users (or reflectors), multipaths will lead by Doppler effect to different frequency shifts ν_n, depending on the different angle of arrivals to the UE and resulting in a phenomenon called Doppler spread, which also design the range of these frequency shifts. The channel coherence time is the time interval over which the frequency response of the channel can be considered static. As it depends on the speed at which the UE moves from a region of destructive interference between multipaths to the other, coherence time is in-

versely proportional to the Doppler spread, and for a constant UE velocity the Doppler frequency shift linearly increase with the carrier frequency. As a result, in mmWaves the coherence time is smaller than in regular cellular bands. This means that passing from 2 GHz to 28 GHz carriers, causes channel coherence times passing from 500 μs to 35 μs. For the Channel State Information at the Transmitter (CSIT) to be useful, it must cover a time-frequency block where the channel is approximately static, so it should be updated at each coherence time. That's one of the reasons why full CSI feedback in mmWaves becomes more challenging than at lower frequencies, additionally to the lower capabilities in uplink with respect to the downlink, if we want to concentrate complexity at the BS. Nevertheless, the use of a narrow beam can allow to eliminate much of multipaths and simplify the channel response by considerably reducing the delay spread and Doppler spread (by reducing the angular spread of incoming waves), thus also increasing coherence time [2]. By channel sparsity we indicate the fact that the range of the parameters (like angle of arrival, delay, Doppler shift) associated to non-negligible received power is very limited.

1.2 Fundamental models for a multiuser downlink MISO in mmWaves bands

Given the high directivity of mmWaves, ray-tracing-based models have been successfully used to find the best path with minimal loss and maintain a low interference level [22]. These models relay on a knowledge of the surrounding environment geometry and propagation characteristics, which could be obtained through an initial set-up operation or a routine training phase. In the 5G NR context, very sophisticated models have been designed, we can cite the ABG model from 3GPP/ITU. But the aim of this thesis is to show how in a mmWaves multiuser downlink scenario, smart stochastic beamforming schemes can achieve almost optimal performances using only a partial CSI feed-back. Our result aim to be of general extent, and to overlook the specificities of the individual settings. We will thus adopt tractable analytical models, with a fixed set of parameters (number of paths, average path loss, blockage probability etc.) which will be used to simulate the channel of each user, making abstraction of the physical environment leading to that channel characteristics, and representing them by appropriate random variables. Nonetheless, in order to apply the beamforming concept, and to implement schemes based on the distance of the UE from the BS, we will also have to assign a position to users around the BS. The details will be explained in the simulation scenario description part.

1.2.1 Rayleigh and Ricean fading models

The well-known Rayleigh and Ricean fading models have been used to represent wireless channels in a compact way by taking advantage of the central limit theorem. They indeed consider the interference at the receiver of a very large number of refracted or reflected multipaths. By considering that the individual attenuations and delays of each of the N_p paths are independent of frequency and by neglecting Doppler effect, we can model a multipath fading channel between the BS and UE as a linear time-variant system, where the signal at the receiver becomes:

$$\boldsymbol{r}(t) = \sum_{n=1}^{N_p} a_n(t) x\left(t - \tau_n(t)\right) \tag{1.2}$$

with $x(t)$ being the transmitted signal and $a_n\left(t\right)$ the complex valued overall attenuation factor that includes all the effects of the propagation and antennas gains. By using Shannon-Nyquist sampling theorem we get a discrete convolutional model:

$$r[m] = \sum_{l=0}^{\infty} h_l[m] x[m-l] \tag{1.3}$$

where the contributes of the different paths have been grouped as:

$$h_l[m] = \sum_{n=1}^{N_p} a_n\left(\frac{m}{W}\right) e^{-j2\pi f \tau_n\left(\frac{m}{W}\right)} \operatorname{sinc}\left[l - \tau_n\left(\frac{m}{W}\right) W\right] . \tag{1.4}$$

Each Fourier component will thus suffer from a phase shift of $e^{-j2\pi\tau_n f}$ due to travel time. If we look at the wavelength λ of the component and the path length d_n we have that $\tau_n f = \frac{d_n}{\lambda} \gg 1$, so for each path the phase shift can basically be considered uniformly distributed between 0 and 2π, thus each of the addends in the previous sum can be modelled as an independent circular symmetric complex random variable. As the number of paths becomes very large, if we consider all their gains to be of close orders of magnitude, we get that the distribution of the sum $h_l[m]$ converges by the central limit theorem towards a circular symmetric complex gaussian (see [23] for a more rigorous and complete development). This defines the Rayleigh fading channel, which is usually associated to additive white Gaussian noise (AWGN) $n[m] \sim CN(0, \sigma_m^2)$ to form the following input-output model:

$$y[m] = \sum_{l=0}^{L} h_l[m] x[m-l] + n[m] \tag{1.5}$$

note than in practice the impulse response will be of finite length L as longer delays correspond to longer paths which in turns are associated to stronger attenuations

and become negligible. In the narrowband approximation all the $\tau_n(t)$ are much lower than the symbol period so we can assume the channel to be memoryless, or equivalently frequency flat fading (FFF), in that case we only have a multiplicative channel:

$$y[m] = h^*[m]x[m] + n[m]\,. \tag{1.6}$$

For a Multiple-Input Single-Output (MISO) system instead, we need to consider the superposition at the k^{th} user's receiver of the signals $x_i[m]$ coming from each of the M transmitting antennas, and that each of them experiment a different fading $h_{k,i}[m]$:

$$y_k[m] = \sum_{i=1}^{M} h^*_{k,i}[m]x_i[m] + n_k[m] = \mathbf{h}_k^\dagger[m]\mathbf{x}[m] + n_k[m]\,. \tag{1.7}$$

In the Rayleigh frequency-flat fading MISO channel, if the separation between the transmitting antennas is sufficiently large, we consider that each of the fading coefficients is independent from the others. The Rayleigh model is fitted for scenarios with only NLOS paths and diffuse scattering or many small reflectors rather than few strong reflections. From the considerations made in the first section of this chapter on mmWaves propagation characteristics, we easily understand that the Rayleigh model isn't particularly adapted. The Rician channel model is an extension of the Rayleigh model which includes an additional dominant LOS or specular path with known attenuation $\sqrt{\frac{k}{k+1}}\sigma_1$ where k, called K-factor, balance the ratio of energy among the main path and the scattered paths:

$$h_l[m] = \sqrt{\frac{k}{k+1}}\sigma_1 e^{j\theta} + \sqrt{\frac{1}{k+1}} CN\left(0, \sigma_l^2\right)\,. \tag{1.8}$$

Although the inclusion of a dominant path makes the Rician model better for LOS mmWaves links both these models have been developed in the context of traditional communications at lower frequencies where the power of scattered waves is more relevant, and don't take into account of the direction of the transmission, nor the distance from the BS, but only small-scale fading effects.

1.2.2 Models including directivity and distance

As anticipated, we will make use of beamforming to focus power toward the users through a good path and avoid interferences. We consider that our transmitting BS is equipped with a large uniform linear array (ULA) of M antennas spaced of a distance d. If the receiving UE are sufficiently far away from the BS (more than the Fraunhofer distance), the far-field approximation holds and the waves can be considered planar. Then, we can note from 1.5 that depending on the receiver

azimuthal angle with respect to the array, it will receive the signal produced by each antenna with a delay due to the difference in the wave path lengths.

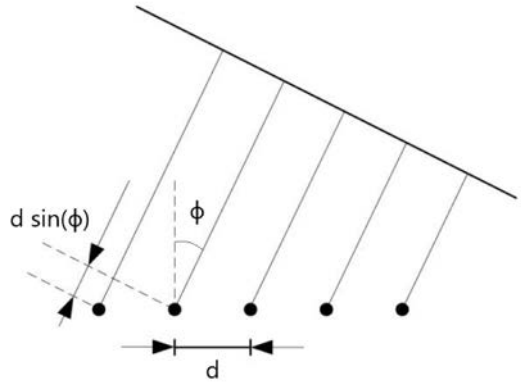


Figure 1.5: Uniform linear array with inter-antenna distance d. Note the additional path distance $d \sin \phi$ for each antenna before the right side one [4]

The signals are received with a constant time delay $\tau = \frac{d \sin \phi}{c}$, where c is the speed of light, between each consecutive antenna, which is equivalent to a phase delay of $2\pi f_c \tau = 2\pi \frac{d \sin \phi}{\lambda_c}$ where f_c and $\lambda_c = \frac{c}{f_c}$ are the carrier frequency and wavelength. We will define the normalized normalized angle of departure (AoD) from the physical AoD $\phi \in [-\pi, \pi]$ and the carrier wavelength λ_c as follows:

$$\theta = \frac{2d \sin \phi}{\lambda_c} . \tag{1.9}$$

If we start counting from the antenna on the closest end of the array, the i^{th} antenna will have a cumulative phase delay of $(i-1)\pi\theta$. We can thus model the received signal from this path as $\mathbf{a}(\theta)^\dagger \mathbf{x}[m]$ where $\mathbf{a}(\theta)$ is called steering vector and is defined as:

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{M}} \begin{pmatrix} 1 \\ e^{-j\pi\theta} \\ e^{-j\pi 2\theta} \\ \vdots \\ e^{-j\pi(M-1)\theta} \end{pmatrix} \tag{1.10}$$

Our channel models are assumed to be non-selective in frequency in order to focus only on the spatial aspects, so we'll use as input output model:

$$y_k[m] = \mathbf{h}_k[m]^\dagger \mathbf{x}[m] + n_k[m] \tag{1.11}$$

The strength of the received signal will mainly depend on how much energy is transmitted in the directions that result in a good propagation channel (in LOS or through reflections) to the UE. Such a channel can be modelled as:

$$\mathbf{h}_k[m] = \sqrt{\frac{M}{N_p}} \sum_{i=1}^{N_p} \alpha_{k,i}[m]\mathbf{a}\left(\theta_{k,i}\right) \tag{1.12}$$

Where the $\alpha_{k,i}$ are the path gains that will change for each transmitted symbol in a fast fading scenario. The high directivity and attenuation of mmWaves allows us to limit ourselves to a small number of paths N_p with non-negligible received power. By modelling the gains as independent standard complex Gaussian variables, that is $\alpha_{k,i}[m] \sim CN(0,1)$, and the AoD $\theta_{k,i}$ as independent random variables with uniform distribution over $[-1,1]$, we obtain the so called Uniform Random Multi-Path (UR-MP) channel model, whose energy is normalized. In the extreme case of only one relevant direction, we reduce to the so called "Uniform random single path" (UR-SP) model:

$$\mathbf{h}_k[m] = \alpha_k[m]\sqrt{M}\mathbf{a}\left(\theta_k\right) \tag{1.13}$$

The UR-SP model has been introduced in [24] and extensively used to simplify the task of analytically studying beamforming in highly directional mmWaves channels, while the UR-MP was introduced in [25] in order to capture a larger variety of scenarios. Indeed, as N_p grows we pass from a single path case to a few dominant reflections and then reach high scattering scenarios, asymptotically converging to the Rayleigh fading scenario. Finally, one may want to explicit the part of attenuation depending on the distance between the UE and the base station, to be able to use the distance information to optimize the transmission strategy, so we have:

$$\mathbf{h}_k[m] = \sqrt{\frac{M}{N_p}} \frac{\alpha_{k,0}[m]\mathbf{a}\left(\theta_{k,0}\right)}{\sqrt{1+d_k^{\beta_{LOS}}}} + \sqrt{\frac{M}{N_p}} \sum_{i=1}^{N_p-1} \frac{\alpha_{k,i}[m]\mathbf{a}\left(\theta_{k,i}\right)}{\sqrt{1+d_k^{\beta_{\mathrm{NLOS}}}}} \tag{1.14}$$

which allows us to distinguish between the loss exponents of the line-of-sight path β_{LOS} and the loss exponent of the non-line-of-sight paths β_{NLOS} . One option adopted in [26] is to focus on the dominant LOS path and neglect the others, obtaining the simplified model:

$$\mathbf{h}_k[m] = \sqrt{M} \frac{\alpha_k[m]\mathbf{a}\left(\theta_k\right)}{\sqrt{1+d_k^{\beta_{LOS}}}} \tag{1.15}$$

then we can assume that this LOS path will be available only with a certain probability $P(LOS) = e^{-\phi d_k}$ with ϕ determined by the building density, their shape and material etc... If the LOS path isn't available, we will simply consider it

as a blockage situation, excluding that user from the transmission. These models will allow us to analytically study and simulate random beamforming schemes based on the position of users, which could be obtained by the BS through feedback or estimation techniques. However, the distance can also be seen as a proxy for any factor determining a long term heterogeneity of the fading between users: such as the presence of highly attenuating medias in the transmission path.

Chapter 2

Optimal performances of the MIMO Broadcast Channel

The scenario considered by this thesis consists in the downlink from a base station to multiple non-cooperative receivers, which is referred to in information theory as broadcast channel. Note that differently from the meaning given to "broadcasting" in networking terminology, we consider that the transmitter wants to deliver independent data to the different users, so our scenario is indeed the one of cellular networks rather than radio or television broadcasting. In order to evaluate the performances of the proposed limited feedback schemes, we need an ideal benchmark as reference. Thus we will now briefly introduce the main capacity results for MIMO broadcast channels and the main concepts that need to be exploited to achieve the capacity region: spatial multiplexing, interference cancellation and multiuser diversity.

2.1 Capacity region of the MIMO BC

As we are dealing with a multiuser system, where we want to transmit independent information to each of the K users, we cannot simply talk about the capacity of the channel, but we have to use the concept of capacity region.

Definition 2.1.1. *Consider a set $(R_1, R_2, ..., R_{K-1}, R_K)$ of simultaneously achievable rates between the base-station and K users, the capacity region of the channel is defined as the the closure of the union of all achievable rate sets.*

As studying the region of achievable sets of rates using a specific communication strategy may become cumbersome in some situations, we may focus on some other simplified metrics. The most important one is the sum-rate R^{Sum} which is the maximal sum of rates simultaneously achievable by a communication strategy, and

is upper-bounded by the sum-rate capacity C^{Sum} of the channel. The sum-rate offers a view on the global performance of the system. Instead, if we are concerned about the balance between the rates of each users a simple solution is to use the symmetric-rate R^{Sym} which is the highest rate that can be simultaneously allocated to every receiver in the system, and whose upper bound is the symmetric capacity C^{Sym} of the channel. The condition of having identical rates for each user being quite restrictive, in the methodology chapter we will see that other metrics can be used to evaluate the fairness of the system.

2.1.1 Capacity region of the SISO Broadcast Channel

Let's first consider the case where there is only one transmitting antenna and one receiving antenna per user: the Single-Input Single-Output (SISO) Broadcast Channel:

$$y_k = h_k^* x + n_k, \quad k = 1, 2, \ldots, K \tag{2.1}$$

where $h_k \in \mathbb{C}$ is the downlink channel between the BS and user k, $x \in \mathbb{C}$ is the transmitted signal and the complex Gaussian noise terms $n_k \sim CN(0, 1)$ are independent. Furthermore, we set a limit on the transmitted power: $\boldsymbol{\sigma}_x = \mathbb{E}_x [xx^*] \leq P$.

Unless all users have the same channel gain, the full capacity region can't be achieved by orthogonal multiple access schemes like Time Division Multiple Access (TDMA) or Frequency Division Multiple Access (FDMA) which totally avoid interference between signals intended to different users. In the general case the capacity-region achieving communication strategy is to use a Non-Orthogonal Multiple Access (NOMA) scheme, composed of Superposition Coding at transmitter and Successive Interference Cancellation (SIC) at receivers [27], [28]. Superposition Coding consists in simply transmitting a weighted sum of the signals intended to each user. Note that the rate of the superposed streams will be different in general. Then, each users is able to decode the data intended to users with channel gain lower than its own, and can thus reconstruct the corresponding transmitted signals and subtract them from the received superposition before decoding its own message. In this way each user only suffers from the interference caused by signals intended to users with a better channel gain.
By varying at the transmitter the power assigned to each user's signal all points of the capacity region are achieved. For instance, supposing to have two users, if we want to help the user which has the channel with the lower gain we can increase the power dedicated to its transmission without significantly deteriorating the performances of the user with the better channel as it will not be affected by interference anyway. The sum-capacity, for any K, is nevertheless achieved by transmitting only to the user with the highest channel gain. If we order users by

decreasing channel gains, that is $|h_1| \geq |h_2| \geq \cdots \geq |h_K|$, the boundary of the capacity region of the SISO BC with AWGN of power N_0 and total transmitted power P is given by the sets of rates $\{R_k\}_{1 \leq k \leq K}$ of the following form:

$$R_k = \log\left(1 + \frac{P_k |h_k|^2}{\left(\sum_{j=1}^{k-1} P_j\right) |h_k|^2 + N_0}\right) \qquad \text{subject to } \sum_{k=1}^{K} P_k = P\,. \tag{2.2}$$

The sum-capacity is then obtained as:

$$\begin{aligned} \mathcal{C}_{SISO-BC}^{Sum} &= \max_{\{P_k\} | \sum_{k=1}^{K} P_k = P} \quad \sum_{k=1}^{K} R_k \\ &= \max_{\{P_k\} | \sum_{k=1}^{K} P_k = P} \quad \sum_{k=1}^{K} \log\left(1 + \frac{P_k |h_k|^2}{\left(\sum_{j=1}^{k-1} P_j\right) |h_k|^2 + N_0}\right) \\ &= \log\left(1 + \frac{P |h_1|^2}{N_0}\right)\,. \end{aligned} \tag{2.3}$$

To apply SIC, one needs to be able to order the channels of the different users, we did this intuitively by considering an ordering of the channel gains $|h_1| \geq |h_2| \geq \cdots \geq |h_K|$ but we should stress the fact that SIC is capacity achieving for a class of broadcast channels designed as degraded.

Definition 2.1.2. *A broadcast channel with K users is said to be degraded if the r.v. X representing the transmitted signal and the r.vs. Y_i with $1 \leq i \leq K$ representing the signal received by each user constitute a Markov chain of the form: $X \to Y_1 \to Y_2 \to ... \to Y_K$.*

This implies for instance, that a higher rate is achievable between the BS and user i than between the BS and user $j > i$ by the data processing inequality which states that $I(X;Y_j) \leq I(X;Y_i)$ for any input distribution $p(x)$, we say in that case that the channel of user i is *more capable* than the channel of user j. We can see that our Gaussian Broadcast Channel GBC falls in the class of degraded channels by first considering the equivalent channel where all channels have equal attenuation but a rescaled noise power $N_i = \frac{N_0}{|h_i|^2}$, which will be ordered as follows: $N_1 \leq N_2 \leq ... \leq N_K$. Then the received signal by user $l > 1$ can be expressed as the sum of the signal received by user $l-1$ and some independent additional Gaussian noise of power $N_l - N_{l-1}$. In this way Y_l depends on X only through Y_{l-1} and we have a Markov chain.

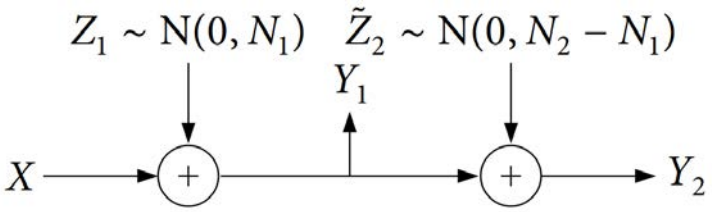


Figure 2.1: Equivalent degraded model of a 2-users Gaussian BC [5]

This description of the SISO BC was important not only as limiting case of the more general MISO BC but because it allowed us to introduce SIC which, despite not being directly applicable or capacity achieving in the MISO case, will turn out useful in the developments part to efficiently multiplex on a single directional beam many users with disparate channel qualities (i.e. when the near-far effect is predominant).

2.1.2 Capacity region of the MIMO Broadcast Channel

By allowing multiple antennas at the transmitter, it's not anymore so obvious how to rank the channels of the different users by their quality: the channel is in general *non-degraded*. We now have a channel vector (or matrix for $N > 1$) for each user, and as the effective channel gain will depend on its scalar product with the transmit vector $\mathbf{x}$, having a higher norm isn't a sufficient condition to get a higher gain, the correspondence between the directions of the two vectors also has to be taken into account, as well as the phase matching. Thus it is impossible to state generally that one user will be able to decode the data intended to other users because its channel is better, making SIC not applicable. Nevertheless, another way to reduce the multiple access interference, taking advantage of a coding technique developed in 1983 which is named "dirty-paper coding" (DPC) or Costa precoding from the name of its inventor [29]. This technique takes advantage of the fact that with full CSIT, the BS knows how the signal intended to different users will interfere with each-other and it can pre-cancel part of the interference at transmission. DPC was first applied to the Gaussian Broadcast Channel in 2000 [30], starting from a $M = 2, K = 2, N = 1$ setting it has successively been extended to the most general case of arbitrary M, K and N. The rates region achieved by DPC was the largest ever discovered but it took a few years to prove that it effectively coincided with the capacity region of the GBC [31]. Although very interesting, a full account of the research process leading to the final proof would fall beyond the scope of this thesis, but it can be found in [32]. We will nevertheless mention the fact that many progresses were allowed by the discovery of a fundamental equivalence

between the Broadcast Channel and the Multiple Access Channel (MAC), which represents the uplink case of the same scenario with multiple users and one base station. For instance the sum-rate capacity of the GBC was upper bounded by the so-called Sato's bound and by passing through the dual MAC channel it has been possible to prove that the maximal sum-rate point of the DPC region coincided with it [33]. As we will see later, this equivalence allows to compute much more easily the bounds of the capacity region of the GBC, for which no closed-form solution exist.

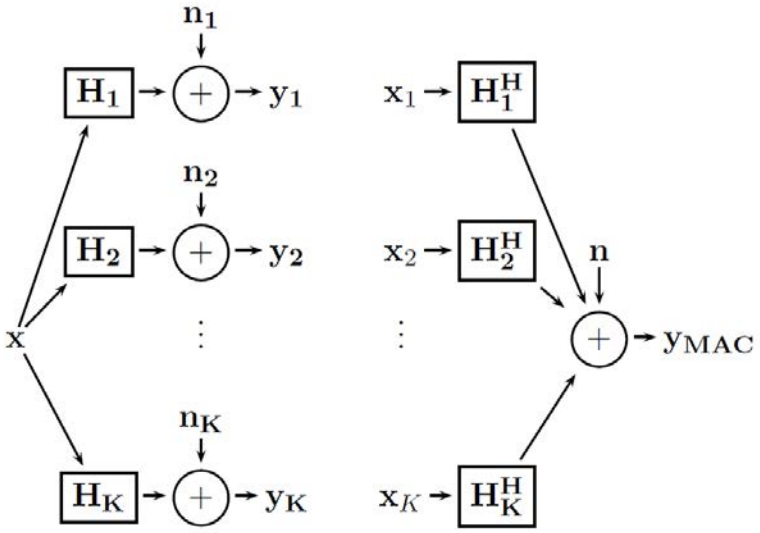


Figure 2.2: System models of the MIMO BC (left) and the MIMO MAC (right) channels [6]

Let's now define the general MIMO Broadcast Channel model, where one base-station with M antennas transmits to K users having each N antennas. Following our narrow-band model we can express the received signal by user k as:

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x} + \mathbf{n}_k, \quad k = 1, 2, \ldots, K \tag{2.4}$$

where $\mathbf{H}_k \in \mathbb{C}^{N\times M}$ is the (known and fixed) downlink channel matrix between the BS and user k, $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is the transmit vector, and the circularly symmetric complex Gaussian noise terms $\mathbf{n}_k \sim CN(0, \mathbf{I})$ are independent. We define the transmit covariance matrix of the input signal as $\mathbf{\Sigma_x} = \mathbb{E}_\mathbf{x}\left[\mathbf{x}\mathbf{x}^\dagger\right] \succeq 0$. Then, the average power constraint of the transmitter can be expressed as: $\mathrm{Tr}\left(\mathbf{\Sigma_x}\right) \leq P$. The fundamental result of dirty paper coding is that when the noise is Gaussian and if the channel to each user is known at the transmitter, it allows to sequentially

encode the information for each user and transmit it simultaneously so that when it receives the superposition each user experiences no interference from the users encoded earlier. This is furthermore achieved without increasing the transmitted power. As a glimpse on the functioning of DPC let's suppose that we want to transmit simultaneously to two different users. We first choose a code-word $\mathbf{w}_1 \in \mathcal{B}_1$ for the first user, then code for the second user by considering $\mathbf{w}_1$ as known interference. The optimal way to do so for the AWGN case $\mathbf{y}_2 = \mathbf{w}_1 + \mathbf{w}_2 + \mathbf{n}_2$ is to find the intended information code word $\mathbf{p}$ which is closest to $\alpha\mathbf{w_1}$ in a code-book $\mathcal{B}_2$ formed by replicating $\mathcal{B}_1$ to uniformly cover all the domain over which the sum of the signal and the interference could lie, and then transmit $\mathbf{w}_2 = \mathbf{p} - \alpha\mathbf{w}_1$. The coefficient $\alpha = \frac{P}{P+N_0}$ will allow to minimize the error of the estimation of $\mathbf{p}$ at the receiver, which is done by looking in $\mathcal{B}_2$ for the nearest code-word to $\alpha\mathbf{y}_2$. In this way, the second receiver error probability is the same as if the transmission to the first user was absent. Similarly if we had one more user, we could choose its code-word by presubtracting the interference caused by the first two user's signals to its transmission. Still, it would cause interference to users one and two as user two causes interference to user one. We can apply the procedure to any number of user but, differently from SIC, the choice of the ordering of users is arbitrary and we will lead to different performances. That's why we will make use of the set Π of all the ordering vectors π over the set $\{1, 2, ..., K\}$, such that $\pi(1)$ is encoded first. By extending DPC to our general MIMO BC model the capacity region is then defined as:

$$\mathcal{C}_{\mathrm{BC}}\left(P, \mathbf{H}_{1\ldots K}\right) \triangleq \\ \mathrm{Convex}\left\{ \bigcup_{\pi\in\Pi,\ \mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K} \begin{pmatrix} R_1^{\mathrm{DPC}}\left(\pi, \mathbf{\Sigma}_{1\ldots K}, \mathbf{H}_{1\ldots K}\right), \ldots, \\ R_K^{\mathrm{DPC}}\left(\pi, \mathbf{\Sigma}_{1\ldots K}, \mathbf{H}_{1\ldots K}\right) \end{pmatrix} \right\} \tag{2.5}$$

where $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$ are the individual power allocation matrices such that $\mathbf{\Sigma}_i \succeq 0, \forall i$ and $\mathrm{tr}\left\{\sum_{i=1}^{K} \mathbf{\Sigma}_i\right\} = \mathrm{Tr}\left(\mathbf{\Sigma}_{\mathbf{x}}\right) \leq P$ and Convex$\{\cdot\}$ represents the convex envelope operator. The rates R_k^{DPC} achievable by DPC are then given by:

$$R_{\pi(i)}^{\mathrm{DPC}} = \log \frac{\left|\mathbf{I} + \mathbf{H}_{\pi(i)}\left(\sum_{j\geq i} \mathbf{\Sigma}_{\pi(j)}\right)\mathbf{H}_{\pi(i)}^{\dagger}\right|}{\left|\mathbf{I} + \mathbf{H}_{\pi(i)}\left(\sum_{j>i} \mathbf{\Sigma}_{\pi(j)}\right)\mathbf{H}_{\pi(i)}^{\dagger}\right|} \tag{2.6}$$

This equation shows explicitly the fact that the rate achieved by the i^{th} user to be encoded is obtained by considering only the power of transmissions to users $\pi(j)$ such that $j \geq i$ as noise or signal contributors. Nevertheless it doesn't help us a lot to obtain the bounds of the capacity region as the maximization of the sets of rates is a non-convex problem. We will thus show how it can be obtained from

the capacity region of the dual MIMO MAC which is much simpler to compute. First, note that if the channel matrix in downlink is $\mathbf{H}$, the uplink channel matrix is $\mathbf{H}^{\dagger}$. The main difference is given by the fact that we have, in general, individual power constraints on each user's covariance matrix $\boldsymbol{Q}_i \succeq 0$ which must be such that $\mathrm{Tr}\left(\boldsymbol{Q}_i\right) \leq P_i$.

