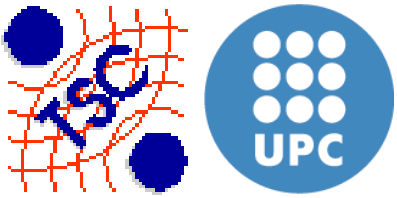


Fairness Analysis of Wireless Beamforming Schedulers

Ph.D. Dissertation by

Diego Bartolomé Calvo
E-mail: dbartolome@ieee.org

Thesis Advisor: Dr. Ana I. Pérez-Neira, Associate Professor
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)
E-mail: anuska@gps.tsc.upc.es

10^{th} November 2004

Estas páginas que empiezan son sólo una travesía que no cesa ...

A mis padres, a mi hermano, y a Montse.

Abstract

This dissertation is devoted to the analysis of fairness at the physical layer in multi-antenna multi-user communications, which implies a new view on traditional techniques. However, the degree of equality/inequality of any resource distribution has been extensively studied in other fields such as Economics or Social Sciences. Indeed, engineers usually aim at optimizing the total performance, but when multiple users come into play, the overall optimization might not necessarily be the best thing to do. For instance in wireless systems, the user with a bad channel condition might suffer the consequences from the selective choice based on the instantaneous channel quality made by a centralized entity. In this sense, the problem has four different perspectives: antenna processing, power allocation, bit allocation, and combination of space diversity (SDMA) with multiple subcarriers (OFDM).

Before the technical content, the landscape where this dissertation is contained is described in detail. In order to put the basis for the following work, a review on broadcast and multiple access channels is conducted, since issues such as the existing duality among them are interesting for the development of multi-antenna techniques. Moreover, dirty paper coding, which attains the capacity region of the broadcast channel, shall be put in context, since it is compared to other traditional schemes in a following chapter. Then, the alternatives for the transmit processing are discussed, after which the scheduling problem is addressed, primarily focusing on how multiple antennas might affect the schedulers. Once a multi-antenna strategy is chosen, power and bit allocation can be done in order to adapt the system to the instantaneous channel conditions and obtain a high gain. Finally, an overview of fairness is given, ranging from modern portfolio selection to the Gini index as a measure of the degree of inequality of a resource allocation.

The technical contribution of the author starts with the analysis of fairness conducted not only for transmit processing, but also for the upper bound that represents the cooperative strategy between the transmitter and the receiver. The SNR analysis for zero forcing, dirty paper, and the cooperative scheme, is based on portfolio theory, and basically consists of the computation of the mean and the variance of each scheme. Interestingly, a higher mean performance comes at the expense of a higher variance in the resource allocation. Whereas in these antenna array techniques, the fairness is implicit, it is made explicit afterwards by the selection of a

power allocation technique with a zero forcing beamforming. The traditional objective functions available in the literature are here compared in terms of fairness, i.e. not only the mean or sum value are analyzed, but also the minimum and the maximum. It can be stated that optimizing the global performance of a cell (e.g. a minimum sum BER or maximum sum rate techniques) comes at the expense of an uneven distribution of the resources among the users. On the other hand, max-min techniques tend to distribute the resources more equally at the expense of loosing in global performance.

Moreover, the game-theoretic power allocation is compared to traditional techniques, and it is shown that the widespread utility function in this context yields an unacceptable BER. Therefore, the optimizing criterion shall be carefully chosen to avoid undesirable operating consequences. Another interesting problem is the admission control, that is, the selection of a subset of users that are scheduled for transmission. Usually, this selection shall be done because the QoS requirements of the communications, e.g. in terms of delay or error rate, prevent all the users from being served. A new algorithm is proposed that balances between the traditional techniques on the extremes of the fairness axis, the uniform power allocation and the equal rate and BER scheme.

After that, the fairness analysis is conducted for the integer bit allocation. First, the traditional approach of the maximization of the sum rate is opposed to the maximization of the minimum rate technique, which ultimately assigns an equal number of bits for all the users. Again, the centralized controller shall balance between the global performance and the individual needs. Nevertheless, an algorithm is proposed, which yields an intermediate behavior among the other traditional schemes. Then, an extension is developed in order to combine the spatial diversity with frequency diversity, that is, SDMA/OFDM systems are analyzed and the initial algorithms for SDMA are extended for such a case. Since the objective functions are NP-complete and very hard to solve even with moderate number of users and antennas, several suboptimal solutions are motivated. Moreover, practical issues such as signaling or a reduction in complexity are faced from a clear engineering point of view.

The final conclusion is that the choices in a multi-antenna multi-user wireless systems are not straightforward, since there exist several trade-offs, among others: performance vs. complexity, global performance vs. individual needs, and performance vs. signaling. The ultimate behavior of the system deeply depends on the design made by the manufacturers, which is not trivial. Moreover, things would get even more complicated if the higher layers of the protocol stack were taken into account, which should be subject of further work.

Resum

Aquesta tesi es dedica a l'anàlisi de la justícia a la capa física en entorns de comunicacions amb múltiples antenes i diversos usuaris, cosa que implica un nou punt de vista sobre problemes tradicionals. Malgrat això, el grau d'equitat o desigualtat en la distribució de recursos ha estat estudiat en profunditat en altres camps com Economia o Ciències Socials. En el fons, el enginyers tendeixen a optimizar les prestacions globals, però quan hi ha múltiples usuaris en escena, aquella optimització no és necessàriament la millor opció. En sistemes mòbils, per exemple, l'usuari amb unes males condicions de canal pot patir les conseqüències d'un controlador central que basi les seves decisions en la millor qualitat instantània del canal. En aquest sentit, el problema s'encara des de quatre perspectives diferents: processament d'antenes, assignació de potència, assignació de bits, i combinació de diversitat en espai (SDMA) amb múltiples subportadores (OFDM).

Abans del contingut tècnic, es descriu en detall l'entorn on s'emmarca aquesta tesi. Per tal de posar les bases pel treball posterior, es comenten els trets característics dels canals *broadcast* i d'accés múltiple, ja que temes com la dualitat existent entre ells són interessants pel desenvolupament de tècniques multi-antena. A més, la codificació *dirty paper* s'ha de posar en context, donat que aconsegueix la capacitat del canal *broadcast* i es compara amb esquemes més tradicionals en un capítol posterior. Després, es discuteixen les alternatives disponibles al transmisor, així com el problema de l'assignació de recursos, on l'autor es concentra en veure com les múltiples antenes els hi afecten. Un cop s'ha seleccionat una tècnica multi-antena, s'ha de fer la distribució de potència i de bits per tal d'adaptar-se a les condicions instantànies del canal i obtenir-ne un gran guany. Finalment, també es fa una revisió de la justícia, anant des de la selecció de carteres fins a l'índex de Gini com una mesura del grau de desigualtat d'una assignació de recursos.

La contribució tècnica de l'autor com a tal comença amb l'anàlisi de la justícia no només pel processament al transmissor, però també pel límit superior que representa la tècnica cooperativa entre el transmissor i el receptor. L'anàlisi de SNR pel forçador de zeros, el *dirty paper* i l'estratègia cooperativa entre transmissor i receptor està basada en la teoria de carteres, i consisteix bàsicament a calcular la mitja i la variància de cada esquema. Es veu que una mitja superior ve donada per una major variància en l'assignació de recursos. Així com a aquestes

tècniques d'antenes, la justícia hi és implícita, es fa totalment explícita en la tria d'una tècnica de distribució de potència amb un conformador forçador de zeros. Llavors, les funcions objectiu tradicionals a la literatura es comparen en termes de justícia, això és en termes del màxim i el mínim, a més de la mitja o la suma. Aquí es pot veure que optimitzar les prestacions globals d'una cel·la (p.ex. tècniques de mínima suma de BER o màxima suma de *rate*) implica una distribució més desigual dels recursos entre els usuaris. Per una altra banda, les tècniques max-min tendeixen a fer una distribució dels recursos més paritària entre els usuaris, alhora que perden en prestacions globals.

A més, l'assignació de potència basada en teoria de jocs es compara a les tècniques tradicionals, i es mostra que la funció d'utilitat àmpliament utilitzada en aquest context té una taxa d'error inacceptable. Llavors, la funció a optimitzar s'ha de triar de forma acurada, per tal d'evitar possibles conseqüències indesitjables. Un altre problema interessant és el control d'admissió, és a dir, la selecció d'un subconjunt d'usuaris que han de ser servits simultàniament. Normalment, el control d'admissió és necessari per complir els requeriments de les comunicacions, en termes de retard o taxa d'error, entre d'altres. Es proposa un nou algoritme que està entre mig de les tècniques tradicionals a l'eix de la justícia, l'assignació uniforme de potència i l'esquema que dóna igual *rate* i BER a tots els usuaris.

Després d'això, l'anàlisi de la justícia es fa per l'assignació de bits. Primer, el punt de vista tradicional de la maximització de la suma de *rates* es contraposa a la maximització de la mínima *rate*, que finalment assigna a tots el usuaris un número igual de bits. Un altre cop, el controlador central ha de balancejar les necessitats individuals amb les prestacions globals. Malgrat això, es proposa un algoritme que té un comportament intermig entre els esquemes tradicionals. A més, s'estudien una extensió per tal de combinar la diversitat en espai amb la freqüencial, per tant, s'analitzen sistemes SDMA/OFDM, pels quals s'extenen els algoritmes inicialment dissenyats per SDMA. Com que les funcions objectiu són NP-completes i molt difícils de resoldre fins i tot amb un nombre moderat d'usuaris i antenes, les solucions subòptimes són clarament bones candidates. A més, temes pràctics com la senyalització i la reducció en complexitat són tractats des d'un clar punt de vista d'enginyeria.

La conclusió final és que les decisions no són trivials en sistemes mòbils multi-antena multi-usuari, perquè existeixen bastants compromisos, entre d'altres: prestacions/complexitat, prestacions globals/necessitats individuals, i prestacions/senyalització. El comportament final del sistema depèn fortament del disseny fet pel fabricant, cosa que no és fàcil. A més, les tries es complicarien encara més si es tinguessin en compte les capes superiors de la pila de protocols, però això forma ja part del treball futur.

Acknowledgements

En primer lugar debo expresar mi mayor agradecimiento a mi directora de tesis, Ana I. Pérez Neira, por su continuo apoyo y motivación sin límites, sin los cuales no se habría llegado a este resultado final. Valoro enormemente sus comentarios que han llevado a buen puerto un buen cúmulo de ideas a veces desordenadas. Realmente, Ana, ha sido un placer trabajar a tu lado durante este tiempo, y no sólo estoy feliz por tu calidad técnica, sino también, y muy especialmente, por como eres en la corta distancia, en el trato personal. Espero que tú también hayas disfrutado en este periodo.

No puedo continuar sin nombrar al CTTC[1], que me acogió para acabar la tesis doctoral. Las gracias se dirigen en primer lugar al director Miguel Ángel Lagunas y al subdirector Carles Antón, por contar conmigo en numerosas ocasiones y haber tenido siempre la puerta abierta para cualquier cosa. Sin Patricia, Aurora, ni la Carme todo hubiera sido mucho más complicado. Agraït, és clar, ho estic a tothom del CTTC, però en especial al Jordi M., al Jordi C., a la Marta, al Marc R., a la Patricia V., i al Josep. També al Jordi S. i al Raül. Y qué decir, los días hubieran sido mucho más aburridos sin mis compañeros de despacho, Miquel, Ricardo, José y Toni M., y sin el resto de ingenieros pre-doctorales Francisco, Pavel, y Luis. Gracias también a los colegas de proyectos, Stephan, Mònica, Christian, y Bader.

Quisiera también agradecer a dos personas que, sin quizás ellos saberlo, tienen gran parte de *culpa* de que esta tesis se haya escrito, Ferran Marqués y Sebastià Sallent. Al Ferran, no només per *portar-me* a Hildesheim i a París, sinó també perquè durant aquell mes de Novembre del 2000, quan et vaig preguntar si hi havia possibilitats de marxar fora, tu em vas dirigir a l'Ana. I al Sebastià, durant les teves classes de doctorat vaig començar a interessar-me pels temes de justícia. També, anònimament, estic agraït als meus professors d'Economia, i al Simó Aliana, per algunes converses molt fructíferes, així com a l'Albert Sitjà.

Je veux remercier aussi tout le group BSTL à Motorola Labs, qui sont mes prèmiers *steps* dans la recherche. Je suis très heureux d'avoir travaillé avec le *Papa* Seb et Marc. Je ne peux pas m'oublier de ceux à qui j'ai trouvé dans quelques confs. Bedanken möchte ich mich auch bei

[1]This dissertation has been partially supported by the CTTC and by the Catalan Government under grant 2003FI 00190.

Frank Hofmann und Hendrik Fuchs bei Bosch Hildesheim.

Un agradecimiento muy especial va también para Toni P., quien fue fundamental en los primeros pasos del doctorado, con sus consejos y apoyo. Asimismo, tampoco me puedo olvidar de otros compañeros de *papers*, el mestre Xavi, i evidentment no puc deixar-me al *crack* Dani. Je veux aussi remercier Christine, Patrick, et Laurent du LETI pour la très interessant collaboration. Finalmente, agradezco la colaboración del Pijoan y del Carles Vilella de La Salle, así como a mis *sufridos* proyectistas Toni, Ismael, Carlos, y recientemente Javier.

Gracias también a toda la gente de la UPC, en especial a los compañeros de despacho, el D5-118, y del anexo 117. Además, no me puedo olvidar de los excelentes compañeros de viaje Carles y Hugo, así como de mi *pareja de hecho* Maribel, y Luis. I està clar, Joel Solé, merci per estar sempre allà, quants anys ja, el que hem canviat sense deixar de ser nosaltres.

Debo dar gracias también a todas las personas que han estado durante la tesis de algún modo soportándome, y eso tiene mérito. Pese a que nos veamos poco con algunos, sois parte de las páginas que siguen. Al Bernet, al Vinagre, a la Sus, a la Marta L., i a la resta de gent del Penyafort. A la gent de Paris i Hildesheim, als companys de carrera a la UPC, a los *telecos* de Sabadell, a la penya del Casablancas, a los currantes de la Sony en Viladecavalls, a la gente del PIE. A los buenos amigos heredados de la universidad, pese a que la vida haga que nos veamos poco, os quiero: Emi, Jordi, Héctor, Marc Aldea, Xavi Egozcue, Xavi *Lobo*, Víctor, Óscar, Mofi, Gino, Frank, y al papá Pichón. La vida sense els amics perdria el sentit. I a la Fanny, i a la Mariona, i al Xavi Pérez, de Heidelberg, i a la ya doctora Marta Racamonde, gente genial todos. En un espai privilegiat està la gent de la colla de Sabadell, tots, però en especial el Marc, el Carles, l'Oriol, l'Albert, el Jesús, el Xavi, el Jordi Caralt, el Roger, el Óscar, el Villegas, el Ruchi ... i xicotes. Quants moments viscuts plegats, i els que queden per passar, us estimo, amics. Ho celebrarem bé, us ho prometo, al nou piset. I també, és clar, el Young, ahora en New Jersey, you cannot imagine how much I miss you, man.

Y ya acabar, recuerdo a toda la lejana familia, primos, tíos, abuelos, pero en especial, por todo, mis padres, mi gran y excepcional hermano, y muy cerca de mí, Montse, la bogeria em dominaria sense tu. T'estimo, os quiero.

A todos estos, i a la resta de gent que ha estat amb mi pel camí, un gran abrazo.

Diego Bartolomé

Contents

List of Figures

List of Tables

Acronyms

ADSL	Asymmetric Digital Subscriber Line
AP	Access Point
ARQ	Automatic Repeat reQuest
AWGN	Additive White Gaussian Noise
BC	Broadcast Channel
BGU	Best Group to the User
BLAST	Bell Labs Layered Space-Time architecture
BS	Base Station
BUG	Best User to the Group
CDF	Cumulative Density Function
CDMA	Code Division Multiple Access
CSI	Channel State Information
DFE	Decision Feedback Equalization
DLC	Data Link Control Layer
DMT	Discrete Multi-Tone
DOA	Direction of Arrival
DP	Dirty Paper
DPC	Dirty Paper Coding
DRM	Digital Radio Mondiale
DS-CDMA	Direct Sequence CDMA
DSL	Digital Subscriber Line
DVB	Digital Video Broadcasting
EPS	Equal Proportional SNR
ERB	Equal Rate and BER
FCFS	First Come First Served
FDMA	Frequency Division Multiple Access
FF	First Fit
FFT	Fast Fourier Transform

FSR	Frame Success Rate
GPS	Generalized Processor Sharing
GT	Game Theory
HDR	Qualcomm's High Data Rate system
IEEE	Institute of Electrical and Electronics Engineers
IF	Index of Fairness
IFFT	Inverse FFT
KKT	Karush-Kuhn-Tucker
LAN	Local Area Network
LQHR	Longer Queue Higher Rate
MAC	Multiple Access Channel
MC-CDMA	Multi-Carrier CDMA
MISO	Multiple Input Single Output
MIMO	Multiple Input Multiple Output
MMMR	Multi-carrier Maximization of the Minimum Rate
MMR	Maximum Minimum Rate
MMSE	Minimum Mean Squared Error
MMSR	Multi-carrier Maximization of the Sum Rate
MRC	Maximum Ratio Combining
MSB	Minimum Sum BER
MSE	Mean Squared Error
MSR	Maximum Sum Rate
MSU	Maximum Sum of Utilities
NE	Nash Equilibrium
NP	Non-deterministic Polynomial-time
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PER	Packet Error Rate
PF	Proportional Fair
PHY	Physical Layer
QAM	Quadrature Amplitude Modulation
RF	Radio Frequency
rms	root mean square
SAC	Spatial Admission Control
SDMA	Space Division Multiple Access
SINR	Signal to Interference and Noise Ratio
SIR	Signal to Interference Ratio

SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
THP	Tomlinson Harashima Precoding
UPA	Uniform Power Allocation
QoS	Quality of Service
WLAN	Wireless LAN
ZF	Zero Forcing

Notation

$\mathbf{a}$	Vector.
$\mathbf{A}$	Matrix.
$[\mathbf{A}]_{i,j}$	Element at row ith and column jth of $\mathbf{A}$.
$\mathbf{A}^T, \mathbf{A}^*, \mathbf{A}^H$	Transpose, conjugate, and complex conjugate transpose of matrix $\mathbf{A}$.
$\|\mathbf{a}\|$	Norm-2 of vector $\mathbf{a}$, that is $\sqrt{\mathbf{a}^H\mathbf{a}}$.
$det(\mathbf{A})$, $\|\mathbf{A}\|$	Determinant of matrix $\mathbf{A}$.
$tr(\mathbf{A})$	Trace of the square matrix $\mathbf{A}$.
$\mathbf{1}_k$	All-zeros vector but at position kth, where it is one.
$diag(a_1, \ldots, a_n)$	Square matrix with diagonal elements given by $a_1, \ldots, a_n$.
$\mathbf{I}$	Identity matrix, a subscript might indicate the dimension.
$\lambda_{\max}, \lambda_{\min}, \lambda_k$	Maximum, minimum, and kth eigenvalue, respectively.
$\mathbb{C}^{n\times m}$	The set of $n \times m$ matrices with complex random entries.
$\|a\|$	Absolute value of a.
$log(a)$	Natural logarithm of a.
$log_c(a)$	Base-c logarithm of a.
a^+	$\max(a, 0)$.
$\mathbb{E}_x(c)$	Expectation of c over the random variable x.
$\sigma_x^2(c)$	Variance of c over the random variable x.
$\mathcal{K}$	Set.
$\|\mathcal{K}\|$	Cardinality of the set $\mathcal{K}$.
$\propto$	Equal up to a scaling factor.
$\cup$	Union
i.i.d.	independent and identically distributed.
$\sim$	Distributed according to.
s.t.	Subject to.
$(\cdot)^*$	Selected value.

Chapter 1

To start: the background

This dissertation is inspired by the recently-created term *cross-layer design*. According to [1], the cross-layer philosophy implies adapting the resource allocation to the channel conditions. In a sense, this has analogies to the traditional *impedance matching* in the implementation of circuits. Although the cross-layer philosophy impinges all the layers of any communications system, in this dissertation the main focus is at the two lower layers, i.e. the PHY and the DLC, and in general the optimization procedures are adapted instantaneously to the quality of the links. Even though it seems clear that the schedulers at the DLC might benefit from the knowledge of the instantaneous channel condition, it is not always so obvious that a joint design will outperform traditional layered approaches [2]. It is argued that the perhaps surprising success of the Internet is mainly due to the layered architecture, since when networks grow larger, architectural issues play a key role. In any case, in wireless systems the knowledge of the channel has an impact on the mechanisms of the second layer, and even on the third layer, e.g. routing strategies in ad-hoc networks [3]. Obviously, there are substantial differences between wired and wireless networks, such as the variation of the channel, either in terms of fast fading or long-term path-loss, which might confirm the fact that the cross-layer interaction is beneficial in wireless. Analogously in information theory, the cross-layer design refers to the fact that the randomness of the packet arrivals shall be included in the model, and its discussion comes from a long time ago, see [4]. Although it is not the primary focus of the dissertation, in the pioneering paper [5] queuing theory is combined with information theory in a MAC.

With a consensus on the fact that a cross-layer design impacts on the design of the (wireless) scheduler and might be included in information-theoretic models, other relevant topics shall be covered. For instance, the ever-increasing frequency band in wireless networks makes it possible to include more than a single antenna not only at the BS or AP, but also at the battery-scarce terminals. Indeed, multiple antennas deeply enhance the system performance [6], but the achievable gain depends on the type of channel knowledge that transmitters and receivers

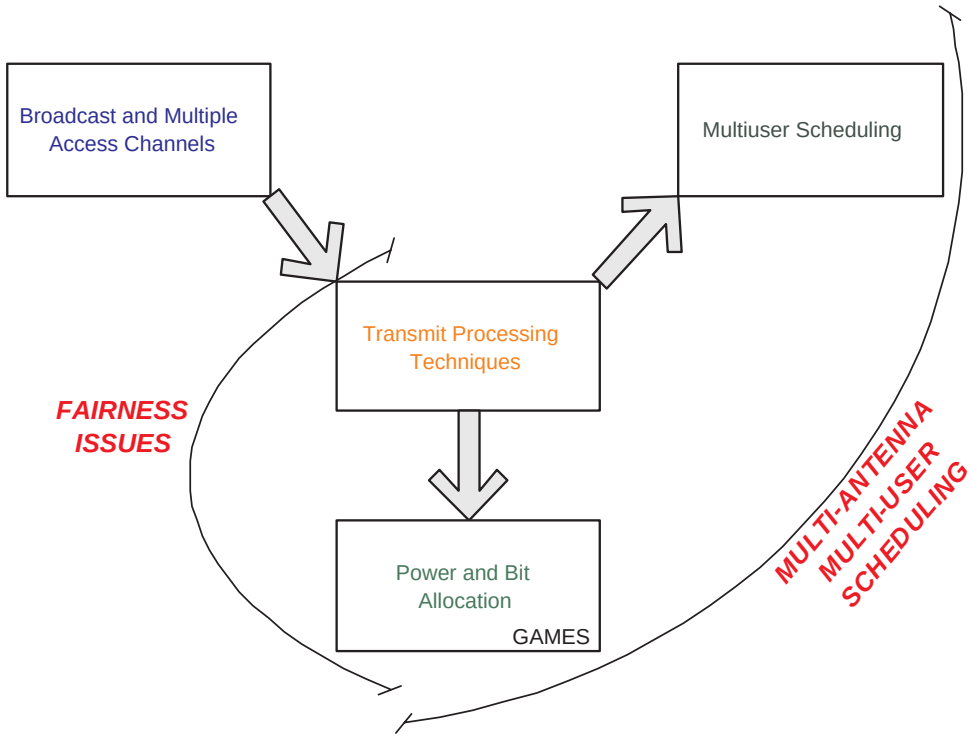


Figure 1.1: Overview of the topics that will be covered in this background chapter. After a review of some important information-theoretic aspects, a survey on practical transmit processing techniques is done. This leads both to the explanation of techniques for power and bit allocation, as well as multi-user scheduling at the physical layer. Finally, fairness considerations are explained, since they cover all aspects of the work that has been performed.

have. Although spatial diversity is a promising technique to increase the capacity of wireless networks [7], the global complexity is increased (and also the cost of the devices). Furthermore, the degrees of freedom shall be efficiently used since if the system optimization is not properly made, the overall performance might even degrade with respect to the use of a single antenna. Reference [8] gathers important results and future challenges for spatial diversity in order to provide QoS, noting that the challenge today lies on real-time applications in heterogeneous networks. Additionally, concerning wireless networks, the research challenges quoted in [9] are: scheduling mechanisms that interact well with TCP at the transport layer, scheduling in multi-carrier systems, fundamental research on properties of multi-user diversity, fast deployment of new algorithms, multi-hop networks, and modeling network performance. In this sense, wireless multi-carrier communications are the best-positioned technique, with its various variants either in wired or wireless environments, among others, Discrete Multi-Tone (DMT), the OFDM modulation in Wireless LAN standards such as the *old* Hiperlan/2 or IEEE 802.11a, Digital Video Broadcasting (DVB), or the standard for digital AM Digital Radio Mondiale (DRM). The advantages include simplicity and the ability to transform the frequency-selective channel

impulse response into a set of parallel flat-fading subchannels [10].

Before going deeper into details, the objective of this chapter is to put the basis for the dissertation contained in the following chapters. This task is not straightforward, because a mixed background is necessary. Since research is needed on fundamental properties of multi-user diversity and multi-carrier systems [1], the basic topic of this dissertation is multi-antenna multi-user communications, with a brief inclusion into OFDM systems. Even more, how fairness is seen in the degrees of freedom included in the joint physical layer and DLC optimization is the main issue throughout the dissertation. The aspects that will be covered are basically those depicted in Figure 1.1. First, an overview of the information-theoretic framework underlying the multi-antenna broadcast channel is given. The basic ideas are presented, as well as key concepts such as duality, which drive the reader into recent results that prove that dirty paper coding strategies achieve not only the sum capacity but also the whole capacity region of the broadcast channel [11]. After that, it is outlined why beamforming techniques might be well-suited in realistic implementations of communication systems, and then the alternatives that have been proposed in the literature are described in detail. If multiple users are active in the cell, the scheduler might distribute the scarce resources according to a certain criterion. In this sense, the cooperation between the DLC and the physical layer plays a very important role in current and future wireless systems. Typically, the instantaneously-constrained power shall be shared among users according to the *multi-dimensional* scheduler, which also impacts the (spatial) bit allocation in practical scenarios. In this sense, Game Theory [12] is relevant as it is confirmed by the recent literature on CDMA systems. On top of all this, fairness issues influence all the choices and will be the main discussion topic throughout the dissertation.

First, a **summary** of the contents of this chapter shall be given.

- **Section 1.1** deals with the optimization of the transmit covariance matrices of information-theoretic models such as the Broadcast Channel (BC) or the Multiple Access Channel (MAC). This constitutes the background of the technical work conducted in this dissertation.

- Then, **Section 1.2** looks into the choice of the realistic transmit beamforming matrices. Since the link covered in this dissertation is the downlink, the topics are the design of the precoding schemes and the transmit beamforming matrix, see (1.5). It shall be noted that the beamforming is a design assumption from Chapter 3, particularly Zero Forcing has been selected. In this section, the literature is reviewed to justify this concrete choice.

- After that, the scheduling procedures in **Section 1.3** try to select the best subset of users that shall be served at any time instant. In other words, provided a practical multi-antenna technique, the best subset of users (thus the number of users) among those that are active shall be selected in order to obtain the best system performance.

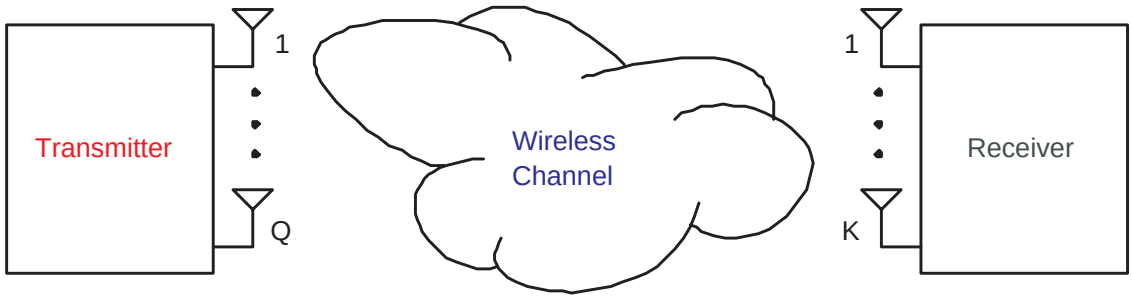


Figure 1.2: A typical single-user MIMO channel, where the transmitter is provided with Q antennas, and the receiver has K antennas. When spatial diversity is deployed, the multi-path inherent at the wireless channel might be a beneficial source rather a distortion.

- The resource management in terms of power and bit allocation is described in detail in **Section 1.4**. Once the beamforming technique is selected, the scarce resource, usually the power (and the number of bits per symbol), shall be distributed among the active users.

- After that, fairness considerations are described in detail in **Section 1.5** because it is the main topic throughout the dissertation. Finally, in **Section 1.6**, an overview of the dissertation is given, together with the research results from the Ph.D. period.

1.1 A review on broadcast channels

It is worth mentioning that information theory is not the main field where the research of this dissertation has been conducted, which is rather in the signal processing part. However, it seems that the information-theoretic limits are a necessary step into more practical strategies, since concepts such as uplink-downlink duality are also present in signal processing techniques. Moreover, the Dirty Paper strategy has been compared to other classical schemes in Chapter 2.

1.1.1 From single-user MIMO to multi-user MIMO

First of all, it is necessary to explain the signal model for a single-user MIMO channel, which is depicted in Figure 1.2. In this model, it has been shown that capacity scales linearly with $\min(Q, K)$ in the high SNR regime, which are the number of transmit and receive antennas respectively [13]. However, to achieve these gains it is mandatory that the channel matrix has full rank and independent entries, independent noise entries, and that perfect estimates of these gains are available at the receiver [14]. In practical systems, the actual procedure would be to use the Hermitian right/left matrix of eigenvectors of the channel matrix at the transmitter/receiver, and then perform a classical water-filling over the eigenvalues of the equivalent channel, see e.g. [15]. However, the perfect CSI assumption might not always hold in practice, and there exists a

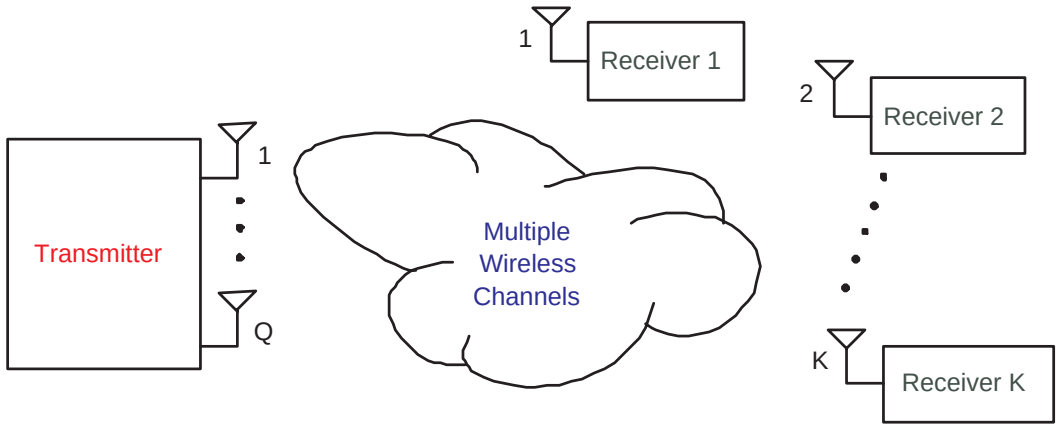


Figure 1.3: A typical multi-user MIMO BC channel. In this case, Q antennas are available at the transmitter side, whereas the K receivers have a single antenna (although they could have more than one in a general case). By reversing the roles of the transmitters and receivers, the MAC channel would be depicted. The BC and its degrees of freedom are the focus of this dissertation.

vast variety of situations depending on the quality and quantity of channel state information. For details, an excellent review of fading channels is conducted in [16], both giving the information-theoretic and the communication points of view.

In multi-user environments, several additional questions arise, because the receivers are no longer a single entity, but rather distributed in the space, see Figure 1.3 for an example of a broadcast channel where the receivers have a single antenna. For instance, in this scenario it is less natural to assume that they have full channel knowledge, since communication among all the terminals would be needed in such a case. This would increase the cell load and required signaling, so the throughput would decrease accordingly. Furthermore, issues related to a fair distribution of the resources should be carefully studied, since some users might be penalized for the sake of the global performance. The focus of the whole dissertation is the Broadcast Channel (BC) depicted in Figure 1.3, however, the dual Multiple Access Channel (MAC) shall also be described because of the existing duality between the BC and the MAC, which is a key step in the recent characterization of the capacity region of the Gaussian MIMO BC [11]. Note that duality also exists in other fields, such as the celebrated duality between downlink and uplink beamforming, see e.g. [17]. In Figure 1.3, the multiple access channel can be obtained by reversing the roles from the transmitter and the receivers.

1.1.2 On broadcast and multiple access channels

This subsection will be devoted to some of the basics in broadcast and multiple-access channels, a field that has dramatically evolved since [18], where at that time *recent* advances on broadcast channels were discussed. Six years later, the capacity region of the Gaussian MIMO broadcast

channel is fully known [11]. The general case would be a single-cell scenario, where it is assumed that there are K receivers with $T \geq 1$ antennas each, and that the transmitter is provided with Q transmit antennas. The signal model for each user k in this BC can be expressed as

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{w}_k, \quad k = 1, \ldots, K, \tag{1.1}$$

where the covariance matrix of the transmitted signal $\mathbf{x} \in \mathbb{C}^{Q \times 1}$ is $\mathbf{\Sigma}_x \triangleq \mathbb{E}[\mathbf{x}\mathbf{x}^H]$. The elements of each channel matrix $\mathbf{H}_k$ are independent and identically distributed complex Gaussian random variables with zero mean and unit variance. The $T \times 1$ received signal for the kth user is $\mathbf{y}_k$, and the noise is circularly symmetric complex Gaussian, i.e. $\mathbf{w}_k \sim \mathcal{N}_{\mathbb{C}}(0, \mathbf{I})$. The AP is allowed to transmit with a maximum power of P_T, that is, $tr(\mathbf{\Sigma}_x) \leq P_T$.

Some remarks are needed before proceeding. If the transmitter has a single antenna, the Gaussian BC is physically degraded, i.e. the users can be absolutely ranked by their channel strength. Therefore, in a single antenna degraded broadcast channel the maximum achievable rate is obtained by transmitting to the best user in the system, that is, the one with a highest channel strength. However, the multiple transmit antenna BC is non-degraded, which means that matrices can be only partially ordered because the users receive different signal strengths at each antenna [19]. This has a severe consequence in the computation of the capacity region, which turns into a non-convex non-linear optimization problem.

For the case where the receivers have a single antenna, an achievable region for the MIMO BC is given by [20] or [21], and it is based on the *writing on dirty paper* principle, which is first described in [22] for a single transmitter and a single receiver in a Gaussian channel, although it has previous (and less known) roots according to that paper. The basic idea behind Dirty Paper Coding (DPC) is that if the transmitter (but not the receiver) has perfect non-causal channel state information regarding an additive interference source, the capacity of the channel remains the same as if there was no interference. In other words, the capacity of the interference channel is equal to the interference-free channel. The key point is that the transmitter shall subtract that interference prior to the transmission of the desired signal. Costa called it *Writing on Dirty Paper* because it models a transmitter which attempts to encode information on a piece of paper partially corrupted by dirt that is seen at the transmitter but it is not known at the receiver.

Logically, these results can be extended to the MIMO BC. First, the transmitter chooses a codeword $\mathbf{x}_K$ for receiver K. After the codeword choice for the second user $(K-1)$, the transmitter can perform a pre-subtraction of the data belonging to the first encoded user, so that the second user is free of interference from the first user. This procedure is repeated for the K users, and it might be foreseen that the performance deeply depends on the encoding order. If user $\pi(K)$ is encoded first, followed by $\pi(K-1)$, and so on, the achievable rate is [23]

$$R(\pi, \mathbf{\Sigma}_i) = \log \frac{|\mathbf{I} + \mathbf{H}_{\pi(i)}(\sum_{j \leq i} \mathbf{\Sigma}_{\pi(j)})\mathbf{H}_{\pi(i)}^H|}{|\mathbf{I} + \mathbf{H}_{\pi(i)}(\sum_{j < i} \mathbf{\Sigma}_{\pi(j)})\mathbf{H}_{\pi(i)}^H|}, \quad i = 1, \ldots, K,$$

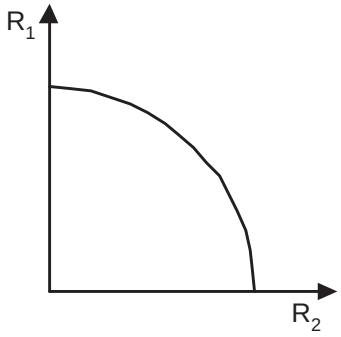


Figure 1.4: Example of a capacity region for the MIMO BC when the receivers have a single antenna.

so that the capacity of the dirty paper region, and also the capacity of the BC, is defined as the convex hull of the union of the rate vectors over all positive-definite covariance matrices such that the power constraint is fulfilled, and over all permutations $(\pi(1), \ldots, \pi(K))$ [23]

$$C_{BC}(P_T; \mathbf{H}) \triangleq Co\left(\bigcup_{\pi, \mathbf{\Sigma}_i} R(\pi, \mathbf{\Sigma}_i)\right),$$

where it shall be noted that the rate equations are in general neither a concave nor a convex function of the covariance matrices, thus the direct numerical finding of the dirty paper region is difficult. For a two user channel with a single antenna each, the rate region is depicted in Figure 1.4. One of the first steps towards the complete characterization of the capacity region of the broadcast channel can be found in [24], which is extended in [25]. The most important concept for the needs of this dissertation is the duality between BC and MAC in information theory. This greatly simplifies the computation of the difficult BC capacity region in (1.2), because the MAC expression is much more simple, as it is shown next. In fact, the duality when transmitters and receivers interchange their roles had been previously shown in the literature in different fields. For instance, [13] shows that the capacity is unchanged when this role interchange is done in a single-user MIMO channel. In the context of a downlink of a multiple antenna system employing simple linear beamforming strategies followed by single-user receivers, [26] and [27] show that the optimal choice of the transmit beamvectors is closely related to a virtual uplink problem, see next section for details. Moreover, the reader is also referred to other existing dualities in the literature such as cyclic prefix and zero padded OFDM, or MC-CDMA and DS-CDMA.

Before proceeding further, the received signal $\mathbf{v}$ for the MAC is analogous to (1.1), i.e.

$$\mathbf{v} = \sum_{k=1}^{K} \mathbf{H}_k^H \mathbf{u}_k + \mathbf{n} = \mathbf{H}^H \mathbf{u} + \mathbf{n}, \tag{1.2}$$

where the transmitted signal $\mathbf{u} = [\mathbf{u}_1^T \ \mathbf{u}_2^T \ \ldots \ \mathbf{u}_K^T]^T$ is now formed by stacking the transmitted vectors for the K users. The channel matrix gathers the individual channel matrices for the users

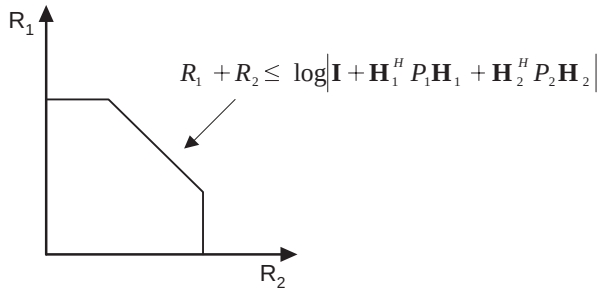


Figure 1.5: Example of a capacity region for the MIMO MAC when the receivers have a single antenna.

in $\mathbf{H}^H = [\mathbf{H}_1^H \ \mathbf{H}_2^H \ \ldots \ \mathbf{H}_K^H]$, and the assumptions for the noise $\mathbf{w}$ hold also for $\mathbf{n}$. In this case, the covariance matrices are denoted by $\mathbf{Q}_k \triangleq \mathbb{E}[\mathbf{u}_k \mathbf{u}_k^H]$, which are subject to an individual power constraint $tr(\mathbf{Q}_k) \leq P_k$, because regulatory authorities usually limit the power per transmitting entity. It shall be noted that this power constraint differs from that of the BC, since in such a case the constraint is global and not individual. In the following, the focus will be the constant channel, since for fading channels with perfect CSI at the transmitters and the receiver, the only difference in terms of notation is an average over the channel statistics in the expression of the capacity region. Moreover, the focus of this dissertation is mainly on the adaptation to the instantaneous channel conditions and not in average. The capacity of any MAC as (1.2) can be written as the convex closure of the union of the rate regions corresponding to every product input distribution $p(\mathbf{u}_1) \ldots p(\mathbf{u}_K)$ [28]. For the Gaussian MIMO MAC, however, it is sufficient to consider only Gaussian inputs and that the convex hull operation is not needed [29]. For any set of powers $\mathbf{P} = (P_1, \ldots, P_K)$, the capacity region of the MIMO MAC is defined as

$$C_{MAC}(\mathbf{P}; \mathbf{H}^H) \triangleq \bigcup_{\substack{\mathbf{Q}_i > 0, \\ tr(\mathbf{Q}_i) \leq P_i, \, \forall i}} \left\{ \begin{array}{l} (R_1, \ldots, R_K) : \\ \sum_{i \in S} R_i \leq \log \left| \mathbf{I} + \sum_{i \in S} \mathbf{H}_i^H \mathbf{Q}_i \mathbf{H}_i \right| \ \forall S \subseteq \{1, \ldots, K\} \end{array} \right\}, \quad (1.3)$$

where each user should transmit a zero-mean Gaussian signal with covariance matrix $\mathbf{Q}_i$. As it is shown in (1.3), each set of covariance matrices determines a K-dimensional polyhedron. Then, the MAC capacity region, which is convex, is equal to the union (over all covariance matrices satisfying the trace constraints) of all such polyhedrons, in which the corner points of each polyhedron can be achieved by successive decoding [14]. Indeed, successive decoding reduces the complex multi-user detection problem into a series of single-user detection steps. If the transmitters have a single antenna, the covariance matrix is simply a scalar equal to the transmitted power, and each user shall transmit at full power to achieve capacity. For the two-user case, the region is a pentagon, see Figure 1.5, the boundaries of which can be achieved by maximizing a weighted sum of the rates assigned to the users [30]. It is important to note that since the MAC capacity region is convex, efficient tools exist [31] e.g. to compute the maximum sum-rate of a MAC when the weighting priorities are equal [29]. This technique is based on the

Karush-Kuhn-Tucker (KKT) conditions, which indicate that the sum rate maximizing covariance matrix of any user should be the single-user water-filling of its own channel with noise equal to the actual noise and interference from the other $K-1$ transmitters.

As stated, the dual MAC is formed by reversing the roles of the transmitters and the receivers in the BC, so the main contribution to the duality concept of [23] is the proof that the achievable rates in the dual MIMO MAC with power constraints whose sum equals the BC power constraint are also achievable in the MIMO BC and *vice versa*. In other words, the BC capacity region is equal to the capacity region of the dual MIMO MAC with sum power constraint P_T, i.e.

$$C_{BC}(P_T, \mathbf{H}) = \bigcup_{\mathbf{P}:\sum_{i=1}^{K} P_i = P_T} C_{MAC}(\mathbf{P}; \mathbf{H}^H),$$

which is a concave expression[2]. In fact, this is the multiple antenna extension of the duality between multiple access and broadcast channels [32]. Although out of the scope of the technical work of the dissertation, the difficulty in the characterization of the broadcast channel appeared to be in proving that Gaussian inputs are optimal for non-rate-sum points. In fact, as it is shown e.g. in [33], the DPC region is the BC capacity region if an additional Gaussianity assumption is made. Nonetheless, it has been recently shown that the sum rate capacity, and in fact the whole capacity region, is achieved with dirty paper techniques [11]. To solve the MAC problem, efficient numerical algorithms inspired by the iterative waterfilling algorithm in [29] exist, see [34], where under a total power constraint, the sum rate optimal covariance matrices are obtained using standard convex optimization techniques. Then, they can be converted into the corresponding optimal BC covariance matrices using the MAC-BC transformations. A specialized algorithm to compute the MAC covariances according to a weighted sum rate criterion is found in [35], so that it is possible to determine the operating points on the boundary of the rate region.

1.1.3 Brief comments on *cross-layer issues*

Recently, there has been an effort to combine information theory with queuing theory, provided the traffic sources are in general bursty, see [5]. Some recent examples are commented here, without the objective of completeness. Among them, the power/delay trade-off in [36] yields a characterization of the different operating points over fading channels with delay constraints, because it is indeed important to consider how data arrives at the PHY from higher layers. In [37], multiaccess communications are studied taking into account the queue states. The authors propose the Longer Queue Higher Rate (LQHR) scheduling strategy so as to minimize the system delay of packets. There, a fundamental lower bound is given on the performance for multiaccess coding schemes which seek to meet any given level of decoding error probability.

[2]Note that if user 1 is decoded first in the Gaussian MAC, this user should be encoded last in the BC.

Finally, a unified cross-layer analytical framework for BC and MAC channel is given in [38], where controlled queuing systems aim to maximize the throughput subject to delay guarantees.

1.2 Multi-antenna transmit processing

After the review of the information-theoretic bounds of the BC and the MAC, this section focuses on the practical alternatives and aims to summarize both classical designs and more recent strategies at the transmitter. Generally, this refers to the downlink communication where a BS is provided with multiple antennas (Q) whereas the K terminals have a single antenna. Due to the less stringent battery constraints, more complexity is usually allowed at the BS. Note that classical linear processing techniques such as ZF or MMSE [39] can be applied both at the transmitter or the receiver, whereas the successive interference cancellation schemes at the receiver (e.g. DFE) have also their counterparts at the transmitter side with the DPC strategies.

1.2.1 A practical review on precoding

This subsection is as a link between practical implementations for multi-antenna multi-user systems that are treated in following subsections, and the information-theoretic point of view that has been reviewed up to now. As an initial warning, [40] states that the design of precoding schemes is still in its infancy due to the difficulties in the implementation. To begin, a theoretical and practical study on the achievable throughput of a multi-antenna Gaussian broadcast channel is conducted in [20], which is commented next. In a $Q \times K$ multi-antenna broadcast channel, where the transmitter has Q antennas and the K receivers have a single antenna. In such a case, the signal model in (1.1) simplifies to

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{1.4}$$

where each row of matrix $\mathbf{H}$ contains the channel vector of a user. By applying a Gram-Schmidt orthogonalization of the rows of the channel matrix, the channel can be decomposed into $\mathbf{H} = \mathbf{R}\mathbf{Q}$, where $\mathbf{R}$ is a $K \times K$ lower-triangular matrix, and $\mathbf{Q}$ is a $K \times Q$ matrix with orthonormal rows. Then, user k would suffer only interference from previously encoded users, and ideally, if it is non-causally known at the transmitter, it can be completely removed without a rate penalty. Therefore, it is a successive interference cancellation at the transmitter side. This scheme is called Zero Forcing-Dirty Paper (ZF-DP), which achieves asymptotically optimal throughput for high SNR if the channel has full row rank [20].

Practical implementations of DPC techniques include Tomlinson-Harashima Precoding (THP), which is dual to a Decision Feedback Equalization (DFE) [41]. THP is no longer ideal and suffers from power, modulo, and shaping losses. While the shaping loss is due to the fact that Shannon's capacity formula for AWGN channel demands a Gaussian input distribution, the

power loss reflects that the transmitted signal might have more power than the intended signal, and it is significant for low constellations. The modulo loss comes from the modulo operation in THP, and it is more pronounced for small constellations. Whereas at high SNR the loss is dominated by the shaping loss (about 1.53 dB, which can be mostly recovered), at low SNR, the power and modulo losses are dominant (about 3-4 dB). To recover some of these losses, pre-subtraction at the transmitter is combined with trellis and convolutional codes [21]. This dissertation will though concentrate on the theoretical bounds for DPC, see Chapter 2.

Since perfect information cannot be always reasonable, schemes that employ only partial CSI are of interest. However, throughout the dissertation perfect channel knowledge is assumed, which yields the upper bound in performance but could sometimes be too optimistic. For instance, in a single-user MIMO channel, independently of the type of knowledge at the transmitter, the capacity scales like $\min(Q,K)\log\mathrm{SNR}$ at high SNR, whereas if perfect CSI is not available at the receivers, this capacity scaling is $\min(Q,K)\left(1+\frac{\min(Q,K)\log\mathrm{SNR}}{T_c}\right)$, where T_c is the coherence time of the channel [42]. To end this subsection, some interesting results of [43] include that the capacity of a single-user channel with Q transmit antennas in an *unbiased* channel is the same as the sum rate capacity of a Q-user channel with a single antenna per user. When channel correlation is intense, a multi-user system is inherently superior to a single-user system because of the multi-user diversity [43]. The term *multi-user diversity* refers to the fact that independent channels can be obtained by a proper selection of the users that are scheduled, see Chapter 3. In other words, the scheduler at the AP benefits from the fact that channels from the users vary independently, so that it can select the best one at every time instant. In fact, channel fading in multi-user communications is a source of randomization that shall be exploited rather than a drawback, see Section 1.3.1 for further details.

