UNIVERSITÉ DE MONTRÉAL

EFFICIENT METHODS FOR RESOURCE ALLOCATION IN MULTI-ANTENNA ORTHOGONAL FREQUENCY-DIVISION MULTIPLE ACCESS (OFDMA) SYSTEMS

DIEGO ENRIQUE PEREA
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE ÉLECTRIQUE)
FÉVRIER 2013

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

EFFICIENT METHODS FOR RESOURCE ALLOCATION IN MULTI-ANTENNA ORTHOGONAL FREQUENCY-DIVISION MULTIPLE ACCESS (OFDMA) SYSTEMS

présentée par : PEREA Diego Enrique
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

Mme SANSÒ Brunilde, Ph.D., présidente
M. FRIGON Jean-François, Ph.D., membre et directeur de recherche
M. GIRARD André, Ph.D., membre et codirecteur de recherche
M. NERGUIZIAN Chahé, Ph.D., membre
M. LE Long, Ph.D., membre

La pluma es la lengua del alma ;
cuales fueren los conceptos que en ella se engendraren,
tales serán sus escritos.

The pen is the tongue of the soul ;
as are the thoughts engendered there,
so will be the things written.

Miguel de Cervantes Saavedra (1547-1616).

A mis padres : Cecilia y Raúl.

ACKNOWLEDGEMENTS

As I come to the end of my Ph.D. program, I would like to extend my gratitude to those people whose positive influence, both personal and academic, have helped me during these last five years in Montréal. I have had the good fortune of having two great research supervisors. Dr. Jean-François Frigon supported me during all the stages of my research, both with his knowledge of telecommunications and through his natural disposition to help. Dr. André Girard helped me start navigating the field of optimization in telecommunication networks, and to correctly formulate problems and find solutions. By sharing their knowledge and enthusiasm for research, they gave me the motivation and persistence needed to finish my studies ; their support and influence are reflected in this thesis. I also benefited from my association with some very talented graduate students and friends, Vida Vakilian, Ali Torabi and Xavier Kuhn. I would also like to thank the people who read and recommended text modifications to this thesis, my supervisors, fellow graduate students, and my very good friends, Carlos Cadena and Kimberly Nash.

But my educational journey and passion for research and school would not have started without the support of my parents, to whom this thesis is dedicated. They always encouraged me to play, create and study, and despite adverse circumstances shared my dreams. I also warmly remember several teachers from my elementary and secondary schools in Colombia, who inspired my love of learning, which I, in turn, now hope to pass on to my son, Diego Eduardo. Finally, from my engineering school years, my friends, Ivan Ladino and Rafael Adames, together with professors, Dr. Pedro Vizcaya and Dr. Edgar Ramos, likewise played an important role in my nascent career.

RÉSUMÉ

Dans cette thèse, nous proposons une solution au problème d'allocation de ressources d'un système MISO (Multiple Input Multiple Output)-OFDMA (Orthogonal Frequency Division Multiplexing Access) supportant des usagers requérant un débit de transmission minimal. Ce problème peut être modélisé comme une optimisation non-linéaire mixte avec variables entières. Nous nous sommes intéressés dans cette thèse à diverses méthodes permettant la résolution d'un tel problème.

La première approche étudiée utilise une méthode hors-ligne et permet d'obtenir une solution quasi-optimale qui peut être utilisée comme références pour évaluer la performance d'algorithmes heuristiques pouvant être réalisés en temps réel. Pour ce faire, nous cherchons une allocation réalisable en se basant sur la solution optimale du problème dual. Nous obtenons la fonction duale et trouvons la solution à l'aide d'un algorithme itératif à sous-gradient. Cette solution permet d'otenir une borne supérieure à la solution optimale. D'autre part, nous développons une heuristique basée sur la solution du problème dual pour obtenir une solution réalisable du problème primaire qui constitue une borne inférieure à la solution optimale. Ces bornes nous permettent d'établir que l'écart de dualité est petit pour les configurations étudiées et elles peuvent servir de référence pour l'évaluation de performances des algorithmes heuristiques. La formulation duale nous fournit aussi une meilleure compréhension du sujet en établissant un lien entre la réalisabilité de l'allocation de ressources et les débits minimaux requis par les usagers.

Afin d'obtenir des méthodes de résolution plus pratiques pouvant être réalisées en temps réel, nous proposons deux heuristiques ayant une faible complexité et permettant d'atteindre des performances assez près des performances optimales. Les performances ainsi obtenues sont légèrement moins bonnes que celles d'autres algorithmes qu'on retrouve dans la littérature, mais supportent une plus grande plage de valeurs de débit minimal tout en réduisant la complexité de l'algorithme d'allocation de ressources de plusieurs ordres de grandeur. L'écart entre la solution trouvée par ces algorithmes heuristiques et la borne supérieure duale est relativement faible. Par exemple, l'écart est de 10.7% en moyenne pour toutes les configurations étudiées. L'augmentation dans la plage de débits minimaux supportés comparés avec les méthodes disponibles dans la littérature est de 14.6% en moyenne. Cette amélioration est obtenue en considérant les variables duals de contrainte de débits minimaux dans l'allocation de puissance aux usagers. L'algorithme heuristique proposé sélectionne un ensemble d'usagers pour chaque sous-porteuse, mais contrairement aux autres méthodes proposées précédemment, l'algorithme considère l'ensemble des usagers avec des contraintes de débits

minimaux dans la réassignation des sous-porteuses pour s'assurer que le niveau de service requis est rencontré. Suite à la sélection des ensembles d'usagers, un problème d'allocation de puissance convexe est résolu. Des algorithmes permettant de résoudre efficacement et en un temps moindre les problèmes d'assignation des sous-porteuses aux usagers et d'allocation de puissance sont proposées dans cette thèse.

Finalement, nous étudions aussi de quelle façon ces algorithmes peuvent être utilisés pour résoudre le problème d'allocation de ressources dans une cellule utilisant la téchnologie LTE (Long Term Evolution)-Advanced. Les méthodes étudiées dans cette thèse font partie d'un nouvel ensemble d'algorithmes nécessaires pour supporter des applications temps réel à haut débit et à l'efficacité spectrale requise dans les prochains réseaux d'accès sans-fil de quatrième génération.

ABSTRACT

In this dissertation, we solve the Resource Allocation (RA) problem of a Multiple Input Single Output (MISO)– Orthogonal Frequency Division Multiplexing Access (OFDMA) system supporting minimum rates. This problem can be modelled as a non-linear Mixed Integer Program (NLMIP). We are interested in various kinds of methods to solve this problem.

First, our focus is on an off-line method that gives near-optimal solutions that serve as benchmark for more practical methods. For this purpose, we propose a method based on the optimal solution of the dual problem. We obtain a dual function and solve the dual problem through subgradient iterations. Then, we find upper and lower bounds for the optimal solution and verify that the duality gap is small for the system configurations studied. Therefore, the dual optimal serves as a reference for any feasible solution produced by the heuristic methods. The dual formulation also gives a better insight into the problem, as it shows us the relation between the problem's feasibility and the minimum rate requirements.

To obtain more practical methods, we propose two heuristics that have very low computational complexity and give performances not far from the optimal. We compare their performance against other methods proposed in the literature and find that they give a somewhat lower performance, but support a wider range of minimum rates while reducing the computational complexity of the algorithm by several orders of magnitude. The gap between the objective achieved by the heuristics and the upper bound given by the dual optimal is not large. For example, in our experiments this gap is 10.7% averaging over all performed numerical evaluations for all system configurations. The increase in the range of the supported minimum rates when compared with the method reported in the literature is 14.6% on average. This increase is achieved by considering the rate constraint dual variables in the user power allocation stage. The proposed heuristics select a set of users for each subcarrier, but contrary to other reported methods used to solve the throughput maximization problem, they consider the set of real-time (RT) users to ensure that their minimum rate requirements are met. Then, they solve a power allocation problem for fix subcarrier assignment, which is a convex problem that is simpler to solve. We use efficient algorithms for the subcarrier assignment and power allocation stages to solve the problem much quicker.

Finally, we adapt the algorithms to solve the RA problem in a single cell using LTE (Long Term Evolution)–Advanced technology. The methods examined in this dissertation are part of the new set of algorithms needed to support the high rate applications and spectral efficiency required in the wireless access of upcoming 4G networks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AMPL	A Modelling Language for Mathematical Programming
BB	Branch and Bound
BER	Bit Error Rate
BS	Base Station
CBP	Codebook Based Precoding
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CSI	Channel State Information
DMRS	Demodulation Reference Signal
IEEE	Institute of Electrical and Electronic Engineers
IP	Internet Protocol
LTE	Long Term Evolution
MAC	Media Access Control
MIP	Mixed Integer Program
MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
MMSE	Minimum Mean Squared Error
MU	Multi-User
NLP	Non Linear Program
NLMIP	Non Linear Mixed Integer Program
NP	Non-deterministic Polynomial time
nRT	non Real Time
OFDMA	Orthogonal Frequency Division Multiplexing Access
PA	Power Allocation
PDCP	Packet Data Converge Protocol
PHY	Physical layer
PMI	Precoding Matrix Indicator
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
4G	Fourth Generation
RA	Resource Allocation
RB	Resource Block

RE	Resource Element
RLC	Radio Link Control
RRC	Radio Resource Control
RT	Real Time
SDMA	Spatial Division Multiple Access
SISO	Single Input Single Output
SNR	Signal to Noise Ratio
SU	Single-User
UMTS	Universal Mobile Telecommunications System
ZF	Zero Forcing

MATHEMATICAL CONVENTION

$\mathbf{A}$	Matrix
$\mathbf{A}^T$	Matrix transpose
$\mathbf{A}^H$	Matrix Hermitian
$\mathbf{A}^\dagger$	Matrix pseudoinverse
$\mathbf{I}$	Identity matrix of compatible dimensions
$p_{i,j}$	Element i, j of matrix $\mathbf{P}$
q_k	Element k of vector $\mathbf{q}$
$[\mathbf{A}]_{j,j}$	Matrix diagonal element
$\mathbf{b}, \boldsymbol{\mu}$	Vectors
$\|\mathbf{b}\|$	Vector norm
x, U, θ, Θ	Scalars
$\|x\|$	Absolute value
$[x]^+$	$= \max\{0, x\}$
$\mathcal{S}$	Sets
$\mathcal{S}(x)$	Elements of set $\mathcal{S}$ are a function of x
$\|\mathcal{S}\|$	Cardinality of set $\mathcal{S}$
$\mathbb{R}^K$	Real numbers set of K dimension
$\mathbb{R}_+$	Set of non-negative real numbers
$\mathbb{C}$	Complex numbers set
$\mathcal{CN}(a, b)$	Complex Gaussian distribution with mean a and variance b

CHAPTER 1

INTRODUCTION

In today's world we use telecommunication networks in most of our daily activities. For example, we receive or make phone calls, and send or receive instant messages or e-mails for our work, school and social affairs. The vision for fourth generation (4G) networks is that people and applications will communicate with each other through appropriate networks, fixed or mobile, with greater ease. However, with the ubiquitous use of smart phones, tablets, laptops and real-time (RT) applications, traffic demand on the wireless access network is increasing exponentially [1]. Therefore, there is a real need to create the infrastructure that supports traffic with higher rates and strict time deadlines.

Due to bandwidth scarcity, as the number of users per cell increases and user applications require higher rates, it becomes critical to augment the system spectral efficiency, i.e., the achieved bit rate per unit of bandwidth. Resource allocation (RA) mechanisms that handle different types of traffic and adapts to the channel conditions would best use system resources. Both an increase in spectral efficiency and intelligent radio resource allocation are key enabling factors of the implementation of 4G wireless networks.

To increase spectral efficiency, we can use systems with multiple antennas. For example, consider a single user multiple input multiple output (MIMO) system with M antennas at the transmitter and M antennas at the receiver. In this configuration, the channel capacity increases approximately linearly when Rayleigh fading is considered [2]. Multi-user diversity, on the other hand, can take advantage of the channel fluctuations by selecting users with good channel conditions at every time interval. In this case the capacity growth as a function of the number of users is logarithmic. Frequency diversity can be attained by using orthogonal frequency division multiplexing access (OFDMA), where coding is performed over independently faded subcarriers. Combining these techniques in a multi-user MIMO-OFDMA system would provide us with multi-user, spatial and frequency diversity, which give us a high spectral efficiency. These systems are proposed in current 4G standards, Long Term Evolution (LTE) and IEEE 802.16 [3, 4].

There is, however, a price to pay. Multi-antenna systems require more hardware and software resources to process the multiple spatial layers. In addition, by increasing the degrees of freedom in the system for transmission, the RA process becomes more complex because we have more possibilities from which to choose. As an example of the increase in computational complexity in the RA process, consider the method we use in chapter 2 to obtain a dual

optimal solution to the resource allocation problem for a system with M transmit antennas, N subchannels and K single antenna users. The computational complexity of the RA algorithm is $O(NK^M M^3)$, which grows exponentially with the number of antennas, whereas in the single antenna case there is only one possible signal path from the transmitter to each receiver so the computational complexity is linear with the number of subchannels and users, $O(NK)$ [5]. This shows that increasing the number of spatial channels, users and subcarriers, increases the spectral efficiency, but it also greatly increases the computational complexity of the RA process. The problem we deal with in this dissertation is a common one in practical applications of optimization theory: how to devise RA algorithms that provide us with solutions not too far from the optimal and that are efficient to compute. In addition, how to obtain a near optimal solution to RA problems using off-line methods to benchmark the devised efficient methods. These two general questions, applied to the RA problem in a Multiple Input Single Output (MISO)-OFDMA system supporting minimum rates, are the subject of this dissertation.

The methods discussed in this dissertation try to answer these questions. More specifically, the set of algorithms proposed are part of the RA algorithms needed to support the high rate applications and spectral efficiency required in the wireless access of upcoming 4G networks.

1.1 Communications System to Optimize

We consider in this dissertation a MISO-OFDMA system in a single cell of a cellular network. This system provides frequency diversity by dividing the available spectrum into a number of subchannels and using OFDMA access techniques. Spatial diversity is achieved by placing multiple antennas at the BS spaced far enough so that the multiple paths from the antennas to each receiver are uncorrelated. We do *not* consider a general MIMO system where multiple antennas are placed at transmitter and receivers because multiple receive antennas are very complex to implement in small devices. Thus, we consider a scenario that is likely to be implemented in the short term.

We only take into account transmission in the downlink direction where spatial diversity is achieved by performing SDMA beamforming [28]. We restrict ourselves to linear beamforming where the vector of symbols to transmit is simply multiplied by a beamforming matrix. The main reason to select this beamforming technique is that it is supported by the current wireless access technologies IEEE 802.16 and LTE-Advanced. In the general beamforming problem described in in chapter 2, the user selection and the beamforming vectors are the variables to optimize. On the other hand, the objective to maximize is the weighted sum of the users rate. This is in contrast with more general utility functions found in the literature

[6]; we use this linear function because a closed-form of the dual function can be found in this case, which is the basis of the proposed method. The user weights can be adjusted from frame to frame, thus fairness among users can be implemented in the system. Strict QoS control is implemented by selecting a number of users at each frame for which hard minimum rate constraints are enforced. Therefore, real-time (RT) QoS is supported for the selected RT users.

For this system, we solve the following optimization problem: for a given time slot, find the user selection and beamforming vectors that maximize a linear utility function of the user rates, given a total transmit power constraint and minimum rate constraints for RT users. A total power constraint is chosen instead of more practical per-antenna power constraints to ease the analysis.

1.2 Thesis Objectives

In the research of methods to solve the RA problem for MISO-OFDMA systems supporting minimum rates, we are interested in solutions with several degrees of accuracy and computational complexity. Due to the combinatorial nature of the problem it is almost impossible to obtain optimal solutions for realistic problem sizes. Thus, we are initially interested in *near-optimal* solution methods that can be executed off-line and serve as references for more efficient on-line methods. It is important to evaluate the accuracy of these near-optimal solutions by finding limits on the difference between the optimal and near-optimal points. In addition, in optimization theory the problem formulation influences the complexity of the solution method and the set of techniques one has available. Therefore, the first two objectives of this dissertation are to find an adequate formulation to the problem and to devise and evaluate an off-line near-optimal solution method to solve it.

The subsequent objectives look at the design of RA algorithms that can be implemented in real-time systems. The third objective is the design of heuristic algorithms that are computationally efficient and that produce sub-optimal feasible points. In this case, it is also important to evaluate how far these points are from optimality. The fourth objective is to study the applicability of the proposed algorithms in current 4G wireless access technologies. We choose LTE-Advanced technology [3] to perform this study because of the expected prevalence in the industry in the upcoming years.

In summary, the objectives of this dissertation are to

1. Formulate an optimization problem for the allocation of physical layer resources in a MISO-OFDMA system considering RT and non Real-Time (nRT) traffic, where RT users have minimum rate requirements.

2. Devise an off-line solution method to solve the problem with near optimality and evaluate the accuracy of the solutions.
3. Devise heuristic methods to solve the problem efficiently, and to evaluate their accuracy and computational complexity against the off-line method and other methods proposed in the literature.
4. Study the application of the proposed heuristic methods to the current LTE-Advanced technology used in 4G wireless access networks.

1.3 State of the Art

We first discuss in subsection 1.3.1 the most important papers for SISO-OFDMA systems. These papers contain the optimization principles later used in the more sophisticated MISO-OFDMA systems presented in subsection 1.3.2. In subsection 1.3.3, we review extensions to the MISO-OFDMA RA problem where multiple cells are considered.

1.3.1 Early Research Work on SISO-OFDMA Systems

In one of the first reported papers [7], the total transmitted power is minimized under user rate constraints. The problem is formulated as a Mixed Integer Program (MIP) where the user rates are modelled as a continuously increasing function of the power, and the subcarrier assignment variable is binary. The integer constraint is relaxed and an stationary point of the dual Lagrange function is obtained to perform subcarrier assignment and power allocation. This solution point is not necessarily primal feasible. Therefore, a simple heuristic is used to find a primal feasible point from the dual optimal. Kim et al. [8] formulated the same problem as a MIP, but solve it using a Linear Program (LP) relaxation. The linear program is less computationally expensive that the original but it is still unmanageable. Therefore, the authors devised a suboptimal heuristic where subcarrier allocation and power assignment are performed separately. Their solution is close to the optimal.

In [6], a general utility function of the rates is maximized under a total power constraint. The utility function is assumed continuous and concave. The purpose of this function is to balance the trade-off between spectral efficiency and fairness. First, the transmission power is assumed uniformly distributed over the entire available frequency band. The optimal subcarrier assignment is found for the constant power case. Later, for the fixed subcarrier assignment case the resulting optimal power is found to have the water-filling form. An optimal condition is derived by performing iteratively constant-power subcarrier assignment and water-filling power allocation. In [9], more efficient algorithms are proposed to solve the problem suboptimally.

Other way to model fairness is by enforcing a proportionality among the user rates. In [10], a set of proportional fairness constraints is imposed to assure that each user can achieve a portion of the total sum rate. Since the optimal solution to the constrained fairness problem is extremely computationally complex to obtain, a low-complexity suboptimal algorithm that separates subchannel allocation and power allocation is proposed. In the proposed algorithm, subchannel allocation is first performed by assuming an equal power distribution. An optimal power allocation algorithm then maximizes the sum rate while maintaining proportional fairness. The proposed algorithm reduces the complexity from exponential to linear in the number of subcarriers.

Several of these early papers use the dual Lagrange approach to solve the RA problem for SISO-OFDMA systems, but the optimality of the solutions is not justified mathematically. However, the intuition behind the dual Lagrange approach and the results obtained favor this method. Proof of optimality was more recently provided by later work such as [5] for the ergodic capacity case.

1.3.2 Work on MISO-OFDMA Systems

For the MISO-OFDMA system the RA problem we want to solve is a nonlinear, non-convex integer program, which makes it almost impossible to solve directly for any realistic number of subcarriers, users and antennas. For this reason, most research work focuses on developing heuristic algorithms. In this context, an important question is always how accurate are the results, i.e., how close to the optimum the objectives are. For an RA problem with rate constraints, it turns out that there are very few results of that kind, as we shall see.

Traffic in the system can be divided into two main groups: delay-sensitive RT services and delay-insensitive nRT services. Early work on OFDMA systems focuses on solving the RA problem for nRT services only, where the objective is to maximize the total throughput with only power constraints and possibly minimum Bit Error Rate (BER) constraints. In [11], the complete optimization problem is divided into two sub-problems: selection of users for each carrier and allocation of power to these users, which are both solved by a heuristic. A similar approach using Zero Forcing (ZF) beamforming is reported in [12]. The work of [11, 12] does not solve the complete optimization problem; instead, it separates it into uncoupled subproblems that provide suboptimal solutions.

There is a definite need to benchmark the performance of these heuristic algorithms. For the RA problem with nRT traffic only, several methods have been proposed to compute near-optimal solutions. For example, genetic algorithms are proposed in [13] while [14, 15, 16, 17] provide methods to compute a near-optimal solution based on dual decomposition. In addition to providing a benchmark, near-optimal algorithms can also lead to the design

of efficient RA methods as shown in [17], where heuristic algorithms derived from the dual decomposition method are proposed.

Several heuristic methods have been used to solve the RA problem for OFDMA-SDMA systems with both RT and nRT traffic. In [18], the objective is to maximize the sum of the user rates subject to per-user minimum rate constraints that model the priority assigned to each user at each frame. The optimization problem is solved approximately for each frame by minimizing a cost function representing the increase in power needed when increasing the number of users or the modulation order. The advantages of this approach are that it handles user scheduling and RA together and supports RT and nRT traffic. Its weaknesses are that no comparison is made against a near-optimal solution and the method used to determine user priorities at every frame is very complex.

In [19], both RT and nRT traffic are supported. Priorities are set according to the remaining deadline time for RT users and to the difference between the achieved rate and the desired rate required for nRT users. The user with the highest priority is paired with the subchannel with the highest vector norm and semi-orthogonal users are multiplexed on the same subchannel. Comparisons against the algorithm of [18] show that the packet drop rate for RT users is significantly reduced. The algorithm's complexity is also reduced because of the semi-orthogonality criteria used to add users. However, as in [18], a performance comparison with a near-optimal solution is not provided.

In [20], the objective is to maximize a utility function without any hard minimum rate constraints for the RT users. The channel quality information is added to the utility function to favor users with good channel conditions and priorities are set by increasing user weights in the utility function. The advantage of this method is that the per-frame optimization problem has only a power constraint and no rate constraints, which makes it simpler to solve. The disadvantage with this *reactive* approach is that RT users with poor channel conditions are backlogged until their delay is close to the deadline, causing an increase in the average delay and jitter.

In [21], a heuristic algorithm is proposed for the sum rate maximization problem with proportional rates among the user data rates, i.e., the ratio among allocated user rates is predetermined. The criteria to form user groups includes semi-orthogonality as in [19], but also fairness through proportional rate constraints. This method is extended to include hard minimum rates in [22]. Again, there is no reported method to evaluate the accuracy of these heuristics, except by comparing them to each other.

Another approach is to maximize the sum rate subject to constraints on the average rate delivered to a user [23]. However, unlike the work presented in [24], where an optimal solution is provided for the single antenna OFDMA RA problem with average rate constraints,

the algorithm presented in [23] is an approximation. Note also that with average rate constraints, RT users tend to be served when they have good channel conditions which can create unwanted delay violations and jitter.

1.3.3 Extension to Multi-cell MISO-OFDMA Systems

The problem we formulate and solve in this dissertation considers only a single cell. We will *not* address the interference coming from other cells using the same channel bandwidth. Systems where the neighboring cells share the same channel bandwidth are important because spectrum re-use is necessary in dense areas. In addition, cell-edge users can greatly benefit from inter-cell interference reducing techniques. One avenue of future research is to extend the concepts and methods we develop in this dissertation to the multi-cell scenario; this scenario has been lately considered in [25], where a system with M' cells is considered for a single transmit and receive antenna. The objective is to maximize the users weighted sum rate under a power constraint per BS. The variables to optimize are the users to subcarrier assignment and the users' power allocation. The resulting problem is not convex and three methods are proposed to solve it. All three methods require that each user reports M' channel gains per subcarrier to a central control unit that processes all the data and makes the subcarrier assignment and power allocation decisions. The numerical results reported provide similar sum rates for all three cases and outperform uncoordinated transmitting strategies. To limit the signaling overhead, a reduced-feedback implementation of these methods is investigated, where the scheduling decisions are made locally by the BS and only the power allocation problem is jointly solved. The sum rate achieved in this case is close to the full feedback case.

In [26], downlink coordinated transmission is considered in a multi-cell OFDM system, where the BSs have multiple antennas and the users have single antennas. This is a direct extension of the MISO-OFDM studied in this dissertation to the inter-cell interference coordinated case. Users are divided in two groups: users that receive data only from a serving BS and consider the interference coming from other BS as noise, and users that are served in a coordinated manner by multiple BSs. The objective is to maximize a monotonically increasing function of the users' SNIR under linear power constraints. The optimization variables are the correlation matrices of the data symbol vectors along users and BSs, which model both the beamforming strategy and the user scheduling. Several properties of the optimal solution are derived to design heuristics that solve the problem efficiently. Two methods are proposed, one centralized algorithm and a distributed strategy. Numerical evaluations based on real channel measurements show that the solutions are close to the optimal and have sig-

nificant performance gains over single-cell processing schemes. The performance evaluation also shows that joint transmission is very sensitive to synchronization errors.

1.4 Thesis Contribution

None of the previous work listed in subsection 1.3.2 provides a near-optimal solution to the MISO-OFDMA RA problem with minimum rate requirements. This is important not only as a benchmark for any heuristic algorithms, but also to get a better insight into the problem and to help devise efficient real-time algorithms. We propose in this dissertation an off-line dual based method that provides a near-optimal solution to the problem.

In addition, we propose two efficient heuristic methods with much lower computational complexity than the methods proposed in the literature. The performance obtained with these heuristic method is within 10.7% of the optimal averaging over all performed numerical evaluations. They have lower computational complexity than the method proposed in [22]. The computational complexity reduction is several orders of magnitude depending on the algorithm used and the problem parameters. The proposed heuristic methods can also be adapted to be used in a LTE-Advanced system. To adapt these algorithms, however, we need to provide an explicit CSI feedback.