Theorem 2.1.1. *The dirty paper region of a MIMO BC channel with power constraint P is equal to the capacity region of the dual MIMO MAC with sum power constraint P.*

$$\mathcal{C}_{\mathrm{BC}}\left(P, \mathbf{H}_{1\ldots K}\right) = \bigcup_{\{P_i\} \mid \sum_{i=0}^{K} P_i \leq P} \mathcal{C}_{\mathrm{MAC}}\left(P_1, \ldots, P_K; \boldsymbol{H}_{1\ldots K}^{\dagger}\right) \tag{2.7}$$

Then we only need to know that the capacity region of the MIMO MAC channel is given by the union (over the possible covariance matrices $\boldsymbol{Q}_j \succeq 0$ which respect the individual power constraints) of rates following a sum-rate bound for each possible subset of the users:

$$\begin{aligned} &\mathcal{C}_{\mathrm{MAC}}\left(P_1, \ldots, P_K, \boldsymbol{H}_{1\ldots K}^{\dagger}\right) \\ &\quad \triangleq \bigcup_{\{\boldsymbol{Q}_i\} \mid \mathrm{Tr}(\boldsymbol{Q}_i) \leq P_i} \Big\{ (R_1, \ldots, R_K) : \textstyle\sum_{i \in S} R_i \\ &\qquad \leq \log \left| \boldsymbol{I} + \textstyle\sum_{i \in S} \boldsymbol{H}_i^{\dagger} \boldsymbol{Q}_i \boldsymbol{H}_i \right| \forall S \subseteq \{1, \ldots, M\} \Big\} \end{aligned} \tag{2.8}$$

For each set of covariance matrices $(\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K)$ we will obtain a K-dimensional polyhedron of achievable rates. The MIMO MAC rates are a concave function of the covariance matrices, so the boundary points of the capacity region and the corresponding covariance matrices that the boundary points of the sum power MIMO MAC capacity region (and the corresponding covariance matrices) can be found by standard convex optimization algorithms. The duality of the MAC and BC channel also allows to find more easily the optimal covariance matrices to be used in dirty paper coding by a simple transformation of those used in the dual MAC [33]. Concerning the achievable sum-rate, we have:

$$\begin{aligned} \mathcal{C}_{\mathrm{BC}}^{\mathrm{Sum}}\left(P, \boldsymbol{H}_{1\ldots K}\right) &= \mathcal{C}_{\mathrm{MAC}}^{\mathrm{Sum}}(P, \boldsymbol{H}_{1\ldots K}^{\dagger}) \\ &= \max_{\{\boldsymbol{Q}_j\} \mid \sum_{i=1}^{K} \mathrm{Tr}(\boldsymbol{Q}_i) \leq P} \log \left| \boldsymbol{I} + \sum_{i=1}^{K} \boldsymbol{H}_i^{\dagger} \boldsymbol{Q}_i \boldsymbol{H}_i \right| \end{aligned} \tag{2.9}$$

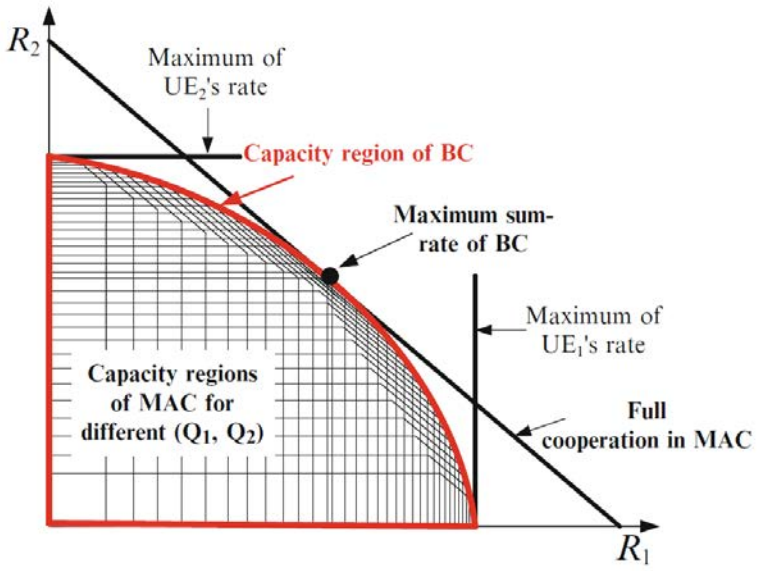


Figure 2.3: Dual capacity regions of the MAC and BC for two-users [7]

Note that till here we have always considered for simplicity deterministic and fixed channel matrices, but the results we have found can be extended to fast fading channels in the sense of ergodic capacity. For instance the ergotic sum-rate capacity of the MIMO-BC with full CSIT can be found in [34] to be:

$$\begin{gathered} C_{BC}^{Sum} = \max_{\boldsymbol{Q}_k(\mathbf{H})} \mathbf{E}_{\mathbf{H}} \left[\log \left(\left| \mathbf{I} + \sum_{k=1}^{K} \mathbf{H}_k^H \boldsymbol{Q}_k(\mathbf{H}) \mathbf{H}_k \right| \right) \right] \\ \text{subject to:} \\ \sum_{k=1}^{K} \mathrm{Tr}\left(\boldsymbol{Q}_k(\mathbf{H})\right) \leq P \\ \boldsymbol{Q}_k(\mathbf{H}) \succeq 0, \text{ for } k = 1, 2, \ldots, K \end{gathered} \tag{2.10}$$

which is an immediate extension of (1.34) taking in account that the covariance matrix can be adapted as a function of the realizations of the channel matrix.

2.2 Spatial multiplexing and multiuser diversity in MIMO Broadcast Channels

In the analysis of the achievable rate of a telecommunication system, one major factor is represented by the number of degrees of freedom of the channel. Degrees of freedom can be seen as the number of independent dimensions over which we can transmit our signal or, more precisely, as we are concerned with

distinguishability at the receiver, as the number of dimensions of the received signal space. For instance by using Quadrature Amplitude Modulation (QAM) instead of Pulse Amplitude Modulation (QAM) we double the number of degrees of freedom, but this can also be achieved by doubling the bandwidth and transmitting two independent symbols over orthogonal carriers. But degrees of freedom can also be increased by taking advantage of the physical separation between multiple receiving antennas if we manage to transmit different non-interfering data streams to each of them. In this case we talk about spatial degrees of freedom, which in turn allow for spatial multiplexing. The number of degrees of freedom of the channel will determine how many additional bits per channel use we obtain by doubling the transmission power at high Signal to Noise Ratio (SNR). Considering only the variability of spatial degrees of freedom, we can define the spatial multiplexing gain as follows.

Definition 2.2.1. *The spatial multiplexing gain r corresponding to a transmission rate R obtained by using a total power P is defined as:*

$$r = \lim_{P \to \infty} \frac{R(P)}{\log P}. \tag{2.11}$$

We will now see how the maximal spatial multiplexing gain can be computed and obtained in a MIMO point-to-point channel, and then understand how this concept can be extended to the broadcast channel.

2.2.1 Spatial multiplexing in MIMO point-to-point systems

As explained spatial multiplexing is very important because it allows to obtain an increase of the degrees of freedom and not only a power gain. This is particularly convenient in high-SNR regime because of the concavity of the logarithm. It is a well-known result that in a point-to-point MIMO system with M transmit an N receive antennas the channel can be decomposed in up to $min(N, M)$ virtual channels over which it is possible to transmit independent data streams. This can be achieved through Singular Value Decomposition (SVD), which requires CSIT, or, when channel matrices are asymptotically circular, by Inverse Discrete Fourier Transform (IDFT) which doesn't require CSIT but necessitate the addition of a cyclic prefix at the end of codewords, thus introducing an overhead. Indeed, the point-to-point narrowband AGWN MIMO channel can be expressed in matrix form as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{2.12}$$

where $\mathbf{x} \in \mathbb{C}^M, \mathbf{y} \in \mathbb{C}^N$ and $\mathbf{n} \sim \mathcal{CN}(0, N_0\mathbf{I}_N)$ are the transmitted, received and noise signals while $\mathbf{H} \in \mathbb{C}^{M \times N}$ is the channel matrix containing the complex multiplicative factors of the channel between each of the transmitting and receiving

antennas, and that we will consider deterministic and constant for simplicity (see [35] for the general case and ergodic capacity computation). The channel matrix be can rewritten by SVD as $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger$ where $\mathbf{U}$ is $M \times M$ and unitary, $\mathbf{V}$ is $N \times N$ and unitary, and Σ is a $M \times N$ rectangular diagonal matrix which contains the so-called singular values σ_i of $\mathbf{H}$ on its diagonal. The matrix $\mathbf{H}$ has exactly R_H positive singular values σ_i, where R_H is the rank of $\mathbf{H}$, which by basic principles satisfies $R_H \leq \min(M, N)$. By applying the linear transformation V to our codewords $\tilde{\mathbf{x}}$ before transmission and $\mathbf{U}^\dagger$ at the receiver, we obtain

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{U}^\dagger(\mathbf{H}\mathbf{x} + \mathbf{n}) \\ &= \mathbf{U}^\dagger \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger(\mathbf{V}\tilde{\mathbf{x}}) + \mathbf{n}\right) \\ &= \mathbf{\Sigma}\tilde{\mathbf{x}} + \mathbf{U}^\dagger\mathbf{n} \\ &= \mathbf{\Sigma}\tilde{\mathbf{x}} + \tilde{\mathbf{n}} \end{aligned} \tag{2.13}$$

that constitutes a system of R_H parallel sub-channels as $\mathbf{\Sigma}$ is diagonal:

$$\tilde{y}_i = \sigma_i \tilde{c}_i + \tilde{n}_i \text{ for } i = 1, \ldots, R_H \,. \tag{2.14}$$

We can then apply a waterfilling policy to optimize the power distribution among channels:

$$P_i = \left(\mu - \frac{1}{\sigma_i^2}\right)^+ \text{ for } 1 \leq i \leq R_H \text{ and subject to } \sum_{i=1}^{R_H} P_i = P \,. \tag{2.15}$$

Then, the capacity can be obtained as the sum of rates achievable on each virtual sub-channel. At high SNR, given that an equal power distribution is optimal we find:

$$\begin{aligned} C_{\text{MIMO}}(P, \mathbf{H}) &= \sum_i^{R_H} \left(\log\left(1 + \frac{P_i|\sigma_i|^2}{N_0}\right)\right) \\ &\approx \sum_i^{R_H} \left(\log\left(1 + \frac{P|\sigma_i|^2}{R_H N_0}\right)\right) && \text{at high SNR} \\ &\leq R_H \log\left(1 + \frac{P}{R_H N_0}\left(\frac{1}{R_H}\sum_i^{R_H}|\sigma_i|^2\right)\right) && \text{by Jensen's inequality} \\ &= R_H \log\left(1 + \frac{P}{R_H N_0}\left(\frac{\text{Tr}[\mathbf{H}^\dagger\mathbf{H}]}{R_H}\right)\right) \,. \end{aligned} \tag{2.16}$$

We can notice that as the transmitted power grows we get a R_H-fold increase of the rate with respect to a SISO system, this kind of gain in degrees of freedom is

called (spatial) multiplexing gain and can be upper bounded by:

$$\begin{aligned}\lim_{P\to\infty}\frac{C_{MIMO}}{\log(P)} &\leq \lim_{P\to\infty}\frac{R_H\log\left(1+\frac{P}{R_H N_0}\left(\frac{Tr[\mathbf{H}^\dagger\mathbf{H}]}{R_H}\right)\right)}{\log(P)}\\ &= \lim_{P\to\infty}\frac{R_H\log\left(\frac{P}{R_H N_0}\left(\frac{Tr[\mathbf{H}^\dagger\mathbf{H}]}{R_H}\right)\right)}{\log(P)}\\ &= \lim_{P\to\infty}\frac{R_H\left(\log(P)+\log\left(\frac{Tr[\mathbf{H}^\dagger\mathbf{H}]}{R_H^2 N_0^2}\right)\right)}{\log(P)}\\ &= R_H \leq \min(M,N)\,.\end{aligned} \tag{2.17}$$

Note that the capacity of a MIMO system is equivalently expressed in matrix form using the covariance matrix $\boldsymbol{Q} = \mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\dagger\right] \succeq 0$ as:

$$C_{MIMO} = \max_{\{\boldsymbol{Q}\}|\,\mathrm{Tr}(\boldsymbol{Q})\leq P} \log\left|\mathbf{I}_N + \mathbf{H}\boldsymbol{Q}\mathbf{H}^\dagger\right|\,. \tag{2.18}$$

We can understand that this expression is equivalent to the one in (21) by remembering that the determinant of a matrix is equal to the product of its eigenvalues, that the non-zero singular values of $\mathbf{H}$ are equal to the eigenvalues of $\mathbf{H}\mathbf{H}^\dagger$(or $\mathbf{H}^\dagger\mathbf{H}$) and that the logarithm of a product is equal to the sum of the logarithms of the factors.

2.2.2 Spatial multiplexing in the MISO Broadcast Channel

We have seen that in a point-to-point MIMO system we can achieve a multiplexing gain of up to the minimum between the number of transmitting and receiving antennas. But what happens with K multiple users, each one having multiple antennas? Can we globally consider the system as a M to $N \times K$ MIMO system and reach a multiplexing gain of $\min(M, N \times K)$? The difference consists in the fact that in the BC the antennas belonging to different users cannot cooperate in the decoding of their messages. Despite this limitation it is actually possible, with perfect CSIT, to obtain the same spatial multiplexing gain $\min(M, N \times K)$ of the equivalent cooperative MIMO [36]. Actually it has been proven that as transmitted power goes to infinity the difference between the sum-capacity of the BC and that of the equivalent cooperative MIMO tends to zero [37].

Let's now consider the MISO Broadcast Channel where each receiver only has $N = 1$ antennas to try to get some insights on the implications of the sum-rate capacity formula. We can express the received signal by user k as:

$$y_k = \mathbf{h}_k\mathbf{x} + n_k, \quad k = 1, 2, \ldots, K \tag{2.19}$$

where $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ is the downlink channel between the BS and user k, $\mathbf{x} \in \mathbb{C}^M$ is the transmit vector and the complex Gaussian noise terms $n_k \sim CN(0,1)$ are independent. We keep the transmit covariance matrix $\mathbf{\Sigma_x} = \mathbb{E}_\mathbf{x}\left[\mathbf{x}\mathbf{x}^\dagger\right]$ and the average power constraint $\mathrm{Tr}\left(\mathbf{\Sigma_x}\right) \leq P$. The sum-rate capacity formula is then simplified to:

$$\mathcal{C}^{Sum}_{MISO-BC}(P,\mathbf{H}) = \max_{\{P_i\}:\sum_{i=1}^K P_i \leq P} \log\left|\mathbf{I} + \sum_{i=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right| \tag{2.20}$$

Indeed, for the $N = 1$ case, duality indicates that rank-one covariance matrices (i.e., beamforming) allow to reach the BC capacity. This fact follows from the transformations to the dual MAC channel where we have a set of covariance matrices which are scalars in the $N = 1$ case [35].
Let's have a look to the spatial multiplexing gain of the MISO BC. First imagine that we have $K = M$ and $\mathbf{h}_i = \mathbf{e}_i \ \forall i$, which are row vectors with all zeros excepted in position i. Then for symmetry reasons the best power allocation is $P_i = \frac{P}{K}$ and we get:

$$\begin{aligned}
\log\left|\mathbf{I} + \sum_{i=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right| &= \log\left|\mathbf{I} + \frac{P}{K}\sum_{i=1}^K \mathbf{e}_i \mathbf{e}_i^\dagger\right| \\
&= \log\left|\mathbf{I} + \frac{P}{K}\mathbf{I}\right| \\
&= \log\prod_{j=1}^M \left(1 + \frac{P}{K}\right) \\
&= M\log\left(1 + \frac{P}{K}\right)
\end{aligned}$$

so we get as expected a multiplexing gain of M. The same principle can be extended to any set of channel vectors $\{\mathbf{h}_k\}$. Note that for any matrix $\mathbf{A}$ if $\mathbf{A}v = \lambda v$, then $(\mathbf{A} + \mathbf{I})v = (\lambda + 1)v$, so $\mathrm{eig}_j\left(\mathbf{I} + \sum_{i=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right) = 1 + \mathrm{eig}_j\left(\sum_{i=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right)$ and we get:

$$\begin{aligned}
\log\left|\mathbf{I} + \sum_{j=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right| &= \log\prod_{i=1}^M \mathrm{eig}_j\left(\mathbf{I} + \sum_{i=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right) \\
&= \log\prod_{i=j}^M 1 + \mathrm{eig}_j\left(\sum_{i=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right) \\
&= \sum_{i=j}^M \log\left(1 + \mathrm{eig}_j\left(\sum_{i=1}^K P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right)\right).
\end{aligned} \tag{2.21}$$

Let's define $R_h \triangleq \text{rank}\left(\sum_{i=1}^{K} P_i \mathbf{h}_i \mathbf{h}_i^\dagger\right)$, as $\forall \quad \mathbf{h}_i \neq \mathbf{0}$, the outer products $\mathbf{h}_i \mathbf{h}_i^\dagger$ will result in rank-1 positive semi-definite matrices which will be independent if the channel vectors are, R_h will coincide with the number of linearly independent channel vectors. The number of non-zero terms in the external sum is also given by R_H as $\log(1+x) = 0 \iff x = 0$.

Considering that we have K of these channel vectors and at most M of them can be linearly independent we get the spatial multiplexing gain upper bound of $\min(M, K)$, which can be reached even without cooperation between the antennas of the K users.

Now we can even go further in the approximation with a cooperative MIMO channel by considering the channel matrix formed in the following way $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \cdots \mathbf{h}_K]$ to state the following theorem from [37]:

Theorem 2.2.1. *If* $\mathbf{H}$ *is full row rank (which implies* $M \geq K$*), then:*

$$\lim_{P \to \infty} \left[\mathcal{C}_{MIMO}(P, \mathbf{H}) - \mathcal{C}_{MISO-BC}^{Sum}(P, \mathbf{H}) \right] = 0 \tag{2.22}$$

We can then use the affine approximation of the point-to-point MIMO sum-rate capacity in [38]:

$$\begin{aligned} \mathcal{C}_{MISO-BC}^{Sum}(P, \mathbf{H}) &\cong \mathcal{C}_{MIMO}(P, \mathbf{H}) \\ &\cong \sum_{i=1}^{K} \log\left(\frac{P}{K} \lambda_i\right) \\ &= K \log(P) + \log\left|\mathbf{H}\mathbf{H}^\dagger\right| - K \log K \end{aligned} \tag{2.23}$$

where $\lambda_1, \ldots, \lambda_K$ are the eigenvalues of $\mathbf{H}\mathbf{H}^\dagger$ and $\cong$ means that the difference between the two sides tends to zero as $P \to \infty$. Another interesting results is that at high-SNR the sum-rate capacity converges to the sum-rate obtained by using the same power for each user:

$$lim_{P \to \infty} \left[\mathcal{C}_{MISO-BC}^{Sum}(P, \mathbf{H}) - \log\left| \mathbf{I} + \sum_{i=1}^{K} \frac{P}{K} \mathbf{h}_i \mathbf{h}_i^\dagger \right| \right] = 0\,. \tag{2.24}$$

2.2.2.1 Do we need multiple antenna at receivers?

As we said, in mmWaves the transmitting BS must have multiple antennas to focus the transmitted power, but what about the users? In present cellular systems like 3G and LTE, cellphones already have from two to four receive antennas. As explained it will be even easier to fit multiple mmWave antennas into small-size devices, for instance Qualcomm already produced a mmWaves antenna module (named QTM052) which consists of an array of 4 antennas and their 5G modem

Snapdragon X50 supports up to 4 of these modules (one on each side of the phone to decrease hand-blockage probability) for a total of 16 antennas. We can thus expect that mmWaves systems will definitely take advantage of multiple antennas at the receiver. Which will anyway have to act as transmitter during the uplink if we consider a duplex system or any form of feedback. Nevertheless, from a system point of view, and as the number of users grows, the major advantage in downlink is given by increasing the number of antennas at the transmitting base station. Indeed in the broadcast channel the maximum multiplexing gain will be $\min(M, N \times K)$ which doesn't depend on N as long as we are in a typical situation where $N < M < K < NK$, so by increasing the number of antennas at receivers we don't gain in spatial degrees of freedom. Obviously, having many antennas at each receiver may allow both a power gain and a diversity gain equal to the number of antennas, but it will not affect significantly the two main aspects on which we focus: spatial multiplexing and multiuser diversity (we can note that the variability of the channel gain decreases with more antennas, so exploiting multi-user diversity becomes less impactful). In conclusion we will consider that the number of transmitting antennas M is large, and for simplicity receivers will have only one antenna.

2.2.3 Multi-user diversity in MIMO Rayleigh broadcast channels

Multi-user diversity consists in the advantage that can be taken from the presence of multiple users whose channels fade independently. With CSIT it is possible to schedule for transmission users when their channel quality is good, and delay their transmission when it is bad. With homogeneous channels among users, the simplest approach consists in transmitting only to the user with maximal channel gain. Or if we want to exploit spatial multiplexing at the same time, we can imagine to transmit the $S < K$ best users, where $S \leq M$ in order to be able to multiplex their signals. With heterogeneous users, i.e. users whose channel qualities have different averages, we should schedule one user when its channel is good with respect to its own distribution, in order to maintain the fairness of the system. In this way nevertheless we are not maximizing the system sum-rate. The effectiveness of multiuser diversity depends on the dynamic range of the channel, and in particular the tail of the channel gain probability distribution. If the tail of the distribution is heavy, we may expect to get a high multiuser diversity gain as some of the users will probably have a peak value which is very high. Obviously channel tracking is necessary to obtain a multi-user diversity gain but some reduced forms of feed-back (and not full CSIT) are sufficient.
One can actually quantify the multi-user gain. For instance in a SISO Broadcast

channel with Rayleigh fading, the $|h_i|^2$ are all i.i.d. exponential random variables with unit mean and it can be shown that for large K the maximum of K such random variables behaves with high probability as $\log K$ [39]. Thus, for large K, we have

$$C_{SISO-BC}^{Sum} = \log\left(P \log K\right) + o(1) \tag{2.25}$$

where $o(1)$ represents terms that vanish as K tends to infinity. The multiuser diversity gain result in a $\log K$ factor for SNR and therefore in a $\log \log K$ behaviour of the rate.