1.2.2 On optimal transmit beamforming

Up to this point, the previous precoding schemes are mainly devoted to increase the system rate. In the literature, practical schemes such as those in this section concentrate on the diversity advantage, that means, the increase of the effective SNR at the receivers. Moreover, they are not based on information-theoretic issues, but rather strategies that could be currently implemented. Otherwise stated, the general signal model that is valid for this section is the same as (1.4). However, the transmitted signal $\mathbf{x}$ has been substituted by a beamforming matrix $\mathbf{B}$ multiplied by the (generally QAM) vector of transmitted symbols $\mathbf{s}$ according to

$$\mathbf{y} = \mathbf{HBs} + \mathbf{w}, \tag{1.5}$$

where it is usually assumed that the symbols for each user are different, $s_i \neq s_j, \forall i \neq j$, and the beamformers $\mathbf{b}_k$ for the K users are gathered at the columns of $\mathbf{B}$. Note that the kth row of matrix $\mathbf{H}$ contains the transpose of the channel vector for the kth user $\mathbf{h}_k^T$, with covariance

matrix $\mathbf{R}_k$, and the usual assumptions hold for the noise. The general SINR for the kth user γ_k in this multi-user MISO system can be expressed as

$$\gamma_k = \frac{\mathbf{b}_k^H \mathbf{R}_k \mathbf{b}_k}{\sum_{i \neq k} \mathbf{b}_i^H \mathbf{R}_k \mathbf{b}_i + \sigma_k^2},$$

where the noise power σ_k^2 is different for each user in general. The optimal beamforming and power control minimizes the transmitted power subject to SINR requirements. Power control in this case refers to the fact that a certain SINR requirement γ^t shall be fulfilled, that is

$$\begin{aligned} &\min_{\mathbf{b}_k} \sum_{k=1}^{K} \mathbf{b}_k^H \mathbf{b}_k \\ &s.t. \frac{\mathbf{b}_k^H \mathbf{R}_k \mathbf{b}_k}{\sum_{i \neq k} \mathbf{b}_i^H \mathbf{R}_k \mathbf{b}_i + \sigma_k^2} \geq \gamma^t, \ k = 1, \ldots, K, \end{aligned} \tag{1.6}$$

where if the transmit power p_k is added to the problem with unitary beamvectors, it can finally be expressed in matrix form as

$$\begin{aligned} &\min \sum_{k=1}^{K} p_k \\ &s.t. \left(\mathbf{I} - \mathbf{F}\right) \mathbf{p} \geq \mathbf{u}, \end{aligned} \tag{1.7}$$

where the matrices involved are

$$[\mathbf{F}]_{i,j} = \begin{cases} \gamma^t \frac{\mathbf{b}_i^H \mathbf{R}_j \mathbf{b}_i}{\mathbf{b}_i^H \mathbf{R}_i \mathbf{b}_i}, & \text{if } i \neq j, \\ 0, & \text{if } i = j, \end{cases}$$

and the kth element of $\mathbf{p}$ is p_k. The kth position of vector $\mathbf{u}$ is

$$[\mathbf{u}]_k = \frac{\gamma^t \sigma_k^2}{\mathbf{b}_k^H \mathbf{R}_k \mathbf{b}_k}.$$

This problem in (1.7) is feasible if, and only if, the greatest eigenvalue of $\mathbf{F}$ is strictly smaller than 1 [44]. It is shown in [26] that the solution to the downlink problem and that of

$$\begin{aligned} &\min_{\mathbf{a}_k, \rho_k} \sum_{k=1}^{K} \rho_k \\ &s.t. \frac{\rho_k \mathbf{a}_k^H \mathbf{R}_k \mathbf{a}_k}{\mathbf{a}_k^H \left(\sum_{i \neq k} \rho_i \mathbf{R}_k + \mathbf{I} \right) \mathbf{a}_k} \geq \gamma^t, \ k = 1, \ldots, K, \end{aligned} \tag{1.8}$$

are equivalent for some power levels p_k and ρ_k. Therefore, the solution to (1.6) can be computed in an iterative fashion using the *virtual uplink problem* in (1.8), which is shown to have a unique solution [45]. For further details, the reader is referred to [46], but it is important to note that the duality between the uplink and downlink schemes appears again as in information-theoretic

models. Interestingly, it is shown in [27] that under a proper scaling, the algorithm in [45] converges to the globally optimum downlink beamforming weights for the problem in (1.6). The idea is to normalize the channel vector by the standard deviation of the noise according to

$$\widetilde{\mathbf{h}}_k = \frac{\mathbf{h}_k}{\sigma_k},$$

and then apply the iterative algorithm in [27]. Note that all the previous optimum beamforming algorithms assume that a solution exists, otherwise they might diverge.

Facing the problem from another perspective, in [47] the optimization conducted is the *equalization* of the Signal to Interference and Noise Ratio (SINR) for the users, in other words, the BS aims to maximize the minimum SINR subject to a power constraint of P_T according to

$$\begin{aligned} &\max_{\mathbf{b}_k} \min_k \gamma_k, \\ &\sum_{k=1}^{K} p_k \leq P_T, \end{aligned} \tag{1.9}$$

which yields a set of coupled problems without a closed-form solution. This optimization is chosen because the user with lowest SINR determines the overall performance of the system, since it is the one where more effort shall be put to accomplish the SINR requirement. The previous optimizations become simpler by the use of a ZF transmit beamforming, see next subsection, and allow the thorough study of fairness issues, see Chapter 3 for further details. In [47], the authors deal with the problem in absence and in presence of noise. If the ZF conditions for complete interference nulling cannot be achieved, the problem is decomposed into two steps, a power assignment and a beamvector optimization. Related to this problem, in [17] the same type of optimization is addressed. However, the authors impose certain QoS requirements, i.e. the SINR shall be maintained over a certain threshold γ^t

$$\gamma_k \geq \gamma^t \Rightarrow \frac{\gamma_k}{\gamma^t} = \gamma'_k \geq 1, \tag{1.10}$$

so that the problem in (1.9) can be rewritten by substituting γ_k with γ'_k. By means of this change, it can be verified whether the problem is feasible or not. Related to these issues, it is shown in [48] that with slight imperfections in the CSI, the degradation might allow only a few users to obtain an acceptable link quality. Therefore, admission control mechanisms such as those implicit in (1.10) are needed to select the users. Note also that admission control is commonly performed by the scheduler at the AP, because it is essential to obtain the best subset of users to achieve a good performance. Therefore, it is an important topic throughout the dissertation, see for instance Chapter 3 and Chapter 4.

Referring to a power minimization problem subject to SINR constraints, which is analogous to (1.9), in [17] the duality between the uplink and downlink solutions is formalized, so that the more

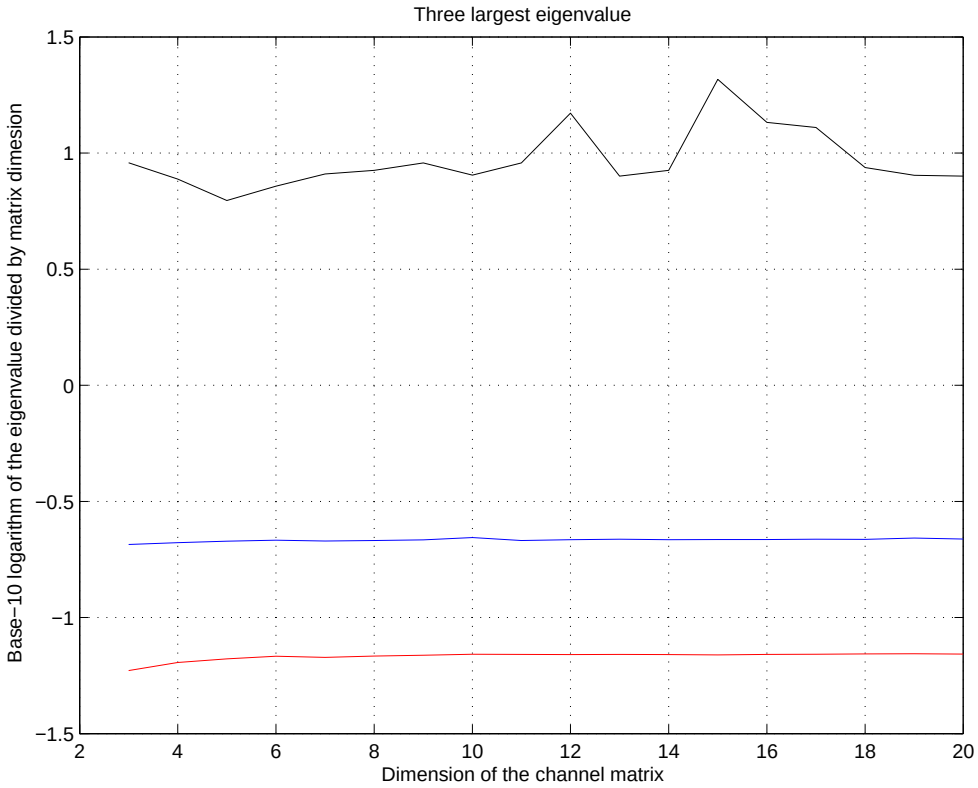


Figure 1.6: Three largest eigenvalues of the matrix $(\mathbf{H}\mathbf{H}^H)^{-1}$ averaged over 5000 trials, when $Q = K$. The largest eigenvalue has a erratic plot because it has mean infinite.

complicated downlink problem can be solved efficiently by the dual uplink problem. This concept is logically linked to the MAC-BC duality. Feasibility conditions are extended in [49], where it is remarked that traditional downlink beamforming algorithms might diverge if SINR requirements are too demanding, see also [50]. In this sense, continuing with [51] and [52], in [53] the SINR achievable regions according to a certain SNR requirements for a two-user downlink beamforming problem are shown. Again, concepts developed in the information-theoretic models are shown in a similar way at the signal processing part. However, this type of iterative solutions could also be seen by an scheduler as too complex, thus techniques providing a closed-form expression as the following are relevant. Moreover, because of the fairness implications studied in Chapter 2, the ZF methods presented next seem to be well-suited for multi-user communications.

1.2.3 Zero Forcing techniques and related issues

When compared to more complicated DPC, or with the previous schemes for joint beamforming and power control, the traditional Zero Forcing (ZF) seems to be an adequate technique since it might offer a significant fraction of the sum capacity of the broadcast channel, especially when the number of users exceeds the number of antennas and a well-suited user selection mechanism

is employed for the rate maximization [54]. Note that ZF (or channel inversion) techniques are unpopular in the uplink because of noise enhancement, whereas in the downlink it might cause wild variations in the transmitted or received power. However, with a low-complex closed-form solution, ZF achieves a large fraction of the sum capacity, thus it has been proposed both for single-user and multi-user systems [55]. Essentially, the ZF for the model in (1.5) is

$$\mathbf{B} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}, \tag{1.11}$$

but the matrix $\mathbf{H}$ shall be formed by the best users to maximize a certain criterion, e.g. sum capacity in [54]. The algorithm the authors propose attains an acceptable fraction of the sum capacity of the broadcast channel with a significant reduction in the computational burden. Moreover, with ZF it is more simple to perform several power allocation techniques because the channels become parallel and orthogonal, thus users see no inner-cell interference. That means that each user sees a flat-fading channel corrupted only by AWGN and not by the interference signals for the other simultaneously-transmitting users. This leads also to a better understanding of fairness issues, and to the development of higher layer mechanisms with less complexity. Note that the diversity order of the ZF transmitter is $Q-K+1$ [56], which is the same as for MMSE. Moreover, ZF is equivalent to MMSE not only in the high SNR regime, but also when a low number of users is served [57]. For all these reasons, ZF seems an adequate criterion and thus has been chosen as the transmit beamforming technique throughout the dissertation, although the results and conclusions have a straightforward extension to DPC ideal schemes.

Related to the MMSE schemes, according to [58] the performance of ZF might be improved by a regularization of the inverse in (1.11), which can be expressed as

$$\mathbf{B} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H + K\sigma^2\mathbf{I})^{-1},$$

where it is assumed that the noise power is equal for all the users. This modification helps to improve the bad behavior of the eigenvalues of the matrix $(\mathbf{H}\mathbf{H}^H)^{-1}$ when $\mathbf{H}$ is square. Figure 1.6 shows the mean of the three largest eigenvalues of $(\mathbf{H}\mathbf{H}^H)^{-1}$ when $K=Q$ for an increasing size of the matrix. The maximum eigenvalue has an erratic plot due to its infinite mean [58], whereas the rest of the eigenvalues have a much better behavior. Therefore, ZF might not perform well when the number of users is equal to the number of antennas, and any regularization scheme such as (1.12) should reduce the effects of the maximum eigenvalue. In fact, the sum rate for $K=Q$ users is constant with K as it tends to infinity, and with the regularization the growth is linear. A drawback is that the channels are no longer orthogonal, instead there is some inner-cell interference, which makes it more difficult to study higher-layer issues such as fairness.

1.2.4 Comparisons

Up to this point, several techniques have been described. However, from Chapter 3 until the end of this dissertation, Zero Forcing is the selected scheme because of several reasons. Some of them are stated in Chapter 2, and some others will be outlined in this section, where the comparisons will be addressed. In [14] it is stated that non-DPC multi-user transmission schemes for the downlink are of practical relevance, such as transmit beamforming. A first question that shall then be addressed is: what advantages does beamforming provide vs. precoding?

Basically, the powerful characteristics of beamforming over the optimum DPC include

1. that it greatly simplifies the vector BC by limiting the rank of the covariance matrices to unity instead of the generally full-rank matrices within (1.2).

2. In the uplink it is sum-capacity achieving with a large number of users in the cell [59].

3. Beamforming and DPC have the same sum rate scaling when Q is fixed and K goes to infinity [60]. Indeed, beamforming is equivalent to DPC both at high and at low SNR [19].

Available techniques for the beamforming include ZF and MMSE, which are equivalent as the SNR increases, although the performance of MMSE is certainly better at low SNR [61]. A clear advantage of ZF over MMSE is that the equivalent channels that are created are parallel and orthogonal, which allows a better study of mechanisms such as the power allocation, as well as higher-layer tasks as the scheduling; MMSE destroys this orthogonality. An ideal implementation of precoders for the BC, the Zero Forcing-Dirty Paper (ZF-DP) scheme explained after (1.4) is relevant. However, these theoretical bounds might be reduced due to the implementation constraints that have already been explained [62]. For instance, it is stated in [63] that practical schemes for THP might have low performance at low SNR due to the loss caused by the modulo operation. According to [35], ZF is a competitive alternative to DPC if implementation complexity is considered with four transmit antennas and one receive antenna. Moreover, ZF-DP reduces to MRC beamforming to the best user for low SNR [20]. For the selected ZF strategy, it shall be noted that it yields the same optimal throughput slope at high SNR, but pays a fairly high throughput loss with respect to the previous ZF-DP. However, when the number of users exceeds the number of antennas, the gap with respect the sum capacity can be negligible [54].

Interestingly enough, in [64] the authors show that under certain conditions in a vector broadcast channel with K users and the BS equipped with 2 transmit antennas, the number of users that can be simultaneously served can be higher than 2. Particularly, if the channel vector norms and angles are such that three users cover more than 90° in space, then it is optimal to serve those three users in the high power regime to fully exploit the spatial channel. The power allocated to the kth user is no longer a water-filling procedure, it is rather found using the KKT conditions for sum rate maximization. Although simulations in [65] show that

typically, the number of active users is four times the number of BS antennas in the high SNR regime with the optimum covariance matrices, restricting the number of transmitting users to the number of antennas with rank-one covariance matrices might only loose a small fraction of the capacity. However, this is a rather theoretic point of view, since with the selected ZF beamforming criterion, only as many users as antennas might be served.

In Chapter 2, ZF is compared to DPC and to the cooperative bound in terms of fairness, which completes the comparisons in this section. This is due to the observed fact that the mean (or sum) throughput loss might not reflect accurately the behavior of the techniques in a multi-user scenario. In this case, the performance variation among users is a key measure. For more details, see Section 1.5. However, for a fair comparison among the techniques, it is worth mentioning a fundamental trade-off in multi-antenna and multi-user channels, the diversity and multiplexing trade-off, see [66] and [67]. It will be shown in a practical situation in Chapter 3.

Diversity vs. multiplexing trade-off

Traditionally, multiple antennas are used to increase the **diversity**, i.e. to combat fading more efficiently. For instance, in an scenario with K receive antennas and a single transmit antenna, the average probability of error might decay theoretically like $1/\mathrm{SNR}^K$ at high SNR. If Q transmit antennas are added to the system, the maximum achievable *diversity gain* is KQ, assuming that the channels between each pair of antennas are i.i.d. Rayleigh faded. Besides diversity, another issue is that if the channel matrix is well-conditioned, **multiple parallel spatial channels** can be created, so that several data streams can be transmitted simultaneously. This effect is called *spatial multiplexing* r, and this gain is bounded by $\min(Q, K)$ at high SNR, where the system is limited by the degrees of freedom and not by the power. More in detail, a scheme is said to have a spatial multiplexing gain r and a diversity advantage of d if the rate of the scheme scales like $r \log \mathrm{SNR}$ and the average error probability decays like $1/\mathrm{SNR}^d$. Formally, the diversity gain is

$$\lim_{\mathrm{SNR}\to\infty} \frac{\log P_e(\mathrm{SNR})}{\log \mathrm{SNR}} = -d,$$

where $P_e(\mathrm{SNR})$ is the average error probability of the system, and the multiplexing gain is

$$\lim_{\mathrm{SNR}\to\infty} \frac{R(\mathrm{SNR})}{\log \mathrm{SNR}} = r,$$

where $R(\mathrm{SNR})$ is the number of bits per symbol of the scheme (the rate). When the block length is higher than $K + Q - 1$, the optimal diversity gain $d^*(r)$ achievable by any coding scheme of block length L and multiplexing gain r (integer) is precisely given by the following equation [66]

$$d^*(r) = (Q - r)(K - r), \tag{1.12}$$

which reflects the trade-off between the error probability and the data rate of a system. The

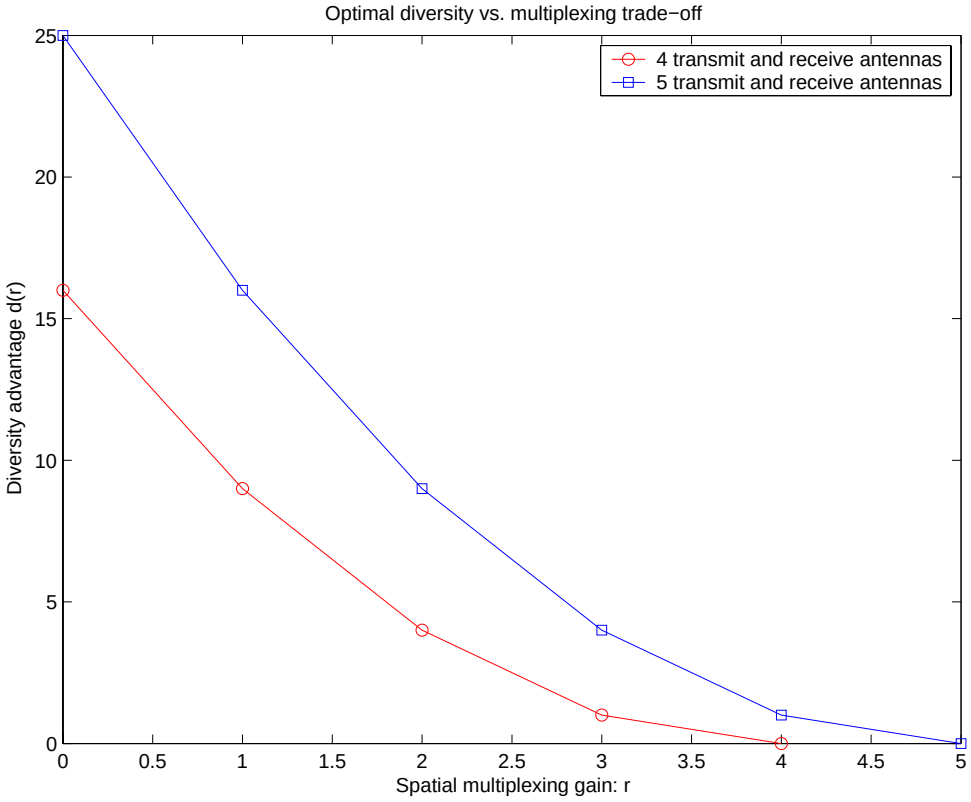


Figure 1.7: Optimal diversity vs. multiplexing trade-off in a single-user multi-antenna system.

optimal operating points of (1.12) are depicted in Figure 1.7, where the diversity gain is plotted vs. the spatial multiplexing gain. The curves plotted refer to the cases with 4 and 5 transmit and receive antennas, so that any practical strategy is contained in the region between them and the axes. In the original paper [66] the reader might find the trade-off curves for other techniques such as Alamouti or BLAST-based schemes. This is explained here not only because in a sense it reflects the existing trade-offs at the PHY, but also because it has been extended to the multiple access channel in [67]. The underlying ideas are the same, but with multiple users, there exists also the multiple access gain, meaning that several signals from different users can be spatially separated. The trade-off curves are given for successive cancellation and rate splitting, showing that there exists a significant gap between them and the joint decoding. In fact, the resource allocation is shown to have an impact on both the diversity advantage and the multiplexing gain [67]. Recently, [68] introduces the notion of a diversity gain region for a multi-user channel, which specifies the set of diversity gain vectors that are simultaneously achievable by all users. The diversity-multiplexing trade-off is closely related to the results in [69] and Chapter 3, where it is stated that the power allocation techniques providing a better diversity gain imply that they serve less users simultaneously (worse multiplexing gain).

To conclude this part, it seems clear that the variety of the techniques and optimization

criteria make it difficult for the AP to select the best scheme according to the needs of the system. Besides the trade-off between performance and complexity, there exist others such as that between performance and signaling, see Chapter 5, which might even increase the difficulty of the choices at the AP, as it will be shown throughout the dissertation. In any case, the selected strategy might be a balance of the goals of the AP. From a complete system point of view, the implemented technique might not necessarily be the optimum PHY technique.

1.2.5 Extensions to MIMO

Before going into the details of the scheduling procedures, related literature concerning the extension to multiple antennas at both sides of the communication link shall be roughly discussed. The author notes here that the extension is not straightforward, especially concerning the scheduling mechanisms that are presented in the next section.

In the uplink of a multi-user MIMO wireless system, [70] proposes an iterative algorithm to find the transmit precoder and the receive decoder in order to minimize the total MSE at all the receivers simultaneously, which differs from the single link MIMO. On the other hand, [71] proposes to maximize a lower bound for the product of SINR of the users, which yields a closed-form expression for the antenna weights at all users. A whole dissertation is devoted to the MIMO beamforming design and power allocation [72]. Particularly, the multi-user MIMO beamforming optimization is a highly non-linear problem, thus quasi-optimum solutions such as simulated annealing are motivated. However, they are still too complex for a realistic implementation.

In [73], two suboptimal solutions are developed to lower the computational complexity of the downlink beamforming problem when there are multiple antennas at both sides of the communication link, namely the block diagonalization (especially suited at high SNR and for maximum sum capacity) and a successive optimization scheme (for power minimization, especially in the low SNR regime). The generalization comes with [74], where a balance of the trade-off between performance and complexity defines the design of the transceiver structures. Concerning power control and beamforming, [75] concludes that significant power savings can be obtained by adapting instantaneously the transmitted power. Moreover, in a linear antenna array, inter-element spacings greater than a quarter of a wavelength are sufficient to achieve close to the minimum average transmit power. In [76], a minimum transmit power subject to SINR constraints is developed borrowing some of the ideas of optimum transmit beamforming. Finally, two options for multi-user precoders for fixed receivers are proposed in [77].

1.3 Multi-user scheduling

Whereas the previous sections deal with the optimization of the transmitter, a different approach is taken here. Instead of the input covariance matrices, the precoding design, or the beamforming

technique, this section is devoted to the choice of the subset of users that shall be served. Up to this point, the concern has been the design of the covariance matrices in (1.2). In this section, the author concentrates on the selection of the permutation of the users within (1.2). Indeed, the performance of the system might be severely degraded depending on the subset of users to which the AP transmits simultaneously. Scheduling in a general case refers to the resource distribution among a set of terminals/users. Particularly in this dissertation, the term scheduling refers to the fact that the users in the cell shall be selected (or divided into groups) so as to be served simultaneously by a given SDMA scheme to achieve the best performance. Moreover, usually the instantaneous output power shall be efficiently shared among users, see Chapter 3.

Therefore, the author tries to summarize the most relevant approaches to the scheduling of multiple users in a system with multiple antennas. As stated, this issue is usually not taken into account when optimizing only physical layer procedures as it has been done up to this point, but it increases the relevance when realistic system implementations are sought, and especially when the multiple dimensions of a communications system are to be exploited jointly. In Figure 1.8 the idea is plotted. Whereas traditionally users are not overlapping (left subplot), with the addition of the frequency and space axes, several users can be served simultaneously at a given slot (right subplot). Note that each user is located at a certain time-frequency-space point, but in order to fully exploit diversity, the remaining points in this space should be filled with the users (not done for the sake of understanding of the plot). However, since in a single antenna multi-user environment the capacity is maximized by transmitting to the best user, opportunistic communications shall be firstly addressed. Moreover, note that in a MIMO Gaussian MAC it is optimal to support only the best user in the low SNR regime [78].

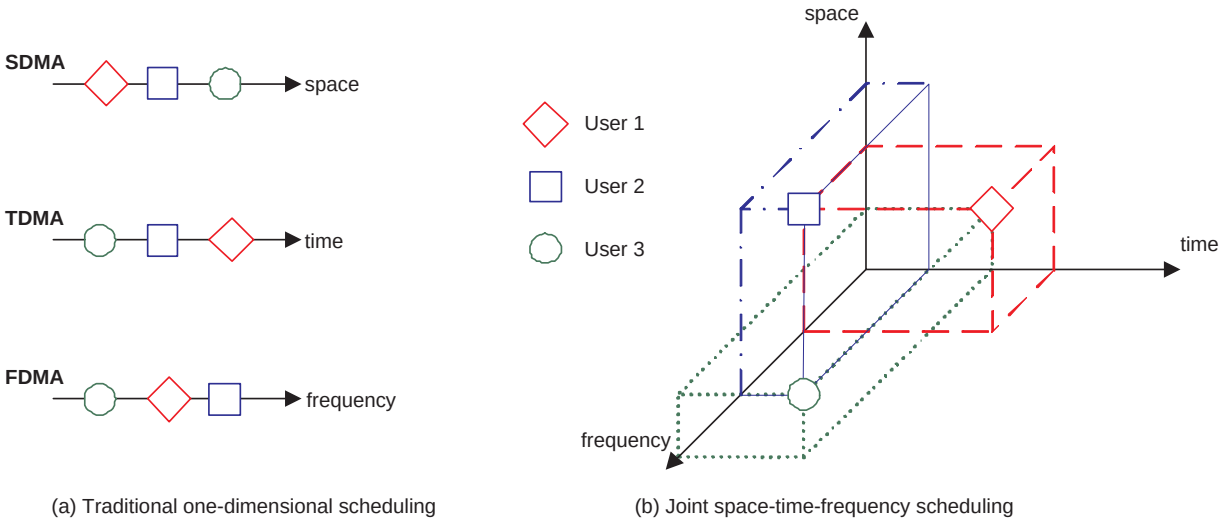
1.3.1 Opportunistic communications

In opportunistic communications, the basic idea is to benefit from multi-user diversity, that means, the best user should be scheduled at any time instant. Mathematically, the AP shall select the user k^* with best rate R_k at time instant t,

$$k^* = \max_k R_k(t), \tag{1.13}$$

which yields the highest throughput of the system. However, the users with worse channels might never be allocated for transmission, especially if there are substantial differences among channel gains. Therefore, instead of (1.13) the AP might schedule user k^* when the instantaneous channel is above its mean (or near its peak), which is called the Proportionally Fair scheme, i.e.

$$k^* = \max_k \frac{R_k(t)}{T_k(t)}, \tag{1.14}$$



(a) Traditional one-dimensional scheduling

(b) Joint space-time-frequency scheduling

Figure 1.8: (a) Traditional one-dimensional scheduling exploits only one diversity scheme, either time, frequency, or space, for each user in a non-overlapping way. (b) Three-dimensional scheduling, where three users share the slots that involve the three available dimensions.

where $T_k(t)$ is the average throughput that can be updated (only for the scheduled user) using an exponentially weighted low-pass filter, i.e.

$$T_{k^*}(t+1) = (1 - \frac{1}{t_c})T_{k^*}(t) + \frac{1}{t_c}R_{k^*}(t).$$

If t_c is much larger than the correlation time scale of the fading, the average throughput of the users $T_k(t)$ converges to the same quantity. Obviously, this type of mechanisms is well-suited whenever the quantity and quality of the feedback is limited, e.g. SNR feedback. This scheme has been developed e.g. in Qualcomm's High Data Rate (HDR) system.

With multiple antennas at the transmitter, these ideas have been applied in [79], where the multi-antenna AP provokes fluctuations in the channel gains, basically by multiplying the signals going out from the antennas by random amplitudes and phases, see Figure 1.9 for an example. This makes sense especially in situations where there is little scattering and/or mainly slow fading. Furthermore, the proposed scheme requires very limited channel feedback, and performs the additional task of an *opportunistic nulling* of the interference affecting adjacent cells. After [79], a number of papers have been published following this approach, see e.g. the multi-user MIMO system in [80], where thanks to an induced fading, it is more likely that some user is near the waterfilling configuration. Opportunistic schemes might be selected especially when there is limited feedback or a high number of users, otherwise they might be clearly outperformed by SDMA techniques, as it is demonstrated e.g. in Chapter 5. For instance, beamforming achieves a higher sum capacity than a TDMA using (1.14), i.e. the user with best channel is scheduled at any time [81]. Although the inefficiency of such a TDMA scheme is studied in [82], TDMA

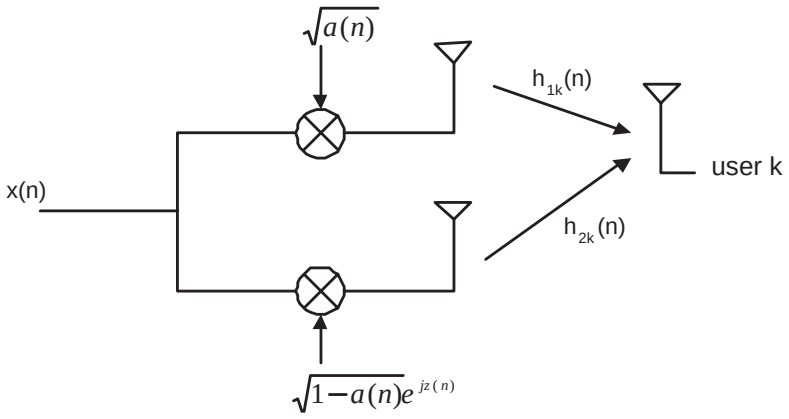


Figure 1.9: Sketch of the opportunistic beamforming: the same signal is transmitted over the two antennas with time-varying phase and powers.

converges to the maximum sum rate as the power decreases. Extending [19], it is proven in [60] that the gain of DPC over time-sharing is substantial in any case.

1.3.2 Scheduling in the beamforming domain

Independently of the multi-antenna technique that is chosen, the scheduler might try to select the best subset of users to achieve the optimum performance. As it has been shown, whenever the quality and quantity of the feedback is high, it is better to transmit simultaneously to several users rather than transmitting to the best one. This is related to the classification of the MIMO schedulers in [83], namely, i) the *spatially-greedy*, which assign all resources to a single user in an opportunistic way, and ii) the *spatially-spread*, where multiple users are allowed to transmit simultaneously. The benefits of the spatially-spread strategies is stated in a number of papers:

- The authors of [83] compare both schemes in terms of throughput and delay by taking into account that the user arrival is a random process, coming up to the conclusion that the spatially-spread approach leads to significantly lower delay than the spatially-greedy.

- Moreover, with a ZF transmit beamforming, it is shown [55] that there exists an optimal number of antennas that shall be used per user, or what is equivalent, there is an number of users that shall be scheduled for a fixed number of antennas. In other words, assigning transmit antennas independently (to several users) offers significant improvements over the allocation of all the antennas to a single user [84].

- The authors in [40] consider the downlink of a multi-user MIMO system, where the receivers have a single antenna, and also show that it might be beneficial to transmit to several users at a time even with incomplete CSI at the transmitter.

- Scheduling for maximum capacity in a MIMO system is studied in [85]. The authors first consider scheduling the transmissions of the users for fixed transmit beamformers, and then present a low complexity algorithm to approach the sum capacity. The capacity-achieving strategy is indeed a combinatorial problem of the choice of the users. An interesting point is that the authors investigate on how the transmit beamvectors shall be chosen depending on the degree of feedback.

A remark is needed before proceeding further: although the sum rate of the spatially-spread is higher than that of the spatially-greedy, the throughput per user is lower [83], which might also have fairness implications. These issues imply that if the CSI allows it, it is better to select a subset of users to optimize both the throughput and the delay. Therefore, the choice of the users that shall be served becomes extremely important in the design of a system, see especially Chapter 5 for a practical perspective on the combination of OFDM schemes with spatial diversity.

For the information-theoretic techniques, scheduling is a rather new research topic, which refers basically to the selection of the encoding order for DPC techniques in order to attain the optimum performance. For instance in [35], with a ZF-DP strategy, two schemes are compared to the First Come First Served (FCFS) case, the first one encodes users in increasing order of the power they would require to achieve their target data rate as if they were alone in the cell, whereas the second one uses the optimal encoding order and the optimal choice of the covariance matrices. Certainly, a significant gain might be achieved by using the optimum encoding order. The authors in [86] propose a greedy algorithm to order the users for a ZF-DP approach which is extremely close to the maximum sum capacity of the broadcast channel.

Another aspect covered by this subsection is the application of traditional DLC scheduling procedures to the multi-antenna physical layer. Through simulations, it is observed in [87] that the multi-user diversity gains might be reduced in presence of any kind of diversity, and spatial diversity in particular[3]. Due to the multi-user diversity, the performance of a Proportionally Fair (PF) scheduler is superior to that of the classical round robin. However, the superiority of the PF decreases with increasing transmit diversity order, due to the fact that the variance in the capacity decreases when the order of the transmit diversity increases. In the limit of high number of antennas and users, the maximum throughput achieved by any optimal scheduler in the presence of transmit diversity under SNR-only feedback (no full CSI at the transmitter) can be infinitely worse than that of a system with no diversity.

As a conclusion of this subsection, as many aspects as possible shall be taken into account in the design of any communications system so as not to increase complexity and worsen the performance. Indeed, the interaction between the PHY and the DLC shall be seen as a synergy rather than a drawback, especially when the number of dimensions increases, see the next section.

[3]Provided a perfect channel knowledge at the receiver but generally unknown at the transmitter.

1.3.3 Combination of schemes

Another interesting issue within the multi-user scheduling is the combination of several diversity dimensions, for which the literature is rather scarce. This might be so because of the exponential complexity of the problem, as it will be shown. In [88], the authors introduce a dynamic slot allocation technique for SDMA, where the basic idea is to construct *intelligent* space-time frames, basically by assuring a predetermined SINR for each user. Since the problem is NP-complete [89], heuristic suboptimal algorithms are motivated[4]. The only implementation constraint is that the channel shall remain invariant from the measurement time to the transmission of the TDMA/SDMA frame. The authors focus on the uplink, and assume a maximum SINR beamformer at the receiver (the BS). Four slot allocation algorithms are compared in increasing degree of complexity, and the key issue is the compatibility measure between users i and j as

$$\mathrm{Cmp}_{i,j} = \min\left(\frac{\|\mathbf{h}_i\|^2}{\|\mathbf{h}_j\|^2}, \frac{\|\mathbf{h}_j\|^2}{\|\mathbf{h}_i\|^2}\right),$$

where $\mathbf{h}_j$ is the channel vector of the jth user. With it, the authors propose an algorithm that tries to place users with similar received powers at the same slot. However, this metric does not take into account that the users might have close channel vectors, thus it does not perform well. Indeed, more power is needed to serve simultaneously two users that have similar channel vector, i.e. they come from the same zone of space if a signal model based on the DOA is assumed. To overcome this problem, in [88] a best fit strategy is proposed, which allocates the user in the time slot where the minimum SINR is the largest. Certainly, a pre-ordering of the users improves the overall performance. An extension to a polling protocol with SDMA is developed in [90]. First, the BS polls the terminals, and after their spatial signatures are obtained, the BS constructs the SDMA/TDMA frames, e.g. according to previously exposed techniques, see Figure 1.10 for an explanation. In [91] the authors take the best fit strategy proposed in [88] and extend their greedy algorithm to take into account not only the spatial characteristics of the users, but also several QoS parameters such as the packet timeout and the packet loss rate, since the traffic characteristics for the terminals might not be the same. A problem in this type of algorithms is that due to the NP-completeness, it is rather difficult to show how far from optimality they are.

In [92], the focus is set on the impact of smart antennas at the second layer of the protocol stack when the transceivers have limited resources. In particular, there is a finite set of beamforming vectors. This issue is significant since it might not be possible to perform adaptive beamforming if real-time implementation and backwards compatibility is sought. The authors combine the space dimension with OFDM, proposing two heuristic algorithms to allocate channels to users, adjust beamforming vectors, and assign users and channels in beams, with the

[4] An NP-complete combinatorial problem is that belonging to a class that cannot be solved in polynomial time, in other words, the complexity increases exponentially with the number of variables.

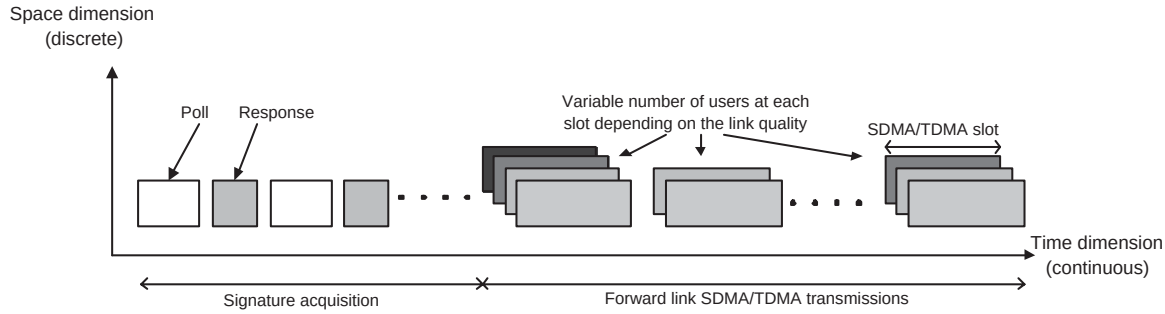


Figure 1.10: Typical SDMA/TDMA frame. First, the AP performs polling in order to acquire the signature of the terminals, then it constructs the SDMA/TDMA slots.

final objective of maximizing the throughput and provide QoS to users. Interestingly enough, the authors introduce the concept of *spatial separability* on a per subcarrier basis. Two or more users are *spatially separable* if there exist a beamforming vector for each user, such that the minimum SIR requirements are satisfied. However, this concept is rather different from the separability strategy based on the normalized scalar product that will be proposed in Chapter 5. The authors identify the trade-off between the number of assigned users in the same subcarrier and spatial separability, which in the end affects the achievable rate. The main idea behind limited transceiver resources is that the maximum SIR beamvectors are combined in order to reduce complexity. With unlimited transceiver resources, the basic ideas are contained in [93].

Within the context of a combination of frequency diversity (OFDM) and space diversity (SDMA), the authors of [94] investigate the possible options in the uplink regarding multi-antenna techniques, ranging from a MMSE solution, passing through the successive interference cancellation procedures, to the maximum likelihood receiver. However, their approach is concerned about the complexity and the multi-antenna implementation, which differs from the point of view of Chapter 5. From a rather different perspective, clustering of users into groups is treated in [95] to alleviate the inaccuracies of the DOA estimation due to the angular spread in wireless environments. The basic idea is to group the users according to power (and angular) classes before the maximum SINR downlink beamforming. Then, there will be a single beamformer for each group, which shall be used in a different time slot for each user in that group, with the consequent reduction in complexity.

To end with this part, there is a number of papers of the authors of [96], which refer to the uplink of an scenario where the users have a single antenna, whereas the AP is provided with multiple antennas. In this paper in particular, the general framework is based on utility functions based on the capacity, and the authors propose two fast practical scheduling techniques, namely a heuristic algorithm and a technique that relies on genetic algorithms. Although the framework the authors want to show is formal, the final resolution for the problem is heuristic. Due to

practical implementation constraints, the authors also perform a ZF linear processing.

1.3.4 Brief comments on DLC aspects

Before dealing with more physical aspects such as the power and bit allocation, a brief explanation on the issues concerning the second layer of the protocol stack shall be given. Indeed, the dissertation focuses on how the scheduling is made at the multi-antenna physical layer, but some assumptions are made in order to make the problems tractable. For instance, it is assumed that there is perfect channel knowledge available at the transmitter. This might be quite close to reality in a TDMA/TDD system, although the feedback channel plays a key role in order to acquire a better CSI. Moreover, the terminals have always data to transmit, which allows the AP to schedule instantaneously the active users. With these considerations, the terminals are assumed to have an infinite buffer to deal with packets that have no delay constraints.

The considered system is a wireless network, for which the definition of fairness might be ambiguous [97]. One of the key aspects is the selection of the granularity of fairness, since several options are available from an instantaneous point of view to a long-term basis. This choice is affected by the variability of the channel and the traffic conditions, and the system performance is dramatically affected by such issues. Moreover, further considerations might be critical in wireless, such as that between effort fair and outcome fair [98]. The system performance might differ significantly if fairness is guaranteed at the effort, e.g. in terms of number of time slots, or if it is evaluated at the output, e.g. according to a throughput metric. Related to these issues, the idealized (wired) scheduling procedure that serves as a first step for the wireless mechanism is the Generalized Processor Sharing (GPS), which serves each user k with rate R_k such that

$$R_k = C_{TOT} \frac{\phi_k}{\sum_j \phi_j},$$

where C_{TOT} is the total link capacity. Although GPS is only a theoretical bound with which schedulers dealing with packet transmissions compare performances, ideally it provides delay guarantees as well as equal normalized service to all sessions [99]. Nevertheless, the scheduling strategies proposed in this dissertation are concerned about the instantaneous choice of the users in order to fulfill certain requirements, which in the end is related to the choice of the ϕ_k. However, it shall be stated that the objective is not to design a full scheduler for the DLC layer.

1.4 Power and bit allocation

In this section, the author concentrates on resource management issues, basically the traditional power allocation and a recent game-theoretic framework, see Chapter 3, together with the available strategies in bit allocation (extended to the multi-antenna scenario in Chapter 4).

1.4.1 Power allocation

In fact, some of the power allocation criteria have already been introduced in Section 1.1 from an information-theoretic point of view, i.e. usually the objective is to maximize the sum rate of a MIMO channel. Assume a set of K parallel and independent channels h_k for the users, for which the sum rate optimization subject to a power constraint of P_T, that is,

$$\max_{p_k} \sum_{k=0}^{K-1} \log\left(1 + \frac{p_k |h_k|^2}{\sigma_k^2}\right)$$
$$s.t. \sum_{k=0}^{K-1} p_k \leq P_T,$$

yields the well-known water-filling procedure for the power allocated to the kth channel,

$$p_k = \left(\mu^{-1} - \frac{\sigma_k^2}{|h_k|^2}\right)^+, \tag{1.15}$$

where σ_k^2 denotes the noise power at the k receiver, and μ^{-1} is such that the power constraint is fulfilled with equality. This kind of optimization is interesting from the moment that it yields the maximum sum rate of the multi-user channel, however, the differences among the users might be significant because the worst user is penalized for the sake of the total achievable rate.

Under the umbrella of a convex optimization framework, the author of [100] develops a number of power allocation algorithms for several techniques trying to optimize different quality measures such as the MSE, the SNR, the BER, or the rate. This dissertation covers mostly single-user MIMO schemes, with particular emphasis on OFDM systems. A related paper is [101], where several power allocation strategies are compared within a single-user OFDM-MIMO system. The strategies aim to maximize the harmonic SINR or to maximize the minimum SINR among the subchannels. The authors also analyze the techniques based on the SINR asymptotically for infinite power. In both works, the power allocation is explicit according to a given cost function.

Other power allocation criteria are implicit in Section 1.2.2. Indeed, there the optimal beamformers (and consequently the power associated to each of them) are obtained with the objective of minimizing the transmitted power or the fulfillment of a QoS constraint for the users. Besides, other works that perform a power allocation will be commented afterwards in Section 1.4.3, where the bit allocation is described. Certainly, as it will be stated, bit loading mechanisms yield an implicit power allocation depending on the objectives at the AP.

1.4.2 Game-theoretic power control for CDMA

Game-theoretic power control was in vogue in the late nineties and in the beginning of this new century, not only because of its benefits in implementing distributed algorithms for computing the power allocation, but also for its appealing mathematical framework. Indeed, it provides a

powerful pragmatic solution that can be calculated independently at the terminals depending on the degree of available information. However, as it will be shown in Chapter 3, there are some points that remain unclear. A remark is needed before proceeding: the objective is not to fully describe Game Theory (GT), but only the specific issues used in the game-theoretic power control, which seems to have exploded some years ago, but currently the number of publications has decreased substantially. For the initial steps on GT, please refer to [102], which gives the basic concepts, as well as a number of examples (mostly based on Economics but quite graphic). For more details, [12] can be considered as the Bible of games.

Without loss of generality, consider that the system under study is CDMA. Following the notation of [103], the SINR for the k user can be expressed as

$$\gamma_k = \frac{W}{R}\frac{|h_k|^2 p_k}{\sum_{i\neq k}|h_i|^2 p_i + \sigma^2},$$

where W is the available spectrum bandwidth, R is the fixed transmission rate, σ^2 the AWGN power, which is assumed to be the same for al the receivers, h_k is the complex channel gain from the terminal k to the BS, and p_k is the transmitted power. This notation assumes conventional matched filter receivers and pseudo-random signature sequences. One of the main issues in game-theoretic power control is the definition of a convenient utility function g_k. In a number of papers, after some curious elaborated rationale, the widespread selected strategy is

$$g_k(p_k, \mathbf{p}_{-k}) = \frac{L_i R(1 - 2\mathrm{BER}(\gamma_k))^L}{L p_k}, \tag{1.16}$$

where L_i is the number of information bits transmitted in a frame of L bits. The objective of this expression is to balance the trade-off between probability of correct reception of a packet and the transmitted power, but it is arbitrary not only because of the insertion of the constant 2 in the numerator of (1.16) for the sake of the fulfillment of some properties, but also because other options could also be valid and even better behaved. For instance, a weighted addition of the power and the probability of correct reception. Anyway, this utility function has appealing properties such as increasing utility with respect the SIR when the power is fixed, or decreasing utility with increasing power, when the SIR is kept invariable. Another interesting issue is that $g_k(p_k, \mathbf{p}_{-k})$ expresses the fact that the utility function depends not only on the individual choice of the transmitted power p_k, but also on the transmitted powers for the other users in the cell, which is denoted by $\mathbf{p}_{-k}$, and contains the power from all users but the kth.

From the basics of GT, the celebrated concept of the Nash Equilibrium (NE) shall be first described. Consider that a game is expressed as $G = [\mathcal{K}, \{P_k\}, \{g_k\}]$, where $\mathcal{K}$ is the set containing the players (or users in the cell), $\{P_k\}$ is the strategy space where the powers p_k are contained, and $\{g_k\}$ is the set of utility functions, which might differ among users. Then, [103] states that

Definition 1.1 *A power vector* $\mathbf{p} = (p_1, \ldots, p_k)$ *is a NE of the game* $G = [\mathcal{K}, \{P_k\}, \{g_k\}]$ *if, for every* $k \in \mathcal{K}$, $g_k(p_k, \mathbf{p}_{-k}) \geq g_k(p'_k, \mathbf{p}_{-k})$ *for all* $p'_k \in P_k$,

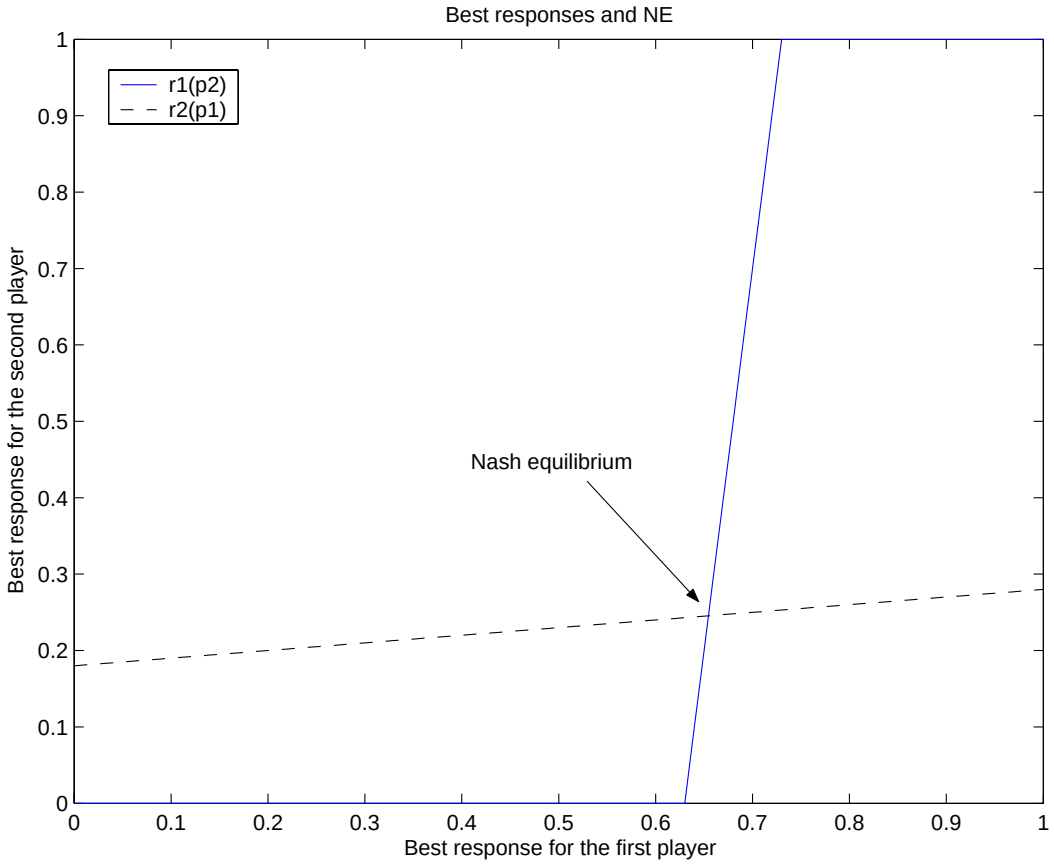


Figure 1.11: Best response functions and NE in a fictitious game with two players.

which means that at the NE, no player can improve its utility by making individual changes in its power. In other words, a NE is a point where no user can increase its own utility function by changing its own transmitted power, given the transmitted power from the other users. The basic assumption in GT is that the players are rational, so the power level is chosen by the best response. The best response for player k is a function that relates the powers from all the other players with its own power, i.e. $p_k = f(\mathbf{p}_{-k})$. The illustrative example for two users in Figure 1.11 clarifies how a NE is obtained. One shall obtain the best response for a given player, which is a function of the power from the other player, e.g. $r1(p2)$ is the best response of the first player given the power from the second player. Then, the NE is obtained at the intersection of the those functions $r1(p2)$ and $r2(p1)$, which might not be unique, and more importantly, the NE might not be Pareto optimal.