In summary, the main contributions of this dissertation are:

1. A dual formulation of the RA problem for a MISO-OFDMA system supporting minimum rate requirements, whose optimum gives us an upper bound to all feasible solutions to the primal problem. The dual function also provides us with a better insight of the RA problem regarding the dependance of its feasibility respect to the minimum rate constraints. In addition, it guides us to the design of the heuristic algorithms.
2. An off-line method that provides a near-optimal solution to the problem. The feasible point obtained by this method and the upper bound are used to find limits on the duality gap.
3. A heuristic method to solve the problem with lower computational complexity than methods reported in the literature. This method also extends the support of the minimum rate requirements.
4. A second heuristic method to solve the problem with much lower computational complexity when the number of subcarriers is large. It uses per subchannel power constraints instead of a total power constraint. As a drawback, this method supports lower minimum rates support than the heuristic method in item 3.

5. An efficient near-optimal power allocation heuristic method for the rate-constrained case. The method is used in this dissertation for an OFDM-SDMA system, but it can be used for power allocation problems in multicarrier SISO and MIMO systems.
6. A general procedure to adapt the proposed heuristic methods to the RA problem in one cell using LTE-Advanced technology.

As results of our research efforts, contributions 1 and 2 resulted in paper [17] for the rate unconstrained problem and [27] for the rate constrained one. On the other hand, contributions 3 to 5 are part of a paper being prepared at the date of submission of this thesis.

1.5 Thesis Organization

This dissertation is divided into six chapters. This first chapter, the thesis introduction, provides a general overview of the problem and justifies the research topic.

In the second chapter, we mathematically formulate the problem we want to solve. We will see that this is a non-convex problem, which is in general difficult to solve. We list several approaches we can use to find a solution method, and we restrict ourselves to transmission in the downlink direction, i.e., from the Base Station (BS) to single antenna users, implementing Space Division Multiple Access (SDMA) and linear precoding.

In the third chapter, we choose the dual Lagrange method from the approaches listed in chapter 2, and propose an off-line method that provides an upper bound to the solution of the problem. Based on this dual solution, we propose a simple off-line heuristic algorithm to compute a feasible point. This point constitutes a lower bound to the optimal solution. Then, these bounds are used to limit the duality gap. We compare results using the upper bound, the proposed off-line method and a weight adjustment heuristic.

In the fourth chapter, we propose two heuristic methods to solve the problem more efficiently. We compare their performance against the upper bound obtained in chapter 3 and against other methods proposed in the literature. The difference between the objective achieved by the heuristics and the upper bound is not large. In addition, the proposed heuristics have much lower computational complexity than the methods proposed in the literature.

In chapter 5, we describe a general procedure to use the algorithms proposed in chapter 4 to solve the RA problem in a single cell using LTE-Advanced technology. The algorithm's input and output parameters are mapped to LTE-Advanced parameters, assuming that explicit Channel State Information (CSI) is available at the base station.

In the final chapter, we summarize the main findings and propose future work for research. These concluding remarks provide the thesis general conclusions. Detailed conclusions and discussions about the proposed algorithms and results are given at the end of each chapter.

CHAPTER 2

PROBLEM FORMULATION AND SOLUTION METHODS

2.1 Introduction

The main objective of this chapter is to formulate and analyze the problem we want to solve in the dissertation. We start by describing in section 2.2 the system under consideration and by defining a signal model to compute the user rates in terms of the physical layer parameters we want to optimize. We restrict the transmission strategy to linear precoding beamforming and consider the cases of beamforming with and without Zero Forcing (ZF) constraints. Linear precoding beamforming does not achieve the full channel capacity but it is widely used because of its simpler implementation [28].

In section 2.3, we formulate a general optimization problem without ZF constraints for an arbitrary utility function and arbitrary rate constraints and we study the benefits of this problem formulation. In this dissertation, we focus on linear utility functions and minimum rate constraints using ZF beamforming. However, we initially formulate a more general problem to illustrate that the problem in hand belongs to a wider category.

In section 2.4, we add ZF constraints to the problem formulation which results in a non-linear mixed integer program (NLMIP). We examine the structure of the resulting subproblem after fixing the binary variable and find that this can be approximated to a convex one. Thus, the ZF problem formulation presents us a structure that we can later use to design efficient heuristics.

Also in section 2.4, we present two approaches to solve the ZF-constrained problem. The first one — the direct method — is only practical for small problem sizes, but it is illustrative of the transformation of the original non-convex problem into smaller convex problems. The second approach is the dual Lagrange method, which we use in chapter 3 to solve the problem with near optimality. The dual Lagrange method has a much lower complexity, but its dual optimal differs from the primal optimal due to the problem's non convexity. In section 2.4, we give a high level description of the dual method which is presented in detail in chapter 3. Along the chapter we provide an overview of other methods available to solve the problems and we illustrate the reasons behind the choice of the problem formulation and solution method we use. Finally, in section 2.5 we give the chapter conclusions.

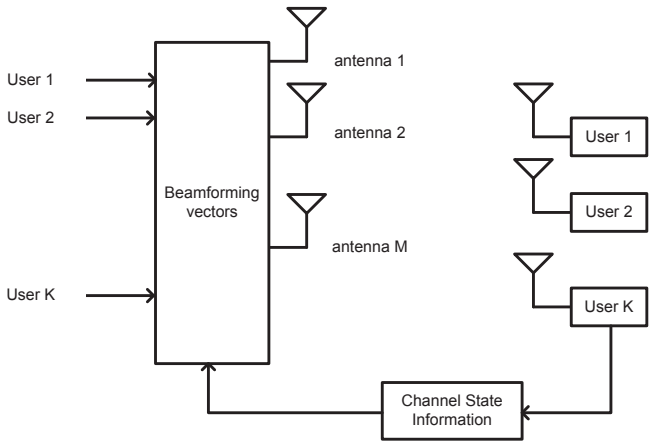


Figure 2.1 System Diagram

2.2 System Description

We consider the resource allocation problem for the downlink transmission in a multi-carrier multi-user multiple input single output (MISO) system with a single base station (BS). There are K users, some of which have RT traffic with minimum rate requirements while the others have nRT traffic that can be served on a best-effort basis. The BS is equipped with M transmit antennas and each user has one receive antenna. Figure 2.1 shows the system diagram. The system's available bandwidth W^o is divided into N subchannels whose coherence bandwidth is assumed larger than W^o/N, thus each subchannel experiences flat fading and OFDMA is effectively used.

In this configuration, the BS can transmit data in the downlink direction to different users on each subcarrier by performing linear beamforming precoding. At each OFDM symbol, the BS can change the beamforming vector for each user on each subcarrier to maximize some performance function. We assume that we use a channel coding that reaches the channel capacity. The data rate are in units of bits per OFDM symbol, or equivalently bits per second per Hertz (bps/Hz).

2.2.1 Signal Model

First, we describe the model used to compute the bit rate received by each user. Define

$\tilde{s}_{k,n}$ the symbol transmitted by the BS to user k on subcarrier n. We assume that the $\tilde{s}_{k,n}$ are independently identically distributed (i.i.d) random variables with $\tilde{s}_{k,n} \sim \mathcal{CN}(0,1)$.

$\mathbf{w}_{k,n}$ an M-component column vector representing the beamforming vector for user k on subcarrier n. Unless otherwise noted, we denote $\mathbf{w}$ the vector made up by the column stacking of the vectors $\mathbf{w}_{k,n}$.

$\mathbf{x}_n$ an M-component column vector representing the signal sent by the array of M antennas at the BS for each subcarrier n.

$\mathbf{h}_{k,n}$ an M-component row vector representing the channel between the M antennas at the BS and the receive antenna at user k for each subcarrier n.

$z_{k,n}$ the additive white gaussian noise at the receiver for user k on subcarrier n. The $z_{k,n}$ are i.i.d. Gaussian random variables and, without loss of generality, we assume that they have unit variance, that is $z_{k,n} \sim \mathcal{CN}(0,1)$.

$y_{k,n}$ the signal received by user k on subcarrier n.

$r^0_{k,n}$ the rate of user k on subcarrier n. We denote $\mathbf{r} \in \mathbb{R}^K$ the column vector of all user rates r_k, where $r_k = \sum_n r^0_{k,n}$.

The signal vector $\mathbf{x}_n$ is built by a linear precoding scheme which is a linear transformation of the information symbols $\tilde{s}_{k,n}$:

$$\mathbf{x}_n = \sum_{k=1}^{K} \mathbf{w}_{k,n} \tilde{s}_{k,n}. \tag{2.1}$$

The signal received by user k on subcarrier n is then given by

$$\begin{aligned} y_{k,n} &= \mathbf{h}_{k,n}\mathbf{x}_n + z_{k,n} \\ &= \mathbf{h}_{k,n}\mathbf{w}_{k,n}\tilde{s}_{k,n} + \sum_{k \neq j} \mathbf{h}_{k,n}\mathbf{w}_{j,n}\tilde{s}_{j,n} + z_{k,n}. \end{aligned} \tag{2.2}$$

The second and third terms in the right side of (2.2) correspond to the interference and noise terms. Since the signals and noise are Gaussian, the interference plus noise term is also Gaussian and the data rate of user k for subcarrier n is given by the Shannon channel capacity for an additive white Gaussian noise channel [29]:

$$r^0_{k,n}(\mathbf{w}) = \log_2 \left(1 + \frac{|\mathbf{h}_{k,n}\mathbf{w}_{k,n}|^2}{1 + \sum_{k \neq j} |\mathbf{h}_{k,n}\mathbf{w}_{j,n}|^2} \right). \tag{2.3}$$

In (2.3), beamforming vectors $\mathbf{w}_{k,n}$ with high norms and aligned to the channel vector $\mathbf{h}_{k,n}$, produce higher rates. In addition, if the other users' beamforming vectors are quasi-orthogonal to the channel vector $\mathbf{h}_{k,n}$, the inter-user interference in the denominator is low. The beamforming vectors completely determine the user rates in (2.3). Therefore, they can be used as the problem optimization variables.

In the problem formulation section 2.3, we assume that we know the parameters listed below

Known problem parameters

M Number of antennas at the BS.

N Number of subcarriers available.

K Number of users in the cell.

$\mathcal{K}$ Set of users in the cell: $\{1, \ldots, K\}$.

$\check{d}_k$ Minimum rate requirement for user k. We denote by $\check{\mathbf{d}} \in \mathbb{R}_+^K$ the vector made up of all minimum rates $\check{d}_k$.

$\mathcal{D}$ A subset of $\mathcal{K}$ containing the users that have minimum rate requirements $\check{d}_k > 0$. We define $D = |\mathcal{D}|$.

$\check{P}$ Total power available at the base station for transmitting over all channels.

c_k Weight of the user rates in the objective function. These could be computed by the scheduler to implement prioritization or fairness. We denote by $\mathbf{c} \in \mathbb{R}_+^K$ the vector made up of all weights c_k.

$\mathbf{h}_{k,n}$ the M-component row vector representing the channel between the M antennas at the BS and the receive antenna at user k for each subcarrier n.

We keep these definitions throughout the dissertation.

2.3 General Beamforming Problem

2.3.1 Problem Formulation

In general, the optimization problem consists of maximizing some utility function of the user rates. The user rates are determined by the beamforming vectors $\mathbf{w}$ through (2.3). The physical layer imposes constraints on these beamforming vectors. A common constraint is that the power sum of the beamforming vectors must be lower or equal than the total available power $\check{P}$. Then, the problem power constraint is given by

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \|\mathbf{w}_{k,n}\|^2 - \check{P} \leq 0, \tag{2.4}$$

i.e. the sum of the square norms of all beamforming vectors for all subcarriers must be lower or equal than the total power constraint $\check{P}$.

The utility maximization problem consists of maximizing an utility function u of the rates under certain rate constraints $\mathbf{g}$ over the set of power feasible beamforming vectors.

$$\max_{\mathbf{w}} u(\mathbf{r}(\mathbf{w})) \tag{2.5}$$

Table 2.1 Examples of utility functions

Weighted sum rate	$\mathbf{c}^T\mathbf{r}$
Maximum fairness	$\min_k r_k$
Proportional fairness	$\sum_k \ln(r_k)$
Sigmoidal	$\sum_k (1+\exp(-a(r_k-b)))^{-1}$

Table 2.2 Examples of rate constraint functions

Minimum rates	$-\mathbf{r}+\check{\mathbf{d}} \leq \mathbf{0}$
Maximum rates	$\mathbf{r}-\check{\mathbf{e}} \leq \mathbf{0}$
Proportional rates	$\frac{r_k}{\gamma_k} = a, \quad \forall k$

$$\mathbf{g}(\mathbf{r}(\mathbf{w})) \leq 0 \tag{2.6}$$

$$\sum_{k=1}^{K} \|\mathbf{w}_{k,n}\|^2 - \check{P} \leq 0, \tag{2.7}$$

where $\mathbf{r}$ is the vector of stacked user rates $r_k = \sum_n r_{k,n}$ and the utility function $u : \mathbb{R}^K \to \mathbb{R}$ is an increasing function of the user rates that represents the benefit we get out of the system. Examples of utility functions include weighted sum rate and sum of logarithm of the rates. Some are listed in table 2.1, where $\mathbf{r}$ is the vector of user rates $\{r_k\}$ and $a\,, b\,, \mathbf{c}$ indicate function parameters

The rate constraint functions $\mathbf{g} : \mathbb{R}^K \to \mathbb{R}^K$ are used as mechanisms to guarantee QoS to the users. They include limiting the minimum and maximum user rates, maintaining proportionality among rates, etc. Some examples are listed in table 2.2, where $\check{\mathbf{d}}$, $\check{\mathbf{e}}$ indicate box rate constraints, $\{\gamma_k\}$ the pre-determined proportionality values and $a > 0$ is any constant.

For the particular case where the utility function is linear and minimum rate constraints are enforced,

$$u(\mathbf{r}(\mathbf{w})) = \mathbf{c}^T\mathbf{r} \tag{2.8}$$

$$\mathbf{g}(\mathbf{r}(\mathbf{w})) = -\mathbf{r}+\check{\mathbf{d}}, \tag{2.9}$$

where $\mathbf{c} \in \mathbb{R}_+^K$ is the vector of user weights $\{c_k\}$ and $\check{\mathbf{d}} \geq \mathbf{0}$ is the vector with minimum rate requirements $\{\check{d}_k\}$ for the RT users. The resulting problem formulation is

$$\max_{\mathbf{w} \in \mathcal{C}^{KNM}} \sum_{n=1}^{N} \sum_{k=1}^{K} c_k r_{k,n}^0(\mathbf{w}_{k,n}) \tag{2.10}$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \|\mathbf{w}_{k,n}\|^2 - \check{P} \leq 0 \tag{2.11}$$

$$-\sum_{n=1}^{N} r_{k,n}^0(\mathbf{w}_{k,n}) + \check{d}_k \leq 0, \quad k \in \mathcal{D}. \tag{2.12}$$

This is the general problem we aim to solve in this dissertation, we later modify it in section 2.4 to include ZF constraints that ease the computation of the beamforming vectors. Let's study what we achieve with the solution to problem (2.10–2.12). The solution to this problem is the set of optimal beamforming vectors for all users at each subcarrier $\{\mathbf{w}_{k,n}^*\}$. In the case of interest when the number of users is higher than the number of antennas, i.e. $K > M$, the beamforming vectors of some users will be exactly zero at some subcarriers indicating that these subcarriers are not allocated to such users. Then, the first useful result is the per-subcarrier user selection.

Problem (2.10–2.12) maximizes a weighted sum rate by taking advantage of multi-user diversity and favoring the combination of users with good channel conditions at each subcarrier. Concurrently, the solution guarantees that the selected RT users are assigned rates higher or equal than the minimum rate constraints. In an actual system, the RA process sits below a scheduler that determines which users and with how much rate they need to be served at each frame. In addition, the user weights c_k can be varied from frame to frame by the scheduler to implement fairness among users. Problem formulation (2.10–2.12) gives us a mechanism to optimally assign resources per time-slot to both RT and non-RT users in an SDMA-OFDMA system. This has practical applications in current LTE-Advanced systems as we will see in chapter 5.

In summary, the benefits of problem formulation (2.10–2.12) are that it

– Exploits multi-user diversity
– Supports minimum rates for the selected RT users
– Can implement fairness among users.

Solving (2.10–2.12) is overly complex because of the problem's non-convexity as we explain in the following.

2.3.2 Non Convexity of the General Beamforming Problem

A non-linear program, where we maximize the objective function, is convex and therefore relatively easy to solve if the objective is a concave function of the optimization variables — the beamforming vectors in our case — and if the constraints are convex functions when written with inequalities of the form ≤ 0 [30]. In the problem formulation (2.10–2.12), constraint (2.11) is a convex function of the beamforming vectors. Define $\boldsymbol{\chi}_0$ as the set of beamforming vectors that satisfy the power constraint

$$\boldsymbol{\chi}_0 = \left\{ \left\{ \mathbf{w}_{k,n} \in \mathbb{C}^M \right\} : \sum_{n=1}^{N} \sum_{k=1}^{K} \|\mathbf{w}_{k,n}\|^2 \leq \check{P} \right\} \tag{2.13}$$

$\boldsymbol{\chi}_0$ is a convex set, it consists of a hypersphere of radius $\sqrt{\check{P}}$ in the $\mathbb{C}^{NKM}$ space.

On the other hand, the objective function (2.10) is not concave, and the rate constraints (2.12) are not convex due to the form of the rate expression (2.3). Therefore, problem (2.10–2.12) is a non-linear non-convex problem and generally hard to solve.

2.3.3 Solution Methods

In [31], Brehmer proposed solving problem (2.5–2.7) using the monotonic optimization method for the one-subcarrier case. This method has the advantage of supporting any type of utility function and rate constraints with the only requirement of being continuous and increasing, this allows to support utility functions like the sigmoidal in table 2.1. The same method can be used to solve problem (2.10–2.12) since it is a particular case. However, the computational demand of this method is extremely high which makes it impractical for realistic problem sizes. In addition, this method does not provides us with any insight into the problem's structure and ideas on how to design efficient heuristics.

Other way to solve problem (2.10–2.12), that gives us more insight into the problem's structure, is using a local non-linear programming (NLP) solver initialized at multiple starting points. Since the objective and constraints functions in problem (2.10–2.12) are smooth, we can use a derivative-based method. We explored this approach for a small problem size using the Minos solver with the AMPL modelling software [32]. Minos employs a projected Lagrangian algorithm [33] and assumes that the objective and constraint functions are twice differentiable. The solver finds one local optimal point, whose location depends on the given starting point. We used multiple starting points and compare all the local optima found, then we picked the largest objective. Figure 2.2 illustrates one example of the local maxima found for a single carrier system with three users and three antennas. In the figure, the local maxima are ordered from largest to smallest. The global solution is the first one in

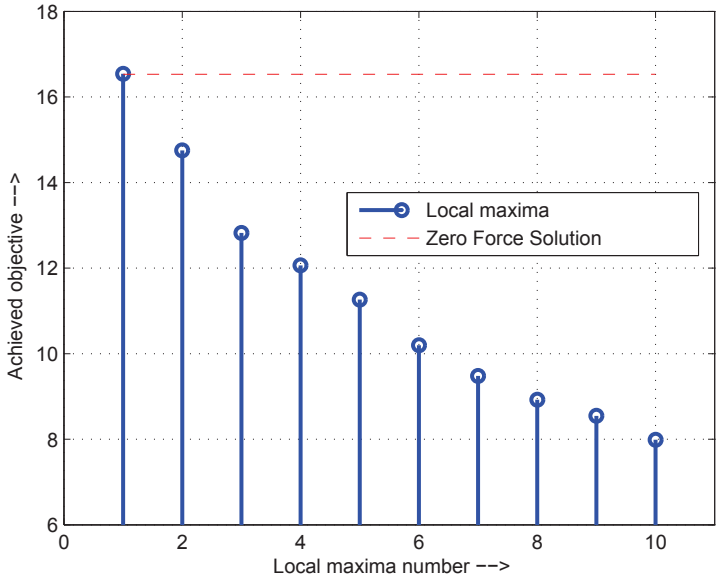


Figure 2.2 Example of local maxima found for a system with $K = 3, M = 3, N = 1, \check{P} = 10$, $\check{\mathbf{d}} = \mathbf{0}$.

the figure. Notice that even for a small problem size, there are many local optima which makes this method computationally unfeasible for practical problem sizes. In addition, global optimality is not guaranteed without scanning all local maxima. However, this method shows us the beamforming vector characteristics of the largest local maxima found; we observe that in most cases some of the beamforming vectors are exactly zero, and that the beamforming vectors that are different than zero are close to the beamforming vectors obtained when ZF beamforming is applied.

After studying the outcome of multiple optimization experiments for different channel realizations we obtain the following observations for a system with $M = K$:

1. When some user channels are close to co-linearity, the global maxima occurs when one or more beamforming vectors are zero. Thus, the number of selected users per subcarrier is between 1 and $M - 1$ and the system does not exploit all degrees of freedom.
2. When user channels are *not* close to co-linearity, the global maxima corresponds to a local optimal solution where none of the resulting beamforming vectors is zero. In this case the number of selected users per subcarrier is equal to M and the system exploits all degrees of freedom.

For the case $K > M$ there is also a user selection process. The optimal solution is achieved by a combination of users that is always between 1 and M. On the other hand, the objectives achieved by solving (2.10–2.12) with and without ZF constraints are very close when the

SNR is high. For example, in figure 2.2 we indicate the ZF solution by a dashed line, the difference between both objectives is very small, i.e., 16.540 vs. 16.529.

It is well known that ZF beamforming is a sub-optimal solution to the sum rate maximization problem. But when the SNR is high or the selected channel vectors are semiorthogonal, the ZF beamforming solution is very close to the optimal [28]. Since in the multi-user multi-carrier scenario with Rayleigh fading, the probability of finding a set of users with good channel conditions increases rapidly with the number of users, it is likely that we obtain a solution using ZF beamforming that is not too far from the optimal. We then will be satisfied with the zero force solution specially if it is much less expensive to compute. This motivates us to look for an alternative problem formulation and solution method using ZF constraints in section 2.4.

2.4 Zero Forcing Constrained Problem Formulation

Global optimization methods could serve our purpose of finding an optimal solution to problem (2.10–2.12) to compare against more efficient methods. However, they have two major drawbacks. First, they require a long time to compute even for an off-line method. Second, some methods do not give us any insight into the problem structure and how we can take advantage of it to design efficient heuristics. In chapter 3 we show that adding ZF constraints to the users selected for each subcarrier, eases the computation of the beamforming vectors $\mathbf{w}$ and provides us with a structure that we can exploit to design efficient heuristics.

Zero-forcing beamforming is a strategy that completely eliminates interference from other users. For each subcarrier n, we choose a set ϕ of $g \leq M$ users which are allowed to transmit. This is called an *SDMA* set. We then impose the condition that for each user k in this set, the beamforming vector of user k must be orthogonal to the channel vectors of all the other users in the set. This amounts to adding the orthogonality constraints

$$\mathbf{h}_{k,n}\mathbf{w}_{j,n} = 0 \quad j \neq k,\ j,k \in \phi \tag{2.14}$$

and the user k data rate for subcarrier n (2.3) then simplifies to

$$r_{k,n}^{(1)}\left(\mathbf{w}_{k,n}\right) = \log_2\left(1 + |\mathbf{h}_{k,n}\mathbf{w}_{k,n}|^2\right). \tag{2.15}$$

With ZF beamforming, the problem is now made up of two parts. We need to select a SDMA set $\phi(n)$ for each subcarrier n and, for each selected SDMA set, we must compute the beamforming vectors in such a way that the total rate received by all users is maximized. Because of this, we need to add another set of decision variables

$\alpha_{k,n}$ a binary variable that is 1 if we allow user k to transmit on subcarrier n and zero otherwise. We denote the collection of $\alpha_{k,n}$ by the vector $\boldsymbol{\alpha}$.

Replacing the user rate expression (2.15) and adding the new optimization variable $\boldsymbol{\alpha}$ and ZF constraints results in the following optimization problem

$$\max_{\mathbf{w},\boldsymbol{\alpha}} \sum_{n=1}^{N}\sum_{k=1}^{K} c_k r_{k,n}^{(1)}(\mathbf{w}_{k,n}) \tag{2.16}$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \|\mathbf{w}_{k,n}\|^2 - \check{P} \leq 0 \tag{2.17}$$

$$-\sum_{n=1}^{N} r_{k,n}^{(1)}(\mathbf{w}_{k,n}) + \check{d}_k \leq 0, \quad \forall k \in \mathcal{D} \tag{2.18}$$

$$\sum_{k=1}^{K} \alpha_{k,n} \leq M, \quad \forall n \tag{2.19}$$

$$|\mathbf{h}_{k,n}\mathbf{w}_{j,n}|^2 \leq B\left[(1-\alpha_{k,n}) + (1-\alpha_{j,n})\right], \quad \forall n, \forall k, \forall j,\ k \neq j \tag{2.20}$$

$$\|\mathbf{w}_{k,n}\| \leq A\alpha_{k,n} \tag{2.21}$$

$$\alpha_{k,n} \in \{0,1\} \tag{2.22}$$

where A and B are some large positive constants. Constraints (2.17) and (2.18) are the same power and rate constraints as in the general beamforming case. Constraint (2.19) guarantees that we do not choose more than M users for each subcarrier, constraint (2.20) guarantees that if we have chosen two users k and j, they meet the ZF constraints while for other users the constraint is redundant, and constraint (2.21) guarantees that the beamforming vector is zero for users that are not chosen.

Problem (2.16–2.22) is a non-linear mixed integer program (NLMIP). The vector of binary variables $\boldsymbol{\alpha}$ determines the set of users that are assigned to each subcarrier. On the other hand, the vector of continuous variables $\mathbf{w}$ need to comply with the ZF constraints (2.20–2.21) which depend on the user selection binary variables $\boldsymbol{\alpha}$. The objective function (2.16) depends only on the set of variables $\mathbf{w}$.

There are many off-the-shelf software packages available to solve NLMIPs, see [34] for a survey. They use different methods with different levels of accuracy and speed. However, the current NLMIP solvers do not automatically exploit the specific structure of our problem. For this reason, we devise in chapter 3 an off-line near optimal method that serves as a benchmark for more efficient methods to solve problem (2.16–2.22).