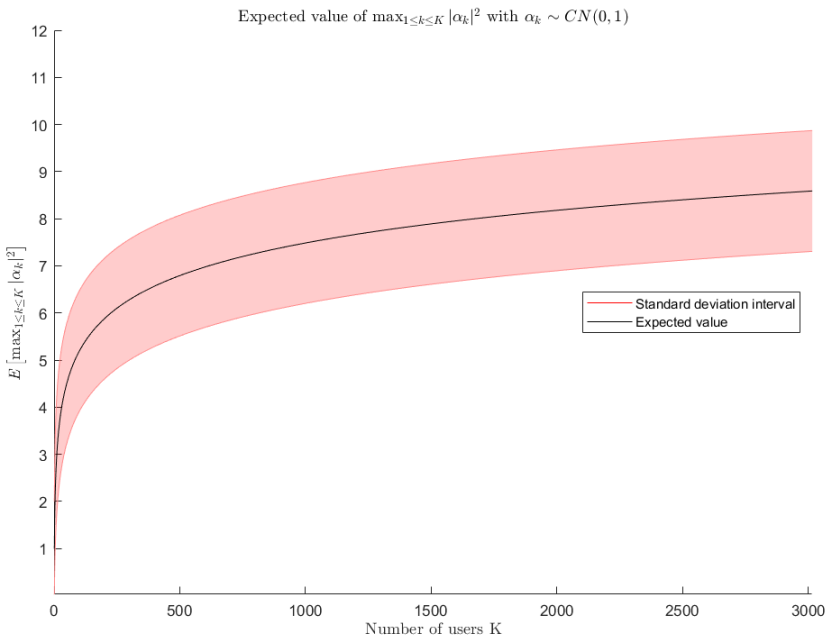


Figure 2.4: Expectation and variance of the maximum gain among K users

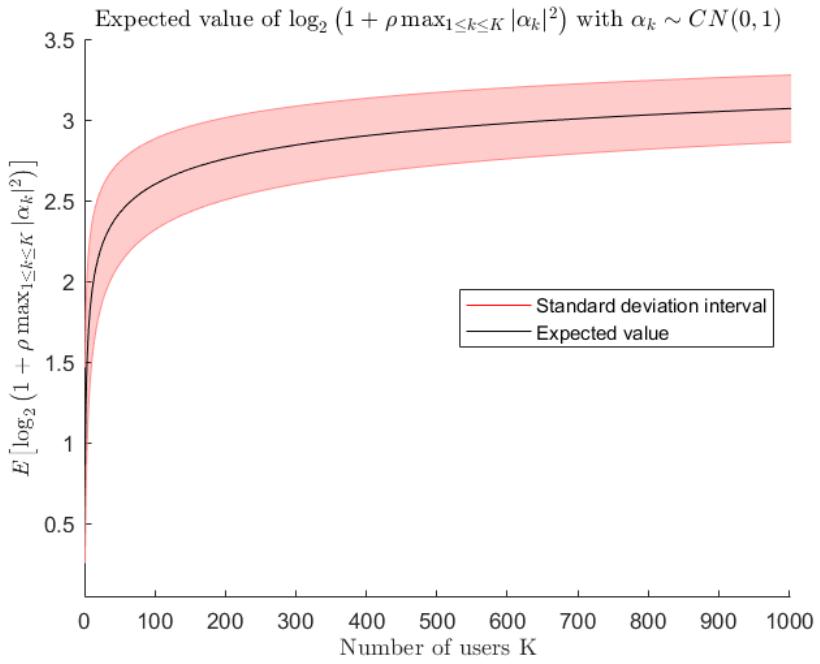


Figure 2.5: Expectation and variance of the maximum rate among K users

Turning to the general MIMO-BC case, we have the following result which includes multiuser diversity and spatial multiplexing:

Theorem 2.2.2. *For M and P fixed and any N, we have:*

$$\lim_{K\to\infty} \frac{\mathrm{E}\left[C^{Sum}_{MIMO-BC}\right]}{M\log(\log(KN))} = 1 \tag{2.26}$$

In [39] we can find an asymptotical expression for the sum-rate capacity of the MIMO-BC in high number of users regime. For fixed P and M we have:

$$C^{Sum}_{MIMO-BC} = M\log(\log K) + M\log\frac{P\log(N)}{M} + o(1) \tag{2.27}$$

where $o(1)$ represents terms that vanish as K tends to infinity.

This results conclude the current section and Chapter 2 as a whole. This chapter will allow us to tackle our particular problem of random beamforming in mmWaves channels starting from the general perspective of the capacity results for MIMO Broadcast Channels. We have seen that to approach capacity, two ingredients will be essential: spatial multiplexing and multiuser diversity. Theorem 2.2.2 is summing up the upper bound on both of them given by the sum-rate capacity.

The lack of already implemented systems for cellular downlink in mmWaves deprives us from any possible benchmark on which evaluating our random beamforming scheme, but now we know that we obtain with our opportunistic beamforming scheme a rate scaling like $M \log\log K + M \log \frac{P \log N}{M}$ we can be satisfied!

Chapter 3

Dealing with hardware and feedback limitation

The capacity achieving strategies presented in the last section unfortunately present a great complexity and require full channel state information at the transmitter, incomplete CSIT fatally degrades the performances of Dirty Paper Coding which is based on the possibility to predict interference at the receivers to pre-remove it. Some efforts have been made to obtain more flexible schemes still based on similar approaches like Ranked Known Interference [34], lattice based DPC [40], Vector Perturbation precoding [41] or Tomlinson-Harashima precoding [42]. These schemes still approach very tightly the capacity of the channel, nevertheless they remain more unpractical to implement than linear techniques which are much simpler and thus computationally cheaper. Transmission at mmWave frequencies involves another practical issue: to digitally precode the signal for each antenna and install a dedicated radio-frequency (RF) chain for each antennas may become prohibitively complex, expensive and energy consuming. That's why hybrid digital-analog precoding architectures are investigated. Finally we will introduce the so-called directional beamforming schemes which only necessitates a simple precoding architecture and can reach very good performances with limited feedback.

3.1 Linear precoding and beamforming

Linear precoding, sometimes referred to as transmit beamforming, consists in simultaneously transmitting each of the user data streams in an unique "transmit direction" (size N complex transmit vector). In this case we only have to decide what are the beamforming vectors optimizing the SNRs: the best solution is the so called Minimum Mean Squared Error (MMSE) beamforming, but its also common

to implement some sub-optimal versions that we will investigate. Then, we will illustrate the dependence of linear precoding from the quantity of feedback, and introduce opportunistic random beamforming as a limited feedback alternative.

3.1.1 Linear precoding with full CSIT

We are now going to focus on the main linear precoding techniques, namely MMSE and its two sub-optimal versions: Maximal Ratio Transmission (MRT) and Zero Forcing (ZF) precoding which respectively maximize the received signal power and minimize the interference power. Note that the MMSE, MRT and ZF linear precoders used in downlink were originally developed as filters for the receiver in the dual uplink MAC. In that context they were respectively called Wiener filter, matched filter and decorrelating filter. Linear precoding is characterized by the fact that our transmitted signal will be of the form:

$$\mathbf{x} = \sum_{k=1}^{K} x_k \sqrt{P_k} \mathbf{w}_k = \mathbf{W}\mathbf{x} \tag{3.1}$$

Where x_k is the information word intended for user i, that we associate to a unit norm transmit vector $\mathbf{w}_k \in \mathbb{C}^{M\times 1}$ (also called beam or spatial signature) and a power gain $P_k \geq 0$, such that $\sum_{i=1}^{K} P_i \leq P$. In the matrix formulation we use the precoding matrix $\mathbf{W} \triangleq \left[\sqrt{P_1}\mathbf{w}_1, \ldots, \sqrt{P_K}\mathbf{w}_K\right]$ and the vector of all users information words $\mathbf{x} \triangleq [x_1, \ldots, x_K]^T$. Then, considering a MISO BC scenario for simplicity (see [43] for the MIMO BC case), we get that the received signal at user k is:

$$y_k = \sqrt{P_k}\left(\mathbf{h}_k^{\dagger}\mathbf{w}_k\right) x_k + \sqrt{P_k} \sum_{j=1, j\neq k}^{K} \left(\mathbf{h}_k^{\dagger}\mathbf{w}_j\right) x_j + n_k \tag{3.2}$$

with, as usual $\mathbf{h}_k \in \mathbb{C}^{1\times M}$ representing the channel vectors and $n_k \in \mathcal{CN}(0, \sigma^2)$ being the independent noise terms. The SINR of user k is thus given by:

$$\mathrm{SINR}_k = \frac{\left|\sqrt{P_k}\mathbf{h}_k^{\dagger}\mathbf{w}_k\right|^2}{\sum_{i=1, i\neq k}^{K}\left|\sqrt{P_i}\mathbf{h}_k^{\dagger}\mathbf{w}_i\right|^2 + \sigma^2} \tag{3.3}$$

These SINRs obviously have to be jointly maximized as the terms $\left|\sqrt{P_i}\mathbf{h}_k^{\dagger}\mathbf{w}_i\right|^2$ aren't null in general. The MMSE approach aims to minimize the mean square errors which can be defined as:

$$\mathrm{MSE}_k = \mathbb{E}_{x_k, \mathbf{h}_x}\left[\frac{1}{P_k}\left|y_k - \sqrt{P_k}\cdot x_k\right|^2\right]$$

leading to the following convex minimization problem:

$$\left(\mathbf{w}_1^{MMSE}, \ldots, \mathbf{w}_K^{MMSE}\right) = \underset{\tilde{\mathbf{w}}_k : \|\tilde{\mathbf{w}}_k\|^2 = 1}{\arg\min} \sum_{k=1}^{K} \mathrm{MSE}_k \tag{3.4}$$

whose solution can be found using the Lagrangian multipliers technique or through iterative algorithms (see [44] for one example). For equal power distribution $P_k = \frac{P}{K} \quad \forall k$ we get [8]:

$$\begin{aligned} \mathbf{w}_k^{MMSE} &= \underset{\tilde{\mathbf{w}}_k : \|\tilde{\mathbf{w}}_k\|^2 = 1}{\arg\max} \frac{\frac{P}{K\sigma^2} \left|\mathbf{h}_k^\dagger \tilde{\mathbf{w}}_k\right|^2}{\sum_{i \neq k} \frac{P}{K\sigma^2} \left|\mathbf{h}_i^\dagger \tilde{\mathbf{w}}_k\right|^2 + 1} \\ &= \frac{\left(\mathbf{I}_N + \sum_{i=1}^{K} \frac{P}{K\sigma^2} \mathbf{h}_i \mathbf{h}_i^\dagger\right)^{-1} \mathbf{h}_k}{\left\| \left(\mathbf{I}_N + \sum_{i=1}^{K} \frac{P}{K\sigma^2} \mathbf{h}_i \mathbf{h}_i^\dagger\right)^{-1} \mathbf{h}_k \right\|} \end{aligned} \tag{3.5}$$

If we look at the SINR formula, two simplified approaches naturally arise: maximizing the numerator (done in MRT precoding) and minimizing the denominator (done in ZF precoding). Starting with the MRT approach, in order to maximize the received signal power, given the Cauchy–Schwarz inequality, one must simply choose:

$$\mathbf{w}_k^{MRT} = \underset{\mathbf{w}_k : |\mathbf{w}_k|^2 = 1}{\arg\max} \quad \left|\mathbf{h}_k^\dagger \mathbf{w}_k\right|^2 = \frac{\mathbf{h}_k}{\|\mathbf{h}_k\|} \tag{3.6}$$

This scheme coincides with the MMSE solution for $K = 1$ and it tends to be optimal when the users suffer from strong noise, which makes the interference negligible.

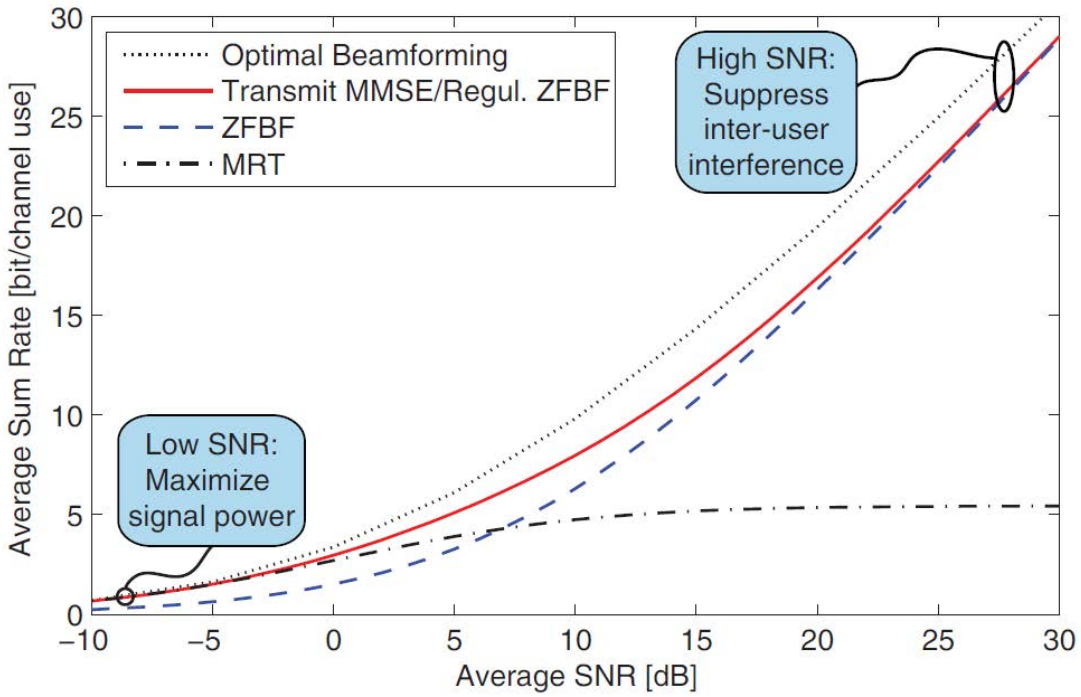


Figure 3.1: Average sum rate of the different linear precoding methods for K = 4 users as a function of the average SNR [8].

On the other hand, in high SNR regime the main problem becomes interference, that we can try to null by using spatial signatures which are orthogonal to all the users' channels but the intended one. We will thus choose the projection of $\mathbf{h}_k$ on the subspace which is orthogonal to the channel vectors of all the other users:

$$\mathbf{w}_k^{ZF} = \text{Proj}\left(\{\mathbf{h}_1, \ldots, \mathbf{h}_{k-1}, \mathbf{h}_{k+1}, \ldots, \mathbf{h}_K\}^{\perp}\right)\mathbf{h}_k \tag{3.7}$$

we can note that this projection is non-null only if $\mathbf{h}_k$ is linearly independent from all the other channel vectors. A necessary condition for this is that $K \leq M$, and if we should have more users than transmitting antennas we should limit ourselves to simultaneous transmission to only M of them. Remember that the maximal spatial multiplexing gain is of the MISO BC is $\min(M, K)$, with ZF precoding we can reach it but obviously not exceed it. The achievable sum-rate is given by:

$$R_{MISO-ZF}^{Sum} = \sum_{i=1}^{M} \log\left(1 + \frac{P_i \left|\mathbf{h}_k^{\dagger}\mathbf{w}_k^{ZF}\right|^2}{\sigma^2}\right)$$

which can be maximised using a waterfilling policy as in the point-to-point MIMO case. The projection operation can be expressed in matrix form and is thus a linear

operation. If the channel vectors of all the scheduled users are linearly independent the ZF vectors are simply given by the normalized columns of the left-inverse of the matrix $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \cdots \mathbf{h}_K]$:

$$\mathbf{w}_k^{ZF} = \frac{\mathrm{col}_k\left((\mathbf{H}^\dagger)^{-1_L}\right)}{\left\|\mathrm{col}_k\left((\mathbf{H}^\dagger)^{-1_L}\right)\right\|}. \tag{3.8}$$

where $\mathrm{col}_k(\cdot)$ selects the k^{th} column of the matrix to which it is applied. If we wanted to solve the more general problem:

$$\begin{array}{ll} \underset{\mathbf{w}_1,\ldots,\mathbf{w}_K}{\text{maximize}} & f\left(\mathrm{SINR}_1,\ldots,\mathrm{SINR}_K\right) \\ \text{subject to} & \sum_{k=1}^{K} P_k \left\|\mathbf{w}_k\right\|^2 \leq P \end{array} \tag{3.9}$$

with for instance $f\left(\mathrm{SINR}_1,\ldots,\mathrm{SINR}_K\right) = \sum_{k=1}^{K} \log_2\left(1+\mathrm{SINR}_k\right)$, the problem would be NP-hard, but we know that the general solution is (geometrically) somewhere between MRT (approached at low SNR) and ZF precoding (approached at high SNR).

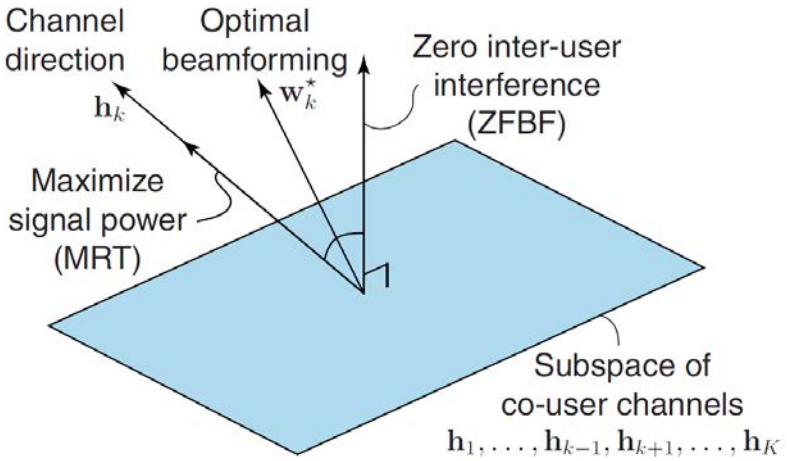


Figure 3.2: Geometrical representation of the optimal, ZF and MRT transmit vectors with respect to the channel vectors [8]

When the number of users becomes very high with respect to the number of transmitting antennas, while limiting the transmission to M simultaneous users we can fully exploit multiuser diversity, we have the following theorem from [45].

Theorem 3.1.1. *Let M, N, and the total average transmit power be fixed, and R_{BF} denotes the sum rate of optimal beamforming. Then*

$$\lim_{K\to\infty} \frac{E\left\{R_{BF}\right\}}{M \log\log KN} = 1 \tag{3.10}$$

We can see that linear precoding, or beamforming, can achieve the full spatial multiplexing and multi-user diversity gains when the number of users goes to infinity: the limit is the same than for dirty paper coding. But till this point we have still considered the availability of full CSIT, which only is an ideal hypothesis. What happens with limited feedback?

3.1.2 Linear precoding with limited feedback

As anticipated, having full CSIT might not be possible or at least practical in mmWave broadcast channels. Channel reciprocity allows to estimate CSI of the downlink channel through the CSI of the uplink in time-division-duplex (TDD) channels, but this isn't possible with frequency-division-duplexing (FDD) as the channel may be different at different frequencies. In this case, CSI is estimated by transmitting from the base station a sequence of pilots which are previously known by the receiver and allow it to infer the channel matrix from the received signal. The number of complex values needed to be fed-back from each users is then equal to $M \times N$, the size of the channel matrix, and can thus cause a large overhead when large arrays are used and the number of users increases. Furthermore a smaller wavelength corresponds to a shorter coherence time, so the training pilots must be sent more frequently to track changes of channel conditions, and this is even more problematic when users are highly mobile. That's why many alternative schemes using only a reduced feedback (often referred to as finite-rate feedback) have been studied (see [46] for an overview). For instance "antenna selection", "channel vector quantization", "per antenna phase quantization", "quantized equal gain codebooks" and "random vector quantization" are some classical feedback systems but we can also find intelligent processing techniques adapted to the nature of mmWave channel, like compressed sensing for sparse channels. Concerning beamforming, one usual approach is to feed-back only a quantized version of the channel state vector. This can be done for instance by using a finite set of vectors (codebook) which span all the directions of the channel vector space. This set of vectors must be known beforehand both by the base station and the users equipment, so it can either be predetermined by the standard or transmitted during the communication set-up (having each user independently generate random quantization codebooks have some advantages, as using the same quantization vectors may reduce the spatial multiplexing gain). Then, the UEs only need to feed-back an index which identifies the closest channel vector from the codebook, with B bits it can identify 2^B different channel vectors. It has been shown in [47] that:

Theorem 3.1.2. *The throughput achieved by finite-rate feedback-based zero-forcing with arbitrary quantization codebooks of fixed size is bounded as the SNR is taken*

to infinity.

So by using a fixed sized codebook we cannot arbitrarily increase the rate by increasing the power. In the same paper another very interesting result is provided:

Theorem 3.1.3. *A finite-rate feedback-based zero-forcing system in which per-user feedback is scaled at rate $B = \alpha \log(P)$ can achieve a multiplexing gain no larger than αM. Therefore, a necessary condition for achieving the full multiplexing gain of M is to scale feedback at least as $B = \log(P)$.*

This last theorem clearly shows us that for systems having a massive number of users also feedback strategies based on quantization of the channel vectors may not be able to scale and maintain optimal performances. We will see that opportunistic beamforming doesn't require CSIT but only some channel quality indicator which doesn't scale with the number of antennas and may allow beamforming gain (almost equal to coherent beamforming one), spatial multiplexing (by use of orthogonal beams) and multiuser diversity. Furthermore by its directivity it seems very adapted to the mmWave channel.

3.1.3 Opportunistic random beamforming

When full CSI is not available at the base station it will not be possible to directly align our beam with users' AoD. So the idea is to transmit randomly one or multiple transmit beams and then ask a feedback about the received SINR to users. As this feedback allows the base station to gain some knowledge on the channels quality and implement strategies which harness multiuser diversity we talk about opportunistic beamforming. The simplest of these schemes consists of having each users feeding-back its best SINR and the corresponding beam index to the base-station. Users can estimate their SINR by evaluating the received signal when predetermined pilots are sent by the base-station using each beam. We will suppose that the base station obtains the true SINR value but in practical systems we can imagine that the receiver only produces an estimate and directly sends a quantized version of it or uses it to derive a Channel Quality Indicator (CQI) to be transmitted like in 3GPP mobile communication standards. Then, the base-station can associate to each beam the user which reported the highest SINR, and transmit its data using that beam. If we want to obtain the full multiplexing gain, the base-station should choose, at each coherence interval, a set of M orthogonal vectors $\mathbf{w}_i^{RBF} \in \mathbb{C}^M \times 1$ such that $\left\|\mathbf{w}_i^{RBF}\right\|^2 = 1 \quad \forall 1 \leq i \leq M$ drawn from an isotropic distribution. Then we get for each users k a SINR corresponding to each different beam j:

$$\mathrm{SINR}_{k,j} = \frac{\left|\sqrt{P_k}\mathbf{h}_k^\dagger \mathbf{w}_j^{RBF}\right|^2}{\sum_{i=1,i\neq j}^{M}\left|\sqrt{P_i}\mathbf{h}_k^\dagger \mathbf{w}_i^{RBF}\right|^2 + \sigma^2} \tag{3.11}$$

and, if the number of users is very large, we can hope to obtain good performances by finding one for each of the beams one user aligned whose channel is aligned with it. The $\mathrm{SINR}_{k,j}$ are random variable depending both on the random choice of the beamforming vectors and on the distribution of the channels vectors. In the Rayleigh channel, having channel vectors with i.i.d. circularly-symmetric complex Gaussian distribution: $\mathbf{h}_k \sim \mathcal{CN}(0, \mathbf{I}_N)$, we have the following asymptotical results from [39]. If we keep M and P fixed but let the number of users K go to infinity we get that the achievable rate $\mathcal{R}_{RBF}$ tends to the BC sum-rate capacity:

$$\lim_{K\to\infty}\left(\mathcal{R}_{RBF} - \mathcal{C}_{MISO-BC}^{Sum}\right) = 0 \tag{3.12}$$

or equivalently, by remembering the high user regime analysis of the MISO BC, we also have for the random beamforming scheme that:

$$\mathcal{R}_{RBF} = M\log\log n + M\log\frac{P}{M} + o(1)\,. \tag{3.13}$$

We can imagine that with infinite users each beam is perfectly matched to one user. However, for fixed K and M we have:

$$\lim_{P\to\infty}\frac{\mathcal{R}_{RBF}}{\log P} = 0 \tag{3.14}$$

so the multiplexing gain of the scheme is formally null. Indeed, as we don't have perfect interference cancellation in high power regime the transmission becomes interference dominated and the SINRs are bounded to a constant value, which implies a bounded sum-rate and the hereinabove result. It seems therefore necessary to understand how many users are necessary to start to see the asymptotic behaviour of high user regime, with respect to the number of antennas. This question and its extension to the mmWave channel models will be one of the main focus of the thesis and is left to the next chapter. To conclude this section we mention that random beamforming is possible also with multiple receiving antennas at each user ($N > 1$). In this case we can choose between two different approaches: treating each of the antennas as a different user (thus involving an higher feedback, proportional to $N \times K$) or assigning only one beam per user and consider the total user's SINRs defined as:

$$\mathrm{SINR}_{k,j} = \frac{\mathbf{w}_j^{RBF\dagger}\mathbf{H}_k\mathbf{H}_k^\dagger\mathbf{w}_j^{RBF}}{\sum_{i=1,i\neq j}^{M}\mathbf{w}_i^{RBF\dagger}\mathbf{H}_k\mathbf{H}_k^\dagger\mathbf{w}_i^{RBF} + \sigma^2} \tag{3.15}$$

where $\mathbf{H}_k \in \mathbb{C}^{M\times N}$ is the channel matrix of user k.

3.2 Directional beamforming for mmWave channels

This section aims to show how directional beamforming appears to be a very adapted scheme in the mmWaves frequency range which entails strong constraints on the complexity of the transmitter architecture and displays a very high channel directivity, captured by the UR-SP and UR-MP models.

3.2.1 Digital, analog and hybrid beamforming

Till now we have used the base-band equivalent model and assumed that the BS was able to process digitally the signal without any limitation apart from computational complexity.

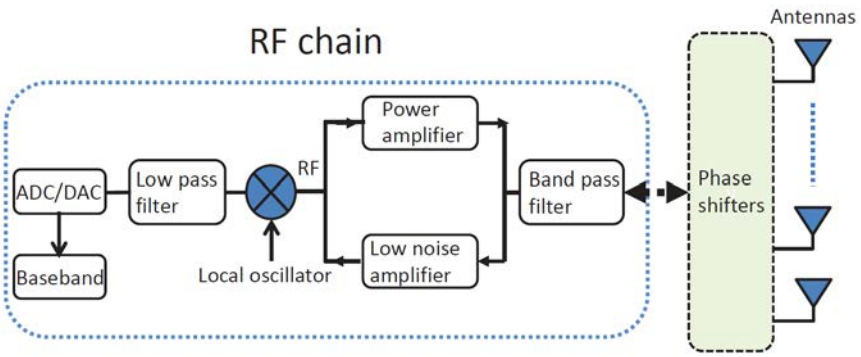


Figure 3.3: Block diagram of a RF chain connected to a phased array [9]

In practice, this digital beamforming architecture implies that after processing the signal at base-band one Radio Frequency (RF) chain is used for each antenna which is impractical for BS equipped with large mmWaves arrays. Indeed, RF chains (composed of filter, amplifier, mixer, DAC/ADC etc...) have in general a high cost and energy consumption, which are even more accentuated for mmWaves frequencies.