Definition 1.2 *A power vector* $\mathbf{p}'$ *Pareto dominates another vector* $\mathbf{p}$ *if,* $\forall k \in \mathcal{K}$, $u_k(\mathbf{p}') \geq u_k(\mathbf{p})$ *and for some* $j \in \mathcal{K}$, $u_j(\mathbf{p}') > u_j(\mathbf{p})$. *Furthermore,* $\mathbf{p}^*$ *is Pareto optimal if there exists no other power vector* $\mathbf{p}$ *such that* $u_k(\mathbf{p}) > u_k(\mathbf{p}^*), \forall k \in \mathcal{K}$ *and* $u_j(\mathbf{p}) > u_j(\mathbf{p}^*)$ *for some* $j \in \mathcal{K}$.

Again, it is further clarified with Figure 1.12, where the utility space for two users is plotted.

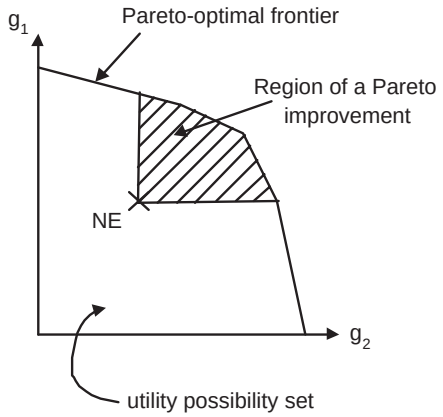


Figure 1.12: Utility space, Pareto efficient frontier, and Pareto region of improvement over a NE.

A Pareto-optimal point is such that no user can increase its own utility without decreasing the utility of some other user, i.e. the *Pareto-optimal frontier* in the figure. A NE as the one shown in the figure is not at all Pareto-efficient, thus a region of improvement becomes available, which is the marked area. In that region, the utility of any given user might be increased without decreasing that of any of the other users.

One of the first papers concerning game-theoretic power control was [104], and the authors of this paper and posterior works, e.g. see [103] or [105], aim to increase the Pareto-efficiency of the NE in the power control. The basic options encompass pricing mechanisms and repeated games. Although the mathematical framework underlying these papers is certainly appealing, the practical implication is that the utility effectively increases, the transmission power decreases, but surprisingly not shown in the papers, the BER also increases [106]. Therefore, care shall be taken when applying this kind of techniques in a real system, and a well-suited optimization criterion shall be selected to avoid undesirable consequences. For a thorough study particularized to the multi-antenna multi-user scenario, please refer to Chapter 3. The advantages of the solutions based on GT are the distributed computation of the transmitted power levels and the fairness provision, however, the necessary signaling might be quite high, since complete information is needed at the terminals. For a detailed study, see [107], and for a practical discussion about competition and cooperation using GT in ADSL, please refer to [108].

1.4.3 Bit allocation

Before explaining the schemes for bit allocation in spatial diversity systems, a background is needed on *traditional* bit allocation, which has been applied to multi-carrier systems, either in its wired version (DSL) or for wireless environments (OFDMA). In fact, a number of mechanisms have been developed since the seminal patents [109]. It is not the objective to give a throughout

review of all the published papers on multi-carrier bit allocation, rather to explain the basics.

Bit allocation problems are naturally derived in a power minimization (or bit rate maximization) under QoS constraints. Usually, the gap approximation is used, which essentially introduces a factor Γ_n to maintain the probability of error under a desired target. Then, the number of bits per symbol m_n that can be transmitted through subchannel index n is

$$m_n = \log_2\left(1 + \frac{p_n|h_n|^2}{\Gamma_n \sigma_n^2}\right),\ n = 1, \ldots, N,$$

where at subchannel n, p_n is the allocated power, h_n is the flat fading complex channel, and σ_n^2 is the noise power. The two main strategies in the literature are the following:

- **Bit rate maximization**: In this problem, the power is limited and the objective is to maximize the total delivered bit rate. Mathematically,

$$\max_{m_n} \sum_{n=0}^{N-1} m_n$$
$$s.t. \sum_{n=0}^{N-1} p_n \leq P_T;$$

- **Power minimization**: This scheme is very similar to the previous one, but now the goal is to minimize the total power, i.e. the sum of powers at the N subcarriers according to

$$\min_{m_n} \sum_{n=0}^{N-1} p_n(m_n)$$
$$s.t. \sum_{n=0}^{N-1} m_n = M,$$

  where M is the fixed number of bits per symbol that are transmitted in the whole band.

The optimal solutions to these problems lead to a non-integer bit allocation, but the system is in practice restricted to integer mappings, which are usually limited by a maximum number of bits per symbol per subcarrier. In [110] the essence of the single-user bit allocation algorithms is very well presented, namely the *bit removal* and *bit filling* procedures. The bit filling techniques are classical, and imply that a bit should be added to the subchannel where it is most efficient, meaning that a bit should be added into the subchannel that requires less power for that purpose, which are shown to yield the optimum allocation [111]. On the other hand, bit removal techniques assign the maximum number of bits at every subchannel and start decreasing the number of bits at the subcarriers where it is most efficient, which means in this case, at the subchannels where more power is saved. Note that in this case, the bit allocation determines the power needed at any given subcarrier. In fact, the rounding of the continuous water-filling solution is shown to

provide the optimal discrete solution, see [112] and references therein. In [113], the problem of the rate maximization subject to a power constraint is solved by means of lookup tables and a Lagrange multiplier bisection method. Very similarly, the power minimization can also be solved using this type of techniques since it is dual to the bit rate maximization. In [114], an optimal method based on the Lagrange multiplier is proposed to minimize the maximum BER under a bit rate constraint and with a limited power budget. Since the literature is quite extensive, for a recent review of bit loading algorithms for wired transmission, please refer to [115].

Previous algorithms are *classical* and involve single-user multi-carrier modulations. However, there has been an increasing interest to adapt them to the multi-user scenario. Perhaps the pioneering paper is [116], where the minimization of the total transmit power in a multi-user OFDM system is performed first by assigning each user a set of subcarriers, and then computing the number of bits $m_{k,n}$ and power $p_{k,n}$ for the user k at the tone n. It is important to note that only one user might use a given subcarrier because no space diversity is added, differently to the approach in Chapter 5. The focus are practical algorithms that can be implemented in real time. A new definition is added in order to distinguish which user has the nth subcarrier,

$$\rho_{k,n} = \begin{cases} 1, & \text{if } m_{k,n} \neq 0, \\ 0, & \text{if } m_{k,n} = 0, \end{cases}$$

and assuming that $f_n(m_{k,n})$ denotes the required received power to maintain a desired BER for user k and subcarrier n, as well as a Lagrangian relaxation of the integer variables $m_{k,n}$ and $\rho_{k,n}$, which are now real, the problem can be finally expressed as

$$\begin{aligned} \min_{m_{k,n},\rho_{k,n}} \quad & \sum_{n=0}^{N-1} \sum_{k=1}^{K} \frac{\rho_{k,n}}{|h_{k,n}|^2} f_n(m_{k,n}) \\ s.t. \quad & \sum_{n=0}^{N-1} \rho_{k,n} m_{k,n} = M, \ \forall k, \\ & \sum_{k=1}^{K} \rho_{k,n} = 1, \ \forall n. \end{aligned} \tag{1.17}$$

The authors in [116] solve (1.17) by using standard optimization techniques. Recently, there has been a renovated interest in these issues: in [117], two heuristic approaches are presented that overcome the implementation complexity of [116]. The dissertation [118] contains a good review of previously proposed alternatives. In [119], simplicity, fairness and efficiency are sought, so that after applying the best allocation, the authors propose two bit swapping techniques, namely horizontal and vertical. Horizontal bit swapping tries to smooth the bit distribution among the same user if there is a power reduction gain, and a vertical bit swapping refers to interchanging the bits among users. Linear programming techniques are used in [120] for dynamic subchannel and bit allocation in multi-user OFDM, which consists essentially of a relaxation of the original

integer programming, similarly to [116]. Finally, the authors in [121] have extensively studied multi-user multi-carrier integer bit loading. Particularly in [121], the Levin-Campello algorithm is extended in order to minimize the total transmitted power with a target sum rate for all the users and a total power budget. For a general mathematical framework, see [122], where the Lagrangian relaxation and the linear relaxation are proposed as approximations of the optimal solution to the Knapsack problem, into which the bit allocation problems can be cast.

Within the context of a combination of frequency diversity (OFDM) and space diversity (SDMA), the authors of [57] investigate bit loading techniques in the uplink and in the downlink, showing the effective throughput increase that can be obtained. They propose a maximum SIR beamformer, which should be equal on all subcarriers for each user, subject to a rate constraint and a per-band (all the subcarriers) power budget per user. Additionally, the authors present a simplified version of the adaptive loading such that all the users use the same constellation size at a given subcarrier, so that the overall complexity of the algorithm is reduced. Differently to Chapter 5, fairness is not addressed since the simplification is just due to complexity issues, but the previous concept is related somehow to the coherence grouping in [94]. Regarding other implementations, bit allocation is applied for Vertical BLAST in [123].

To end with this section, the use of **multiple antennas to both sides** of the communication link shall be (at least) briefly commented. For instance, bit allocation is treated in [124], where the MIMO channel is decomposed into a set of parallel equivalent subchannels through a SVD, so as to apply afterwards water-filling over the eigenmodes. In order to perform integer bit allocation, the authors propose a rounding off solution. So as to increase the efficiency of the used power, a QoS-based solution is proposed, which essentially tries to redistribute the power among the eigenmodes as efficiently as possible. This method is shown to outperform other methods in the literature. Finally, in [125] a minimum power strategy subject to BER constraints is developed. The authors show that the optimal number of operating subchannels is insensitive to the type of CSI, but rather greatly dependent on the data rate requirements for the users.

1.5 The boundary: an insight into fairness issues

In fact, this section is a link between the background chapter and the first contribution, Chapter 2. Although fairness concepts have already appeared somehow during this chapter, an explicit section seems necessary since the tools and concepts that are explained here serve as an inspiration for the subsequent work, or are in some sense part of the work.

1.5.1 Fairness definitions

If in the multi-antenna techniques in Section 1.2 the fairness is implicit in the presented techniques, e.g. by the fulfillment of certain QoS requirements, in this section the fairness is

made explicit by the selection of a given cost function. Therefore, the first question that shall be asked might be: is there a unique definition of fairness? According to [126], there is not, and at least three definitions in the allocation of (wired) bandwidth to users are given. In general, one should define a *utility* function of the resource r_k assigned to the kth user, i.e. $g_k(r_k),\ k = 1, \ldots, K$, generally by the AP. Some remarks are needed: i) as it has been shown, this utility might reflect an ad-hoc function such in the game-theoretic power control, the SNR, or the rate (capacity), among others; ii) the resources that shall be shared among the users are limited, e.g. the instantaneous output power as in this dissertation, although this is implicit in the following description. The AP might then select in principle among three possibilities, see [126] and references therein:

- **Proportional fairness**: This criterion is usually the preferred one in signal processing or information theory, because it optimizes the global performance, see e.g. the sum rate maximization of the Gaussian MIMO BC. A proportional fair allocation is such that the optimization problem can be written as

$$\max_{r_k} \sum_{k=1}^{K} g_k(r_k), \tag{1.18}$$

  which incurs in a performance penalty to the worst user. This means, for instance, that if this optimization is conducted instantaneously in a wireless channel, the user with worst channel condition will get only a small fraction of the resources for the sake of the collective revenue. The interesting property of such a scheme is that for any other feasible allocation $g_k^*(r_k)$, the aggregate of proportional changes is non-positive, i.e.

$$\max_{r_k} \sum_{k=1}^{K} \frac{g_k^*(r_k) - g_k(r_k)}{g_k(r_k)} \leq 0,$$

  which confirms the fact that this scheme *prefers* to serve the users with better conditions. Clearly, this option is preferred by the operator, since e.g. it maximizes the total data rate delivered to higher-layer applications, but other options should be also evaluated.

- **Max-min fairness**: This strategy is the opposite of the previous one, thus it might be the preferred option for the terminals in a bad condition, since it assures that all the users receive the same resource sharing. Mathematically, as long as the utility function is concave (with negative second derivative), it might be expressed as

$$\max_{r_k} \min_{k} g_k(r_k), \tag{1.19}$$

  which has the property that for all k, the utility of the kth user $g_k(r_k)$ cannot be increased without simultaneously decreasing $g_j(s_j)$ for some j with $g_j(s_j) \leq g_k(r_k)$. If the utility function is concave or convex, convex optimization theory [31] might be useful to prove that this scheme finally assigns the same resource to everyone. For more details, see [127].

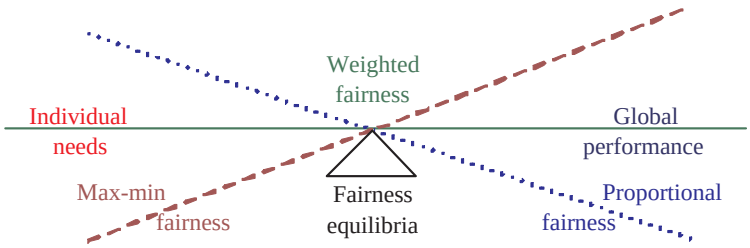


Figure 1.13: The equilibrium points in the *fairness game*.

- **Weighted fairness**: Obviously, there is a *valley between two hills*. In fact, both max-min fairness and proportional fairness can be described using this weighted fairness, which assigns a weight c_k to each utility function $g_k(r_k)$ in (1.18) or (1.19). Essentially, one might then have a weighted proportional fairness

$$\max_{r_k} \sum_{k=1}^{K} c_k g_k(r_k),$$

or a weighted max-min fairness

$$\max_{r_k} \min_{k} c_k g_k(r_k),$$

in which the weights reflect the relative importance of that user in the optimization, since $\sum_{k=1}^{K} c_k = 1$, but how these weights are distributed among users depends on the particular criterion of the scheduler at the assigning entity.

Figure 1.13 illustrates the difference between the fairness definitions, where the weighted fairness is the aristotelic equilibrium point between the extreme positions, namely the max-min fairness and the proportional fairness. In fact, the weighted fairness might include all the intermediate operating points between the total worry about the global performance and the stringent thought on the individual needs. It shall be noted that the fairness criterion is totally subjective, and it is not clear which is the best option, since depending on the burstiness of the traffic, the number of users, the time scale in the system, etc, the scheduling procedure at the assigning entity might select one option or another. This choice (among others) determines the overall system (and service) performance, thus the price consumers are willing to pay.

As stated, fairness issues are usually a matter of the DLC. For instance, according to [97], the major issues in wireless scheduling are, among others, i) the variability of the channel conditions both in time, frequency, and space, ii) fairness issues, which shall consider the fact that a scheduled packet might see the channel in an error state, iii) the provision of QoS, and iv) power constraint and simplicity. In a sense, the scheduling algorithms that will be presented in the following chapters borrow some of these ideas, but it is out of the scope of the dissertation to get into other details such as the queue length or the arriving time and the timeout. For instance,

a good scheduler might provide short-term fairness due to the burstiness of the traffic, as well as long-term fairness, without forgetting the required bounded delay for real-time applications. The main issue in this work is to perform an instantaneous PHY scheduler that helps that at the DLC in the traditional tasks. Therefore, this dissertation might be considered as a first step into the realistic development of a complete cross-layer scheduler.

A common fairness definition in wired and wireless networks, see [128], when regarding scheduling mechanisms is to assure that for every pair of users i and j

$$\left| \frac{R_i(t_1,t_2)}{\phi_i} - \frac{R_j(t_1,t_2)}{\phi_j} \right| \leq \delta, \tag{1.20}$$

where δ can be made as small as wanted, $R_i(t_1,t_2)$ denotes the service (e.g. rate, utility, time slots) user i receives between times t_1 and t_2, and ϕ_i denotes the assigned weight for user i. With the GPS described previously, δ in (1.20) is equal to 0. In advance of further results, assuming a particular format for the weights for the users, the Equal Proportional SNR (EPS) algorithm in Chapter 3 fulfills (1.20) with $\delta = 0$.

1.5.2 Fairness issues at the physical layer

Before presenting the measures of the degree of fairness/unfairness of any resource allocation, a *random walk* down the physical layer might give consistency to the fact that *PHY people* are starting to look into fairness issues in a variety of problems in order to get a broader view of communications, e.g. in order to evaluate not only the sum rate. One of the first papers that brought the writer into fairness issues was [129], where multiple antennas are not only studied at the lowest layer of the protocol stack, but also the analysis is extended to the scheduling mechanism. Some conclusions can be drawn from this paper. The use of diversity antennas has the effect of dampening the variations in the channel conditions, which might reduce the capacity gains of opportunistic scheduling mechanisms that try to exploit fluctuations in the transmission rates. Moreover, diversity antennas may produce substantially smaller gains or even have a negative impact on capacity when combined with scheduling. Finally, as it has been already stated, the gain of multiple antennas might widely vary depending on the fairness notion.

An effort to take into account fairness in a wireline multiaccess channel is done in [130]. There, the traditional algorithms and techniques commented in Section 1.1 are extended to take into account other design possibilities than the traditional sum rate or single user rate. For instance, they deal with the optimization of the common rate, where every user in the system ultimately gets the same rate, i.e. $R_1 = R_2 = \ldots = R_K$. More interestingly, the main focus of [130] is the balanced rate, that is, the ratio between the actual achieved rate in a multi-user environment and the potential rate achieved only by its own channel. It shall be noted that this concept is closely related to the equal proportional rate developed and solved by the author of

this dissertation in [131] for a concrete multi-antenna multi-user scenario. Interestingly in [19], the authors argue that it is difficult to obtain a meaningful metric that quantifies the difference between two K-dimensional rate regions for $K \geq 2$. This issue is first addressed in [35], where it is stated that the sum rate might be a good metric when the difference in terms of SNR among users is not very large, and it is significant from the fact that it quantifies the total data flow [19]. For this reason, the framework in Chapter 2 analyzes the behavior of the multi-antenna schemes which not only takes into account the sum value, but also the dispersion around it.

It shall be noted that the physical layer is concerned about fairness since quite recently, other papers include [132], where a minimum power problem is added a long-term fairness constraint, or [133], where it is shown that with the hard fairness constraint of assigning a subchannel to every user instantaneously, multi-user diversity can still be achieved. To end with these subsection, a random beamforming technique is proposed in [42], whose sum rate throughput scales like DPC with perfect CSI. The basic idea is to construct Q beamformers, and assign them to the users with highest SNR, which is the only necessary feedback. For the purposes of this subsection, the scheme is fair in the sense that when Q is large enough, the probability of transmitting to any user converges to $\frac{1}{K}$ irrespective of the path-loss.

1.5.3 From an index of fairness to portfolio selection

After this brief immersion into the PHY, it shall be recalled that fairness is usually a matter treated at the DLC, in resource allocation it has been studied since the eighties [134]. Prior to this paper, fairness indices were usually qualitative, or quantitative measures too specific for certain applications or with a lack of well-suited properties. The merit of the Index of Fairness (IF) in [134] is that it is i) applicable to any allocation problem, ii) independent of the amount of the resource, and iii) bounded between 0 and 1. Without further preambles, if the resource allocated to user k is r_k and there are K users in the system, the index is expressed as

$$IF = \frac{\left(\sum_k r_k\right)^2}{K \sum_k r_k^2}, \tag{1.21}$$

for which the upper bound ($IF = 1$) reflects an scheme which is totally fair. Note that there is a variety of fairness measures that were previously proposed in the literature, among others the variance, the ratio variance vs. mean, or the min-max ratio. However, they lack from the desirable properties for a fairness index [134]: population size independence, scale and metric independence, boundedness, and continuity. A key point is the selection of the metric to be compared in terms of fairness, since it can be the SNR, the throughput, the power, the length of the queue, etc. It depends on the specific topic the researcher is dealing with. A clear drawback of this type of indices is that they only measure relative performance, thus only serve to compare allocations, but not to design a system. For instance, if $\tilde{r}_k = cr_k$, it is obvious that the IF in (1.21) remains unchanged, although the actual allocation has logically changed.

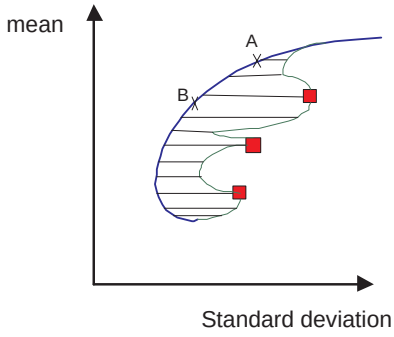


Figure 1.14: Mean vs. variance plot in portfolio selection.

To overcome this problem, the author of this dissertation has been inspired by modern portfolio theory, see e.g. [135] or [136]. Details on the concrete application to multi-antenna multi-user channels can be found in the next chapter, while here, the useful concepts of portfolio theory will only be given. Note that, one would like to obtain the highest possible expected outcome with null risk of a certain investment. However, this cannot happen in markets, where usually more expected outcome comes at the expense of a higher risk [137]. The red boxes in Figure 1.14 reflect a sample of the stocks that might be available. Then, by the adequate portfolio selection of those stocks and other that might be in the market, the area between the green line and the blue line (which is dashed) can be obtained. However, not all the portfolios in this area are efficient, i.e. for any given standard deviation of the portfolio (x axis, which is the risk) one might obtain different values for the mean (y axis, expected profit). Therefore, the efficient combinations of values (efficient frontier) are those on the blue line, e.g. A and B in that case. All the portfolios along this frontier are efficient in the sense that a higher expected value of the outcome cannot be obtained with the same risk. The basic idea behind is that if one selects *firms* that have different behavior in the markets, one might decrease the risk of the individuals. The chosen point along the frontier depends on risk one would like to *suffer*. As a curiosity, such mean vs. variance analysis has been used for the classification of voice signals in [138]. In fact, such an analysis is useful not only for the evaluation but also for the design of multi-antenna multi-user channels. Even more, the mean vs. variance analysis can be deployed whenever a resource sharing problem appears. With the following example, it is emphasized that multi-user communications can benefit from the economic concepts that measure differences in the distribution of a scarce resource; this dissertation is only the first step.

1.5.4 The Gini index as a measure of inequality

Indeed, one can arrive to a very interesting fairness index starting at [139], where several schedulers are compared in terms of fairness. The authors explain briefly the Gini index, which

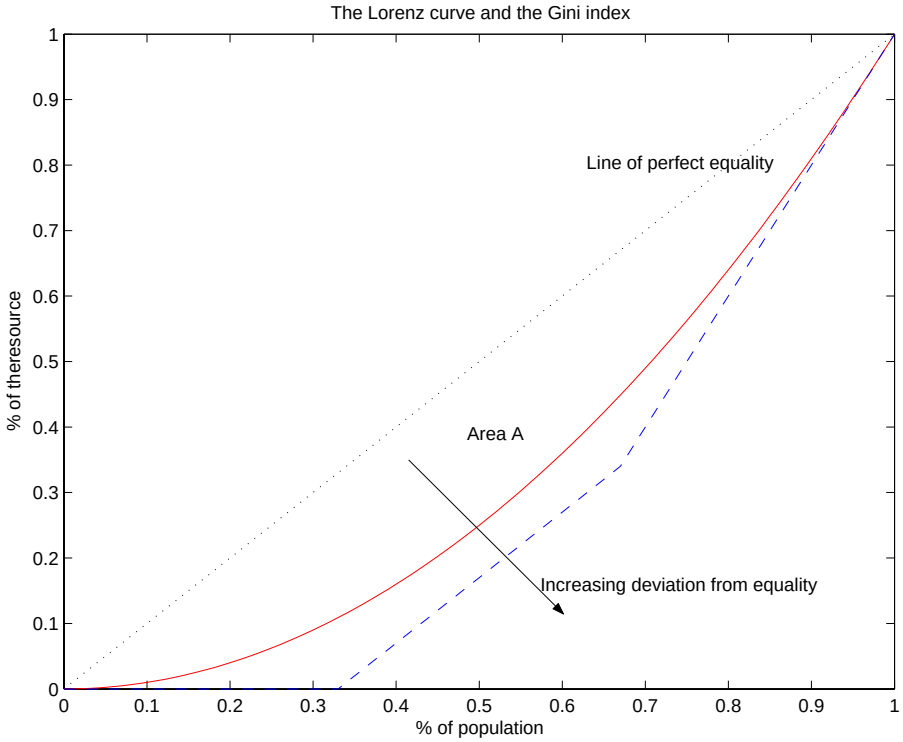


Figure 1.15: Illustrative plot of the Gini index.

can serve as a measure of the degree of inequality in the wealth distribution within a society, which was first proposed in 1921. Essentially, the Gini index measures the degree of fairness/unfairness of a resource allocation $g_k(r_k)$ for the K users such that $g_1(r_1) \leq g_2(r_2) \leq \cdots \leq g_K(r_k)$. In Figure 1.15, the percentage of the resource is plotted as a function of the percentage of the population. If there is perfect equality in these quantities, $g_i(r_i) = g_j(r_j), \forall i, j$, the Lorenz curve in Figure 1.15 will be the 45-degree line starting at the origin, i.e. for any percentage of the population (users) the resource is shared equally among all of them. This might be the distribution that is socially the most fair. As it is depicted with the continuous red line and the dashed line, other Lorenz curves might exist within this unit box, which correspond to different resource allocations reflecting that the income/resource share grows at much slower rate as the population share increases, thus there is a higher degree of resource concentration within the population [140]. The area between the perfect equality and any other Lorenz curve (area A in Figure 1.15) corresponds to the Gini index. Note that this index is one of the most used indicators of social and economic conditions, since the more area among the curves, the more concentrated is the wealth. There are variants of this index which could be quite useful, for instance, the subgroup and source decomposition, but are rather a matter of social theories [140]. One drawback of this method when applied to communications is that the degree of inequality disregards the fact that two distributions might have the same mean, which is analogous to the

problems already described for (1.21). For more details, the reader is referred to the Chapter 3.

1.6 Overview of the dissertation

In this section, an overview chapter by chapter of the dissertation is given, as well as the list of the papers the author has published during the Ph.D. period, which started in December 2000 at the Technical University of Catalonia (UPC), but ended at the Telecommunications Technological Center of Catalonia (CTTC), where the author has been since January 2003.

1.6.1 Chapter 2

The first technical chapter is devoted to fairness issues in multi-antenna techniques, in particular, to the transmit processing, namely Zero Forcing and Dirty Paper Coding, but they are also compared to the cooperative strategy so as to have an upper bound. In order to study the distribution of the resources, in this case the SNR, a mean vs. standard deviation analysis is conducted, which is an extension of the techniques used for portfolio selection. This analysis serves for evaluation in this chapter, but it could also be used for the design of wireless systems. The basic conclusion that can be drawn is that without an explicit cost function, techniques that optimize the global performance tend to be unfair in the resource distribution. Therefore, the choice of the transmit technique at the AP might not be so straightforward.

This chapter has generated basically two publications:

- D. Bartolomé, A. I. Pérez-Neira, *Cross-Layer Design in Multi-Antenna Multi-User Channels: A Unified Framework for Fairness*, submitted to IEEE Transactions on Wireless Communications, April 2004 (revised September 2004).
- D. Bartolomé, A. I. Pérez-Neira, *A Unified Fairness Framework in Multi-Antenna Multi-User Channels*, in Proceedings of the 11^{th} IEEE International Conference on Electronics, Circuits and Systems (ICECS), Tel-Aviv, Israel, December 2004.

1.6.2 Chapter 3

Once the transmit technique has been chosen, more fairness issues come with the power allocation. In that case, there is an explicit cost function to control the final behavior of the cell, and the balance between the global performance and the individual needs can be better controlled. Therefore, traditional power allocation techniques are compared in terms of fairness, that is, not only showing the mean results, but also the behavior of the best and the worst user. For instance, the well-known waterfilling scheme tends not to serve some users for the sake of the collective performance. After an analysis of the techniques in the high SNR, the schemes performing best for each used metric (rate and BER) are compared to a widely deployed view

of communication systems based on Game Theory. It is concluded that the objective function shall be carefully chosen if one does not want to have unacceptable error rates. The third topic within the chapter is the admission control procedure, i.e. the user selection according to their QoS constraints. Two extreme techniques for the power allocation are extended and compared to a newly proposed strategy that balances the performance among them. Results are shown both theoretically and by means of simulations.

This chapter is based on several publications:

- I. Gutiérrez, D. Bartolomé, C.Vilella, *Study of the different power control methods for CDMA systems based on Game Theory* (original title in Spanish), in Proceeding of the XIX Simposium Nacional de la Unión Científica Internacional de Radio (URSI), Barcelona, Spain, September 2004.
- D. Bartolomé, A. I. Pérez-Neira, *BER-based vs. Game-theoretic Power Allocation Strategies for Multiuser MISO Systems*, in Proceedings of the X European Signal Processing Conference (EUSIPCO), Viena, Austria, September 2004.
- D. Bartolomé, A. I. Pérez-Neira, *Performance Analysis of Scheduling and Admission Control for Multiuser Downlink SDMA*, in Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, May 2004.
- D. Bartolomé, A.I. Pérez-Neira, *Spatial Scheduling in Multiuser Wireless Systems: from Power Allocation to Admission Control*, submitted to IEEE Transactions on Wireless Communications, October 2003.
- D. Bartolomé, D.P. Palomar, A.I. Pérez-Neira, *Real-Time Scheduling for Wireless Multiuser MISO Systems under Different Fairness Criteria*, in Proceedings of the Seventh International Symposium on Signal Processing and Applications (ISSPA), Paris, France, July 2003.

Moreover, the author has been the advisor of two Final Degree Projects dealing with power control based on game theory for CDMA systems.

1.6.3 Chapter 4

This chapter is devoted to the bit allocation strategies that could be implemented. Again, the problem can be approached from (at least) two perspectives, the satisfaction of the individual needs and the achievement of the optimum performance. In this case, signaling requirements also have something to say. The author proposes a mechanism between both, which obtains a Pareto improvement over the satisfaction of the individual needs. In this chapter, it is treated for the first time throughout the dissertation the fact that the number of users might be higher than the number of antennas. Since the problem is too complicated to obtain the optimum solution, three greedy intelligent solutions are compared. Basically, the conclusion is that better performance comes at the expense of more complexity.

This chapter is an extension of the conference paper

- D. Bartolomé, A. I. Pérez-Neira, *Multiuser Spatial Scheduling in the Downlink of Wireless Systems*, in Proceedings of the 3^{rd} IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), Sitges, Spain, July 2004.

1.6.4 Chapter 5

Finally, this chapter is devoted to the extension of such bit allocation schemes to the multi-carrier scenario. This chapter is innovative, since there is scarce literature dealing with such a hard problem of space-frequency diversity. After the extension of the algorithms in the previous chapter to the multi-carrier scenario, the problem is approached from a practical perspective. Then, several suboptimal solutions are motivated and compared, showing that the normalized scalar product among channel vectors might yield a reasonable trade-off between performance and complexity. After that, the practical considerations such as complexity reduction and signaling are addressed, proposing solutions for each problem.

This chapter is based on

- D. Bartolomé, A.I. Pérez-Neira, *Practical Implementation of Bit Loading Schemes for Multi-Antenna Multi-User Wireless OFDM Systems*, submitted to IEEE Transactions on Communications, October 2004.

- D. Bartolomé, A.I. Pérez-Neira, *Practical Bit Loading Schemes for Multi-Antenna Multi-User Wireless OFDM Systems*, in Proceedings of the Asilomar Conference on Signals, Systems, and Computers, November 2004.

- D. Bartolomé, A. Pascual-Iserte, A.I. Pérez-Neira, *Spatial Scheduling Algorithms for Wireless Systems*, in Proceedings of the 2003 International Conference on Acoustics, Speech and Signal Processing, (ICASSP), Hong Kong, China, April 2003.

1.6.5 Other publications

During the first period of the dissertation at the UPC, the author was involved in the design of multi-antenna techniques for Wireless LAN, where he could learn the basics of OFDM and of multiple antennas. From this period, other collaborations, and before, several papers have been published that are not (explicitly) contained in the dissertation, namely,

- C. Hennebert, P. Rosson, D. Bartolomé, A. Pascual-Iserte, Ana I. Pérez-Neira, *Practical Implementation of Space-Diversity Receivers in OFDM Systems: Structure, Performance, and Complexity*, in Proceedings of the 13^{th} IST Mobile Communications Summit, Lyon, France, June 2004.

- D. Bartolomé, A. Pascual-Iserte, A.I. Pérez-Neira, and P. Rosson, *From a Theoretical Framework to a Feasible Hardware Implementation of Antenna Array Algorithms for WLAN*, in Proceedings of the 12^{nd} IST Mobile Communications Summit, Aveiro, Portugal, June 2003.

- D. Bartolomé, A.I. Pérez-Neira, *MMSE Techniques for Space Diversity Receivers in OFDM-based Wireless LANs*, IEEE Journal on Selected Areas in Communications, February 2003.

- D. Bartolomé, A.I. Pérez-Neira, *Reconfigurable Antenna Array Architecture for OFDM Receivers*, in Proceedings of 2^{nd} IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Marrakech, Morocco, December 2002.

- D. Bartolomé, A.I. Pérez-Neira, *Exploiting the Cyclic Prefix for Beamforming in OFDM Receivers*, in Proceedings of the IX European Signal Processing Conference (EUSIPCO), Toulouse, France, September 2002.

- D. Bartolomé, A.I. Pérez-Neira, A. Pascual-Iserte, *Blind and Semiblind spatio-temporal diversity for OFDM systems*, in Proceedings of the 2002 International Conference on Acoustics, Speech and Signal Processing, (ICASSP), Orlando, USA, May 2002.

- D. Bartolomé, A.I. Pérez-Neira, *Pre- and Post-FFT SIMO Array Techniques in Hiperlan/2 Environments*, in Proceedings of the 2002 Spring Vehicular Technology Conference (VTC), Birmingham, USA, May 2002.

- D. Bartolomé, A.I. Pérez-Neira, *Modified SMI Techniques for Frequency Selective Channels in OFDM*, in Proceedings of the 2002 Spring Vehicular Technology Conference (VTC), Birmingham, USA, May 2002.

- D. Bartolomé, X. Mestre, A.I. Pérez-Neira, *Single Input Multiple Output techniques for Hiperlan/2*, in Proceedings of the 10^{th} IST Mobile Communications Summit, Sitges, Spain, September 2001.

- S. Simoens, D. Bartolomé, *Optimum performance of link adaptation in Hiperlan/2 networks*, in Proceedings of the 2001 Spring Vehicular Technology Conference (VTC), Rodes, Greece, May 2001.

Chapter 2

Fairness in multi-antenna processing

In multi-user communications, the Access Point (AP) has several alternatives for distributing the scarce resources among users. Since there exists a trade-off between the global performance and the individual needs, an analytical framework to study fairness is derived in this chapter, which overcomes the relative nature of the fairness indexes in the literature by obtaining closed-form expressions for a wide range of situations. The framework is inspired by portfolio selection, and basically analyzes the trade-off between the mean and the standard deviation in multi-user communications. With this relationship, it can be verified that more mean comes usually at the expense of a higher variance. Instantaneously in a cell, this means that the users with worse channels obtain less with the method that maximizes the cell performance and might not be allocated for transmission. On the other hand, if an equal behavior is imposed for all the users, the best users are not completely satisfied, thus the global outcome is penalized.

The target application is a multi-antenna AP transmitting simultaneously to several single-antenna terminals within a cell, i.e. a multi-antenna broadcast channel. However, the framework proposed is valid to analyze other procedures in multi-user communications. The novelty of such an approach comes from basically three aspects. First, usually the performance is maximized as a whole, without considering the interactions among the users. In multi-user communications they might have paramount importance and should be included in the design of (especially wireless) systems. Second, the fairness considerations that are studied in this chapter have rarely been studied at the physical layer, since they are usually treated as part of the DLC. Third, the framework that is here proposed overcomes some of the *inefficiencies* of previous index of fairness that were proposed in the literature, as it will be shown next.

This chapter is organized as follows. First, a brief overview on multi-antenna processing and fairness is given, after which the reader finds the system modeling in Section 2.2. Then, it is proposed in Section 2.3 a general framework based on a mean vs. variance (or standard deviation) analysis. Such analysis has been widely used in modern portfolio selection [137], and

briefly, the basic idea is that one can assume a certain risk (expressed by the variance) in order to attain a desired expected profit (that is, the mean). Obviously, everyone would like to invest in a portfolio with a low risk and high profit, but this is not always possible as it is clear by the behavior of financial markets. Usually, one pays the price of a higher expected outcome at the expense of a higher risk [141]. Analogously in multi-antenna multi-user communications, the preferred option would provide a higher mean value together with low variance, but it is shown in Section 2.4 that this might be difficult to obtain in practice, since more mean usually comes at the expense of more variance. Finally, conclusions are drawn.

2.1 Introduction

Although fairness is usually studied at higher layers, it is observed as a trade-off between the global performance and the individual needs at the physical layer, see e.g. [129]. Generally speaking, whenever the sum value among the users is maximized, the differences among the maximum and minimum becomes higher. However, if an equal performance is forced for all the users, the global performance is penalized [131]. This has implications in the system design: if all the users are homogeneous and generate bursty traffic, the AP shall provide an equal performance for all the terminals in the cell. On the other hand, this might severely degrade the total throughput, thus the revenues of the operator. In this sense, the choices at the AP are, at least, not straightforward, not only because of the performance vs. complexity trade-off, but also due to the trade-off between the global performance and the individual needs.

As it has been seen in the previous chapter, in the literature of the second layer of the protocol stack [142], one finds essentially three types of fairness [126], namely i) **max-min fairness**, which gives all the users the same performance, ii) **proportional fairness**, which maximizes the sum performance of all the users, and iii) **weighted fairness**, which is a modification of the previous two strategies that includes a different weight for each user. Although it might be foreseen that proportional fairness might yield better performance than max-min fairness, these objective functions do not clarify which will be the exact behavior of the resource sharing. As it has been stated, the maximization of a sum among several users yields higher differences among them than the max-min criterion. In the literature, the main index of fairness in [134], which is cited in (1.21), is closely related to the framework it is presented here, and has appealing properties such as population size independence, scale and metric independence, boundedness, and continuity, but again, it measures only relative performance. Moreover, it is a global measure for the K users. Therefore, it does not give a clear idea on the specific user behavior, e.g. the individual outcomes (equal for everyone) could be extremely low.

Clearly, the fairness index in (1.21) depends on the distribution of the resource r_k, but does not provide an idea on how these resources are shared. As an example, one could find the *socially*

most fair distribution, i.e. equal resource to all the agents, however, it might assign only a very small fraction of the resource to everyone, e.g. $\bar{r}$. On the other hand, one could find another distribution which allocates a different fraction to everyone, but no user might get less than $\bar{r}$. This will be seen by the index of fairness as less fair than the first one, although the no user might get less than with the first strategy. Furthermore, another distribution that assigns $100\bar{r}$ to everyone is as fair as the one that assigns only $\bar{r}$, even though the global behavior is much better. In a sense, the framework in this chapter is another way of looking at this fairness index. The trade-off it is characterized gives not only the mean value but also how the resources might be distributed. Interestingly, the mean vs. variance analysis has also been used in the literature of speech recognition for the classification of voice signals [138].

Inspired by the cross-layer philosophy, the fairness analysis is particularized to the system studied in this dissertation, i.e. a multi-antenna broadcast channel [23], where the transmitter is provided with multiple antennas, and the receivers have only one. Recently, the trade-off between diversity and multiplexing has been characterized for multiple-antenna systems [66], and also for the extensions in multiple-access channels [67]. Although these excellent information-theoretic trade-offs curves characterize the rate vs. bit error rate (reliability) of a system with concrete techniques, they do not deal with the differences among the users. Regarding multiple antennas, relative fairness indexes have been proposed in [131] relating the actual rate achieved by a given user in a multi-antenna multi-user system, with the rate a user would achieve if all the antennas were dedicated to him/her, i.e. as if the desired user were alone in the cell. More recently, this relative (or proportional) fairness has been proposed taking the SNR as performance metric instead of the rate [69]. These relative fairness indexes, though, might not be sufficient for a practical design of communication systems. Moreover, the focus of this chapter is at the multi-antenna processing, and the measurable differences are under consideration.

It is meaningful to separate the problem at the AP into the transmit beamforming and the power allocation. Since the focus of this chapter is the transmit beamforming, a Uniform Power Allocation (UPA) is chosen without loss of generality. For details on fairness issues at the power allocation, with and without QoS constraints, please refer to the next chapter. Although it would not yield optimum results in terms of rate or error rate performance, the UPA is shown to be asymptotically equivalent to the maximum sum rate power allocation (waterfilling) at high SNR (see Chapter 3). With these assumptions, two strategies at the transmitter are analyzed in terms of fairness, namely, the widely-deployed Zero Forcing (ZF) [84], Dirty Paper Coding (DPC), see e.g. [20] or [86], and they are compared the cooperative processing if both transmitter and receivers have full channel knowledge [100]. Other options include e.g. the matched filtering or the MMSE transmit pre-equalization, see [61] for a comparison by means of simulations. In any case, the cooperative scheme is an upper bound, since it is assumed that the receivers have also perfect CSI, which might be unrealistic in this type of communications. The interesting

point is that with ZF, the AP creates parallel and orthogonal channels at each subcarrier, thus at most Q users can be served at each subcarrier. However, it is well-known that there exists a number of antennas per user that optimizes the sum rate with ZF beamforming [55], i.e. as many users as antennas is suboptimum. Although this is out of the scope of this chapter, it will be an important issue in the next.

The performance of ZF might be enhanced by DPC [20], which yields asymptotically optimum throughput for high SNR if the channel matrix has full row rank and it is perfectly non-causally known at the transmitter. Based on [22], the multi-antenna transmitter might perform a pre-cancellation of the interference (ideally) without a rate penalty, see also e.g. [35]. In the literature, the broadcast channel is usually studied in terms of sum rate, see e.g. [21] or [34], or [29] for the dual MAC. However, it has not been shown in the literature how the transmit processing might affect the fairness among the users, which is an important issue when bursty traffic dominates the type of established links. Compared to [51] or [52], any SNR requirement for the users is tried to be fulfilled, but instead, the focus lies on the difference among the performance of the users without constraints. It is indeed interesting to see how the transmit processing techniques determine the fairness of the system. In this sense, fairness should be seen as a cross-layer issue, since concepts of higher layers are adapted to the physical layer.

2.2 Problem statement

The focus is the downlink, where a Q-antenna AP communicates simultaneously with K single-antenna terminals, under the assumption that $K \leq Q$. In a practical situation $K > Q$, thus some kind of user grouping should be performed, but this will be treated in following chapters. The analysis presented here is valid for each group of users that would be formed in those cases. In fact, the general fairness framework could even be useful for the clustering of users, since it can be part of a design criterion. Generally speaking, it can be considered that both the transmitter and the receivers perform some kind of linear processing [100], so that at any time instant, the **signal model** is expressed as

$$\mathbf{y} = \mathbf{AHBs} + \mathbf{w} \in \mathbb{C}^{K\times 1}, \tag{2.1}$$

where the kth position of vector $\mathbf{y}$ ($\mathbf{s}$) is the received (transmitted) signal for user k. $\mathbf{H}$ is the $K\times Q$ complex flat-fading channel matrix, whose ith row contains the $1\times Q$ vector of the channel gains for the ith user, i.e. $\mathbf{h}_i^T$. The channel matrix elements are independent and identically distributed complex Gaussian random variables with zero mean and unit variance, and is assumed to be perfectly known at the transmitter. The noise vector is complex Gaussian, i.e. $\mathbf{w} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sigma^2\mathbf{I}_K\right)$. The transmit beamvectors are gathered in the matrix $\mathbf{B} = [\mathbf{b}_1\ \mathbf{b}_2 \ldots \mathbf{b}_K] \in \mathbb{C}^{Q\times K}$, and the receive processing is included in the matrix $\mathbf{A} \in \mathbb{C}^{K\times K}$, the kth row of which contains the receive filter

for user k, $\mathbf{a}_k^T$. Assuming that the transmitted signal $\mathbf{s}$ has unitary mean energy and the power budget is P_T, then $tr(\mathbf{B}^H\mathbf{B}) \leq P_T$ should be fulfilled instantaneously.

2.2.1 Cooperative scheme

With the signal model in (2.1), if the channel is perfectly known not only at the transmitter but also at the receivers, then all of them can cooperate [30]. In fact, cooperation might yield the Pareto-efficient operating point, where no user can further increase its performance without decreasing the performance of any other user [12]. Related to this, the Coase theorem for profit maximization states that a set of agents might always try to obtain the highest overall profit, since there always exist a set of payment functions which would yield a higher profit than any other solution. However, it is rather difficult in communications to design efficient protocols for the necessary exchange of such an amount of information. Anyway, provided a SVD of the channel matrix $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$, the transmitter beamforming matrix $\mathbf{B}$ can be decomposed into $\mathbf{B} = \widetilde{\mathbf{B}}\mathbf{D}_p$, where $\widetilde{\mathbf{B}}$ is unitary and $\mathbf{D}_p$ is a diagonal matrix that contains the power allocated to the K users, that is, $\mathbf{D}_p = diag\left(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K}\right)$. The power constraint can be expressed as $tr(\mathbf{D}_p^2) = \sum_{k=1}^{K} p_k \leq P_T$. In the cooperative scheme, the unitary beamforming matrix is $\widetilde{\mathbf{B}} = \mathbf{V}$, and the receiver k would use the kth row of the unitary matrix $\mathbf{U}^H$ as receive filter. In that case, the received signal for each terminal k is given by

$$y_k^{CO} = \lambda_k \sqrt{p_k} s_k + w_k,$$

where λ_k denotes the kth eigenvalue of the channel matrix $\mathbf{H}$. If one would like to obtain the sum rate of this channel, the well-known water-filling comes up as the solution of the design of the power allocation factors p_k [28]. Since a water-filling scheme would increase the differences among the users, and provided that is equivalent to the UPA at high SNR, it is advanced that the latter is used for the fairness analysis in this chapter.

2.2.2 Transmit-only processing

In a TDMA/TDD system, if terminals send/receive a periodic training sequence to/from the AP, but not to/from other terminals, they might only be aware of their own channel response, while the AP (transmitter) might then have full channel knowledge. According to its computational capabilities (higher than terminals), the AP is able to perform beamforming and power allocation. The objective of this subsection is to present two transmitter techniques that are be capable of eliminating completely the inner-cell interference provided perfect channel knowledge at the transmitter. Note that full channel knowledge is fairly unrealistic, at least some kind of channel prediction seems to be necessary in a realistic implementation.

In this case, it is assumed that $\mathbf{A}$ in (2.1) is the identity matrix because the receivers do not perform any filtering. At the transmitter, the beamforming matrix should be designed not

to change the original power from the channel, which allows a clean separation of the transmit processing technique from the particular resource distribution in terms of power allocation. In the most general case the received signal for the kth user would have interference from the rest of the users according to

$$y_k = s_k \mathbf{h}_k^T \mathbf{b}_k + \sum_{i \neq k} s_i \mathbf{h}_i^T \mathbf{b}_i + w_k, \tag{2.2}$$

but with ZF or DPC it can be completely removed. The former is a linear processing technique, which is equivalent to the MMSE when a low number of users and also for high SNR [57]. Moreover, it yields a reasonable degradation with respect to the optimum sum capacity as it is shown in [54]. ZF implies that the K users see parallel and orthogonal fading channels corrupted only by Additive White Gaussian Noise and not by interference signals from other users, although it might not work well in general when the mobiles have also multiple antennas. For such a case, in [74] the authors present two constrained solutions, namely a block diagonalization and successive optimization, which yield closed-form expressions to maximize the sum capacity and provide a reasonable trade-off between performance and complexity. Besides, they provide extensions if these two methods cannot be supported. In [143], the multi-user MIMO channel is decomposed into several parallel independent conventional single-user MIMO channels, in which if a transmit antenna is added, the number of spatial channels for each user is increased also by one. Finally, iterative algorithms are considered for uplink MIMO systems in [70] to find the optimum transmit and receive linear filters according to a mean squared error criterion, assuming an error-free low-delay feedback channel. On the other hand, optimum joint transmit-receive filter design with the computationally expensive Simulated Annealing is treated in [144]. In this chapter and in the whole dissertation, single-antenna terminals are considered taking into account the current state of technology [79]. In any case, the proposed system provides a well-suited framework to analyze more practical issues such as realistic spatial scheduling. On the other hand, it is shown e.g. in [20] that DP yields asymptotically optimum throughput for high SNR if the channel matrix has full row rank and it is perfectly non-causally known at the transmitter. Dirty Paper is due to Costa [22], who showed that the capacity of a single-user additive white Gaussian noise channel is unchanged in the presence of an independent additive white Gaussian interferer, provided that this signal is non-causally known at the transmitter. Based on this, the multi-antenna transmitter can perform a pre-cancellation of the known interference without a rate penalty, see also e.g. [35]. In [145], a combination of the Tomlinson-Harashima precoding with scheduling is studied in the long term, whose final objective is to select the subset of users that maximizes the sum rate. Next, these two strategies are further developed and analyzed.