In subsection 2.4.1, we study the structure of the resulting subproblem after fixing the binary variables $\boldsymbol{\alpha}$ in problem (2.16–2.22). We will see that the resulting subproblem can be

approximated to a small convex problem. A complete enumeration on the binary variables would require to solve I sub-problems, where $I \sim K^{MN}$. This is not practical for realistic problem sizes, but it is illustrative of the problems' structure.

In subsection 2.4.2 we outline a dual Lagrange method that reduces the number of sub-problems to $I \sim NK^M$, so we can use this method to obtain a solution in a reasonable amount of time. The methods in subsections 2.4.1 and 2.4.2 give optimal and near optimal solutions to problem (2.16–2.22).

2.4.1 Structure of Subproblem for Fixed $\boldsymbol{\alpha}$ and Direct Method

After fixing $\boldsymbol{\alpha}$ in problem (2.16–2.22), the resulting subproblem over variable $\mathbf{w}$ is

$$\max_{\mathbf{w}} \sum_{n=1}^{N} \sum_{k=1}^{K} c_k r_{k,n}^{(1)}(\mathbf{w}_{k,n}) \tag{2.23}$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \|\mathbf{w}_{k,n}\|^2 - \check{P} \leq 0 \tag{2.24}$$

$$-\sum_{n=1}^{N} r_{k,n}^{(1)}(\mathbf{w}_{k,n}) + \check{d}_k \leq 0, \quad \forall k \in \mathcal{D} \tag{2.25}$$

$$|\mathbf{h}_{k,n}\mathbf{w}_{j,n}|^2 \leq B\left[(1-\alpha_{k,n}) + (1-\alpha_{j,n})\right], \quad \forall n, \forall k, \forall j,\ k \neq j \tag{2.26}$$

$$\|\mathbf{w}_{k,n}\| \leq A\alpha_{k,n} \tag{2.27}$$

The objective (2.23) is not concave and the rate constraints (2.25) are not convex, thus the problem is not convex. However, subproblem (2.23–2.27) can be approximated to a convex problem. We perform this approximation in section 4.2 in detail. Here, we outline the basic principle through a small example.

Let's consider the case of one subcarrier, two users, two transmit antennas and the channel vectors listed in figure 2.3. The user channel vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ are illustrated in the figure. We drop the subindex n for this example because there is only one subcarrier. There are three possible values for $\boldsymbol{\alpha}$, i.e., three possible SDMA sets. Let's focus on the SDMA set where both users are picked, this is equivalent to fix $\alpha_1 = 1$ and $\alpha_2 = 1$. The ZF constraints (2.26) impose the beamforming vectors to be orthogonal to the other user channel vector. This is illustrated by the thick lines in figure 2.3. The vectors in set Ω_1 are orthogonal to $\mathbf{h}_2$, and the vectors in set Ω_2 are orthogonal to $\mathbf{h}_1$, i.e.,

$$\Omega_1 = \{\mathbf{w}_1 \in \mathbb{C}^M : \mathbf{h}_2\mathbf{w}_1 = 0\},$$
$$\Omega_2 = \{\mathbf{w}_2 \in \mathbb{C}^M : \mathbf{h}_1\mathbf{w}_2 = 0\}.$$

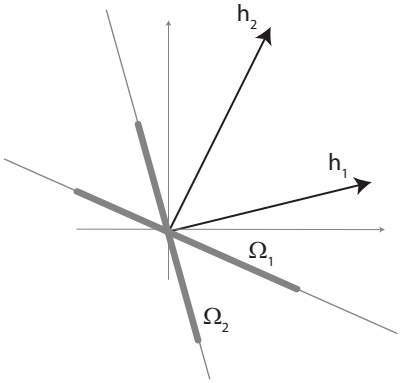


Figure 2.3 Beamforming vectors support of ZF constrained prob. for SDMA set with 2 users, $K = 2, M = 2, N = 1, \mathbf{h}_1 = (2, 1), \mathbf{h}_2 = (1, 2)$

Each set is convex but the union of sets: $\Omega = \Omega_1 \cup \Omega_2$ is not convex. The sum of the norms is limited by the power constrained $\check{P}$. The resulting subproblem can be approximated to a convex problem by fixing the direction of the beamforming vectors using the ZF constraints and changing the optimization variable $\mathbf{w}_k$ to a scalar variable $q_k \geq 0$. The new optimization variable is related to the beamforming vectors by $q_k = |\mathbf{h}_k \mathbf{w}_k|^2$. Then, after replacing in (2.15) the user rate becomes a concave function of the user power

$$r_k^{(1)} = \log(1 + q_k) \tag{2.28}$$

and the power constraint becomes linear (c.f. problem (4.13–4.15))

$$\sum_{k=1}^{K} \beta_k q_k - \check{P} \leq 0, \quad \beta_k \geq 0. \tag{2.29}$$

where β_k is given by (4.10) for any given subcarrier. With this approximation solving each subproblem is easy since it is convex and the set of active users is small ($\leq M$). Moreover, the heuristic method we present in chapter 4 iteratively solves one subproblem for fixed $\boldsymbol{\alpha}$ and then changes the subcarrier assignment until the minimum rate constraints are satisfied. Thus, both the direct and the heuristic methods solve the same convex sub-problem.

A direct method where we perform a complete enumeration of the binary variables would result in an extremely high number of sub-problems. For each subcarrier n, the number of possible values for variable $\alpha_{k,n}$ equals the number of possible SDMA sets

$$I = \sum_{m=1}^{M} \binom{K}{m} \sim K^M \tag{2.30}$$

and we have to include all possible combinations of SDMA sets and subcarriers. Thus, the number of subproblems to solve is

$$I_{direct} = I^N = \left(\sum_{m=1}^{M} \binom{K}{m} \right)^N \sim K^{MN}. \tag{2.31}$$

In the direct method after solving all subproblems with each feasible permutation of $\boldsymbol{\alpha}$, we simply pick the largest objective. In section 3.3, we show some results using this method for small problems, but for larger cases this method is not practical even for off-line purposes. To avoid the N exponent in the number of subproblems to solve (2.31), we outline the dual Lagrange decomposition method in subsection 2.4.2.

2.4.2 Dual Lagrange Method

Dual Lagrange decomposition has been used in a wide range of problems in telecommunications. For example in [24, 35], it is used to perform optimal resource allocation in SISO OFDMA systems. The main advantage of the dual Lagrange method is that constraints that are coupled in the primal domain can be decoupled in the dual domain. For OFDM systems, this is of particular interest because power and minimum rate constraints are coupled across subcarriers and solving the problem in the primal domain implies considering all dimensions concurrently. In contrast, the dual Lagrange method solves independent subproblems across subcarriers and links the constraints via the dual variables. Therefore, the exponential dependency with N in the number of subproblems to solve in (2.31) can be converted to a linear dependency.

We come back to the complete ZF-constrained problem (2.16–2.22). Here we give a high level description of the dual method which is presented in detail in chapter 3. The first step is to form a Lagrangian dualizing the power and rate constraints (2.24–2.25), for which we define dual variable $\lambda \geq 0$ and the vector of dual variables $\boldsymbol{\mu} \in \mathbb{R}_+^K$ obtaining

$$\begin{aligned} \mathcal{L}_1(\mathbf{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\mu}) = &- \sum_{k=1}^{K} \sum_{n=1}^{N} c_k r_{k,n}^{(1)} - \lambda \left(\sum_{k=1}^{K} \sum_{n=1}^{N} \|\mathbf{w}_{k,n}\|^2 - \check{P} \right) \\ &+ \sum_{k=1}^{K} \mu_k \left(\check{d}_k - \sum_{n=1}^{N} r_{k,n}^{(1)} \right) \end{aligned} \tag{2.32}$$

Then, the Lagrangian is minimized over the primal variables to obtain the dual function

$$\Phi(\lambda, \boldsymbol{\mu}) = \min_{\boldsymbol{\alpha}, \mathbf{w}} \mathcal{L}_1(\mathbf{w}, \boldsymbol{\alpha}, \lambda, \boldsymbol{\mu}) \tag{2.33}$$

$$= -\lambda\check{P} + \boldsymbol{\mu}^T\check{\mathbf{d}} - \min_{\boldsymbol{\alpha}\in\mathcal{A}_1}\min_{\mathbf{w}\in\mathcal{W}_1(\boldsymbol{\alpha})}\sum_{n=1}^{N}\sum_{k=1}^{K}(c_k+\mu_k)r_{k,n}^{(1)}$$
$$+\lambda\sum_{k=1}^{K}\sum_{n=1}^{N}\|\mathbf{w}_{k,n}\|^2, \tag{2.34}$$

where $\mathcal{A}_1$ is the set of feasible $\boldsymbol{\alpha}$ that fulfill constraints (2.19) and (2.22), and $\mathcal{W}_1$ is the set of feasible vectors for each $\boldsymbol{\alpha}$ that complies with the ZF constraints (2.20) and (2.21).

The advantage of the dual formulation is that it allows to decouple the minimization problem because the variables in (2.34) are independent across subcarriers. Therefore, the sum operator with n as index in (2.34) can be taken outside the minimization, giving

$$\Phi(\lambda,\boldsymbol{\mu}) = -\lambda\check{P} + \boldsymbol{\mu}^T\check{\mathbf{d}} - \sum_{n=1}^{N}\left(\min_{\boldsymbol{\alpha}\in\mathcal{A}_1}\min_{\mathbf{w}\in\mathcal{W}_1(\boldsymbol{\alpha})}\sum_{k=1}^{K}(c_k+\mu_k)r_{k,n}^{(1)} + \lambda\sum_{k=1}^{K}\|\mathbf{w}_{k,n}\|^2\right) \tag{2.35}$$

which translates to N independent doubled minimization problems for each value of $\lambda, \boldsymbol{\mu}$. In section 3.2.2, we will see that each minimization problem over $\mathbf{w}$ can be approximated to a convex problem and the optimum can be expressed in terms of the dual variables $\lambda, \boldsymbol{\mu}$. The dual problem associated to the dual function is

$$\max_{\lambda,\boldsymbol{\mu}} \Phi(\lambda,\boldsymbol{\mu}) \tag{2.36}$$
$$\lambda \geq 0 \tag{2.37}$$
$$\boldsymbol{\mu} \geq 0, \tag{2.38}$$

where Φ is a concave function of the dual variables. Therefore, an iterative process on $\lambda, \boldsymbol{\mu}$ can solve the problem to dual optimality. If p^* is the primal optimum of problem (2.16–2.22) and Φ^* is the optimum of the dual problem, weak duality [30] establishes that

$$p^* \leq -\Phi^*. \tag{2.39}$$

If the primal problem is convex and it satisfies the Slater condition, there is no difference between the primal and dual optimal and (2.39) is satisfied with equality, meaning that we have strong duality. When the dual problem is much easier to solve, then the dual problem formulation provides a more efficient way to find the optimal.

Solving dual problem (2.36–2.38) implies three steps. First, minimize over the primal variables $\mathbf{w}$ in (2.35). This is approximately a convex problem of small size as we will see in chapter 3. Second, minimize over the user selection variables $\boldsymbol{\alpha}$ in (2.35). This is done by enumeration since the search space can be reduced to a small size for practical problem sizes.

And third, solving the dual problem (2.36–2.37), which can be done using the subgradient method since a subgradient is already provided by the difference between the total power and the power constraint, and the difference between the required rates $\check{\mathbf{d}}$ and the rates achieved by the current $\lambda, \boldsymbol{\mu}$.

However, problem (2.16–2.22) is not convex. Therefore, the optimal primal objective attained differs from the dual optimal by an amount lower or equal than the duality gap, i.e., (2.39) is satisfied with inequality. If the gap is small, let's say a few percent, we can still use such solution as a benchmark for more efficient heuristic methods. Assuming that the duality gap is small, the main motivation for using the dual method for non-convex problems is that the solution can be obtained much quicker than using other methods. The number of subproblems to solve is

$$I_{\text{dual}} \sim (NK^M) \tag{2.40}$$

which does not have the the exponential dependency K^N that (2.31) does. This is a substantial reduction in complexity since in practical LTE-Advanced systems N can be as large as 550. We still have the exponential dependency K^M in (2.40). However, M is not very large in practical systems (up to 8 in LTE-Advanced) and the SDMA search size can be reduced to a manageable size by a number of techniques, see [17]. Results in chapter 3 show that the duality gap is small for different system configurations. Thus, the dual method can provide us with an efficient off-line method to find a near-optimal solution to the problem.

2.4.3 Advantages of the ZF-constrained Over the General Beamforming Problem Formulation

In this chapter we have formulated two problems, namely the general beamforming problem (2.10–2.12) and the ZF-constrained problem (2.16–2.22). The difference between these two problems is the set of ZF constraints (2.20–2.21), which imply that the optimum of the ZF-constrained problem is lower or equal than the optimum of the general beamforming problem. For a particular channel realization, the difference between the solutions depends on both the SNR and the spatial distribution of the selected user channels. If the selected user channel vectors are close to orthogonality, their projection onto the null space of the other channel vectors is close to the original channel vector. Thus, the difference between the achieved sum rate by both problem solutions will be very small. On the other hand, ZF beamforming produces low rates when the SNR is low or there is a high spatial correlation among the user channels. In [2], a comparison is made between the MMSE and ZF receivers for an 8×8 MIMO system considering Rayleigh fading. At high SNR, both receivers achieved approximately 75% of the channel capacity, with the MMSE receiver being just slightly bet-

ter than the ZF receiver. For low SNR, however, the MMSE receiver achieves approximately 85% of the channel capacity while the ZF receiver only achieves 20%. In this comparison, however, multiuser diversity is not considered. When considering many users and Rayleigh fading, the probability that all SDMA sets have low SNR or are close to colinearity *simultaneously*, decreases rapidly with the number of users. Then, for high K and uncorrelated channels the probability that *all* SDMA sets do not present good channel conditions is very small. In [28], it is proved that when the number of users is large, ZF beamforming combined with user selection achieves a sum rate that has the same scaling law as that of Dirty Paper Coding (DPC) which is the optimal strategy. For these reasons, when considering Rayleigh fading the difference between the solutions of the general beamforming problem (2.10–2.12) and the ZF-constrained problem (2.16–2.22) is small for most channel realizations.

The fact that the optimization of the beamforming vectors for fixed $\boldsymbol{\alpha}$ can be approximated to a convex problem can be exploited to design more efficient methods. For example in a practical on-line heuristic method, we do not need to scan all SDMA sets since some will be formed by channels that have bad channel conditions and can be discarded beforehand. We would need to scan only the channels that have good channel conditions and select one combination that seems suitable, resulting in a convex subproblem that gives us a suboptimal solution. Therefore, a global non-convex method that exploits the fact that the problem can be decomposed into multiple convex subproblems would be useful when designing these heuristics. Because the ZF-constrained problem formulation allows such problem decomposition, from now on in this dissertation we focus on solving this problem only and discard the general beamforming problem (2.10–2.12).

2.5 Chapter Conclusion

From the two problems formulated in this chapter: the general beamforming problem (2.10–2.12) and the ZF-constrained problem (2.16–2.22), we chose to solve the ZF-constrained problem and discard the general beamforming problem because the former one provides us with a structure that we can exploit to find efficient solution methods.

We explored two approaches to solve the ZF-constrained problem. The first approach is a direct method that performs enumeration on the binary variable. It is not efficient, but is illustrative of the approximation of the complete non-convex problem into multiple smaller convex problems. The second approach — the dual Lagrange method — allows us to decouple the problem across subcarriers reducing the computational complexity with respect to the number of them. However, this method produces a duality gap caused by the problem's non convexity. Thus, we need to evaluate such duality gap to validate the method's applicability.

In chapter 3, we use the dual Lagrange approach to find an off-line near-optimal solution and we evaluate the duality gap for several system configurations. In chapter 4, we design efficient sub-optimal heuristics that use the dual Lagrange method and exploit the structure of the ZF-constrained problem.

CHAPTER 3

DUAL-BASED BOUNDS

3.1 Introduction

In the previous chapter we formulated the problem we solve in this dissertation. It is a non-linear mixed integer program (NLMIP) that is in general hard to solve. We saw that the problem can be decomposed into a very large number of multiple convex sub-problems of smaller size, but using this approach is *not* practical for typical problem sizes. We also saw that the dual approach can reduce the problem complexity, but that for non convex problems — like ours — the dual method provides a non zero duality gap. We cannot use the dual bound as a benchmark of more efficient heuristic methods if the duality gap is large. Therefore in this chapter, we estimate this duality gap for typical system configurations to validate the usefulness of the dual method to solve the problem.

There is a direct way to compute the duality gap, which is to find the difference between the dual and primal solutions. However, finding the primal optimum by a direct method is impractical for a typical problem size. Instead, we will estimate limits on the duality gap in the following way. We find the dual optimum that establishes an upper bound on all feasible points. If after primal recovery, the point obtained from the dual optimal is feasible, we use this as a our primal solution. If it is not feasible, we search in the dual space around the dual optimal until we find a primal feasible point. This will give us a lower bound on the primal optimum. The difference between the upper and lower bounds will establish an upper limit to the duality gap.

To follow this approach, we need to find a method to solve the dual problem efficiently. In this chapter, we devise such a method and compare its performance against a weight adjustment heuristic that serves as a reference point to find better heuristics in chapter 4.

The main contribution of this chapter is a method that provides an upper bound to the solution to problem (3.1–3.7). A second contribution is a simple off-line heuristic algorithm to compute a feasible point based on the dual solution. This point is a lower bound for the optimal solution and, in conjunction with the upper bound, will be used to bound the optimality gap in larger cases where an optimal solution is not available.

We then study several cases where we compare the performance of the upper and lower bounds. The results show that they are tight when the number of RT users is small. We find that their difference increases when the number of RT users increases but that it stays small.

Thus, the dual method provides a good approximation to the optimal solution. We also compare the performance of the weight adjustment algorithm versus the upper bound. The results indicate that adjusting the user weights to prioritize RT users leads to significantly sub-optimal solutions.

We present in Section 3.2 the dual-based method and two algorithms: one that finds the dual solution (the upper bound) and the other that finds a dual-based primal feasible solution (the lower bound). In the same section, we study the dual function and relate its shape to the activation of the rate constraints and the problem feasibility. In Section 3.3, we present numerical results showing the accuracy of the upper and lower bounds and of the weight adjustment algorithm for different scenarios. Finally, we present our conclusions in Section 3.4.

For the reader's convenience, we repeat here the ZF problem formulation from section 2.4 since we make many references to it in this chapter.

$$\max_{\mathbf{w},\boldsymbol{\alpha}} \sum_{n=1}^{N}\sum_{k=1}^{K} c_k r_{k,n}^{(1)}(\mathbf{w}_{k,n}) \tag{3.1}$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \|\mathbf{w}_{k,n}\|^2 \leq \check{P} \tag{3.2}$$

$$\sum_{n=1}^{N} r_{k,n}^{(1)}(\mathbf{w}_{k,n}) \geq \check{d}_k, \quad \forall k \in \mathcal{D} \tag{3.3}$$

$$\sum_{k=1}^{K} \alpha_{k,n} \leq M, \quad \forall n \tag{3.4}$$

$$|\mathbf{h}_{k,n}\mathbf{w}_{j,n}|^2 \leq B\left[(1-\alpha_{k,n})+(1-\alpha_{j,n})\right], \quad \forall n, \forall k, \forall j,\ k \neq j \tag{3.5}$$

$$\|\mathbf{w}_{k,n}\| \leq A\alpha_{k,n} \tag{3.6}$$

$$\alpha_{k,n} \in \{0,1\} \tag{3.7}$$

3.2 Dual-Based Solution Method

We cannot solve problem (3.1–3.7) fast enough to use it for a real time algorithm because it is NP-complete [11], the actual computation time becomes quickly prohibitive for realistic problem sizes even for off-line computations. We present in this section two off-line solution techniques that are tractable for problems of realistic size based on the Lagrange relaxation of the primal.

Solving the ZF problem will require some form of search over the $\alpha_{k,n}$ variables. Note that this ranges over all subsets with a number of users smaller than or equal to M, so

that the search space is going to be fairly large. Our first transformation is thus to separate the problem into single-subcarrier subproblems. For this, we dualize the constraints (3.2) and (3.3) since they are the ones that couple the subcarriers. Define the dual variables

λ Lagrange multiplier for power constraint (3.2).

μ_k Lagrange multipliers for minimum rate constraint (3.3) of user k. The collection of μ_k is denoted $\boldsymbol{\mu}$.

In order to simplify the derivation, we also define the dual variables μ_k for all users $k \in \mathcal{K}$. For users with no minimum rate requirements ($k \notin \mathcal{D}$), we have $\mu_k = 0$ and $\check{d}_k = 0$. In what follows, we use the standard form of Lagrangian duality which is expressed in terms of minimization with inequality constraints of the form $\leq$. Under these conditions, the multipliers $\lambda, \boldsymbol{\mu} \geq 0$. We get the partial Lagrangian

$$\begin{aligned}\mathcal{L} &= -\sum_{n=1}^{N}\sum_{k=1}^{K} c_k r_{k,n}^{(1)}(\mathbf{w}_{k,n}) + \lambda\left[\sum_{n=1}^{N}\sum_{k=1}^{K}\|\mathbf{w}_{k,n}\|^2 - \check{P}\right] \\ &\quad + \sum_{k=1}^{K}\mu_k\left[\sum_{n=1}^{N} -r_{k,n}^{(1)}(\mathbf{w}_{k,n}) + \check{d}_k\right] \\ &= -\lambda\check{P} + \sum_{k=1}^{K}\mu_k\check{d}_k + \sum_{n=1}^{N}\left\{-\sum_{k=1}^{K}(c_k+\mu_k)r_{k,n}^{(1)}(\mathbf{w}_{k,n}) + \lambda\sum_{k=1}^{K}\|\mathbf{w}_{k,n}\|^2\right\}. \qquad (3.8)\end{aligned}$$

The value of the dual function Θ at some point $(\lambda, \boldsymbol{\mu})$ is obtained by minimizing the Lagrange function over the primal variables

$$\Theta(\lambda, \boldsymbol{\mu}) = \min_{\mathbf{w},\boldsymbol{\alpha}} \mathcal{L}(\lambda, \boldsymbol{\mu}, \mathbf{w}, \boldsymbol{\alpha}) \qquad (3.9)$$

and the dual problem is

$$\max_{\lambda,\boldsymbol{\mu}} \Theta(\lambda, \boldsymbol{\mu}) \qquad (3.10)$$

$$\lambda, \boldsymbol{\mu} \geq 0 \qquad (3.11)$$

which we can solve by the well known subgradient algorithm [36]. From now on, we concentrate on the calculation of subproblem (3.9).

3.2.1 Subchannel Subproblem

Because of the relaxation of the carriers coupling constraints (3.2–3.3), the subproblems in (3.9) decouple by subcarrier since the objective (3.8) is separable in n and so are constraints (3.4–3.6). Computing the dual function then requires the solution of N independent

subproblems. For each subcarrier n, this has the form

$$\min_{\mathbf{w}_n,\boldsymbol{\alpha}_n} \; -\sum_{k=1}^{K}(c_k+\mu_k)r_{k,n}^{(1)}(\mathbf{w}_{k,n})+\lambda\sum_{k=1}^{K}\|\mathbf{w}_{k,n}\|^2 \tag{3.12}$$

$$\sum_k \alpha_{k,n} \le M, \tag{3.13}$$

$$|\mathbf{h}_{k,n}\mathbf{w}_{j,n}|^2 \le B\left[(1-\alpha_{k,n})+(1-\alpha_{j,n})\right], \quad \forall k,\, \forall j,\, k\neq j \tag{3.14}$$

$$\|\mathbf{w}_{k,n}\| \le A\alpha_{k,n} \tag{3.15}$$

$$\alpha_{k,n} \in \{0,1\}$$

where $\mathbf{w}_n$ is the vector made up by the column stacking of the vectors $\mathbf{w}_{k,n}$ for subcarriers n and $\boldsymbol{\alpha}_n$ denote the collection of $\alpha_{k,n}$ for subcarrier n. Problem (3.12–3.15) is still a mixed NLP, albeit of a smaller size.

3.2.2 SDMA Subproblem

A simple solution procedure is to enumerate all possible choices for $\alpha_{k,n}$ that meet constraint (3.4). This is called the *extensive* formulation of the problem. Each choice defines a SDMA set s and $\kappa = |s|$. Furthermore, for each SDMA set s, the problem separates into κ independent problems to compute the optimal beamforming vector $\mathbf{w}_{k,n,s}$ for each user $k \in s$. For each user $k \in s$, we know the set of channel vectors for the other members of s and we collect these vectors in the $(\kappa-1)\times M$ matrix $\mathbf{H}_{k,n,s}$. Problem (3.12–3.15) can then be rewritten as

$$\min_s \quad f_{n,s} \tag{3.16}$$

$$f_{n,s} = \sum_{k\in s} f^*_{k,n,s} \tag{3.17}$$

$$f^*_{k,n,s} = \min_{\mathbf{w}_{k,n,s}} -c'_k \log_2\left(1+|\mathbf{h}_k\mathbf{w}_{k,n,s}|^2\right)+\lambda\|\mathbf{w}_{k,n,s}\|^2 \tag{3.18}$$

$$\mathbf{H}_{k,n,s}\mathbf{w}_{k,n,s} = 0 \tag{3.19}$$

where $c'_k = c_k+\mu_k$, $\mathbf{w}_{k,n,s}$ is the beamforming vector for user k on subcarrier n for SDMA set s, and $\mathbf{w}_{n,s}$ is the vector made up by the column stacking of the vectors $\mathbf{w}_{k,n,s}$ for the κ users in s. Note that constraint (3.13) is automatically satisfied by the construction of s, constraint (3.15) simply drops out since $\mathbf{w}_{k,n,s}=0$ for $k\notin s$ and constraint (3.14) remains only for $k\in s$, but we write it as (3.19) because we are considering only users that belong to SDMA set s.