On the other hand we have fully analog beamforming which is performed by using one single RF chain to up-convert the signal from base-band to pass-band and then feed its output through independent phase-shifters and power amplifiers to each antenna. The phase shift and power amplification correspond to the multiplication by a complex coefficient in the digital domain. Unfortunately with this architecture the problem is that to have multiple beams carrying different streams we need to reproduce for each beam the whole structure from a new RF chain to a new set of antennas. This scheme is not flexible at all and cause a disadvantageous

trade-off between the multiuser spatial multiplexing gain limited by the number of RF chains and the beamforming gain limited to the fraction of antennas dedicated to each RF chain.

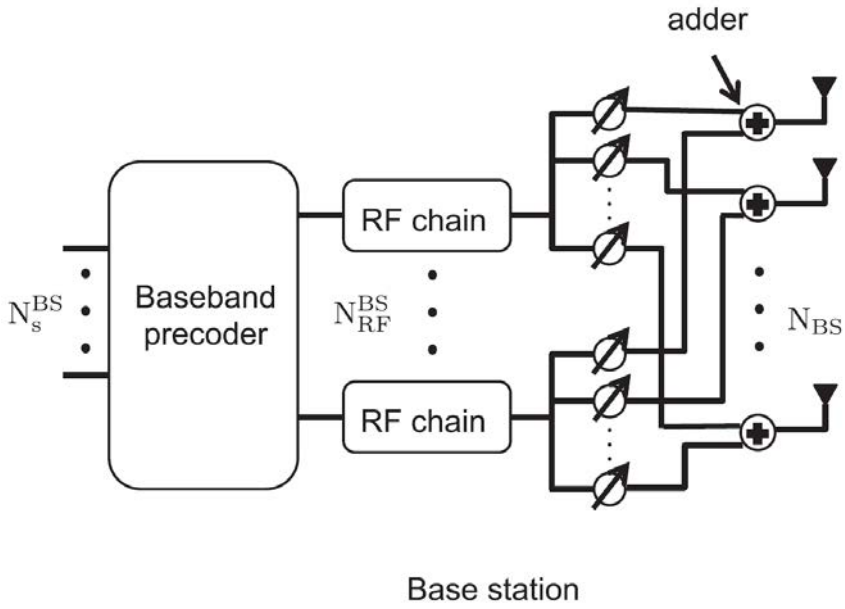


Figure 3.4: Fully connected hybrid beamforming architecture [10]

In order to find a balance between digital and analog beamforming some hybrid architectures have been developed. The most common hybrid beamforming architecture first perform some digital precoding and then, by connecting each RF chain to all the antennas through dedicated phase-shifters, allows to produce a number of beams equal to the number of RF chains while maintaining a beamforming gain equal to the total number of antennas [48]. With a reduced cost, power consumption and complexity, we can still approach the performances of digital beamforming, making hybrid architectures the main solution for the implementation of large mmWave beamforming arrays, independently from the chosen beamforming strategy. Furthermore directional beamforming schemes are particularly adapted to hybrid structures as they only need phase shifting (to compensate for the additional path-lengths caused by the AoD and obtain in-phase reception) but no single-antenna power control.

3.2.2 Single beam model

As anticipated, we can make use of beamforming to focus power towards the users through a good path and avoid interferences. We consider that our transmitting BS

is equipped with a large uniform linear array (ULA) of M antennas. The beam is produced by transmitting the desired signal through all the antennas with a constant phase delay $\pi\theta$ between each consecutive antenna, so that the i^{th} antenna will have a cumulative phase delay of $(i-1)\pi\theta$. This kind of transmitter is named phased array. Let's remember that the steering vector for a given normalized angle of departure (AoD) $\theta \in [-1, 1]$ is:

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{M}} \begin{pmatrix} 1 & e^{-j\pi\theta} & e^{-j\pi 2\theta} & \dots & e^{-j\pi(M-1)\theta} \end{pmatrix}^T . \tag{3.16}$$

Then, we simply obtain the transmitted beam as the product of the data stream $x[m]$ and the steering vector: $\mathbf{p}[m] = \mathbf{a}(\vartheta)\, x[m]$, leading to a received power for user k of: $\left|\mathbf{h}_k^\dagger \mathbf{p}\right|^2$. With full CSIT the optimal beamforming strategy is to beamform in the direction of the user which would have the highest SNR. In the UR-SP model we have that:

$$\mathbf{h}_k[m] = \alpha_k[m]\sqrt{M}\mathbf{a}(\theta_k) \tag{3.17}$$

and so the BS will transmit to the i^{th} user such that $i = \underset{k}{\operatorname{argmax}} |\alpha_k[m]|$ using $\mathbf{p}[m] = \mathbf{a}(\theta_i)\, x[m]$ and obtaining at the receiver a signal-to-noise ratio:

$$SNR = \rho M \max_k |\alpha_k[m]|^2 \left|\mathbf{a}(\theta_k)^\dagger \mathbf{a}(\theta_k)\right|^2 = \rho M \max_k |\alpha_k[m]|^2 = \rho M |\alpha_i[m]|^2$$

which leads, for constant ρ, to a maximum reliable transmission rate of:

$$\mathcal{R}_{SB} = \mathbb{E}\left[\log\left(1 + \rho \max_k |\alpha_k|^2 M\right)\right] . \tag{3.18}$$

To increase performances we can also adopt a waterfilling power distribution policy over time if the average SNR of future best users can be predicted.

As we were beamforming in the perfect direction in the previous case we had $\left|\mathbf{a}(\theta_i)^\dagger \mathbf{a}(\vartheta)\right|^2 = 1$ but it is interesting to see how this expression behave when the channel and the transmitting steering vectors don't coincide.

$$M\left|\mathbf{a}(\theta_k)^\dagger \mathbf{a}(\vartheta)\right|^2 = \frac{1}{M}\left|\sum_{n=0}^{M-1} e^{-\imath\pi n(\vartheta-\theta_k)}\right|^2 = \frac{1}{M}\left|\frac{\sin\frac{\pi(\vartheta-\theta_k)M}{2}}{\sin\frac{\pi(\vartheta-\theta_k)}{2}}\right|^2 \triangleq F_M(\vartheta-\theta_k) \tag{3.19}$$

which is called Féjer kernel of order M. This kernel defines a pattern composed of one central main lobe and $M-2$ secondary lobes. It has unitary mean, which reflects the fact that no power gain is obtained by beamforming totally randomly (on

the contrary the capacity decreases, due to Jensen's inequality and the concavity of the logarithm), and a peak value of M in $\vartheta - \theta_k = 0$, where full beamforming gain is obtained. So as we increase the number of antennas in the array we proportionally increase the potential beamforming gain but we reduce the width of the main lobe, which means that we'll need a better alignment between user's channel vector and the beam.

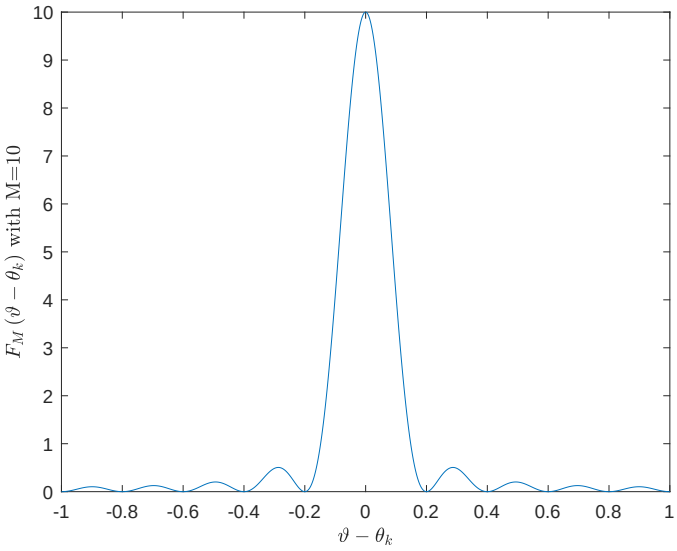


Figure 3.5: Féjer kernel of order 10, note zeros at multiples of $\frac{2}{M}$

If we apply a random beamforming approach we will not be able to aim at some user, instead the direction will be chosen randomly and then users will feed-back their SNR for that beam. Random beamforming schemes count on the fact that the number of users is sufficiently high to have some users in the main lobe with high confidence for any arbitrary AoD. Hopefully we will have many of them, and we will be able to select to which of them to transmit. To maximize sum-rate by harnessing multiuser diversity the base-station can transmit only to the user with the highest SNR for that beam. The maximum reliable transmission rate is then given by:

$$\mathcal{R}_{SB} = \mathbb{E}\left[\log\left(1 + \rho \max_k |\alpha_k[m]|^2 M \left|\mathbf{a}\left(\theta_k\right)^\dagger \mathbf{a}(\vartheta)\right|^2\right)\right]. \tag{3.20}$$

3.2.3 Multiple beams model

In order to exploit the spatial degrees of freedom resulting from the elevated number of antennas at the base station and the many receiving users, it is profitable to

transmit many beams simultaneously. As we have seen this kind of transmission schemes can be effectively implemented through hybrid digital-analog architectures. We know that in a general setting with M transmitting antennas and K single-receiver users the potential spatial multiplexing gain is given by $\min(M, K)$. Let's see what happens now that the structure of beams is limited to steering vectors. Considering that the total transmit power is equally distributed among each of the S beams we can express the signal to noise plus interference ratio for user k and beam b as:

$$SINR_{k,b} = \frac{|\alpha_k[m]|^2 \frac{\rho}{S} M \left| \mathbf{a}(\theta_k)^\dagger \mathbf{a}(\vartheta_b) \right|^2}{1 + \sum_{b' \neq b} |\alpha_k[m]|^2 \frac{\rho}{S} M \left| \mathbf{a}(\theta_k)^\dagger \mathbf{a}(\vartheta_{b'}) \right|^2} \tag{3.21}$$

From the behaviour of the Féjer kernel it is intuitive that we can only get M interference-free streams if we find M users equispaced in the angular domain, so that when it is beamformed each user channel is parallel to its beam and orthogonal to the other ones. In this ideal case we use the following equispaced beam transmission scheme with a number $S = M$ of beams:

$$\mathbf{p}_b[m] = \mathbf{a}(\vartheta_b)\, x_b[m] = \mathbf{a}\left(\vartheta + \frac{2(b-1)}{S}\right) x_b[m] \tag{3.22}$$

and we obtain for the user k in the direction $\theta_k = \vartheta_b$ a beamforming gain of:

$$M \left| \mathbf{a}(\theta_k)^\dagger \mathbf{a}(\vartheta_b) \right|^2 = F_M(\vartheta_b - \theta_k) = F_M(0) = M$$

while the received power from other beams is null:

$$M \left| \mathbf{a}(\theta_k)^\dagger \mathbf{a}(\vartheta_{b'}) \right|^2 = F_M(\vartheta_{b'} - \theta_k) = F_M\left(\frac{2(b'-b)}{M}\right) = 0\,.$$

We thus have a $SINR_{k,b} = \alpha_k[m] \frac{\rho}{S} M = \alpha_k[m]\rho$ and a sum-rate of:

$$\mathcal{R}_{MB} = \sum_{1 \leq k \leq M} \mathbb{E}\left[\log\left(1 + |\alpha_k|^2 \rho\right)\right] = M \times \mathbb{E}\left[\log\left(1 + |\alpha_k|^2 \rho\right)\right]. \tag{3.23}$$

The selected M users can achieve full beamforming gain and interference free reception only because we consider that they are perfectly aligned with the equispaced beams, this happens almost surely if we let $K \to \infty$ while keeping fixed the other parameters, but what we may expect in a real scenario is to have users only close to the beams directions. Obviously the higher is the number of users the more likely it is to find users which are almost in optimal directions. If we have enough users we will also have the choice, for each beam, to select the user with

the best fading fluctuation, or to perform superposition coding and SIC among users in the same sector in order to increase fairness. Researchers have thus tried to understand how many users are enough to get the full spatial multiplexing gain and exploit the multiuser diversity, in which ways we can keep good performances with partial feedback and what kinds of scheduling and multiple access schemes are more appropriate to increase fairness in asymmetric scenarios.

Chapter 4

Some open issues in opportunistic beamforming

The first studies on opportunistic beamforming schemes are quite old we can cite the seminal work of P. Visawanath and D. Tse [49] where all the fundamentals are already laid out. However, they were focusing on Rayleigh and Rician channels while only recently the attention has been focused on mmWave frequencies and results have been extended to the corresponding channel models. We will first see how one critical aspect to evaluate the opportunistic beamforming scheme is the necessary number of users for the system to work properly. Furthermore the simple opportunistic beamforming scheme exposed in the past chapter can be considered a basis over which many proposals of sophistication have been made. These alternatives generally aim at two objectives. First, try to guarantee a high fairness in the transmission when users qualities are heterogeneous. Secondly reduce the feedback containing information on the channels quality as much as possible. All these topics are still active lines of research, and we will report hereinafter some of the main contributions to conclude the first part of this thesis.

4.1 Necessary number of users for full multiplexing gain

As the number of users grows the probability to find at least one user whose channel is close to each of the orthogonal beams increases. On the other hand, the more transmitting antennas we have the more beams we have and the thinner they get, so it's harder to find properly aligned users. How must K scale with respect to M in order to still have a full multiplexing gain? In [50] the random beamforming performances in mm Wave channels have first been investigated analytically in asymptotic terms, by considering the variation of achievable rates depending on the

relationship between the number of transmitting antennas and the number of users. As an anticipation, the results were actually very promising if the number of users scales linearly with the number of antennas, the sum rate will also scale linearly with the number of antennas. But these results were obtained by considering only the one single path channel model (UR SP). The variable channel sparsity in the mm Wave band have been taken into account in a successive study [25], which demonstrates that the more the channel is rich in multipaths and scattering the more we need a large number of users to maintain the same performances. After reviewing the previous result for Rayleigh channels, we will report the main obtained results.

4.1.1 Rayleigh channel

The issue of the necessary number of users so that random beamforming achieves full multiplexing gain in Rayleigh fading channel scenario is exposed in [51]. What has been found, to put it in a nutshell, is that M cannot asymptotically grow faster than $\log K$ if we want to maintain the full multiplexing gain. From [39] we indeed have the following theorem.

Theorem 4.1.1. *Suppose the transmitter has M antennas, each receiver is equipped with a single antenna, and that we use random beamforming to users with the highest SINRs. Then we have that*

$$\text{if} \quad \frac{M}{\log K} = c_1 > 0 \quad \text{then} \quad \frac{C_{sum}}{M} = c_2 > 0 \tag{4.1}$$

where c_1 and c_2 are two positive constants independent of K. Whereas,

$$\text{if } \lim_{K\to\infty} \frac{M}{\log K} = \infty, \text{ then } \lim_{K\to\infty} \frac{R}{M} = 0 \tag{4.2}$$

So in Rayleigh channels the number of users must grow at least exponentially with the number of transmitting antennas if we want to have a linear scaling of the rate with M. This is one of the reasons why random beamforming schemes were not largely adopted in classical microwave mobile systems.

4.1.2 Directional beamforming in UR-SP channel

Suppose the base station has M antennas, the K receivers have only one antenna each and they are a fractional power of the number of transmitting antennas ($K = M^q$ with $q \in (0, 1)$), then it has been proved in [50] that the following results hold.

4.1.2.1 Single beam, single user

We start with the OBF scheme where one single beam in a random direction is used to transmit a known pilot, and after users report their SINRs it is used to transmit only to the user with the best SINR. The expected achievable rate $\mathcal{R}_1$ is given by

$$\mathcal{R}_1 = \mathbb{E}\left[\log\left(1 + \max_{1\leq k\leq K} |\alpha_k|^2 M \left|\mathbf{a}\left(\theta_k\right)^\dagger \mathbf{a}(\vartheta)\right|^2\right)\right] \tag{4.3}$$

and we have two theorems which bound its asymptotic behaviour in the high number of users and antennas scenario. In the first Theorem 4.1.2 no fading is considered but only the beamforming gain.

Theorem 4.1.2. *For $K = M^q$ and any given $q \in (0,1)$ we have asymptotic upper and lower bounds for $\mathcal{R}_1$ when $|\alpha_k| = 1$ for all k given by*

$$\log\left(1 + M^{2q-1-\epsilon}\right) \lesssim_M \mathbb{E}\left[\log(1+Z)\right] \lesssim_M \log\left(1 + M^{2q-1+\epsilon}\right) \tag{4.4}$$

for any sufficiently small $\epsilon > 0$, where $Z = \max_k M \left|\mathbf{a}\left(\theta_k\right)^\dagger \mathbf{a}(\vartheta)\right|^2$ and $x \lesssim_M y$ means that $\lim_{M\to\infty} x/y \leq 1$.

Then in Theorem 4.1.3 fading is taken in account and the previous result is confirmed.

Theorem 4.1.3. *For $\alpha_k \overset{i.i.d.}{\sim} \mathrm{CN}(0,1)$ and $q \in \left(\frac{1}{2}, 1\right)$ we have*

$$\lim_{M\to\infty} \frac{\mathcal{R}_1}{\mathbb{E}\left[\log\left(1 + \rho M \max_k |\alpha_k|^2\right)\right]} = 2q - 1 \tag{4.5}$$

and on the other hand when $q \in \left(0, \frac{1}{2}\right), \mathcal{R}_1 \to 0$ as $M \to \infty$.

These two theorems are basically saying that if the number of antennas grows like a fractional power of the number of antennas, we will still asymptotically have a full multiuser diversity gain and a beamforming gain behaving like $\frac{K^2}{M}$. We know that the multiuser diversity gain is given by

$$\mathbb{E}\left[\log\left(1 + \rho M \max_k |\alpha_k|^2\right)\right] \sim_M \log\left(M \log\left(K\right)\right) = \log\left(M \log\left(M^q\right)\right) \tag{4.6}$$

so we get $\mathcal{R}_1 \sim_M (2q-1)\log\left(M \log\left(M^q\right)\right)$.

4.1.2.2 Multiple beams, best user-beam pair selected

We now consider to sequentially transmit $S = M^l$ with $l \in (0,1)$ equispaced beams during the training phase and then transmit only for the best user-beam pair. The expected rate $\mathcal{R}_S$ of the multiple training beams scheme is given by

$$\mathcal{R}_S = \mathbb{E}\left[\log\left(1 + \max_{1\le k\le K}\max_{1\le b\le S} |\alpha_k|^2 M \left|\mathbf{a}(\theta_k)^\dagger \mathbf{a}(\vartheta_b)\right|^2\right)\right] \tag{4.7}$$

This time, as stated by the following theorem the many beams will allow us to compensate for sparsity of users. No fading is considered in this scenario.

Theorem 4.1.4. *For $K = M^q, S = M^l$ and any $l, q \in (0,1)$ such that $l + q < 1$, we have asymptotic lower and upper bounds on $\mathcal{R}_S$ in the case of $|\alpha_k| = 1, \forall k$, given, for any sufficiently small $\epsilon > 0$, by*

$$\log\left(1 + M^{2q+2l-1-\epsilon}\right) \lesssim_M \mathbb{E}\left[\log(1+Z')\right] \lesssim_M \log\left(1 + M^{2q+2l-1+\epsilon}\right) \tag{4.8}$$

where $Z' = \max_k Z'_k$ and $Z'_k = \max_b M\left|\mathbf{a}(\theta_k)^\dagger \mathbf{a}(\vartheta_b)\right|^2$.

Thus we achieve a fraction of the full multiplexing gain given by the following theorem.

Corollary 4.1.4.1. *For $K = M^q, S = M^l$ and any $l, q \in (0,1)$ such that $\frac{1}{2} < l + q < 1$ we have*

$$\lim_{M\to\infty} \frac{\mathbb{E}\left[\log(1+Z')\right]}{\log(1+M)} = 2(q+l) - 1 \tag{4.9}$$

4.1.2.3 Multiple beams, multiple users

In this version of the scheme, after the training period where a known pilot is transmitted through every beam in order for the user to get an estimate of the $SINRs$ to feedback, the base station will simultaneously transmit on each beam b the data corresponding to the user κ_b which has the highest $SINR$ for that beam

$$\begin{aligned}\mathcal{R}_{\kappa_b} &= \mathbb{E}\left[\log\left(1 + \max_{1\le k\le K} \mathrm{SINR}_{k,b}\right)\right] \\ &= \mathbb{E}\left[\log\left(1 + \frac{\rho M \left|\mathbf{a}(\theta_{\kappa_b})^\dagger \mathbf{a}(\vartheta_b)\right|^2}{1 + \sum_{b'\neq b} \rho M \left|\mathbf{a}(\theta_{k_b})^\dagger \mathbf{a}(\vartheta'_{b'})\right|^2}\right)\right]\end{aligned} \tag{4.10}$$

where we assumed that $|\alpha_k| = 1, \forall k$ for simplicity. So the achievable sum-rate with S equispaced beams transmitted simultaneously is given by:

$$\mathcal{R}_M = \sum_{b=1}^{S} \mathcal{R}_{\kappa_b} \tag{4.11}$$

Considering no fading we have the following bound on the per-beam rate.

Theorem 4.1.5. *For $K = M^q, S = M^l$ with $q \in (0,1)$ and l in $\left(0, q - \frac{\epsilon}{2}\right)$, asymptotic upper and lower bounds on the per-user rate $\mathcal{R}_{\kappa_b}$ of selected user κ_b for fixed total transmit power $P_t = 1$ are given by*

$$\log\left(1 + M^{2q-1-l-\epsilon}\right) \lesssim_M \mathcal{R}_{\kappa_b} \lesssim_M \log\left(1 + M^{2q-1-l+\epsilon}\right) \tag{4.12}$$

for any sufficiently small $\epsilon > 0$.

This implies that the sum-rate scales like $M^l \log\left(1 + M^{2q-1-l}\right)$, so in the limit of infinite transmission antennas and users we are both maximally exploiting the multiplexing capability of the channel and getting a beamforming gain.

4.1.3 Directional beamforming in UR-MP channel

To have only one path would be ideal for opportunistic beamforming purposes but in practice it is more realistic, even at mmWaves frequencies, to have multiple reflections or scattering clusters. We are thus going to present an extension of the results of the previous subsection to the UR-MP channel. These results and their proofs can be found in [25].

The number of multipaths L will be expressed as a function of M, as this allows to get asymptotical results on the rate behaviour for a range of channels going from the single-path to the Rayleigh channel. Indeed we have that the UR-MP channel converges to a Rayleigh channel when $L = M^\beta$ with $\beta > 1$ in the sense described by the following theorem.

Theorem 4.1.6. *Under the UR-MP model when $L = M^\beta$ we have*

$$\mathcal{L}\left(\mathrm{h}_k | \theta_{k,1}, \cdots, \theta_{k,L}\right) = \mathcal{CN}\left(0, \mathbf{R}\left(\theta_{k,1}, \cdots, \theta_{k,L}\right)\right) \tag{4.13}$$

where $\mathcal{L}\left(\mathbf{h}_k | \theta_{k,1}, \cdots, \theta_{k,L}\right)$ is the distribution of $\mathbf{h}_k$ conditioned on $(\theta_{k,1}, \cdots, \theta_{k,L})$. Furthermore we have for any $\epsilon > 0$,

$$\Pr\left(\left|\mathbf{R}\left(\theta_{k,1}, \cdots, \theta_{k,L}\right) - \mathbf{I}\right|_E \geq \epsilon \mathbf{1}\mathbf{1}^T\right) \to 0 \tag{4.14}$$

as $M \to \infty$, if $\beta > 1$ where the probability is computed according to $\theta_k|_i \overset{i.i.d.}{\sim}$ Unif$[-1,1]$, $|\cdot|_E$ represents the element-wise absolute value and $\mathbf{1}$ is a vector of ones. That is, the conditional channel covariance matrix $\mathbf{R}\left(\theta_{k,1}, \cdots, \theta_{k,L}\right)$ converges to $\mathbf{I}$ uniformly in elements in probability.

We already knew that in Rayleigh channel an exponential scaling of the number of users with respect to the number of antennas was necessary to have a full

multiplexing gain. The following results show us what happens in between the Rayleigh channel and the fixed number of paths models. These scenario can be modelled by a logarithmic, fractional power and linear (or faster) growth of the number of users with respect to the number of antennas. Let c_u and c'_u be positive constants, the following table summarizez the necessary number of users depending on the scattering richness.

Scattering richness	Sufficient number of users
L fixed	$K \simeq M$ (linear)
$L = \log(M)$	$K \simeq M^{1+c_u}$ (polynomial)
$L = M^{\beta}, 0 \leq \beta < 1$	$K \simeq e^{c'_u M^{\beta}}$ (sub-exponential)
$L = M$ or faster	$K \simeq e^{c'_u M}$ (exponential)

Table 4.1: Sufficient number of users for linear sum-rate scaling with M

4.2 Achieving fairness in opportunistic beamforming

In a multiuser scenario, the distribution of data transmission rates among users is a crucial issue. In many practical contexts, the base-station must try to guarantee to all users, which experiment different channel conditions, some minimal performances. On the other hand from a total system throughput point of view, it may be profitable to give priority (i.e. assign more time, power or degrees of freedom) to users with a better channel, thus increasing the disparity between data rates of different users. Opportunistic beamforming in its simplest version consists in simply select the user with the best SINR for each beam, which may result in a particularly unfair distribution of rates. That's why it is generally associated to a user scheduling or radio resource management protocol which takes care of balancing the rates of users. In order to compare different schemes under the fairness aspect, it should first be well defined and dedicated metrics are needed.