Particularly with **Zero Forcing**, the interference signals in (2.2) can be completely eliminated by creating parallel and orthogonal spatial channels, thus the signal received for the kth user is only corrupted by noise. The equivalent channel is captured by α_k, so that the beamforming

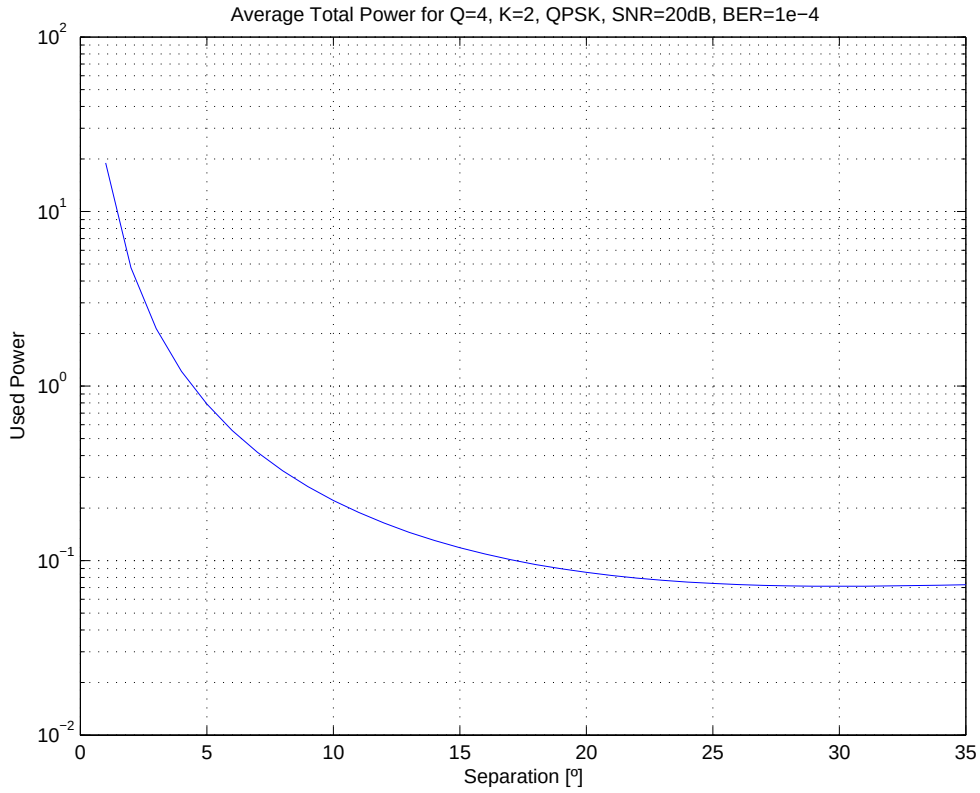


Figure 2.1: Example of used power for a concrete scenario.

criterion becomes $\mathbf{H}\widetilde{\mathbf{B}} = \mathbf{D}_\alpha$, where $\mathbf{D}_\alpha = diag\left(\alpha_1, \alpha_2, \ldots, \alpha_K\right)$ [69]. Since

$$\alpha_k^2 = 1/\left[\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\right]_{k,k}, \tag{2.3}$$

it is essential that the channel covariance matrix is well-conditioned, i.e. the channels from the users should not arrive from close directions in the spatial domain. The α_k^2 behave like independent central Chi-Square random variables with $2(Q-K+1)$ degrees of freedom, i.e. $\alpha_k^2 \sim \frac{1}{2}\chi^2_{2(Q-K+1)}$. However, they concentrate the effect of the choice of the users that are simultaneously served, because they depend on the inverse of the matrix $\mathbf{H}\mathbf{H}^H$. If the channel vectors are highly correlated, then the determinant of the matrix $\mathbf{H}\mathbf{H}^H$ tends to 0, and consequently, the equivalent spatial channels in (2.3) tend also to be null out. As an extreme example, if two rows were exactly the same, the previous matrix would be ill-conditioned and the inverse would not exist. Relaxing this, if the rows of the matrix $\mathbf{H}$ are highly correlated, i.e. the channels refer to a close spatial zone, more power is needed to fulfill the requirements of any user. This is shown in an illustrative example in Figure 2.1, where the used power of serving two users is evaluated in terms of the separation among them in degrees, with $Q = 4$, SNR = 20 dB, QPSK symbols, and the objective of a target BER of 10^{-4}. The effect of the relative angular position of two users into the power used in the cell can be clearly appreciated.

Therefore, an intelligent scheduler might try to separate those users with similar channel vectors. In this particular case, if the power constraint were 1, a higher separation than 5° between the two users would be needed.

As it has been shown, the beamforming matrix contains the power allocation, i.e. $\mathbf{B} = \widetilde{\mathbf{B}}\mathbf{D}_p$, so the signal model in (2.2) finally reduces to

$$\mathbf{y} = \mathbf{D}_\alpha \mathbf{D}_p \mathbf{s} + \mathbf{w} \Rightarrow y_k^{ZF} = \alpha_k \sqrt{p_k} s_k + w_k. \tag{2.4}$$

As it has been stated, the performance of the ZF can be dramatically improved by the application of **Dirty Paper** techniques, which attain the maximum sum rate of the broadcast channel by a pre-cancellation of the interference signals for every user, see e.g. [20] and references therein. Theoretically, a QR decomposition of the channel matrix can be performed, that is $\mathbf{H} = \mathbf{R}\mathbf{Q}$, where $\mathbf{R}$ is a $K \times K$ lower triangular matrix, and $\mathbf{Q}$ is a $K \times Q$ matrix with orthonormal rows. Therefore, the unitary beamforming matrix is $\widetilde{\mathbf{B}} = \mathbf{Q}^H$, so that the signal model becomes

$$y_k = [\mathbf{R}]_{k,k} \sqrt{p_k} s_k + \sum_{i<k} [\mathbf{R}]_{k,i} \sqrt{p_i} s_i + w_k, \tag{2.5}$$

but this interference can be cancelled *a priori* at the AP, and the maximum sum rate of the channel is obtained if the corresponding power allocation is performed [20]. Certainly, this interference pre-subtraction implies power, modulo, and shaping losses, but there are efficient moderately-complex methods to reduce the rate penalty induced in practical implementations, see e.g. [21]. Therefore, in this chapter it is assumed that DP might achieve the theoretical bound without degradation due to the imperfections in the realistic implementation. A remark is needed: although the ordering of the users has some impact on performance, a random order is selected in this chapter, which already yields much better results than ZF. With this strategy, the signal model in (2.5) simply reduces to

$$y_k^{DP} = d_k \sqrt{p_k} s_k + w_k,$$

where $d_k = [\mathbf{R}]_{k,k}$ [20]. Next, the mean vs. standard deviation analysis is performed for the three techniques that have been concisely described.

2.3 Fairness analysis

In this section, the previous three multi-antenna techniques are compared in terms of fairness. Instead of an index such as (1.21), the necessary information might be collected in a plot showing both the mean and the standard deviation, although other similar alternatives might also be good well-suited. As stated, this analysis has been inspired by portfolio selection, and basically allows the AP to evaluate and select a multi-antenna technique in a multi-user scenario. The

relevant fairness index (1.21) is proposed in [134], and others that take into account multiple antennas have been recently proposed, see [131] or [69]. Nevertheless, they measure only relative performances and serve just for evaluation, thus they might not be sufficient in a practical cross-layer design of multi-user communications. Moreover, they are global measures for the K users, thus do not give a clear idea on the specific user behavior, e.g. the individual outcomes (equal for everyone) could be extremely low. Therefore, a plot reflecting the individual behavior and the global outcome seems to be a good alternative. An additional novelty of the framework in this chapter is the evaluation of fairness at the PHY antenna processing.

Without loss of generality, the Uniform Power Allocation among the K users is chosen, i.e.

$$p_k = \frac{P_T}{K}, \quad k = 1, \ldots, K,$$

which is shown to be equivalent to the maximum sum rate solution in the high SNR range e.g. in [69], although it might not yield optimum solutions neither in terms of sum capacity, sum BER, nor any other performance criterion. In the work conducted in [131] or [69], which will be thoroughly developed in the next chapter, it is stated that the differences between the best and the worst user is higher if the mean is higher. This shows the uneven distribution of the resources given by the schemes that seek an optimum global (sum) performance. For the characterization of this trade-off, the techniques are analyzed first by calculating the expectations over the channel matrix, and then over the users. Indeed, this procedure gives clarifying results for the intuition the author aims to show in this chapter. The relevant results are collected in Table 2.1.

2.3.1 Cooperative scheme

Assuming a UPA, the SNR at the kth receiver with the cooperative scheme is given by

$$\gamma_k^{CO} = \frac{P_T}{\sigma^2}\frac{\lambda_k^2}{K} = \gamma^n \frac{\lambda_k^2}{K},$$

where $\gamma^n = \frac{P_T}{\sigma^2}$ only reflects a SNR scaling, thus it is omitted in the analysis. Note that for the three techniques studied in this section, the author concentrates on the equivalent channels that are created after the multi-antenna processing, thus K appears only in the summary of the results depicted in Table 2.1, but not in the derivations. With this the cooperative scheme, the mean and the variance of the square of the eigenvalues λ_k over the random channel matrix are given by $\mathbb{E}_{\mathbf{H}}(\lambda_k^2) = Q$, and $\sigma^2_{\mathbf{H}}(\lambda_k^2) = QK$ [146]. Clearly, they are also the expectation over the K users[5], i.e.

$$\mathbb{E}_{\mathbf{H},K}(\lambda_k^2) = \mathbb{E}_{\mathbf{H}}(\lambda_k^2) = Q,$$

and

$$\sigma^2_{\mathbf{H},K}(\lambda_k^2) = \sigma^2_{\mathbf{H}}(\lambda_k^2) = QK.$$

[5] In the following, the expectation or variance over the users is denoted by the subscript K.

With the values in Table 2.1, the index of fairness in (1.21) can be computed as[6]

$$IF^{CO} = \frac{\mathrm{E}^2_{\mathbf{H},K}(\lambda_k^2/K)}{\mathrm{E}^2_{\mathbf{H},K}(\lambda_k^2/K) + \sigma^2_{\mathbf{H},K}(\lambda_k^2/K)} = \frac{Q}{Q+K} = \frac{1}{1+\xi},$$

where $0 \leq \xi = \frac{K}{Q} \leq 1$. This IF depends exclusively on the ratio $\xi = \frac{K}{Q}$, and $0.5 \leq IF^{CO} \leq 1$.

2.3.2 Zero Forcing

With **Zero Forcing** (ZF) and considering a UPA, the SNR for the kth receiver is obtained as

$$\gamma_k^{ZF} = \gamma^n \frac{\alpha_k^2}{K}.$$

Therefore, taking into account that the α_k^2 are distributed as central independent Chi-squared random variables with $2(Q-K+1)$ degrees of freedom[7], i.e. $\alpha_k^2 \sim \chi^2_{2(Q-K+1)}$, it is straightforward to obtain the mean and variance, i.e. $\mathbb{E}_{\mathbf{H}}(\alpha_k^2) = \sigma^2_{\mathbf{H}}(\alpha_k^2) = Q-K+1$, which again yield the same average among the users, i.e.

$$\mathbb{E}_{\mathbf{H},K}(\alpha_k^2) = \sigma^2_{\mathbf{H},K}(\alpha_k^2) = Q-K+1.$$

As it is shown in [56], this maximum diversity might be though reduced in presence of correlation in the channels from the users. However, for the goal of this chapter it is considered that the channels from the users are completely uncorrelated. With the previously computed mean and variance, the index of fairness in (1.21) can be obtained as

$$IF^{ZF} = \frac{\mathrm{E}^2_{\mathbf{H},K}(\alpha_k^2/K)}{\mathrm{E}^2_{\mathbf{H},K}(\alpha_k^2/K) + \sigma^2_{\mathbf{H},K}(\alpha_k^2/K)} = \frac{1-\xi+1/Q}{1-\xi+2/Q},$$

which converges to 1 as both the number of users K and antennas Q grow without bound but their ratio remains fixed to $\xi = \frac{K}{Q}$.

2.3.3 Dirty Paper

Although a uniform power allocation is assumed, **Dirty Paper** (DP) still outperforms other transmit processing techniques such as ZF, but note that this power allocation strategy is not optimal in terms of sum rate. The SNR for the kth user with DP is given by

$$\gamma_k^{DP} = \gamma^n \frac{d_k^2}{K},$$

[6] Note that the IF in (1.21) can be expressed as $IF = \mathbb{E}^2_K\{r_k\}/\mathbb{E}_K\{r_k^2\}$.

[7] A Chi-squared random variable with r degrees of freedom χ_r^2 has mean r and variance $2r$. Since the channel is complex, each Gaussian random variable of the channel matrix has variance of $1/2$ instead of 1. Therefore, the mean of the Chi-squared random variables is given by $r/2$ and the variance is also $2r/4 = r/2$.

Technique	Gain	Mean	Standard Deviation	Asymptotic IF
Cooperative	λ_k^2/K	Q/K	$\sqrt{Q/K}$	$1/(1+\xi)$
Dirty Paper	d_k^2/K	$(2Q-K+1)/2K$	$\sqrt{Q+\frac{1}{12}(K-5)(K-1)}/K$	$(2-\xi)^2/\left[(2-\xi)^2+\xi^2/3\right]$
Zero Forcing	α_k^2/K	$(Q-K+1)/K$	$\sqrt{Q-K+1}/K$	1

Table 2.1: Mean and standard deviation of the SNR (without the scaling γ^n) for the three schemes: cooperative bound, Zero Forcing transmit beamforming, and Dirty Paper encoding.

where the d_k^2 are distributed as central Chi-Squared random variables with $2(Q-k+1)$ degrees of freedom, i.e. $d_k^2 \sim \chi^2_{2(Q-k+1)}$ [20]. The difference with ZF shall be commented here. Whereas DP-like techniques achieve an increasing diversity for the users, with ZF the system is restricted to the minimum of those diversity values, thus the global performance is penalized. Then, the average statistics over the channel are $\mathbb{E}_{\mathbf{H}}(d_k^2) = \sigma^2_{\mathbf{H}}(d_k^2) = Q-k+1$, with which the average over the K users is given by

$$\mathbb{E}_{\mathbf{H},K}(d_k^2) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{H}}(d_k^2) = \tfrac{1}{2}(2Q-K+1), \tag{2.6}$$

where the equality $\sum_{k=1}^{K} k = \frac{1}{2}K(K+1)$ has been used. In order to obtain the variance over the users, some previous computations are needed. Since $\sigma^2_{\mathbf{H}}(d_k^2) = \mathbb{E}_{\mathbf{H}}([d_k^2]^2) - \mathbb{E}^2_{\mathbf{H}}(d_k^2)$, one can obtain

$$\mathbb{E}_{\mathbf{H}}([d_k^2]^2) = \sigma^2_{\mathbf{H}}(d_k^2) + \mathbb{E}^2_{\mathbf{H}}(d_k^2) = (Q-k+1)(Q-k+2).$$

Using this and some manipulations, $\mathbb{E}_{\mathbf{H},K}([d_k^2]^2) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{H}}([d_k^2]^2)$ becomes

$$\mathbb{E}_{\mathbf{H},K}([d_k^2]^2) = (Q+1)(Q+2) + (K+1)\left(\tfrac{1}{6}(2K+1) - \tfrac{1}{2}(2Q+3)\right), \tag{2.7}$$

where $\sum_{k=1}^{K} k^2 = \frac{1}{6}K(K+1)(2K+1)$ shall be recalled, so that finally

$$\sigma^2_{\mathbf{H},K}(d_k^2) = \mathbb{E}_{\mathbf{H},K}([d_k^2]^2) - \mathbb{E}^2_{\mathbf{H},K}(d_k^2)$$

can be found with (2.6) and (2.7). After some manipulations, the expression in Table 2.1 is obtained. As with the other techniques, the IF can be computed,

$$IF^{DP} = \frac{\mathrm{E}^2_{\mathbf{H},K}(d_k^2/K)}{\mathrm{E}^2_{\mathbf{H},K}(d_k^2/K) + \sigma^2_{\mathbf{H},K}(d_k^2/K)} = \frac{(2-\xi+1/Q)^2}{(2-\xi+1/Q)^2 + 4/Q + (\xi-5/Q)(\xi-1/Q)/3},$$

which converges to

$$IF^{DP} \to \frac{(2-\xi)^2}{(2-\xi)^2+\xi^2/3}$$

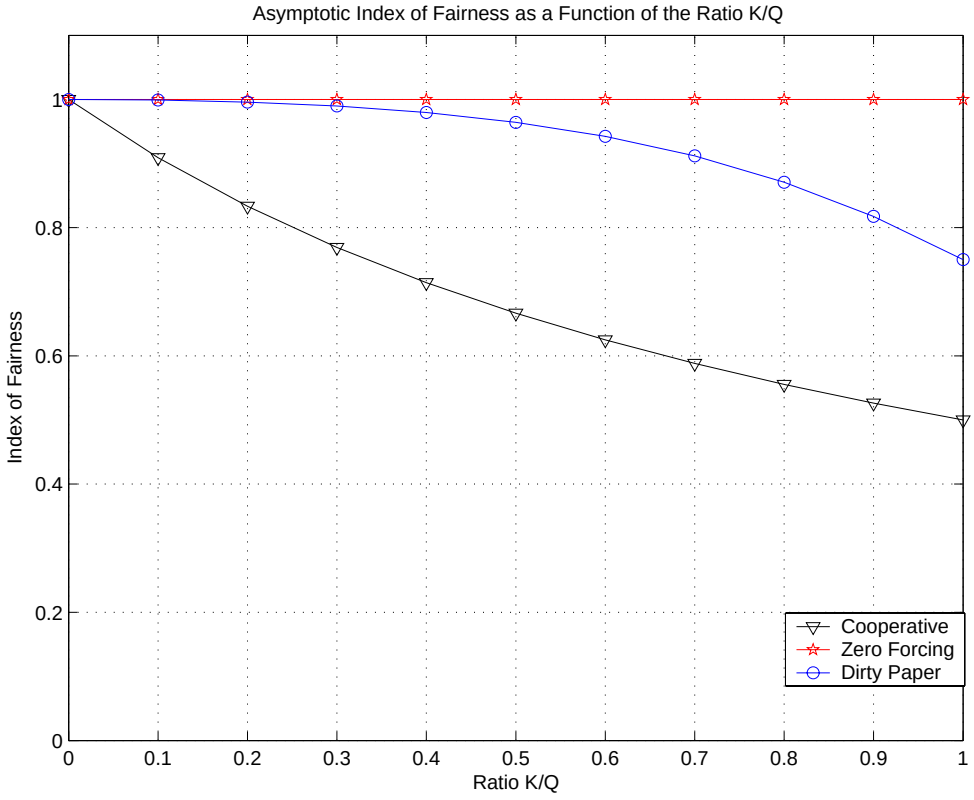


Figure 2.2: Asymptotic index of fairness for the three multi-antenna techniques.

as both the number of users K and antennas Q grow without bound but their ratio remains fixed. In this case, $0.75 \leq IF^{DP} \leq 1$. The results in this section might already shed some light into the behavior of the strategies, but next section explicitly discusses them with the aid of some revealing plots. Moreover, the asymptotic indexes of fairness are compared to show that they might not be sufficient for the evaluation and design of wireless systems, which confirms the usefulness of the proposed mean vs. standard deviation plot as a measure of fairness in multi-user scenarios.

2.4 Results and comparison

First, Fig. 2.2 shows the IF in (1.21) as a function of the ratio K/Q for the three described multi-antenna techniques. In the following, the line with triangles refers to the cooperative scheme, the one with stars to ZF, and the one with circles to DP. According to this index of fairness, ZF is the most fair technique. This is due to the fact that the inner-cell interference is cancelled in the same way for all user, consequently, it yields the same distribution for all of them. Then, DP is less fair because of the successive interference cancellation at the transmitter: some users get a better performance than others, although no one gets less than for ZF. Finally, the cooperative

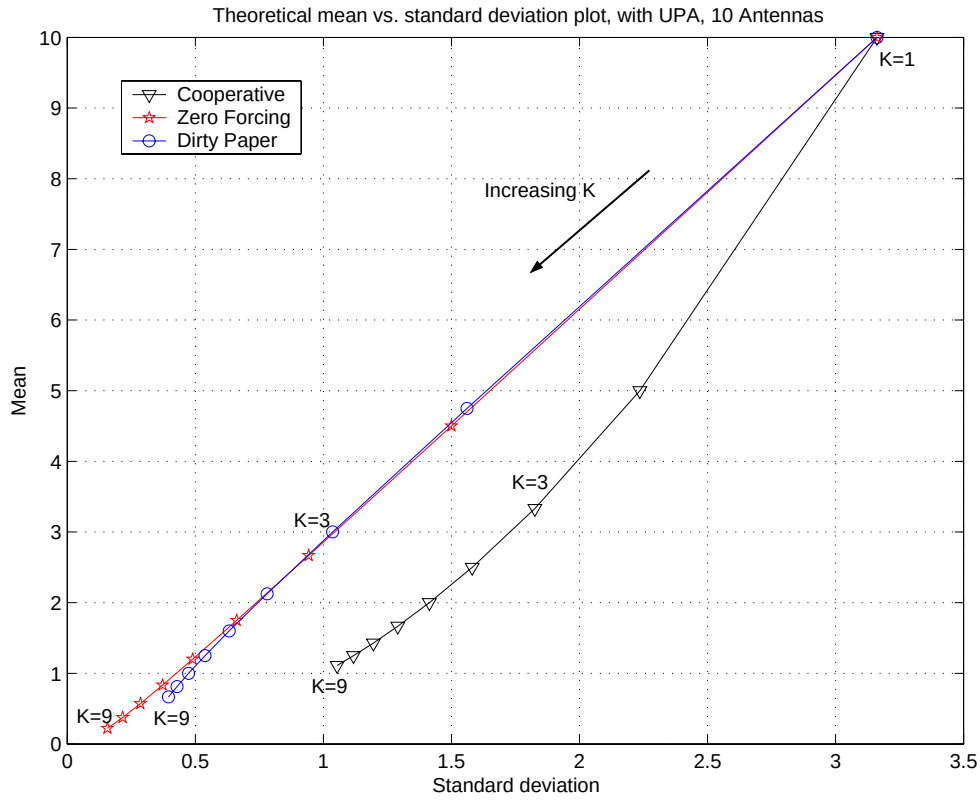


Figure 2.3: Mean vs. standard deviation plot for the Cooperative scheme, the Dirty Paper, and the Zero Forcing, with 10 antennas and varying number of users.

scheme is the most unfair multi-antenna strategy, it creates even higher differences among the users because it is penalizing the worse channel modes.

Unfortunately, this index of fairness does not give any information about the achievable performance of the techniques, thus the scheduler at the AP might not be able to decide on the best-suited technique to use according to the needs of the users. Therefore, the mean vs. standard deviation figures are shown next and their usefulness is discussed, either for the comparison of techniques or for the system designer. These curves are inspired by portfolio selection, so the mean and the standard deviation are plotted, although it shall be noted that the numerator and the denominator in (1.21) could have also been used, since one objective is to show that an index is not sufficient. The summary of the results in the previous section is found in Table 2.1, now taking K into account and without the SNR scaling γ^n. The preferred option would be the one having as low as possible standard deviation and as high as possible mean, but this might not possible to obtain in practice.

The mean vs. standard deviation in a system with $Q = 10$ antennas and varying number of users is plotted in Fig. 2.3, where each point (star, circle, or triangle) reflects a different number of users, starting from $K = 1$ at the upper right corner, to $K = 9$ at the lower left corner. This

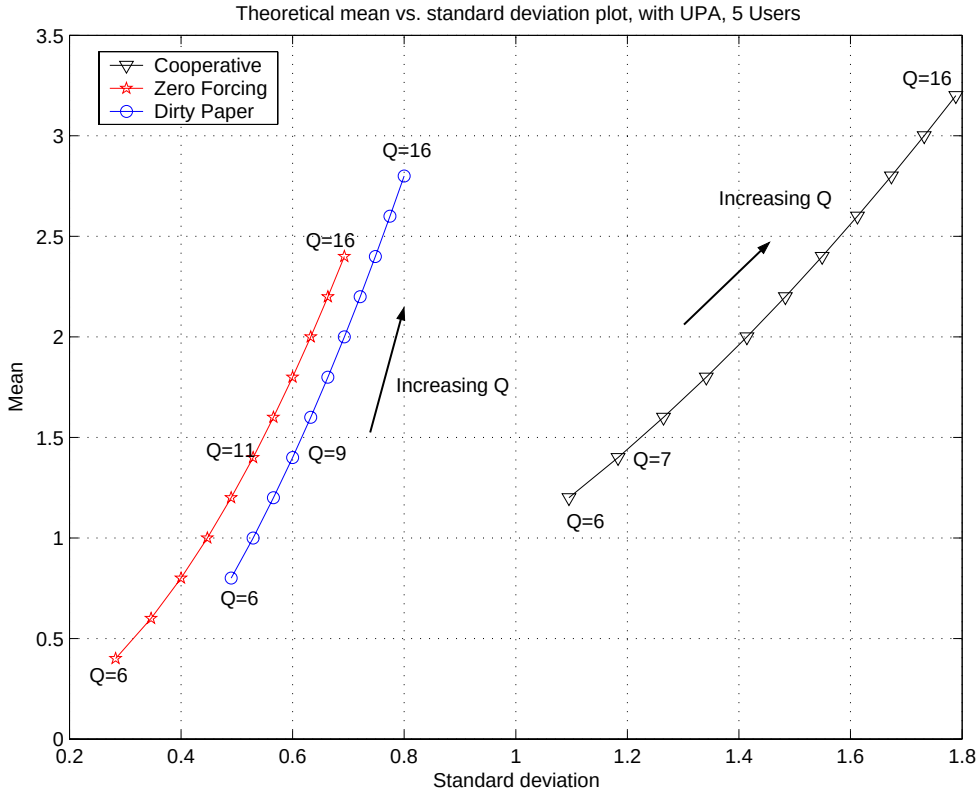


Figure 2.4: Mean vs. standard deviation plot for the Cooperative scheme, the Dirty Paper, and the Zero Forcing, with 5 users and varying number of antennas.

figure can be seen in two ways. First, for a fixed K, say 3, ZF yields the worst mean performance but has also a lower standard deviation. DP has a slightly higher mean, but also higher standard deviation, and the cooperative system increases slightly the mean at the expense of a much higher standard deviation. Consequently, a higher mean implies a higher variance for a fixed number of users, thus the choice at the AP may not be straightforward, and it seems to depend on the fairness criterion of the manufacturer or on the operator pricing policy. The second way of interpreting Fig. 2.3 is the following. Imagine one wants a mean of, say at least 2. With DP, $K = 4$ users should be allowed, whereas $K = 3$ with ZF, and $K = 5$ with the cooperative scheme. For an equal mean, the techniques involving more users yield a higher variance. This has an impact on the scheduler strategy, so it is part of the cross-layer issues in wireless.

In Fig. 2.4 each point (star, circle, or triangle) is a different number of antennas, and the mean vs. standard deviation is evaluated with fixed number of users, $K = 5$. A higher mean comes at the expense of a higher variance in the distribution of the resources, see e.g. the case where $Q = 6$. The same procedure as in the previous figure applies e.g. with a required mean slightly lower than 1.5. Now, the cooperative scheme requires less number of antennas ($Q = 7$) than DP ($Q = 9$) and ZF ($Q = 11$). It shall be noted here that a different selection of the

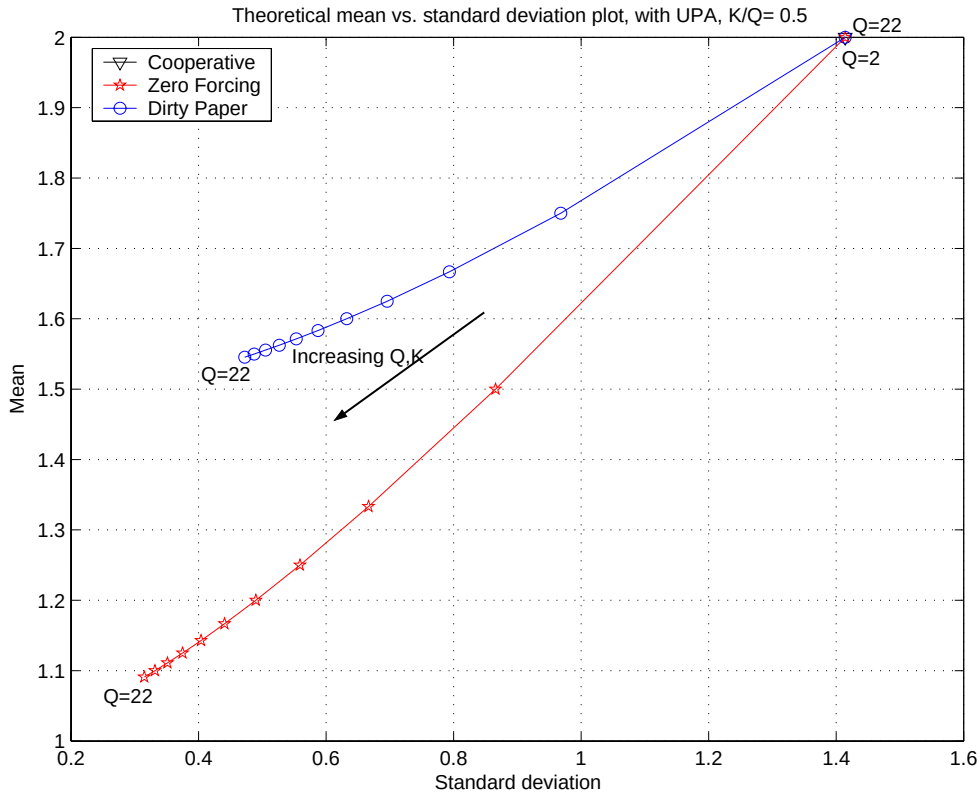


Figure 2.5: Mean vs. standard deviation plot for the Cooperative scheme, the Dirty Paper, and the Zero Forcing, with K/Q ratio fixed and varying number of antennas (and users).

power allocation technique might vary the shape and position of these curves in the plot, as it is observed in the next chapter. Finally, the mean vs. standard deviation trade-off is depicted in Fig. 2.5 when the ratio K/Q remains fixed and equal to 0.5 in this case, and both K and Q are increased without bound (actually, here until $Q = 22$). Each point (star, circle, or triangle) is a different value for Q, in steps of 2. Certainly, with one user, the performance coincides for the three schemes. According to Table 2.1, the performance of the cooperative scheme does not vary, and the same interpretation as in previous figures applies here. Additionally in Fig. 2.5, it is observed that increasing Q over a threshold ($Q = 14$) does not change significantly nor the mean nor the standard deviation, thus it might not make sense to increase Q over a certain value.

2.5 Conclusions

In this chapter, a new framework has been proposed to evaluate the fairness of multi-user communications in general, and of multi-antenna systems in particular, which completes the information provided by previously proposed indexes. It consists in analyzing the mean vs. the standard deviation trade-off at the receivers, which is inspired by portfolio selection and is a

new way of looking at previously proposed fairness indexes: not only the relative performance is important. Moreover, it confirms the fact that more mean comes at the expense of an *unfair* distribution of the resources (more variance). Although this framework could be deployed to analyze and even to design any type of resource allocation in multi-user communications, it has been particularized for three well-known schemes in a multi-antenna broadcast channel, namely the cooperative scheme, zero forcing, and dirty paper. It has been shown that the differences among the users not only exist with an explicit objective function which balances the assignments for the users, such as in the fairness definitions of the previous chapter, but also with implicit sharing cost functions such as the multi-antenna techniques.

The framework provides a powerful tool not only for the evaluation but also for the cross-layer design of multi-user systems. Moreover, it exemplifies that choices at the AP are not straightforward when fairness is taken into account. This involves not only the transmit antenna array processing as in this chapter, but also the power allocation (see following chapter) or any other degrees of freedom existing in multi-user systems, e.g. at the DLC layer. For these reasons, in the remainder of the dissertation, the zero forcing transmit beamforming is the selected technique because it provides a reasonable trade-off between performance (could be quite close to maximum sum capacity) and complexity with a low-complex closed-form solution. Moreover, it provides a valid framework to analyze fairness issues and cross-layer scheduling aspects, which are indeed the main focus of this dissertation.

Chapter 3

Power allocation and admission control

In this chapter, multiple antennas are used to enhance the performance of a centralized scheduler at the base station or access point. Since it has been stated in the previous chapter that the transmit beamforming technique is also affected by fairness issues, a Zero Forcing beamforming criterion is selected for the downlink. Moreover, several papers conclude that it is a scheme that provides appealing properties in a practical implementation, and it allows to study the fairness implications of the power allocation, as well as the admission control, which is a mechanism balancing between the PHY and the DLC in wireless scenarios.

Indeed, the problem that is addressed is how the multi-antenna AP distributes the scarce resource among the single-antenna terminals. In this case, the scarce resource is the limited output power, which is shared among several users and determines both their bit rate and Bit Error Rate (BER). Since there is a clear trade-off between the satisfaction of the individual needs and the global performance of the cell, several criteria are proposed, ranging from a classical physical layer point of view of capacity (rate) maximization to closer-to-DLC BER-based cost functions. Between two traditional techniques, namely the Uniform Power Allocation (UPA) and the Equal BER and Rate (ERB), a new one is proposed, which ultimately provides an intermediate performance. Furthermore, motivated by the extensive use of the game-theoretic formulation for the uplink power control in CDMA, a new strategy is studied, which is based on the widespread utility function used in the literature. Instead, the focus is on the downlink of a communication system, but it is shown that if the game-theoretic framework is selected, a convenient utility function shall be chosen. Nevertheless, pricing mechanisms are without doubt a useful mechanism in the design of wireless systems. Depending on the objectives of the scheduler, the best-suited technique would vary. As it is foreseen, it will be shown that the strategies maximizing the utility or the capacity (rate) lead to a higher error rate than the BER-based schemes. Furthermore, there is also a trade-off between the global performance and the individual needs in the power control.

The power allocation problem is extended by the addition of BER (or SNR) constraints, which is the admission control mechanism, usually a matter of the DLC, i.e. the AP shall decide which users within the cell can be served as well as both their rate and BER. Among traditional options, in this chapter a new mechanism is proposed to balance the above-mentioned trade-off between the total performance and the particular behavior. Throughout the chapter, the synergies among the physical layer and the DLC are exploited so as to improve the scheduling performance. With this interaction, mechanisms at the two layers can be improved. It is currently assumed that any system cannot be fully optimized if cross-layer issues are not taken into account [4], although physical layer problems have often forgotten the implications at higher layers. In this sense, the admission control is a traditional DLC mechanism and the fairness implications have been issues for higher layers.

This chapter is organized as follows after the introduction and problem statement that is presented next. In Section 3.2, several power allocation techniques are proposed for this multi-antenna multi-user system, and the author evaluates the implications of the power allocation within the scheduling at both the PHY and the DLC, especially in terms of fairness. Then, section 3.3 introduces a new power allocation strategy to better balance the trade-off between the global performance and the individual needs. After that, in Section 3.4, the power allocation techniques minimizing the sum BER or maximizing the sum rate are compared to the maximization of the sum of utilities, which resorts to game theory. Finally, in Section 3.5, the admission control mechanisms are compared, just before the final conclusions of this chapter.

3.1 Introduction

As in the whole dissertation, this chapter deals with the simultaneous downlink communication of a multi-antenna BS or AP with single antenna terminals. In this case, the spatial diversity is used to enhance the scheduling, which consists in the assignment of a certain rate to all the users, or in the denial of service if some minimum requirements cannot be fulfilled, i.e. the admission control mechanism. Basically, the scheduling consists of dividing the limited available resources among the active users. Instead of the link bandwidth, the scarce resource is the instantaneous output power, which is usually specified by regulatory authorities. For instance, in the upper band dedicated to 5 GHz wireless local area networks in Europe, the maximum instantaneous output power is 30 dBm. To solve the problem, the BS has to cope with the channel variations, the moving nature of the users, and the usually scarce battery life, among others [99].

The solution of the scheduling at the physical layer is conceptually divided into two steps, namely the transmit beamforming and the power allocation. As in [79], terminals are dumb, so that all the intelligence is located at the multi-antenna AP. The AP performs the transmit beamforming, so that the symbols the terminals receive are only corrupted by noise and not

by the signals transmitted for the other users, which differs from the approach taken in e.g. [92]. Then, terminals shall not perform any filtering, their computational load is reduced, thus their battery life is increased. Besides, they do not need to be aware of the channels from the other users. This seems a well-suited implementation for the Spatial Division Multiple Access (SDMA), since the resources granted for the users do not overlap. Essentially, the same idea holds for other schemes such as Time Division Multiple Access (TDMA) or Frequency Division Multiple Access (FDMA). The beamforming criterion that matches those requirements is ZF, which provides a low-complex closed-form solution to create parallel and orthogonal equivalent channels for the users that are being served, without inner-cell interference [84]. As it has been shown, compared to optimum downlink beamforming [46], the main powerful characteristics are the low complexity and the closed-form solution. Indeed, these are characteristics the DLC layer demands for a real-time scheduler. In fact, in this chapter it is shown that some functions of the scheduling task, which is traditionally part of the DLC, could also be performed by the physical layer. In this sense, simplicity is a key feature.

Once the spatial architecture has been established, the AP has to distribute the available power among the users. Since there is a clear tradeoff between the satisfaction of the individual needs and the global performance of the cell [129], the fairness criterion determines the cost function of the problem. On the one hand, it is well-known that optimizing the global performance implies an asymmetric distribution of the resources, i.e. some are given more than others. On the other hand, max-min or min-max schemes [127] distribute the resources equally, i.e. the gains are cell-wide at the expense of loosing in global performance. In this chapter, as well as in the whole dissertation, the author concentrates on the instantaneous fairness, i.e. the implications of the resource allocation in the short-term. This viewpoint is especially suited for applications with hard delay constraints or bursty transmissions. It is concluded here that the choices at the BS or AP are, at least, not straightforward concerning the power allocation.

One of the aspects studied in this chapter is the game-theoretic power control, which has been widely studied in the literature not only in the context of Code Division Multiple Access (CDMA), see e.g. and [107] and references therein, but also for digital subscriber lines [108]. Some advantages of the game-theoretic formulation for the uplink are its easy scalability and the fairness guarantees among the users, since they are always granted their maximum satisfaction. This degree of satisfaction is expressed mathematically in terms of a convenient utility function, which is a key issue. If data is transmitted, the utility should be increasing with respect to the Signal to Interference Ratio (SIR) if the transmit power is fixed, or it should be decreasing with power if the SIR is kept constant, among other properties [103]. Therefore, it is sensible to use a ratio between the Frame Success Rate (FSR), that is, the probability that the frame is correct, and the transmitted power, as the authors suggest [105]. It is shown in this chapter that although the utility-based optimization certainly maximizes this metric within the cell, while

minimizing the power, the FSR is penalized, or equivalently, the BER is higher than for other methods. This result reflects the difficulty in choosing convenient utility functions.

Even with the optimal power allocation, it might happen that all the active users could not be scheduled. If only a subset can be served simultaneously, the multi-antenna BS has to decide which users are selected, that is the Spatial Admission Control (SAC) mechanism [147]. Besides, the SAC shall fulfill the QoS requirements of the users, e.g. in terms of BER. In the proposed SDMA system, the interactions among the users play a very important role and thus deeply impact on the selection of the users in the optimization of a certain criterion. Indeed, the performance varies significantly depending on the users that are being served, because more power is required if users with correlated channels are scheduled, as shown in the previous chapter.

3.2 Power allocation techniques

In this section, several alternatives are proposed to allocate the total available power P_T to the users, which are assumed to be homogeneous, i.e. their requirements are the same in terms of delay, throughput, and BER. Given a number of users, the BS tries to obtain the best resource sharing. Nevertheless, the BS does not optimize the number of users nor selects the best users to serve (see Section 3.5 for details on these issues). Besides, a single mapping is available, for details on adaptive modulation, see next chapters. In the downlink, cost functions usually aim at optimizing the BER or rate while constraining the power, although other meaningful options minimize the total power subject to a BER or rate constraint [148].

Although the analysis that is presented in this section might also hold for DPC techniques, ZF is taken as the beamforming technique, thus the applied signal model is (2.4), namely

$$\mathbf{y} = \mathbf{D}_\alpha \mathbf{D}_\beta \mathbf{s} + \mathbf{w} \Rightarrow y_k = \alpha_k \beta_k s_k + w_k,$$

in which the equivalent channels α_k are affected by the channels of the other users by means of the inverse of the matrix $\mathbf{H}\mathbf{H}^H$, as it has already been shown. The SNR for the kth user is

$$\gamma_k = \frac{\alpha_k^2 \beta_k^2}{\sigma^2},$$

where it has been assumed that the symbols have unitary mean energy, particularly, normalized QAM symbols, and it is considered without loss of generality that the noise power σ^2 is equal for the K users. Since simplicity is an important feature for schedulers, it is necessary to use an easy-differentiable BER expression. Therefore, it is meaningful to use the approximate BER for QAM signals in Rayleigh fading channels corrupted only by AWGN given in [149], i.e.

$$\mathrm{BER}(\gamma) \approx c_1 \exp(-c_2 \gamma), \tag{3.1}$$

where $c_1 = 0.2$, $c_2 = \frac{1.6}{2^m - 1}$, and m is the number of bits in the constellation, fixed for the purpose of this chapter. This expression is valid within 1.5 dB of error for a BER $\leq 10^{-3}$. Since it is

assumed that there is no channel coding at the transmitter, the channel is time-invariant, and the noise is Gaussian, the Packet Error Rate (PER) can be expressed as a function of the BER and the packet length L in bits as $\mathrm{PER} = 1 - (1 - \mathrm{BER})^L$. If several signal mappings are available, the throughput (rate) depends on the number of bits of the symbols and on the BER. At the physical layer, rate is the maximum number of bits per symbol m that can be transmitted while fulfilling a target BER, BER_t, and it can be obtained using (3.1) as

$$m = \log_2\left(1 + \frac{\gamma}{\Gamma}\right), \tag{3.2}$$

where the constant Γ is given by $\Gamma = \frac{\log(c_1/\mathrm{BER}_t)}{c_2}$. In fact, $\Gamma = 1$ can be interpreted as the classical Shannon's limit to error-free bit rate (capacity) [150]. Typically, m is a real number, and an spatial waterfilling can be performed in order to achieve the maximum sum rate of the SDMA channel, as it is shown in the next section. Other schemes may provide a lower sum rate but ensure a more fair resource distribution. In practical systems only a finite set of mappings is usually available, thus m is an integer, see next chapter for details on these issues.

3.2.1 Uniform Power Allocation (UPA)

Without any channel knowledge, the best option reduces to the well-known UPA. In that case, the AP divides equally the whole power among the active users in the cell, so that it does not care about their actual channel gain nor how the performance might be improved. Basically, the power allocated to the kth user is

$$\beta_k^2 = \frac{P_T}{K},$$

thus the SNR for the kth user is given by

$$\gamma_k = \frac{P_T}{\sigma^2}\frac{\alpha_k^2}{K}, \tag{3.3}$$

which leads to a higher rate (and lower BER) for the users having a better channel. Since it is assumed that the BS has perfect channel knowledge, a more efficient power allocation criterion might be applied. In fact, the fairness criterion determines the power allocation. A first option assigns the same amount of resource to the users, which is translated into the same BER and rate. However, other fairness considerations state that the users with a higher mean SNR during a certain time window shall be provided a higher rate than those having a poorer link quality[8]. Then, a second option optimizes the global performance regardless of the users with worse channel conditions, including the maximum sum rate technique and the strategy minimizing the sum BER that are presented next. This SDMA scheme differs from opportunistic communications such as [151], since several users share the spatial channel and not only the one with a proportionally

[8] This might also be applied if the price of the service varies depending on the desired QoS.

better channel. A last remark is that the techniques that are presented can be considered as a best-effort type of service, since the BS optimizes the BER or the rate regardless of the individual QoS achieved by the active users.

3.2.2 Equal Rate and BER (ERB)

A possible optimization criterion consists of assigning the same rate and BER to all users, regardless of their channel quality. It will be seen in this subsection that it finally reduces to assigning the same BER to all users. The cost function might be expressed as

$$\begin{aligned} &\min_{\beta_k^2} \max_k \mathrm{BER}_k \\ &s.t. \sum_{k\in\mathcal{K}} \beta_k^2 \leq P_T, \end{aligned}$$

where it has been implicitly assumed that the β_k^2 are non-negative, since they are power allocation factors. Note also that this problem is in fact the same as a max-min SNR or max-min rate strategy. However, in this chapter it is solved with the previous formulation, which is a convex problem because the BER approximation is an exponential and the constraints are linear, see [31]. In order to properly solve this optimization, one should express it according to the convex formulation. Recalling (3.1), the *convex way* is

$$\begin{aligned} &\min_{\beta_k^2} \ t \\ &s.t. \sum_{k\in\mathcal{K}} \beta_k^2 - P_T \leq 0, \\ &\quad c_1 \exp(-c_2 \frac{\alpha_k^2 \beta_k^2}{\sigma^2}) - t \leq 0, \forall k \in \mathcal{K}, \\ &\quad -\beta_k^2 \leq 0, \forall k \in \mathcal{K}, \end{aligned}$$

to which the Karush-Kuhn-Tucker (KKT) conditions apply [31]. Using these and the equality

$$\sum_{k\in\mathcal{K}} \frac{1}{\alpha_k^2} = tr\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right],$$

one can find the following solution

$$\alpha_k^2 \beta_k^2 = \frac{P_T}{tr\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right]},$$

so that the SNR is the same for the K users, i.e. $\gamma_k = \gamma, \forall k \in \mathcal{K}$, and it is given by

$$\gamma = \frac{P_T}{\sigma^2} \frac{1}{tr\,(\mathbf{H}\mathbf{H}^H)^{-1}}. \tag{3.4}$$

With this technique, a high amount of power is used for the users that have a poor channel quality, which degrades the performance of the better users, thus the global performance of the

cell. As it is shown next, the ERB always achieves a lower sum rate than the UPA, but it has the powerful property that all the users are granted the same service. Indeed, it penalizes the global performance for the sake of the individual revenues, which might be considered as a fair power allocation.

Taking the approximation of a high SNR, it can be easily proven that **the ERB achieves always a lower sum rate than the UPA**. In the high SNR regime, the sum rate of the ERB R_T^{ERB} can be expressed as

$$R_T^{ERB} \approx K \log\left(\frac{P_T}{\sigma^2}\frac{1}{tr\left(\mathbf{H}\mathbf{H}^H\right)^{-1}}\right) = K \log\left(\frac{P_T}{\sigma^2}\frac{1}{\sum_{k\in\mathcal{K}}\frac{1}{\alpha_k^2}}\right) = CT + K \log H_\alpha,$$

where H_α is the harmonic mean of the α_k^2 and CT is a constant, i.e.

$$\frac{K}{H_\alpha} = \sum_{k\in\mathcal{K}}\frac{1}{\alpha_k^2},$$
$$CT = \sum_{k\in\mathcal{K}} \log\left(\frac{P_T}{K\sigma^2}\right) = K \log\left(\frac{P_T}{K\sigma^2}\right).$$

On the other hand, at high SNR the sum rate of the UPA R_T^{UPA} can be approximated as

$$R_T^{UPA} \approx \sum_{k\in\mathcal{K}} \log\left(\frac{P_T}{\sigma^2}\frac{\alpha_k^2}{K}\right) = CT + \log\prod_{k\in\mathcal{K}}\alpha_k^2 = CT + K\log G_\alpha,$$

where G_α is the geometric mean of the α_k^2, i.e.

$$G_\alpha = \left(\prod_{k\in\mathcal{K}}\alpha_k^2\right)^{1/K},$$

which means that at high SNR the sum rate of the UPA (R_T^{UPA}) is always greater or equal to the sum rate of the ERB (R_T^{ERB}). This comes from the fact that the geometric mean is always greater or equal than the harmonic mean, i.e.

$$G_\alpha \geq H_\alpha \Rightarrow R_T^{UPA} \geq R_T^{ERB},$$

in which equality holds when the α_k are all the same. This will unlikely occur in this case, because of the distributed nature of the users.

3.2.3 Maximum Sum Rate (MSR)

Instead of guaranteeing the same SNR for all users, thus BER and rate, another option is to maximize the sum rate of the cell, without considering a possibly uneven resource partitioning. Some users might not be scheduled, allowing others to have a higher rate and a lower BER. It is important to note that since perfect information is available at the AP, not only the best user

will be scheduled for transmission, but rather a subset of all the terminals that are active. The AP penalizes the users with poorer channels, and increase the performance of the better users, thus the global performance of the cell. This cost function is expressed as

$$\max_{\beta_k^2} \sum_{k\in\mathcal{K}} m_k \tag{3.5}$$

$$s.t. \sum_{k\in\mathcal{K}} \beta_k^2 \leq P_T, \tag{3.6}$$

where m_k is a real number reflecting the maximum achievable rate for the kth user, recall (3.2). After the application of the KKT conditions to this problem, the following modified water-filling algorithm is obtained:

$$\beta_k^2 = \left(\mu^{-1} - \frac{\sigma^2}{\alpha_k^2}\right)^+, \tag{3.7}$$

where μ is chosen to satisfy the power constraint in (3.6) with equality. Besides, this technique is explicitly designed for the maximization of the sum rate, and thus always outperforms the ERB in those terms, as it is shown later in this subsection. The SNR for the kth user is then

$$\gamma_k = \left[\left(\mu^{-1} - \frac{\sigma^2}{\alpha_k^2}\right)\frac{\alpha_k^2}{\sigma^2}\right]^+.$$

It is shown that **the MSR is equivalent to the UPA if the SNR is high**. If σ^2 is low, then $\beta_k^2 \approx \mu^{-1}$, so by applying the constraint in (3.6), the power allocation reduces to $\beta_k^2 = \frac{P_T}{K}$. Approaching the problem from another perspective, if the number of users K is low, or if the SNR is very high, all of them might be allocated for transmission. Under these circumstances, it is easy to verify that the power factors tend to be those from UPA using a result in [55]: if Q and K are let grow without bound, but their ratio $\zeta = Q/K$ remains fixed, then [55]

$$\lim_{K,Q\to\infty} \mathbb{E}\left\{\frac{1}{\alpha_k^2}\right\} = \frac{1}{Q-K}. \tag{3.8}$$

Introducing (3.7) into (3.6) to obtain μ^{-1} using the expression in (3.8), and then substituting μ^{-1} back into (3.7), the power allocation factors reduce to $\mathbb{E}\left\{\beta_k^2\right\} = \frac{P_T}{K}$. By the use of two approximations and a large matrix result, it has been shown that at high SNR, the MSR tends to the UPA. Besides, since it was proven in the previous section that the UPA outperforms the ERB in terms of sum rate, the MSR provides also a higher sum rate than the ERB, as it is obvious due to the objective function. However, the distribution of the resources is asymmetric and some are given more than others.

The implementation of this waterfilling procedure is detailed in Table 3.1. First, the scheduler tries to serve all the users that are active, steps 1 and 2, and then calculates the equivalent channels after the beamforming α_k (step 3), after that it computes the water level for that

1. Set $\mathcal{K} = \{1, \ldots, K\}$.
2. Build matrix $\mathbf{H}$ for the users in the set $\mathcal{K}$.
3. Compute $\alpha_k^2 = 1/\left[\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\right]_{k,k}, \forall k \in \mathcal{K}$.
4. Compute $\mu^{-1} = \frac{P_T + \sigma^2 \sum_{k\in\mathcal{K}} \alpha_k^{-2}}{|\mathcal{K}|}$.
5. Compute the power allocation factors as in (3.7).
6. If $\beta_k^2 > 0, \forall k \in \mathcal{K}$, then the algorithm finishes.
   Otherwise, remove the worst user having zero power $\mathcal{K} \leftarrow \mathcal{K} - \left\{k \in \mathcal{K} : \min_k \alpha_k^2 \text{ and } \beta_k^2 = 0\right\}$, and go to step 2.