This is certainly not a feasible real-time algorithm, but for realistic values of K and M, the number of SDMA sets is still manageable and the optimization sub-problem (3.16–3.19) is a small nonlinear program. There are M variables and $\kappa - 1$ linear constraints. It can thus be solved quickly by a number of techniques. Still, the overall computation load can be quite large. There will be κ such problems to solve for each SDMA set, and there are $S = \sum_{i=1}^{M} \binom{K}{i}$ such sets for each of the N subcarriers so that we have to solve the problem $\kappa \times S \times N$ times, and this for each iteration of the subgradient algorithm. Clearly, any simplification of the beamforming subproblem can reduce the overall computation time significantly.

3.2.3 Approximate Solution to the Beamforming Problem

This can be done by the following construction. Instead of searching in the whole orthogonal subspace of $\mathbf{H}_{k,n,s}$ as defined by (3.19), we pick a direction vector in that subspace and search only on its support. This will give a good approximation to the extent that the direction vector is close to the optimal vector. The choice of direction is motivated by the fact that the objective function depends only on the product $\mathbf{h}_k \mathbf{w}_{k,n,s}$. Let's introduce a new independent variable

$$q_{k,n,s} = \mathbf{h}_k \mathbf{w}_{k,n,s} \tag{3.20}$$

and because this variable is not independent of $\mathbf{w}_{k,n,s}$, we add (3.20) as a constraint. We then get from (3.18–3.19) the equivalent problem

$$\max_{\mathbf{w}_{k,n,s}, q_{k,n,s}} c'_k \log_2 \left(1 + |q_{k,n,s}|^2\right) - \lambda \|\mathbf{w}_{k,n,s}\|^2 \tag{3.21}$$

$$\mathbf{h}_k \mathbf{w}_{k,n,s} = q_{k,n,s} \tag{3.22}$$

$$\mathbf{H}_{k,n,s} \mathbf{w}_{k,n,s} = 0. \tag{3.23}$$

Constraints (3.22) and (3.23) can then be rewritten in the standard form $\mathbf{G}_{k,n,s} \mathbf{w}_{k,n} = \mathbf{b}_{k,n,s}$ where the $\mathbf{G}_{k,n,s}$ matrix is the concatenation of $\mathbf{h}_k$ and $\mathbf{H}_{k,n,s}$ and $\mathbf{b}_{k,n,s} = [q_{k,n,s}, 0, 0 \ldots 0]^T$.

Since we are proposing to transform the constrained optimization over the κ variables into an unconstrained optimization over $q_{k,n,s}$ only, we must be able to express $\mathbf{w}_{k,n,s}$ as a function of $q_{k,n,s}$. The linear system being under-determined, this is obviously not unique. We propose to use $\mathbf{G}^+_{k,n,s}$, the pseudo-inverse of $\mathbf{G}_{k,n,s}$, for the back-transformation $\mathbf{w}_{k,n,s} = \mathbf{G}^+_{k,n,s} \mathbf{b}_{k,n,s}$. The pseudo-inverse picks the vector of minimum norm compatible with the linear system. In other words, choosing this transformation will *minimize* $\|\mathbf{w}_{k,n,s}\|$ so that it is minimizing the power term in the objective function in (3.21). Because $\lambda \geq 0$, this has the effect of contributing to the maximization of $f^*_{k,n,s}$. Note that this technique provides only an approximate solution of the beamforming problem; we cannot invoke Theorem 1 from [37]

which shows that in certain cases, the pseudo-inverse transformation is optimal. A strong assumption for the theorem is that the objective function depends only on the $q_{k,n,s}$ variable, which is not the case here since (3.21) also depends on $\|\mathbf{w}_{k,n,s}\|^2$. However, we observed from numerical results that the difference between the pseudo-inverse solution and the optimal solution is not significant. With this approximation we fix the direction of the beamforming vectors to

$$\begin{aligned}\mathbf{w}_{k,n,s} =& \mathbf{G}^{+}_{k,n,s}\mathbf{b}_{k,n,s} \\ =& q_{k,n,s}[\mathbf{G}^{+}_{k,n,s}]_1\end{aligned}$$

where $[\mathbf{G}^{+}_{k,n,s}]_1$ denotes the first column of $\mathbf{G}^{+}_{k,n,s}$. Now, we can obtain a problem formulation in terms of the user powers only by replacing the following expression in (3.21):

$$\|\mathbf{w}_{k,n,s}\|^2 = \gamma^2_{k,n,s} p_{k,n,s}, \tag{3.24}$$

where $\gamma_{k,n,s} = \|[\mathbf{G}^{+}_{k,n,s}]_1\|$ and $p_{k,n,s} = |q_{k,n,s}|^2$. Adding the constraint $p_{k,n,s} \geq 0$, we get the equivalent problem

$$\max_{p_{k,n,s}} c'_k \log(1 + p_{k,n,s}) - \lambda \gamma^2_{k,n,s} p_{k,n,s} \tag{3.25}$$

$$p_{k,n,s} \geq 0 \tag{3.26}$$

which has the well-known water-filling solution

$$p_{k,n,s} = \max\left\{0, \frac{c'_k}{\lambda \gamma^2_{k,n,s}} - 1\right\} \tag{3.27}$$

so that the computation time is basically the evaluation of $\mathbf{G}^{+}_{k,n,s}$. Also note that using $\mathbf{G}^{+}_{k,n,s}$ we can also find the optimal beamforming vectors for all users in s, the only difference being that $\gamma_{k,n,s}$ is computed using the column of $\mathbf{G}^{+}_{k,n,s}$ corresponding to the channel vector of this user.

3.2.4 Solving the Dual Problem

To summarize, the dual function $\Theta(\lambda, \boldsymbol{\mu})$ is obtained for the current values of the multipliers by finding for each subcarrier $n = 1, \ldots, N$ the optimal SDMA set $s^*(n)$ to the

minimization problem in (3.16), where

$$f_{n,s}(\lambda, \boldsymbol{\mu}) = -\sum_{k \in s} \left[c'_k \log\left(1 + p_{n,s,k}\right) - \lambda \gamma^2_{k,n,s} p_{k,n,s} \right] \tag{3.28}$$

and $p_{k,n,s}$ is given by (3.27). Substituting back in (3.9), the dual function is

$$\Theta(\lambda, \boldsymbol{\mu}) = -\lambda \check{P} + \sum_{k=1}^{K} \mu_k \check{d}_k + \sum_{n=1}^{N} \min_s f_{n,s}(\lambda, \boldsymbol{\mu}) \tag{3.29}$$

with $f_{n,s}$ given by (3.28). For any value of the dual variables $(\lambda, \boldsymbol{\mu})$ we can determine the primal variables $(\boldsymbol{\alpha}, \mathbf{w})$; $\boldsymbol{\alpha}$ is obtained by the optimal subcarrier assignment vector $s(n)$ after performing the minimization over s in (3.29), and the optimal beamforming vectors $\mathbf{w}^*_{n,k}$ for the users $k \in s^*(n)$ are given by

$$\mathbf{w}^*_{n,k} = \mathbf{G}^+_{k,n,s^*(n)} [p^{1/2}_{k,n,s^*(n)}, 0, \ldots 0]^T. \tag{3.30}$$

The largest part of the computation to evaluate the dual function is the calculation of $\mathbf{G}^+_{k,n,s}$ which has to be done for each subchannel and each possible SDMA set. The number of evaluations can become quite large but the size of each matrix is relatively small so that the calculation remains feasible for medium-size problems. Furthermore, while solving the dual problem requires multiple subgradient iterations, the calculation of the pseudo-inverses is *independent* of the value of the multipliers. This means that the calculation of $\mathbf{G}^+_{k,n,s}$ can be done only once in the initialization step of the subgradient procedure and not at each iteration.

Finally, algorithm 1 finds the optimal dual variables $(\lambda^*, \boldsymbol{\mu}^*)$ that solve the dual problem (3.10) using the subgradient method [36] with a fixed step size δ and provides an upper bound to problem (3.1–3.7), as discussed in subsection 3.2.5. Note that this algorithm can be used to solve the beamforming problem or equivalently the power allocation problem for a fixed SDMA set assignment. The only difference is that, in the latter case (3.16) becomes trivial since only the pre-assigned SDMA set per subcarrier needs to be considered.

3.2.5 Analysis of the Dual Function

Lets denote Θ^* the maximum of the dual function $\Theta(\lambda, \boldsymbol{\mu})$ over $(\lambda, \boldsymbol{\mu}) \geq 0$. If U^1 is the weighted sum-rate objective function achieved by any feasible point in the primal problem and U^* its optimum value, then the following inequalities hold [30]

$$U^1 \leq U^* \leq -\Theta^* \leq -\Theta(\lambda, \boldsymbol{\mu}) \tag{3.31}$$

Construct the set $\mathcal{S}$ of all subsets of users of size $1 \leq \kappa \leq M$
for all $n = 1 \dots N$ **do**
  for all $s \in \mathcal{S}$ **do**
    for all $k \in s$ **do**
      Compute the pseudo-inverse $G^{+}_{k,n,s}$ and $\gamma_{k,n,s}$
    end for
  end for
end for
Choose an initial value λ^0 and $\boldsymbol{\mu}^0$
Subgradient iterations. We set a limit of I_m on the number of iterations
for all $i = 1 \dots I_m$ **do**
  Solve the N subproblems (3.16) to compute the dual function
  Compute the subgradients:
  $g_{\mu}^{(k)} = \check{d}_k - \sum_n r_{k,n}$
  $g_{\lambda} = \sum_n \sum_{k \in s^*(n)} \|\mathbf{w}^*_{k,n}\|^2 - \check{P}$
  if $\|\boldsymbol{g}_{\mu}\| \leq \epsilon$ and $\|g_{\lambda}\| \leq \epsilon$ **then**
    Break
    Exit. A dual feasible solution has been found
  else
    Update the multipliers
    $\lambda^{i+1} = [\lambda^i + \delta g_{\lambda}]^+$
    $\boldsymbol{\mu}^{i+1} = [\boldsymbol{\mu}^i + \delta \boldsymbol{g}_{\mu}]^+$
  end if
end for
Exit. A dual feasible solution was NOT found.

Algorithm 1 Calculation of the Dual Problem

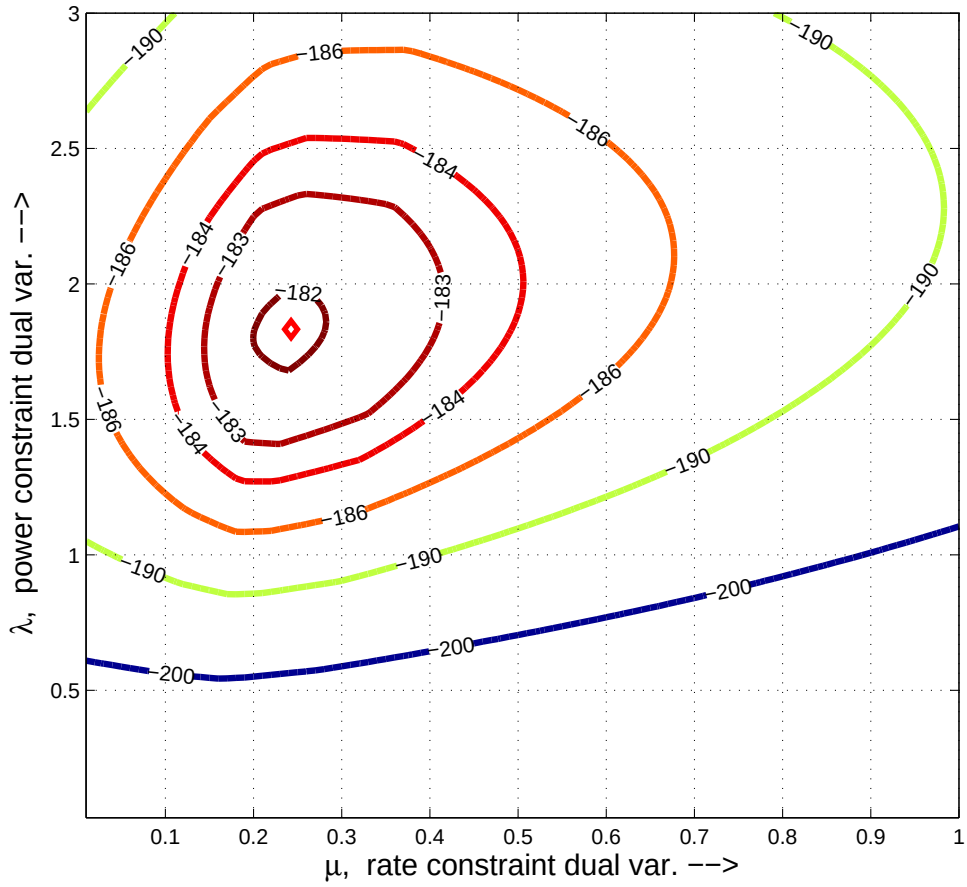


Figure 3.1 Contours of Dual Function, Single RT User. Parameters $N = 8, K = 8, M = 3, \check{P} = 20$, $\check{d}_1 = 50$ bps/Hz

The value $-\Theta^*$, or any feasible approximation $-\Theta(\lambda, \boldsymbol{\mu})$ to it, is thus an *upper bound* to the optimum value of the primal problem U^*.

Figure 3.1 shows a contour plot of the dual function $\Theta(\lambda, \boldsymbol{\mu})$ for the case of one RT user and parameters in the figure title. The diamond marker shows the maximum.

We can get some insight on the shape of the dual function from figure 3.2 where we plotted the function with respect to μ for a fixed value of λ. The solid line curve corresponds to the same dual function as in figure 3.1, where the rate constraint is active, $\check{d}_1 = 50$ bps/Hz. We see that the dual function goes through a maximum at $\mu = 0.24$. We have also shown the case where we increase the minimum rate constraint so much that the problem becomes infeasible, e.g., we make $\check{d}_1 = 100$ bps/Hz. As expected from duality theory, the dual function has no maximum since $\lim_{\mu \to \infty} \Theta(\lambda, \boldsymbol{\mu}) = \infty$ as shown by the dash-dotted curve. Finally, the dashed

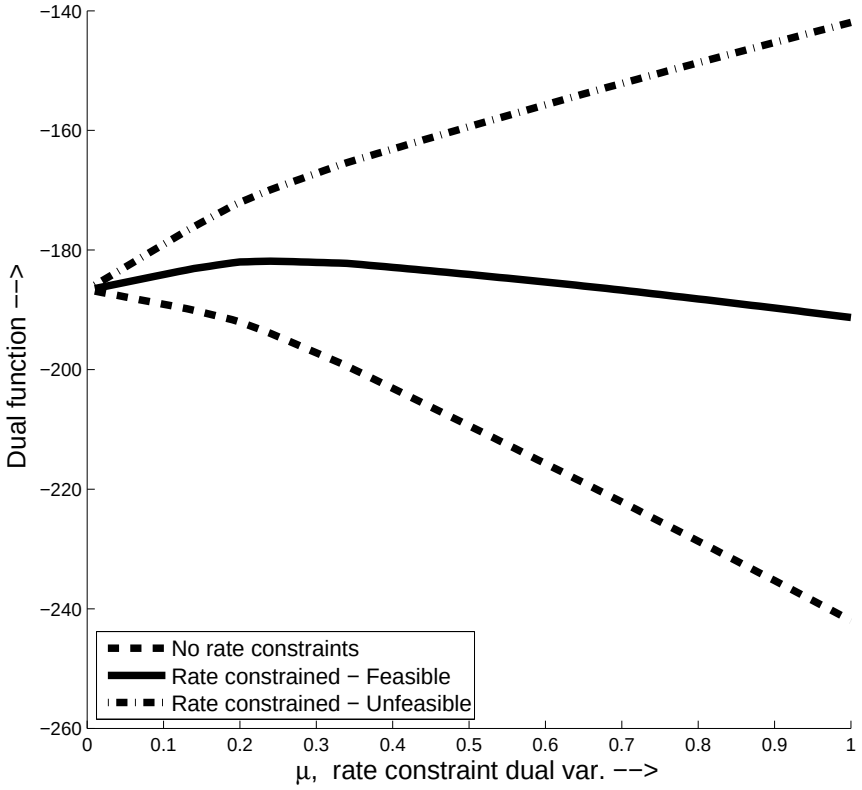


Figure 3.2 Dual Functions for Different Rate Constraints, $\lambda = 1.83$

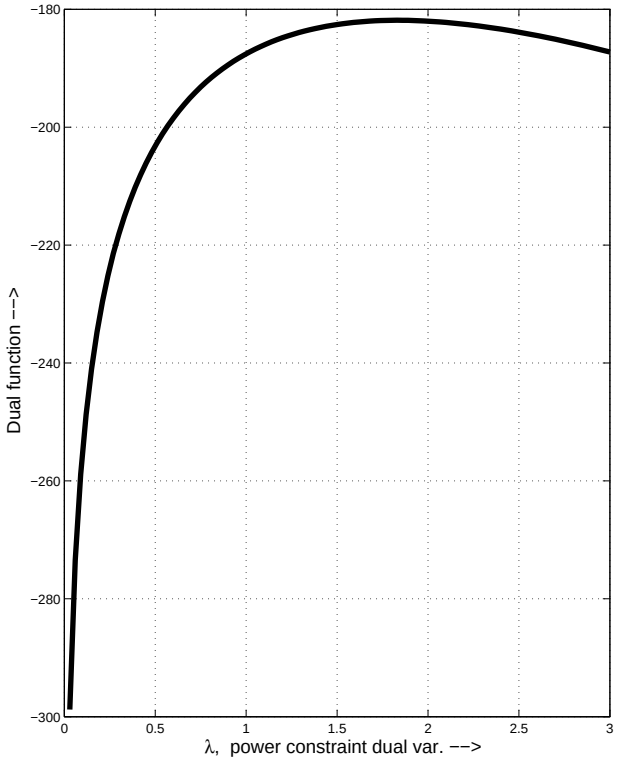


Figure 3.3 Dual Function vs. λ, $\mu = 0.24$

curve at the bottom corresponds to $\check{d}_1 = 0$ bps/Hz such that the constraint is inactive and the solution where the maximum occurs is located at $\mu_1 = 0$.

For completeness, we show in Figure 3.3 the dual function as a function of λ for the rate constrained feasible case. The dual function increases rapidly and reaches a maximum at $\lambda = 1.83$.

3.2.6 Dual-Based Primal Feasible Method

The SDMA set selection and beamforming vectors found by Algorithm 1 do not always provide a primal feasible solution. The rate or power constraints might be violated whenever the algorithm stops because the number of iterations has been reached before the convergence rule is met. In this subsection we propose a simple procedure to obtain a feasible point to problem (3.1–3.7) from the dual solution found with Algorithm 1. This point is not optimal, but because we start from the dual optimal solution, we expect that it will be near the primal

optimal solution. Obviously, this will give us a value U^1 which is a lower bound to the optimal primal solution (c.f. Eq. (3.31)).

Solve the dual problem (3.10) using algorithm 1. This yields the optimal dual variables $\lambda^*, \boldsymbol{\mu}_k^*$ and a SDMA set assignment vector $s^*(n)$ for each subchannel n.
Set $s_0^o(n) = s^*(n)$
Evaluate total power and user rate constraints (3.2–3.3)
if All constraints are met **then**
  Exit. A feasible solution has been found.
end if
Compute power allocation problem for $s_0^o(n)$ and evaluate total power and user rate constraints (3.2–3.3)
if All constraints are met **then**
  Exit. A feasible solution has been found.
end if
Compute the multipliers μ_k for users k such that $r_k < \check{d}_k$
for $j = 1$ to $\overline{J}$ **do**
  $\mu_k = \mu_k + \delta$
  Find $s_j^o = \arg\ \min_s\{f_{n,s}\}$ where $f_{n,s}$ is given by (3.17) for the current dual variables $\lambda, \boldsymbol{\mu}$
  Let $s_j^o(n)$ be the SDMA assignment found
  if $s_j^o(n) \neq s_{j-1}^o(n)$ **then**
    We have found a new SDMA assignment
    Compute power allocation problem for $s_j^o(n)$ and evaluate total power and user rate constraints (3.2–3.3)
    if All constraints are met **then**
      Exit. A feasible solution has been found.
    end if
  end if
end for
Exit. A feasible solution was not found.

Algorithm 2 Calculating a Feasible Point from the Dual Solution

---

Algorithm 2 summarizes this method. The algorithm begins by solving the dual problem (3.10) using Algorithm 1. If the solution is not feasible either directly or by recomputing the power allocation for the SDMA set assignment found in the dual problem, the algorithm performs a search by increasing the dual variables associated to the users whose QoS constraints are not met until a new SDMA set assignment is found. It then solves the power allocation problem for this new SDMA set assignment and checks the solution feasibility with regards to the minimum rate constraints. The search for new SDMA sets continues using this

method until a feasible SDMA set assignment is found or a maximum number of iterations is reached.

In contrast to the method described in section 2.4.1, which performs an enumeration of all possible SDMA set assignments, the dual-based algorithm 2 is a method that finds new candidate SDMA set assignments close to the dual optimal and then uses them to solve a simple power allocation problem until the rate and power constraints are met. This makes the search for a near-optimal feasible point much faster than finding the exact solution.

3.3 Performance Analysis

In this section, we present some numerical results to study the performance of the dual-based algorithm and the accuracy of the upper and lower bounds. To show how they can be used to evaluate heuristic algorithms, we also compare those bounds with the solution provided by a weight adjustment method we describe in section 3.3.2.

3.3.1 Convergence of the Dual Algorithm

To show the convergence properties of the upper bound computation, we first present in Figure 3.4 the value of the dual function and Lagrange multipliers as a function of the number of iterations for a given channel realization. The corresponding transmit power and the rate received by the RT user are shown in Figure 3.5. The parameters used for the calculation are listed in the figure titles. We see that the algorithm converges very quickly to a solution that is both close to the minimum value and feasible. This is typical of several other configurations, except that the number of iteration increases with the number of RT users.

3.3.2 Weight Adjustment Heuristic

Several RA algorithms provide support for users with RT traffic by increasing the user weights in the utility function until they receive enough transmission resources [19, 20]. In this section, we describe a generic weight adjustment method which will be used to show that this technique leaves much room for improvement.

In the weight adjustment method, we want to find a set of weights in the utility function (3.1) such that the rate requirements of the RT users are met when we solve problem (3.1–3.7) without the rate constraints (3.3). Also, the set of weights must not be very different among users to maximize the multi-user diversity gain. Algorithm 3 implements a generic method for weight adjustment that aims to do this. It increases the user weights for RT users until enough resources are allocated to meet the minimum rate requirements. The

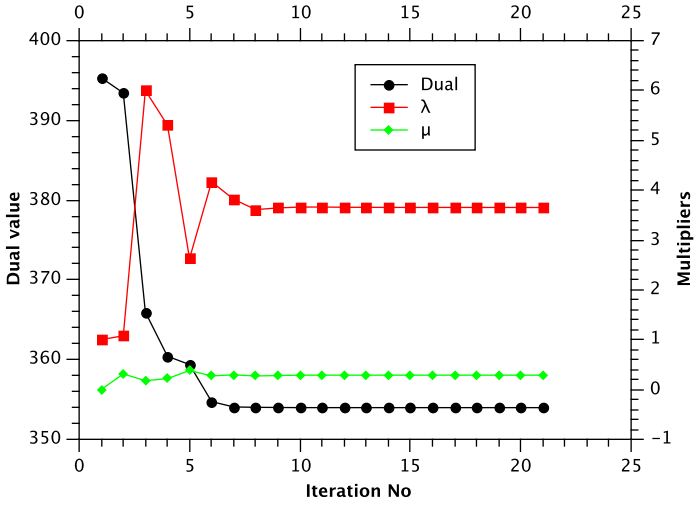


Figure 3.4 Dual function and multipliers for $M = 3$, $K = 16$, $N = 16$, $\check{P} = 20$, $D = 1$ and $\check{d}_1$ =80 bps/Hz.

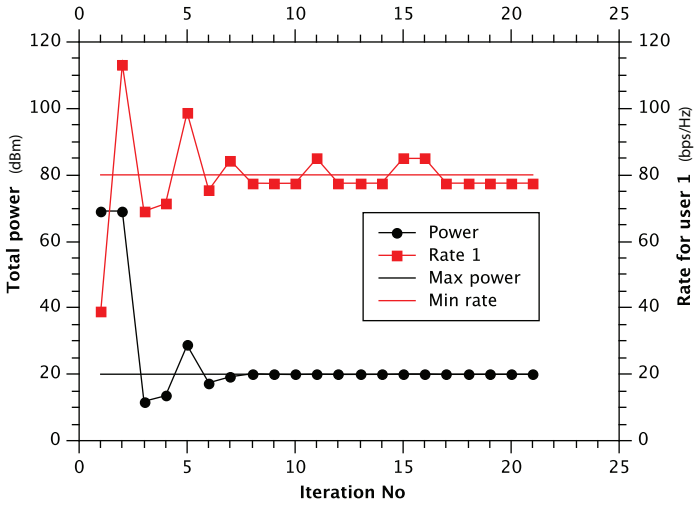


Figure 3.5 Power and rate constraints for $M = 3$, $K = 16$, $N = 16$, $\check{P} = 20$ dBm, $D = 1$ and $\check{d}_1$ =80 bps/Hz.

Solve RA problem (3.1–3.7) without minimum rate constraints constraints (3.3)
$\mathbf{c}' \leftarrow \mathbf{c}$
Let r_k be the achieved rate for user k at every iteration
iteration $\leftarrow 1$
while ($r_k < \check{d}_k$ for one or more users $k \in \mathcal{D}$) AND (iteration $\leq \bar{I}$) **do**
  Increase user weight using $c'_k = c'_k + \epsilon\left(\check{d}_k - r_k\right)$ for users in need, where $0 < \epsilon \leq 1$
  Solve RA problem (3.1–3.7) without minimum rate constraints (3.3) using user weights $\mathbf{c}'$
  iteration $\leftarrow$ iteration $+1$
end while

Algorithm 3 Weight Adjustment Algorithm

parameter ϵ controls how much the weights are increased with respect to the rate bounds. The rates achieved by Algorithms 3 and 2 are different since they solve different problems. Algorithm 3 can be seen as solving problem (3.1–3.7) by using a linear penalty method for constraints (3.3) of the form

$$P_k = \min\left\{0, r_k - \check{d}_k\right\}.$$

The modified objective function is then

$$\begin{aligned} U_P &= \sum_k c_k r_k + P_k \\ &= \sum_k c_k r_k + \epsilon \sum_{k|r_k<\check{d}} (r_k - \check{d}). \end{aligned} \tag{3.32}$$

At each iteration of the penalty method, whenever rate constraints are active, the solution of (3.32) cannot be smaller than that of (3.1–3.7) since it is a relaxation. Notice that problem (3.32) is quite simple since it has a single constraint for the transmit power but it has to be solved many times to adjust the weights of the real time users. In weight adjustment algorithms such as [20], the user weights are increased at each time slot using an increasing function of the packets delay, so the computation task is distributed over time. However, this distributed approach does not guarantee that the rate requirements are met in a given time slot which can lead to delay violations and jitter.