4.2.1 Fairness concept and metrics

We already saw in Chapter 2 the concept of achievable rates region. From the rate region it is possible to find the Pareto optimal point of a given scheme and compare it to another one but with large number of users rate regions are difficult

to obtain as we should vary all the parameters involved in the protocol (like power distribution, number of beams etc.). In practice network operators generally face the fairness problem by scalarization, which consists in selecting a network utility function that tries to balance the sum-rate and fairness and concentrate both aspects of performance in one single figure. We already introduced the symmetric rate, which is the maximum rate simultaneously achievable by all users but it becomes unadapted in scenarios where the fatally low performances of users with really bad channels may totally hide the ones of the other users. The fairness problem being of great interest in many different study disciplines, starting from economics, some very precise and significative fairness optimality criteria have been derived, we are going to state the most important ones as reported in [52]. Let be $\mathbf{R} = (R_1, \ldots, R_K)$ the allocation of rates among the users, the set of possible allocations $\mathcal{R} \subseteq \mathbb{R}_+^K$ is a subset of the capacity region of the channel which will depend on the utilized transmission scheme.

Definition 4.2.1. *An allocation $\mathbf{R} \in \mathcal{R}$ is called*

- *(globally) optimal if it maximizes the sum-rate $\sum_{k=1}^{K} R_k$*
- *(strictly) Pareto optimal if there is no solution $\mathbf{R}' \in \mathcal{R}$ dominating it, i.e. such that $R'_k \geq R_k$ for all $k = 1, \ldots, K$ and $R'_{k_0} > R_{k_0}$ for some $k_0 \in \{1, \ldots, K\}$*
- *max-min fair if for each $k \in \{1, \ldots, K\}$ increasing R_k must be at the expense of decreasing R_l for some l such that initially $R_l < R_k$. If a max-min fair allocation exists, then it is unique and strictly Pareto optimal.*
- *proportionally fair if for each other allocation $\mathbf{R}' \in \mathcal{R}$ we have*

$$\sum_{k=1}^{K} \frac{(R'_k - R_k)}{R_k} \leq 0.$$

  If a proportionally fair allocation exists on $\mathcal{R}$, then it is unique and it is the solution of the following maximization problem

$$\max_{\mathbf{R} \in \mathcal{R}} \sum_{k=1}^{K} \log R_k \,.$$

- *α-fair optimal if it solves the following maximization problem where α is a real number:*

$$\max_{\mathbf{R} \in \mathcal{R}} \sum_{k=1}^{K} \frac{R_k^{1-\alpha}}{(1-\alpha)} \,.$$

  We have that an α -fair optimal policy is globally optimal when $\alpha \to 0$, proportionally fair when $\alpha \to 1$, and max-min fair when $\alpha \to \infty$.

4.2.2 Scheduling the proportional-fair policy

One first approach to the fairness problem is to solve it from a higher control layer by scheduling users by following a policy which may take into account the state of the channel (channel aware scheduling) or not (channel unaware scheduling). The latter are often the simplest and may consist in a trivial first-in first-out rule or Round-Robin scheduler (where each user is given a fixed time, after which it is put at the end of the users' queue). Channel aware protocols require feedback but are much more efficient as they allow to include the expected rate obtained by scheduling one user in the decision.

Single user opportunistic communication is the simplest version of channel aware scheduling, as it requires channel qualities to select the user who would lead to the highest rate. This may obviously lead users with a poorer channel to never be scheduled. Proportionally Fair (PF) scheduling, which is one of the most used schedulers in telecommunications, aims to solve this issue by transmitting to the user whose potential rate is the highest with respect to the average rate he has achieved in the past. This is done by transmitting to user k^* such that

$$k^* = \arg\max_{1\leq k\leq K} \frac{R_k^\alpha(t)}{T_k^\beta(t)} \tag{4.15}$$

where R_k is the rate potentially achievable in the incoming time slot by user k, T_k is its historical average data rate and α and β are two parameters allowing to tune the balance between fairness and current rate maximization to the desired value. T_k can be computed as

$$T_k(t+1) = \begin{cases} \left(1-\frac{1}{t_e}\right) T_k(t) + \frac{1}{t_c} R_k(t), & k = k^* \\ \left(1-\frac{1}{t_c}\right) T_k(t), & k \neq k^* \end{cases} \tag{4.16}$$

With t_c defining the time interval over which we want to achieve fairness. In a scenario where we transmit simultaneously to many users we have to select at each transmission slot a set of users I^*. As the rates of the different users are in general interdependent, we cannot simply maximize the rates-throughput ratios individually, nor the sum of the ratios. In [53] it is shown that the criterion

$$I^* = \arg\max_I \prod_{k\in I} \left(1 + \frac{R_{k|I}(t)}{(t_c-1)\,T_k(t)}\right) \tag{4.17}$$

is the one maximizing the sum over all users of the logarithms of the throughputs

$$\sum_{1\leq k\leq K} \log\left(T_k(t+1)\right) = \log\left(\prod_{1\leq k\leq K} T_k(t+1)\right). \tag{4.18}$$

If compared to the Definitions 4.2.1, we can notice that this means that a form of long term proportional fairness will be achieved. From an intuitive point of view this criterion guarantees than none of the average throughput is too low, otherwise the product would be largely reduced. This scheme can be applied to opportunistic beamforming with quasi-orthogonal beams, by scheduling one single user per beam. When the rates of the different users are independent from each other, Eq. 4.17 is simplified to:

$$I^* = \arg\max \sum_{k \in I} \left(\frac{R_{k|I}(t)}{T_k(t)} \right) \tag{4.19}$$

Another notable policy based on proportional fair scheduling is the Modified Largest Weighted Delay First which guarantees to all users a minimum bit-rate when used with a token bucket system.

4.2.3 Orthogonal and non-orthogonal schemes

One usual approach to the general fairness problem is, in systems based on the subdivision of time and frequency resources among users (TDMA, FDMA, etc...), to assign a large portion of bandwidth or transmission time to users with poor channel quality. This solution is adopted in 3GPP standards like 4G LTE and 5G where spectrum and time are divided in Physical Resource Elements which consist in one carrier sub-band and one OFDM symbol transmission time. Scheduling policies, where the transmission to one user is prioritized depending on some objective function may be considered as intelligent forms of TDMA. Although very practical, this kind of orthogonal schemes are not optimal when the channels of the different users are heterogeneous (for instance if some of them are far away and others are closer to the BS). Indeed we saw that techniques based on superposition of signals intended to the different users resulted in better sets of rate and it is proven from an information theoretic point of view [54]. Enforcing fairness through NOMA has also another advantage over TDMA as by simultaneous transmission it allows to keep a low transmission latency even if the bit-rate is low while TDMA systems may involve large delays for weak users who would have few assigned slots. That's why the application of NOMA to opportunistic beamforming schemes in mmWaves gave birth to a flourishing line of research [55]. The main idea is to divide the users in different sectors, each corresponding to one dominant beam which will be used to transmit the superposition of the signals intended to the users of the sector. Then SIC is performed among the users belonging to the same sector.

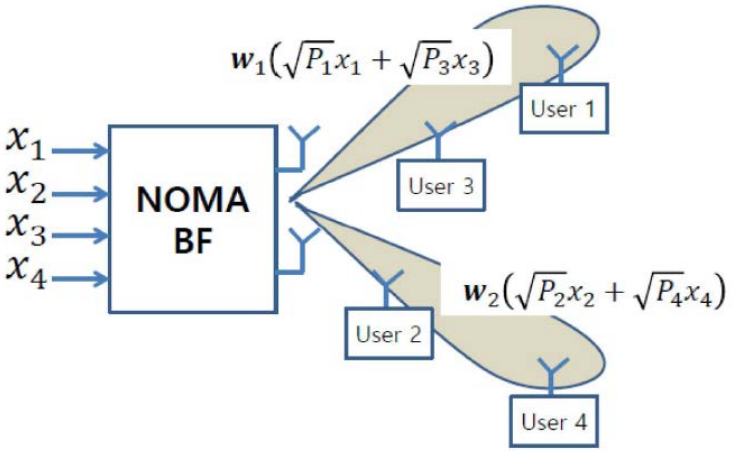


Figure 4.1: Scheme of superposition coding combined to beamforming [11]

In Chapter 2 we saw that SIC required ordering among channel qualities in order to be performed, and that the absence of an unambiguous ordering in the MISO (and MIMO) BC context didn't allow us to apply it directly. However, with respect to each single beam b, the channels of the users can be ordered according to their $SINR_{k,b}$. Then we can enforce users to fully decode and cancel interference created by transmissions to other weaker users in the same sector. To avoid putting too much computational burden on the users, some strategies consider to perform SIC one couple of users at a time. In this way we can associate to one user with a bad channel quality one user with a good channel quality through SIC guarantee decent performances to the user with the bad channel almost without sacrificing the rate of the user with the good channel. More sophisticated and optimal pairing schemes have been developed [56]. Comparison of beamforming in mmWaves channels using OMA and NOMA have been performed showing the advantages of the latter [57]. The problem of optimal power allocation between users is treated in [58]. And we can find results on NOMA combined to hybrid precoding in mmWaves in [59] and [60].

Part II

Computational analysis and numerical results

Chapter 5

Methodology and approach

The remaining part of the thesis will be carried out with the help of computer simulations and numerical methods. Before presenting and analysing the results in the next chapter we are going to first state the objective of the analysis and then describe and motivate the chosen methodology to tackle each problem. Some details will also be given about utilized software and the simulation parameters.

5.1 Tackled problems and objectives

Now that the framework and the main relevant literature have been exposed, let's see which open problems we can try to solve.

In the previous chapters many valuable asymptotical bounds on the performances of OBF schemes have been stated. As we have seen, asymptotical results in large number of users regime showing that opportunistic beamforming schemes are particularly adapted to the mmWave channel have been published.The main results being that in a mmWave channel with finite number of transmission paths, as the number of antennas in the system tends to infinity a linear scaling of the number of users with respect to the number of antennas M is sufficient to obtain a linear scaling of the sum-rate with M, contrary to the Rayleigh channel case where an exponential scaling of users is needed. Despite this fact a complete analysis for finite number of users is still missing. This statement exposes a promising property of mmWaves channels, but is valid only as the number of users and antennas tends to infinity, and we are not sure that, although large, the size of these parameters in real networks will be sufficient to approximate the asymptotical behaviour. We will thus try to investigate the problem over plausible ranges of the parameters. In this view we will first introduce an efficient way to actually compute the expected rates of the simplest schemes from the parameters probability distributions, without passing through simulations whose complexity scales with the number of users

and antennas. This allows a complete characterization of the performances over the parameter space for the simpler schemes. Then, we will try to understand what are the actual number of users, as a function of the number of antennas that are necessary to approach optimal performances in multi-beam OBF. This time the problem will not be approached from an asymptotic point of view but by concretely investigating the whole range of plausible parameters. On one hand by using the direct computation of the expected value of the relevant quantities when their expression is simple enough and in the other through Monte Carlo simulations where an analytical solution would be too complicated. Finally, we will step aside from sum-rate analysis to focus on the distribution of rates among users with heterogeneous channels. The fairness in simulated heterogeneous scenarios will be compared between some intuitive policies and the proportional-fair one. The possibility to simultaneously schedule multiple users per beam through OMA or NOMA should also be considered.

5.2 Stochastic characterization of the multiuser and beamforming gains

5.2.1 Beamforming only

As we have seen, the beamforming gain is given by the Féjer kernel function computed on the difference between the beam and the user normalized angles.

$$F_M\left(\vartheta - \theta_k\right) = \frac{1}{M}\left|\frac{\sin\frac{\pi(\vartheta-\theta_k)M}{2}}{\sin\frac{\pi(\vartheta-\theta_k)}{2}}\right|^2 .$$

In our random beamforming scenario, the beam angle and users' angles are independent and uniformly distributed over $[-1,1]$. So the difference $\delta_k = \vartheta - \theta_k$ is also a random variable uniformly distributed over $[-1,1]$. We are now interested in the distribution of $F_M\left(\delta_k\right)$. Unfortunately the Féjer function is not invertible, so we cannot directly apply the following theorem on invertible transformations of random variables (that we state hereafter as we'll need it later).

Theorem 5.2.1. *Let X be a continuous random variable with probability density function f_X and support I where $I = [a,b]$. Let $g : I \to \mathrm{Re}$ be a continuous monotonic function with inverse function $h : J \to I$ where $J = g(I)$. Let $Y = g(X)$ be our transformed random variable. Then the probability density function f_Y of Y satisfies:*

$$f_Y(y) = \begin{cases} f_X(h(y))\left|h'(y)\right| & \text{if } y \in J \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

One could note that the Féjer kernel is piece-wise monotone so inverses exist over injective subsets of the support and we could sum-up the densities coming from each subset but these inverses still doesn't have an a closed form expression so this approach doesn't lead to any solution. Hence we will try to obtain the distribution numerically in MATLAB.

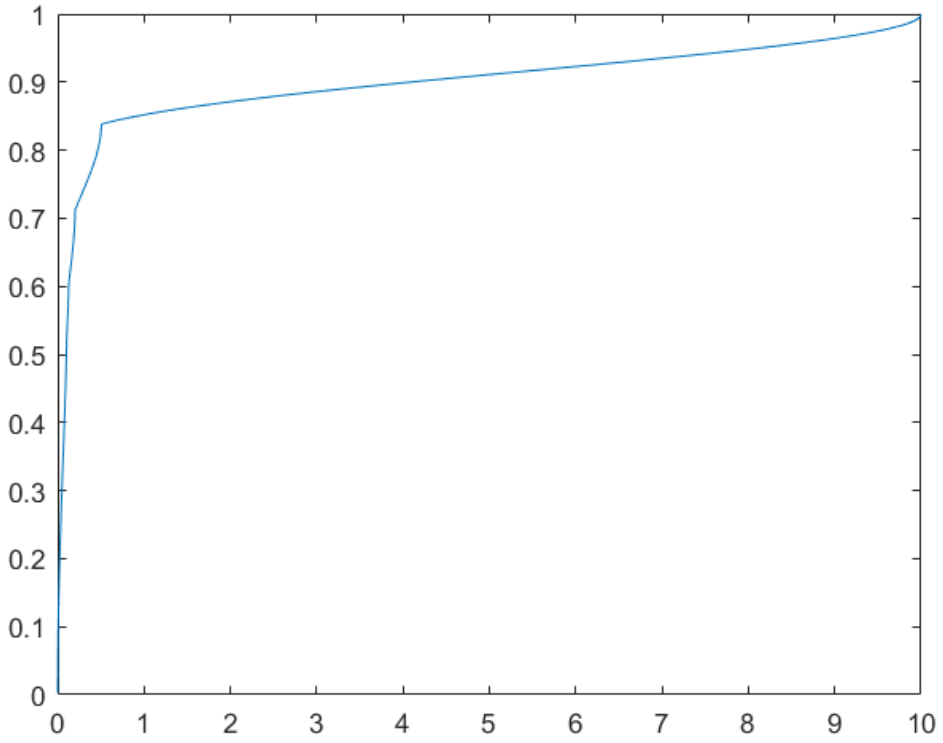


Figure 5.1: CDF of the Féjer kernel of a uniformly distributed angle for M=10

The probability density function can be approximated through a normalized histogram with bins of minimal size. But MATLAB also contains specific built-in functions to create non-parametric distributions from empirical data: **ecdf()**, which stands for empirical Cumulative Distribution Function (CDF), allows to create a discrete CDF from samples, that we will simply obtain by computing the Féjer kernel over $[-1, 1]$ as δ_k is considered to be uniformly distributed. If we desire a continuous CDF we can create with **makedist()** a "Piecewise Linear Distribution" object which computes as before the discrete CDF for the points present in the sample vectors and then linearly connect the CDF values to form a continuous curve. In this way it was possible to obtain the following plot of the

CDF of the Féjer kernel of order ten applied to a uniform angle distribution. We can see from Figure 5.1 that with high probability (over 0.8) we get a gain inferior to 1, which can be associated to the high probability of being outside of the main lobe. Then, the CDF increases almost uniformly for values from 1 to 10, which implies that the Probability Distribution Function (PDF) is almost constant over these values. If we compute the PDF we can actually note a pick around 10, but also around the secondary lobes peak values as the Féjer kernel has null derivative around these values.

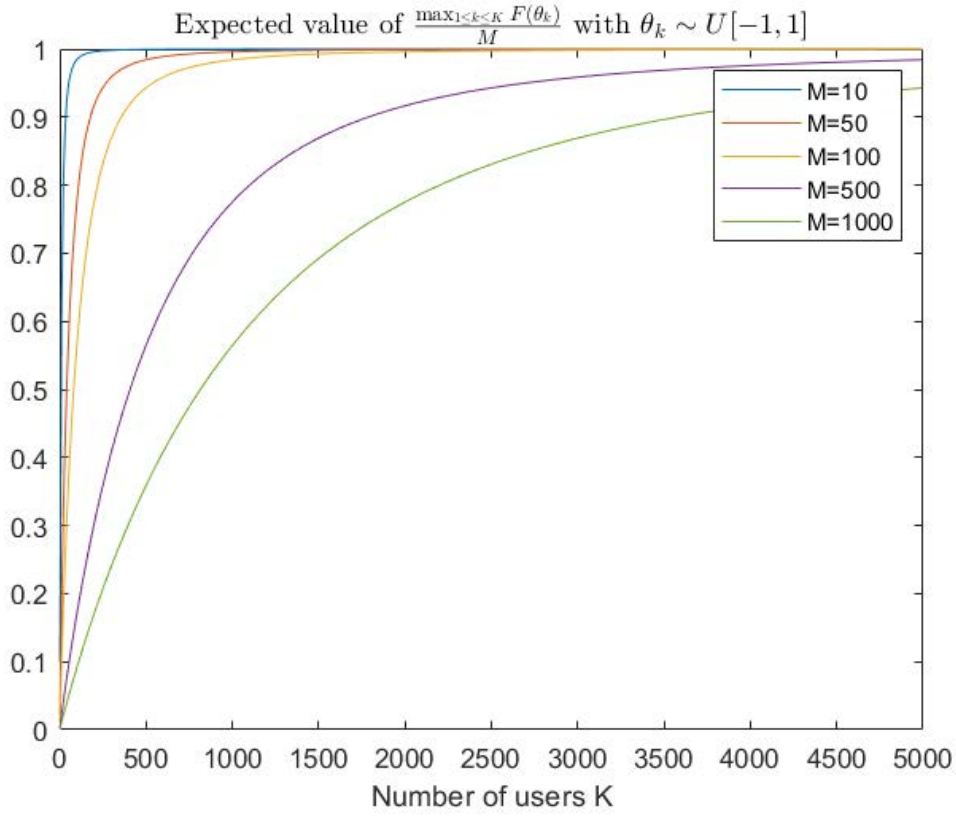


Figure 5.2: Fraction of the maximal beamforming gain achieved in average by Single Beam OBF with no fading as a function of K.

This CDF can be used to compute the CDF of the maximum of K r.v. with the same distribution by simply elevating it at the K^{th} power. Finally, the distribution of the maximum can for instance be used to compute their expected value and see how it evolves as a fraction of its upper bound M as done in Figure 5.2.

5.2.2 Beamforming and fading

We now consider both components of the gain: $g_{\max,K} = \max_{1\leq k\leq K} |\alpha_k|^2 F_M(\delta_k)$ indeed depends on the beamforming gain and on the fading coefficient whose square modulus follows an exponential distribution. This follows from:

$$\alpha_k \sim CN\left(0,\sigma^2\right) \Rightarrow |\alpha_k| \sim \text{Rayleigh}\left(\frac{\sigma}{\sqrt{2}}\right) \iff |\alpha_k|^2 \sim \text{Exp}\left(\frac{1}{\sigma^2}\right) \tag{5.2}$$

To obtain the distribution of the product we could make use of the following theorem.

Theorem 5.2.2. *If X and Y are two independent, continuous random variables, described by probability density functions f_X and f_Y then the probability density function of $Z = XY$ is:*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx \tag{5.3}$$

As we don't have an analytical expression for the distribution of $F_M(\delta_k)$ in order to get the product variable CDF shown in the next figure we used the MATLAB toolbox CUPID [61] which allows to effectively manipulate random variables.

Again, once we have the CDF of $|\alpha_k|^2 F_M(\delta_k)$ it's immediate to get the CDF of $\max_{1\leq k\leq K} |\alpha_k|^2 F_M(\delta_k)$ by elevating it to the k^{th} power.

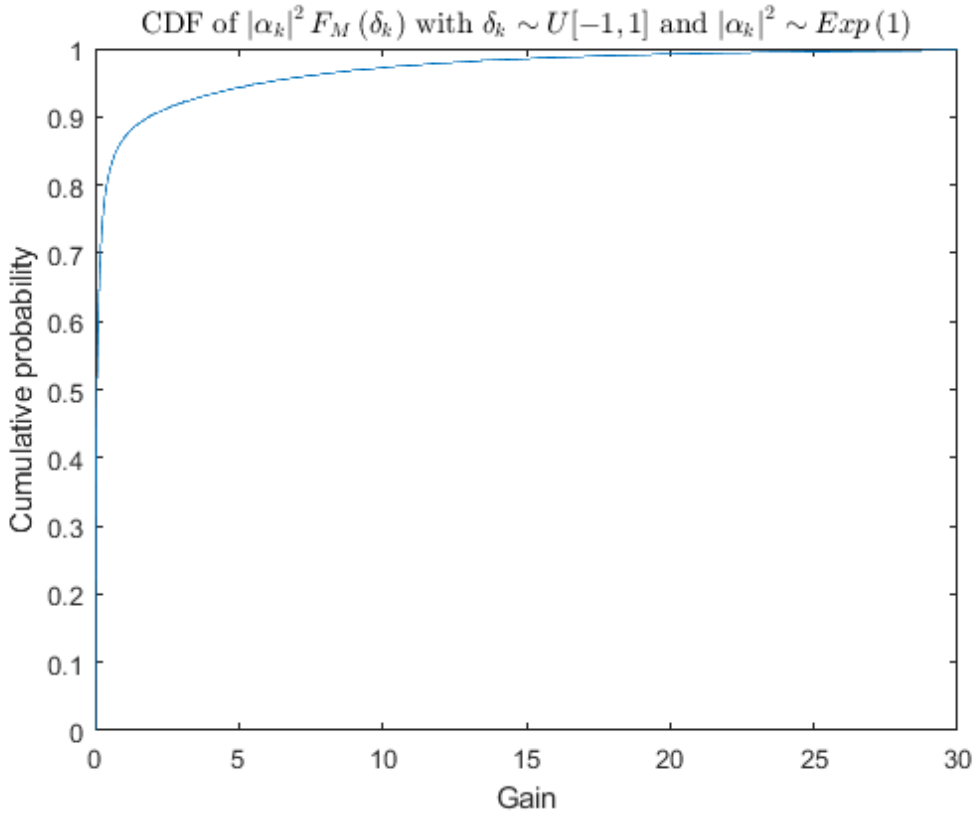


Figure 5.3: CDF of the product of the beamforming and Rayleigh fading gains for M=10

The expected value can be computed through a single integration from the CDF using the formula following formula which is valid for non-negative random variables. Fortunately the random variables of interest, our gains and rates are positive by definition.

$$E[X] = \int_0^\infty (1 - F_X(x))\, \mathrm{d}x \tag{5.4}$$

that can also be used to obtain the second order moment as:

$$\begin{aligned} E\left[X^2\right] &= \int_{x^2=0}^{\infty} Pr[X^2 \geq x^2] dx^2 \\ &= \int_{x^2=0}^{\infty} Pr[X \geq x] dx^2 \\ &= \int_{x^2=0}^{\infty} \left(1 - F_X(x)\right) dx^2 \\ &= 2\int_{x=0}^{\infty} x(1 - F_X(x)) dx \end{aligned} \tag{5.5}$$

where in the first step we used the fact that the square function is monotonously increasing for positive arguments and the last step is obtained by changing the variable of integration from x^2 to x and noting that $dx^2 = 2xdx$. The mean and second order moment lead to the variance using the basic decomposition

$$\mathrm{Var}(X) = E\left[X^2\right] - E\left[X\right]^2. \tag{5.6}$$

This new approach allows us to obtain the expected rates using only one integral for any K and M whereas a simulation would require to perform at least K products between the beam-vector and each channel vector of length M in order to select the maximum of them. Furthermore this whole simulation should be repeated a lot of times in order to estimate the average behaviour and variance, leading to large simulation times. Hence, when the number of users and antennas is very large but, maybe not enough to use asymptotical analysis results this semi-analytical method turns-out very helpful.