Table 3.1: Spatial waterfilling algorithm.

configuration (step 4). With it, it obtains the power allocation factors in step 5. If some users cannot reach the water level, the worst user is removed from the set of active users (step 6). It is important to note here that, since the equivalent spatial channels change depending on the subset of users, it is not possible to remove all users having zero power. At this point, it is essential that the equivalent channels α_k^2 are recomputed (step 2) since they depend on the users that are being served. This procedure is repeated until all the users in the set of active users are assigned a non-zero power (finishing condition in step 6).

3.2.4 Minimum Sum BER (MSB)

Instead of rate-based methods, another possibility is to minimize the total BER. In fact, with a single constellation, since a lower BER implies a higher throughput, this option might be preferred in real systems. Besides, it provides a direct link to the DLC, due to the relation between the BER and the PER. With this minimum sum BER technique, the objective is to minimize the sum BER of all the users in the cell subject to the power constraint, i.e.

$$\begin{aligned} &\min_{\beta_k^2} \sum_{k\in\mathcal{K}} \mathrm{BER}_k \\ &s.t. \sum_{k\in\mathcal{K}} \beta_k^2 \leq P_T, \end{aligned}$$

which in convex formulation it can be expressed as

$$\min_{\beta_k^2} \sum_{k\in\mathcal{K}} \mathrm{BER}_k \tag{3.9}$$

$$s.t. \sum_{k\in\mathcal{K}} \beta_k^2 - P_T \leq 0, \tag{3.10}$$

$$-\beta_k^2 \leq 0, \forall k \in \mathcal{K}. \tag{3.11}$$

The KKT conditions [31] can be applied because the problem is convex, and one can see that the solution is similar to a waterfilling:

$$\beta_k^2 = \frac{\sigma^2}{c_2\alpha_k^2}\left[\log\left(\frac{c_1c_2\alpha_k^2}{\sigma^2}\right) - \log\mu\right]^+, \forall k \in \mathcal{K}, \tag{3.12}$$

where $\log\mu$ is obtained in order to fulfill (3.10) with equality. The same remark about implementation of the MSR shall be made here, i.e. since the α_k^2 change (increase) when the number of users is reduced, they shall be recomputed if there is any user j for which $\beta_j^2 = 0$. Then, user j is removed from the active set $\mathcal{K}$, thus the jth row is eliminated from $\mathbf{H}$. Therefore, the solution in (3.12) shall be computed again. Since this scheme can be seen again as a modified waterfilling, it is implemented in a similar way as the procedure in Table 3.1, thus the algorithmic implementation is not explicitly detailed.

By construction, it is clear that this scheme will provide a lower BER than the ERB, but the drawback is that for the sake of the collective revenue, some users might not even be allocated for transmission. Similarly to the MSR, the packets in the queues from the not-allocated users will be either lost or deferred depending on the delay constraints of the application. It is important to note that **the performance of the MSB in terms of BER tends to that of the ERB at high SNR**, as it is proven next. This might have computational implications, since the MSB is more complex than the ERB. To achieve the goal, note that if all the users in the active set $\mathcal{K}$ are served, $\log\mu$ in (3.12) can be calculated as

$$\log\mu = \frac{\sum_{j\in\mathcal{K}}\frac{\sigma^2}{c_2\alpha_j^2}\log\left(c_1c_2\frac{\alpha_j^2}{\sigma^2}\right) - P_T}{\sum_{j\in\mathcal{K}}\frac{\sigma^2}{c_2\alpha_j^2}},$$

so at high SNR, the first term in the numerator tends to zero, and the power allocation factors can then be approximated as

$$\beta_k^2 \approx \frac{\sigma^2}{c_2\alpha_k^2}\left(\log\left(c_1c_2\frac{\alpha_k^2}{\sigma^2}\right) + \frac{P_T}{\sum_{j\in\mathcal{K}}\frac{\sigma^2}{c_2\alpha_j^2}}\right).$$

At a high SNR, the linear term grows faster than the logarithmic term. Therefore, it can be assumed that the first term in the addition can be disregarded at high SNR, so these power allocation factors finally reduce to

$$\beta_k^2 \approx \frac{\sigma^2}{c_2\alpha_k^2}\frac{P_T}{\sum_{j\in\mathcal{K}}\frac{\sigma^2}{c_2\alpha_j^2}} = \frac{P_T}{\alpha_k^2}\frac{1}{\sum_{j\in\mathcal{K}}\frac{1}{\alpha_j^2}} = \frac{P_T}{\alpha_k^2}\frac{1}{tr\left(\mathbf{H}\mathbf{H}^H\right)^{-1}}, \tag{3.13}$$

which leads to a SNR for the kth user given by

$$\gamma_k = \frac{\alpha_k^2\beta_k^2}{\sigma^2} = \frac{P_T}{\sigma^2}\frac{1}{tr\left(\mathbf{H}\mathbf{H}^H\right)^{-1}},$$

which is the same as in (3.4) for the ERB. Both BER methods tend to attain the same performance in terms of BER at high SNR, but note that this convergence does not occur with the rate-based schemes. However, note that the approximation in (3.13) is rather strict, and the convergence in behavior with respect to the SNR might be slow for the BER-based techniques. It can be also shown that **an upper bound of the sum BER is minimized by the UPA**, which might occur if the AP has no knowledge about the channel or if the quality of the estimation is very low. In fact, in that case the UPA is the best it can be done. Essentially, the equivalent channel gain is lower- and upper-bounded by

$$\frac{1}{\lambda_{max}} \leq \alpha_k^2 \leq \frac{1}{\lambda_{min}},$$

where λ_i denotes the ith eigenvalue of the matrix $\left(\mathbf{H}\mathbf{H}^H\right)^{-1}$. If one assumes that all the users are transmitting the same modulation, the SNR for each user k in the system is bounded by

$$\gamma_k = \frac{\alpha_k^2 \beta_k^2}{\sigma^2} \geq \frac{\beta_k^2}{\lambda_{max}\sigma^2},$$

so that the lower bound on the SNR is translated into an upper bound for the BER. Then, if this upper bound on the sum BER is minimized, the real sum BER is also minimized. The Lagrangian of this modified problem using (3.1) is expressed as

$$\mathcal{L} = \sum_{k\in\mathcal{K}} c_1 \exp\left(-c_2 \frac{\beta_k^2}{\lambda_{max}\sigma^2}\right) - \mu\left(P_T - \sum_{k\in\mathcal{K}} \beta_k^2\right).$$

where it can be seen after some algebra that taking derivatives of $\mathcal{L}$ with respect to the power allocation factors β_k^2 and to the Lagrange multiplier μ, the optimum power factors are given by those from the UPA, i.e. $\beta_k^2 = \frac{P_T}{K}$.

3.2.5 Simulation results

Up to this point, several traditional power allocation techniques have been presented, and it has been theoretically shown that asymptotically in the **high SNR regime**,

1. the sum rate of the ERB is always lower than that of the UPA;
2. the sum rate of the MSR tends to be the same as that of the UPA;
3. the performance in terms of BER of the BER-based techniques, namely the MSB and the ERB, tends to be the same.

In this section, these results will be shown through simulations, as well as a new perspective to show the fairness of the proposed power allocation techniques will be given. Indeed, it is important to see graphically the behavior of the proposed strategies.

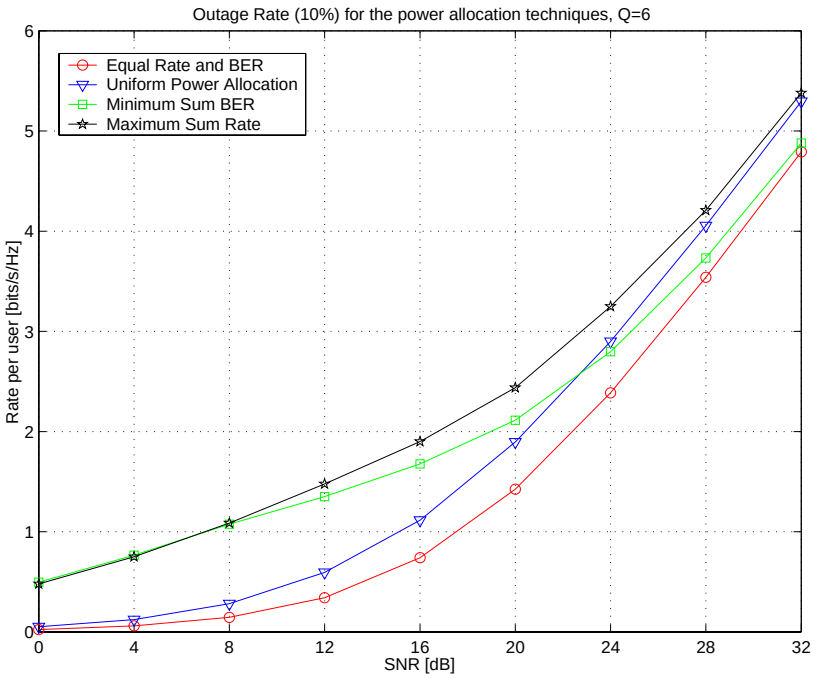


Figure 3.1: For the proposed power allocation techniques, the figure shows the outage mean rate at 10% vs. SNR. The equivalences at high and low SNR are clear.

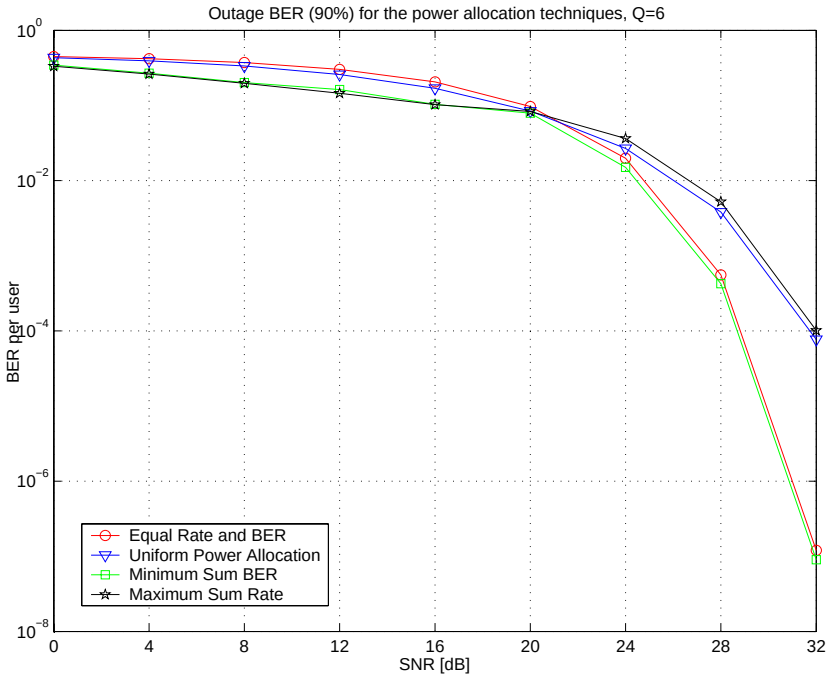


Figure 3.2: For the proposed power allocation techniques, the figure shows the outage mean BER at 90% vs. SNR. The equivalences at high and low SNR are clear.

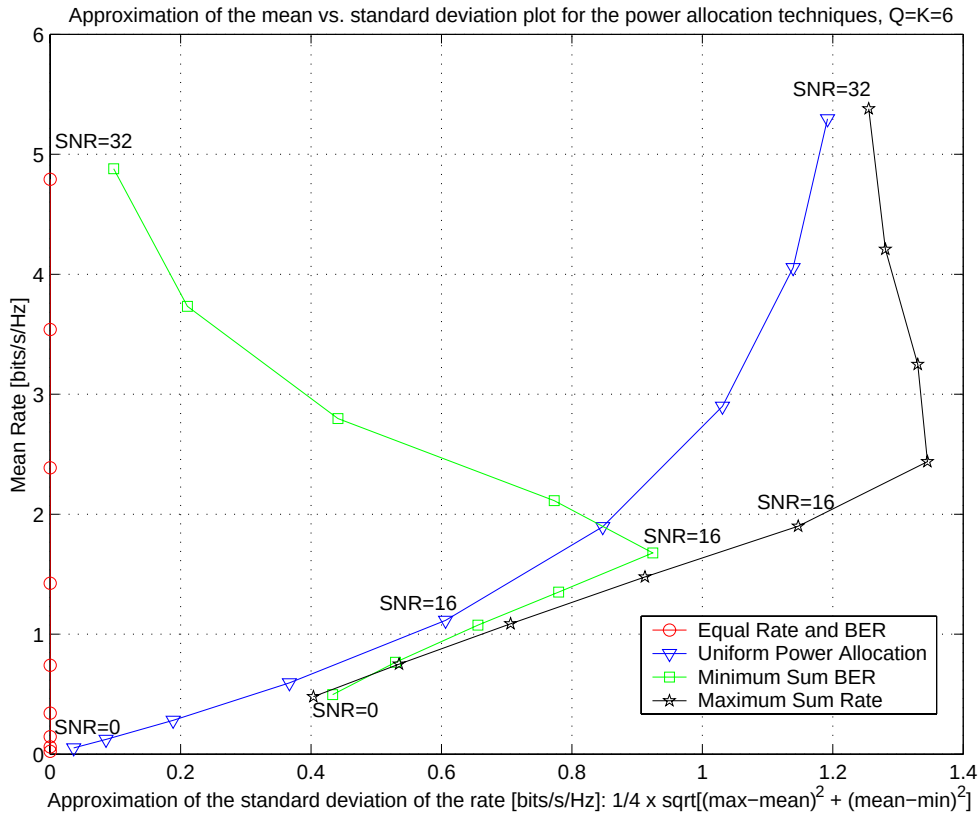


Figure 3.3: For the proposed power allocation techniques, the figure shows the outage mean rate vs. the approximation of the standard deviation at 90% SNR.

An AP provided with $Q = 6$ antennas is the transmitter, and the SNR in the figures refers to the ratio $\frac{P_T}{\sigma^2}$. In order to compare the power allocation strategies, the number of users is equal to the number of antennas, i.e. $K = Q$, and the gap in (3.2) is set to $\Gamma = 1$ because these methods do not take into account the BER constraints for the moment. The signal mapping that has been assumed is QPSK without loss of generality. In Figure 3.1, one finds the outage rate at 10% vs. the SNR. With an outage rate of R at x% it is meant that x% of the time the rate is below R, or equivalently, that a minimum rate of R is guaranteed 100-x% of the time. Conversely for the BER, the outage is usually 90%, which means that the BER is 90% of the time below the plotted value. On the other hand, Figure 3.2 compares the proposed techniques in terms of BER vs. the SNR. Several observations can be made from the two figures. As stated, the ERB and the MSB tend to achieve the same performance both in terms of BER and rate in the high SNR regime, whereas the UPA and the MSR tend to obtain the same average performance at high SNR. As expected, the rate is maximized by the MSR, whereas the BER is minimized by the MSB. Another interesting comparison is at the low SNR regime. There, the ERB and the UPA achieve very similar performance both in terms of BER and rate. On the other hand, the MSR and the MSB are equivalent at low SNR, since it is in that region where the number of served

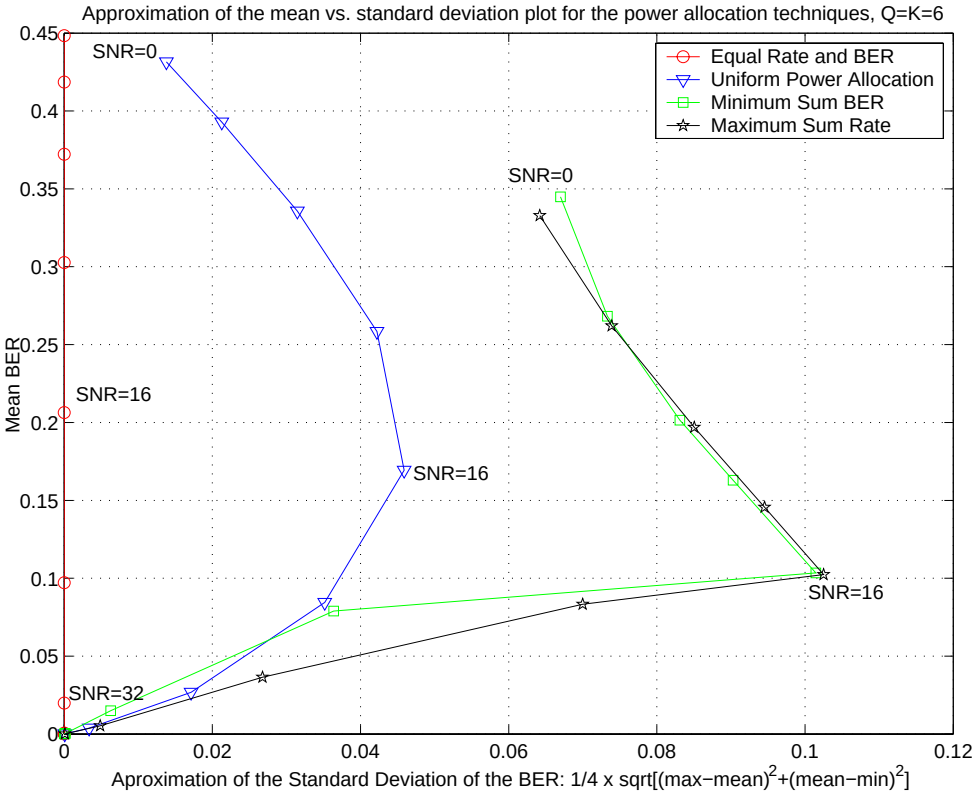


Figure 3.4: For the proposed power allocation techniques, the figure shows the outage mean BER vs. the approximation of the standard deviation at 90% SNR.

users is very low, typically at low SNR only the best user shall be scheduled for transmission. As a final remark, it should be noted that if bit allocation strategies were used in this case, the BER methods would always choose robust constellations because they obtain the lowest BER, whereas the methods based on rate tend to choose the constellations with a higher number of bits since the highest rate is then achieved. For further details on bit allocation strategies, please refer to Chapter 4. The figures that have been shown up to now are traditional in the literature, however, they do not reflect how the resources are distributed among the users, which is a key part of the dissertation.

Among other options, the preferred plots are equivalent to the fairness framework proposed in the previous chapter, that is, the mean vs. an approximation of the standard deviation for each of the metrics (rate and BER). First, Figure 3.3 shows the mean rate vs. the standard deviation, and Figure 3.4 is devoted to the same plots referring to the BER. It shall be noted here that each point in the figures refers to a SNR point, ranging from 0 dB to 32 dB in steps of 4 dB, so there is a total of 9 points per method. Moreover, the ERB curve is denoted by circles, the MSB by squares, the MSR by stars, and the UPA by triangles. In both Figure 3.3 and Figure 3.4, it can be stated that the ERB provides an equal performance for all the users, since it is a line at the

coordinate axis. The equivalences in terms of rate at high are also clear: the MSB tends to the ERB at high SNR, and the UPA converges in rate to the MSR. An interesting performance plot is that of the MSB, since at low SNR it tends to the MSR, whereas at high SNR it approaches the rate of the ERB. This behavior has been already identified in the previous comparisons, but it is rather interesting to reflect it in a figure. Regarding the BER plot in Figure 3.4, the results are not so clear as for the rate metric, although the same comments as before shall be made. Here, one does not distinguish so well the behavior of the MSB. Finally, the methods that have a better global performance tend to distribute the resources in an uneven way in both figures, see e.g. the MSR in Figure 3.3. For each SNR point, the mean performance is the best, but the variance among the users is also higher than for the other methods.

3.3 The fair balance: the Equal Proportional SNR (EPS)

From the previous results, the DLC might select the ERB and the UPA for simplicity, in order to deal with other mechanisms such as the admission control. However, before going into deeper details on these issues, a new method providing an intermediate performance among them shall be described, the EPS. In fact, the EPS is both a power allocation and an admission control mechanism, see Section 3.5. A first presentation and comparison with respect to the UPA and the ERB is needed in order to show the benefits of such a technique in terms of fairness. The EPS is based on the fact that the users might agree to loose the same proportion δ_k of their maximum achievable SNR, γ_k^a, which is obtained as if they were served alone in the cell, i.e.

$$\gamma_k^a = \gamma^n \|\mathbf{h}_k\|^2,$$

where the channel $\mathbf{h}_k$ is the kth row of the complete matrix $\mathbf{H}$. In fact, δ_k can be seen as the price paid for the collective satisfaction and could be computed according to the traffic requirements. If the terminals belong to the same network, e.g. at home, this might be a criterion to determine the access to the core, and it can be classified into a new metric for fairness. Mathematically, the fraction of the maximum achievable SNR is given by

$$\delta_k = \frac{\gamma_k}{\gamma_k^a}, \forall k \in \mathcal{K}.$$

If all the users are homogeneous and allow the same loss in proportion to their maximum SNR, i.e. $\delta_k = \delta, \forall k \in \mathcal{K}$, the cost function of this problem is expressed as

$$\begin{aligned} &\max \delta \\ &s.t. \sum_{k \in \mathcal{K}} \beta_k^2 \leq P_T, \end{aligned}$$

which has the nice property of yielding a closed-form solution for δ,

$$\delta^{-1} = \sum_{k \in \mathcal{K}} \frac{\|\mathbf{h}_k\|^2}{\alpha_k^2}, \tag{3.14}$$

but ultimately attains a different SNR for each user given by

$$\gamma_k^{EPS} = \delta\gamma_k^a = \gamma^n \delta \|\mathbf{h}_k\|^2 = \gamma^n \frac{\|\mathbf{h}_k\|^2}{\delta^{-1}}. \tag{3.15}$$

The suitability of this strategy is shown next by means of a fairness analysis of the UPA, the ERB, and the newly proposed EPS.

3.3.1 Fairness analysis of the UPA, the ERB, and the EPS

Before proceeding further, the authors wishes to recall the SNR after each of the proposed power allocation techniques that will be here analyzed. First, the UPA yields

$$\gamma_k^{UPA} = \gamma^n \frac{\alpha_k^2}{K},$$

whereas the ERB achieves an equal SNR for all the active users given by

$$\gamma_k^{ERB} = \gamma^n \frac{1}{tr\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right]},$$

whereas the EPS is more recent and can be found just before in (3.15). In the analysis, the term γ^n is disregarded since it is a common factor of all the techniques. The objective is to plot a figure similar to that showing the measure of inequality called the Gini index, therefore, a maximum, mean, and minimum analysis among the users is required. In fact, it is not the exact plot showing the percentage of the resource with respect the percentage of the population, but rather a decomposition into the main three groups.

Analysis of the mean

In this subsection, an analysis of the mean SNR is conducted, and it is assumed that the cardinality of the set $\mathcal{K}$ is K. For the **UPA**, since the α_k^2 behave like central Chi-squared random variables with $2(Q-K+1)$ degrees of freedom, where each random variable has variance 1/2 i.e. $\alpha_k^2 \sim \chi^2_{2(Q-K+1)}$, the mean value is given by

$$\mathbb{E}\left\{\frac{\alpha_k^2}{K}\right\} = \frac{Q-K+1}{K}. \tag{3.16}$$

Using a result of large random matrices, it has been already shown that for the **ERB**

$$\mathbb{E}\left\{\frac{1}{tr\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right]}\right\} = \frac{Q-K}{K}, \tag{3.17}$$

whereas for the EPS, a first analysis is needed on the value δ^{-1}. Using properties of block matrices and simple algebra, it can be verified that

$$\frac{\|\mathbf{h}_k^2\|}{\alpha_k^2} = \frac{\mathbf{h}_k^H \mathbf{h}_k}{\mathbf{h}_k^H \left(\mathbf{I} - \widetilde{\mathbf{H}}^H (\widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^H)^{-1} \widetilde{\mathbf{H}}\right) \mathbf{h}_k} = \frac{\mathbf{h}_k^H \mathbf{P}_\perp \mathbf{h}_k + \mathbf{h}_k^H \mathbf{P} \mathbf{h}_k}{\mathbf{h}_k^H \mathbf{P}_\perp \mathbf{h}_k} = 1 + \frac{\mathbf{h}_k^H \mathbf{P} \mathbf{h}_k}{\mathbf{h}_k^H \mathbf{P}_\perp \mathbf{h}_k}, \tag{3.18}$$

where the matrix $\widetilde{\mathbf{H}}$ is the matrix $\mathbf{H}$ without the kth row, $\mathbf{P} = \widetilde{\mathbf{H}}^H(\widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^H)^{-1}\widetilde{\mathbf{H}}$, and $\mathbf{P}_\perp = \mathbf{I} - \mathbf{P}$. As it has been stated $\alpha_k^2 = \mathbf{h}_k^H \mathbf{P}_\perp \mathbf{h}_k \sim \chi^2_{2(Q-K+1)}$, whereas $\mathbf{h}_k^H \mathbf{P} \mathbf{h}_k \sim \chi^2_{2(K-1)}$. With all this, since $\mathbf{P}\mathbf{P}_\perp = 0$, the second term in the summation in (3.18) (with the corresponding scaling) behaves like an $\mathcal{F}$ random variable [152] with $2(K-1)$ degrees of freedom at the numerator and $2(Q-K+1)$ at the denominator, i.e. $\mathcal{F}_{2(K-1),2(Q-K+1)}$. Using the fact that each term in the summation at (3.14) is independent, after some algebra it yields

$$\mathbb{E}\left\{\delta^{-1}\right\} = \mathbb{E}\left\{\sum_{k\in\mathcal{K}} \frac{\|\mathbf{h}_k\|^2}{\alpha_k^2}\right\} = K\mathbb{E}\left\{1 + \frac{K-1}{Q-K+1}\mathcal{F}_{2(K-1),2(Q-K+1)}\right\} = K\frac{Q-1}{Q-K},$$

so that assuming independence, the mean value for the **EPS** can be finally approximated as

$$\mathbb{E}\left\{\frac{\|\mathbf{h}_k\|^2}{\delta^{-1}}\right\} \approx \frac{Q}{Q-1}\frac{Q-K}{K}.$$

Since $\frac{Q}{Q-1} > 1$ the mean value for the EPS is always greater than for the ERB in (3.17). On the other hand, after some algebra, it can be verified that if $K-1 > 0$, the EPS yields always a lower mean than the UPA in (3.16). Note that if $K = 1$, then the three methods yield the same mean values. Therefore, it has been verified that the EPS might be a well-suited technique because it yields a mean value which is in between the concern about the global performance (the UPA at high SNR) and the fulfillment of the individual needs (the ERB). The results in this subsection are in fact the diversity order of the proposed techniques. Particularly, it has been stated that **the diversity order of the UPA is higher than that of the ERB**, see (3.16) and (3.17). The analysis conducted in this subsection reflect that the diversity order of the EPS is in between both methods. After that, it shall be proven that the dispersion between the maximum and minimum values is lower than for the UPA. The analysis is only conducted for the UPA and for the EPS, because the maximum and minimum values among the users are the same as the mean for the ERB, for which all the users are granted the same fraction of the resource, recall (3.17).

Analysis of the maximum and the minimum

The procedure to obtain the behavior of the maximum of any of the techniques is the following. First, it shall be noted that the Cumulative Density Function (cdf) of the maximum of K i.i.d. random variables with cdf $F(x) = P(\mathcal{X} \leq x)$ is given by

$$F_K^{\max}(x) = \left(F(x)\right)^K,$$

and since the interest of this section is to evaluate the mean values, it is sufficient to solve the equation $F_K^{\max}(x) = 0.5$. On the other hand, the cdf of the minimum of the K i.i.d. random variables is

$$F_K^{\min}(x) = 1 - \left(1 - F(x)\right)^K,$$

which shall be again made equal to 0.5 in order to obtain the average performance. Furthermore, the cdf of a Chi-squared random variable with r degrees of freedom is needed, which is given by the regularized Gamma function

$$P(r/2, x/2) = \frac{\gamma(r/2, x/2)}{\Gamma(r/2)},$$

where $\gamma(r/2, x/2)$ is the incomplete Gamma function, and $\Gamma(r/2)$ is the complete Gamma function. With all this, it is sufficient to obtain the mean of the maximum and the minimum values of the EPS and the UPA.

For the **UPA**, the equations that shall be fulfilled for the maximum $x_{\max}^{\text{UPA}}$ and minimum values $x_{\min}^{\text{UPA}}$ are the following

$$\begin{aligned} P(Kx_{\max}^{\text{UPA}}, Q-K+1) &= \sqrt[K]{0.5}, \\ P(Kx_{\min}^{\text{UPA}}, Q-K+1) &= 1-\sqrt[K]{0.5}. \end{aligned}$$

For the **EPS** it shall be noted that the value δ^{-1} is constant for any realization, thus the maximum and minimum values depend only on the channel norm $\|\mathbf{h}_k\|^2$, which is a Central Chi-squared random variable with $2Q$ degrees of freedom. Then, the maximum $x_{\max}^{\text{EPS}}$ and minimum values $x_{\min}^{\text{EPS}}$ shall fulfill

$$\begin{aligned} P(\frac{Kx_{\max}^{\text{EPS}}(Q-1)}{(Q-K)}, Q) &= \sqrt[K]{0.5}, \\ P(\frac{Kx_{\min}^{\text{EPS}}(Q-1)}{(Q-K)}, Q) &= 1-\sqrt[K]{0.5}. \end{aligned}$$

A plot to clarify everything

The figure that is shown in this part is based on the Gini index. In fact, it is like a Lorentz curve within the plot it was shown in the first chapter (Figure 1.15). However, instead of computing how the resources are shared among all the population, only the key values of the minimum, mean, and maximum are needed to have an idea of the fairness of the resource allocation. In this case, the resource is the value for which the author has computed the maximum, mean, and minimum values. Then, the approximate cdf (among users) can be seen in the plot.

It is shown in Figure 3.5 that the ERB yields an equal distribution of the resource, i.e. for a given percentage of the population (x axis), the same fraction of the resource (y axis) is obtained. Increasing the distance with respect to this Lorentz curve increases the differences among users, since in that case the slope is small for low percentage of the population, whereas the slope starts to increase when the percentage of the population becomes higher. Then, the most unfair solution is the UPA (solid blue line), which yields higher differences among users, as it has been shown, as well as it yields the higher area between its line and that of the ERB (dotted black curve). Between the UPA and the ERB one finds the EPS (dashed red line), which yields not

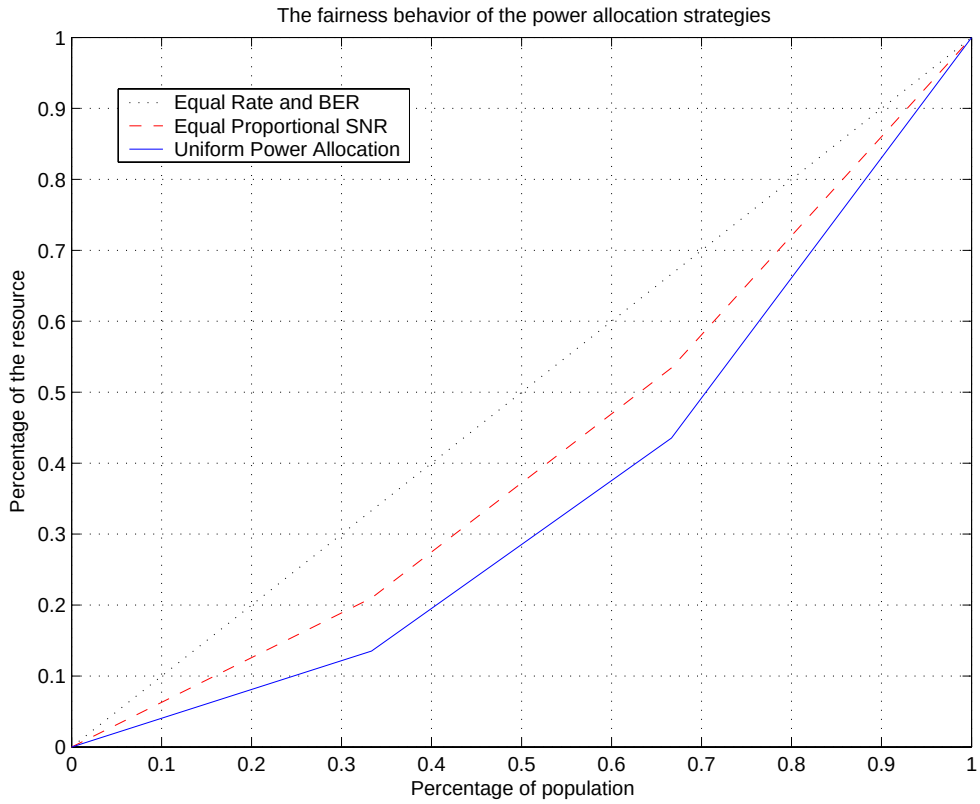


Figure 3.5: For $Q = 20$ antennas and $K = 16$ users, the fairness curves for the UPA (solid blue), the ERB (dotted black), and the EPS (dashed red).

only an intermediate mean value, but also average maximum and minimum values. With the EPS, a good balance between the global performance and the individual needs is obtained.

3.4 A comparison of the best technique for each metric

Before going into the details of the admission control, the widespread power allocation strategy based on utility functions shall be presented and discussed, see Chapter 1 for an overview. For the comparison, only the techniques that optimize the global performance will be taken as benchmarks, namely the Maximum Sum Rate (MSR) and the Minimum Sum BER (MSB). The basic objective of this section is to show that ad-hoc cost functions might not be well-suited for a realistic system optimization, and that a fair comparison shall encompass all the used metrics.

3.4.1 Maximum Sum of Utilities (MSU)

Differently to related papers in the literature, see e.g. [104], [105], or [107], the utility-based downlink power control is solved using tools from convex optimization [31]. Although a game is generally competitive [108], where the users try to obtain a fraction of the resource, the MSU in

this section is analogous to a refereed game in which a cooperative strategy is sought, see [105] and references therein for more details. Whereas in the refereed game the AP would tell the terminals the uplink power at which they shall transmit, in this case the AP allocates a certain power to the users for downlink transmission. Note that if the system is TDMA/TDD and the channel is quasi-static, the same power could be used for the uplink, although there might be some imperfections due to the non-symmetry of the RF chains.

For the problem treated in this dissertation, minor modifications are required in the utility function given e.g. in [105]. Note that a simple flat-fading channel is assumed, whereas CDMA communications are usually the focus of the game-theoretic formulation of the power control. Without further delay, the utility perceived by the kth user can be expressed as

$$u_k = \frac{\left(1 - \frac{\mathrm{BER}}{c_1}\right)^L}{\beta_k^2} = \frac{\left(1 - \exp(-c_2 \frac{\alpha_k^2 \beta_k^2}{\sigma^2})\right)^L}{\beta_k^2}, \tag{3.19}$$

in which, in agreement with e.g. [103], the Frame Success Rate (FSR) in the numerator has been slightly modified. Briefly, the FSR indicates the probability that a packet is successfully received, and it is assumed here that there is no error correction capability, although the basic underlying idea would not change in that case. The modification is the division of the BER by the constant c_1, so as to guarantee that the utility tends to zero as the power goes either to zero or to infinity, that is,

$$\lim_{\beta_k^2 \to 0} u_k = 0, \text{ and } \lim_{\beta_k^2 \to \infty} u_k = 0,$$

If the AP had not proceeded so, at null power, $\beta_k^2 = 0$, the utility would be infinity, and the terminal would choose not to transmit. This modification does not have a deep impact in the trend of the FSR, as it is shown e.g. in [104]. In the literature for the uplink power control, each user is willing to maximize its own utility, given the actions from other players. Then, the first derivative of the utility function in (3.19) shall be obtained. The maximizing solution is the β_k^2 whose equilibrium SNR γ_k^* satisfies

$$\exp(-c_2 \gamma_k^*)(1 + L c_2 \gamma_k^*) - 1 = 0, \tag{3.20}$$

so that the power allocation factors can be obtained using (3.3) as

$$\beta_k^2 = \frac{\sigma^2}{\alpha_k^2} \gamma_k^* = c. \tag{3.21}$$

This point c is a maximum of the utility function in (3.19). Therefore, since there exists a point c such that u_k is non-decreasing for $t \leq c$, and non-increasing for $t > c$, the function in (3.19) is quasi-concave [31]. Moreover, this point constitutes a Nash Equilibrium (NE) for the

uplink power control game, which is taken as a benchmark in existing literature of the game-theoretic power control, see e.g. [103] and [105]. A NE is a point where no user can increase its own utility function by changing its own transmitted power, given the transmitted power from the other users [12]. With all this, the equilibrium SNR is denoted by $\gamma^{NE} = \gamma_k^*$. It is important to note that, with unbounded power, this equilibrium point would be the same for all users as long as they used the same utility function.

Since the focus is the downlink, the AP shall distribute the limited instantaneous power among the users in the cell. This constitutes a difference with respect to existing literature, see e.g. [153] and references therein, where the focus is usually the uplink, and the computation is performed distributed at all the terminals. For this multi-user communication, the AP has several alternatives involving fairness issues, as it has been already seen up to this point. In this section, the AP wishes to maximize the sum of utilities of all the users in the cell, which means that the total perceived satisfaction would be maximum. Again, this maximization increases the differences among the users, but the good thing is that all the scheduled users are granted the maximum satisfaction. Since the objective function (sum of utilities) is quasi-concave because it is obtained by a sum of quasi-concave functions [31], minus a sum of quasi-concave functions is quasi-convex. Therefore, the optimization can be cast in convex form according to

$$\begin{aligned} &\min_{\beta_k^2} -\sum_{k\in\mathcal{K}} u_k \\ &s.t. \sum_{k\in\mathcal{K}} \beta_k^2 - P_T \leq 0, \\ &\quad -\beta_k^2 \leq 0, \forall k \in \mathcal{K}. \end{aligned}$$

Applying the KKT conditions [31], the solution β_k^2 might be in the set

$$\beta_k^2 \in \left\{0, \frac{\sigma^2}{\alpha_k^2}\gamma^{NE}\right\}, \forall k \in \mathcal{K}. \tag{3.22}$$

If the power were unbounded, the utility maximization would yield the same performance as the ERB, since all the users would get the same equilibrium SNR given by γ^{NE} in (3.20). However, since the power is limited, either the user is allocated at a point such that its own utility is maximized or it is not scheduled. The key question is which users will not be allocated for transmission. If the power factors β_k^2 obtained in (3.21) are added as if all the users were active, the total power is $\sigma^2\gamma^{NE} tr\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right]$. Therefore, the maximum sum of utilities problem serves all the users with the SNR of the NE, γ^{NE}, if

$$tr\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right] \leq \frac{P_T/\sigma^2}{\gamma^{NE}}. \tag{3.23}$$

In such a case, the problem is considered to be feasible, whereas in any other case, the AP should decide which users are allocated null power. If one substitutes the equilibrium β_k^2 obtained

1. Set $\mathcal{K} = \{1, \ldots, K\}$.
2. Build matrix $\mathbf{H}$ with the users in the set $\mathcal{K}$, and compute $\alpha_k^2 = 1/\left[\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\right]_{k,k}, \forall k \in \mathcal{K}$.
3. If the condition in (3.23) is satisfied, go to step 5.
4. Otherwise, select the user $k^* : \min_k \alpha_k^2$, and remove it from the active set, $\mathcal{K} = \mathcal{K} - k^*$. Go to step 2.
5. Compute the power for the users in $\mathcal{K}$ according to (3.21). For the users not in $\mathcal{K}$, set $\beta_k^2 = 0$.

Table 3.2: Maximization of the Sum of Utilities (MSU) algorithm.

in (3.21) in the utility function in (3.19), the utility for the kth user at the NE is $u_k^{NE} = k_{NE}\alpha_k^2$, where k_{NE} is a constant. Therefore, the user with lower α_k^2 (worst channel) is the selected candidate to be allocated null power, since it is the user that penalizes the performance of the rest of the users. With these issues, Table 3.2 summarizes the algorithm that yields the highest sum of utilities. First, it tries to allocate all the users, but if the problem is not feasible, the best strategy is to remove the user with worst channel, see step 4 in Table 3.2. Step 5 reflects (3.22). Note that if a user is allocated null power, the α_k^2 shall be recomputed because they increase when less users are served. In most cases, $\sum_{k\in\mathcal{K}} \beta_k^2 < P_T$ because the power is determined by (3.21), which constitutes a power inefficient method. The reason behind is that using more or less power than the NE for any user would imply in any case a lower utility, which is not the ultimate objective of the algorithm.

3.4.2 Simulation results and discussion

The objective of this subsection is to show that a correctly-made comparison shall encompass all the compared metrics. For instance, when comparing rate-based vs. BER-based techniques, the two figures of merit shall be presented. This seems not to be so obvious, because in the game-theoretic literature of power control, the papers did not show how the FSR was degraded whenever the authors proposed a new mechanism to increase the utility (and thus the Pareto efficiency) of the system, and there have been a number of papers on the topic.

An AP provided with $Q = 4$ antennas is the centralized agent of the cell, which tries to serve $K = 4$ active users. Again, the SNR in the figures refers to the ratio P_T/σ^2, and the range is from 0 dB to 32 dB in steps of 4 dB. $M = 2$ bits of symbol are assumed without loss of generality, i.e. QPSK (or 4-QAM) mapping. Although simulations have been conducted to evaluate the utility, the BER, the rate, and the power, here only the BER and the utility will be shown for the sake of

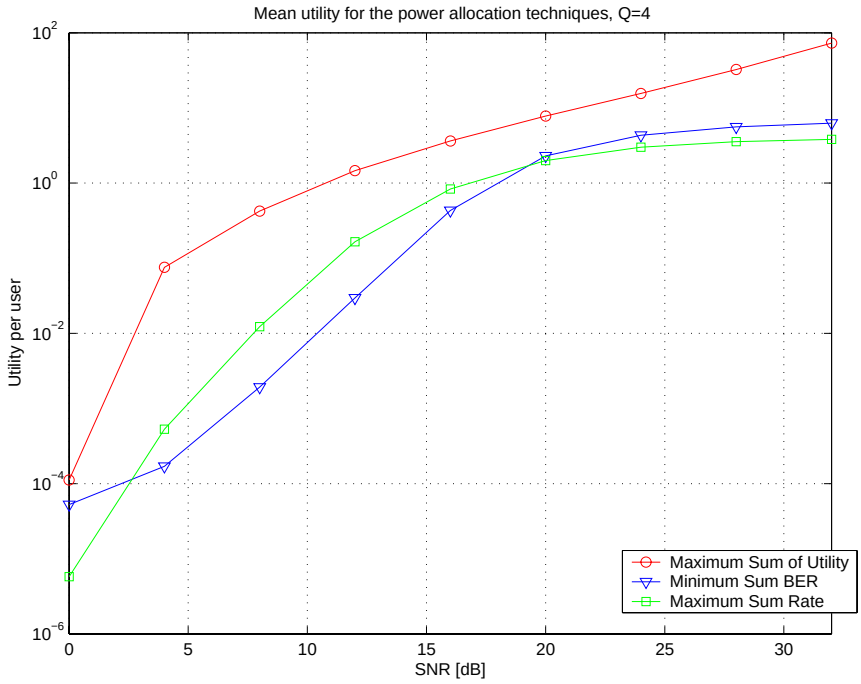


Figure 3.6: Mean utility per user for the optimum techniques for each metric.

conciseness, but note that the rate is minimized by the MSR, while the MSB provides a slightly worse performance, and the MSU degrades significantly, especially in the high-SNR regime. In terms of power, the MSR and the MSB use all the available instantaneous output power, whereas the MSU takes profit of a decreasing percentage of the power with increasing SNR due to the needs of the objective function, which are focused on the utility.

First, it is plotted in Figure 3.6 the utility per user with respect to the SNR. It is clear that the technique based on the maximization of the utility yields the best results compared to the MSR and the MSB. The performance in terms of utility of these techniques does not provide much relevant information. It can be concluded that for the MSU the utility is maximized while the used power is the lowest among the studied methods. However, it is shown in Figure 3.7 the mean BER per user vs. the SNR for the proposed methods, so that the BER is set to 0.5 to any user if he/she is not scheduled for transmission. As it could be foreseen, the MSB yields the optimum performance since it is designed for that purpose, closely followed by the MSR technique. It is important to see that the maximization of the sum of utilities does not yield a good behavior in terms of mean BER. The performance loss is about 9 dB for a BER of 10^{-2}. The final remark is that this SDMA system based on ZF fully exploits the multiplexing gain because it is serving the maximum number of users, i.e. $K = Q$, but the diversity gain is penalized [66]. The author shows in the next section that even when $K = Q - 1$, the BER decreases in more

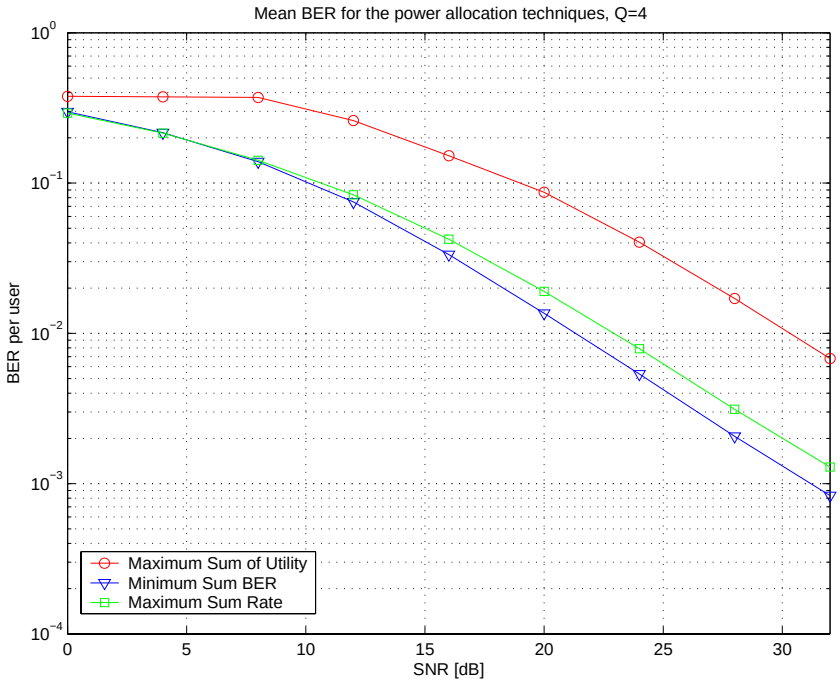


Figure 3.7: Mean BER per user for the optimum techniques for each metric.

than one magnitude order for moderate SNR.

The alternatives in the literature based on game theory, e.g. pricing [103] or repeated games [153], would increase the utility while reducing even more the power. These options are studied in order to overcome the Pareto deficiency of the NE. Briefly, a Pareto optimum point means that no user can increase its own utility without decreasing the utility obtained by other users [12]. To the best of our knowledge, it is not shown in existing papers how the BER (or FSR) performance degrades in such cases, see e.g. [105]. In practical systems, it is not relevant to increase the utility if the error rate is also increased, thus in the end the efficiency of the system is reduced. Therefore, constraints on the SNR or on the BER should be added to the problem in order to fulfill the real traffic requirements from the users. In any case, game theory provides an attractive mathematical framework, and concepts such as pricing can be useful for future communication systems. Essentially, the pricing factor can be effectively set by the AP in order to force the terminal to transmit at a certain power level in the uplink. For instance, it is often assumed that each selfish terminal wishes to maximize the modified utility function

$$\tilde{u}_k = u_k - c_k \beta_k^2, \tag{3.24}$$

where c_k is a different linear pricing factor for each user. The pricing is usually designed in order to increase the sum of utilities for all the users in the cell. However, this has the undesirable effect

of increasing the BER for some users, and the sum BER in particular. Therefore, a better-suited option would be to choose the pricing factor c_k in a way such that when the terminal optimizes individually $\tilde{u}_k$ with respect to β_k^2, the selected power would be the one previously computed by the AP in order to optimize a certain cost function with a clearer physical sense. In other words, if (3.24) is derived with respect to the power allocation β_k^2, then

$$\frac{\delta \tilde{u}_k}{\delta \beta_k^2} = \frac{\delta u_k}{\delta \beta_k^2} - c_k = 0,$$

shall be fulfilled, so that finally the power allocation factor is dependent on the pricing, i.e. $\beta_k^2 = g(c_k)$. Therefore, the AP could select among one of the proposed techniques, and compute the c_k such that β_k^2 corresponds to that criterion. Since the AP has all the necessary information and computational capabilities, after these simple operations it can communicate the pricing value to the terminals, so that the power allocation is computed in a distributed manner. However, note that some information is needed at the terminals, which shall be provided by the AP, but it is limited to the pricing factor, since the equivalent spatial channels α_k can be estimated through an appropriate training sequence.

3.5 Admission control

In fact, some kind of admission control or user selection is already being made in the previous algorithms whenever a user shall be removed from the active set because the power constraint cannot be fulfilled. However, the previous strategies for the power allocation obtain the solution for a given number of antennas Q and users K, and the BS does not control the individual QoS for the users in a best-effort type of service. In the SDMA system that has been proposed, a maximum number of Q users can be allocated for transmission, that is, *one antenna, one user*, but the optimum number of users might be lower than Q. This fact leads the BS to allocate the best users in order to obtain the best performance. The optimization at the BS station shall determine both the number of users and, more concretely, which users, since the interactions among them are crucial for the performance of the scheduler (as it has already been shown).