3.3.3 Parameter Setup and Methodology

We now present the method and parameter values used to compare the performance of the different methods to solve problem (3.1–3.7). We used a Rayleigh fading model to generate the user channels such that each component of the channel vectors $\mathbf{h}_{k,n}$ are i.i.d. random variables

Table 3.1 Average performance gap against the dual optimal upper bound

Method	Minimum rate (bps/Hz)		
	13.33	16.66	20
Dual-based upper bound (bps/Hz)	49.13	47.12	40.8
Primal enum. gap (%)	0.57	0.55	0.10
Dual-based feas. gap (%)	0.57	0.59	0.04
Weight mod. gap (%)	0.68	0.71	0.15

distributed as $\mathcal{CN}(0,1)$. We also assumed independent fading between users, antennas and subcarriers. Unless otherwise noted, we used a configuration with $M = 3$ antennas, $K = 16$ users, $N = 16$ subcarriers, and one RT user. The minimum rate constraint was set at 40 bps/Hz unless otherwise stated. We also fixed the power constraint to $\check{P} = 20$ and used a large-scale attenuation of 0 dB for all users. The user weights in (3.1) were set to $c_k = 1$ for all users. The results are the average over the feasible cases from 100 independent channel realizations.

We compared the performance of the different methods for various scenarios where we increased the resource requirements for the RT users until the minimum rate requirements can no longer be met for all RT users. For each scenario and channel realization, the upper bound was computed from the dual solution using Algorithm 1 described in Section 3.2.4. For small systems, we also found the exact solution using the primal enumeration method given in Section 2.4.1. We also computed the lower bound given by dual-based primal feasible Algorithm 2 and the heuristic solution provided by the weight adjustment Algorithm 3 described in Section 3.2.6 and 3.3.2, respectively. We used the upper bound given by the dual optimal solution as the reference point when computing the gap when the exact solution is not available.

3.3.4 Single User, Increasing Minimum Rate

In this first scenario, we have a single RT user and we increase its minimum rate $\check{d}_1$. First we consider a small system with $K = 4$ users and $N = 2$ subcarriers where it is possible to compute a primal solution using an enumeration method over the binary variable $\boldsymbol{\alpha}$. We present in Table 3.1 the average gap in percent between the three methods used to find feasible solutions against the dual upper bound for a small system configuration. As the required minimum rate increases from 13.33 to 20 bps/Hz, the upper bound decreases as more resources need to be assigned to the RT user until the problem is no longer feasible. For this small configuration, we see that all methods give excellent results and the duality gap is very small.

Table 3.2 Average total rate gap as a function minimum rate requirement

Method	Minimum rate (bps/Hz)		
	80	100	120
Total rate gap against the upper bound (%)			
Dual-based feas.	0.24	0.23	0.21
Weight mod.	9.49	7.30	3.36

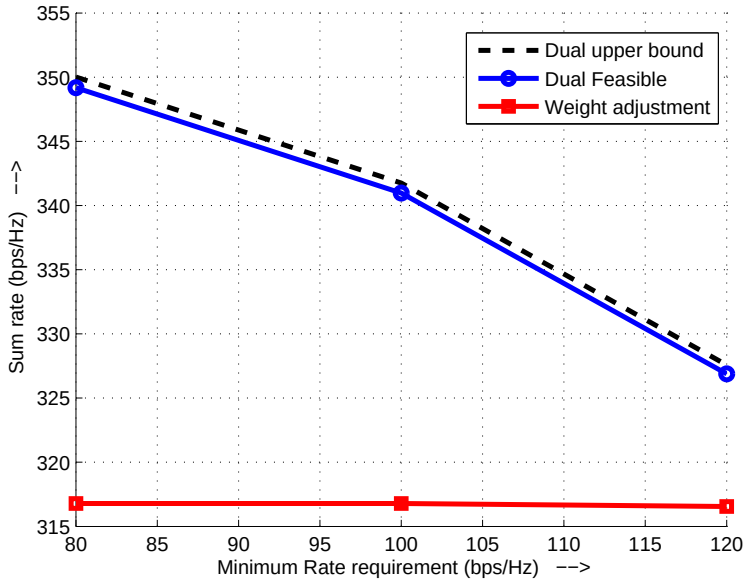


Figure 3.6 Average total rate as a function of the minimum rate requirements

Table 3.3 Average total rate gap as a function of RT user large-scale channel attenuation

Method	RT user attenuation (dB)		
	0	5	10
Total rate gap against the upper bound (%)			
Dual-based feas.	0.16	0.70	0.82
Weight mod.	9.53	32.95	52.35

In the remaining results, we use a larger system with $K = 16$ users and $N = 16$ subcarriers. With these values, it is no longer computationally feasible to find a primal solution using the enumeration method. We present in Table 3.2 the difference in percentage between the upper bound and the solutions of the dual-based feasible and the weight adjustment algorithms. The dual-based feasible algorithm provides a lower bound solution within 0.25% of the dual upper bound, the primal solution lies inside this small interval. On the other hand, the weight adjustment solution difference against the upper bound can be almost 10%. As discussed in Section 3.3.2, this is due to the fact that the weight adjustment algorithm stops as soon as it finds a feasible solution and does not have the option of finding a better assignment. As a result, the objective does not change much when the minimum rate is increased. This can be seen in figure 3.6 which shows the sum rate achieved by the dual-based feasible algorithm and the weight modification method against the minimum rate requirement.

3.3.5 Single User, Increasing Attenuation

Figure 3.7 shows the average total rate when the large-scale channel attenuation of the RT user varies from 0 to 15 dB. As the user moves away from the BS and the channel attenuation increases, the RA algorithm dedicates more resources to the RT user until the problem is unfeasible. The results show that for all SNR, the dual-based lower bound provides a tight solution with the upper bound while the weight adjustment method shows a large performance gap. Table 3.3 shows the error in percentage between the objective and the upper bound. For an attenuation of 15 dB, neither method is able to find a feasible solution; the problem is feasible because the dual upper bound is around 140, but the algorithms cannot find a solution.

3.3.6 Increasing Number of RT Users

Finally, figure 3.8 shows the upper dual bound, the lower bound and the solution given by weight adjustment methods as a function of the number of RT users. Table 3.4 lists the performance gap against the dual bound in percentage. The dual feasible lower bound is again

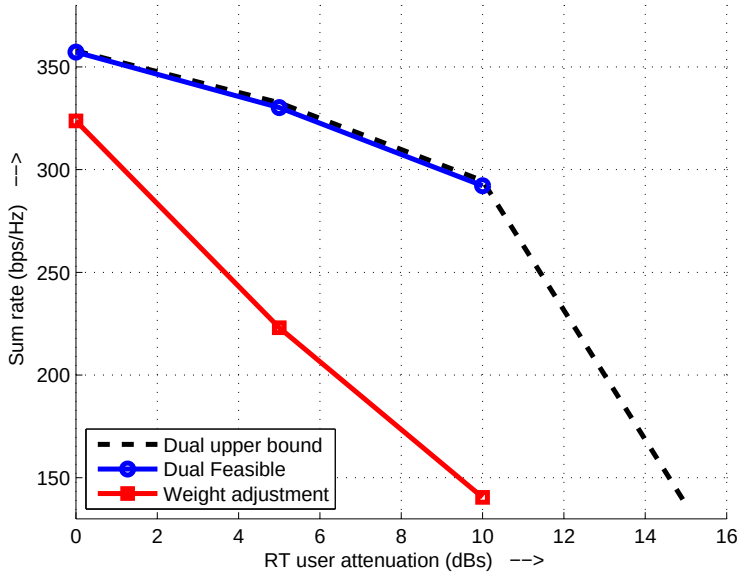


Figure 3.7 Average total rate as a function of RT user large-scale channel attenuation.

Table 3.4 Average total rate gap as a function of the number of RT users

Method	Number of RT users						
	1	2	3	4	5	6	7
	Total rate gap against the upper bound (%)						
Dual Feas.	0.16	0.61	2.09	2.41	3.52	3.20	3.43
Weight mod.	3.5	3.5	6.52	13.86	22.71	-	-

very close to the upper bound. Meanwhile, we can see that the performance of the weight adjustment method quickly degrades when the number of RT users increases. It cannot find feasible points when the number of RT users is 6 or 7 while the dual-based feasible algorithm yields solutions for these values within 3.52% of the upper bound.

For a single RT user, we have seen in tables 3.2 and 3.3 that the difference between the upper and lower solution is small. In figure 3.8, we see that this difference increases for three or more RT users. Still, this growth is not large and we can consider that a 3.52% is an acceptable error tolerance. Based on this, we can claim that it is possible to find a near-optimal solution to problem (3.1–3.7) with the proposed method, albeit with an off-line algorithm.

Furthermore, the results show that the weight modification method has a large performance gap which becomes more significant as the number of RT users increase. Also, the dual method can find feasible solutions for cases where the weight adjustment method can-

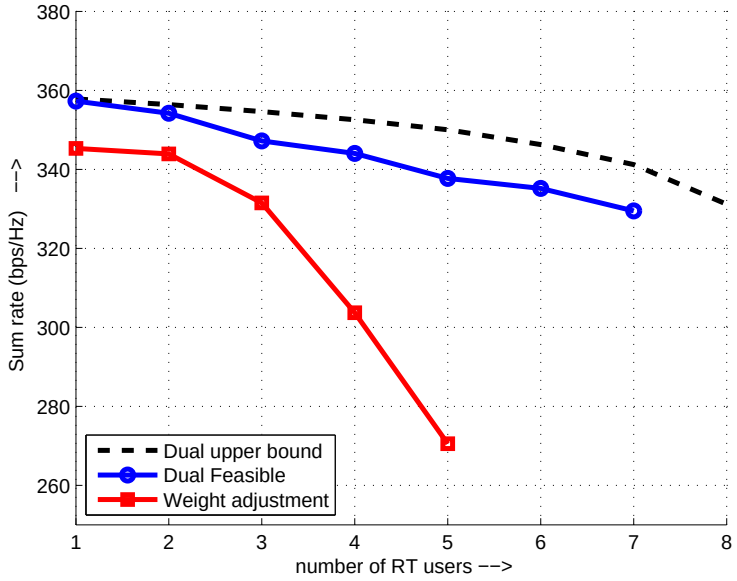


Figure 3.8 Average total rate as a function of the number of RT users.

not. This shows that the weight modification method should be used carefully for RA in OFDMA-SDMA systems with RT users and that more efficient heuristics should be developed to approach the performance of the dual-based feasible solution.

3.4 Chapter Conclusion

In this chapter, we proposed a method to compute the beamforming vectors and the user selection in an OFDMA-SDMA MISO system with minimum rate requirements for some RT users. We used a Lagrangian relaxation of the power and rate constraints to solve the dual problem using a subgradient algorithm. The Lagrange decomposition yields subproblems separated per subcarrier, SDMA sets and users which substantially reduces the computational complexity. We obtained a simple expression of the dual function for the beamforming problem for a given SDMA set based on a pseudo-inverse condition on the beamforming vectors. The dual optimum can then be used as a benchmark to compare against any other solution methods and heuristics. The dual function also gives us a better understanding of the problem. Its shape is related to the rate constraint activation and problem feasibility, and it also justifies the splitting of the subcarrier assignment and power allocation processes used in several heuristic methods.

We then proposed an algorithm which finds a feasible point by starting from the dual-based optimal solution and searches around the dual variables of the rate constraints. Nu-

merical results indicate that the two bounds are close. These upper and lower bounds provide a very useful benchmark to compare the performance of any heuristic method.

As a point of comparison, we also evaluated the performance of a weight adjustment method which adjusts the user weights in the objective function to achieve the required rates. Our results show that the performance gap of this approach is large and grows when the SNR of a single RT user decreases or when the number of RT users increases.

In addition, the weight adjustment method requires many time slots to adjust the weights and schedule real time users. The dual-based method explicitly includes the minimum rate constraints which allows RT users to be scheduled in the current slot decreasing the overall packet delay and jitter.

The significant gap between the weight adjustment algorithm and the optimal RA solution suggests that there is a need to find better heuristics. The dual approach looks promising to guide the design of efficient novel heuristics. However, to implement the RA algorithm in real time, we need to design fast heuristic methods that reduce the number of SDMA sets to be searched. In the next chapter, we design such heuristic algorithms.

CHAPTER 4

EFFICIENT HEURISTIC METHODS

4.1 Introduction

The dual-based algorithm proposed in chapter 3 requires a search over all SDMA sets while performing the subgradient iterations to solve the dual problem. The size of this search grows as K^M, where K is the number of users and M the number of antennas, and the algorithm requires the pseudo-inverse computation of all SDMA sets per subcarrier at initialization. These are the main causes of the algorithm's high computational complexity.

In this chapter, we propose heuristic methods to reduce this search size and solve the problem more efficiently. We are interested in feasible solutions, i.e., points that satisfy the rate and power constraints, and that are not too far from the optimal solution. In the dual-based algorithm 1 proposed in chapter 3, power allocation and subcarrier assignment are jointly performed; the dual variables determine the user power allocation and the subcarrier assignment through equations (3.16) and (3.27). If the problem is feasible, at the end of the subgradient iterations we obtain the near-optimal user power allocation and subcarrier assignment given by the optimal dual variables. Except for some trivial cases, we cannot separate the subcarrier allocation and power allocation processes. For heuristic methods, however, we separate these processes in order to reduce computational complexity. In the first stage, we find a subcarrier assignment that has enough subcarriers assigned to the real-time (RT) users, and in the second stage, we allocate power among users using the fixed subcarrier assignment. This approach has been used in [12] for the RA problem without RT minimum rate constraints, and in [22] where they are considered. We also follow this approach in this chapter.

For the subcarrier assignment stage, we make use of the well known Semiorthogonal User Selection (SUS) algorithm [28] to select user channels that have high norms and are semiorthogonal to each other. But contrary to the throughput maximization case, we include the RT users to satisfy their minimum rates when selecting the user set for each subcarrier. For the power allocation stage, we propose a method that finds feasible points and is much quicker than solving the complete power optimization problem. The subcarrier assignment algorithm and the power allocation algorithm constitute the proposed heuristic method.

We evaluate three key aspects of the proposed method: the performance gap against the dual upper bound found in the previous chapter, the range of the supported minimum

rates, and the method's computational complexity. In summary, the numerical results and theoretical analysis done in this chapter show that:

1. The gap between the objective achieved by the heuristics and the upper bound is not large. For example, in our experiments this gap is 10.7% averaging over all performed numerical evaluations for all system configurations.
2. The proposed algorithm increases the range of the supported minimum rates when compared with the method proposed in [22]. For the same case above, the increase in the rate range is 14.6% on average. This increase is achieved by considering the rate constraint dual variables in the user power allocation stage.
3. The heuristics have significantly lower computational complexity than the method proposed in [22]. The computational complexity reduction is several orders of magnitude depending on the algorithm used and the problem parameters.

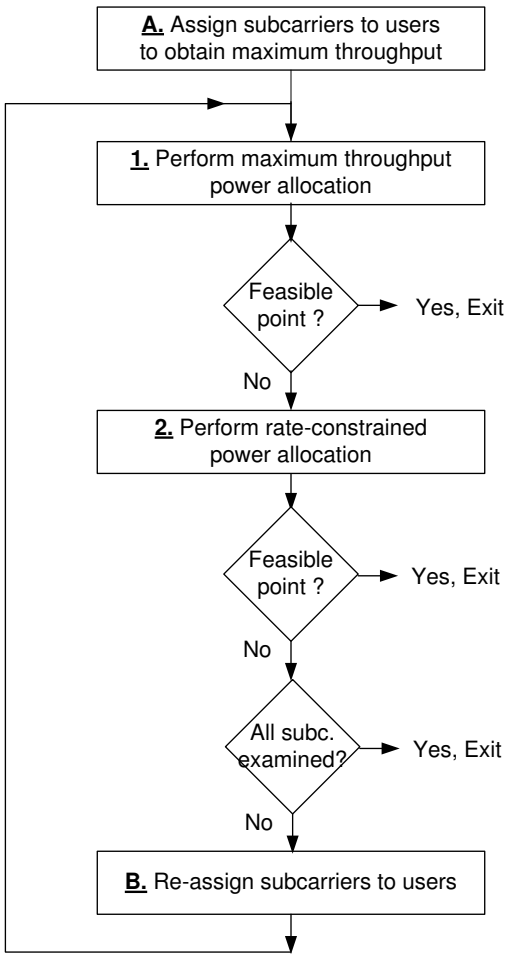
The augmented supported minimum rates and the reduced computational complexity are important characteristics of the proposed algorithm. They provide the rate requirements that real-time applications need and allow a practical implementation. The objective of this chapter is to present and examine the proposed heuristic method. To guide the reader through the several algorithms that constitute the heuristic method, and that are discussed in this chapter, we give a general description of the proposed method in the following subsection.

4.1.1 General Description of the Proposed Heuristic Method

We start by solving a maximum throughput problem. Let's name $r_k^{(a)}$ the rates obtained when solving problem (3.1–3.7) without considering rate constraints (3.3). If the required rates $\check{d}_k$ are lower or equal than the obtained rates $r_k^{(a)}$, we have an optimal solution and the algorithm finishes. To obtain the maximum throughput solution efficiently, we use a heuristic method to assign subcarriers to users and then perform *maximum throughput power allocation*, which consists of finding the user power allocation that satisfies the power constraint with equality disregarding the rate constraints. These correspond to the first two blocks in the diagram of figure 4.1 – maximum throughput subcarrier assignment corresponds to block A and power allocation to block 1. If the required rates are met, we exit, otherwise, we need to assign more resources to the users in need.

Starting from the maximum throughput solution, there are two mechanisms to reassign resources to users. The first mechanism — *subcarrier reassignment* — takes away subcarriers assigned to users that do not require them, because they are not RT users or they have more resources than needed, and assigns them to the users in need. The second mechanism — *rate-constrained power allocation* — takes into account the user rate constraints to reallocate

Figure 4.1 Heuristic general algorithm



power between users. Subcarrier reassignment has a much larger effect because users in need are given subcarriers that they did not have before; the rates increase substantially with every subcarrier added. Rate-constrained power allocation has a lower effect because the rate increase dependency against power is logarithmic. However, this mechanism proves to be crucial in finding feasible points when the minimum rate requirements increase as we will see in section 4.2. In addition, recomputing the users power using the proposed rate-constrained PA is quicker than finding a new subcarrier and inverting the new channel matrix.

If the rate constraints are not met after maximum throughput power allocation, we perform rate-constrained power allocation as indicated by block 2 in figure 4.1.

We perform subcarrier *re*-assignment when the maximum throughput subcarrier assignment plus rate-constrained power allocation does not support the required minimum rates. A heuristic method that groups semiorthogonal user vectors is used to assign more subcarriers to the users in need as indicated by block B in figure 4.1. Then, maximum throughput power allocation is performed and, if the minimum rates are not met, rate-constrained power allocation is performed. If the minimum rates are still not met, we have to assign more subcarriers to the users in need. Therefore, in the block diagram of figure 4.1, we perform iterations adding subcarriers to users in need (block B) and performing power allocation (blocks 1 and 2) until the user minimum rates are met or there are no more subcarriers to reassign and the problem is declared unfeasible by the heuristic.

4.1.2 Chapter Description

In section 4.2, we reformulate the original problem (3.1–3.7) for the case of fixed subcarrier assignment. We present algorithms 4 and 5 to perform power allocation. These correspond to blocks 1 and 2 in figure 4.1. The first algorithm performs exact power allocation to maximize throughput, while the second considers the rate constraints and finds a feasible point very efficiently. In section 4.3, we present the heuristic method for subcarrier assignment. We analyze the main approaches that have been followed for the maximum throughput case. The proposed method is based on the well known SUS algorithm [28] , which we use to perform maximum throughput subcarrier assignment and corresponds to block A in figure 4.1. We then propose an algorithm for subcarrier *re*-assignment in section 4.3 which invokes the other algorithms presented in this chapter. This corresponds to block B in figure 4.1. The performance of the proposed algorithm is studied in sections 4.5 and 4.6. Finally, we give the chapter conclusions in section 4.7.

4.1.3 Chapter Contribution

The key contributions of this chapter are

- An efficient power allocation heuristic method for the rate-constrained case — algorithm 5 in subsection 4.2.4. The method is used here for an OFDM-SDMA system, but it can also be used for power allocation problems in SISO and MIMO systems.
- Efficient subcarrier reassigning methods for the rate-constrained case — algorithms 9 and 10 in subsection 4.3.3. These algorithms group users based on the channel spatial characteristics and can be combined with other power allocation methods.
- An overall RA method that extends the supported minimum rates and has a computational complexity that is several orders of magnitude lower than existing methods.

4.2 Power Allocation for Fixed Subcarrier Assignment

The problem we deal in this section is to find the user power allocation for a fixed subcarrier assignment. Assume that we have chosen a vector $\boldsymbol{\alpha}^{(n)}$ for each subcarrier n satisfying Eqs. (3.4) and (3.7) [1]. The vector $\boldsymbol{\alpha}^{(n)}$ determines a fixed SDMA set of users, S_n defined as

$$S_n \doteq \{k \in \mathcal{K} : \alpha_{n,k} = 1\}, \tag{4.1}$$

$$g_n \doteq |S_n|, \quad \forall n. \tag{4.2}$$

Sets S_n contain the indexes of the users assigned to subcarrier n. We first reformulate problem (3.1–3.7) using these known sets. Then, we apply a dual method to solve it. For this purpose, we arrange the channel vectors of selected users in the rows of a $g_n \times M$ matrix

$$\mathbf{H}_n \doteq \begin{bmatrix} \mathbf{h}_{n,S_n(1)} \\ \vdots \\ \mathbf{h}_{n,S_n(g_n)} \end{bmatrix}, \quad \forall n, \tag{4.3}$$

where $S_n(j)$ is the j-th user in the set S_n. We also arrange the corresponding beamforming vectors in the columns of a $M \times g_n$ matrix for each subcarrier

$$\mathbf{W}_n \doteq \left[\mathbf{w}_{n,S_n(1)}, \ldots, \mathbf{w}_{n,S_n(g_n)}\right], \quad \forall n. \tag{4.4}$$

Then, the ZF constraints (3.5) can be written as

$$\mathbf{H}_n \mathbf{W}_n = \mathrm{diag}(\sqrt{\mathbf{q}^{(n)}}), \quad \forall n \tag{4.5}$$

1. We explain the heuristic method to obtain such a vector in section 4.3.

where $\mathbf{q}^{(n)} = \{q_{n,j}\}$ is the users power vector comprised of

$$q_{n,j} = \mathbf{h}_{n,S_n(j)}\mathbf{w}_{n,S_n(j)}, \quad j \in \{1,\ldots,g_n\} \tag{4.6}$$

Beamforming vectors for users k not belonging to S_n are set to zero. Restricting the direction of $\mathbf{W}_n$ to the pseudo-inverse of matrix $\mathbf{H}_n$ as done in section 3.2.3, we obtain from (4.5)

$$\mathbf{W}_n = \mathbf{H}_n^{\dagger}\,\mathrm{diag}(\sqrt{\mathbf{q}^{(n)}}), \quad \forall n \tag{4.7}$$

the power constraint can now be written as

$$\sum_{n=1}^{N} \mathrm{tr}\,(\mathbf{W}_n^H\mathbf{W}_n) - \check{P} \leq 0 \tag{4.8}$$

and replacing (4.7) in (4.8) we obtain

$$\sum_{n=1}^{N}\sum_{j=1}^{g_n} \left[(\mathbf{H}_n^{\dagger})^H\mathbf{H}_n^{\dagger}\right]_{j,j} q_{n,j} - \check{P} \leq 0 \tag{4.9}$$

Let's define the entries of the $N \times K$ matrices $\boldsymbol{\beta}$ and $\mathbf{p}$ as

$$\beta_{n,k} = \begin{cases} \left[(\mathbf{H}_n^{\dagger})^H\mathbf{H}_n^{\dagger}\right]_{j,j} & \text{if } k = S_n(j), \quad \forall j \in \{1,\ldots,g_n\} \\ 0, & \text{otherwise} \end{cases} \tag{4.10}$$

$$p_{n,k} = \begin{cases} q_{n,j} & \text{if } k = S_n(j), \quad \forall j \in \{1,\ldots,g_n\} \\ 0, & \text{otherwise.} \end{cases} \tag{4.11}$$

The power constraint can now be expressed as

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \beta_{n,k}p_{n,k} - \check{P} \leq 0. \tag{4.12}$$

From the original model (3.1–3.7), we do not need constraints (3.4) and (3.7) because we choose $\alpha_{n,k}$ satisfying these conditions. Constraints (3.5) and (3.6) are implicit in the reformulated model because the beamforming vectors satisfy (4.5) and we set to zero all beamforming vectors for which $\alpha_{n,k} = 0$. Therefore, only constraints (3.2) and (3.3) remain. We also changed the problem optimization variables from the vectors $\mathbf{w}_{n,k}$ to the scalars $p_{n,k}$ because the vector directions are now fixed by (4.7).

Replacing $(\mathbf{h}_{n,k}\mathbf{w}_{n,k})^2$ by $p_{k,n}$ in (3.1),(3.3) and replacing (3.2) by (4.12), we obtain the problem formulation

$$\max_{p_{n,k}} \quad \sum_{n=1}^{N}\sum_{k=1}^{K} c_k \log_2(1+p_{n,k}) \tag{4.13}$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \beta_{n,k}\ p_{n,k} - \check{P} \leq 0 \tag{4.14}$$

$$-\sum_{n=1}^{N} \log_2(1+p_{n,k}) + \check{d}_k \leq 0, \quad k \in \mathcal{D} \tag{4.15}$$

$$p_{n,k} \geq 0, \quad \forall n,k. \tag{4.16}$$

4.2.1 Optimal Power Allocation

Problem (4.13–4.15) is convex since we maximize a concave function over a convex set formed by constraints (4.14) and (4.15). We can solve this problem optimally using a dual approach. Our objective is to derive a closed-form expression of the dual function and solve the dual problem optimally. We choose the dual method to solve the problem instead of any other optimization technique, because later we design heuristic methods that operate on the dual domain.