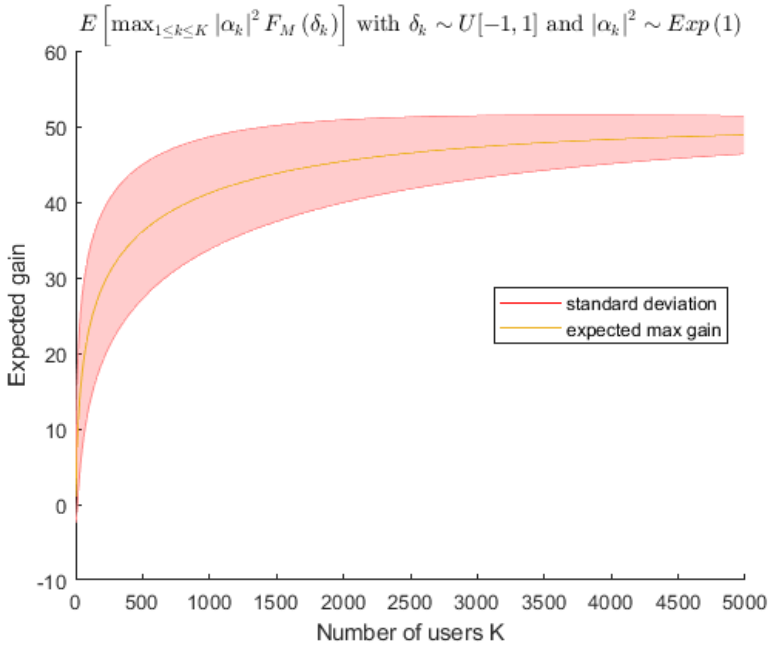


Figure 5.4: Mean and standard deviation of the maximal random gain among K users for M=10

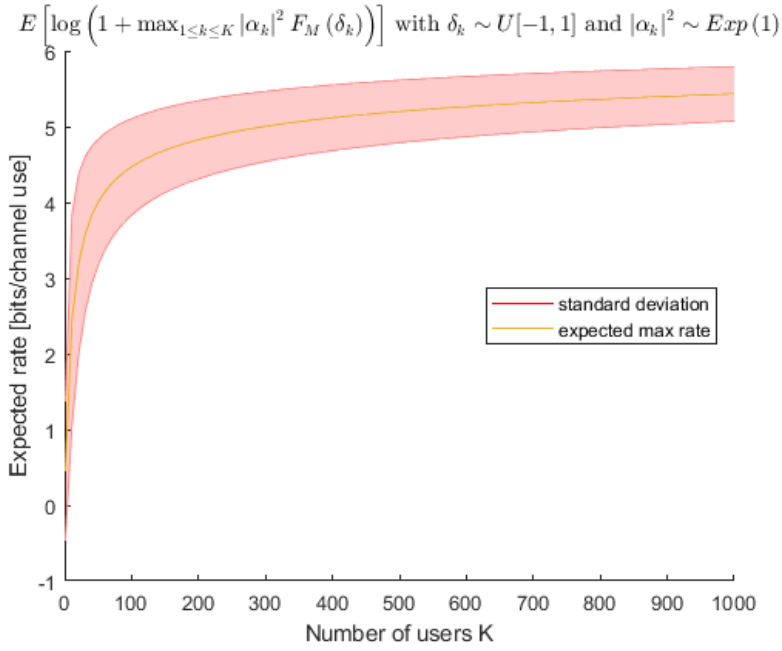


Figure 5.5: Mean and standard deviation of the rate achievable by the user with maximal random gain among K users for M=10

From Figure 5.4 and 5.5 we can note that the presence of the Féjer kernel doesn't change the global shape of the curve which still looks logarithmic for the gain and doubly logarithmic for the rate, with respect to the number of users. This implies that after an important initial rise the rate tends to grow very slowly with the number of users.
This method is adapted to quickly compute interest values for high numbers of users, but as the number of users becomes extremely high we may face problems due to numerical computation. It would then be more appropriate to use the fact that our product distribution tail has an exponential decay to apply the Fisher–Tippett–Gnedenko theorem of extreme value theory and approximate the distribution of the maximum as an appropriately fitted Gumbel distribution, as done by P. Viswanath and D. Tse to get their first results in their seminal work on opportunistic beamforming [49].

5.3 Simulation methods and considered scenarios

When the considered scenario was too complicated for direct computation we used simulations based on Monte Carlo methods. In Monte Carlo simulations the initial systems parameters are generated according to their respective distributions, then they are deterministically processed and results are averaged over the simulations. By the law of the unconscious statistician and the law of large numbers the sample averages should approximate the statistical expected values, with a variance decreasing with the number of averaged elements. The number of repetitions ranges from a minimum of one hundred for the more computationally demanding, to a thousand in the general case. All the simulations were performed in MATLAB R2018a.
The channel model used in the simulations of the first Section of Chapter 6 is the UR-SP described in Subsection 1.2.2. The considered numbers of users per cell are arbitrarily chosen and range from zero to a few thousands. Considering that mmWaves generally don't offer a long range coverage, these number of users may be reached only in really crowded environments. The random parameters generated at each run of the simulation are, for all `K` users, their complex channel gain generated using `alpha wgn(K,1,0,'complex')` (function coming from the Communication toolbox) and their channel AoD angles generated using `rand(K,1)*2-1`. The random beam angles are generated in the same way.

For the heterogeneous case the model including distance described at the end of Subsection 1.2.2 is used. This time the initial number of users `K_tot` is generated according to a Poisson distribution by using `random('Poisson',mu)` with `mu` being the chosen mean. Their polar coordinates are generated in order to have a uniform distribution over a disk of fixed radius `Disk_Radius`. This can be achieved by

choosing the angle from a uniform distribution as before and also the square of the distance as $d_k^2 \sim \mathcal{U}[0, d_{Disk-Radius}^2]$. Then, following the rationale of [62], we consider a total blockage probability for each user depending on its distance d_k from the BS as:

$$\mathrm{P(Blockage)} = 1 - e^{-\phi d_k} \tag{5.7}$$

where the coefficient ϕ represents the building density and is set to 0.1. Finally, users without LOS are excluded from transmission.

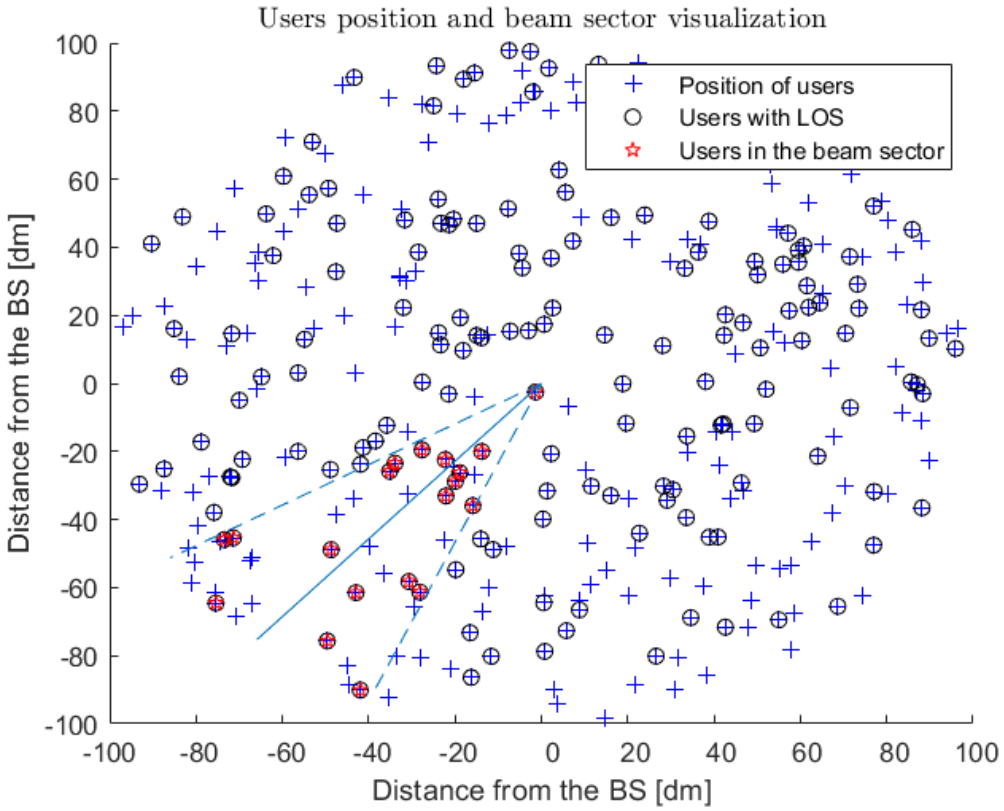


Figure 5.6: Scatter plot of users positions in the azimuthal plane with respect to a central BS. Users suffering blockage and those having a direct path (LOS) are distinguished and those in the sector corresponding to one of the beams are highlighted.

One example of the final obtained users distribution is shown in Figure 5.6. The figure also shows a sector, indeed for each beam only a subset of users is considered for transmission according to their angular distance with respect of the beam direction. This allows to speed-up simulations by optimizing each beam transmission policy only over a fraction of the total number of users. The sector

width is a set to be inversely proportional to the number of transmitting antennas, and thus directly proportional to the main lobe width.

Chapter 6

Simulations and results

In this section we will tackle the two issues introduced in Chapter 4. The first section will be focused on determining the necessary number of users for OBF to approach optimal performances while the second section will deal with the trade-off between fairness among users and sum-rate maximization.

6.1 Evolution of the OBF rates with the number of users

We have seen at the end of Chapter 3 that if we let the number of users in the system tend to infinity, OBF becomes sum-rate capacity achieving. In the first section of Chapter 4 we saw that the necessary number of users is asymptotically related to the number of transmitting antennas. More precisely, if the number of antennas M tends to infinity, we have that K must grow at least linearly with M in order to get a sum-rate linearly scaling with M (i.e. full multiplexing gain). This result is derived from the mmWaves channel UR-SP model, while in the classical Rayleigh channel model the number of users had to grow at exponentially with M in order to get the full multiplexing gain. Unfortunately these results are valid only for infinite numbers of antennas and users, we are thus going to see if they still hold with a large, but finite, number of users and antennas.

6.1.1 Single Beam OBF

We start from the simulation of the simplest case, the single beam OBF scheme, which is recapitulated in Alogrithm 1 box.

We know that as the number of users goes to infinity, fading is beneficial as it allows to obtain a multiuser diversity gain, but is it true also when the number of

Algorithm 1: Single Beam Opportunistic Beamforming (SB-OBF)

1 BS generates a beam direction $\vartheta \sim \text{Unif}\,[-1,1]$
2 BS transmits a known pilot in that direction: $\rho x_{pilot}\mathbf{a}\,(\vartheta)$
3 UEs compute and feed-back their signal-to-noise ratios:

$$\text{SNR}_k = \rho M |\alpha_k|^2 \left|\mathbf{a}\,(\theta_k)^\dagger\,\mathbf{a}\,(\vartheta)\right|^2$$

4 BS selects the user $i = \underset{1\le k\le K}{\text{argmax}}\,\text{SNR}_k$ for transmission
5 BS transmits $\rho x_i \mathbf{a}\,(\vartheta)$
6 Repeat from 1 after the transmission time is over

users is small? Let's compare the rate achieved with beamforming gain only and with both beamforming and Rayleigh fading.

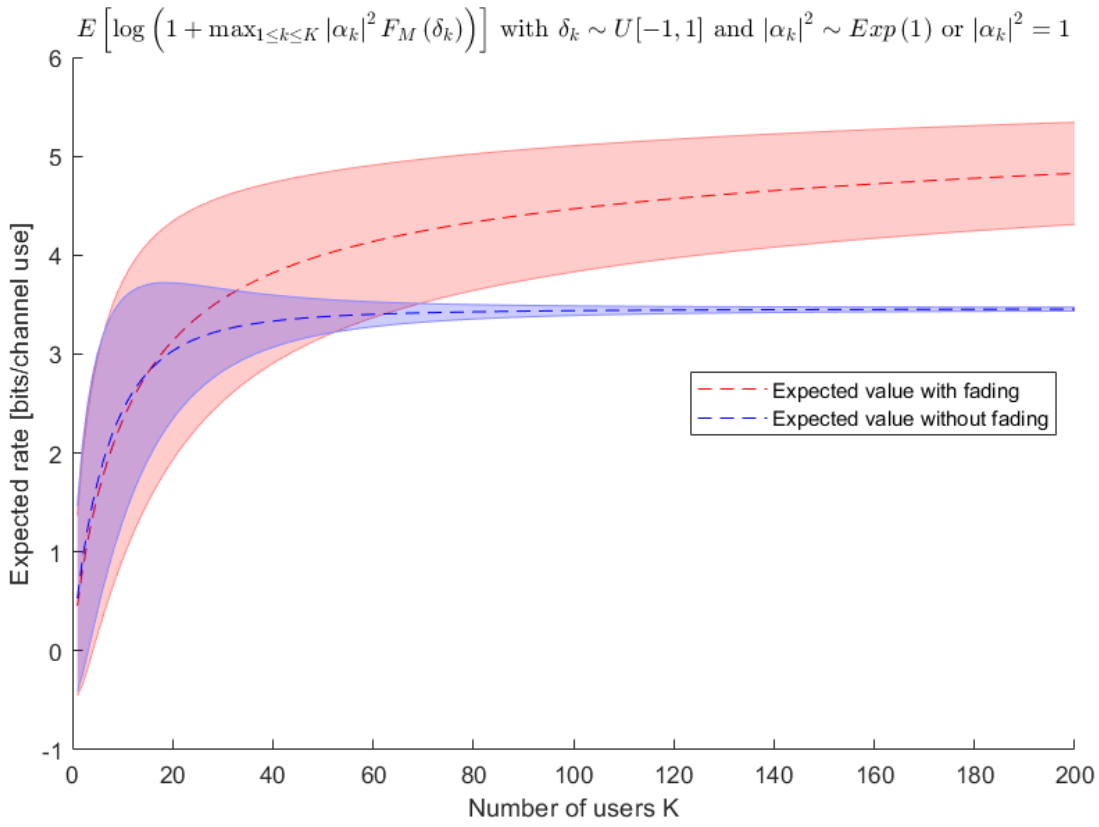


Figure 6.1: Comparison of the mean and standard deviation of the rate achievable by SB OBF with K users for M=10 in the fading and unitary channel gain cases

We can note that the achieved rate is almost the same in both cases for $K \leq 20$ while afterwards the rate with fading continues to increase with the number of users while the no-fading gain follows a horizontal asymptote of value $\log(1+M)$ corresponding to perfectly matched beamforming.

Now let's have a look at the achievable rate as a function of the number of users for different number of transmitting antennas. The results obtained through Monte Carlo simulation and from the numerical CDF extrapolation reported in the next figure coincide, thus confirming the validity of both approaches.

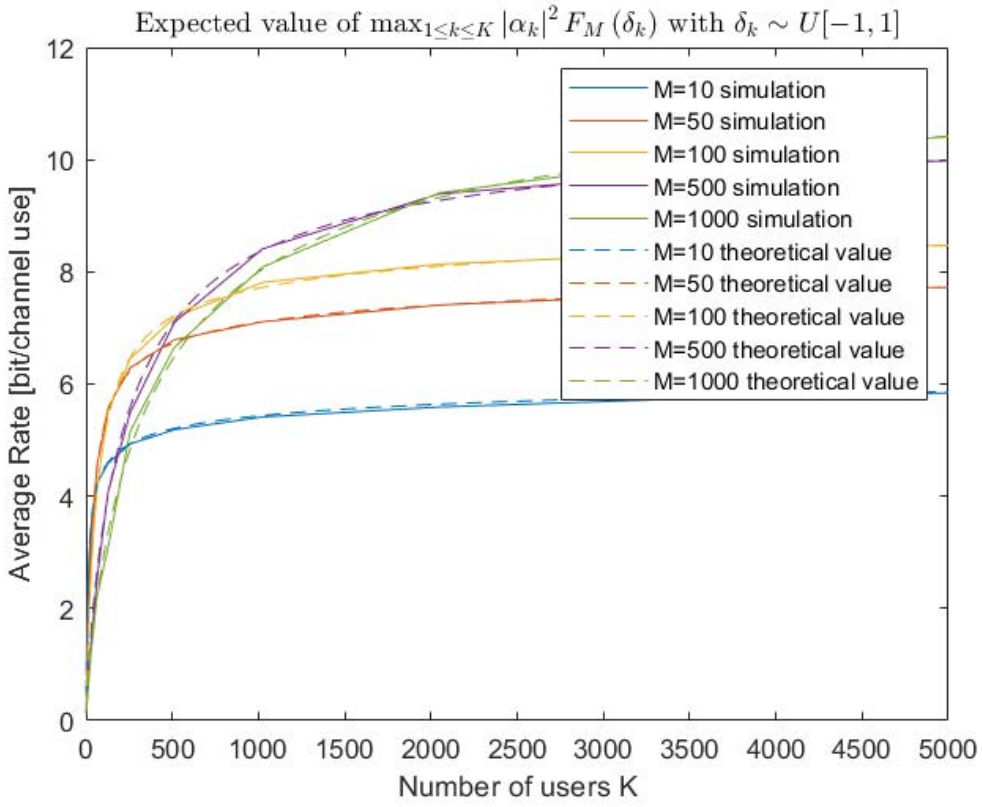


Figure 6.2: Single user OBF achievable rate as a function of the number of users for different values of M. Theoretical values obtained using the CDF.

From the plots we also notice that for small K we initially obtain the highest rate with small arrays but as K grows larger arrays become optimal. We are interested in the necessary number of users to approach the almost stagnating part of the curve. In this optic we consider the threshold of 0.95% of $\log(1+M\rho)$, with varying M, and see how many users are necessary in order to overcome the threshold.

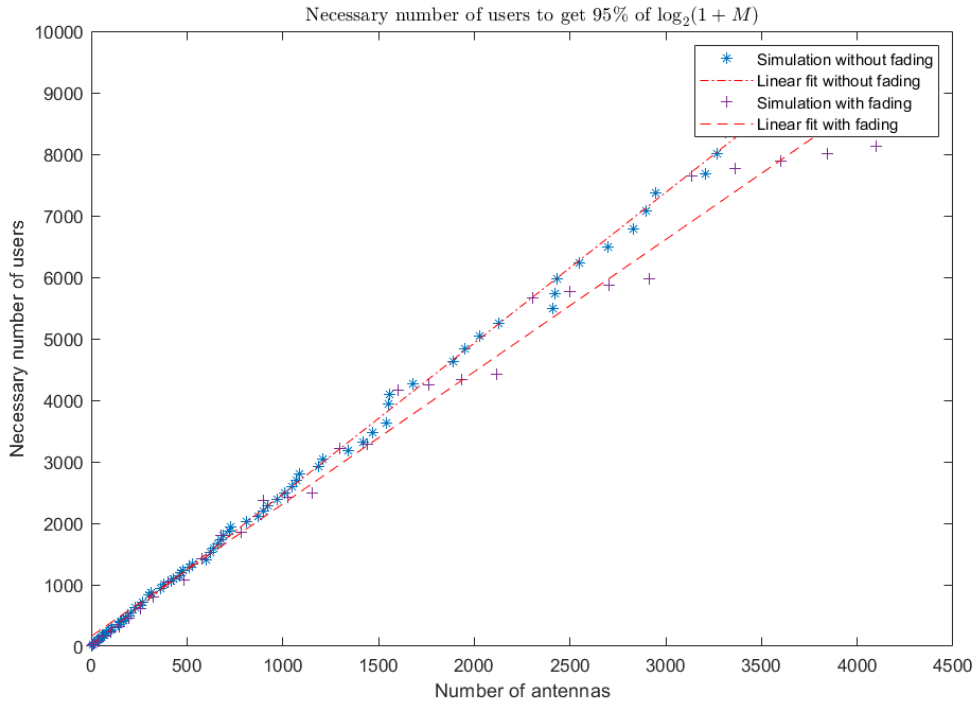


Figure 6.3: Necessary number of users in SB-OBF to have a 95% of the rate $\log(1+M)$ with and without fading

By looking at the figure a linear relation between the necessary number of users and the number of antennas is evident. We repeated the test with and without fading but there are no significant differences, the linear coefficients are both close to 2.5 but we have to keep in mind that this figure depends on the transmission SNR. If we now consider M equispaced beams with power equally splitted, and we neglect inter-beam interference we would obtain a total rate greater than $\sum_{i=1}^{M} 0.95 \log(1+\rho) = 0.95M \log(1+\rho)$ so we would approach a full multiplexing gain by increasing the number of users linearly w.r.t. the number of antennas. Obviously inter-beam interference cannot be neglected *a priori* so we will now study the multi-beam case through a dedicated simulation.

6.1.2 Multiple Beams OBF

In order to exploit spatial diversity and allow quasi-orthogonal simultaneous transmission to M users we now consider the OBF scheme with equispaced beams, which is summarized in the following box.

Again, we start by looking in Figure 6.4 at the behaviour of the rate as the number of users increases, for different numbers of transmitting antennas. This

Algorithm 2: Equispaced Beams Opportunistic Beamforming (EB-OBF)

1 BS generates $\vartheta_1 \sim \text{Unif}\,[-1, 1]$

2 BS computes the angles $\vartheta_b = \vartheta_1 + \frac{2(b-1)}{S}$ for $b = 1, ..., S$

3 BS transmits sequentially a known pilot using $S = M$ different beams:

$$x_{pilot}\mathbf{a}\left(\vartheta_b\right) \quad b = 1, ..., S$$

4 UEs compute and feed-back for all beams:

$$\text{SINR}_{k,b} = \frac{\left|\alpha_k\right|^2 \frac{\rho}{S} M \left|\mathbf{a}\left(\theta_k\right)^\dagger \mathbf{a}\left(\vartheta_b\right)\right|^2}{1 + \sum_{b' \neq b} \left|\alpha_k\right|^2 \frac{\rho}{S} M \left|\mathbf{a}\left(\theta_k\right)^\dagger \mathbf{a}\left(\vartheta_{b'}\right)\right|^2}$$

5 BS selects the user $i_b = \underset{1 \leq k \leq K}{\text{argmax}}\, \text{SINR}_{k,b}$ for transmission on each beam

6 BS transmits

$$\sum_{1 \leq i \leq S} \frac{\rho}{S} x_{i_b} \mathbf{a}\left(\vartheta_b\right)$$

7 Repeat from 1 after the transmission time is over

Symbol	Definition
K	number of users
M	number of antennas at the BS
S	number of beams
$\mathbf{a}\left(\cdot\right)$	steering vector function
ρ	SNR at the transmitter
α_k	fading coefficient of user k
θ_k	AoD towards user k

Table 6.1: Algorithm 1 (SB-OBF) and 2 (EB-OBF) parameters

time we can notice that, thanks to the parallel transmission to many users, the sum-rate is always higher with larger transmission arrays, independently from the number of users.

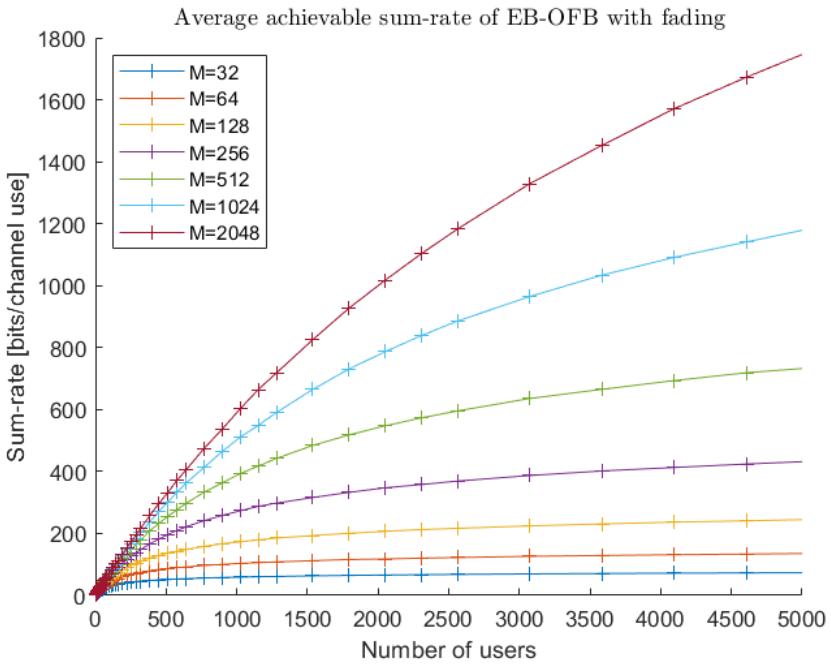


Figure 6.4: EB-OBF achievable sum-rate as a function K for $\rho = 1$

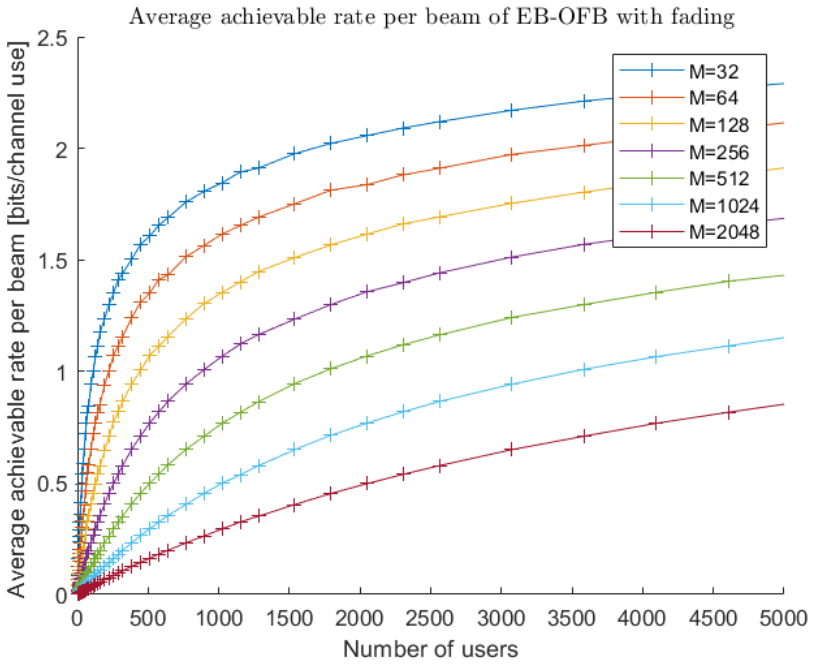


Figure 6.5: Achievable rate per beam as a function of K for $\rho = 1$

Then we consider in Figure 6.5 the average rate per beam, this time we find an opposite result: the more antennas and beams we use, the lower is the average rate per beam. Indeed, the larger beamforming gain is compensated by the power split and the beams become narrower as M increases so its harder to find users in the main lobe and the multiuser diversity can exploit only a smaller number of users. The rate per beam could be increased by using a number of beams $S < M$ but this would reduce the sum-rate, so from a system view it's a bad option.

We are now going to focus on the necessary number of users to approach a full multiplexing gain. As before we set the threshold at 95% of $M \log(1+\rho)$ which is the rate we would obtain without fading if we were able to find M perfectly placed users. In order to avoid confusion between the multiplexing gain and the multiuser-diversity gain due to fading we also perform the simulation without fading. The plot in Figure 6.6 displays an almost perfect linearity between the necessary number of users and the number of antennas, thus extending the asymptotical results of Chapter 4 to the finite number of users case.