Linked to the choice of the optimum number of users, it has been recently reported that there exists an optimum number of antennas that should be dedicated to the users if ZF is used as the transmit beamforming scheme [55]. Equivalently, the best global performance is achieved when less than Q users are served simultaneously. If the number of antennas Q is higher than the number of active users K, and they both grow without bound, i.e. $K, Q \to \infty$, but their ratio $\zeta = \frac{Q}{K}$ remains constant, the sum rate R increases linearly with respect to the number of antennas not only for the ERB technique [55], but also for the UPA and the MSR, i.e.

$$\lim_{K,Q\to\infty} \frac{R}{Q} = \frac{1}{\zeta} \log\left(1 + \frac{P_T}{\sigma^2}(\zeta - 1)\right).$$

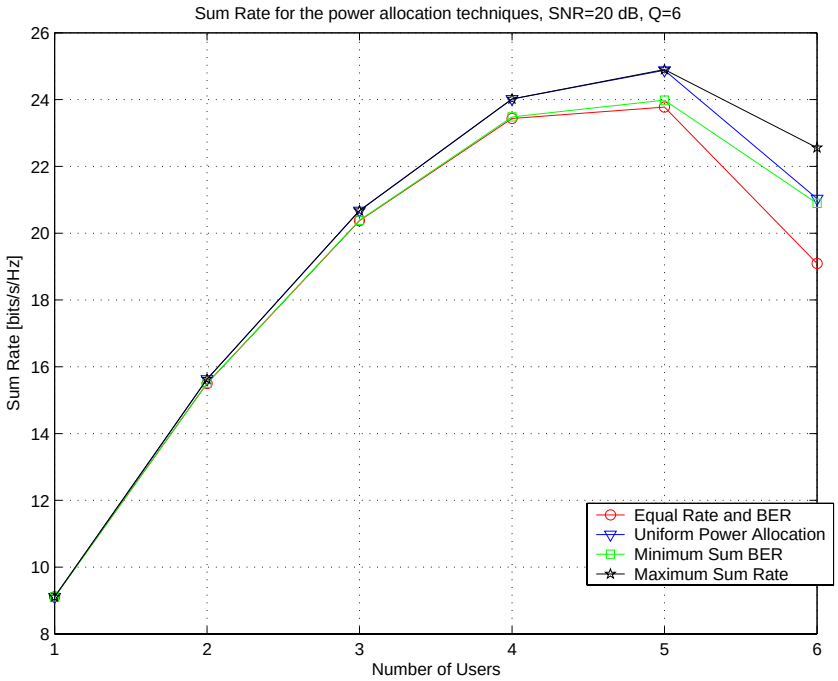


Figure 3.8: For $Q = 6$ antennas, mean sum rate vs. number of users for the power allocation techniques: the ERB, the UPA, the MSB, and the MSR.

Looking at the previous equation, one should note that there exists a number of users K that optimizes the sum rate for any given number of antennas Q. If the number of antennas Q is fixed, the BS has to select the optimum number of users K. Besides, the sum rate can differ significantly depending on the choice of these K users, since if their channels are correlated, then more power is needed. Figure 3.8 depicts the outage sum rate at 10% vs. the number of users being simultaneously served at the array, for the ERB, the UPA, the MSB, and the MSR. In any case, the sum rate is maximized if the number of users is lower than the number of antennas. In fact, serving as many users as antennas penalizes the performance. However, note that if one looks at the rate per user, with 5 users, the rate for the MSR is about 5 bits/s/Hz, whereas for a single user, this value is nearly doubled. Therefore, the performance decreases in terms of rate per user although the sum rate might be better. Differently to the rate methods, for the BER techniques the sum BER always decreases as the number of users decreases. It is shown in Figure 3.9 the BER performance vs. the SNR for the ERB, the UPA, the MSB, and the MSR with 4-QAM (QPSK) mapping. The BER can be dramatically reduced if the scheduler serves two users less than the number of antennas.

In practical scenarios the BS aims at optimizing the global performance of the cell while trying to cope with the individual QoS requirements of the users. For this purpose, the Spatial

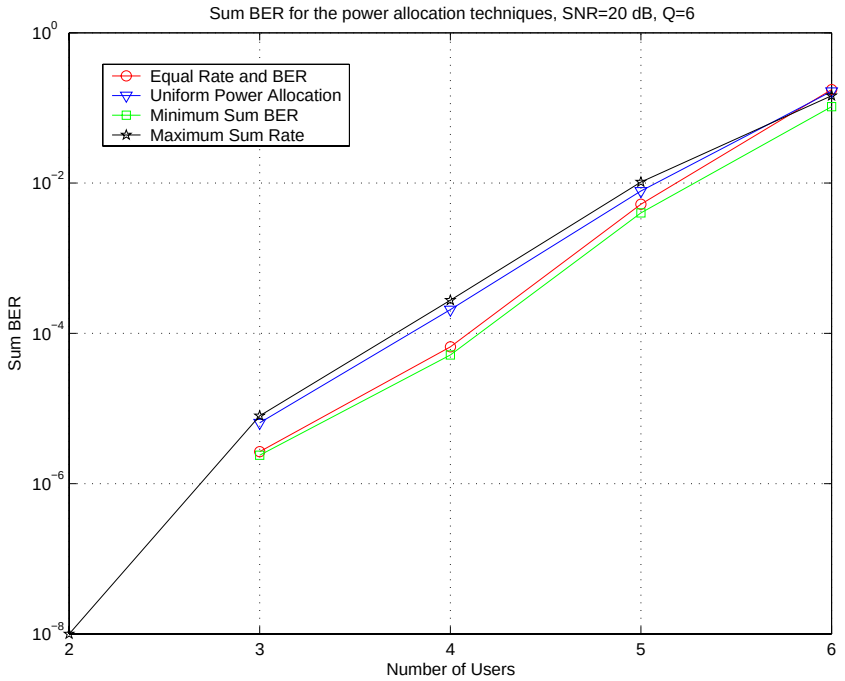


Figure 3.9: For $Q = 6$ antennas, mean sum BER vs. number of users for the power allocation techniques: the ERB, the UPA, the MSB, and the MSR.

Admission Control (SAC) mechanism decides which users cannot be scheduled while fulfilling the requirements of the selected ones and optimizing the global performance at the same time. In this section, it is assumed that the packets from the users that are not scheduled are lost because their due date is stringent, such as it happens for voice and real-time video applications. Another possibility would be that users transmit older packets in subsequent slots if the delay constraints are relaxed, e.g. for data transfers. In any case, the basics of the algorithms presented next are valuable. Due to the asymptotic behavior discussed previously in this chapter, only the UPA and the ERB technique are taken as benchmarks for a comparison with the newly proposed EPS. It has been shown in Section 3.3 that the diversity order of the UPA is always higher than that of the ERB, and that the EPS provides an intermediate performance among them. In this section, it will be shown that the number of served users (multiplexing gain) is higher for the ERB than for the UPA, and the EPS also provides an intermediate behavior.

3.5.1 The addition of SNR constraints

The main goal of the PHY scheduler is to reduce the amount of information that shall be processed by the traffic scheduler at the DLC. Particularly, the PHY scheduler performs the admission control. Due to the interactions in this SDMA system, a crucial point is which subset

1. Set $\mathcal{K} = \{1, \ldots, K\}$.
2. Build matrix $\mathbf{H}$ for the users in the set $\mathcal{K}$.
3. Compute $\alpha_k^2 = 1/\left[\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\right]_{k,k}, \forall k \in \mathcal{K}$.
4. If the condition (depending on the technique) in Section 3.5.1 is satisfied, go to step 7.
5. Otherwise, remove the active user having the worst channel $\mathcal{K} \leftarrow \mathcal{K} - \left\{k^* \in \mathcal{K} : \min_k \alpha_k^2\right\}$, and go to step 2.
6. If $|\mathcal{K}| = \emptyset$, the algorithm finishes.
7. Compute the power allocation according to one of the criteria in Section 3.5.1, and finish.

Table 3.3: Spatial Admission Control.

of users $\mathcal{K}$ is served. This shall be decided taking into account the BER or rate requirements, which can be mapped into a target SNR γ^t. The feasibility conditions for the UPA, the ERB, and the EPS with SNR constraints are provided, after which some simulation results are given.

Feasibility conditions

First, for the **UPA**, the SNR for each user k shall be above the threshold γ^t, i.e.

$$\gamma_k^{UPA} = \gamma^n \frac{\alpha_k^2}{K} \geq \gamma^t,$$

so that the equivalent channel for each user shall fulfill

$$\alpha_k^2 \geq |\mathcal{K}| \frac{\gamma^t}{\gamma^n}, \ \forall k \in \mathcal{K},$$

where it is emphasized that the set of users $\mathcal{K}$ that shall be served is optimized. On the other hand, for the **ERB**

$$tr\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right] \leq \frac{\gamma^n}{\gamma^t}$$

shall be fulfilled, which is a single constraint for all the users although it requires the same computational complexity. Finally, for the **EPS** the constraints are again individual, so that

$$\|\mathbf{h}_k\|^2 \geq \delta^{-1} \frac{\gamma^t}{\gamma^n}, \ \forall k \in \mathcal{K},$$

shall be fulfilled.

The admission control mechanism is summarized in Table 3.3. It tries to fit all the users (steps 1 to 4), but if the feasibility condition for the selected technique is not satisfied (step 4)

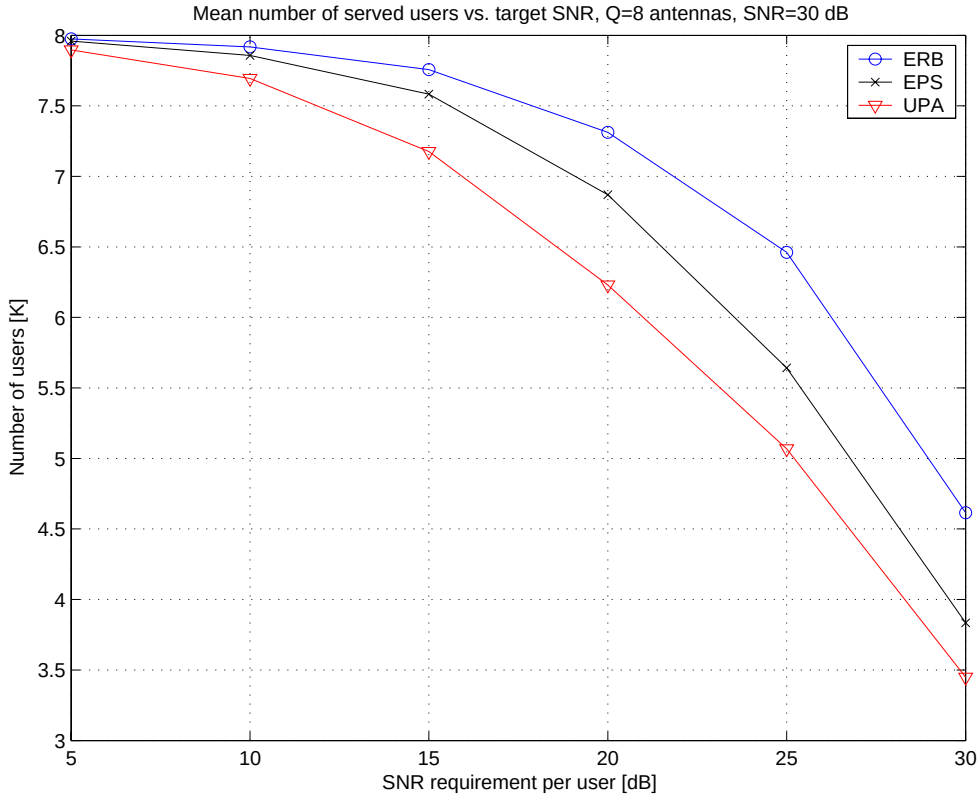


Figure 3.10: For $Q = 8$ antennas, mean number of served users vs. the SNR requirement.

the user with worst channel is removed from the set of active users $\mathcal{K}$ (step 5), since it is the one that worsens the performance of the other users, as stated by the solvability conditions. When the solvability condition is fulfilled, the power allocation factors are computed, see step 7. On the other hand, if the condition cannot be fulfilled with any user, the algorithm finishes and does not serve any user (step 6). Since the AP starts from the maximum number of users and then drops the worst user out at each iteration, the optimum distribution of the users might be found.

In order to evaluate the performance of this spatial admission control mechanism, the following setup is built. The cell is governed by an AP with $Q = 8$ antennas, and there is a maximum of $K = 8$ users in the cell. However, since the SNR constraints shall be fulfilled, not all of them will be served. It is assumed that the SNR is $\gamma^n = \frac{P_T}{\sigma^2} = 30$ dB, so that the system operates in the high SNR regime, and the target SNR is varied from 5 to 30 dB. Figure 3.10 reflects the results that have been previously exposed from another perspective, namely the mean number of served users with respect to the SNR requirement, which reflects in some sense the multiplexing gain. It is stated that the UPA gives service to the lowest number of users so as to improve the global performance by not serving the poorer users. On the other hand, the ERB serves the highest number of users, but the global performance is penalized, as it has been seen before. Finally, the EPS strategy provides again an intermediate solution between them.

Therefore, recalling the results concerning the diversity gain, it can be stated that more diversity gain implies a lower multiplexing gain, which agrees with the trade-off in [66].

3.6 Conclusions

This chapter is devoted to the power allocation techniques in a multi-antenna broadcast channel. First, several traditional criteria have been formulated and analyzed, especially in terms of fairness, which is an issue that is usually *forgotten* in the physical layer literature. Depending on the optimization goal of the PHY-DLC scheduler, several options are available to balance the individual needs and the cell needs. Moreover, their correspondences in the high SNR regime have been evaluated, so that finally the Uniform Power Allocation (UPA) and the scheme providing an equal SNR (thus rate and BER) for every user (ERB) are the selected techniques for the admission control. They are asymptotically equivalent to their counterparts: the Maximum Sum Rate (MSR) and the Minimum Sum BER (MSB).

After that, the ERB and the UPA are compared to a new strategy, the Equal Proportional SNR (EPS), which balances in an intermediate way the trade-off among the global optimization and the fulfillment of the individual constraints. Moreover, the admission control procedure reflects the fundamental trade-off between the diversity gain and the multiplexing gain for these power allocation techniques. A theoretical comparison is conducted, and simulations results are given to validate the results. The EPS is shown to provide an intermediate behavior among the UPA and ERB.

Finally, the best technique for each of the proposed metrics, i.e. the maximum sum rate and the minimum sum BER, is compared to a widely deployed cost function in the game-theoretic of the power control for CDMA: the maximum sum of utilities. It is shown that the objective function shall be carefully chosen, otherwise undesirable effects such as a dramatic (and unacceptable) BER increase might be suffered. However, pricing mechanisms may have an importance in future wireless systems and shall be considered as part of the design of a system.

Chapter 4

Spatial bit allocation

In this chapter, the focus is exclusively the spatial bit allocation. As it has been stated up to this chapter, optimizing the resources in a communication of a multi-antenna AP with several single antenna terminals is not an easy problem, even in the case when the number of users does not exceed the number of antennas. Indeed, the AP has several alternatives for the tasks involving both physical layer issues e.g. beamforming and power or bit allocation, and DLC aspects such as scheduling. As it has been shown, when several users are to be served simultaneously, the selection of the technique might be more complicated than for single-user communications.

The purpose of this chapter is the extension of traditional bit allocation strategies so as to take into account the spatial dimension whenever the number of users which is not higher than the number of antennas. Moreover, three proposed techniques are analyzed in terms of fairness. This extension is not straightforward, because the interactions among the users that are being simultaneously served is crucial for the performance of the algorithms. For instance, whenever a user is not scheduled for transmission, the equivalent channels from the others change. Two traditional approaches are taken to tackle the problem, as well as a modification that is shown to provide the best trade-off between performance and complexity based on simulation results.

The rest of the chapter is organized as follows. Next section provides the introduction, immediately after which a comparison of the spatial bit allocation algorithms for a lower number of users than antennas is conducted in Section 4.2. Then, conclusions are given.

4.1 Introduction

The Q-antenna AP shall now distribute the K single antenna users in such a way that the bit allocation in the spatial domain is also performed. When $K \leq Q$, with multiple signal mappings and multiple antennas, the rate optimization under BER and power constraints leads to an spatial bit allocation problem [109], which is also linked to the selection of the users that will be served (admission control [154], see also previous chapter). However, there are differences

between traditional bit allocation and the same problem when the spatial dimension is added. As it will be shown, the user with best channel norm might not even be among the better users when several users are to be served simultaneously. This is due to the special characteristics of the spatial domain, an in particular, to the ZF beamforming criterion that is used.

After the beamforming, the AP shall distribute the total available instantaneous power among the users. In a practical system, the communications from the users require some QoS in terms of error rate (e.g. BER) or SNR, certainly under the assumption that an integer number of bits per symbol is used. If the objective is to maximize the rate of the system, an spatial bit allocation problem is naturally formulated therefrom. In the literature, the bit allocation/loading problem has been extensively studied since [109], where an optimal algorithm for single-user discrete bit loading in multi-carrier systems is proposed. On the other hand, computationally-efficient suboptimum schemes for DMT are developed in [111], [110], and in the references they include. Essentially, two strategies can be found, namely *bit filling* and *bit removal*. The former adds a bit to the user/subcarrier providing the lowest increase in total power, and *bit removal* schemes remove the most penalizing bit until the power constraint is fulfilled. Both approaches yield the optimum solution whenever the equivalent channels for the users do not change according to the subset of users that is served, as it happens also in [155]. Nevertheless, this is no longer valid for multi-antenna systems because the spatial channel gains gather the influence of the users that are being simultaneously served. In such a case, bit removal algorithms shall be used to attain the optimum solution. Bit filling techniques in the spatial domain could be useful when the number of users exceeds the number of antennas, but the solution might be rather inefficient in general because some effects among the channel vectors might be masked. There also exists a clear trade-off between performance and complexity.

Even without the spatial dimension, in the literature the fairness implications of the bit allocation strategies are rarely evaluated. Some exceptions include [132], where the authors optimize the sum of rates while ensuring an equal long term throughput for all the users. On the other hand, opportunistic communications usually aim at *equalizing* the performance of the users in the long term, see [156] for an example in the uplink. Instead, the focus, as in the whole dissertation, is primarily on the implications in the short-term, which is especially suited for communications with hard delay constraints. Note that the fairness implications might vary substantially according to the different perspectives of the AP on the time scale. In this case, taking into account that the number of bits per symbol (rate) is an integer, the problem is solved according to two perspectives, namely the Maximization of the Sum Rate (MSR), which is discrete in contrast to that in the previous chapter, and the Maximization of the Minimum Rate (MMR). The former pursuits the best global performance although the resources are distributed in an asymmetric way. On the other hand, max-min schemes distribute the resources equally at the expense of loosing in global performance [129]. In this chapter, the MMR is modified so as to

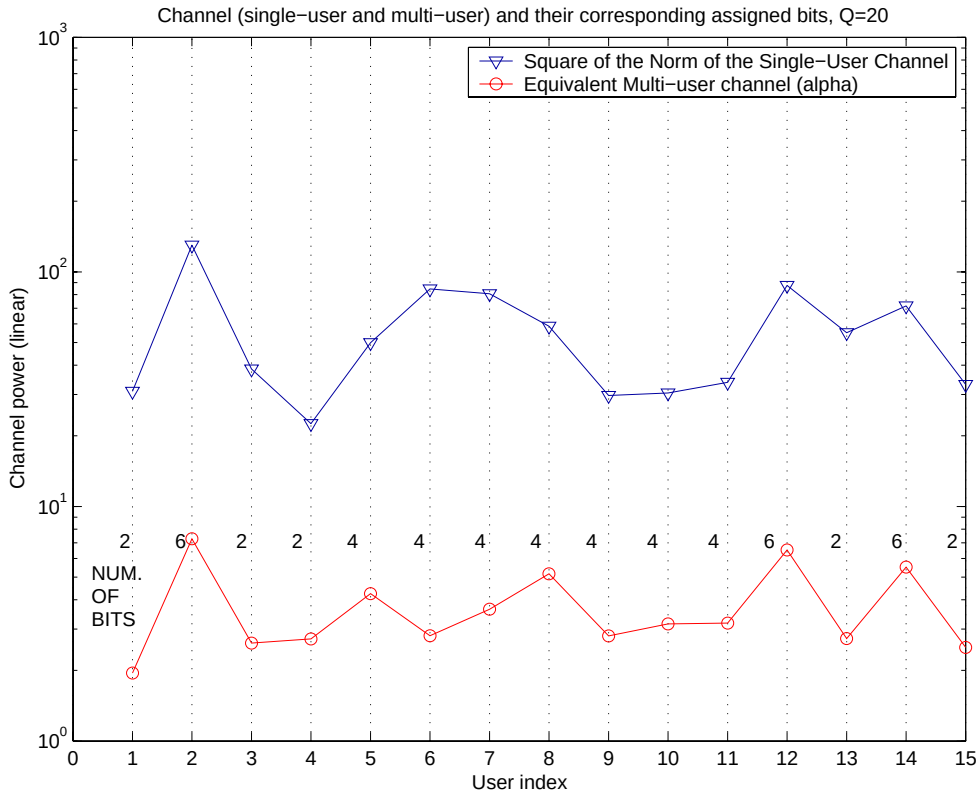


Figure 4.1: Relationship between the channel as if each user were alone in the cell (top blue curve), together with the equivalent channel, α_k in (4.1), after the beamforming (bottom red curve). The number of bits per symbol for each user is also given for the MSR algorithm that will be presented next.

improve its behaviour, and it is shown that the throughput can be improved without decreasing the number of bits assigned by the MMR to any user. This means that the modified scheme yields a Pareto improvement [12] over the traditional MMR.

4.2 Spatial bit allocation strategies

Before proceeding, it seems necessary to recall the final signal model with the deployed Zero Forcing (ZF) transmit beamforming, see (3.1),

$$y_k = \alpha_k \beta_k s_k + w_k, \forall k \in \mathcal{K}, \tag{4.1}$$

so that the SNR for the kth user (assuming equal noise power σ^2 for all of them without loss of generality) is given by

$$\gamma_k = \frac{\alpha_k^2 \beta_k^2}{\sigma^2},$$

where QAM symbols have been assumed, as in the previous chapter. However, note that it is possible to formulate the algorithms that will be proposed in this chapter for more general

mappings. The signal model in (4.1) provides parallel and orthogonal channels for the users, thus there is no inner-cell interference. If the noise is Gaussian, the BER requirement translates directly into a target SNR. Finally, the author reminds the reader that the approximate BER for QAM signals given in [149] will also be used in this chapter

$$\mathrm{BER}(\gamma) \approx c_1 \exp(-\frac{c_2 \gamma}{2^m - 1}),$$

where m is the number of bits in the constellation, $c_1 = 0.2$, and $c_2 = 1.6$. In a real system with multiple signal mappings, bit and power allocation shall be done together so as to obtain a high spectral efficiency. The maximum achievable rate is bounded by the error-free rate of each mode, which rarely occurs in wireless channels. Therefore, the use of multiple signal mappings allows the scheduler to fulfill the QoS requirements in a practical situation due to the better adaptation to the environmental conditions. Realistic spatial bit allocation considers the fact that the number of points in the constellation is always an integer number, usually a power of 2 as in the signal mappings on-the-market systems deploy. In that case, the power allocation is univocally determined by the constellation size and the target BER. If the target BER is fulfilled, a higher number of bits per symbol implies a higher throughput.

Particularly in the SDMA system under consideration, the interactions among the users shall be carefully considered because they have a deep impact on the system performance, differently to other existing literature. In fact, the upper curve in Figure 4.1 are the channels from the users as if they were alone in the cell, i.e. the $\|\mathbf{h}_k\|^2$, so that all the antennas are dedicated to a given user. The lower red curve from Figure 4.1 reflects the equivalent channels as if all the users were served together by the SDMA scheme, i.e. the α_k^2. Without loss of generality, this situation includes an AP with $Q = 20$ antennas and $K = 15$ users in the cell. A higher channel gain $\|\mathbf{h}_k\|^2$ does not necessarily imply a high value of the equivalent spatial channel α_k^2 when several users are served simultaneously, see e.g. the loss in the equivalent channel from user 6. Moreover, it is observed that the trends in the channel gains are not the same, which reflects that the interactions among the users are crucial. Therefore, traditional *orthogonal* bit allocation algorithms shall be modified. In that figure, it is also plotted how many bits per symbol are assigned to the users when the available constellations are 4-QAM (QPSK), 16-QAM, and 64-QAM, see Section 4.2.4 for further details. In any case, it is verified that the users with better equivalent channels α_k achieve the highest number of bits.

The bit allocation algorithms shall apply the fairness criterion of the AP in order to determine which users are served (admission control) and their rate. In fact, the fairness criterion determines the bit allocation strategy in multi-user communications, since the interactions among the users shall be carefully considered. On the one hand, the AP could assign the same rate (number of bits per symbol) to all the users. On the other hand, the AP could choose to optimize the global performance regardless of the users with worse channel conditions. The former is the Maximum

Minimum Rate strategy (MMR) whereas the latter is the Maximum Sum Rate (MSR) scheme. Both of them are mostly based on *bit removal* strategies because of the particular aspects that arise with the use of the spatial domain, and in particular with the ZF beamforming, which have been exposed up to this point. Between them, this chapter proposes the Modified MMR, which yields a close performance to the MSR by a combination of a *bit removal* and a *bit filling* strategy. Indeed, using some unused power the MMR naturally *wastes*, it might obtain a Pareto improvement over the MMR, that means that some users might obtain a higher rate than with the MMR without decreasing that of any other user. This Modified MMR is especially well-suited for the case where the number of users exceeds the number of antennas, as it will be shown.

4.2.1 Maximization of the Sum of Rates (MSR)

This viewpoint is the most used in the literature, see Chapter 1, and has been extensively studied. However, it has not been directly applied to the case where multiple antennas are deployed, an in particular, with a ZF beamforming. This forces the system designer to take into account that the equivalent channel gains α_k with ZF beamforming reflect the interactions among the users that are simultaneously served, as it has been already stated. Moreover, this kind of cost function reflects the optimization of the global performance, without considering that there might be some users that could not even transmit. In any case, the MSR can be expressed as:

$$\max_{m_k} \sum_k m_k \tag{4.2}$$

$$s.t. \quad \sum_{k\in\mathcal{K}} \beta_k^2 \leq P_T, \tag{4.3}$$

$$\mathrm{BER}_k \leq \mathrm{BER}_t, \forall k, \tag{4.4}$$

$$m_k \in \widetilde{\mathcal{M}}, \forall k, \tag{4.5}$$

where, due to algorithmic issues, the set $\widetilde{\mathcal{M}}$ is defined as the union of the possible constellations together with 0 (no transmission), that is, $\widetilde{\mathcal{M}} = \{0\} \cup \mathcal{M}$. A difference with the MSR algorithm in the previous chapter is that the number of bits per symbol is now an integer and not a real value. This imposes some constraints on the algorithmic solution as it will be shown.

The optimum solution to this problem is complex since it requires an exhaustive search among all possible combinations of users and number of bits. If the number of users K is equal to the number of antennas Q, the total number of combinations is $|\widetilde{\mathcal{M}}|^Q$, which is not a negligible search space when $Q = K = 6$ (4096 combinations if $|\widetilde{\mathcal{M}}| = 4$). However, for moderate values of K and Q the exhaustive search could be feasible. Moreover, if the knowledge of the SNR is added to the problem, the algorithm could benefit from the fact that at low SNR a single user is usually scheduled for transmission, whereas at high SNR, as many users as antennas can be in general simultaneously served. In this chapter, the objective is to compare the bit allocation

1. Set $\mathcal{K} = \{1, \ldots, K\}$.
2. Set $m_k = \max \mathcal{M}, \forall k \in \mathcal{K}$.
3. Build channel matrix $\mathbf{H}$ for the users in $\mathcal{K}$. Compute α_k^2.
4. Compute β_k^2 according to (4.6), and the used power $P_S = \sum_{k \in \mathcal{K}} \beta_k^2$.
5. If $P_S \leq P_T$ or $|\mathcal{K}| = \emptyset$, the algorithm finishes.
6. Compute $p_k \left(m_k^i, m_k^j \right), \forall k \in \mathcal{K}$, where m_k^i is the current mapping and m_k^j the lower one in $\widetilde{\mathcal{M}}$.
7. Select $k^* : \max_k p_k \left(m_k^i, m_k^j \right)$, i.e. the user with a higher power gain, and reduce the number of bits $m_k^i \leftarrow m_k^j$.
   If $m_k^i \in \mathcal{M}$, go to step 4. Else, remove user k^* from the set of active users $\mathcal{K} \leftarrow \mathcal{K} - k^*$, and go to step 2.

Table 4.1: Maximization of the Sum of Rates (MSR).

in terms of fairness, as well as to propose their extension to more realistic cases where the K exceeds Q. Therefore, this kind of *implementation* issues is left as further work.

In case the problem is not feasible with all the active users, the AP has to perform the admission control, i.e. choose the users that will be served. The ultimate goal is to optimize the cell performance regardless of the poorer users, thus the distribution of the resources might be uneven (unfair). The spatial bit allocation algorithm that is proposed next yields a close-to-optimum solution that could be implemented in real time. First, note that the constraint (4.4) fixes the power allocation β_k^2 according to

$$\beta_k^2 = \frac{\sigma^2 \left(2^{m_k} - 1\right)}{c_2 \alpha_k^2} log \left(\frac{c_1}{\mathrm{BER}_t} \right), \tag{4.6}$$

which helps in the definition of a function that reflects the power decrease of using a lower constellation size. With traditional bit filling algorithms, it is a proven fact that if a bit is added where it is most efficient, that is, where it consumes less power, the algorithm yields the optimum solution. Conversely, for bit removal techniques it is optimal to remove the bits that make use of the highest amount of power. Assuming that the constellation size (in bits per symbol) for the kth user decreases from m_k^i to m_k^j, where $m_k^i > m_k^j$, and that the rest of users do not change their modulation order, fixed at m_i for the ith user, the power reduction is approximately

$$p_k \left(m_k^i, m_k^j \right) = \begin{cases} \frac{1}{\alpha_k^2} \left(2^{m_k^i} - 2^{m_k^j} \right) & \text{if } m_k^j \in \mathcal{M}, \\ \sum_{i \in \mathcal{K}} \frac{2^{m_i}}{\alpha_i^2} - \sum_{i \in \widetilde{\mathcal{K}}} \frac{2^{m_i}}{\widetilde{\alpha}_i^2} & \text{if } m_k^j \notin \mathcal{M}, \end{cases} \tag{4.7}$$

where the set $\widetilde{\mathcal{K}}$ gathers all the users but the kth and the equivalent channels $\widetilde{\alpha}_i$ are computed

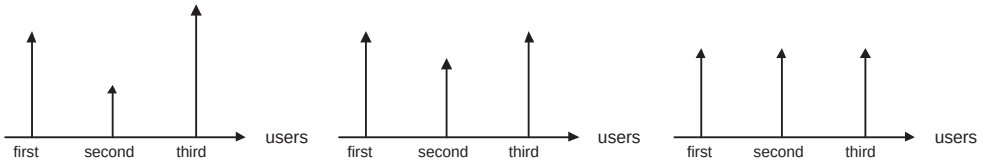


Figure 4.2: Example of a three-step procedure for a max-min optimization.

for the users in $\widetilde{\mathcal{K}}$. In fact, it can be seen this is not the exact power saving that is obtained because if a user is removed, the rest of the users have the chance to increase their modulation index. However, this would require to compute the exact constellation size for each user after the removal. This increases severely the computational burden of the algorithm, so the reduction in complexity of this approximation justifies its use.

The MSR algorithm in Table 4.1 is essentially a *bit removal* technique, but since the spatial channel gains α_k change whenever the set of active users varies, it is combined with a *bit filling* scheme. Briefly, the MSR works as follows. First, it tries to serve all the users with the highest modulation in $\mathcal{M}$ at steps 1-4. If the power constraint in (4.3) is not fulfilled (step 5), the scheduler decides which user should reduce the constellation size or which user should not be served. Since the number of bits shall be reduced, the scheduler selects the user having a maximum incremental cost of using a lower modulation, i.e. the user that saves more power if the bit rate is reduced, see steps 6 and 7. The AP reduces the number of bits of the selected user, and if it belongs to a possible constellation the algorithm goes again to step 4. Otherwise, it drops that user out from the set of active users (step 7), and the constellation size of all the remaining users is set again to the maximum (step 2), so that the power and bit allocation need to be done again. The algorithm finishes when the power constraint is fulfilled or if the set of active users is empty (step 5). This algorithm benefits from the fact of an increasing SNR because in an opportunistic way, the best users are scheduled at any realization of the channel.

4.2.2 Maximization of the Minimum Rate (MMR)

Instead of optimizing the global performance, another option is to serve as many users as possible with the same number of bits per symbol. With this approach, the global performance is penalized, although it is guaranteed that the users being served receive the same rate. If all the users are homogeneous (or pay the same price for the service) this option might be preferred because of its fairness. This type of problems are formulated as a max-min allocation, which yields finally an equal resource to all the users. Intuitively, if the minimum is to be maximized, the algorithm might always reduce the allocation for the maximum user to *give* it to the worst, until all of them are *equalized*. This procedure might be clarified by the Figure 4.2. There, at the first step, all the users are given a different service. The plot in the middle uses some of the

1. Set $\mathcal{K} = \{1, \ldots, K\}$.
2. Build matrix $\mathbf{H}$ for the users in $\mathcal{K}$, and compute α_k^2.
3. Set $m_k = m = \max \mathcal{M}, \forall k \in \mathcal{K}$.
4. Compute β_k^2 according to (4.6).
5. If the condition in (4.8) is not satisfied, reduce the number of bits $m^i \leftarrow m^j$, where $m^j < m^i$ (the lower one in the set). Otherwise, the algorithm finishes.
6. If $m^i \in \mathcal{M}$ go to step 4.
   Otherwise, select the worst user $k^* : \min_k \alpha_k^2$, which is eliminated from the active set, $\mathcal{K} = \mathcal{K} - k^*$.
   Go to step 2.

Table 4.2: Maximum Minimum Rate (MMR).

resources of the third user to increase the performance of the second user. Finally, the first and the third users *give* some of their resource to equalize their behavior of the second user. With convex optimization, this can be shown easily using the KKT conditions, see the ERB in the previous chapter. Then, this alternative can be expressed as a max-min problem, i.e.

$$\begin{aligned} &\max_{m_k} \min_k m_k \\ &s.t. \quad \sum_{k \in \mathcal{K}} \beta_k^2 \leq P_T, \\ &\qquad \mathrm{BER}_k \leq \mathrm{BER}_t, \forall k, \\ &\qquad m_k \in \widetilde{\mathcal{M}}, \forall k. \end{aligned}$$

Again, the optimum solution implies the exhaustive search among all the users and all the number of bits. However, the complexity is lower than the MSR, since now an equal number of bits is imposed to all the scheduled users, thus the number of combinations is reduced. Now, it is guaranteed that all the users receive the same service, but the global performance might be penalized. However, the problem might not be feasible and the AP should decide which users are served. Assuming that the number of bits is $m_k = m, \forall k \in \mathcal{K}$, the problem is feasible if

$$tr\left[\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\right] \leq \frac{P_T}{\sigma^2} \frac{c_2}{log\left(c_1/\mathrm{BER}_t\right)} \frac{1}{2^m - 1}, \tag{4.8}$$

where this condition is obtained by performing the summation of the power allocation factors β_k^2 in (4.6) when the number of bits is equal to all the users. Based on (4.8), an algorithm is proposed in Table 4.2: first, the highest constellation is tried for all the users, steps 1-4. If the feasibility constraint in (4.8) is not fulfilled, it reduces the number of bits for all users, see step

5. When the number of bits is the lowest, the user with a lower equivalent channel α_k is dropped out from the set of active users, step 6, and then the number of bits of all the other users might be increased again. The objective is now to serve as many users as possible, although they might be assigned a lower constellation size. A drawback of this method is that the collective outcome might be significantly reduced due to the objective of serving the users with the same constellation size. An advantage is that this equal mapping reduces the signaling needs.

4.2.3 Modified MMR

In order to alleviate the stringent behavior of the MMR, it is proposed in this section another strategy that yields a Pareto improvement over the MMR. That means that the performance (in terms of number of bits per symbol) of some users is increased without decreasing the performance of other users. This is possible thanks to some unused power the MMR naturally wastes. Briefly, the output in terms of users and number of bits of the MMR is used as input. Recalling (4.7), a bit filling strategy is performed until the unused power is efficiently consumed. As the MMR does not use all the available power, the constellation index for some users might be increased without decreasing any modulation from any user. For this purpose, the user that requires less power to increase its constellation size shall be chosen, as it has been already commented before, and the scheduler tries to increase its mapping size. This is effectively done if the power constraint is fulfilled. The procedure is repeated until no more bits can be added without using more power than the budget P_T. Given the output of the MMR, this algorithm is a *bit filling* technique instead of a *bit removal*. Next section will show the benefits of such a strategy.

4.2.4 Simulation results

Note that the SNR in the figures refers to the ratio P_T/σ^2, and it is assumed that the length of the packets is $L = 1024$ bits with a target PER of 0.1, which is directly translated into a BER. The available constellations are 4-QAM (QPSK), 16-QAM, and 64-QAM.

First of all, it is interesting to evaluate the behavior of the algorithms in a concrete realization of the channel, see Figure 4.3 for a particular case: $Q = 6$ and $K = 5$. There, the solid line with circle reflects the channel gains α_k for each of the five users. Moreover, the bars reflect the number of bits per symbol allocated by each of the algorithms to the users. At each user, the blue left bar is the performance of the MSR, the green middle bar that of the modified MMR, and the brown right bar is that of the MMR. It is observed that the MMR allocates all the users the lowest constellation size (a total of 10 bits per symbol are transmitted), whereas the MSR yields the maximum sum rate of 20 bits per symbol. The Modified MMR outperforms the MMR by increasing the rates of some users with better channels, but it cannot obtain the highest performance of the MSR. Moreover, it is shown in this figure that the modified MMR

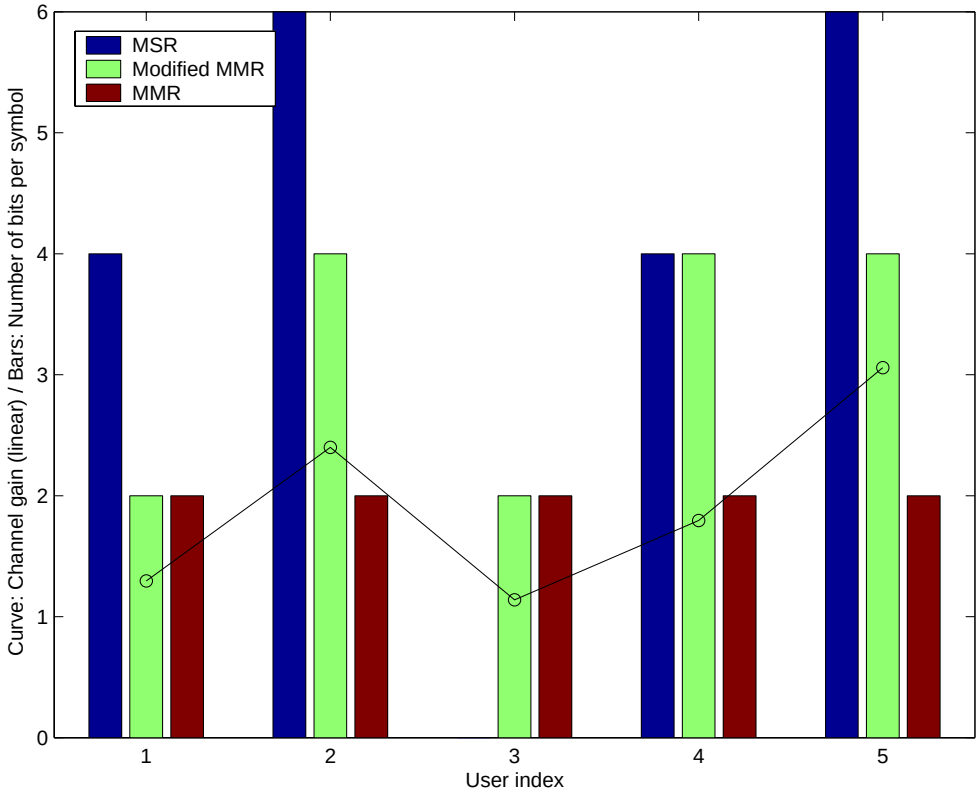


Figure 4.3: Output in terms of number of bits for a realization of the bit allocation strategies, $Q = 6, K = 5$. The solid line shows the channel gain α_k for each user.

yields a Pareto improvement over the simple MMR, that is, the performance of some users might increase without reducing the number of bits for any user. On the other hand, user 3 is only served by the MMR and the modified MMR, whereas the MSR sacrifices that user for the sake of the collective satisfaction. The modified MMR attains a closer performance to the MSR than the MMR.

The bit allocation techniques need to be compared to the use of a single mapping for all the users, which is simply a power distribution among the users. Among several alternatives involving fairness issues, the Uniform Power Allocation (UPA) is chosen. That means that the power is divided equally among the users, i.e. the kth user is assigned an equal fraction of power $\beta_k^2 = \frac{P_T}{|\mathcal{K}|}$, so that the AP does not care about their channel gain nor how it can improve the performance. Since this scheme optimizes only the power, which is limited, it cannot obtain such a high spectral efficiency as the bit allocation strategies. In any case, the algorithm presented next tries to fulfill the BER requirements of the communications and performs also an admission control. Table 4.3 summarizes the procedure, which yields the same results as the UPA with admission control in the previous chapter. First, it tries to fit all the users (steps 1 to 4). If any of them cannot fulfill the BER constraint, the user having a lower equivalent channel gain α_k is

1. Set $\mathcal{K} = \{1, \ldots, K\}$.
2. Build matrix $\mathbf{H}$ for the users in the set $\mathcal{K}$.
3. Compute $\alpha_k^2 = 1/\left[\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\right]_{k,k}, \forall k \in \mathcal{K}$.
4. Compute the power allocation $\beta_k^2 = \frac{P_T}{|\mathcal{K}|}$.
5. If $\mathrm{BER}_k \leq \mathrm{BER}_t, \forall k \in \mathcal{K}$ then the algorithm finishes.
   Else, drop out the active user having the worst channel $\mathcal{K} \leftarrow \mathcal{K} - \left\{k \in \mathcal{K} : \min_k \alpha_k^2\right\}$.
6. If $|\mathcal{K}| = \emptyset$, the algorithm finishes. Else, go to step 2.

Table 4.3: Uniform Power Allocation with BER constraints.

left out from the set of active users $\mathcal{K}$ (step 5). This is repeated until all the active users satisfy the BER constraint (step 5) or the active set is empty (step 6). Note that every time a user is left out from the active set, the equivalent channels α_k have to be recomputed again in step 3 since there are interactions among users. On the other hand, it is also interesting to evaluate the throughput of the single (best) user bit allocation, which can be seen as an opportunistic communication [79]. For this purpose, the best user among those active is selected, and it is provided with the highest number of bits per symbol that fulfill the BER constraint. Moreover, the exhaustive search among all the possibilities for the MSR has also been simulated, which provides the upper bound in performance.

For the first simulation, it is assumed that the AP has $Q = 4$ antennas, and that the number of active users at each time slot is a uniform integer number between 0 and 4. Then, the mean number of active users per slot is $\bar{K} = 2$. It is considered that a packet transmission requires 1 time unit using 4-QAM. Then, 1/2 and 1/3 time units are needed for 16-QAM and 64-QAM respectively, because they use the double and three times the number of bits of 4-QAM. The mean offered throughput can be obtained as the expectation of the ratio between the number of active users and the duration of a packet transmission. Therefore, the mean offered throughput per time slot is 2, 4, and 6, for 4-QAM, 16-QAM, and 64-QAM, respectively. These bounds are the maximum achievable rates for the three transmission modes, and they can be clearly observed in Figure 4.4. Since it is assumed that the communications have tight delay bounds, the packet is lost if any user is not scheduled, thus the instantaneous fairness is evaluated. That means that the author does not deal with possible retransmissions attempts or with the necessary size of the queues, which are left as further work, see Chapter 6.

First, it is depicted in Figure 4.4 the throughput delivered by the PHY in terms of correct packets per unit time vs. the SNR. It is shown that the adaptive schemes outperform those using

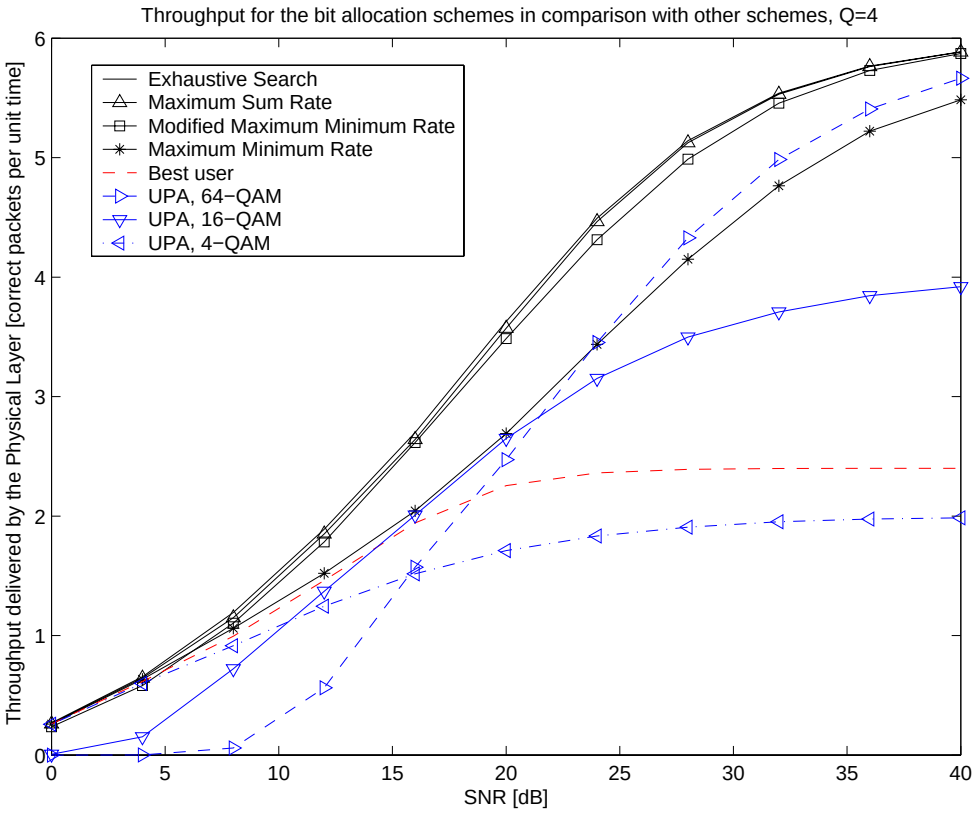


Figure 4.4: Throughput per slot delivered by the physical layer vs. the SNR for the bit allocation schemes compared to the communication with a single user, and the use of a fixed QAM signal mapping (4-QAM, 16-QAM, and 64-QAM) with Uniform Power Allocation.

a fixed signal mapping with the UPA because of the instantaneous adaptation to the channel conditions. The performance of the MSR is close to the exhaustive search with much lower complexity. The MMR looses in global performance because it is aimed at assigning the same number of bits for all the users. The allocation of the best user obtains a throughput between the 4-QAM and 16-QAM, but this depends on the number of active users in the cell, since the gain due to the multi-user diversity increases with the number of users. Particularly in this case, it would ideally achieve a rate of 3, but since there is a probability of 0.2 that no user is active, the throughput is slightly lower. The modified MMR provides the best balance of the trade-off between performance and complexity, as it approaches the exhaustive search while typically serving more users than the MSR.

The asymmetries in the distribution of the resources of the MSR and the MMR bit allocation schemes are depicted in Figure 4.5, where the mean vs. the approximation of the standard deviation of the throughput per user are plotted. Each point in the figures reflects a different situation in the SNR, ranging from 0 dB to 40 dB, in steps of 4 dB. It can be stated that more mean comes usually at the expense of a higher variance in the bit distribution. In fact, the

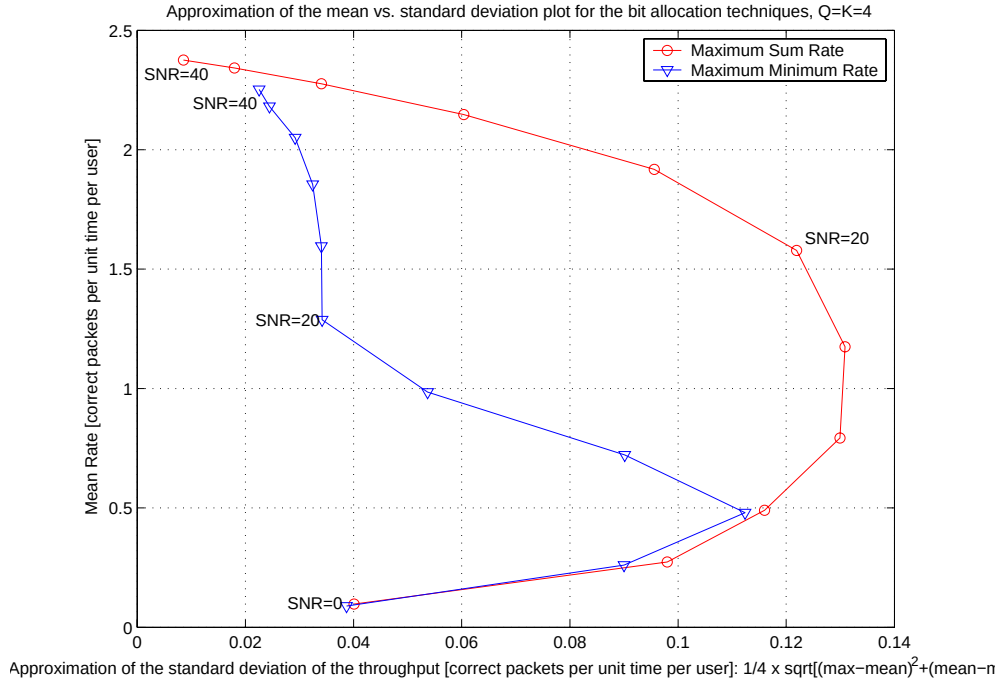


Figure 4.5: Approximation of the mean vs. standard deviation plot for the bit allocation techniques.

MMR tends to select the lower modulations, thus serving as many users as possible recalling the diversity-multiplexing trade-off. In fact, the constraint of assigning an equal number of bits to all the users punishes the global performance. Moreover, at high SNR this limitation yields a lower variance of the MSR with respect to the MMR. It shall be recalled here that the variance is due to the fact that if the BER constraint cannot be fulfilled, the algorithm for both the MSR and the MMR disregard the worst user to compute the corresponding bit allocation. For this reason, the approximation of the standard deviation tends to zero as the SNR increases, all the users will be served in the limit. For the bit allocation, it can also be concluded that cost functions optimizing the global performance yield uneven resource sharing, whereas solutions with an equal resource partitioning do not achieve a global optimization.