Defining Lagrange multipliers $\theta > 0$ for the power constraint (4.14) and $\delta_{\mathbf{k}} \in \mathbb{R}_+^K$ for the rate constraints (4.15), we get the Lagrangian function

$$\mathcal{L}_2(\mathbf{p},\theta,\boldsymbol{\delta}) = -\theta\check{P} + \sum_{k=1}^{K}\delta_k\check{d}_k + \sum_{n=1}^{N}\sum_{k=1}^{K} -(c_k+\delta_k)\log_2(1+p_{n,k}) + \theta\beta_{n,k}p_{n,k}. \tag{4.17}$$

For convenience we have defined dual variables δ_k for all users including the ones with no minimum rate requirements ($k \notin \mathcal{D}$); for these users we set $\check{d}_k = 0$. Minimizing over the primal variables $p_{n,k}$ we obtain the dual function

$$\psi(\theta,\boldsymbol{\delta}) = \min_{p_{n,k}\geq 0} \mathcal{L}_2(\{p_{n,k}\},\theta,\boldsymbol{\delta}) \tag{4.18}$$

The minimization in (4.18) separates over subcarriers and users

$$\min_{p_{n,k}\geq 0} -(c_k+\delta_k)\log_2(1+p_{n,k}) + \theta\beta_{n,k}p_{n,k}, \quad \forall n,k \tag{4.19}$$

which yields the water-filling power allocation

$$\overline{p}_{n,k} = \left[\frac{c_k + \delta_k}{\theta\beta_{n,k}\ln 2} - 1\right]^+, \quad \forall n, \forall k : \beta_{n,k} \neq 0 \tag{4.20}$$

and rate allocation

$$r_{n,k} = \log_2\left(1 + \left[\frac{c_k + \delta_k}{\theta\beta_{n,k}\ln 2} - 1\right]^+\right), \quad \forall n, \forall k : \beta_{n,k} \neq 0 \tag{4.21}$$

Thus, the dual function is

$$\psi(\theta, \boldsymbol{\delta}) = -\theta\check{P} + \sum_{k=1}^{K}\delta_k\check{d}_k + \sum_{n=1}^{N}\sum_{k=1}^{K} -(c_k + \delta_k)\log_2(1 + \overline{p}_{n,k}) + \theta\beta_{n,k}\overline{p}_{n,k} \tag{4.22}$$

where $\boldsymbol{\delta}$ is the vector of dual variables δ_k for $k \in \mathcal{D}$. The shape of $\psi(\theta, \boldsymbol{\delta})$ depends on the constraint parameters $\check{P}, \check{d}_k$, the choice of subcarrier assignment vector $\boldsymbol{\alpha}^{(n)}$ and the current channel realization. For instance, if $\check{d}_k$ is large for a particular k, $\psi(\delta_k)$ will tend to ∞ and the primal becomes infeasible. On the contrary, if $\check{d}_k$ is very small, $\psi(\delta_k)$ always decreases and the maximum occurs at $\delta_k = 0$. For values of $\check{d}_k$ in between these two extrema, a maximum occurs at some $\delta_k > 0$. Also, if feasible the problem dual function presents a maximum w.r.t. θ, at $\theta > 0$.

The associated dual problem of (4.13–4.15) is

$$\max_{\theta>0, \boldsymbol{\delta}\geq 0} \psi(\theta, \boldsymbol{\delta}) \tag{4.23}$$

We can solve it using derivative-free techniques like the subgradient iterations used in section 3.2.4. The use of such methods involves two steps. In the first one, we compute a matrix pseudo-inverse per subcarrier to obtain the inverse of the channel effective gains $\beta_{n,k}$; this has computational complexity $O(NM^3)$. In the second step, we perform subgradient iterations computing the power and rate constraints to obtain the subgradient vector; this has a lower computational complexity. Then, the total computational complexity is $O(NM^3)$. We are interested, however, in approximate methods that produce a primal feasible point of problem (4.13–4.15) and that are more computationally efficient.

4.2.2 Efficient Power Allocation

To solve problem (4.13–4.15) more efficiently we separate it in two stages: maximum-throughput power allocation (PA) and rate-constrained PA. Maximum-throughput PA only

considers power constraint (4.14) making it easier to obtain a solution. After solving the problem, if the achieved rates are feasible they are the optimal ones. In subsection 4.2.3 we use an exact method for this purpose. Exact methods, contrary to iterative methods, find the solution by testing a hypothesis and — in our case — require less iterations. The solution found satisfies the power constraint with equality.

If the rate constraints are not met after maximum throughput PA, we incorporate the rate constraints (4.15). We describe a heuristic method to perform rate-constrained PA in subsection 4.2.4. The proposed method finds a feasible point and does not require any iterations. This feasible point satisfies the power constraint with equality but the rate constraints with inequality which is faster to compute. Maximum throughput and rate constrained power allocation correspond to blocks 1 and 2 in figure 4.1.

4.2.3 Maximum Throughput Power Allocation

This consists of solving problem (4.13–4.14) without rate constraints (4.15), we just need to set $\boldsymbol{\delta}$ to zero and solve (4.23) using subgradient iterations. This will give us an optimal power constraint dual variable θ^*. However, we can get a solution much more quickly by using the following observation

Lemma 1. *The solution point to problem (4.13–4.15), satisfies the power constraint (4.14) with equality.*

Proof. We prove this lemma by contradiction. Assume $\mathbf{p}^{(a)}$ is the optimal point of problem (4.13–4.15) and that it satisfies (4.14) with *strict* inequality attaining objective $U^{(a)}$. We can always find arbitrary $\Delta_{n,k} \geq 0$ that define a new point $\mathbf{p}^{(b)} = \{p_{n,k}^{(a)} + \Delta_{n,k}\}$, such as the point $\mathbf{p}^{(b)}$ satisfies the power constraint with equality, i.e.,

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \beta_{n,k}(p_{n,k}^{(a)} + \Delta_{n,k}) = \check{P}. \tag{4.24}$$

Point $\mathbf{p}^{(b)}$ attains objective $U^{(b)}$. The objective (4.13) is an increasing function of the powers and $p_{n,k}^{(b)} > p_{n,k}^{(a)}$ for some n, k. Therefore, $U^{(b)} > U^{(a)}$ and point $\mathbf{p}^{(a)}$ is *not* the optimal point.

This implies that for a point $\mathbf{p}$ to be optimal, it has to satisfy the power constraint with equality. □

Using lemma 1 we can obtain a solution to problem (4.13–4.14) just by finding the value of the dual variable θ that satisfies (4.14) with equality.

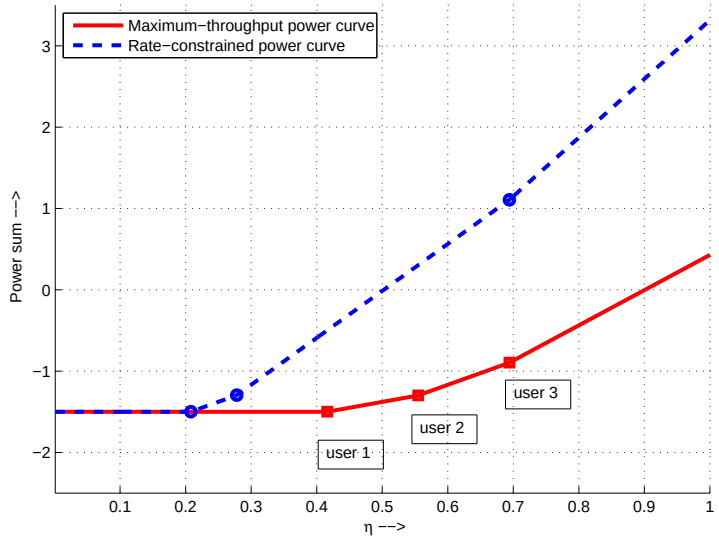


Figure 4.2 Water-filling power sum curves

We start by replacing the optimal powers given by (4.20) with $\delta_k = 0$ in the power constraint (4.14)

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\beta_{k,n}\left[\frac{c_k}{\theta\beta_{n,k}\ln 2}-1\right]^{+}=\check{P}. \tag{4.25}$$

In equation (4.25) only variable θ is unknown. To illustrate this root finding problem, we rewrite (4.25) using variable $\eta=\theta^{-1}$ and $c_k^0=c_k/\ln 2$

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\left[c_k^0\eta-\beta_{n,k}\right]^{+}-\check{P}=0. \tag{4.26}$$

Figure 4.2 illustrates one example for the case of one subcarrier and three users. The left-hand side of (4.26) is a increasing piecewise linear function in η. The function is plotted in solid line as a function of the dual variable η. Each user k becomes active at $\eta=\beta_k/c_k^0$; the function is continuous but not differentiable at these points. User 1 has the best channel (lowest β_k), so it becomes active first, then user 2 and 3. The curve between β_1/c_1^0 and β_2/c_2^0 has slope c_1^0, and between β_2/c_2^0, and β_3/c_3^0 has slope $c_1^0+c_2^0$. The slope increases as more users get active. The curve crosses zero at around $\eta=0.9$ which is the root we seek to find for this particular example.

We can find the root of (4.26) using a derivative-free numerical method. Instead, we devise an efficient method based on the modified water-filling algorithm in [38]. We define sets $\mathcal{B}_k$ containing the subcarriers that have been assigned to each user k by the subcarrier assignment

heuristic and that comply with the condition $(\beta_{n,k}\theta \ln 2) < c_k$, so that the allocated user powers are higher than zero. Let's $\mathcal{N}$ be the set of subcarriers $\{1, \ldots, N\}$, the sets $\mathcal{B}_k$ are thus defined by

$$\mathcal{B}_k(\theta) \doteq \left\{ n \in \mathcal{N} : (k \in S_n) \wedge \left(\beta_{n,k} < \frac{c_k}{\theta \ln 2} \right) \right\}, \quad \forall k \in \mathcal{K}, \tag{4.27}$$

where S_n is the SDMA set associated to subcarrier n. We compute the total power performing the sum over the subcarriers in sets $\mathcal{B}_k$ only. We re-write (4.25) as

$$\sum_{k=1}^{K} \sum_{n \in \mathcal{B}_k(\theta)} \beta_{k,n} \left(\frac{c_k}{\theta \beta_{n,k} \ln 2} - 1 \right) = \check{P}. \tag{4.28}$$

where θ is the power constraint dual variable we seek to compute. Notice that we have removed the function $[.]^+$ because definition (4.27) assures that user powers are higher than zero. To compute sets $\mathcal{B}_k$, we need to know the dual variable θ associated to the power constraint. But this is precisely what we want to find, so instead we use a lower bound $\check{\theta}$. This can be any value $\check{\theta} > 0$ (e.g. $\check{\theta} = 0.1$) that we can assure is lower than θ. Using this lower bound we compute sets $\mathcal{B}_k(\check{\theta})$ using (4.27). After some manipulation of (4.28), the power constraint dual variable is given by

$$\theta^{(i)} = \frac{\sum_{k=1}^{K} |\mathcal{B}_k(\check{\theta})| c_k}{(\check{P} + \sum_{k=1}^{K} \sum_{n \in \mathcal{B}_k(\check{\theta})} \beta_{n,k}) \ln 2} \tag{4.29}$$

where we use the index (i) to denote the current iteration. Then, we recompute sets $\mathcal{B}_k^{(i+1)}$ using this $\theta^{(i)}$. If the sets $\mathcal{B}_k^{(i)}$ and $\mathcal{B}_k^{(i+1)}$ are equal, we have found the solution, the power constraint is satisfied with equality. Otherwise, we iterate recomputing (4.27) and (4.29) until we find identical sets in two consecutive iterations. Algorithm 4 summarizes the steps to perform maximum throughput power allocation using this method. This algorithm is equivalent to the one reported in table I in [38], where removing the channels with negative energies there is equivalent to recomputing sets (4.27) here. In the same work, a variation of this algorithm (table II in [38]) that requires user ordering is proposed. The convergence to the optimum of such an algorithm is proved in [39] for a more general case.

The computational complexity of algorithm 4 is linear with K and $|\mathcal{N}|$, i.e., $O(K|\mathcal{N}|)$. Numerical results show that when applying this method after subcarrier assignment, we require very few iterations computing (4.27) and (4.29). This is because PA is performed after a user selection process that picks the users with good channel conditions, i.e. channels with high vector norms and semiorthogonal to each other. Therefore, the values of the

Input: Subcarrier set $\mathcal{N}$, Subcarrier assignment sets S_n, lower bound $\check{\theta}$, rate constraints dual variables $\{\delta_k\}$
Output: power constraints dual variable θ, user rates $\{r_k\}$
$i \leftarrow 1$; Solution $\leftarrow$ False
$\theta^{(i)} \leftarrow \check{\theta}$
- Compute sets $\mathcal{B}_k^{(i)}(\theta^{(i)})$ in (4.27) for all users and subcarriers in $\mathcal{N}$.
while not Solution **do**
  - Compute power constraint dual variable $\theta^{(i+1)}$ using $\mathcal{B}_k^{(i)}$ in (4.29).
  - Compute sets $\mathcal{B}_k^{(i+1)}(\theta^{(i+1)})$ in (4.27) for all users and subcarriers in $\mathcal{N}$.
  - Compute power constraint (4.28).
  if power constrained is satisfied **then**
    Solution $\leftarrow$ True
  end if
end while
- Compute rates using (4.21).

Algorithm 4 Maximum Throughput power allocation: $[\theta, \{r_k\}]$= Max_Troughput_power ($\mathcal{N}$, S_n, $\check{\theta}$, $\{\delta_k\}$)

effective channel gains are usually close to each other and the number of iterations required is small. This results in a low processing time. We compared the elapsed time needed by algorithm 4 when coded in a Matlab script and by the golden section method implemented in the Matlab function *fminbnd* [40]. The golden section method iteratively narrows the range of values of θ inside which the solution of (4.25) lies. The distance to the solution decreases with the number of iterations until it is negligible. In contrast with this derivative-free iterative method, the proposed algorithm 4 belongs to the hypothesis-based methods as the algorithms in [38, 39].

Figure 4.3-a) and b) show the elapsed time by algorithm 4 and by the golden section method respectively w.r.t. the number of users K. Both vary linearly, but the elapsed time of algorithm 4 is four orders of magnitude lower than that of the golden section method. We used the Matlab default values indicated in table 4.1 for the parameters in the golden section algorithm. For both algorithms, we bounded the optimization variable θ, to the interval $[0.1 - 3.0]$. On the other hand, figure 4.4-a) and b) show the elapsed time by algorithm 4 and the golden section method, respectively, w.r.t. the number of subchannels N. The elapsed time of algorithm 4 is also four orders of magnitude lower than the golden section method. In addition, this has a non-linear dependency with the number of subchannels. In conclusion for this subsection, algorithm 4 presents an efficient approach for solving the power allocation problem (4.13–4.14) optimally when no rate constraints are considered.

Table 4.1 Parameters used for the golden section method

Max. iterations	500
Max. function evaluations	500
Tolerance	10^{-4}

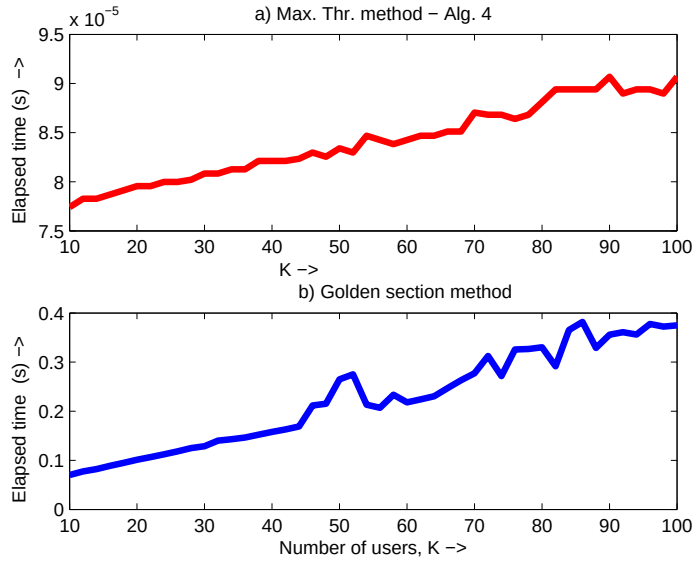


Figure 4.3 Algorithm 4 and golden section method CPU comparison, $N = 128, M = 8$.

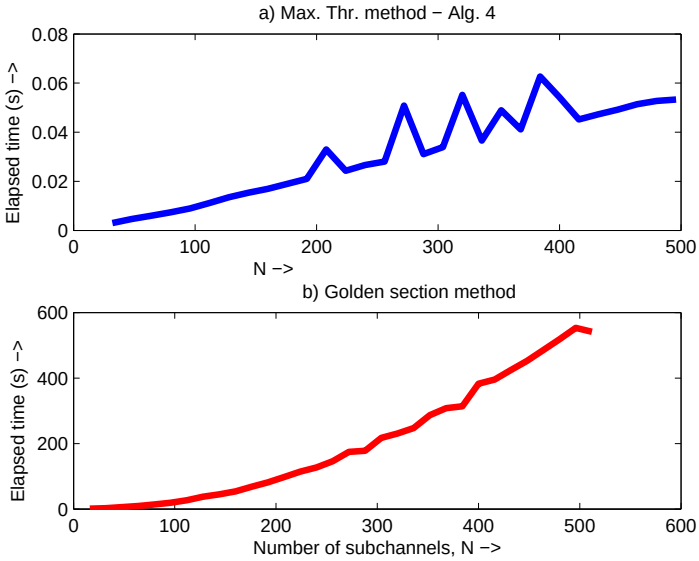


Figure 4.4 Algorithm 4 and golden section method CPU comparison, $K = 128, M = 8$.

4.2.4 Rate-constrained Power Allocation

If rates constraints are not met after power has been allocated following the method presented in subsection 4.2.3, we proceed to the second part of the heuristic where we take the rate constraints dual variables into consideration. This corresponds to the rate-constrained power allocation blocks marked as 2 in figure 4.1. This rate-constrained power allocation method should give dual variables $(\theta^{(2)}, \boldsymbol{\delta}^{(2)})$ that produce a feasible point $\{p_{n,k}^{(2)}\}$ that satisfies both the problem power and rate constraints (4.14) and (4.15).

We assume that the problem is feasible and that we have a maximum throughput solution to the dual problem given by dual variable $\theta^{(1)}$ and rates r_k which we obtained from algorithm 4 in subsection 4.2.3. This point satisfies the power but not the rate constraint for one or more users. We define the set of unsatisfied users $\mathcal{T}$ as

$$\mathcal{T} \doteq \{k \in \{1, \ldots, K\} : r_k < \check{d}_k\} \tag{4.30}$$

$$r_k \doteq \sum_{n=1}^{N} \log_2 \left(1 + \left[\frac{c_k}{\theta^{(1)} \beta_{n,k} \ln 2} - 1\right]^+\right) \tag{4.31}$$

We first find dual variables $\delta_k^{(2)} > 0$ for $k \in \mathcal{T}$ such that the rate constraints are satisfied with inequality. Then, we obtain a power constraint dual variable $\theta^{(2)}$ that satisfies the power constraint with equality. As our results show this approach is much quicker than trying to find a dual optimal through subgradient iterations.

In order to satisfy the rate constraints that were not satisfied previously by the maximum throughput solution, we must increase the dual variables $\delta_k > 0$ for $k \in \mathcal{T}$ such that

$$\sum_{n=1}^{N} \log_2 \left(1 + \left[\frac{c_k + \delta_k}{\theta \beta_{n,k} \ln 2} - 1\right]^+\right) \geq \check{d}_k, \quad k \in \mathcal{T} \tag{4.32}$$

The value of the power constraint dual variable θ in (4.32) is bounded by

$$\theta^{(1)} < \theta < \bar{\theta} \tag{4.33}$$

where $\theta^{(1)}$ is the value given by the maximum throughput solution and $\bar{\theta}$ is an upper limit. This is any value that we can guarantee is higher than our desired power constrained dual variable $\theta^{(2)}$. We use an upper limit $\bar{\theta} = W\theta^{(1)}$, where $W > 1$. We will study the effect of W in subsection 4.2.5.

In the sum over subcarriers n in (4.32), only some subcarriers contribute to user k's rate. We define the set of subcarriers that contribute to user k's rate as

$$\mathcal{A}_k(\theta, \delta_k) \doteq \left\{ n \in \{1, \ldots, N\} : (k \in S_n) \wedge \left(\beta_{n,k} < \frac{c_k + \delta_k}{\theta \ln 2} \right) \right\}, \quad \forall k \in \mathcal{T}. \tag{4.34}$$

The subcarriers in set $\mathcal{A}_k$ satisfy two conditions: they are assigned to user k as indicated by the set S_n; and the user k's allocated power for subcarrier n is strictly higher than zero which is indicated by the inequality in parentheses in (4.34). To evaluate (4.34), we need dual variables θ and δ_k. We initially use $\delta_k = 0$ and $\bar{\theta} = W\theta$. This choice should give enough subcarriers to satisfy the minimum rate $\check{d}_k$ in (4.32). The set of selected subcarriers is then

$$\mathcal{A}'_k = \mathcal{A}_k(\bar{\theta}, 0) = \left\{ n \in \{1, \ldots, N\} : (k \in S_n^0) \wedge \left(\beta_{n,k} < \frac{c_k}{\bar{\theta} \ln 2} \right) \right\} \quad \forall k \in \mathcal{T}. \tag{4.35}$$

We rewrite (4.32) performing the sum only on the subcarriers n given by sets (4.35)

$$\sum_{n \in \mathcal{A}'_k} \log_2 \left(\frac{c_k + \delta_k}{\bar{\theta} \beta_{n,k} \ln 2} \right) \geq \check{d}_k, \quad k \in \mathcal{T}. \tag{4.36}$$

Evaluating at equality and manipulating the previous expression we obtain the minimum value of the dual variable δ_k required to satisfy (4.36),

$$\delta_k^{(2)} = \left[(\bar{\theta} \ln 2) \left(2^{\check{d}_k} \prod_{n \in \mathcal{A}'_k} \beta_{k,n} \right)^{|\mathcal{A}'_k|^{-1}} - c_k \right]^+. \tag{4.37}$$

Once we have computed all $\delta_k^{(2)}$ for $k \in \mathcal{T}$, we compute the power constraint dual variable θ that must satisfy the power constraint

$$\sum_{k=1}^{K} \sum_{n : k \in S_n^0} \beta_{k,n} \left[\frac{c_k + \delta_k^{(2)}}{\theta \beta_{n,k} \ln 2} - 1 \right]^+ = \check{P}, \tag{4.38}$$

where we have defined $\delta_k^{(2)} = 0$ for users $k \notin \mathcal{T}$. We compute θ using algorithm 4. While algorithm 4 gives the power allocation for the rate-unconstrained case, we can replace c_k by $(c_k + \delta_k)$ and use the same algorithm. The resulting power constraint dual variable $\theta^{(2)}$ in conjunction with $\delta_k^{(2)}$ will satisfy the power constraint with equality and the rate constraints with inequality, provided that the problem is feasible.

In figure 4.2 (c.f. subsection 4.2.3) we have plotted in dashed line the sum power curve after arbitrarily adding dual variables $\{\delta_k\} = \{1.0, 1.0, 0.0\}$ to users 1,2 and 3 powers re-

spectively in (4.38). The points where users 1 and 2 become active changed from β_k/c_k^0 to $\beta_k/(c_k^0+\mu_k)$, which are lower than the original ones. In addition, the slopes between points increased from c_k^0 to $(c_k^0+\mu_k)$. The point where user 3 becomes active did not change because $\delta_3 = 0$, user 3 does not get any power now because the root of the curve moved from 0.9 to 0.5, which occurs before user 3 becomes active at $\eta = 0.7$. This example shows that considering the rate constraint dual variables has multiple effects. It can change the order in which users become active, it can make inactive users active or vice versa, and it assigns them different powers. The role of the rate constraints dual variables is to finely adjust the power levels so user rates can be satisfied, while preserving some degree of multiuser diversity.

However, we cannot increase the dual variables δ_k indefinitely since there is a point where the slopes get so high that only the first users get power. If one of the users in $\mathcal{T}$ becomes inactive, its rate constraint will not be met and the heuristic method will declare that it cannot find a feasible point. Notice that a feasible point may exist, but it will require a more extensive search than the one we are willing to afford with limited computation resources.

In summary, the proposed heuristic consists of the following simple steps. Compute power allocation without rate constraints using algorithm 4 and find set $\mathcal{T}$ in (4.30). If minimum rates are not achieved for some users, compute sets $\mathcal{A}'(\bar{\theta}, 0)$ in (4.35) and rate constraint dual variables $\{\delta_k^{(2)}\}$ using (4.37). These dual variables are higher than the optimal ones but will guarantee that rate constraints are satisfied producing a primal feasible point. Afterwards, we compute the power constraint dual variable $\theta^{(2)}$ using dual variables $\{\delta_k^{(2)}\}$ in algorithm 4. Finally, we compute the user rates power vector using dual variables $(\theta^{(2)}, \{\delta_k^{(2)}\})$. The computational complexity of the proposed heuristic is linear with the number of subcarriers and users and can be written as $O(M|\mathcal{N}|)$, where $|\mathcal{N}|$ is the size of the set of subcarriers considered and M the number antennas since this limits the number of users k selected per subcarrier.

Because the heuristic algorithm reduces the power of the users that are not in set $\mathcal{T}$, it is possible that the algorithm makes certain users that are not in $\mathcal{T}$ unfeasible, but whose maximum throughput rates are very close to the feasibility region boundary. This will produce a point that is not feasible because it will not satisfy the rate constraints for all users in $\mathcal{T}$. However, this can be solved in the following way. For users $(k \in \mathcal{D}) \wedge (k \notin \mathcal{T})$ such that $r_k^{(1)}$ is close to $\check{d}_k$, include such users in set $\mathcal{T}$ and run the same procedure outlined above. The algorithm, if successful, will guarantee that the rate constraint is fulfilled for the added users. Therefore, the proposed heuristic method extends to these cases.