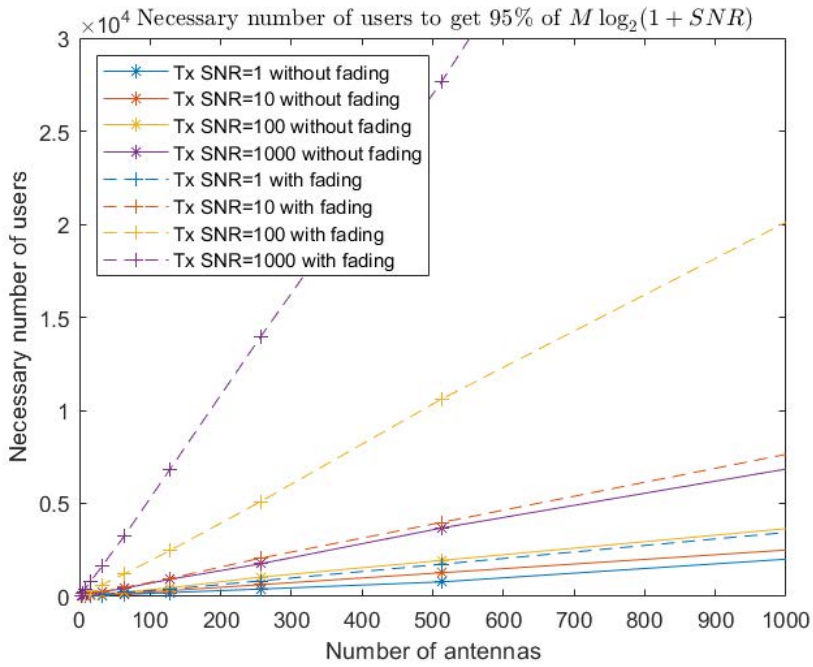


Figure 6.6: Necessary number of users to have a 95% of the rate $M \log(1+\rho)$ with fading and no fading for different transmission SNRs

As a first remark we note that the higher is the transmitted power, the higher is the linear coefficient giving the necessary number of users. We must pay attention to this fact as the multiplexing gain is generally defined as a limit at high SNR, we

should have convergence to a fixed linear coefficient as the transmit SNR increases but this doesn't happen. What we can say is that at fixed transmitting power, a linear increase of the number of users with the number of antennas is sufficient to approach the perfect CSIT sum-rate of beamforming without fading $M\log(1+\rho)$. Finally, by comparing the doted lines versus the full lines of the same colour we see that the presence of fading only increases the slope of the line, which makes sense as it is equivalent to a power gain.

To be fair, the sum-rate obtained by considering $|\alpha_k|^2 \sim Exp(1)$, as it takes benefit from multiuser diversity, should be compared to

$$M \times E\left[\log\left(1+\max_{1\leq k\leq K}|\alpha_k|^2\right)\right]$$

which is the high number of users regime upper bound to sum-rate for our channel, as stated in Theorem 2.2.2.

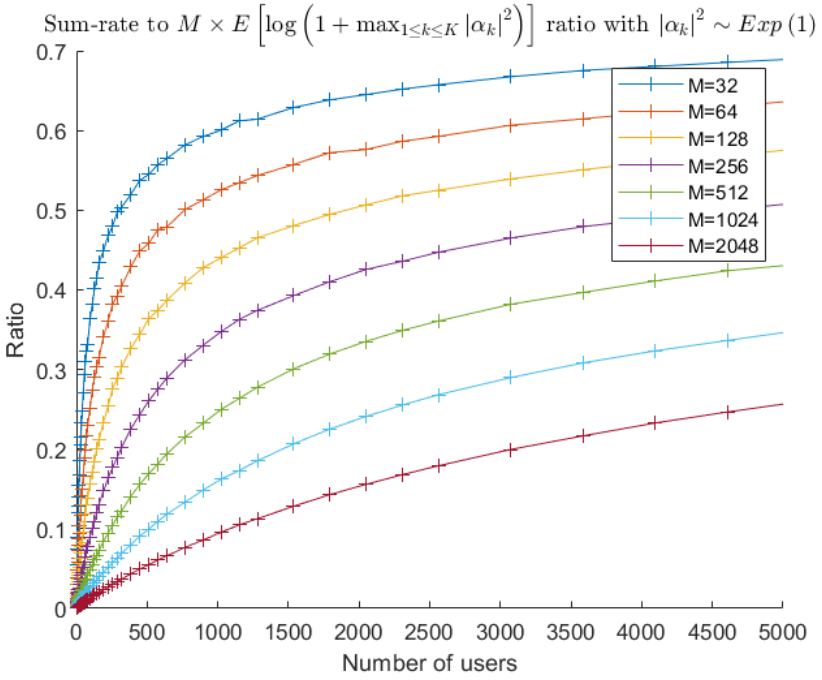


Figure 6.7: Ratio of EB-OPBF sum-rate w.r.t. the full multiplexing and multi-user diversity gains (considering all users) sum-rate, with $\rho = 1$.

The ratio between the two is plotted in Figure 6.7 and unfortunately we see that the performances are pretty poor for large arrays which seems to necessitate much more users to achieve their full potential. Nonetheless, if we think about it, this result was expectable, this sum-rate upper bound could be reached only if all

beams were transmitting to the single user with the best fading gain in the whole system, while being orthogonal to each-other at the same time in order to create interference-free parallel streams. This is obviously impossible and we should rather consider that each beam can only select the best user among those in the sector spanned by its main lobe, so the maximum should be taken over a fraction of K inversely proportional to M. In this way we obtain Figure 6.8, which shows a really quick convergence for every M to around $0.9M \times E\left[\log\left(1+\max_{1\leq k\leq \frac{K}{2M}} |\alpha_k|^2\right)\right]$.

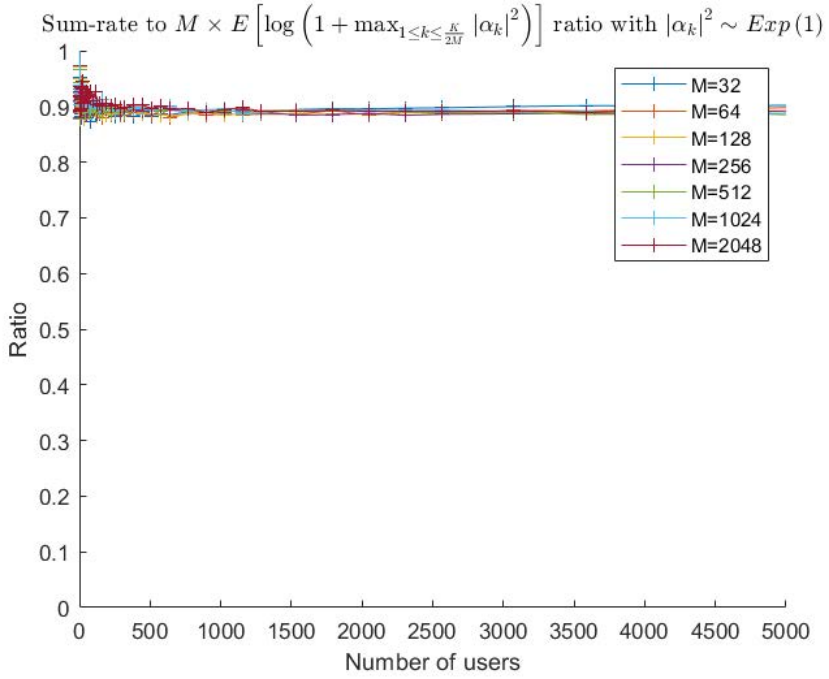


Figure 6.8: Ratio of EB-OPBF sum-rate w.r.t. the full multiplexing and multi-user diversity gain (considering each beam's sector users only) sum-rate, with $\rho = 1$.

6.2 Fairness in non-homogeneous networks

Till now we have only dealt with homogeneous networks, assuming that the channels of the different users had identical statistics. Indeed, for a given beam users can be more or less aligned with its direction, but considering that the beam directions are uniformly distributed over all angles and updated at each transmission, we should obtain in the long term a perfect equality among users' throughputs. And this still holds despite the opportunistic policy of transmitting with each beam to the user with the highest instantaneous SINR with respect to that beam. In this section instead we will deal with heterogeneous networks, that is a network

where the fading statistics of the users are different.

To simulate an heterogeneous network we will use the model including transmission path distance presented in Subsection 1.2.2 of the first chapter and whose channel vector model is presented in Equation 1.15 and reported hereinafter:

$$\mathbf{h}_{\mathrm{k}}[\mathrm{m}] = \sqrt{M}\frac{\alpha_k[\mathrm{m}]\mathbf{a}\left(\theta_k\right)}{\sqrt{1+d_k^{\beta_{LOS}}}}$$

where d_k is the distance of the UE from the BS and β_{LOS} is the loss exponent for LOS paths, as detailed in the methodology chapter, transmissions to users without LOS are assumed to be blocked in this scenario. Given the fixed different distances from the base-station, some users will have a channel which have an average higher path loss than others. In this situation applying the opportunistic policy of the previous section (like in Alg. 1 and Alg. 2) would result in some users being almost never scheduled. In a fairness and minimum guaranteed quality of service perspective, a different transmission strategy should be adopted. In Section 4.2 proportional-fair scheduling was introduced, and will be used as a reference to evaluate the fairness of some alternative policies. Then, we will focus on the fairness gain obtained by applying NOMA instead of OMA.

6.2.1 Single user per beam scheduling

In all the cases of this subsection, the BS will still generate multiple equispaced beams like in EB-OBF but for each beam only users in the corresponding sector will be considered and will need to feed-back their SINR. Then different choices of what to transmit with each beam will be considered. Given that two beams cannot transmit to the same user in our sectorized scenario, inter-beam interference isn't influenced by which user is scheduled on each beam. This implies that the rates and throughputs of users scheduled on different beams are independent, so the proportional-fair scheduling policy can be applied beam by beam and still reach the global optimum. The details of this protocol are reported in the Alg. 3 box. This protocol has been implemented along with four other intuitive scheduling options. Each of these policies will substitute $i_s^{PF}[t]$ of Alg. 3, with a different decision criteria. They are meant to embody the main guiding principles one could adopt while trying to find a policy which maximize the fairness.

- Policy 1: The opportunistic policy of the previous section where for sector s we select:

$$i_s^1[t] = \underset{1\leq k_s\leq K_s}{\mathrm{argmax}} R_{k_s}[t] \tag{6.1}$$

- Policy 2: The policy consisting selecting the user having the current lowest throughput:

$$i_s^2[t] = \underset{1 \leq k_s \leq K_s}{\operatorname{argmin}} T_{k_s}[t] \tag{6.2}$$

- Policy 3: The policy consisting in randomly selecting with equal probabilities each user in the sector for transmission. We will call this this policy the "random user" policy.

$$i_s^3[t] \sim \mathrm{Unif}_{[1:K_s]} \tag{6.3}$$

- Policy 4: An intermediary policy, which, gives to all users an equal probability to be selected over time, but select each user when its fading fast-varying component is higher than the other users' one. Indeed, in this section we consider that the users in the sector feedback before each scheduling update their instantaneous effective channel gains for each beam:

$$\mathrm{Gain}_{k_s,s}[t] = \frac{|\alpha_{k_s}[t]|^2 \, \mathrm{F}_M \left(\vartheta_s[t] - \theta_{k_s}\right)}{1 + d_{k_s}^{\beta_{LOS}}} \tag{6.4}$$

  while the distance d_{k_s}, like the AoD θ_{k_s}, is considered nearly constants over a long period of time. So, we suppose that the distance is known at the BS which can then compute:

$$\begin{aligned} i_s^4[t] &= \underset{1 \leq k_s \leq K_s}{\operatorname{argmax}} \left[\frac{\mathrm{Gain}_{k_s,s}[t]}{1 + d_{k_s}^{\beta_{LOS}}} \right] \\ &= \underset{1 \leq k_s \leq K_s}{\operatorname{argmax}} \left[|\alpha_{k_s}[t]|^2 \, \mathrm{F}_M \left(\vartheta_s[t] - \theta_{k_s}\right) \right] \end{aligned} \tag{6.5}$$

These four policies, along with the SPF are compared in Figure 6.9 in their steady state behaviour.

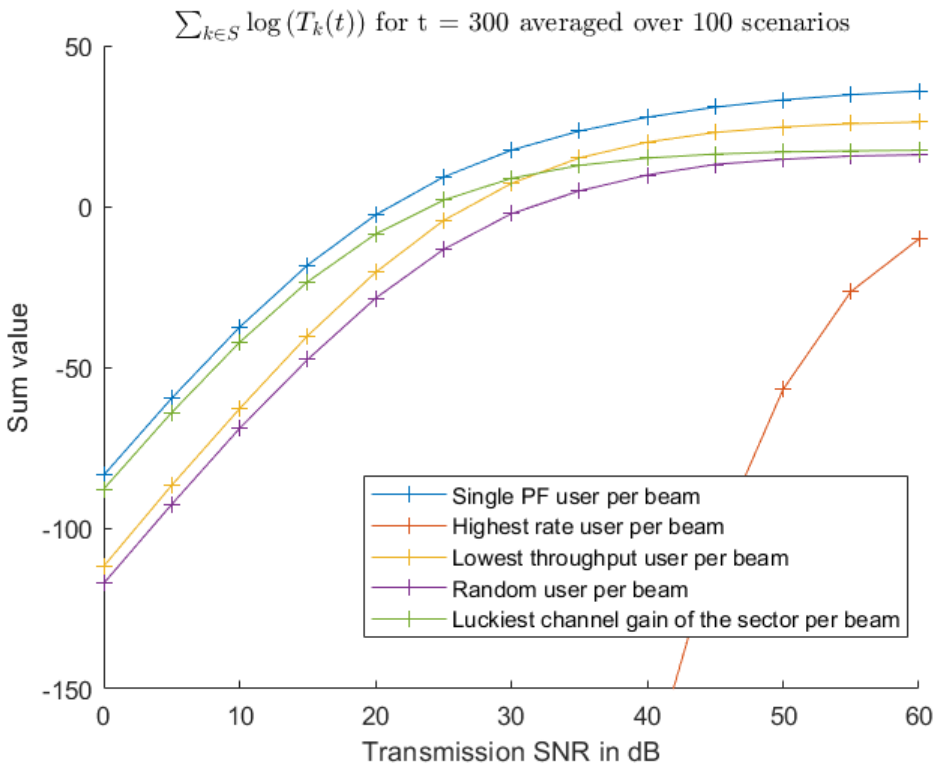


Figure 6.9: Proportional-fairness objective function value for different scheduling policies and transmitted power

We can see that with the opportunistic policy fairness is much lower than for all the others policies. It approaches the other ones only at very high transmission SNR. This can be explained by considering that at high transmission SNR the rate is mainly determined by inter-beam interference, which depends from the alignment of the users with the transmitted beams. However, being the beam directions uniformly distributed, no user has a better angle than the others in the long term, thus the network becomes almost homogeneous.

The second policy, which selects the user having the lowest throughput, performs well in terms of fairness especially at high SNR, and it is always over the random user selection policy.

Concerning the policy 4 we can see that on the long term it reaches is close to optimal proportional fairness for low transmission SNRs. This can be explained in the following way. At low SNR ($\rho << 1$) the inter-beam interference is neglectable and so we can write:

$$R_{i_s[t]} \approx \log\left(1 + \frac{\rho}{S}\,\mathrm{Gain}_{k_n,s}[t]\right) \tag{6.6}$$

$$\approx \frac{\rho}{S}\,\mathrm{Gain}_{k_n,s}[t] \tag{6.7}$$

And, for $t_c >> 1$ by the law of large numbers we have on the long term:

$$\frac{R_{i_s[t]}}{T_{k_*}[t]} \approx \frac{R_{i_s[t]}}{\mathrm{E}\left[R_{i_s[t]}\right]}$$

$$\approx \frac{\rho}{S}\,\mathrm{Gain}_{k_n,s}[t]\frac{1}{\mathrm{E}\left[\frac{\rho}{S}\,\mathrm{Gain}_{k_n,s}[t]\right]}$$

Furthermore as the Féjer kernel and the fast fading coefficient $|\alpha_{k_s}[t]|^2$ are assumed independent and having both mean 1, the expectation of their product has still mean 1. Then, the distance d_{k_s} is not considered a random variable once the scenario has been generated so we have :

$$\mathrm{E}\left[\mathrm{Gain}_{k_n,s}[t]\right] = \mathrm{E}\left[\frac{|\alpha_{k_s}[t]|^2\,\mathrm{F}_M\left(\vartheta_s[t] - \theta_{k_s}\right)}{1 + d_{k_s}^{\beta_{LOS}}}\right] = \frac{1}{1 + d_{k_s}^{\beta_{LOS}}}$$

which in turns leads to:

$$\frac{R_{i_s[t]}}{T_{k_*}[t]} \approx \frac{|\alpha_{k_s}[t]|^2\,\mathrm{F}_M\left(\vartheta_s[t] - \theta_{k_s}\right)}{1 + d_{k_s}^{\beta_{LOS}}}(1 + d_{k_s}^{\beta_{LOS}})$$

$$= |\alpha_{k_s}[t]|^2\,\mathrm{F}_M\left(\vartheta_s[t] - \theta_{k_s}\right)$$

This last expression is precisely the one maximized in policy 4, so at low SNR we are nearly optimizing the proportional fairness objective function. Unfortunately at high SNR its fairness tends to the one of policy 3 where a random user is selected. The latter actually doesn't perform so badly in terms of fairness, we can note an almost constant and not very consistent offset with the optimal one (SPF). On the other hand, if we look at the sum-rates, shown in Figure 6.10, we notice that it's performances are very poor, just above policy 2which is the the worse one in term of sum-rate as users with poor channels keep being scheduled despite their low rate. As expected the higher curve corresponds to the opportunistic policy. One more time the SPF and policy 4 almost coincide for low SNRs and then diverge with SPF keeping close to the opportunistic strategy sum-rate.

This brief comparison allowed us to gain some insights into the proportional fair policy. This policy allows to increase fairness but is also exploits multiuser diversity very well in both low SNR and high SNR. We are now going to see

what happens if we relax the condition of transmitting only to one user with each beam. this will also be the occasion to comapare OMA and NOMA on the fairness achievement issue.

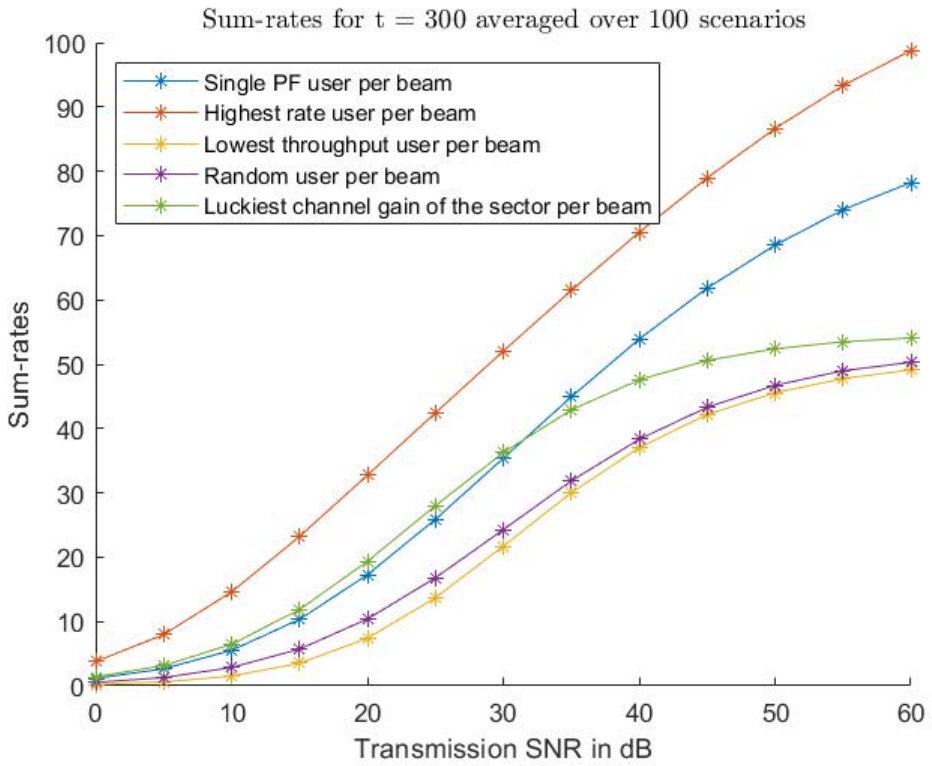


Figure 6.10: Proportional-fairness objective function value for different scheduling policies and transmitted power

6.2.2 NOMA versus OMA for proportional fairness

To increase fairness we could think to allocate simultaneously many users on a single beam, for each beam. To do so we can either choose an Orthogogal Multiple Access approach like TDMA and FDMA or a Non-Orthogonal Multiple Access Approach like superposition coding followed by successive interference cancellation. The aim of this subsection is to compare the two approaches in the two-users per beam case.

Let's start by defining the NOMA approach. For each beam b we superpose in the power domain the signals intended for two distinct users i_b^* and j_b^*:

$$\mathbf{p}_b = \frac{\rho}{S}\mathbf{a}\left(\vartheta_b\right)\left(\beta_b x_{i_b^*} + (1-\beta_b)x_{j_b^*}\right) \tag{6.8}$$

where i_b, j_b and β_b are selected in order to maximize the objective function

$$J(i_b, j_b, \beta_b) = \sum_{1 \leq k_b \leq K_b} \log\left(T_{k_b}[t+1]\right) . \tag{6.9}$$

As here the rates of the scheduled users are not independent, the solution is given by Eq. 4.17 which applied to our sector becomes:

$$I^* = \underset{I=\{i_b, j_b\}}{\arg\max} \left[\max_{\beta_b} \left(1 + \frac{R_{i_b}[t]}{(t_c - 1)\, T_{i_b}[t]}\right) \left(1 + \frac{R_{j_b}[t]}{(t_c - 1)\, T_{j_b}[t]}\right) \right] \tag{6.10}$$

Let's consider that j_b is the user with the best channel, that is the highest inter-beam SINR for beam b. The fact that it's channel is better than the one of user i_b, implies that, if user i_b is able to decode some data transmitted through beam b, then also user j_b will be able to decode that message. Thus, user j_b can decode the message intended to user i_b, re-modulate it in the same way as the base station did, simulate the effect of the channel on that signal and subtract the resulting signal from the received one. In this way, the interference due to the signal intended to i_b is cancelled and user j_b is now only experiencing interference from other beams, so its SINR is:

$$\mathrm{SINR}_{j_b} = \frac{(1 - \beta_b^2)\, |\alpha_{j_b}|^2\, \frac{\rho}{S}\, \mathrm{F}_M\left(\vartheta_b - \theta_{j_b}\right)}{1 + d_{j_b}^{\beta_{LOS}} + \sum_{b' \neq b} |\alpha_{j_b}|^2\, \frac{\rho}{S}\, \mathrm{F}_M\left(\vartheta_{b'} - \theta_{j_b}\right)} \tag{6.11}$$

and it can support a rate up to $R_{j_b} = \log\left(1 + \mathrm{SINR}_{j_b}\right)$.

On the other hand, user i_b will suffer both from inter-beam interference and from the interference due to the signal intended to user j_b. It's global SINR is thus given by:

$$\mathrm{SINR}_{\mathrm{i_b}} = \frac{\beta_b^2\, |\alpha_{i_b}|^2\, \mathrm{F}_M\left(\vartheta_b - \theta_{i_b}\right)}{\frac{s}{\rho}(1 + d_{i_b}^{\beta_{LOS}}) + |\alpha_{i_b}|^2 \left((1 - \beta_b^2)\, \mathrm{F}_M\left(\vartheta_b - \theta_{i_b}\right) + \sum\limits_{b' \neq b} \mathrm{F}_M\left(\vartheta_{b'} - \theta_{i_b}\right) \right)} \tag{6.12}$$

which allows a transmission rate up to $R_{i_b} = log\left(1 + SINR_{i_b}\right)$.

By considering superposition coding applied on individual beams, the sum-rate would be achieved by transmitting only to the user k_b with the higher $\mathrm{SINR}_{\mathrm{k_b,b}}$, so with no need for superposition and SIC. Nonetheless the NOMA approach allows to obtain a capacity region which is larger than the one achieved by OMA. In other words, in a given sector some sets of rates $\{R_{1_b}, ..., R_{K_b}\}$, can be achieved only through NOMA. As we are trying to increase fairness, we can expect that PF scheduling will not aim to always select one single user, but rather to share

the transmission in some cases. In that case superposition coding may give better results than OMA which finally result in an increase of our objective function.

Now, in order to perform a fair comparison, we have to extend the Single Proportional Fair user per beam (Alg. 3) protocol to the possibility of simultaneous transmission to two users. Let's suppose that FDMA is chosen, and performed over users i_b and j_b. Let's say that a bandwidth portion $0 \leq \xi_b \leq 1$ and a share of the power $0 \leq \eta_b \leq 1$ are assigned to user i_b. The formulas giving the rates are:

$$R_{i_b} = \xi_b log\left(1 + \frac{\eta_b}{\xi_b} SINR_{i_b}\right) \tag{6.13}$$

$$R_{j_b} = (1 - \xi_b) log\left(1 + \frac{1 - \eta_b}{1 - \xi_b} SINR_{j_b}\right) \tag{6.14}$$

Then we proceed like for the NOMA case by finding:

$$I^* = \underset{I=\{i_b, j_b\}}{\arg\max} \left[\max_{\eta_b} \left(1 + \frac{R_{i_b}[t]}{(t_c - 1)\, T_{i_b}[t]}\right)\left(1 + \frac{R_{j_b}[t]}{(t_c - 1)\, T_{j_b}[t]}\right)\right] \tag{6.15}$$

The whole transmission protocol is actually the same than in Alg. 4, we just have to replace the three equations of line 8 with the ones hereinabove.