Finally, the number of antennas is set to $Q = 6$ in order to show the behavior of the total delivered throughput in terms of the number of users in the cell. Now, the number of users is fixed and the bit allocation mechanism has to optimize the performance. It is shown in Figure 4.6 the mean throughput per unit time vs. the number of active users. The increase in number of users does not saturate the performance of the exhaustive search, because it obtains the optimum user and bit allocation no matter the number of users. However, when $K = Q = 6$ the complexity could be prohibitive for real-time applications. The MSR saturates at a high number of users, because it suffers some losses due to the sequential removal of the users without testing if the already removed users can be fitted again. On the other hand, the bound on the

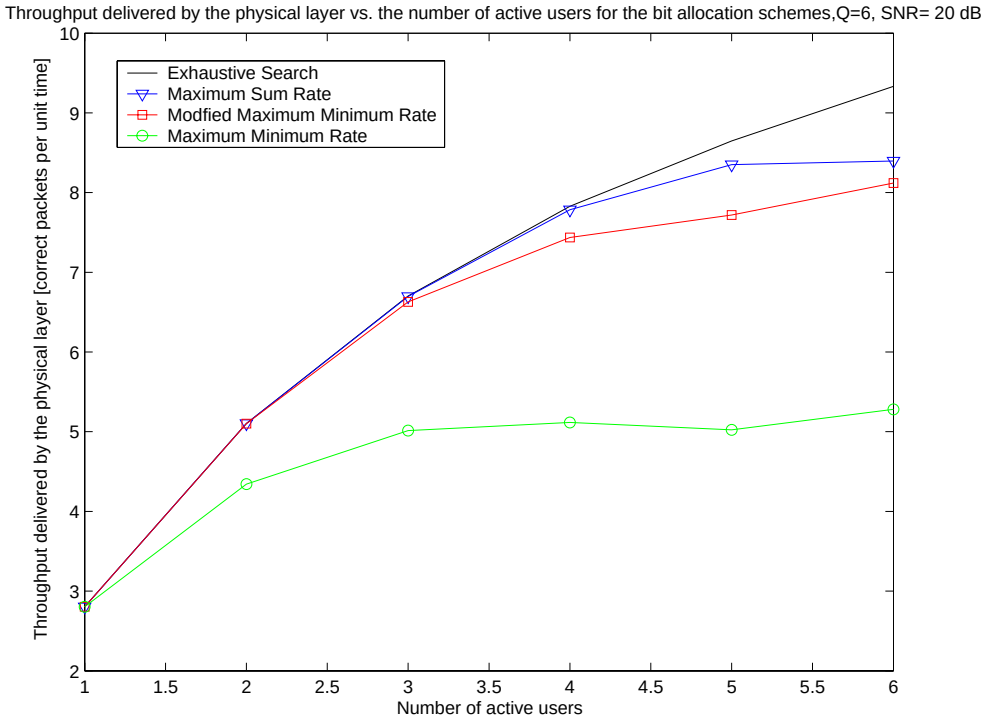


Figure 4.6: Throughput delivered by the physical layer vs. the number of active users for the MSR and MMR bit allocation strategies with $Q = 6$ antennas.

throughput of the MMR is very low since it always assigns the same number of bits to the users and cannot increase the performance over a threshold. The modified MMR yields an intermediate performance with a Pareto improvement over the MMR and reasonable complexity. Note that typically, the operating region of such schemes with realistic BER constraints is approximately lower than six users at this SNR, so the proposed MSR provides a reasonable performance.

4.3 Conclusions

Within the context of an spatial scheduling problem, this chapter has proposed practical algorithms for the case of a lower number of users than antennas. For a more realistic case where there exists a higher number of users than antennas, the reader is referred to the next chapter. The traditional maximization of the sum rate algorithm with integer bits per symbol is compared to other schemes that provide another perspective into fairness. On the one hand, all the users might be assigned the same constellation, which penalizes the global performance but introduces an equal resource sharing. On the other hand, the maximization of the sum rate (MSR) obtains the best global performance without taking much into account the worse users. The MMR has signaling advantages, as it is more precisely evaluated in the next chapter. Between the MMR and the MSR, the author proposes a modified version of the previous maximum minimum rate

algorithm, which approaches the performance bound of the MSR with lower complexity and a better balance of the trade-off between the global performance and the individual needs. This algorithm is based on a combination of *bit removal* and *bit filling* strategies. With the simulation results and the complexity involved, it is not clear which is the best-suited technique. Indeed, the trade-off between performance and complexity is a crucial point in the joint optimization of the tasks involving both the physical layer and the DLC.

Chapter 5

Practical bit loading for a multi-antenna broadcast OFDM

With the degrees of freedom obtained with the use of multiple antennas and multiple subcarriers, the system performance might be further enhanced, at the expense of a exponentially-increasing scheduling complexity. If the spatial scheduling in previous chapter was already complex, with multiple subcarriers it worsens. For instance, the channel was assumed to be constant in the previous chapter during the time slot, whereas the channel is generally frequency selective for OFDM. Since the scheduling with realistic integer signal mappings is NP-complete, suboptimum solutions based on the scalar product are shown in this chapter to be good candidates to yield a fast and realizable practical implementation of the clustering of users into groups, which is a new approach in the literature. This strategy aims to reduce the complexity involved in the computation of the used power that is dependent on the trace of a matrix inverse. Since it is a limiting factor for the system, a reduction in complexity is a key task that should be performed by the scheduler.

Indeed, this chapter is mainly dedicated to the clustering of users into groups, so it shall be noted that the approach that will be taken is based more on the physical layer than the previous chapter, and *new* issues such as complexity and signaling are addressed. Particularly in this chapter, several scheduling strategies are studied ranging from the optimum one with a highest complexity towards simple suboptimal solutions. Indeed, whereas in the previous chapter implementation was not a direct issue, this chapter is devoted to the reduction in complexity for the clustering techniques, and several questions shall be answered for the realistic implementation of an OFDM system with multiple antennas. First, a power reuse strategy to lower the computational complexity is proposed, which reduces the amount of operations that shall be performed. Then, it is shown that the amount of required signaling might be reduced by means of a user-subcarrier clustering or by using an scheme that forces an equal mapping for

all the users at the same subcarrier, which has an impact on the instantaneous fairness. As the reader might think, this scheme is a multi-carrier extension of the MMR in previous chapter. However, the approach is here different, since a lower amount of signaling is sought instead of the application of different fairness criteria. A new aspect is that the proposed strategies are evaluated for typical OFDM-based wireless LAN scenarios.

The remainder of this chapter is organized as follows. First, the introduction is given in the next section and the problem statement comes thereafter in Section 5.2. Then, the space-frequency scheduling problem is studied in detail in Section 5.3, following a different approach to that in the previous chapter. First, the NP-completeness of the objective function is shown, and then three more practical schemes are derived with different degrees of complexity. In Section 5.4, the main power (and bit) allocation strategy is described, together with simple power reuse schemes for comparison and two solutions to reduce the amount of signaling. Before concluding, Section 5.5 evaluates the proposed strategies with realistic simulation scenarios.

5.1 Introduction

Among other reasons, multi-carrier communications are widely deployed due to its ability to transform the frequency-selective channel impulse response into a set of parallel flat-fading subchannels [10]. On top of this, multiple antennas further enhance the system performance [6]. In this chapter, the focus is on the AP in the downlink of a OFDM-based Wireless LAN, such as [157]. As in [79] and throughout the dissertation, the single-antenna terminals are dumb, so that the multi-antenna AP has all the intelligence. With multiple users, allocating these users into subcarriers, antennas, and performing the integer bit allocation is NP-complete [89]. Therefore, the problem can be separated into two main parts, i) the user grouping (or clustering) and ii) the beamforming, power, and bit allocation for each group at all the subcarriers.

In a realistic scenario where the total number of users K exceeds the number of antennas Q, i.e. $K > Q$, related literature is scarce. The AP shall distribute the K users into groups of Q for each subcarrier, serve them by means of a Space Division Multiple Access (SDMA) scheme, and then perform the corresponding power (and integer bit) allocation. Because of the NP-completeness [89], suboptimum solutions are adequate, see e.g. [88] for an example of the uplink of a SDMA/TDMA system or [90] for an extension, as well as the references cited in the introduction of the previous chapter. In [91], the authors extend the best fit strategy proposed in [88] to take into account not only the spatial characteristics of the users, but also several Quality of Service (QoS) parameters. Differently to existing literature, the author derives in this chapter intelligent greedy solutions based on the **normalized scalar product** to form Q-user groups at each subcarrier with several degrees of complexity.

As it has been done throughout the dissertation, for each group at each subcarrier it

is meaningful to separate the problem into the transmit beamforming, and a power (and bit) allocation [131], which enables a cross-layer scheduler. A well-suited transmit processing technique is **Zero Forcing** (ZF) [84], which provides a reasonable performance loss with respect to i) optimum downlink beamforming [46], [26], ii) Dirty Paper coding, which obtains the maximum sum capacity [20], or iii) a regularized channel inversion with additional encoding that yields quasi-optimum performance [58]. Differently to [73], the transmit processing totally eliminates the inner-cell interference. With ZF, the AP creates parallel and orthogonal channels at each subcarrier, thus at most Q users can be served per subcarrier.

On top of the transmit beamforming, the AP shall perform the corresponding **(spatial) power and bit allocation**. Bit allocation is naturally derived in a power minimization (or bit rate maximization) under QoS constraints. Without multiple antennas, several mechanisms have been developed since [109]. For a review, see Chapter 1 or the previous chapter, but some shall be mentioned here: the computationally-efficient suboptimum schemes for Discrete Multi-Tone (DMT) in [111], [110], [115], and [113], and heuristic approaches for Orthogonal Frequency Division Multiple Access (OFDMA) are e.g. [117], [119], or [116]. On the other hand, the authors in [121] have extensively studied multi-user multi-carrier integer bit loading. Particularly in [121], the Levin-Campello algorithm is extended so as to minimize the total transmitted power with a target sum rate for all the users and a total power budget. The use of a ZF transmit beamforming forces the modification of the traditional single-antenna bit loading schemes, basically because the channels from the users change depending on those scheduled, as stated in Chapter 2.

Finally, when dealing with multi-user multi-carrier communications with multiple antennas, the associated signaling might be quite huge, thus practical schemes become relevant. Moreover, the (sometimes) complex solutions could be constraint in order to reduce the computational burden required in a real-time implementation. This is sought in this chapter with the practical insight into the power reuse schemes, as well as the user-subcarrier clustering to reduce the amount of overhead. Indeed, this chapter is devoted to the practical questions that arise in an implementation of a multi-user multi-antenna OFDM system.

5.2 Problem statement

The system is summarized in Figure 5.1. The purpose of the scheduling is to distribute the K users in the cell into groups of Q at every subcarrier, so that they can be served simultaneously by the Q-antenna AP with the selected SDMA scheme. In a realistic case, K is higher than Q. Since multi-carrier modulations are well-known [10], the signal model in this section is devoted exclusively to the frequency domain representation, although one should keep in mind that the channels at the adjacent subcarriers are correlated. Moreover, since the optimization procedures are performed instantaneously, the time index is omitted in (5.1), which is basically an extension

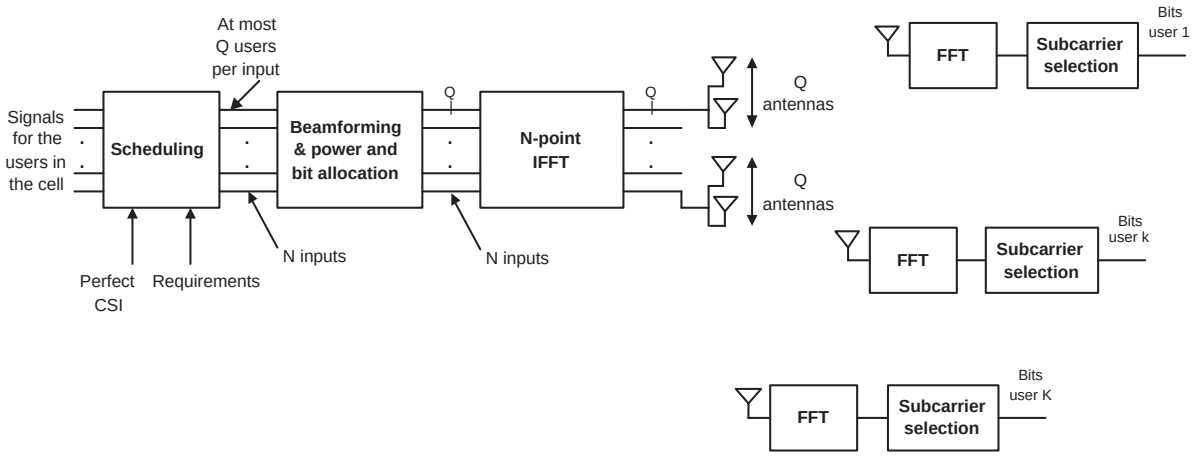


Figure 5.1: Block diagram of the system. With perfect Channel State Information (CSI), the Access Point (AP) clusters the users into groups, i.e. the scheduling task. For each group at each of the N subcarriers, the AP performs the Zero Forcing (ZF) transmit beamforming, and the corresponding power and bit allocation. The N inputs of at most Q users feed the IFFT block of the OFDM system. At the receivers, terminals shall demodulate only the signals at their subcarriers. It is assumed that this information is sent by the AP through a broadcast channel.

of the signal model that has been used up to this chapter. Assuming that the N subcarriers have their particular set of users $\mathcal{K}_n$ to be served, the signal at subcarrier n is given by

$$\mathbf{y}(\mathcal{K}_n) = \mathbf{H}(\mathcal{K}_n)\mathbf{B}(\mathcal{K}_n)\mathbf{s}(\mathcal{K}_n) + \mathbf{w}(\mathcal{K}_n) \ \in \mathbb{C}^{|\mathcal{K}_n|\times 1}, \tag{5.1}$$

where, differently to (2.4), here $\mathcal{K}_n$ emphasizes that the signal model is expressed for subcarrier n and for the users gathered in $\mathcal{K}_n$, which are served simultaneously. This set is in fact the objective of this chapter. The subindex k when needed refers to the kth user in the set $\mathcal{K}_n$, e.g. the kth position of $\mathbf{y}(\mathcal{K}_n)$ ($\mathbf{s}(\mathcal{K}_n)$) is the received (transmitted) signal for user k in the set $\mathcal{K}_n$ at the subcarrier n. $\mathbf{H}(\mathcal{K}_n)$ is the $|\mathcal{K}_n| \times Q$ complex flat-fading channel matrix at the nth subcarrier, the kth row of which contains the $1 \times Q$ vector of the channel gains for the kth user at the nth subcarrier, i.e. $\mathbf{h}_{k,n}^T$. This frequency domain representation is obtained by evaluating the Fourier transform of the channel impulse response of the L-tap channel vector $\mathbf{h}_k^t$ at the n subcarrier, i.e. $\mathbf{f}_n^H\mathbf{h}_k^t$, where $\mathbf{f}_n^H\mathbf{h}_k^t$, where $\mathbf{f}_n^H = [\,1 \ \exp(-j2\pi n/N) \ \ldots \ \exp(-j2\pi n(L-1)/N)\,]$. The channels $\mathbf{h}_k^t$ are supposed to be independent and perfectly known at the AP. Without loss of generality, the noise $[\mathbf{w}(\mathcal{K}_n)]_k$ at each subcarrier and for each user are independent zero-mean complex Gaussian random variables with variance $\sigma_{k,n}^2$. As usual, the transmit beamvectors are gathered in $\mathbf{B}(\mathcal{K}_n) = [\mathbf{b}_1(\mathcal{K}_n)\mathbf{b}_2(\mathcal{K}_n)\ldots\mathbf{b}_K(\mathcal{K}_n)] \in \mathbb{C}^{Q\times K}$.

As in the whole dissertation, with a ZF transmit beamforming, the signal model in (5.1) can be reduced to a very simple expression

$$y_k(\mathcal{K}_n) = \alpha_k(\mathcal{K}_n)\beta_k(\mathcal{K}_n)s_k(\mathcal{K}_n) + w_k(\mathcal{K}_n), \forall k \in \mathcal{K}_n,$$

in which $\alpha_k^2(\mathcal{K}_n)$ behave like independent central Chi-Square random variables with $2(Q-|\mathcal{K}_n|+1)$ degrees of freedom, i.e. $\alpha_k^2(\mathcal{K}_n) \sim \frac{1}{2}\chi^2_{2(Q-|\mathcal{K}_n|+1)}$. As shown in Figure 2.1, an intelligent scheduler should try to separate those users with similar channel vectors, because they deeply penalize the performance. The symbols are normalized QAM, for which the approximate BER is [149]

$$\mathrm{BER}_k(\mathcal{K}_n) \approx c_1 \exp\left(-\frac{c_2\gamma_k(\mathcal{K}_n)}{2^{m_k(\mathcal{K}_n)}-1}\right), \tag{5.2}$$

where $m_k(\mathcal{K}_n)$ is the number of bits per symbol of the mapping used at the nth subcarrier by user k, $c_1 = 0.2$, $c_2 = 1.6$, and $\gamma_k(\mathcal{K}_n)$ refers to the SNR for the kth user at subcarrier n, i.e. $\gamma_k(\mathcal{K}_n) = \alpha_k^2(\mathcal{K}_n)\beta_k^2(\mathcal{K}_n)/\sigma_{k,n}^2(\mathcal{K}_n)$. In an optimistic situation, the number of bits per symbol $m_k(\mathcal{K}_n)$ might change at every subcarrier and for every user. The general problem for this chapter consists of maximizing the total discrete achievable rate for all the users at all the subcarriers, i.e.

$$\max \sum_{n=0}^{N-1} \sum_{k\in\mathcal{K}_n} m_k(\mathcal{K}_n) \tag{5.3}$$

$$s.t. \quad \sum_{n=0}^{N-1} \sum_{k\in\mathcal{K}_n} \beta_k^2(\mathcal{K}_n) \leq P_T,\ 0 \leq n \leq N-1, \tag{5.4}$$

$$\mathrm{BER}_k(\mathcal{K}_n) \leq \mathrm{BER}_t,\ 0 \leq n \leq N-1,\ \forall k \in \mathcal{K}_n, \tag{5.5}$$

$$m_k(\mathcal{K}_n) \in \widetilde{\mathcal{M}},\ 0 \leq n \leq N-1,\ \forall k \in \mathcal{K}_n, \tag{5.6}$$

where, due to algorithmic issues, the set $\widetilde{\mathcal{M}}$ is defined as the union of the possible constellations together with 0 (no transmission), that is, $\widetilde{\mathcal{M}} = \{0\} \cup \mathcal{M}$, and the cardinality of $\widetilde{\mathcal{M}}$ is usually denoted by M. The set $\mathcal{K}_n$ gathers the users that are scheduled at the nth subcarrier, and it is assumed that all the users have the same BER requirements. Other options might include the minimization of the transmit power subject to BER or rate requirements. Besides, note that a combination of constraints is possible here. As an example, one could impose a constraint on the power per subcarrier, per user, or in total for the whole band as in (5.4). In fact, for some modifications proposed in this chapter, the constraint (5.4) is changed and a total power budget per subcarrier is imposed, see Section 5.4.2. The same holds for the BER constraints, and one could choose to limit the BER per subcarrier and user as in (5.5), as a sum for the subcarriers, in total for a given user, or as a whole for all the subcarriers and users. Due to the burstiness of data transmission and for an easy scalability, the user and subcarrier BER constraint in (5.5) is chosen. Moreover, it allows to find a direct expression for the power. If (5.2) is substituted into (5.5), it is easy to see that the power allocation reduces to

$$\beta_k^2(\mathcal{K}_n) = \frac{\sigma_{k,n}^2\left(2^{m_k(\mathcal{K}_n)}-1\right)}{c_2\alpha_k^2(\mathcal{K}_n)} log\left(\frac{c_1}{\mathrm{BER}_t}\right), \tag{5.7}$$

which can be inserted into (5.4) to compute the total used power. In the following section, it is shown that the general optimization objective in (5.3)-(5.6) is an NP-complete combinatorial

problem [89]. For this reason, the problem is divided into i) the user-subcarrier assignment (Section 5.3) which allocates Q users per subcarrier, and ii) the spatial power and bit allocation, which comes afterwards in Section 5.4.

5.3 Space-frequency multi-user scheduling

First, the NP-completeness of the problem in (5.3)-(5.6) is identified. Based on this, three techniques are proposed to allocate the users into groups with different degrees of complexity.

5.3.1 NP-completeness of the objective function

An NP-complete combinatorial problem is that belonging to a class that cannot be solved in polynomial time, in other words, the complexity increases exponentially with the number of variables. A little example serves for an intuitive proof. Simplifying the problem, imagine $N = 1$ subcarriers, $K = 10$ users present, $Q = 3$ antennas at the AP, and 3 available signal mappings, together with no transmission ($M = 4$ possibilities). The total number of combinations is $C_T = \binom{K}{Q} M^Q = 7680$. If N subcarriers are available, the total number of combinations is $(C_T)^N$, which might have unreasonable complexity even for low to moderate number of subcarriers.

It is shown in [88] by theory of graphs that minimizing the length of a SDMA/TDMA frame, while ensuring a minimum Signal to Interference to Noise Ratio (SINR) is NP-complete. The problem in this chapter is closely related to [88], but the NP-completeness follows from linear programming, in which several combinatorial optimizations are known to be NP-complete. Particularly, the problem in (5.3)-(5.6) is analogous to the well-known NP-complete Knapsack problem, then, it is also NP-complete. Therefore, several techniques are proposed next to group the users at the subcarriers with different degrees of complexity. It is shown that the scalar product might provide the best trade-off between performance and complexity in order to find the user-subcarrier allocation.

5.3.2 On the optimum user clustering

First, one shall focus on subcarrier n, and assume that $\mathcal{K}_n$ has a number of users $|\mathcal{K}_n|$ still lower than the number of antennas Q. The final goal is to select the best-suited user to be served together with the users already in the set $\mathcal{K}_n$. Imagine the noise and the number of bits is equal for all the users, then it follows from (5.7) that $\sum \beta_k^2(\mathcal{K}_n) \propto \sum 1/\alpha_k^2(\mathcal{K}_n) = tr(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))^{-1}$. The objective is to find the user m among those not in $\mathcal{K}_n$ such that $tr(\mathbf{H}(\mathcal{K}_n \cup m)\mathbf{H}^H(\mathcal{K}_n \cup m))^{-1}$ is minimum. For this purpose, $\mathbf{H}(\mathcal{K}_n \cup m) = [\mathbf{H}(\mathcal{K}_n)^T \mathbf{h}_m]^T$, then it is easy to see that

$$\mathbf{H}(\mathcal{K}_n \cup m)\mathbf{H}^H(\mathcal{K}_n \cup m) = \begin{pmatrix} \mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n) & \mathbf{H}(\mathcal{K}_n)\mathbf{h}_m^* \\ \mathbf{h}_m^T\mathbf{H}^H(\mathcal{K}_n) & \|\mathbf{h}_m\|^2 \end{pmatrix}.$$

Using the properties of block matrices, the $tr(\mathbf{H}(\mathcal{K}_n \cup m)\mathbf{H}^H(\mathcal{K}_n \cup m))^{-1}$ is equal to

$$tr(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))^{-1} + \frac{1 + \| \left(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n)\right)^{-1} \mathbf{H}(\mathcal{K}_n)\mathbf{h}_m^* \|^2}{\|\mathbf{h}_m\|^2 - \| \left(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n)\right)^{-1/2} \mathbf{H}(\mathcal{K}_n)\mathbf{h}_m^* \|^2}, \tag{5.8}$$

which is extremely useful to draw some conclusions about the user grouping. If the AP shall minimize the power, the exact expression of the trace obtained in (5.8) shall be used to find the user-subcarrier allocation with multiple antennas. According to (5.8), the best user, i.e. that with highest $\|\mathbf{h}_m\|^2$, should initialize the set $\mathcal{K}_n$, and then the AP shall iteratively fill this set with the user m such that (5.8) is minimum, until $\mathcal{K}_n$ contains Q users. The complexity is $\mathcal{O}(|\mathcal{K}_n|^3)$ if the inverse of the matrix is assumed to have this cost, which might be prohibitive even for a moderate number of users. Therefore, complexity reduction mechanisms are proposed.

5.3.3 Towards a simple user clustering

Given a selected candidate user m, the expression in (5.8) is minimized when the channel vector $\mathbf{h}_m^*$ is orthogonal to all the rows of $\mathbf{H}(\mathcal{K}_n)$, that is, $\mathbf{H}(\mathcal{K}_n)\mathbf{h}_m^* = \mathbf{0}$, where $\mathbf{0}$ is an all-zeros vector. This might be quite difficult to obtain in a practical situation, so a possible option is to try to select a candidate whose channel is as orthogonal as possible to the rows of $\mathbf{H}(\mathcal{K}_n)$. For this purpose, the AP might choose for subcarrier n the user m such that

$$m : \min_m \|\mathbf{H}(\mathcal{K}_n)\mathbf{h}_m^*\|^2, \tag{5.9}$$

until Q users fill $\mathcal{K}_n$. Although this approximation reduces complexity to $\mathcal{O}(|\mathcal{K}_n|)$, it might mask the users which use more power due to the implicit sum among the users in (5.9). An illustrative example will clarify the issue. Assume that the candidate to be inserted in $\mathcal{K}_n$ is the mth user, and that the set $\mathcal{K}_n$ already gathers two users. For this example, the elements of the channel vectors are assumed to be real and positive without loss of generality. One the one hand, $\mathbf{H}(\mathcal{K}_n)\mathbf{h}_m$ is $\mathbf{h}_1^T\mathbf{h}_m + \mathbf{h}_2^T\mathbf{h}_m$ in the general case. On the other hand, if $\mathbf{h}_m = \mathbf{h}_2$, then $\mathbf{H}(\mathcal{K}_n)\mathbf{h}_m$ becomes $\mathbf{h}_1^T\mathbf{h}_m + \|\mathbf{h}_m\|^2$. To compare both situations, the AP shall obtain the minimum value of $\mathbf{H}(\mathcal{K}_n)\mathbf{h}_m$ according to (5.9). Therefore, if $\mathbf{h}_1^T\mathbf{h}_m + \|\mathbf{h}_m\|^2 < \mathbf{h}_1^T\mathbf{h}_m + \mathbf{h}_2^T\mathbf{h}_m$, i.e. $\|\mathbf{h}_m\|^2 < \mathbf{h}_1^T\mathbf{h}_m$, then the situation $\mathbf{h}_m = \mathbf{h}_2$ would be preferred instead of the general case. Nevertheless, it is clearly the worst option because two users would have the same channel vector, which would increase the required power to infinity. Similar effects occur when the number of users is higher.

5.3.4 A Simple yet efficient user clustering

With the same complexity $\mathcal{O}(|\mathcal{K}_n|)$, the normalized scalar product of the channel vectors among users is proposed as a powerful tool for the scheduling [131]. In the literature, a concept of user separability is developed in [158], which refers to the fact that there exist beamforming vectors and powers for each user such that the SNR requirements are satisfied. Differently to this paper,

[158] is based on a maximum SNR beamforming and does not depend on the scalar product. Finally, a capture threshold on the SNR is described in [159] as an option for the scheduling.

It shall be proven that the normalized scalar product is a well-suited technique to design the schedulers. For this purpose, assume first that the channels $\mathbf{h}_k$ from the users in $\mathcal{K}_n$ are gathered in $\mathbf{H}(\mathcal{K}_n)$. If all users use the same mapping, it follows from (4.6) that $\sum_{k\in\mathcal{K}_n}\beta_k^2(\mathcal{K}_n) \propto \sum_{k\in\mathcal{K}_n} 1/\alpha_k^2(\mathcal{K}_n) = tr(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))^{-1}$, i.e. the used power depends on the inverse of the matrix $\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n)$. Then, in a general case the trace of this matrix can be expressed as

$$tr(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))^{-1} = \frac{adj(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))}{det(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))},$$

where $adj(\cdot)$ denotes the matrix of adjoints, so that the infinite value of the trace would be determined by the denominator. Consider the decomposition $\mathbf{H}(\mathcal{K}_n) = \left[\mathbf{H}_x^T \;\; \widetilde{\mathbf{H}}^T\right]^T$, where $\mathbf{H}_x = [\mathbf{h}_1\mathbf{h}_2]^T$ contains the users that are coming from closer zones of space (but are not colinear) and $\widetilde{\mathbf{H}}$ is the channel matrix of all other users. Using properties from block matrices,

$$det(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n)) = det(\mathbf{H}_x\mathbf{H}_x^H)det(\widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^H - \widetilde{\mathbf{H}}\mathbf{H}_x^H(\mathbf{H}_x\mathbf{H}_x^H)^{-1}\mathbf{H}_x\widetilde{\mathbf{H}}^H),$$

so if the first multiplying term is 0, then the determinant would also be 0. It can be verified that

$$det(\mathbf{H}_x\mathbf{H}_x^H) = \|\mathbf{h}_1\|^2\|\mathbf{h}_2\|^2 - |\mathbf{h}_1^H\mathbf{h}_2|^2 = \|\mathbf{h}_1\|^2\|\mathbf{h}_2\|^2(1-c_{1,2}^2), \tag{5.10}$$

where $c_{1,2} = \frac{|\mathbf{h}_1^H\mathbf{h}_2|}{\|\mathbf{h}_1\|\|\mathbf{h}_2\|}$ is the *normalized* scalar product between $\mathbf{h}_1$ and $\mathbf{h}_2$ defined accordingly, and its range is $0 \le c_{1,2} \le 1$: the lower bound occurs if $\mathbf{h}_1$ and $\mathbf{h}_2$ are orthogonal, whereas the upper bound when $\mathbf{h}_1 = \mathbf{h}_2$. Therefore, the scheduler should separate those users with similar channel vectors, and (5.10) justifies the use of the *normalized* scalar product $c_{1,2}$ as a powerful tool for this task. Several remarks are needed:

- With $Q = 2$ antennas, the term $c_{1,2}$ reflects the cosinus of the angle between $\mathbf{h}_1$ and $\mathbf{h}_2$.
- Differently to (5.9), the normalization factor in $c_{1,2}$, i.e. $\|\mathbf{h}_1\|\|\mathbf{h}_2\|$, takes into account that some channels might be better than others, so one truly concentrates on the *real* relative position among two channel vectors. As an example, if $\mathbf{h}_1 = p\mathbf{h}_2$, where p is real, then without normalization $\mathbf{h}_1^H\mathbf{h}_2 = p\|\mathbf{h}_1\|^2$, which is dependent on the constant p. If the normalization is included, $c_{1,2} = 1$, colinear vectors are detected no matter the value p.
- The cost $c_{1,2}$ is in fact determined by the users i,j with a highest $c_{i,j}$, because the determinant would be 0 if any two rows of the matrix $\mathbf{H}(\mathcal{K}_n)$ were the same.

A better insight into the problem is given with $K = 2$ users in $\mathcal{K}_n$, i.e. $k = \{1, 2\}$. Their channels are expressed as $\mathbf{h}_k$, and are gathered in $\mathbf{H}(\mathcal{K}_n) = [\mathbf{h}_1\mathbf{h}_2]^T$, so that

$$det(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n)) = \|\mathbf{h}_1\|^2\|\mathbf{h}_2\|^2 - |\mathbf{h}_1^H\mathbf{h}_2|^2, \tag{5.11}$$

1. Set $n = 0$. The users for subcarrier n are collected in $\mathcal{K}_n$, which has been initialized.
2. If $n = N - 1$, the algorithm finishes. Otherwise, do $n \leftarrow n + 1$.
3. Select $k : \min_k c_k(\mathcal{K}_n)$. Add user k to $\mathcal{K}_n$, i.e. $\mathcal{K}_n \leftarrow \mathcal{K}_n + k$.
4. If $|\mathcal{K}_n| < Q$, go to step 3. Otherwise, go to step 1.

Table 5.1: User clustering based on the scalar product.

so that the trace of $(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))^{-1}$, which determines the used power as it has been seen, is

$$tr(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))^{-1} = \frac{\|\mathbf{h}_1\|^2 + \|\mathbf{h}_2\|^2}{det(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))} = \frac{\|\mathbf{h}_1\|^{-2} + \|\mathbf{h}_2\|^{-2}}{1 - c_{1,2}^2}, \tag{5.12}$$

where the last step results from dividing numerator and denominator by $(\|\mathbf{h}_1\|\|\mathbf{h}_2\|)^2$. Then,

$$\lim_{\|\mathbf{h}_1\|\to\infty} tr(\mathbf{H}(\mathcal{K}_n)\mathbf{H}^H(\mathcal{K}_n))^{-1} = \frac{\|\mathbf{h}_2\|^{-2}}{1 - c_{1,2}^2} \tag{5.13}$$

which is bounded if $\mathbf{h}_1 \neq \mathbf{h}_2$. The same situation occurs if the limit is calculated for $\|\mathbf{h}_2\| \to \infty$. However, if one computes the limit when $\mathbf{h}_1 \to \mathbf{h}_2$, or equivalently when $c_{1,2} \to 1$, it yields

$$\lim_{c_{1,2}\to 1} tr(\mathbf{H}(\mathcal{K}_n))\mathbf{H}^H(\mathcal{K}_n))^{-1} = \infty. \tag{5.14}$$

Therefore, it is more critical to separate those users coming from the same zone of space rather than using the norm of their channel vector as a measure to allocate users, see e.g. [88]. These intuitive examples justify the use of the scalar product to allocate the users into the subcarriers.

With the previous results, the author develops a suboptimum but very simple real-time algorithm that allocates Q users per subcarrier, so that all of them could be served simultaneously by the SDMA scheme. The initialization procedure allocates the best user at every subcarrier, thus the set $\mathcal{K}_n$ has one user. Then, the AP shall fill the subcarriers until there are Q users pre-allocated per subcarrier. For this purpose, the associated cost of putting user k into group $\mathcal{K}_n$ is needed, which is determined by the maximum normalized scalar product among the users in $\mathcal{K}_n$ because it is the one that penalizes the performance, i.e.

$$c_k(\mathcal{K}_n) = \max_{k' \in \mathcal{K}_n} \frac{|\mathbf{h}_k^H \mathbf{h}_{k'}|}{\|\mathbf{h}_k\|\|\mathbf{h}_{k'}\|}, \forall k.$$

The procedure is summarized in Table 5.1. For each subcarrier, the AP selects the user having a minimum maximum associated cost among the users that are already pre-allocated to that subcarrier, see step 3. As shown, the highest cost is the one that most impacts the group performance. This procedure is repeated until Q users form the group for subcarrier n, see step

4. When a subcarrier is full, the AP repeats the same mechanism for the next subcarrier, see step 2. When the AP has Q pre-allocated users per subcarrier, it shall perform the bit allocation. A remark is needed: according to (5.8), the initialization of each set is done by selecting the user with highest $\|\mathbf{h}_k\|^2$, which is the one consuming less power if the set is empty.

5.4 Space-frequency power and bit allocation

In this section, the author describes in detail an extension to OFDM of the Maximum Sum Rate (MSR) for single carrier developed in the previous chapter, together with other practical schemes. Additionally, two strategies are proposed to lower the huge signaling needs.

5.4.1 Multi-antenna Multi-Carrier Maximum Sum Rate (MMSR)

It has already been noted that the space-frequency bit allocation has some differences with respect to traditional multi-carrier bit loading, e.g. the channels change when the users that are simultaneously served vary (see Figure 4.1). In realistic scenarios with several users, if the problem is not feasible, the AP has to perform the admission control, i.e. choose the users that will be served. In any case, the ultimate goal is to optimize the cell performance. The spatial bit allocation algorithm proposed yields a real-time close-to-optimum solution. Essentially, two strategies can be found in the single antenna bit loading literature in Chapter 1, namely *bit filling* and *bit removal*. The former adds a bit to the user/subcarrier providing the lowest increase in total power, and *bit removal* schemes remove the most penalizing bit until the power constraint is fulfilled. When the number of bits is an integer, linear programming solutions are convenient [120]. With multiple antennas, it is not possible to do strict bit filling algorithms, since the interactions among the users that are being simultaneously served are crucial, thus the user with better channel might not have a good channel when grouped together in an SDMA scheme.

Prior to the description of the algorithm, the number of bits for user k at subcarrier n is $m_k(\mathcal{K}_n)$, except for the lth, which changes the number to $m_l^j(\mathcal{K}_n)$ instead of $m_l^i(\mathcal{K}_n)$, where $m_l^i(\mathcal{K}_n) > m_l^j(\mathcal{K}_n)$. Then, extending (4.7), the power saving can be obtained approximately as

$$p_{l,n}\left(m_l^i(\mathcal{K}_n), m_l^j(\mathcal{K}_n)\right) = \begin{cases} \frac{\sigma_{l,n}^2}{\alpha_k^2(\mathcal{K}_n)}\left(2^{m_l^i(\mathcal{K}_n)} - 2^{m_l^j(\mathcal{K}_n)}\right) & \text{if } m_l^j(\mathcal{K}_n) \in \mathcal{M}, \\ \sum_{k\in\mathcal{K}_n} \sigma_{k,n}^2 \frac{2^{m_k(\mathcal{K}_n)}}{\alpha_k^2(\mathcal{K}_n)} - \sum_{k\in\widetilde{\mathcal{K}}_n} \sigma_{k,n}^2 \frac{2^{m_k(\widetilde{\mathcal{K}}_n)}}{\widetilde{\alpha}_k^2(\widetilde{\mathcal{K}}_n)} & \text{if } m_l^j(\mathcal{K}_n) \notin \mathcal{M}, \end{cases} \tag{5.15}$$

where the set $\widetilde{\mathcal{K}}$ gathers all the users but the lth and the equivalent channels $\widetilde{\alpha}_i$ are computed for the users in $\widetilde{\mathcal{K}}$. On the one hand, if $m_l^j(\mathcal{K}_n) \in \mathcal{M}$ the power saving comes directly from the subtraction of the power required by the use of $m_l^i(\mathcal{K}_n)$ and $m_l^j(\mathcal{K}_n)$ bits per symbol, recall the expression (4.6). On the other hand, if $m_l^j(\mathcal{K}_n) \notin \mathcal{M}$, then the value is not the exact saving because a user has been removed from the set of active users, thus the other channels $\widetilde{\alpha}_k^2(\widetilde{\mathcal{K}}_n)$

1. The set $\mathcal{K}_n$ is obtained by the user-subcarrier assignment.
2. Set $m_k(\mathcal{K}_n) = \max \mathcal{M}, \forall k \in \mathcal{K}_n, 0 \leq n \leq N-1$.
3. Build $\mathbf{H}(\mathcal{K}_n)$ and compute $\alpha_k^2(\mathcal{K}_n)$.
4. Compute $\beta_k^2(\mathcal{K}_n)$ according to (4.6), and the total used power $P_S = \sum_{n=0}^{N-1} \sum_{k \in \mathcal{K}_n} \beta_k^2(\mathcal{K}_n)$.
5. If $P_S \leq P_T$ or $\mathcal{K}_n = \emptyset, 0 \leq n \leq N-1$, then finish.
6. Compute $p_{k,n}\left(m_k^i(\mathcal{K}_n), m_k^j(\mathcal{K}_n)\right), 0 \leq n \leq N-1, \forall k \in \mathcal{K}_n$ according to (5.15) , where $m_l^i(\mathcal{K}_n)$ is the current mapping and $m_l^j(\mathcal{K}_n)$ the lower one in $\mathcal{M}$.
7. Select $\{n,k\}$: $\max_n \max_k p_{k,n}\left(m_k^i(\mathcal{K}_n), m_k^j(\mathcal{K}_n)\right)$, and reduce the number of bits $m_k^i(\mathcal{K}_n) \leftarrow m_k^j(\mathcal{K}_n)$.
8. **Only for MMMR**: $m_k^i(\mathcal{K}_n) \leftarrow m_k^j(\mathcal{K}_n), \forall k \in \mathcal{K}_n$.
9. If $m_k^i(\mathcal{K}_n) \in \mathcal{M}$, go to step 4. Else, $\mathcal{K}_n \leftarrow \mathcal{K}_n - k$, set $m_k(\mathcal{K}_n) = \max \mathcal{M}, \forall k \in \mathcal{K}_n$, and go to step 3.

Table 5.2: Space-frequency bit allocation: MMSR and MMMR.

become better. In such a case, the rest would have the chance to increase their modulation index. However, the reduction in complexity of this approximation in the cost function justifies its use.

The MMSR algorithm in Table 5.2 is based on a *bit removal* technique, but it is aided by a *bit filling* scheme, performed whenever the spatial channel gains α_k change, that is, whenever the set of active users varies. That table explains also the Multi-Carrier Maximum Minimum Rate (MMMR), see Section 5.4.3. Note that both are **space-frequency** bit allocation schemes. Briefly, the MMSR first tries to serve all the users in the set $\mathcal{K}_n$ obtained by the user-subcarrier clustering method with the highest modulation in $\mathcal{M}$ at steps 2-4. If the power constraint in (5.4) is not fulfilled (step 5), the scheduler decides which user among all the carriers should reduce the constellation size or which user should not be served. Since the number of bits shall be reduced, the scheduler selects the user having a maximum incremental cost of using a lower modulation, i.e. the user that saves more power if the bit rate is reduced, see steps 6 and 7. The AP reduces the number of bits of the selected user, and if it belongs to a possible constellation the algorithm goes again to step 4. Otherwise, it drops that user out from the set of active users (step 9) at the corresponding subcarrier, and the constellation size of all the remaining users in that subcarrier is set again to the maximum (step 2). The algorithm finishes when the power constraint is fulfilled or when the set of active users for all the subcarriers is empty (step 5).

5.4.2 Power reuse

As stated, the optimum solution to (5.3)-(5.6) is extremely complex, since it requires the **Exhaustive Search** (ES) among all the users, and all the possible combinations of bits per symbol for the selected users at all the subcarriers. If instead of the global power constraint as (4.3), a per subcarrier power constraint could be imposed, that is $\sum_{k\in\mathcal{K}_n}\beta_k^2(\mathcal{K}_n) \leq P_T/N,\ 0 \leq n \leq N-1$. In such a case, we could apply the ES for each subcarrier if the number of users K is low. However, since the bit allocation deals with integer signal mappings, some power would be wasted at each subcarrier, i.e. not all the P_T/N would be used. Therefore, the very simple power reuse scheme proposed next might be especially well-suited. The procedure is the following: the ES is performed at each subcarrier sequentially with an available power of $P_T/N + P_n$, where P_n gathers the accumulated unused power for the subcarriers computed previously to the nth. This strategy is related to the iterative (and more complex) technique for the eigenmodes of single link MIMO system in [124]. There, an optimum power redistribution routine is applied, which reallocates the unused power where it is most efficient. Our proposed scheme is yet very simple and already provides a very good performance, see Section 5.5 for details.

The same concept exposed for the ES is valid for the **MMSR**, which would then be performed independently (and sequentially) for each subcarrier according to the proposed power reuse. As a (low) benchmark, the simplest scheme is the pseudo-intelligent **random approach**, which selects a random combination of the set of active users and then applies the proposed MMSR spatial bit allocation for the users at each subcarrier with power reuse. Although the user selection is random, there is some intelligence located at the AP to perform the spatial bit allocation. For the comparisons in Section 5.5, **opportunistic** communications [156] shall also be considered. With multiple antennas, it means that the spatial diversity is used to enhance the receiver SNR, but there is no (user) multiplexing gain. Only the user with the best channel norm, that with $\max_k \|\mathbf{h}_k\|$, is scheduled for transmission at each subcarrier, with the highest number of bits per symbol satisfying the BER constraint. Here, the power reuse strategy plays also a key role.

5.4.3 Reducing the signaling needs

If $b = 2$ bits (3 mappings and no transmission) are required to transmit the desired constellation to the scheduled users, a total number of $b_T = Q \times N \times b$ bits is needed for signaling at every burst, which might be not negligible. As an example, if $Q = 3$ and $N = 64$, then $b_T = 384$ bits are needed for signaling. Note that in a OFDM symbol, a maximum number of $Q \times N \times M_{max} = 1152$ bits are delivered, where $M_{max} = 6$ is the highest constellation size in this particular case. This amount of overhead has relevance especially when the number of transmitted OFDM symbols is low, and might deeply penalize the performance of the system. Therefore, it is sought in this subsection to find practical and simple schemes to reduce the amount of signaling.

Clustering reuse

A possible option is to use the same users and the same number of bits at adjacent subcarriers, which is related to the coherence grouping in [94] and references therein, where the focus is the beamforming. Differently, bit loading is maintained here in clusters, but the beamforming is computed for each subcarrier. Essentially, the AP obtains the users and their power and bit allocation thanks to the MMSR performed on a per subcarrier basis, and the AP applies the same configuration to N_b adjacent subcarriers. Therefore, the space-frequency bit allocation shall be done once every N_b subcarriers, and the bits for signaling are reduced as $b_T^{'} = b_T/N_b$. Following the previous examples, if $N_b = 2$, then $b_T^{'} = 192$, and if $N_b = 4$, then $b_T^{'} = 96$.

Multi-carrier Maximization of the Minimum Rate (MMMR)

Another option to reduce the signaling by a maximum factor of Q is to force an equal signal mapping for all the users at the same subcarrier. It might happen that not all the users are allocated, thus the gain of Q reflects only an upper bound. In fact, by using the same mapping for all the users it is guaranteed that the users being served receive the same rate. This ensures the fairness among users if they are homogeneous (or pay the same price for the service). In some sense, the AP maximizes the number of users that are served but the global performance might be penalized [129]. Fairness issues have been rarely evaluated in the literature, see e.g. [119], where without multiple antennas, the author proposes an iterative technique for OFDMA aiming to provide fairness guarantees among users by swapping allocated subcarriers among users, or bits among the same user. The simplified approach in [57] in fact guarantees fairness among users in terms of rate, but it is proposed for the sake of simplicity in the beamforming. A theoretical study of fairness in multi-antenna multi-user channels has been conducted in Chapter 2.

The alternative in this subsection can be expressed as a max-min problem according to

$$\begin{aligned}
&\max_{m_k(\mathcal{K}_n)} \min_k m_k(\mathcal{K}_n),\ 0 \leq n \leq N-1, \\
s.t.\ &\sum_{n=0}^{N-1} \sum_{k \in \mathcal{K}_n} \beta_k^2(\mathcal{K}_n) \leq P_T,\ 0 \leq n \leq N-1, \\
&\mathrm{BER}(\mathcal{K}_n) \leq \mathrm{BER}_t,\ 0 \leq n \leq N-1,\ \forall k \in \mathcal{K}_n, \\
&m_k(\mathcal{K}_n) \in \widetilde{\mathcal{M}},\ 0 \leq n \leq N-1,\ \forall k \in \mathcal{K}_n.
\end{aligned}$$

The optimum solution for this problem implies the exhaustive search among all the users and all the number of bits. However, the complexity might be slightly lower than the MMSR, since now an equal number of bits is imposed to all the scheduled users. Based on this, the Multi-Carrier Maximization of the Minimum Rate (MMMR) algorithm in Table 5.2 consists essentially of the same steps as the MMSR, but in this case, all the users at subcarrier n lower their modulation size when the selected user is at the same subcarrier, see step 8.

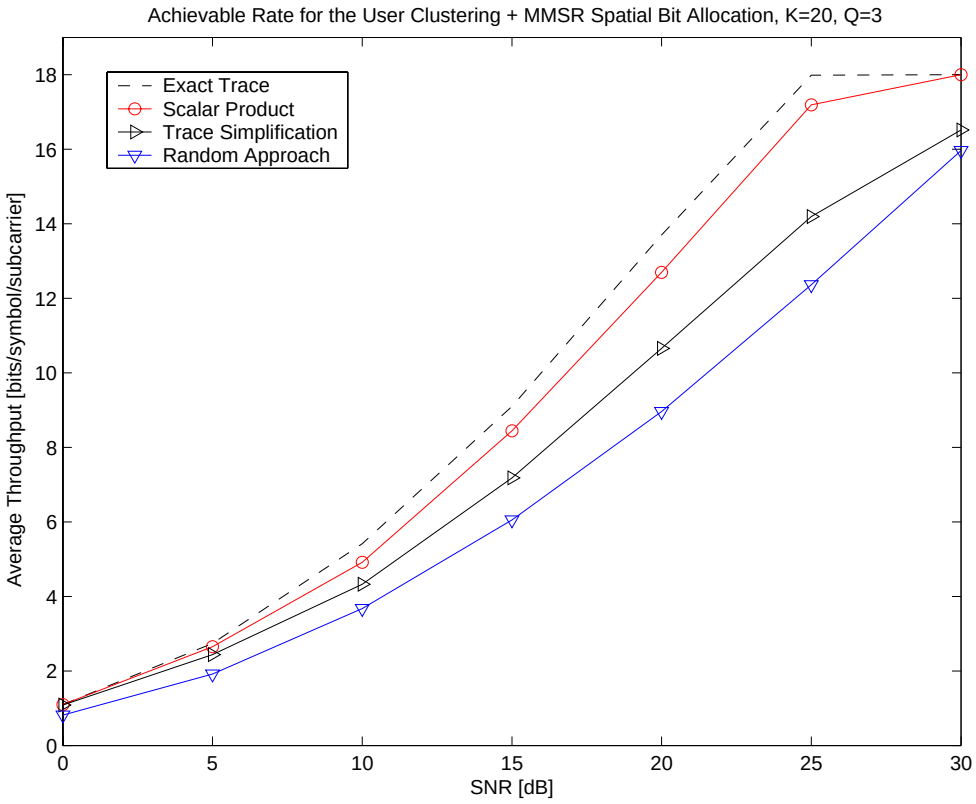


Figure 5.2: Comparison of the proposed user clustering schemes with MMSR spatial bit allocation in terms of throughput vs. the ratio $SNR = P_T/\sigma^2$ with $K = 20$ and $Q = 3$.

5.5 Performance evaluation

In this section, simulation results for the proposed schemes are shown, and particularized for channel model A [160], that is, a typical office environment with 50 ns average rms delay spread for OFDM-based Wireless LAN. Only for simulation purposes, the noise power is equal for all subcarriers $\sigma^2_{k,n} = \sigma^2, \forall n, \forall k$, and the SNR is defined as the ratio P_T/σ^2. Differently to current Wireless LAN standards [157], all the $N = 64$ subcarriers contain useful information. QAM constellations are considered with $\mathcal{M} = \{2, 4, 6\}$ bits per symbol.