Algorithm 5 lists the steps of the proposed heuristic method to obtain a feasible point $\{p_{n,k}^{(2)}\}$ to problem (4.13–4.15). Algorithm 4 is executed in advance, so $W\theta^{(1)}$ is used as an upper bound for θ.

Input: Subcarrier assignment sets S_n, Current rates $r_k^{(1)}$, lower bound $\theta^{(1)}$, upper bound factor W
Output: New user rates $r_k^{(2)}$
- Compute set $\mathcal{T}$ of users in need using (4.30)
- Compute upper bound $\bar{\theta} = W\theta^{(1)}$
if $\mathcal{T} = \emptyset$ **then**
  Exit
else
  Compute sets $\mathcal{A}'_k = \mathcal{A}_k(\bar{\theta}, 0)$ in (4.35) for $k \in \mathcal{T}$.
  Compute rate constraints dual variables $\delta_k^{(2)}$ for $k \in \mathcal{T}$ using (4.37).
  Compute power constraint dual variable $\theta^{(2)}$ and user rates $r_k^{(2)}$ using algorithm 4
  $[\theta^{(2)}, \{r_k^{(2)}\}]$= Max_Troughput_power ($\{1, \ldots, N\}$, S_n, $\theta^{(1)}$, $\{\delta_k^{(2)}\}$)
end if

Algorithm 5 Rate-constrained power allocation: $r_k^{(2)}$= Rate_Constrained_power (S_n, $r_k^{(1)}$, $\theta^{(1)}$, W)

Table 4.2 Parameters for Figures 4.5 and 4.7

K	N	D	$\check{P}$
16	8	1	20

4.2.5 Heuristic vs. Optimal Results

In this subsection, we evaluate how close to the optimal are the points given by power allocation heuristics. Figure 4.5 shows in square markers the sum rate obtained when optimally solving problem (4.13–4.15) for the parameters listed in table 4.2, where only one user has minimum rate requirements. The sum rate is constant until the rate constraint is active for that user at ≈ 32 bps/Hz. Then, it decreases until the problem is unfeasible.

After executing the maximum throughput power allocation method in subsection 4.2.3, the difference between the optimal dual variable θ^* and $\theta^{(1)}$ — the dual variable computed by algorithm 4— is zero. This gives rates equal to the optimal ones and is illustrated by the flat part of the curves before 32 bps/Hz in figure 4.5.

Rate constrained power allocation is used when the rate constraints are higher than the maximum throughput rate. We ran algorithm 5 in subsection 4.2.4 with parameter $W = 3$ and obtained the sum rate over all users shown in figure 4.5 by circle markers. The algorithm is able to find feasible points up to 38 bps/Hz but the gap against the dual upper bound is significant for rates between 32 and 38 bps/Hz. It is possible to reduce this gap by changing

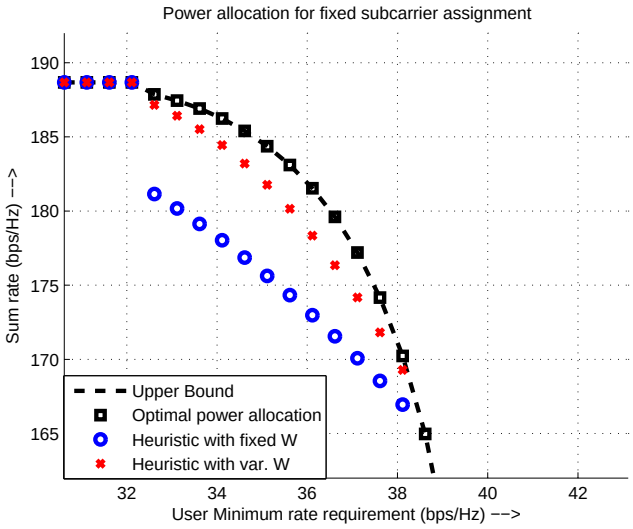


Figure 4.5 Optimal and heuristic power allocation comparison.

W to a lower value, but then the algorithm is not able to find feasible points at rates close to 38 bps/Hz.

To overcome this situation we make the parameter W vary depending on the difference between the required rate and the user rate achieved by the maximum throughput PA i.e. $W_k = (\check{d}_k - r_k)$. We obtain a substantial improvement in the power allocation heuristic algorithm by making $W_k = 2^{\epsilon(\check{d}_k - r_k)}$. Figure 4.5 shows in cross markers the result for $\epsilon = 0.175$. The feasible points give a sum rate that closely follows the optimal solution. This is remarkable considering that we are performing very few operations to obtain these feasible points.

The choice of parameter ϵ has an effect on how close the attained sum rate is to the optimal and the range of the supported rates. Figure 4.6 illustrates one example for the same parameters listed in table 4.2 and three different values for ϵ. As we increase ϵ the range of supported rates increases but the achieved sum rate is farther from the optimal. Higher values of parameter ϵ, e.g. 0.5, will make the algorithm more robust but will produce a bigger gap against the optimal solution.

The power allocation algorithm 4 is exact and always gives the optimal power allocation. On the other hand, the performance of algorithm 5 depends only on the parameter ϵ. We performed multiple experiments for different system configurations and found that the parameter value $\epsilon = 0.2$ works for most cases. We do not present an extensive evaluation of effect of parameter ϵ on the PA algorithms because the overall performance of the heuristic method will depend more on the effectiveness of the subcarriers assignment algorithm, which we present in section 4.3.

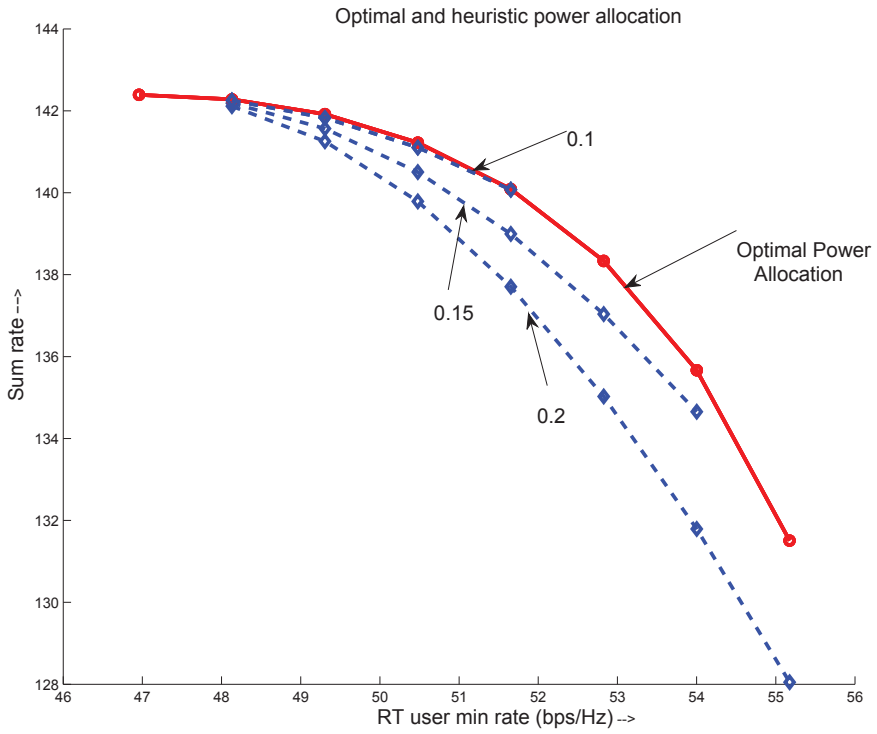


Figure 4.6 Optimal and heuristic power allocation comparison for different values of ϵ.

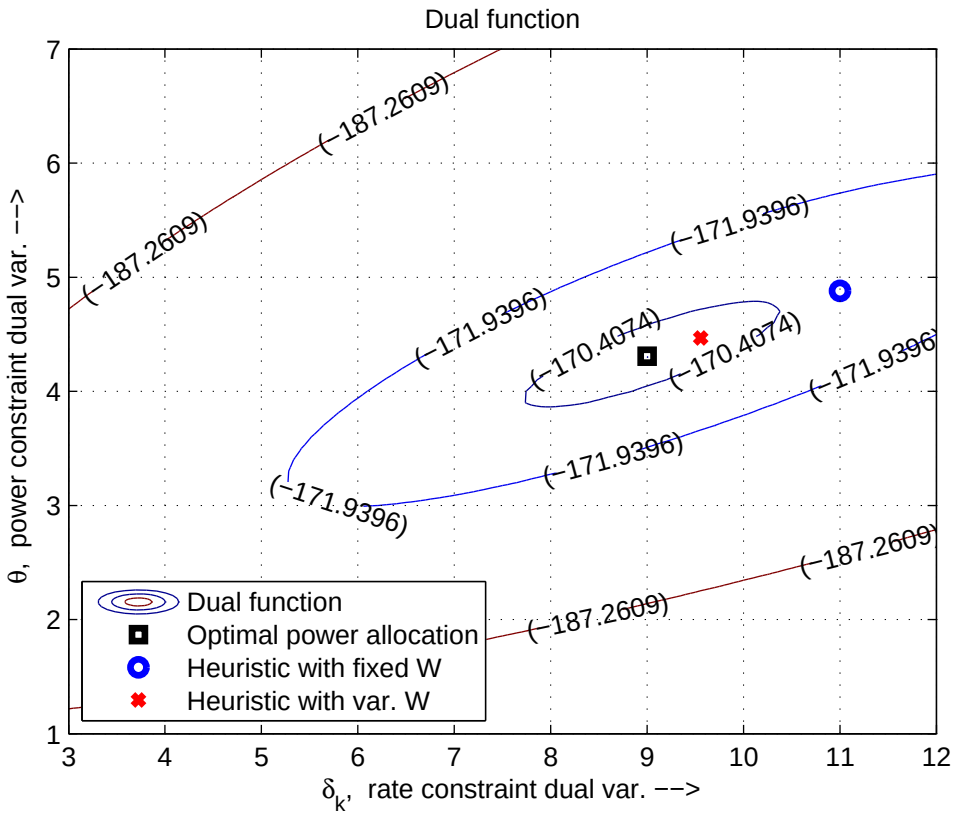


Figure 4.7 Dual function of power allocation problem.

Finally for illustration purposes, figure 4.7 shows the contour map of the dual function for the case when the minimum rate $\check{d}_k = 38.6$ bps/Hz, which corresponds to the highest feasible point found by the heuristic in figure 4.5. The optimal point shown by a square marker indicates dual variables that satisfy the rate and power constraints with equality, while the dual variables corresponding to the feasible points obtained by the heuristic method are shown by circle and cross markers; these dual variables are higher than the optimal because they satisfy the rate constraints with inequality.

In conclusion for this section, the proposed power allocation heuristic method effectively finds feasible points to problem (4.13–4.15). The difference against the optimal solution is very small and these points are obtained with very few operations, justifying its use in the proposed heuristic method.

4.3 Efficient Subcarrier Assignment

In this section, we present the design of the heuristic methods to perform the initial *maximum throughput* subcarrier assignment, and the subsequent *rate-constrained* subcarrier

reassignment corresponding to blocks A and B in figure 4.1. The objectives of blocks A and B differ. For block A, we want to assign subcarriers to get high throughput. This would correspond to finding the optimal dual variable λ associated to the power constraint in the dual formulation in section 3.2 making the dual variable vector associated to the rate constraints $\boldsymbol{\mu}$, equal to zero. For block B, we want to achieve the minimum rates of the RT users, which would correspond to adjusting $\boldsymbol{\mu} \geq 0$. Notice that the subcarrier assignment is controlled by the dual variables through equations (3.16) and (3.27), but here we are interested in obtaining a subcarrier assignment much more efficiently although suboptimally. The approach we follow is to assign subcarriers to users that have the following channel vector characteristics: large norms and quasi-orthogonality among the selected users. The user rates are affected by the effective channel gain $\beta_{n,k}^{-1}$ in (4.21) which increases in these cases. After selecting a subcarrier assignment, we perform power allocation using the efficient heuristic methods designed in section 4.2.

We first discuss the reported methods for maximum throughput subcarrier assignment in subsection 4.3.1 and select the SUS algorithm [28] from the current methods available in the literature. We use the subcarrier assignment given by the SUS algorithm as the starting point of our heuristic method. If the minimum rates are not met after rate-constrained power allocation, we proceed to reassign subcarriers. After subcarrier assignment or re-assignment, the power allocation algorithms 4 and 5 presented in section 4.2 are invoked.

In subsection 4.3.2, we discuss rate-constrained subcarrier *re*-assignment, which corresponds to block B in the block diagram of figure 4.1. We review an algorithm reported in the literature and present our method, which is an adaptation of the SUS algorithm to the rate-constrained case. This further justifies the complete presentation of the SUS algorithm and its comparison to the ZFUS algorithm for maximum throughput subcarrier assignment in subsection 4.3.1. The analysis on the relation between the SUS and ZFUS algorithms is an original contribution of this chapter and it brings a new insight into the algorithms.

4.3.1 Maximum Throughput Subcarrier Assignment

Maximum throughput subcarrier assignment consists of finding a set of users for each subcarrier that, for a given power allocation, will produce the maximum throughput. In the OFDM-SDMA system using ZF beamforming, maximum throughput subcarrier assignment falls in the category of user selection for Zero Force Beamforming (ZFBF) techniques. Early work focused on studying the capacity slope as the number of users grows to infinity. In [28], it is proved that when the number of users is large, ZFBF combined with user selection achieves a sum rate that has the same scaling law as that of Dirty Paper Coding (DPC) which is the optimal strategy; that is the sum rate grows as $M \log(1 + \frac{P}{M} \log K)$, where $\frac{P}{M}$ is the

power equally assigned to all selected users. Also in [28], a heuristic user selection method is proposed to prove this asymptotic slope based on a search of user sets with semiorthogonal channel vectors.

In the block diagram of figure 4.1 the subcarriers are reassigned if the minimum rates are not achieved. This implies that it is not worth spending too much time looking for a near-optimal solution in block A, if the subcarriers assignment is going to be replaced. Therefore, our focus is on the efficiency of the user selection method rather than on its performance. We are interested in an efficient heuristic method for ZFBF user selection that has an acceptable performance not far from the optimal for a practical number of users (< 100). There is an ample literature on user selection for ZFBF, see [17, 41, 42] and references therein. Here we review two main approaches, the Zero Force with user selection (ZFUS) and the semiorthogonal user selection (SUS) algorithms reported in [43] and [28] respectively, which have been widely studied and several improvements have been made to the original algorithms. The main difference is that ZFUS performs full enumeration of the remaining users after a set of users has been selected, while the SUS selects the other user channel vectors that are semiorthogonal to the previously selected users without need of computing the pseudo-inverse matrices.

We describe each of these approaches in the sequel. Our analysis shows that there exists a fundamental relation between them. We compare their computational complexity and choose an efficient implementation of the SUS algorithm over the ZFUS method to perform subcarrier assignment in block A of figure 4.1.

Zero Force with User Selection (ZFUS) algorithm

The off-line near-optimal algorithm 1 proposed in chapter 3 can be used to solve problem (3.1–3.7) without rate constraints (3.3); it suffices to make the rate dual variables $\mu_k = 0$ and to perform subgradient iterations to find the optimal dual variable λ. This will give us the subcarrier assignment and user powers by evaluating (3.16) and (3.27). For the rate unconstrained case, the solution is often primal feasible and there is no need for further processing. However, the computational complexity of this method is high; this is mostly due to the number of matrix pseudo-inverse computations which increases as $I(K) = \sum_{m=1}^{M} \binom{K}{m}$.

To reduce this computational complexity, the ZFUS algorithm [43] makes two main simplifications. First, it selects the channel with largest vector norm for each subcarrier and then iteratively adds one user at a time, picking the user that maximizes the sum rate. This selection method reduces the number of pseudo-inverse matrix computations to $J(K) = \sum_{m=1}^{M-1} (K - m)$ which increases much more slowly than $I(K)$ above. The second simplification is to put one power constraint per subcarrier $\check{P}/N$ instead of a total power

constraint $\check{P}$. Power allocation is thus performed independently for each subcarrier, which leads to solution points that are very close to the optimal when channel gains are balanced among users and subcarriers. However, it yields lower performance when *all* users have very bad channel conditions in certain subcarriers, because the power assigned to those subcarriers with largely attenuated channels is wasted.

The user selection process at each iteration i consists of finding the user π_i that added to the currently selected users, produces the maximum sum rate

$$\pi_i = \arg \max_{k \in \{\mathcal{K} - S_0\}} R(S_0 \cup \{k\}), \tag{4.39}$$

where S_0 is the set of selected users at iteration $(i-1)$, $\mathcal{K}$ is the set of all users and R is the sum rate attained by the solution to the problem

$$\max_{\{p_k\} \geq 0} R(\phi) = \sum_{k \in \phi} c_k \log_2(1 + p_k), \tag{4.40}$$

$$\sum_{k \in \phi} \beta_k^{(\phi)} p_k - \check{P}/N \leq 0, \tag{4.41}$$

$$\beta_k^{(\phi)} \doteq \left[(\mathbf{H}_\phi^\dagger)^H \mathbf{H}_\phi^\dagger \right]_{k,k}, \tag{4.42}$$

where $\mathbf{H}_\phi$ is the channel matrix formed by users $j \in \phi \doteq S_0 \cup \{k\}$ given by (4.3) with set $S_n = \phi$. Problem (4.40) is equivalent to problem (4.13) without considering rate constraints (4.15) and dropping the subcarrier index since the same procedure is applied to all subcarriers independently. The solution to this problem is obtained in subsection 4.2.1 by making $\delta_k = 0$ in (4.20)

$$p_k = \left[\frac{c_k}{\theta_n^0 \beta_k^{(\phi)} \ln 2} - 1 \right]^+, \tag{4.43}$$

where θ_n^0 is the dual variable obtained by solving the associated dual problem or much more efficiently by invoking heuristic algorithm 4 with only one subcarrier as input [2]. The attained rate $R(\phi)$ and the allocated power p_k expressions are somehow similar to Eqs. (21) and (23) in [43], but we use our notation and formulation that is more general since we do not assume that all users in the SDMA set are active. We use the superscript ϕ for β_k to highlight that these factors are computed for set ϕ.

Computing the user selection rule (4.39) for this method requires one matrix pseudo-inverse calculation for every candidate user $k \in \{\mathcal{K} - S_0\}$ in order to obtain $\beta_k^{(\phi)}$. This is the

2. See subsection 4.4 for further explanation on per-subcarrier power constraints.

main drawback of this method since the computation time would be prohibitive for moderately high M and K. In [43] however, the complexity of the pseudo-inverse computations is reduced by computing them incrementally. In addition, [42] used a LQ decomposition of the channel matrix to further reduce the pseudo-inverse computational complexity. The reduced computational complexity of this method is $O(NKM^3)$ considering all subcarriers, which is similar to that of other methods (see table 4.3).

Semiorthogonal User Selection (SUS) subcarrier assignment algorithm

We first rewrite model (4.40–4.41) to introduce the SUS method in relation to ZFUS. We replace the optimization variable p_k by $\beta_k^{-1}q_k$, where $q_k \geq 0$ is the new optimization variable

$$\max_{\{q_k\}\geq 0} R(\phi) = \sum_{k\in\phi} c_k \log_2(1+\beta_k^{-1}q_k), \tag{4.44}$$

$$\sum_{k\in\phi} q_k - \check{P}/N \leq 0. \tag{4.45}$$

The solution to problem (4.44–4.45) is now written as

$$q_k^0 = \left[\frac{c_k}{\theta_n^0 \ln 2} - \beta_k\right]^+, \tag{4.46}$$

and the objective at the solution point is

$$R(\phi) = \sum_{k\in\phi} c_k \log_2\left(1+\beta_k^{-1}q_k^0\right) \tag{4.47}$$

$$= \sum_{k'\in\phi} c_k \log_2\left(\frac{c_k\beta_k^{-1}}{\theta_n^0 \ln 2}\right), \tag{4.48}$$

where in (4.48), the sum includes only the users k' for which the power q_k^0 is greater than zero.

The user selection rule (4.39) in the ZFUS method requires two computations to evaluate which user to add to an existing SDMA set S_0. It first inverts the channel matrix to compute the effective channel gains β_k^{-1}, and then it performs power allocation to obtain the dual variable θ_n^0. We can perform these computations with different efficiency according to the method used, but what we would like is a method that does not require inverting the channel matrix nor performing power allocation to evaluate which user to add. This method can be devised by observing that objective (4.48) increases with the effective channel gain β_k^{-1}. Thus, one possibility is to choose the new user $k \in \{\mathcal{K} - S_0\}$ such that β_k^{-1} is maximized.

The new selection rule is then

$$\pi_i = \arg \max_{k \in \{\mathcal{K}-S_0\}} (\beta_k^{(\phi)})^{-1} = \arg \max_{k \in \{\mathcal{K}-S_0\}} \left[(\mathbf{H}_\phi^\dagger)^H \mathbf{H}_\phi^\dagger \right]_{k,k}^{-1}. \tag{4.49}$$

The ZFUS selection rule (4.39) and rule (4.49) differ because (4.49) only considers the effective gain of the user added, while (4.39) considers the rates of all users in the set. In addition, rule (4.49) does not consider the power allocation dual variable θ_n^0. We will also see later that (4.49) can be manipulated so that it does not require matrix pseudo-inverse computations and it has the potential of being efficiently implemented by limiting the examined users to only those semiorthogonal to the users already selected. We will thus select users based solely on the vector spatial characteristics, making user selection independent of power allocation.

Defining $\mathbf{H}_{(S_0)}$, the matrix formed by arranging in the rows the channel vectors corresponding to the users in S_0, and $\mathbf{G}_{(S_0)}$ an orthogonal base for the subspace spanned by $\mathbf{H}_{(S_0)}$, theorem 1 in [44] relates the effective channel gain $(\beta_k^{(\phi)})^{-1}$ to $\mathbf{G}_{(S_0)}$

$$(\beta_k^{(\phi)})^{-1} = \|\mathbf{h}_k(\mathbf{I} - \mathbf{G}_{(S_0)})\|^2, \tag{4.50}$$

so that we can re-write the selection rule (4.49)

$$\pi_i = \arg \max_{k \in \{\mathcal{K}-S_0\}} \|\mathbf{h}_k(\mathbf{I} - \mathbf{G}_{(S_0)})\|. \tag{4.51}$$

Using this rule, we select the user whose channel vector $\mathbf{h}_k$ has the largest projection to the subspace orthogonal to the channel vectors of users already selected. This is the same selection rule used by the SUS algorithm proposed in [28].

In what follows, we split the SUS algorithm in two parts: an initialization stage implemented in algorithm 6 and a user search stage implemented in algorithm 7. This is done to later adapt the SUS algorithm to the rate-constrained case that we describe in subsection 4.3.2. Algorithm 6 along with algorithm 7 are equivalent to the SUS algorithm in [28]. In the SUS initialization algorithm 6, we choose the user with the highest norm among the available users and build matrix $\mathbf{G}_1$ that spans the vectors parallel to the chosen vector. After the initialization phase, algorithm 7 is executed using user and matrix as input. In the first part inside the *while* loop, the SUS algorithm eliminates the user channels that are approximately colinear to the users already selected. The threshold parameter $\breve{\alpha}$ controls how selective is the algorithm in terms of the orthogonality among users; if the internal product between two user channel vectors is lower than $\breve{\alpha}$, their channel vectors are semiorthogonal and they should be kept. A default value $\breve{\alpha} = 0.3$ is used in our numerical evaluations [28]. This filtering is done to reduce computations in the subsequent steps because users with approximately

Table 4.3 Maximum Throughput ZF User selection algorithms complexity

Algorithm	Method	Complexity
SUS [28]	Semiorthogonal user sel.	$O(KNM^3)$
S-SUS [45]	Simplified SUS.	$O(KNM^2)$
ZFUS [43]	Zero Forcing user sel.	$O(KNM^3)$
SWF [42]	Sequential water-filling for user sel.	$O(KNM^3)$
GWC [41]	Greedy weighted clique ZFBF	$O(KNM)$

colinear vectors will not provide high effective channel gains. Then, we apply selection rule (4.51) to select the best user. The vectors $\mathbf{g}_k$ computed inside the *for* loop in algorithm 7 are the components of the user channels orthogonal to the subspace spanned by the user channel vectors already selected. After the user with the maximum component is picked, the subspace matrix is updated. The procedure to obtain the subcarrier assignment for all subcarriers and the user power and rate allocation is listed in algorithm 8, which invokes algorithms 4, 6 and 7. The computational complexity of algorithm 8 and the ZFUS algorithm after the LQ decomposition simplification in [42], is $O(NKM^3)$ for both algorithms. However, in [45] a simplified SUS algorithm is proposed that has exactly the same performance as SUS and a lower computational complexity, $O(NKM^2)$.

Results comparing the SUS and a variation of the ZFUS methods in [42] show that the objective achieved by the SUS method is lower by 2 bps/Hz at the most, for the configuration used. This is not a substantial difference and, as stated before, our emphasis is on the computational efficiency of the algorithm rather than on performance. For this reason, despite the fact that the SUS performance is slightly lower than the ZFUS method, we select the SUS as a starting point for our heuristic method. The main advantages of the SUS algorithm are its low complexity and adaptability. Table 4.3 lists the computational complexity of several methods for user selection proposed in the literature. Notice that the SUS, ZFUS and SWF have similar computational complexities, only the GWC algorithm has linear complexity but its reported performance is low [41].

We performed numerical evaluations using the parameters listed in table 4.4 to compare the ZFUS, the SUS method and the dual upper bound. The results are illustrated in figure 4.8 and verify the reported performance. The processing parameters required for the SUS algorithm and power allocation are the power dual variable lower bound $\check{\theta}$ and the user filtering orthogonality threshold $\check{\alpha}$ which are also listed in table 4.4. Given that the performance are close, the main reasons to choose the SUS algorithm are low computational complexity and adaptability to the minimum rate case as we will see next.