Let's now see how the two algorithms perform in our heterogeneous scenario. In Figure 6.11 the steady state sums of the throughputs logarithms for both of them are plotted. From it, we can clearly see the superiority of NOMA over OMA, but we have to notice that it's only an increase of about 2%. It has been shown that NOMA particularly outperforms OMA in near-far situations. However our scenario, in order to be realistic, consider a probability of LOS decreasing exponentially with the distance, so there is a limited number of users at large distance. Anyway, from Figure 6.12 this advantage is also present in terms of sum-rate, even if it was not the objective we set.

As a final word on the fairness issue, we have been able to see that thanks to the randomness of the beams angles and to the possibility to actually schedule users on each beam by also taking care of fairness, random beamforming schemes can conserve almost all their capability in term of sum-rate while guaranteeing decent performances to the weaker users.

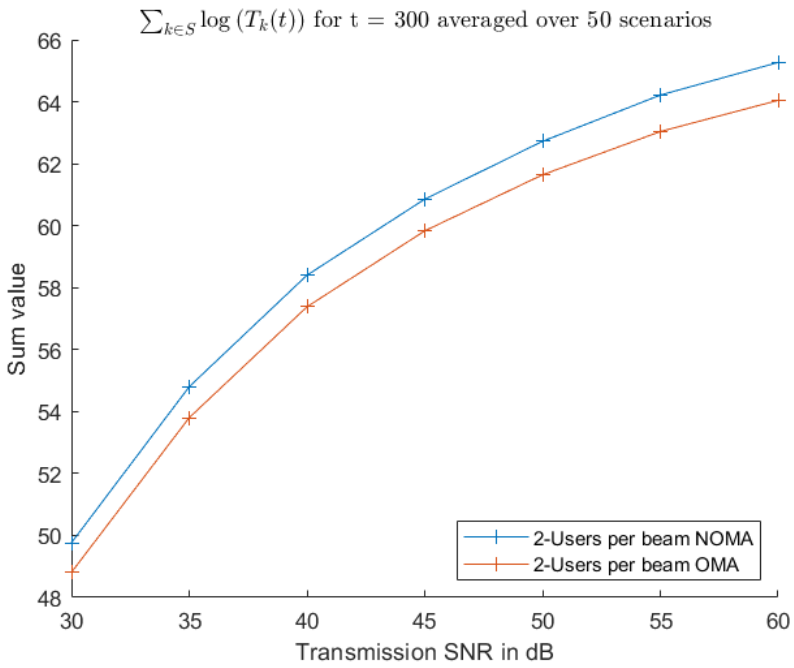


Figure 6.11: Proportional-fairness of our NOMA and OMA policies

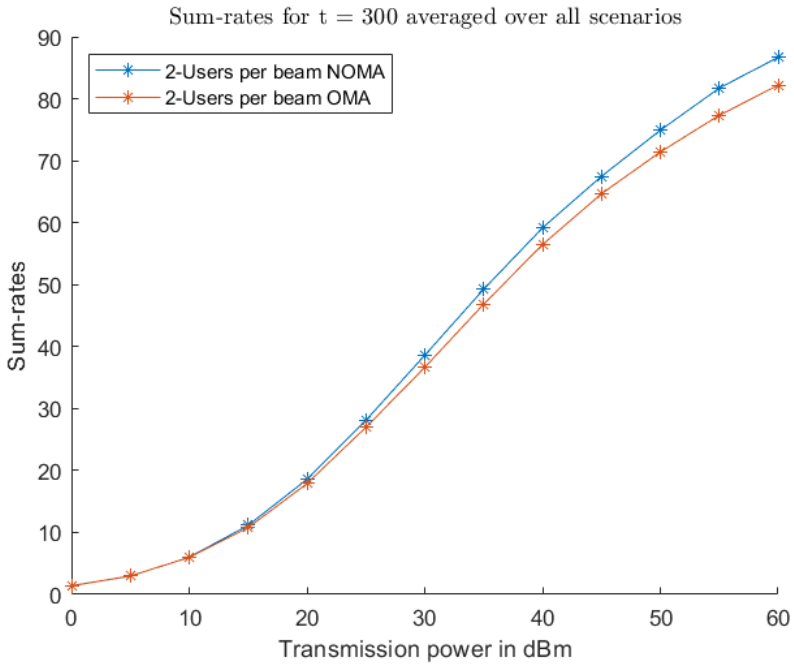


Figure 6.12: Sum-rates of our NOMA and OMA policies

Algorithm 3: Single Proportional Fair user per beam (SPF)

BS generates $\vartheta_1[t] \sim \text{Unif}\,[-1,1]$

BS computes the angles $\vartheta_b[t] = \vartheta_1[t] + \frac{2(b-1)}{S}$ for $b = 1, ..., S$

BS transmits sequentially a known pilot using $S = M$ different beams:

$$x_{pilot}\mathbf{a}\left(\vartheta_b[t]\right) \quad b = 1, ..., S$$

$\forall 1 \leq k \leq K$ UE k is assigned to sector $s = 1, ..., S$ if $|\vartheta_s - \theta_k| \leq \Delta$, the number of UEs in sector s is noted K_s and users are re-indexed in their sector as $1 \leq k_s \leq K_s$.

For $s = 1, ..., S$:

{

UEs $1 \leq k_s \leq K_s$ estimate from the pilots and feed-back their effective channel gains for each beam s:

$$\text{Gain}_{k_s,s}[t] = \frac{|\alpha_{k_s}[t]|^2 \,\text{F}_M\left(\vartheta_s[t] - \theta_{k_s}\right)}{1 + d_{k_s}^{\beta_{LOS}}}$$

The BS computes from the received channel gains the final SINR for scheduling in sector s of the users $1 \leq k_s \leq K_s$:

$$\text{SINR}_{k_s,s}[t] = \frac{|\alpha_{k_s}[t]|^2 \frac{\rho}{S} M \left|\mathbf{a}\left(\theta_{k_s}\right)^\dagger \mathbf{a}\left(\vartheta_s[t]\right)\right|^2}{1 + d_{k_s}^{\beta_{LOS}} + \sum_{s' \neq s} |\alpha_{k_s}[t]|^2 \frac{\rho}{S} M \left|\mathbf{a}\left(\theta_{k_s}\right)^\dagger \mathbf{a}\left(\vartheta_{s'}[t]\right)\right|^2}$$

The BS computes the potential rate achieved by scheduling user k_s on beam s as:

$$R_{k_s}[t] = \log\left(1 + \text{SINR}_{k_s,s}[t]\right)$$

The BS schedules for transmission on beam s the user i_s^{PF} such that:

$$i_s^{PF}[t] = \underset{1 \leq k_s \leq K_s}{\text{argmax}} \frac{R_{k_s}[t]}{T_{k_s}[t]}$$

The throughput of user i_s is updated for next transmission:

$$T_{k_s}[t+1] = \left(1 - \frac{1}{t_c}\right) T_{k_s}[t] + \frac{1}{t_c} R_{k_S}[t]$$

}

BS transmits:

$$\mathbf{x}[t] = \sum_{1 \leq s \leq S} \frac{\rho}{S} x_{i_s[t]}[t] \mathbf{a}\left(\vartheta_s[t]\right)$$

Repeat from 1 after the transmission time is over

Algorithm 4: 2-Users per beam NOMA (2U-NOMA)

BS generates $\vartheta_1[t] \sim \text{Unif}\,[-1, 1]$

BS computes the angles $\vartheta_b[t] = \vartheta_1[t] + \frac{2(b-1)}{S}$ for $b = 1, ..., S$

BS transmits sequentially a known pilot using $S = M$ different beams:

$$x_{pilot}\mathbf{a}\left(\vartheta_b[t]\right) \quad b = 1, ..., S$$

$\forall 1 \leq k \leq K$ UE k is assigned to sector $s = 1, ..., S$ if $|\vartheta_s - \theta_k| \leq \Delta$, the number of UEs in sector s is noted K_s and users are re-indexed in their sector as $1 \leq k_s \leq K_s$.

For $s = 1, ..., S$:

{

UEs $1 \leq k_s \leq K_s$ estimate from the pilots and feed-back their effective channel gain for each beam s:

$$\text{Gain}_{k_s,s}[t] = \frac{|\alpha_{k_s}[t]|^2 \, \text{F}_M\left(\vartheta_s[t] - \theta_{k_s}\right)}{1 + d_{k_s}^{\beta_{LOS}}}$$

Using the effective channel gains and Eq. 6.11, Eq. 6.12 the BS is able to find through optimization:

$$I^* = \underset{I=\{i_s,j_s\}}{\arg\max} \left[\max_{\beta_s} \left(1 + \frac{R_{i_s}[t]}{(t_c - 1)\, T_{i_s}[t]}\right)\left(1 + \frac{R_{j_s}[t]}{(t_c - 1)\, T_{j_s}[t]}\right)\right]$$

and store the corresponding proportional-fair rates $R_{i_s^*}[t]$ and $R_{j_s^*}[t]$.

The throughputs of user i_s^* and j_s^* are updated for next transmission:

$$T_{i_s^*}(t+1) = \left(1 - \frac{1}{t_c}\right) T_{i_s^*}[t] + \frac{1}{t_c} R_{i_s^*}[t]$$

$$T_{j_s^*}(t+1) = \left(1 - \frac{1}{t_c}\right) T_{j_s^*}[t] + \frac{1}{t_c} R_{j_s^*}[t]$$

}

The BS transmits:

$$\mathbf{x}[t] = \sum_{1 \leq s \leq S} \frac{\rho}{S} \mathbf{a}\left(\vartheta_b\right)\left(\beta_b x_{i_s^*} + (1 - \beta_b) x_{j_s^*}\right)$$

Repeat from 1 after the transmission time is over

Symbol	Definition
K	number of users
M	number of antennas at the BS
S	number of beams
$\mathbf{a}(\cdot)$	steering vector function
ρ	SNR at the transmitter
α_k	fading coefficient of user k
d_k	path length to user k
θ_k	AoD towards user k
t_c	PF throughput time-window parameter
Δ	sector half-angle
$\mathrm{F}_M(\cdot)$	Féjer kernel function of order M
β_{LOS}	loss exponent of the line-of-sight components

Table 6.2: Algorithm 3 (SPF) and 4 (2U-NOMA) symbols and parameters

Conclusion and future investigations

In order to get a general view on the thesis, let's now sum-up the results of each chapter. We started this thesis by deriving in Chapter 1 the main features that a downlink transmission scheme should adopt to be adapted to the mmWave channel by studying its characteristics. Its high attenuation needs to be compensated by a high power gain and its high directivity and sparsity in the angular domain suggest to employ directional beamforming. Furthermore the coherence time of the channel is reduced with respect to lower frequencies so users have to transmit more often their feedback, which should thus be limited in some way to reduce the overhead, leading to random directional beamforming approaches.

Then, in order to understand what are the main concepts to be exploited in multiuser downlink communications we have taken in Chapter 2 a bird's eye view on our problem and analysed the capacity of the MIMO Broadcast Channel. The final aim was to obtain the sum-rate capacity of the MIMO Broadcast channel, but on the way the concepts of capacity region and successive interference were presented, as they would have come to our aid in the final developments on fairness. The sum-rate capacity was then analysed in asymptotical terms (scaling w.r.t. the number of antennas and users), so that the concepts of spatial multiplexing and multiuser diversity were highlighted and could be thoroughly explained.

After that, in Chapter 3, we turned our attention to linear precoding, the family of precoding strategies to which directional beamforming belongs. Optimal and almost optimal solutions (i.e. MMSE, ZF and MRT beamforming) were presented but limitations in the transmitter architecture complexity and necessary feedback made it clear that this optimal solutions may not be implementable in mmWaves, nevertheless they provided us some insights in what characteristics should have the beams to obtain high performances: alignment with the intended user's channel and orthogonality to the other users' channels. The asymptotical performances of opportunistic random beamforming in high user regime were derived and they were shown to reach the sum-rate capacity of the MISO channel for infinite users. The question of how many users were necessary to approach these performances naturally arose.

A first answer to this question was given in the first section of Chapter 4 where

asymptotical results linking the necessary number of users to have full multiplexing gain were exposed. The main result being that while in Rayleigh fading channels the number of necessary users scales exponentially with the number of antennas, in the mmWave case only linear scaling is necessary thanks to the directivity of the channel which was modelled by the UR-SP model. The second section of the chapter was dedicated to introduce the main concepts and tools related to another issue: the maintainment of fairness in opportustic communications. To this aim Pareto, max-min, proportional-fair and α-fair policies were introduced and a focus was made on the scheduling algorithm which achieves the proportional-fair policy. Finally NOMA was introduced and associated to opportunistic beamforming.

In Chapter 5 we started by stating the objectives of the simulations to be performed and to give a general description of the way scenarios were generated. But the chapter also included a new approach which allows to compute the CDF of the product of the beamforming and fast fading gains and of its maximum over K such random variables, as well as the CDF of derived quantities like the corresponding rate. CDFs from which one can efficiently compute the expected value or variance with a single integral if needed, whereas a direct computation would involve multiple integration over $2 \times K$ independent random variables.

In the first section of Chapter 6 the Single Beam Opportunistic Beamforming rate was first computed for up to 1000 transmitting antennas and 5000 users. An insufficient and sufficient number of users regime (where the beamforming gain approaches M) were identified, and the fixed linear relation between the necessary number of users and the number of transmitting antennas was pointed out. The same experiment was repeated in the Equispaced Beams OBF case and this time the sufficient number of users regime was identified by a rate approaching the full beamforming and multiplexing gains. The necessary number of users turned out to maintain a linear relation with the number of transmitting antennas but this time the linearity coefficient wasn't fixed but increased with the transmission SNR. Finally a comparison with the asymptotical sum-rate capacity of the channel was performed by considering both the full spatial multiplexing gain and multiuser diversity gains, and it was shown that the EB-OBF scheme approaches for all number of users and antennas 90% of this sum-rate capacity if we only take care of reducing the considered number of users to compute the multiuser diversity gain from the total number of users in the system to the fraction of users in a single beam sector, in order to take into account the impossibility of one user to be the best one for multiple beams at the same time.

After this in the second section of Chapter 6 we moved from the sum-rate analysis to the fairness analysis in non-homogeneous networks. As a metric we used the sum of the logarithms of the time-average throughputs of all the users, which is the one maximized by proportional-fair policies. We considered in our

random equispaced beamforming protocol many policies to determine to which user to transmit. These policies where meant to cover different choices for the trade-off between multiuser diversity and fairness. The proportional-fair policy was used as a reference for fairness and was shown to keep the second best sum-rate after the opportunistic beamforming policy, with a fixed offset from OBF corresponding to a power penalty of around 5dB for the considered scenarios. These results were obtained by considering the possibility of scheduling only one user per beam, but were then extended to the possibility to simultaneously schedule two of them through OMA or NOMA. The best Proportional-Fair policy in the two cases was evaluated and an little advantage of the NOMA over OMA was measured both in terms of fairness and sum-rate.

As a conclusion we can say that Opportunistic Beamforming appears an optimal solution for mmWaves channel for a multitude of reasons: channel directivity, low transmitter and receiver complexity, small required feedback... Furthermore the factors which made it unfeasible at lower frequencies disappear in mmWaves, like the necessary number of users passing from an exponential growth in the number of transmitting antennas to a linear one or the possibility to fit much more mmWaves antenna elements in an array of the same size. Opportunistic beamforming thus result a very promising solution for the downlink of cellular networks in mmWaves frequencies. The analysis carried out in this thesis was of course at high level of abstraction, so simulations with more realistic channel models should be performed to see if the exposed results still hold. Our analysis was also lacking a proper characterization of the coherence time, and the consequent separation of the analysis in fast-fading and low fading regimes which may influence the performances as it may not be possible to have instantaneous channel quality information in fast-fading regime. Finally the influence of multiple antennas at receivers should be checked, as well as the effect of having many neighbouring cells performing OBF, the inter-cell interference may be a problem but we can also imagine some advantages in coordinated transmission or position estimation through triangulation.

Bibliography

[1] M. Sayed, "Millimeter wave tests and instrumentation," in *65th ARFTG Conference Digest, 2005. Spring 2005*, (Long Beach, CA, USA), pp. 20–24, IEEE, 2005. xiii, 11

[2] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, pp. 101–107, June 2011. xiii, 12, 14

[3] 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. xiii, 10, 13

[4] B. Schulz, "LTE Transmission Modes andBeamformingWhite Paper," July 2015. xiii, 17

[5] A. E. Gamal and Y.-H. Kim, "Broadcast communication system," p. 16, 2014. xiii, 24

[6] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity Limits of MIMO Systems," p. 68, 2005. xiii, 25

[7] M. Ding and H. Luo, *Multi-point Cooperative Communication Systems: Theory and Applications*. Signals and Communication Technology, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. xiii, 28

[8] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal Multiuser Transmit Beamforming: A Difficult Problem with a Simple Solution Structure," *IEEE Signal Processing Magazine*, vol. 31, pp. 142–148, July 2014. arXiv: 1404.0408. xiii, 41, 42, 43

[9] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-User Precoding and Channel Estimation for Hybrid Millimeter Wave Systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 1576–1590, July 2017. xiii, 47

[10] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "A Comparison of MIMO Techniques in Downlink Millimeter Wave Cellular Networks With Hybrid Beamforming," *IEEE Transactions on Communications*, vol. 64, pp. 1952–1967, May 2016. xiii, 48

[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," *IEEE Communications Magazine*, vol. 55, pp. 185–191, Feb. 2017. xiii, 62

[12] T. E. Bogale, X. Wang, and L. B. Le, "Chapter 9 - mmWave communication enabling techniques for 5g wireless systems: A link level perspective," in *mmWave Massive MIMO* (S. Mumtaz, J. Rodriguez, and L. Dai, eds.), pp. 195–225, Academic Press, Jan. 2017. 9

[13] C. M. VNI, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper," Feb. 2019. 9

[14] Qualcomm, "Webinar - Breaking the Wireless Barriers to Mobilize 5g NR mmWave," Jan. 2019. 10

[15] M. Shafi, J. Zhang, H. Tataria, A. F. Molisch, S. Sun, T. S. Rappaport, F. Tufvesson, S. Wu, and K. Kitao, "Microwave vs. Millimeter-Wave Propagation Channels: Key Differences and Impact on 5g Cellular Systems," *IEEE Communications Magazine*, vol. 56, pp. 14–20, Dec. 2018. 10, 11, 12, 13

[16] T. S. Rappaport, R. W. H. Jr, R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Prentice Hall, Sept. 2014. Google-Books-ID: _Tt_BAAAQBAJ. 10

[17] Z. Qingling and J. Li, "Rain Attenuation in Millimeter Wave Ranges," in *2006 7th International Symposium on Antennas, Propagation EM Theory*, pp. 1–4, Oct. 2006. 11

[18] I. Shayea, T. A. Rahman, M. H. Azmi, and M. R. Islam, "Real Measurement Study for Rain Rate and Rain Attenuation Conducted Over 26 GHz Microwave 5g Link System in Malaysia," *IEEE Access*, vol. 6, pp. 19044–19064, 2018. 11

[19] I. K. Jain, R. Kumar, and S. Panwar, "Driven by Capacity or Blockage? A Millimeter Wave Blockage Analysis," in *2018 30th International Teletraffic Congress (ITC 30)*, vol. 01, pp. 153–159, Sept. 2018. 12

[20] Qualcomm, "White Paper: 5g NR Millimeter Wave Network Coverage Simulation," Oct. 2017. 12

[21] M. Jacob, S. Priebe, R. Dickhoff, T. Kleine-Ostmann, T. Schrader, and T. Kurner, "Diffraction in mm and Sub-mm Wave Indoor Propagation Channels," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, pp. 833–844, Mar. 2012. 13

[22] Z. Lin, X. Du, H. Chen, B. Ai, Z. Chen, and D. Wu, "Millimeter-Wave Propagation Modeling and Measurements for 5g Mobile Networks," *IEEE Wireless Communications*, vol. 26, pp. 72–77, Feb. 2019. 14

[23] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge: Cambridge University Press, 2005. 15

[24] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *2013 IEEE Global Communications Conference (GLOBECOM)*, pp. 3679–3684, Dec. 2013. 18

[25] G. Lee, Y. Sung, and M. Kountouris, "On the Performance of Random Beamforming in Sparse Millimeter Wave Channels," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 560–575, Apr. 2016. 18, 54, 57

[26] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the Performance of Non-Orthogonal Multiple Access in 5g Systems with Randomly Deployed Users," *IEEE Signal Processing Letters*, vol. 21, pp. 1501–1505, Dec. 2014. 18

[27] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, pp. 2–14, Jan. 1972. 22

[28] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Transactions on Information Theory*, vol. 19, pp. 197–207, Mar. 1973. 22

[29] M. Costa, "Writing on dirty paper (Corresp.)," *IEEE Transactions on Information Theory*, vol. 29, pp. 439–441, May 1983. 24

[30] G. Caire and S. Shamaiz, "On Achivable Rates in a Multi-Antenna Broadcast Downlink," p. 10, 2000. 24

[31] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel," *IEEE Transactions on Information Theory*, vol. 52, pp. 3936–3964, Sept. 2006. 24

[32] H. Bölcskei, ed., *Space-time wireless systems: from array processing to MIMO communications*. Cambridge ; New York: Cambridge University Press, 2006. OCLC: ocm63186217. 24

[33] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49, pp. 2658–2668, Oct. 2003. 25, 27

[34] Y. Xie and C. N. Georghiades, "Some results on the sum-rate capacity of MIMO fading broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 5, pp. 377–383, Feb. 2006. 28, 39

[35] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 684–702, June 2003. 30, 32

[36] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO Broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, pp. 1783–1794, May 2005. 31

[37] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, pp. 1691–1706, July 2003. 31, 33

[38] N. Jindal, "High SNR analysis of MIMO broadcast channels," in *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pp. 2310–2314, Sept. 2005. 33

[39] B. Hassibi and M. Sharif, "Fundamental Limits in MIMO Broadcast Channels," *IEEE Journal on Selected Areas in Communications*, vol. 25, pp. 1333–1344, Sept. 2007. 35, 36, 46, 54

[40] S. Shim, C.-B. Chae, and R. W. Heath Jr, "A Lattice-Based MIMO Broadcast Precoder for Multi-Stream Transmission," *arXiv:cs/0605117*, May 2006. arXiv: cs/0605117. 39

[41] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization," *IEEE Transactions on Communications*, vol. 53, pp. 195–202, Jan. 2005. 39

[42] H. Harashima and H. Miyakawa, "Matched-Transmission Technique for Channels With Intersymbol Interference," *IEEE Transactions on Communications*, vol. 20, pp. 774–780, Aug. 1972. 39

[43] S. Serbetli and A. Yener, "Transceiver optimization for multiuser MIMO systems," *IEEE Transactions on Signal Processing*, vol. 52, pp. 214–226, Jan. 2004. 40

[44] Jinfan Zhang, Yongle Wu, Shidong Zhou, and Jing Wang, "Joint linear transmitter and receiver design for the downlink of multiuser MIMO systems," *IEEE Communications Letters*, vol. 9, pp. 991–993, Nov. 2005. 41

[45] M. Sharif and B. Hassibi, "Scaling laws of sum rate using time-sharing, DPC, and beamforming for MIMO broadcast channels," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, pp. 175–, June 2004. 43

[46] D. Love, R. Heath, V. N. Lau, D. Gesbert, B. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, pp. 1341–1365, Oct. 2008. 44

[47] N. Jindal, "MIMO Broadcast Channels with Finite Rate Feedback," *arXiv:cs/0603065*, Mar. 2006. arXiv: cs/0603065. 44

[48] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid Beamforming for Massive MIMO: A Survey," *IEEE Communications Magazine*, vol. 55, pp. 134–141, Sept. 2017. 48

[49] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, pp. 1277–1294, June 2002. 53, 73

[50] G. Lee, Y. Sung, and J. Seo, "Randomly-Directional Beamforming in Millimeter-Wave Multiuser MISO Downlink," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 1086–1100, Feb. 2016. 53, 54

[51] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, pp. 506–522, Feb. 2005. 54

[52] F. Baccelli, "Stochastic Geometry and Wireless Networks: Volume II Applications," *Foundations and Trends® in Networking*, vol. 4, no. 1-2, pp. 1–312, 2009. 59

[53] M. Kountouris and D. Gesbert, "Memory-based opportunistic multi-user beamforming," in *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pp. 1426–1430, Sept. 2005. 60

[54] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "NOMA: An Information Theoretic Perspective," *arXiv:1504.07751 [cs, math]*, Apr. 2015. arXiv: 1504.07751. 61

[55] Z. Ding, P. Fan, and H. V. Poor, "On the coexistence of non-orthogonal multiple access and millimeter-wave communications," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2017. 61

[56] M. B. Shahab, M. Irfan, M. F. Kader, and S. Young Shin, "User pairing schemes for capacity maximization in non-orthogonal multiple access systems," *Wireless Communications and Mobile Computing*, vol. 16, no. 17, pp. 2884–2894, 2016. 62

[57] W. Yuan, V. Kalokidou, S. M. D. Armour, A. Doufexi, and M. A. Beach, "Application of Non-Orthogonal Multiplexing to mmWave Multi-User Systems," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–6, June 2017. 62

[58] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X. Xia, "Joint Power Allocation and Beamforming for Non-Orthogonal Multiple Access (NOMA) in 5g Millimeter Wave Communications," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 2961–2974, May 2018. 62

[59] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "A Multi-Beam NOMA Framework for Hybrid mmWave Systems," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–7, May 2018. 62

[60] J. Jiang, M. Lei, and H. Hou, "Downlink Multiuser Hybrid Beamforming for MmWave Massive MIMO-NOMA System with Imperfect CSI," 2019. 62

[61] J. Miller, "CUPID: A MATLAB Toolbox for Computations with Univariate Probability Distributions," p. 36, Aug. 2019. 69

[62] Z. Ding, P. Fan, and H. V. Poor, "Random Beamforming in Millimeter-Wave NOMA Networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017. 74