The first simulation evaluates the user-subcarrier assignment schemes in Section 5.3, followed by a MMSR bit allocation. It is plotted in Figure 5.2 the average throughput in terms of number of bits per symbol per subcarrier vs. the SNR, when $K = 20$ and $Q = 3$. It is observed that the computation of the exact trace yields the best performance at the expense of a high computational complexity. The number of operations is reduced with the trace simplification, but one can see in this figure the problems that have been commented in Section 5.3 because of the bad performance. The strategy based on the scalar product clearly outperforms the trace simplification, and there is, in any case, a gain with respect to the random approach. To sum

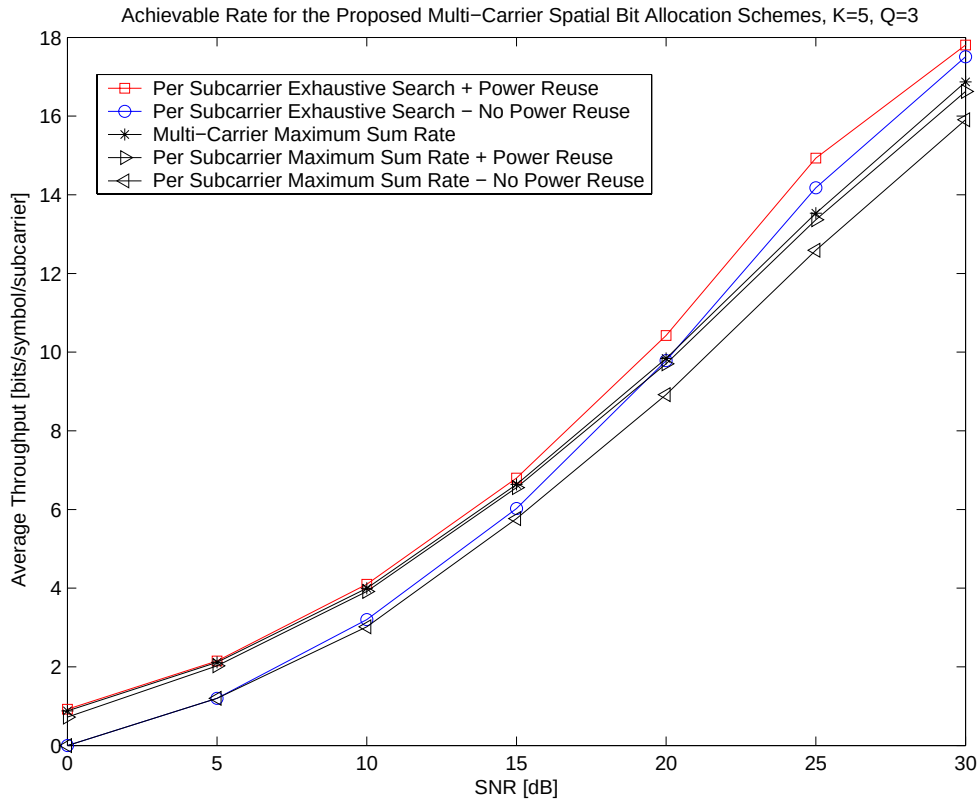


Figure 5.3: Comparison of the proposed SDMA-OFDM bit allocation techniques in terms of throughput vs. the ratio $SNR = P_T/\sigma^2$. In this case, $K = 5$ and $Q = 3$.

up, the scheme based on the scalar product offers the best trade-off between performance and complexity, therefore it is chosen for the next simulations.

Second, the proposed space-frequency maximum sum rate is evaluated with respect to the simpler schemes based on power reuse. It is shown in Figure 5.3 the average throughput at the physical layer in terms of number of bits per symbol per subcarrier vs. the SNR, when $K = 5$ and $Q = 3$ because of complexity for the ES. One observes that the exhaustive search scheme, with or without power reuse, outperforms the other methods in the high signal to noise ratio range at the expense of a prohibitive computational complexity when the number of users increases. The performance of the MMSR has practically no degradation at low SNR, but differences are higher when the SNR increases. Clearly, the simple scheme with power reuse is very close to the globally computed maximum sum rate. Note that if the problem is separated into subcarriers without power reuse, a high amount of power is wasted, and the rate is penalized.

Third, a more realistic scenario is simulated, with $K = 20$ users and the same number of antennas. Figure 5.4 shows the average throughput at the physical layer in terms of number of bits per symbol per subcarrier vs. the SNR. Again, the performance with the globally computed algorithms is very close to the use of the simple power reuse scheme. If the AP imposes equal

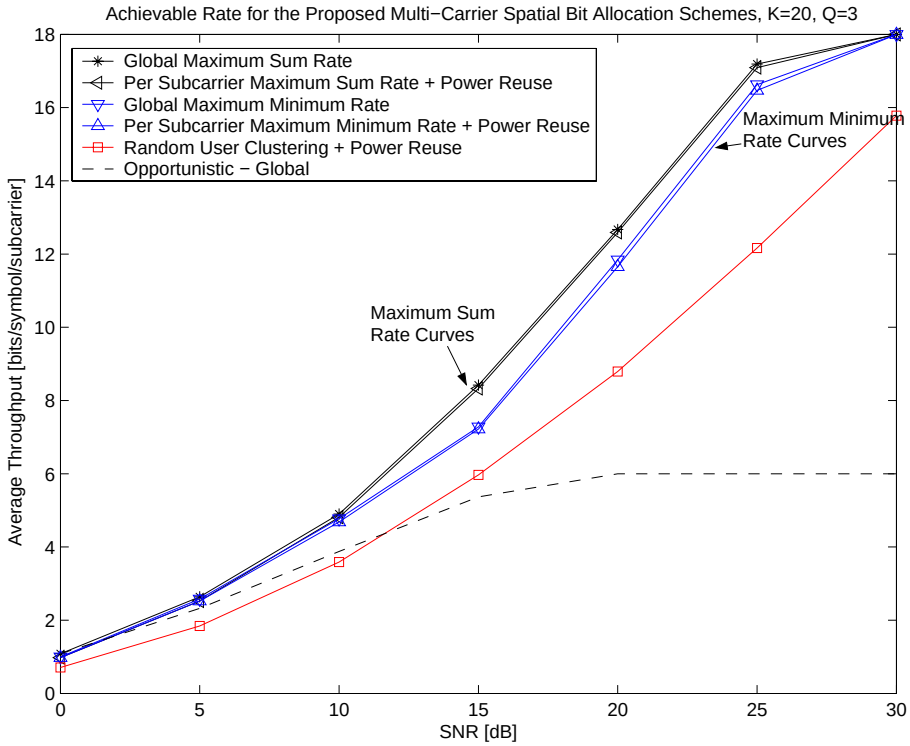


Figure 5.4: Comparison of the proposed SDMA-OFDM spatial bit allocation techniques in terms of throughput vs. the ratio $SNR = P_T/\sigma^2$ in a more realistic case, $K = 20$ and $Q = 3$.

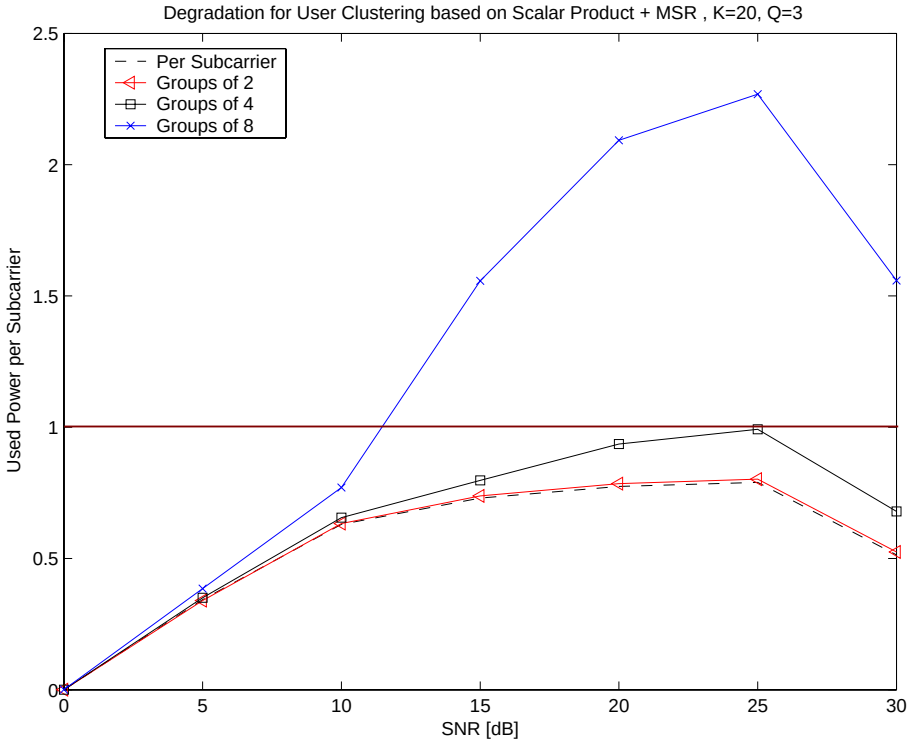


Figure 5.5: Power degradation by the subcarrier clustering for the based on the scalar product with MSR spatial bit allocation, with $K = 20$ and $Q = 3$.

constellation for all users at a certain subcarrier, i.e. the MMMR, the performance is penalized with respect to the MMSR, but the amount of signaling is also reduced and the AP guarantees fairness among the users. The particular choice depends on the AP. In any case, a noticeable gain is achieved with respect to a random selection of users. Finally, one can see that opportunistic communications yield a low throughput at high SNR because of the limitation of the scheduling of a single user. However, at low SNR where the noise is dominates the performance is better, because typically a single user is scheduled per subcarrier at most. Therefore, at low SNR the MMMR is equivalent to the MMSR.

Finally, in Figure 5.5 the average power per subcarrier is evaluated with respect to the SNR, with $K = 20$ and $Q = 3$. The objective is to see the penalty in power by using the same users and number of bits in adjacent subcarriers, which is important for practical considerations. The lowest dashed curve refers to the MMSR performed on a per subcarrier basis. The curves above this show the used power when the same user/mapping configuration is used for $N_b = \{2, 4, 8\}$ subcarriers. Note that in this particular simulation the power constraint per subcarrier is set to 1, so that groups of 4 subcarriers might be used at most. The degradation is negligible when the same setup is applied at the adjacent subcarrier. The optimum size of the grouping depends on the coherence bandwidth of the channel. Therefore, mechanisms to detect the surrounding environment seem adequate [161], which would allow to adapt the grouping instantaneously.

5.6 Conclusions

The practical implementation of bit allocation algorithms has been studied in this chapter, as well as techniques performing a combination of space diversity and frequency diversity introduced by the OFDM modulation. The author has reviewed the particular issues that arise when the spatial dimension is added to the problem, and have shown that the pursued objective in OFDM systems is an NP-complete combinatorial problem. Therefore, suboptimal solutions are adequate, and it is shown that the scalar product might be a good candidate for a fast and realizable cross-layer scheduler. Moreover, it has been compared to other simpler strategies.

After that, the chapter is devoted to the practical issues that arise in a realistic system implementation. First, a simple power reuse scheme is proposed to reduce the complexity of the space-frequency bit allocation algorithms, which yields quasi-optimum performance. After that, two mechanisms have been proposed to reduce the huge signaling needs of the multi-user multi-carrier spatial bit allocation schemes, which have fairness implications in the design of the scheduler. One is to force an equal signal mapping for all the users at a given subcarrier, whereas the other one assumes that the same configuration in terms of users and number of bits can be deployed for the adjacent subcarriers. There is a trade-off between performance and complexity in terms of amount of signaling.

Finally, it has been shown by means of realistic simulations for typical indoor Wireless LAN environments, that simple practical schemes are adequate in the design of such a complicated cross-layer scheduler. Besides the trade-off between performance and complexity, there exists the trade-off between performance and signaling, and between global and the individual needs. As it has been reviewed throughout the dissertation, the choice (and thus the ultimate quality of the communications) strongly depends on the criterion of the Access Point.

Chapter 6

Conclusions and further work

Perhaps the most relevant conclusion throughout the dissertation is that fairness issues should be carefully considered in the design of a system. Moreover, the objective is to develop a fairness framework (and its tools) to analyze multi-user communications. Particularly, the consideration of fairness impacts not only on the design of explicit schedulers at the access point, such as the power or bit allocation, but also on the implicit part which is the beamforming criterion. In this work, fairness has been only considered in the short-term, thus a study and a comparison to long-term fairness should be part of the further work. Next, the main conclusions for each topic are stated, followed by possible further work.

Chapter 2 deals with the multi-antenna techniques, not only at the transmitter side, but also for the cooperative scheme between transmitter and receivers. The main contributions can be summarized into:

- A fairness analysis of three antenna array techniques, which yields the perhaps surprising result that more mean comes at the expense of an uneven distribution of the resources. In other words, if the optimum technique is selected, the variance in the performance will be higher than for a worse-behaved technique. Therefore, in bursty transmissions of homogeneous users, it might not be clear which technique shall be selected.

- The fairness point of view is based on an approach that was originally deployed for portfolio selection, which constitutes a new way of looking at the results especially at the physical layer and overcomes the relative nature of the fairness indices in the literature. In this sense, fairness is no longer an index, but it should rather be described as a plot.

The work concerning multi-antenna techniques could be extended to other non-orthogonal schemes, such as the minimum mean square error approach or the optimal schemes explained in the first chapter, where QoS constraints are added. Moreover, the type of analysis that is conducted could also be deployed for other degrees of freedom in a multi-user scenario, since the

type of plot reflects the mean and the variance of any scheme. Particularly, the evaluation of receiver techniques might be of interest. Related to the Coase theorem that is briefly outlined in this chapter, it would be an interesting subject of research to try to design mechanisms to exchange efficiently information among the terminals so as to swap the allocated resource among the terminals. For instance, one terminal could give to another a part of its allocated resource, provided that this one gives that part back in another situation. The analogies between Economics and Telecommunications should be further explored.

The main contributions of **Chapter 3** in the power allocation and admission control mechanisms are the following:

- An instantaneous fairness comparison among four widely-deployed power allocation techniques in the literature, although the fairness is made explicit by the cost function. The fairness analysis implies not only to show the mean or sum performance, but also the behavior of the best and the worst user.

- An asymptotic analysis (in the high SNR regime) of the power allocation techniques is conducted. It is stated that the uniform power allocation tends to the maximum sum rate at high SNR, and the minimum sum BER tends to behave like the scheme providing equal BER to every user. Moreover, the rate-based methods yield always a higher sum rate than the BER-based methods, although depending on the objectives of the scheduler, a different technique could be selected. By minimizing and upper bound on the sum BER, the UPA comes up as the solution.

- A comparison of the best technique for each of the metrics treated in this chapter, namely rate, BER, and utility (based on game theory). It is stated that the utility-based techniques might yield an unacceptable performance in terms of BER, which might be a hard drawback of such techniques in real systems. To the best of the author's knowledge, such a comparison has not been conducted in the literature. However, pricing mechanisms are shown to be a good tool to control the individual behavior of the terminals in the cell.

- A theoretical and practical study of the admission control in the multi-user system. A new technique is proposed, which is between the uniform power allocation and the equal rate and BER scheme, which lie on the extreme points of the fairness balance. The illustration of fairness is made by the Lorentz curves, which originally serve as a basis for the computation of the Gini index. This is a widespread index to measure the degree of equality of a resource distribution in other fields of research.

This chapter is focused on the instantaneous distribution of the resource, which is the limited output power. As part of the further work, such analysis shall be conducted for a bigger time scale of fairness, e.g. in the long term or a combination of short-term and long-term constraints

to ensure a fair allocation independently of the time scale. Moreover, although the variety of objective functions and useful metrics is waste, for the sake of conciseness only a few have been compared. Therefore, schemes based on the mean squared error, the SNR, and other metrics shall also be compared in terms of fairness. It is a rather challenging work to find a fairness index (or plot) that shows the behavior of the techniques in terms of several metrics, and not only a single one. Regarding game theory, other options for the utility function should be evaluated, since the benefits of such schemes are on the simplicity of the implementation. Moreover, the pricing mechanism shall be investigated as an important part of future communication systems.

After that, the bit allocation strategies are developed in **Chapter 4**. The drawn conclusions are:

- The bit allocation strategies can be extended to the spatial dimension. However, some practical concerns change the traditional schemes developed for multi-carrier transmission. For instance, the instantaneous channel gains for the users vary depending on the subset of users that are scheduled, which impacts the bit distribution mechanism.

- The bit allocation problem can also be seen as a resource distribution, thus fairness issues could also be studied. There are two extreme solutions, the maximization of the sum rate and the maximization of the minimum rate. In this chapter, a new scheme is proposed that yields an intermediate behavior between them, and provides a good balance between performance and complexity.

In this sense, other practical strategies for the bit allocation might be found, that balance in a different way not only the trade-off between performance and complexity, but also between the global performance and the individual needs. For the bit allocation strategies, more knowledge could be added to the algorithms in order to reduce the available search space, which is a matter of realistic implementations.

Finally, **Chapter 5** is devoted to the practical combination of space diversity with frequency diversity. The main contributions are the following:

- The spatial bit allocation strategies are extended to take into account the frequency dimension. Again, there is a trade-off between the global performance and the individual behavior, but also signaling plays here a very important role.

- Among other options, the scalar product is shown to provide a good tool for the separation of users into groups, which attains a quasi-optimum performance with reduced complexity.

- The power reuse strategy might reduce the computational load in practical scenarios, since the bit allocation strategies can be performed on a per subcarrier basis. In such a case, the unused power can be allocated in other frequency bins.

- To lower the signaling needs, two strategies have been proposed, one is the clustering reuse, which selects the same configuration at adjacent subcarriers, and the other one is to assign an equal number of bits to all users at the same subcarrier, which has fairness implications.

This work on space-frequency scheduling is rather new should be further extended in order to take into account as many dimensions as possible among time, frequency, code, and space. Other mechanisms might be necessary to reduce the signaling and computational burden in such a case. Since the problem is NP-complete, suboptimal solutions are adequate and might be subject of subsequent work, although the scalar product has been shown to provide a remarkable tool. In any case, the NP-completeness of the objective function leaves room to a number of proposals.

In this dissertation, perfect channel knowledge has been assumed, therefore, subsequent work might include imperfections, not only due to the channel knowledge, but also from the hardware and other implementation constraints. In this sense, techniques would suffer from additional losses that might be interesting to quantify. Furthermore, the assumption of independence between channel vectors from the users might be too optimistic in indoor scenarios, thus the impact of correlation on the performance of the proposed techniques should be evaluated.

Moreover, the system should include multiple antennas at both sides of the communication link and not only at the transmitter side. This problem is more complicated, especially when the total number of receive antennas is higher than the number of transmit antennas. In such cases, not only a user selection shall be performed, but also the antenna subset selection.

Finally, real cross-layer information shall be treated, such as delay, jitter, throughput with concrete ARQ schemes, length of the queues, due date of the packets, etc. This topic is in its infancy for the moment, thus a number of research contributions could be made if it is studied carefully. However, it is not an easy task because of the high number of degrees of freedom that are available and should be efficiently used.

To conclude (and to start again): in this dissertation the physical layer has benefit from the concepts at the DLC, but a real cross-layer design shall be made to get the most of any wireless system.

Bibliography

[1] S. Shakkottai ,T.S. Rappaport, and P.C. Karlsson, "Cross-layer Design for Wireless Networks," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74 – 80, October 2003.

[2] V. Kawadia and P.R. Kumar, "A Cautionary Perspective on Cross Layer Design," *submitted to IEEE Wireless Communication Magazine*, July 2003, available at http://black.csl.uiuc.edu/∼prkumar/.

[3] C.K. Toh, "Maximum Battery Life Routing to Support Ubiquitous Mobile Computing in Wireless Ad-hoc Networks," *IEEE Communications Magazine*, vol. 39, no. 6, pp. 138 – 147, June 2001.

[4] R.G. Gallager, "A Perspective on Multiaccess Channels," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 124 –142, March 1985.

[5] I.E. Telatar and R.G. Gallager, "Combining Queuing Theory with Information Theory for Multiaccess," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 963–969, August 1995.

[6] G.J. Foschini, M.J. Gans, "On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, March 1998.

[7] A.J. Paulraj, D.A. Gore, R.U. Nabar, and H. Bolcskei, "An Overview of MIMO Communications: a Key to Gigabit Wireless," *Proceedings of the IEEE*, vol. 92, no. 2, pp. 198 – 218, February 2004.

[8] S.N. Diggavi, N. Al-Dhahir, and A. Stamoulis, "Great Expectations: The Value of Spatial Diversity in Wireless Networks," *Proceedings of the IEEE*, vol. 92, no. 2, pp. 219 – 270, February 2004.

[9] S. Shakkottai and T.S. Rappaport, "Research Challenges in Wireless Networks: a Technical Overview," *in Proc. of the 5th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, October 2002.

[10] Z. Wang and G.B. Giannakis, "Wireless Multicarrier Communications," *IEEE Signal Processing Magazine*, vol. 17, no. 3, pp. 29–48, May 2000.

[11] H. Weingarten, Y. Steinberg, and S. Shamai, "The Capacity Region of the Gaussian MIMO Broadcast Channel," *in Proc. of the Conference on Information Sciences and Systems (CISS)*, March 2004.

[12] D. Fudenberg and J. Tirole, *Game Theory*, Cambridge, MA: MIT Press, Ed., 1991.

[13] E. Telatar, "Capacity of Multi-antenna Gaussian Channels," *European Transactions on Telecommunications*, vol. 10, pp. 585 – 595, November 1999.

[14] A. Goldsmith, S.A. Jafar, N. Jindal, and S. Vishwanath, "Capacity Limits of MIMO Channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 864–702, June 2003.

[15] G.R. Raleigh and J.M. Cioffi, "Spatio-Temporal Coding for Wireless Communications," *IEEE Transactions on Communications*, vol. 46, pp. 357–366, March 1998.

[16] E. Biglieri, J. Proakis, and S. Shamai, "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, October 1998.

[17] H. Boche and M. Schubert, "A General Duality Theory for Uplink and Downlink Beamforming," *in Proc. of the IEEE Vehicular Technology Conference (VTC) Fall*, September 2002.

[18] T.M. Cover, "Comments on Broadcast Channels," *IEEE Transactions on Information Theory*, vol. 4, no. 6, pp. 2524–2530, October 1998.

[19] N. Jindal and A. Goldsmith, "Dirty Paper Coding vs. TDMA for MIMO Broadcast Channels," *submitted to IEEE Transactions on Information Theory*, June 2004, available at http://wsl.stanford.edu/Publications/Nihar.

[20] G. Caire and S. Shamai, "On the Achievable Throughput of a Multiantenna Gaussian Broadcast Channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691 – 1706, July 2003.

[21] W. Yu, D.P. Varodayan, and J.M. Cioffi, "Trellis and Convolutional Precoding for Transmitter-Based Interference Pre-Subtraction," *submitted to IEEE Transactions on Communications*, March 2004, available at www.comm.toronto.edu/∼weiyu.

[22] M.H.M. Costa, "Writing on Dirty Paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439 – 441, May 1983.

[23] 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, no. 10, pp. 2658 – 2668, October 2003.

[24] P. Viswanath and D.N.C. Tse, "Sum Capacity of the Vector Gaussian Broadcast Channel and Uplink Downlink Duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912 – 1921, August 2003.

[25] W. Yu and J.M. Cioffi, "Sum Capacity of a Gaussian Vector Broadcast Channel," *accepted in IEEE Transactions on Information Theory*, available at http://www.comm.toronto.edu/∼weiyu.

[26] F. Rashid-Farrokhi, K.J. Ray Liu, and L. Tassiulas, "Transmit Beamforming and Power Control for Cellular Wireless Systems," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1437–1449, October 1998.

[27] E. Visotsky and U. Madhow, "Optimum Beamforming Using Transmit Antenna Arrays," *in Proc. of the IEEE 49th Vehicular Technology Conference (VTC)*, May 1999.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley Series in Telecom., J. Wiley & Sons, 1991.

[29] W. Yu, W. Rhee, S. Boyd, and J.M. Cioffi, "Iterative Water-filling for Gaussian Vector Multiple Access Channels," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 145 – 152, January 2004.

[30] W. Yu, "Competition and Cooperation in Multiuser Communication Environments," *Ph.D. Thesis, Stanford University*, 2002.

[31] S. Boyd and L. Vanderberghe, *Convex optimization.* available online at http://www.stanford.edu/∼boyd/bv_cvxbook.pdf, 2003.

[32] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the Duality of Gaussian Multiple-Access and Broadcast Channels," *IEEE Transactions on Information Theory*, vol. 50, no. 2, pp. 768–783, May 2004.

[33] S. Vishwanath, G. Kramer, S. Shamai, S. Jafar, and A. Goldsmith, "Capacity bounds for Gaussian Vector Broadcast Channels," *DIMACS Workshop on Signal Processing for Wireless Transmission*, 2003.

[34] N. Jindal, S.A. Jafar, S. Vishwanath, and A. Goldsmith, "Sum Power Iterative Water-filling for Multi-Antenna Gaussian Broadcast Channels," *in Proc. of Asilomar Conference on Signals, Systems, and Computers*, November 2002.

[35] H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink Capacity Evaluation of Cellular Networks With Known-Interference Cancellation," *Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 802 – 811, June 2003.

[36] R.A. Berry and R.G. Gallager, "Communication over Fading Channels with Delay Constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[37] E. Yeh, "Delay-optimal Rate Allocation in Multiaccess Communications: a Cross-layer View," *in Proc. of the IEEE Workshop on Multimedia Signal Processing*, December 2002.

[38] E. Yeh and A.S. Cohen, "Information Theory, Queueing, and Resource Allocation in Multi-user Fading Communications," *in Proc. of the Conference on Information Sciences and Systems*, March 2004.

[39] G. Montalbano and D.T.M. Slock, "Spatio-Temporal Array Processing for Matched Filter Bound Optimization in SDMA Downlink Transmission," *in Proc. of URSI International Symposium on Signals, Systems, and Electronics)*, September-October 1998.

[40] D.J. Mazzarese and W.A. Krzymien, "High throughput Downlink Cellular Packet Data Access with Multiple Antennas and Multiuser Diversity," *in Proc. of the 57th IEEE Semiannual Vehicular Technology Conference (VTC)*, April 2003.

[41] R.F.H. Fischer, J.B. Huber, C.A. Windpassinger, "Precoding for Point-to-Multipoint Transmission," *in Proc. of the Eighth International Workshop on Signal Processing for Space Communications*, September 2003.

[42] M. Sharif and B. Hassibi, "On the Capacity of MIMO Broadcast Channel with Partial Side Information," *submitted to IEEE Transactions on Information Theory*, 2004, available at http://www.its.caltech.edu/∼masoud/.

[43] W. Rhee and J.M. Cioffi, "On the Capacity of Multiuser Wireless Channels with Multiple Antennas," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2580–2595, October 2003.

[44] J. Bertrand and P. Forster, "Optimal Weights Computation of an Emiting Antenna Array - The Obèle Algorithm," *IEEE Transactions on Signal Processing*, vol. 51, no. 7, pp. 1716–1721, July 2003.

[45] F. Rashid-Farrokhi, L. Tassiulas, K.J. Ray Liu, "Joint Optimal Power Control and Beamforming in Wireless Networks Using Antenna Arrays," *IEEE Journal on Selected Areas in Communications*, vol. 46, no. 10, pp. 1313–1324, October 1998.

[46] M. Bengtsson and B. Ottersten, "Optimal and Suboptimal Transmit Beamforming," in *Handbook of Antennas in Wireless Communications*, L. Godara, Ed. CRC Press, 2001.

[47] G. Montalbano and D.T.M. Slock, "Matched Filter Bound Optimization for Multiuser Downlink Transmit Beamforming," *in Proc. of IEEE International Conference on Universal Personal Communications)*, October 1998.

[48] J.K. Cavers, "Multiuser Transmitter Diversity through Adaptive Downlink Beamforming," *in Proc. of the IEEE Wireless Communications and Networking Conference (WCNC)*, September 1999.

[49] M. Schubert and H. Boche, "Solvability of Coupled Beamforming Problems," *in Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, November 2001.

[50] ——, "Solution of the Multiuser Downlink Beamforming Problem with Individual SINR Constraints," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 1, pp. 18–28, January 2004.

[51] ——, "Joint Dirty Paper Pre-Coding and Downlink Beamforming," *in Proc. of the IEEE 7th International Symposium on Spread Spectrum Techniques and Applications (ISSTA)*, September 2002.

[52] E. Jorswieck H. Boche, "Rate Balancing for the Multi-Antenna Gaussian Broadcast Channel," *in Proc. of the IEEE 7th International Symposium on Spread Spectrum Techniques and Applications (ISSTA)*, September 2002.

[53] H.Boche and M. Schubert, "Comparison of l_∞ and l_1 Optimization for Multi-antenna Downlink Transmission," *in Proc. of the International Symposium on Information Theory (ISIT))*, June-July 2002.

[54] G. Dimic and N.D. Sidiropoulos, "Low-complexity Downlink Beamforming for Maximum Sum Capacity," *in Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004.

[55] B. Hochwald and S.Wishwanath, "Space-Time Multiple Access: Linear Growth in the Sum Rate," *in Proc. of the 40th Allerton Conference on Communications, Control and Computing*, October 2002.

[56] D.A. Gore, R.W. Heath Jr., and A.J. Paulraj, "Transmit Selection in Spatial Multiplexing Systems," *IEEE Communications Letters*, vol. 6, no. 11, pp. 491–493, November 2002.

[57] S. Thoen, L. Van der Perre, M. Engels, and H. De Man, "Adaptive Loading for OFDM/SDMA-Based Wireless Networks," *IEEE Transactions on Communications*, vol. 50, no. 11, pp. 1798 – 1810, November 2002.

[58] C. Peel, B. Hochwald, and L. Swindlehurst, "A Vector-Perturbation Technique for Near-Capacity Multi-Antenna Multi-User Communication: Part I and II," *submitted to IEEE Transactions Wireless Communications*, June 2003, available at http://mars.bell-labs.com/cm/ms/what/mars/papers/mod_precoding/.

[59] W. Rhee, W. Yu, and J.M. Cioffi, "The Optimality of Beamforming in Uplink Multiuser Wireless Systems," *IEEE Transactions on Wireless Communications*, vol. 3, no. 1, pp. 86–96, January 2004.

[60] M. Sharif and B. Hassibi, "A Comparison of Time-sharing, DPC, and Beamforming for MIMO Broadcast Channels with many Users," *submitted to IEEE Transactions on Communications*, 2004, available at http://www.its.caltech.edu/∼masoud/.

[61] D. Samardzija and N. Mandayam, "Multiple Antenna Transmitter Optimization Schemes for Multiuser Systems," *in Proc. of the IEEE 58th Vehicular Technology Conference (VTC) Fall*, October 2003.

[62] C. Windpassinger, R.F.H. Fischer, T. Vencel, and J.B. Huber, "Precoding in Multi-Antenna and Multi-User Communications," *to appear in IEEE Transactions on Wireless Communications*, July 2004, available at http://www.lnt.de/WORLD/LNT/publ/nt_publ.html.

[63] J. A. Nossek, M. Joham, and W. Utschick, "Transmit Processing in MIMO Wireless Systems," *in Proc. of the 6th CAS Symposium on Emerging Technologies*, May 2004.

[64] D.J. Mazzarese and W.A. Krzymien, "Throughput Maximization and Optimal Number of Active Users on the Two Transmit Antenna Downlink of a Cellular System," *in Proc. of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM))*, August 2003.

[65] W. Yu and W. Rhee, "Degrees of Freedom in Multi-user Spatial Multiplex Systems with Multiple Antennas," *submitted to IEEE Transactions on Communications*, March 2004, available at http://www.comm.toronto.edu/∼weiyu.

[66] L. Zheng and D.N.C. Tse, "Diversity and Multiplexing: A Fundamental Tradeoff in Multiple Antenna Channels," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[67] D.N.C. Tse, P. Viswanath, and L. Zheng, "Diversity-Multiplexing Tradeoff in Multiple Access Channels," *to appear in IEEE Transactions on Information Theory.*

[68] L. Weng, A. Anastasopoulos, and S.S. Pradhan, "Diversity Gain Region for MIMO Fading Broadcast Channels," *in Proc. of the Information Theory Workshop (ITW)*, October 2004.

[69] D. Bartolome and A.I. Perez-Neira, "Performance Analysis of Scheduling and Admission control for Multiuser Downlink SDMA," *in Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004.

[70] S. Serbetli and A. Yener, "Transceiver Optimization for Multiuser MIMO Systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 1, pp. 214 – 226, January 2004.

[71] K.K. Wong, R.D. Murch, and K.B. Letaief, "Performance Enhancement of Multisuer MIMO Wireless Communications Systems," *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 1960–1969, December 2002.

[72] A. Pascual-Iserte, "Cooperative Transmitter-Receiver Design: Towards the Implementation of Multi-Antenna Systems," *Ph.D. Dissertation*, in preparation, 2004.

[73] Q.H. Spencer and M. Haardt, "Capacity and Downlink Transmission Algorithms for a Multi-user MIMO Channel," *in Proc. of the 36th Asilomar Conference on Signals, Systems and Computers*, November 2002.

[74] Q.H. Spencer, A.L. Swindlehurst, and M. Haardt, "Zero Forcing Methods for Downlink Spatial Multiplexing in Multiuser MIMO Channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, February 2004.

[75] R. Knopp and G. Caire, "Power Control and Beamforming for Systems With Multiple Transmit and Receive Antennas," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 638–648, October 2002.

[76] M. Bengtsson, "Pragmatic Multi-user Spatial Multiplexing with Robustness to Channel Estimation Errors," *in Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003.

[77] A. Wiesel, Y.C. Eldar, and S. Shamai, "Multiuser Precoders for Fixed Receivers," *in Proc. of the Inernational Zurich Seminar on Communications (IZS)*, February 2004.

[78] H. Boche, M. Schubert, and E.A. Jorswieck, "Throughput Maximization for the Multiuser MIMO Broadcast Channel," *in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (CASSP))*, April 2003.

[79] P. Viswanath, D.N.C. Tse, and R. Laroia, "Opportunistic Beamforming using Dumb Antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[80] L. Dong, T. Li, and Y.F. Huang, "Opportunistic Transmission Scheduling for Multiuser MIMO Systems," *in Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP*, April 2003.

[81] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR: a high Efficiency-high Data Rate Personal Communication Wireless System," *in Proc. of IEEE Vehicular Technology Conference (VTC)*, May 2000.

[82] G. Caire, D. Tuninetti, and S. Verdú, "Suboptimality of TDMA in the Low-Power Regime," *IEEE Transactions on Information Theory*, vol. 50, no. 4, pp. 608–620, April 2004.

[83] M. Airy, S. Shakkottai, and R.W. Heath Jr., "Downlink Multi-User MIMO Scheduling," *submitted*, 2003.

[84] R.W. Heath, M. Airy, and A.J. Paulraj, "Multiuser Diversity for MIMO Wireless Systems with Linear Receivers," in *in Proceedings of the 35th Asilomar Conference on Signals, Systems and Computers*, November 2001.

[85] S. Serbetli and A. Yener, "Time Slotted Multiuser MIMO Systems: Beamforming and Scheduling Strategies," *accepted in EURASIP Journal on Wireless Communications and Newtworking (JWCN)*, June 2004.

[86] Z. Tu and R.S. Blum, "Multiuser Diversity for a Dirty Paper Approach," *IEEE Communications Letters*, vol. 7, no. 8, pp. 370–372, August 2003.

[87] H. Viswanathan and S. Venkatesan, "The Impact of Antenna Diversity in Packet Data Systems with Scheduling," *IEEE Transactions on Communications*, vol. 52, no. 4, pp. 546–549, April 2004.

[88] F. Shad, T.D. Todd, V. Kezys, and J. Litva, "Dynamic Slot Allocation (DSA) in Indoor SDMA/TDMA Using a Smart Antenna Basestation," *IEEE/ACM Transactions on Networking*, vol. 9, no. 1, pp. 69 – 81, February 2001.

[89] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness.* Freeman, 1979.

[90] R. Kuehner, T.D. Todd, F. Shad, and V. Kezys, "Forward-Link Capacity in Smart Antenna Base Stations with Dynamic Slot Allocation," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 4, pp. 1024 – 1038, July 2001.

[91] H. Yin and H. Liu, "Performance of Space-Division Multiple-Access (SDMA) with Scheduling," *IEEE Transaction on Wireless Communications*, vol. 1, no. 4, pp. 611 – 618, October 2002.

[92] I. Koutsopoulos and L. Tassiulas, "Adaptive Channel Allocation for OFDM-based Smart Antenna Systems with Limited Transceiver Resources," *Technical Research Report*, 2002, available at http://techreports.isr.umd.edu/reports/2002/TR_2002-64.pdf.

[93] ——, "Adaptive Resource Allocation in SDMA-based Wireless Broadband Networks with OFDM Signaling," *in Proc. of the 24th Annual Joint Conference of the IEEE Communication Society (INFOCOM)*, June 2002.

[94] P. Vandenameele, L. van Der Perre, M.G.E. Engels, B. Gyselinckx, and H.J. De Man, "A Combined OFDM/SDMA Approach," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 11, pp. 2312 – 2321, November 2000.

[95] W.J. Huang and J.F. Doherty, "A Spatial Clustering Scheme for Downlink Beamforming in SDMA Mobile Radio," *in Proc. of the Tenth IEEE Workshop on Statistical Signal and Array Processing*, August 2000.

[96] V.K.N. Lau, Y.K. Kwok, "Performance Analysis of SIMO Space-Time Scheduling with Convex Utility Function: Zero-Forcing Linear Processing," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 2, pp. 339–350, March 2004.

[97] Y. Cao and V.O.K. Li, "Scheduling Algorithms in Broad-Band Wireless Networks," *Proceedings of the IEEE*, vol. 89, no. 1, pp. 76 – 87, January 2001.

[98] W. Li, K.K. Yu, L.C. Wing, and V.K.N. Lau, "Channel Capacity Fair Queueing in Wireless Networks: Issues and a New Algorithm," *in Proc. of the IEEE International Conference on Communications*, April-May 2002.

[99] H. Fattah and C. Leung, "An Overview of Scheduling Algorithms in Wireless Multimedia Networks," *IEEE Wireless Communications*, vol. 9, no. 5, pp. 76–83, October 2002.

[100] D.P. Palomar, "A Unified Framework for Communications through MIMO Channels," *Ph.D. Thesis*, May 2003.

[101] A. Pascual-Iserte, A. I. Perez-Neira, M.A. Lagunas-Hernandez, "On Power Allocation Strategies for Maximum Signal to Noise and Interference Ratio in an OFDM-MIMO System," *IEEE Transactions on Wireless Commununications*, vol. 3, no. 3, pp. 808–820, May 2004.

[102] R. Gibbons, *A Primer in Game Theory (Spanish Edition)*, Antoni Bosch, Ed., 1997.

[103] C.U. Saraydar, N.B. Mandayam, and D.J. Goodman, "Efficient Power Control via Pricing in Wireless Data Networks," *IEEE Transactions on Communications*, vol. 50, no. 2, pp. 291–303, Feb. 2002.

[104] D. Goodman and N. Mandayam, "Power Control for Wireless Data," *IEEE Personal Communications*, vol. 7, no. 2, pp. 48–54, April 2000.

[105] A.B. MacKenzie and S.B. Wicker, "Game Theory and the Design of Self-configuring, Adaptive Wireless Networks," *IEEE Communications Magazine*, vol. 39, no. 11, pp. 126–131, November 2001.

[106] T. Polo, "Power Control based on Game Theory. Application to the Uplink in CDMA systems. (original title in Spanish)," *M.Sc. Thesis from the Technical University of Catalonia (UPC), Spain*, January 2004.

[107] E. Altman and Z. Altman, "S-modular Games and Power Control in Wireless Networks," *IEEE Transactions on Automatic Control*, vol. 48, no. 5, pp. 839 – 842, May 2003.

[108] W. Yu, G. Ginis, and J.M. Cioffi, "Distributed Multiuser Power Control for Digital Subscriber Lines," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 5, pp. 1105 – 1115, June 2002.

[109] D. Hughes-Hartogs, "Ensemble Modem Structure for Imperfect Transmission Media," *U.S. Patents 4679227 (July 1987) and 4731816 (March 1988).*

[110] R. V. Sonalkar and R.R. Shively, "An Efficient Bit-loading Algorithm for DMT Applications," *IEEE Communications Letters*, vol. 4, no. 3, pp. 80 – 82, March 2003.

[111] J. Campello, "Practical Bit Loading for DMT," *in Proc. of the IEEE International Conference on Communications (ICC)*, June 1999.

[112] L.M.C. Hoo, J. Tellado, and J.M. Cioffi, "Discrete Dual QoS Loading Algorithms For Multicarrier Systems," *in Proc. of International Conference Communications (ICC)*, 1999.

[113] B.S. Krongold, K. Ramchandran, and D.L. Jones, "Computationally Efficient Optimal Power Allocation Algorithms for Multicarrier Communication Systems," *IEEE Transactions on Communications*, vol. 48, no. 1, pp. 23 – 27, January 2000.

[114] D.W. Lin, "An Optimal Bit Loading For Multitone ADSL," *in Proc. of IEEE International Symposium on Circuits and Systems*, May 2000.

[115] M. Zwingelstein-Colin, M. Gazalet, and M. Gharbi, "Non-iterative Bit-loading Algorithm for ADSL-type DMT Applications," *IEE Proceedings Communications*, vol. 150, no. 6, pp. 414 – 418, December 2003.

[116] C.Y. Wong, R.S. Cheng, K.B. Letaief, and R.D. Murch, "Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747 – 1758, October 1999.

[117] D. Kivanc, G. Li, and H. Liu, "Computationally Efficient Bandwidth Allocation and Power Control for OFDMA," *IEEE Transactions on Wireless Communications*, vol. 2, no. 6, pp. 1150 – 1158, November 2003.

[118] S. Pfletschinger, *Multicarrier Modulation for Broadband Return Channels in Cable TV Networks.* Ph.D. Thesis, 2003.

[119] M. Ergen, S. Coleri, and P. Varaiya, "QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems," *IEEE Transactions on Broadcasting*, vol. 49, no. 4, pp. 362 – 370, December 2003.

[120] K. Inhyong, L.L. Hae, K. Beomsup, and Y.H. Lee, "On the Use of Linear Programming for Dynamic Subchannel and Bit Allocation in Multiuser OFDM," *in Proc. of the IEEE Global Telecommunications Conference (GLOBECOM)*, November 2001.

[121] J. Lee, R. V. Sonalkar, and J. M. Cioffi, "Multi-user Bit-loading for Multi-carrier Systems," *submitted to IEEE Transactions on Communications*, 2003, available at http://www.stanford.edu/∼jungwon/.

[122] A.E. Mohr, "Bit Allocation in Sub-linear Time and the Multiple-choice Knapsack Problem," *in Proc. of the Data Compression Conference (DCC)*, April 2003.

[123] D. Meiri and I. Kalet, "Wide-Band Multitone Water-Pouring Version of V-BLAST," *in Proc. of the 7th International OFDM-Workshop*, September 2002.

[124] X. Zhang and B. Ottersten, "Power Allocation and Bit Loading for Spatial Multiplexing in MIMO Systems," *in Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003.

[125] K.K. Wong, "Adaptive Space-Division-Multiplexing and Bit-and-Power Allocation in Multiuser MIMO Flat Fading Broadcast Channels," *in Proc. of the 58th Vehicular Technology Conference (VTC)*, October 2003.

[126] A.Banchs, "User Fair Queuing: Fair Allocation of Bandwidth for Users," *in Proc. of the 21st Annual Joint Conference of IEEE Computer and Communications Societies (INFOCOM)*, June 2002.

[127] J. Y. Le Boudec, B. Radunovic, "A Unified Framework for Max-Min and Min-Max Fairness with Applications," *in Proc. of the 40th Annual Allerton Conference on Communication, Control, and Computing*, October 2002.

[128] A. Kunz and C. Stepping, "Overview and Implementation of Scheduling Algorithms for Wireless Environments," *in Proc. of the 5th European Personal Mobile Communications Conference*, 2003.

[129] H. Zhu and K.J.R. Liu, "The Use of Diversity Antennas in High-Speed Wireless Systems: Capacity Gains, Fairness Issues, Multi-User Scheduling," *Bell Labs Technical Memorandum*, 2001, available at http://mars.bell-labs.com/cm/ms/what/mars/papers/borst_whiting.

[130] T. Sartenaer and L. Vandendorpe, "Balanced Capacity of Wireline Multiaccess Channels," *in Proc. of the European Signal Processing Conference (EUSIPCO)*, September 2004.

[131] D. Bartolome, D.P. Palomar, A.I. Perez-Neira, "Real-Time Scheduling for Wireless Multiuser MISO Systems under Different Fairness Criteria," *in Proc. of the International Symposium Signal Processing and its Applications (ISSPA)*, July 2003.

[132] H. Zhu and K.J.R. Liu, "Joint Adaptive Power and Modulation Management in Wireless Networks with Antenna Diversity," *in Proc. of the Sensor Array and Multichannel Signal Processing Workshop*, August 2002.

[133] R. Knopp, "Achieving Multiuser Diversity under Hard Fairness Constraints," *in Proc. of the International Symposium on Information Theory (ISIT)*, June-July 2002.

[134] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems," *DEC Research Rep. TR-301*, vol. 2, no. 5, pp. 1017 – 1028, September 2003, available at http://www.cse.ohio-state.edu/~jain/papers/fairness.htm.

[135] H. Markowitz, "Foundations of Portfolio Theory," *The Journal of Finance*, vol. 46, no. 2, pp. 469–477, June 1991.

[136] ——, "Portfolio Selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, March 1952.

[137] R.A. Brealey and S.C. Myers, *Principles of Corporate Finance, 7th Edition.* New York: McGraw-Hill, 2002.

[138] H. Fu, R. Rodman, D. McAllister, D. Bitzer, and B. Xu, "Classification of Voiceless Fricatives through Spectral Moments," *in Proc. of the 5th International Conference on Information Systems Analysis and Synthesis (ISAS)*, 1999.

[139] H. Shi and H. Sethu, "An Evaluation of Timestamp-Based Packet Schedulers Using a Novel Measure of Instantaneous Fairness," *in Proc. of the Performance, Computing and Communications Conference*, April 2003.

[140] K. Xu, "How has the Literature on Gini's Index Evolved in the Past 80 Years?" *Dalhousie University, Economics Working Paper.*, April 2003, available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=423200.

[141] G. Yin and X.Y. Zhou, "Markowitz's Mean-Variance Portfolio Selection With Regime Switching: From Discrete-Time Models to Their Continuous-Time Limits," *IEEE Transactions on Automatics and Control*, vol. 49, no. 3, pp. 349 – 360, March 2004.

[142] L. Massoulie and J. Roberts, "Bandwidth Sharing: Objectives and Algorithms," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 320–328, June 2002.

[143] J. Choi and S. Perreau, "MMSE Multiuser Downlink Multiple Antenna Transmission for CDMA Systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 6, pp. 1564–1573, June 2003.

[144] A. Pascual-Iserte, A. I. Perez-Neira, M.A. Lagunas-Hernandez, "An Approach to Optimum Joint Beamforming Design in a MIMO-OFDM Multiuser System," *to be published in the EURASIP Journal on Wireless Commununications.*

[145] J. Jiang, M. Buehrer, and W.H. Tranter, "Spatial T-H Precoding for Packet Data Systems with Scheduling," *in Proc. of the 58th IEEE Vehicular Technology Conference (VTC)*, October 2003.

[146] B.M. Hochwald, T.L. Marzetta, and V. Tarokh, "Multi-Antenna Channel Hardening and its Implications for Rate Feedback and Scheduling," *submitted to IEEE Trans. on Inform. Theory*, May 2002.

[147] Z. Dongmei, X. Shen, and J.W. Mark, "Efficient Call Admission Control for Heterogeneous Services in Wireless Mobile ATM Networks," *IEEE Communications Magazine*, vol. 38, no. 10, pp. 72–78, October 2000.

[148] D. Bartolome, A. Pascual Iserte, A.I. Perez-Neira, "Spatial Scheduling Algorithms for Wireless Systems," *in Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2003.

[149] S.T. Chung and A.J. Goldsmith, "Degrees of Freedom in Adaptive Modulation: A Unified View," *IEEE Transactions on Communications*, vol. 49, no. 9, pp. 1561–1571, September 2001.

[150] J.M. Cioffi, "A multicarrier primer," *ANSI Doc., T1E1.4 Tech. Subcommittee*, no. 91-157, 1991.

[151] D. Piazza and L.B. Milstein, "Multiuser Diversity-Mobility Tradeoff: Modeling and Performance Analysis of a Proportional Fair Scheduling," *in Proc. of the Global Telecommunications Conference (GLOBECOM)*, November 2002.

[152] C. Rao, *Linear Statistical Inference and its Application.* John Wiley and Sons, 1973.

[153] A.B. MacKenzie and S.B. Wicker, "A Repeated Game Approach to Distributed Power Control in CDMA Wireless Data Networks," *in revision for IEEE Trans. on Wireless Commun.*

[154] C. Comaniciu and H.V. Poor, "Jointly Optimal Power and Admission Control for Delay Sensitive Traffic in CDMA Networks with LMMSE Receivers," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2031 – 2042, August 2003.

[155] A.J. Goldsmith and C. Soon-Ghee, "Variable-rate Variable-power MQAM for Fading Channels," *IEEE Transactions on Communications*, vol. 45, no. 10, pp. 1218 –1230, Oct. 1997.

[156] R. Knopp and P.A. Humblet, "Information Capacity and Power Control in Single-cell Multiuser Communications," *in Proc. of the 1995 IEEE International Conference on Communications (ICC)*, June 1995.

[157] IEEE Std 802.11a-1999, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-speed Physical Layer in the 5 GHz Band, 1999.

[158] I. Koutsopoulos, *Resource allocation issues in broadband wireless networks with OFDM signaling.* Ph.D. Dissertation, 2002.

[159] A.S. Macedo and E.S. Sousa, "Antenna-Sector Time-Division Multiple Access for Broadband Indoor Wireless Systems," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 6, pp. 937 – 952, Aug. 1998.

[160] J. Medbo and P. Schramm, "Channel Models for Hiperlan/2 in Different Indoor Scenarios," *BRAN 3ERI085B*, March 1998.

[161] D. Bartolome and A.I. Perez-Neira, "Reconfigurable Antenna Array Architecture for OFDM Receivers," *in Proc. of the 2nd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, December 2002.