Table 4.4 Parameters for Figure 4.8

K	N	M	$\check{P}$	$\check{\theta}$	$\check{\alpha}$
16	8	4	20 dBm	0.1	0.3

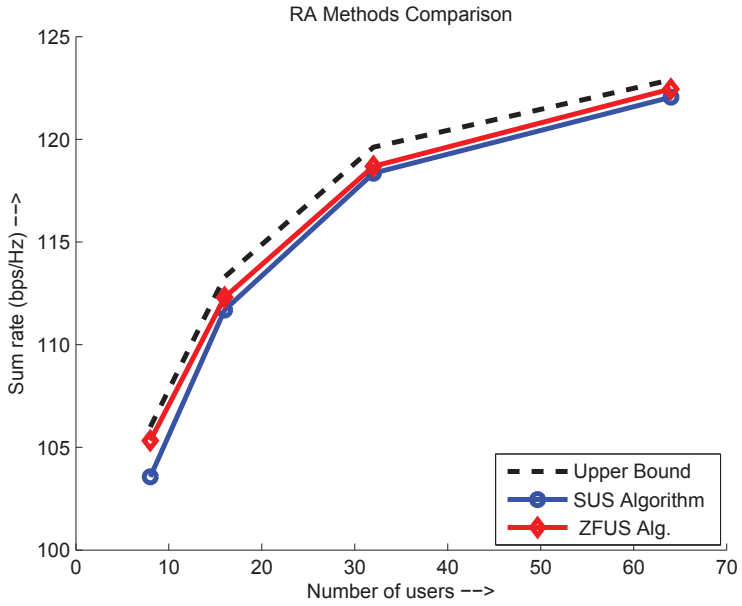


Figure 4.8 Maximum Throughput Optimal and heuristic methods comparison.

Input: Start users set $\mathcal{U}_0$
Output: First selected user S_0, orthogonal subspace $\mathbf{G}_1$ spanned by users in S_0
$\pi_1 = \arg\ \max_{k \in \mathcal{U}_0} \|\mathbf{h}_k\|$
$\mathbf{G}_1 = \mathbf{h}_{\pi_1}^H \mathbf{h}_{\pi_1} / \|\mathbf{h}_{\pi_1}\|$
$S_0 = \{\pi_1\}$

Algorithm 6 SUS initialization algorithm: $[S_0, \mathbf{G}_1]$= SUS_Init($\mathcal{U}_0$)

Input: Search users set U_i, previously selected users set S_0, orthogonal subspace $\mathbf{G}_i$ spanned by S_0 vectors, user filtering orthogonality threshold $\check{\alpha}$
Output: Selected users set S_0, orthogonal subspace $\mathbf{G}_i$ spanned by users in S_0
$i \leftarrow |S_0|$; $i \leftarrow i+1$
while $i \leq M$ **do**
  Eliminate users that are approximately colinear (not semi-orthogonal:)
  $\mathcal{U}_i = \{k \in \mathcal{U}_{i-1} : (k \notin S_0) \wedge \frac{\mathbf{h}_k \mathbf{h}_j}{\|\mathbf{h}_k\| \|\mathbf{h}_j\|} < \check{\alpha}, \forall j \in \mathcal{S}_0\}$
  if $\mathcal{U}_i = \emptyset$ **then**
    Break
  end if
  Compute projection onto the orthogonal subspace:
  for all $k \in \mathcal{U}_i$ **do**
    $\mathbf{g}_k = \mathbf{h}_k \left(\mathbf{I} - \mathbf{G}_{i-1}\right)$
  end for
  Select best user:
  $\pi_i = \arg \max_{k \in \mathcal{U}_i} \|\mathbf{g}_k\|$
  $S_0 \leftarrow S_0 \cup \{\pi_i\}$
  $\mathbf{G}_i = \mathbf{G}_{i-1} + \mathbf{h}_{\pi_i}^H \mathbf{h}_{\pi_i} / \|\mathbf{h}_{\pi_i}\|$
  $i \leftarrow i+1$
end while

Algorithm 7 SUS Users search algorithm: $[S_0, \mathbf{G}_i]$= SUS_Search($U_i, \mathcal{U}_0$, $\mathbf{G}_i$)

Input: Problem parameters
Output: Selected users set per subcarrier S_n^0, User rates $\{r_{n,k}^{(0)}\}$, Power constraint dual variable $\theta^{(0)}$
for all subcarriers $n \in \{1, \ldots, N\}$ **do**
  Initialize SUS algorithm 6 scanning all users $\mathcal{K} = \{1, \ldots, K\}$
  $[\ \bar{S}, \mathbf{G}_1\ ]$= SUS_init ($\mathcal{K}$)
  Find a SDMA set S_n^0 invoking algorithm 7 scanning remaining users $\mathcal{K} - \bar{S}$
  $[\ S_n^0, \mathbf{G}_i\ ]$= SUS_search ($\mathcal{K} - \bar{S}, \bar{S}, \mathbf{G}_1, \check{\alpha}$)
  Compute pseudo-inverse of channel matrix formed by users in S_n^0 and compute $\beta_{n,k}^{(S_0)}$ using (4.10)
end for
Perform Maximum Throughput power allocation using algorithm 4
$[\theta^{(0)}, \{r_k^{(0)}\}]$= Max_Troughput_power ($\{S_n^0\}$, $\check{\theta}$)

Algorithm 8 Maximum Throughput subcarrier assignment algorithm

4.3.2 Papoutsis Rate Constrained Subcarrier Reassignment

We start by summarizing the only other work that has been reported in the literature to solve problem (3.1–3.7). Papoutsis et al [22] used a heuristic method for resource allocation with two stages. In stage 1, subcarriers are assigned and the users power is allocated without considering rate constraints. This is done through a modification of the ZFUS algorithm [43] which has complexity $O(NKM^3)$.

In stage 2, subcarrier reassignment is carried out on a per-user basis for all users whose minimum rate constraints have not been satisfied in stage 1. For each user in need k, a cost matrix $\mathbf{V}^{(k)}$ is computed whose entries $v_{n,j}^{(k)}$ indicate the reduction in the sum rate if user k is to replace user j who is currently assigned to subcarrier n. The dimension of this matrix is $N \times M$ because there are N subcarriers and each one is assigned M users at the most. If subcarrier n is already assigned to user k, such subcarrier is not taken into consideration.

For the subcarrier with the minimum cost, the user with the lowest cost is replaced by user k after verifying certain conditions. The rule for selecting the subcarrier n^0 and the user k^0 to be replaced is

$$[n^0, k^0] = \arg \min_{n \in \mathcal{N}, k^0 \in S_n^0} \mathbf{V}^{(k)}, \tag{4.52}$$

$$\text{where } v_{n,k^0}^{(k)} = \sum_{j \in S_n^0} r_{n,j}^0 - R(\phi_{n,k^0}^{(k)}), \tag{4.53}$$

$$\phi_{n,k^0}^{(k)}) = S_n^0 - \{k^0\} \cup \{k\} \tag{4.54}$$

and $R(\phi_{n,k^0}^{(k)})$ is the solution to problem (4.40). To compute the cost matrix $\mathbf{V}^{(k)}$ we need to calculate the new rates after replacing each of the users in the current SDMA set S_n^0 by user k. We thus need to invert up to M matrices per subcarrier. The computational complexity of such inversion using the LQ simplification in [42] is $O(M^2)$. In addition, to compute $R(\phi_n)$, a power allocation is required which we assume is done using an approximation method with negligible complexity. Thus, the computational complexity of (4.52) is $O(NM^4)$.

The subcarrier reassignment process is repeated for user k until its required minimum rate is met. The number of iterations is bounded by N, and the process is repeated for all users in need which is bounded by K. Therefore, the computational complexity of Papoutsis' method is $O(KN^2M^4)$.

4.3.3 Proposed Subcarrier Reassignment Heuristic

We now describe the proposed rate-constrained subcarrier *re*-assignment algorithm. It builds upon the SUS search and power allocation algorithms presented so far. Algorithm

9 lists the corresponding pseudo-code of the proposed subcarrier reassignment heuristic. In what follows, we explain the pseudo-code and present the computational complexity of each stage.

After executing algorithm 8 in block 1 of the diagram in figure 4.1, we obtain the user rates per subcarrier $r_{n,k}^{(0)}$ with computational complexity $O(NKM^3)$. The user rates are simply computed by $r_k^{(0)} = \sum_{n \in \mathcal{C}_k} r_{n,k}^{(0)}$, where $\mathcal{C}_k$ is the set of subcarriers assigned to user k. In the pseudo-code of algorithm 9, we start by computing the set of users in need $\mathcal{T}$ using (4.30). We first scan the subcarriers in which any of the users in need have good channel conditions, so they can be first reassigned to these users. The computational complexity of the subcarrier ordering is bounded by $O(N^2)$. For each subcarrier n, we build a critical set $\mathcal{E}$ containing the users in the current SDMA set that can not be removed from the SDMA set because that would take the user out of feasibility. We consider two cases: first that the set $\mathcal{E}$ is empty. In this case we invoke the SUS initialization algorithm 6 to select the strongest user in $\mathcal{T}$ as the first element of the SDMA set. Then, we invoke the SUS search algorithm 7 to add users to the SDMA set. We initially scan other users in $\mathcal{T}$ so that they can be added with priority to the SDMA set, and if the SDMA set has not yet been completed, we scan the rest of the users $\{1, \ldots, K\}$.

In the second case, when set $\mathcal{E}$ is not empty, we initialize the SDMA set with all the users in set $\mathcal{E}$, and then add users invoking the SUS search algorithm. Notice that the users in $\mathcal{E}$ were previously selected by the SUS maximum throughput algorithm, and were part of the SDMA set for this subcarrier, so we can directly include them in the initial SDMA set. In the pseudo-code, however, we repeat the initialization and first phase search but this is solely to make the listed code more compact. To add users to this SDMA set, we scan the rest of the users but look first in the set of users in need $\mathcal{T}$. The difference between the cases $\mathcal{E} = \emptyset$ and $\mathcal{E} \neq \emptyset$ is that in the second case, we keep the users in need that were already in the set before trying to add more users.

Notice that all the selected users comply with the semiorthogonality condition of the SUS algorithm; the only change in computations to algorithm 8 is the order in which we examine the users. By changing the order, we are giving priority to the users in need. The computational complexity of this stage is bounded by the complexity of the maximum throughput SUS search algorithm $O(KM^3)$.

After obtaining the new SDMA set for this subcarrier, we perform maximum throughput power allocation invoking algorithm 4. If the resulting rates are feasible we exit the algorithm. Otherwise, we perform rate-constrained power allocation invoking algorithm 5. If the resulting rates are still not feasible, we continue reassigning subcarriers until the rates are feasible or there are no more subcarriers and the algorithm declares that is not able to find a

feasible point. This corresponds to the loop in the lower part of the block diagram in figure 4.1.

The power allocation algorithms have computational complexity $O(KN)$ as described in section 4.2.2. Assuming the worst case where all subcarriers are examined for reassignment, the proposed algorithm's overall computational complexity is

$$O_{\text{alg. 9}} = \begin{cases} O(KN^2), & \text{if } N > M^3 \\ O(KNM^3), & \text{otherwise.} \end{cases} \tag{4.55}$$

This is lower than Papoutsis' method computational complexity $O(KN^2M^4)$ for all N. The main reason for this reduction is not having to compute matrices $\mathbf{V}^{(k)}$ to evaluate the user selection rule (4.52) at each iteration.

4.4 Reduced Complexity Algorithm

In this section, we devise a variation to the subcarrier reassignment algorithm 9 that linearizes the dependency of the computational complexity in expression (4.55) with respect to the number of subcarriers N, for $N > M^3$. Since in LTE-Advanced systems, the maximum number of subchannels is large 550 and the maximum number of antennas is small 8 (c.f. chapter 5), it is utterly important to linearize the complexity with respect to N.

For this purpose, we solve a sum rate maximization problem with one power constraint per subcarrier instead of a total power constraint. Problem formulation (4.13–4.14) is replaced by

$$\max_{\{p_{n,k} \geq 0\}} \quad \sum_{n=1}^{N} \sum_{k=1}^{K} c_k \log_2(1 + p_{n,k}) \tag{4.56}$$

$$\sum_{k=1}^{K} \beta_{n,k} \; p_{n,k} - \check{P}/N \leq 0, \quad \forall n \in \{1, \ldots, N\} \tag{4.57}$$

where constraints (4.57) replace the total power constraint (4.14) and we do not consider the rate constraints (4.15). For well-balanced quality channels across subcarriers, the solutions to problem (4.13–4.14) and (4.56–4.57) are very close.

In the subcarrier iteration loop in algorithm 9, we update the user power corresponding to all subcarriers because the power-constraint dual variable affects them all, which produces the term N^2 in the complexity expression (4.55) for $N > M^3$. The new formulation (4.56–4.57) is useful here because in the subcarrier iteration loop, we would need to update the user power corresponding to only one subcarrier.

Table 4.5 Algorithms complexity

Algorithm	Complexity	Purpose
Max. Throughput PA 4	$O(K\|\mathcal{N}\|)$	Prob. (4.13–4.14)
Rate-constrained PA 5	$O(KN)$	Prob. (4.13–4.15)
Proposed heuristic method 9	Eq. (4.55)	Prob. (4.13–4.15)
Proposed simplified method 10	$O(KNM^3)$	Prob. (4.13–4.15)
Papoutsis' method [22]	$O(KN^2M^4)$	Prob. (4.13–4.15)
Dual bound	$O(NK^MM^3)$	Dual of (3.1–3.7)

We modify algorithm 9 after the SUS search and channel inversion by performing maximum throughput power allocation using algorithm 4 with only subcarrier n as input variable. Notice that in this case, there is one power constraint dual variable θ_n per subcarrier n. Therefore, when computing the power and rates after power allocation, only the ones corresponding to that subcarrier are affected, making the computational complexity of this step $O(K)$ as opposed to the original $O(KN)$.

The pseudocode of algorithm 4 considers the option of invoking the algorithm with a variable subset of subcarriers $\mathcal{N}$. Then, we can invoke this algorithm for one subcarrier at a time to consider the per-subcarrier power constraint case.

Algorithm 10 lists the changes to algorithm 9. We do not perform rate-constraint power allocation in this case. To increase the supported minimum rates, one would like to perform rate-constrained power allocation as algorithm 9 does. However, it is not possible to solve the problem without an increase in the computational complexity because the rate constraints are linked among subcarriers, i.e., a user rate is the contribution of all subcarriers assigned to that user and we cannot independently define a rate-constrained dual variable per subcarrier as we did for the power constraint. Therefore, algorithm 10 cannot support the high minimum rates that algorithm 9 can, but it is a more efficient algorithm when N is large.

The computational complexity of algorithm 10 is

$$O_{\text{alg. 10}} = O(KNM^3), \tag{4.58}$$

which is linear with N. Notice that (4.58) varies linearly with N and since N ranges from 6 to 550 in a LTE-Advanced system with Carrier Aggregation (CA), this results in a much faster algorithm for large N. For comparison, table 4.5 summarizes the computational complexity of the algorithms proposed in this chapter.

Input: Current rates $r_{n,k}$, current subcarrier assignment sets S_n
Output: New subcarrier assignment sets $S_n^{(1)}$, user rates $r_k^{(1)}$ or $r_k^{(2)}$
- Compute users in need set $\mathcal{T}$ using (4.30)
- Order subcarriers according to maximum norm of the users in need, i.e. according to $\max_{k\in\mathcal{T}} \|h_{n,k}\| \quad \forall n$, producing ordered set $\mathcal{N}$
for all $n \in \mathcal{N}$ **do**
- Compute critical user set $\mathcal{E}$ containing users in S_n for which $r_k - r_{n,k} < \check{d}_k$
If $\mathcal{E} = \emptyset$, $\mathcal{Z} \leftarrow \mathcal{T}$
Otherwise, $\mathcal{Z} \leftarrow \mathcal{E}$
$S_0 =$ SUS Init $(\mathcal{Z})$ alg. 6
$S_n^{(1)} =$ SUS search $(S_0, \mathcal{Z}, \mathbf{G})$ alg. 7
if $|S_n^{(1)}| < M$ **then**
$S_n^{(1)} =$ SUS search $(S_n^{(1)}, \{1, \ldots, K\}, \mathbf{G})$ alg. 7
end if
Compute pseudo-inverse of channel matrix formed by users in $S_n^{(1)}$ and compute $\beta_{n,k}$ using (4.10)
Run Maximum Throughput power allocation algorithm 4 obtaining $\theta^{(1)}, r_k^{(1)}$
if $r_k^{(1)}$ satisfy rate constraints **then**
Break
else
$W' = \max_k \epsilon[\check{d}_k - r_k^{(1)}]^+$; $W = 2^{W'}$;
Run rate constrained power allocation algorithm 5 obtaining $r_k^{(2)}$
if $r_k^{(2)}$ satisfy rate constraints **then**
Break
end if
end if
Update users in need $\mathcal{T}$
end for

Algorithm 9 Subcarrier Reassignment Heuristic Algorithm

Input: Current rates $r_{n,k}$, current subcarrier assignment sets S_n
Output: New subcarrier assignment sets $S_n^{(1)}$, user rates $r_k^{(1)}$
In algorithm 9, after SUS search and channel inversion:
Run Maximum Throughput power allocation algorithm 4 with input:
Current subcarrier n, Subcarrier assignment sets S_n^0 and lower bound $\check{\theta}$.
Obtaining $\theta_n^{(1)}, r_{n,k}^{(1)}$
if $\sum_n r_{n,k}^{(1)} \geq \check{d}_k$, $\forall k \in \mathcal{T}$ **then**
Break
end if

Algorithm 10 Per-subcarrier power constraint subcarrier Reassignment Heuristic Algorithm

4.5 Performance Comparison

The focus of this chapter has been devising heuristic algorithms to efficiently solve problem (3.1–3.7). In section 4.3.3, we proposed algorithm 9 and in section 4.4 its simplified version — algorithm 10. Table 4.5 shows that they have reduced computational complexity when compared to other methods. This makes them good candidates to be implemented in practical systems. In this section, we evaluate the objective achieved by these algorithms and the support of the fulfilled minimum rates. Our interest is on answering the following questions: how far from the optimal is the sum rate achieved by these methods; what is the range of the minimum rates supported by the rate-constrained power allocation in algorithm 9, as opposed to the maximum throughput power allocation in algorithm 10; and how these results compare to Papoutsis' method.

To answer these questions, we use a Rayleigh fading channel model to generate independent channels and compare numerically the supported minimum rates and the sum rate achieved by the following methods:

1. Dual-based upper bound using algorithm 1 from chapter 3
2. Papoutsis' algorithm [22] described in subsection 4.3.2
3. The proposed heuristic SUS-based heuristic algorithm 9 described in subsection 4.3.3
4. A simplified SUS-based heuristic algorithm 10 described in subsection 4.4 that performs per-subcarrier constrained power allocation.

Figure 4.9 illustrates one example of the objective achieved by these methods for the parameters listed under the title of table 4.6 and one real-time user, $D = 1$. The plots only show feasible points. Thus, when increasing the minimum required rates (horizontal axis), the curves stop if the methods can no longer find feasible points.

We start by solving the problem without considering minimum rate requirements as indicated by blocks A and 1 in figure 4.1; this gives us user rates $\{r_k^0\}$. If we were to extend the curves of figure 4.9 to zero, they would be flat curves with $\sum_k r_k^0$ as the sum rate. We want to focus on the domain where rate constraints are active. For this purpose, we increase the rate constraints incrementally

$$\check{d}_k = r_k^0 + \Delta_r \quad k \in \{1, \ldots, D\} \tag{4.59}$$

For a number of RT users $D > 1$, we use $\sum_{k=1}^{D} \check{d}_k$ to list the minimum rate constraint in tables and plots. In our numerical evaluations, we increase the rate constraints and try to find feasible points using the heuristics until the dual upper bound becomes negative indicating the problem unfeasibility. In figure 4.9, the upper bound provided by the dual function

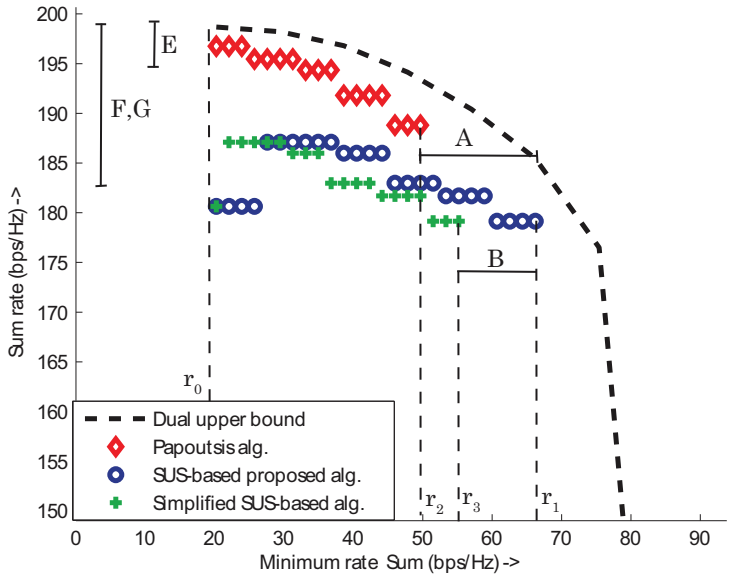


Figure 4.9 Optimal and heuristic methods comparison.

minimum is shown by a dashed line and it is the reference to measure the performance of all heuristic methods. Papoutsis' method is shown in diamond markers and closely follows the upper bound. The proposed SUS-based heuristic (algorithm 9) is shown in circle markers and its simplified version (algorithm 10) in cross markers. They have lower performance than Papoutsis' method, but they increase the range of supported minimum rates. To quantify these observations, we define the following measurements:

A: The difference in percentage between the minimum rate supported by the SUS-based heuristic algorithm 9 r_1, and Papoutsis' method r_2, i.e., $A = 100(r_1 - r_2)/r_1$.

B: The difference in percentage between the minimum rate supported by the SUS-based heuristic algorithm r_1 and its simplified version r_3 in percentage. A and B indicate how much the proposed SUS based algorithm 9 increases the range of supported minimum rates. The larger these measurements are, the better the proposed algorithm 9.

E: The difference in percentage between the upper dual bound u_1 and the sum rate achieved by Papoutsis' method u_2, i.e. $E = 100(u_1 - u_2)/u_1$. To compute u_1 and u_2 we average the sum rates over the minimum rates supported by Papoutsis' method. This corresponds to rates between r_0 up to r_2 in figure 4.9, where r_0 is sum of of the rates at which the D RT users' rate constraints become active. We average the sum rates over the same rate interval for all methods.

Table 4.6 Average measurements for variable D

Measurement	$D=1$	$D=2$	$D=3$	$D=4$
Rate gap:				
E	2.2	1.5	2.1	2.5
F	15.7	10.8	12.3	14.0
G	15.9	10.9	12.8	14.2
Range gap:				
A	21.8	15.2	12.4	11.5
B	20.7	13.6	11.9	9.5

F: The difference in percentage between the upper dual bound and the SUS-based heuristic algorithm 9.

G: The difference in percentage between the upper dual bound and the simplified SUS-based heuristic algorithm 10. E, F and G indicate how far the sum rate is from the upper bound for each method. The smaller this measurement is, the better the algorithm.

Averaging these measurements over 100 channel realizations, we obtain the results listed in table 4.6 for various number of RT users D, other parameters are kept fixed $N = 8$, $K = 8$, $M = 3$, $\check{P} = 20$, $\epsilon = 0.2$. The difference between Papoutsis' method performance and the upper bound is very small (< 2.5 %). This is because Papoutsis' method minimizes the throughput reduction by scanning over all possible users swapping as described in subsection 4.3.2. The proposed heuristic methods have a similar performance gap against the dual bound (≈ 13 %), which is larger than Papoutsis' method. However, they achieve this performance with a much lower computational complexity as listed in table 4.5. In addition, the proposed SUS-based heuristic method with rate-constrained power allocation, supports up to 20% larger minimum rates than the other two methods. As the number of RT users D increases, the difference A decreases since it is harder for the algorithm to find feasible points. Recall that we force all D user rate constraints to be active by using the procedure presented at the beginning of section 4.5 to set the minimum rates.

Table 4.7 lists the results when varying the number of users K from 8 to 32, other parameters remain fixed $N = 8$, $D = 4$, $M = 3$. The performance of the proposed methods improves as the number of users increase, as indicated by the difference between the upper dual bound and the sum rate attained (measurements F and G decrease from 14% to 7%). This is because in the presence of more users, the SUS algorithm is more likely to find semiorthogonal channel vectors, thus increasing the rates and effectively exploiting the multiuser diversity. In contrast, Papoutsis' method slightly deteriorates when the number of users increase (measurement E increases to 3.8 %).

Table 4.7 Average measurements for variable K

Measurement	$K = 8$	$K = 16$	$K = 24$	$K = 32$
Rate gap:				
E	2.5	3.0	3.8	3.7
F	14.0	7.8	7.7	6.9
G	14.2	9.0	8.7	7.7
Range gap:				
A	11.5	15.6	10.8	14.8
B	9.5	10.2	6.7	4.7

4.6 CPU Time

The results in tables 4.6 and 4.7 show that our algorithms provide sum rates that are not far from the optimal and that they increase the range of supported minimum rates. The numbers may not seem spectacular, we get up to 21 % performance gap against the upper bound, 10.7 % in average and the increase in the supported rates is 14.6% in average. Ideally, we would like to have smaller gaps against the upper bound and larger minimum rate support, but the important fact is that the subcarrier assignment and power allocation are obtained with much less computations than in previous methods. For example, algorithm 10 complexity reduction is three orders of magnitude lower with respect to Papoutsis' method for parameters $N = 64$, $K = 16$, $M = 4$. This results from comparing 1.7×10^7 vs. 1.6×10^4 in table 4.5. Algorithm 9 reduction is two orders of magnitude, resulting from comparing 1.7×10^7 vs. 0.7×10^5. Therefore, algorithms 9 and 10 are much less complex than Papoutsis' method.

We now turn to the question of which of the proposed algorithms 9 or 10 is faster for a particular set of problem parameters. The computational complexity measure in table 4.5 is an asymptotic one, it does not show how an algorithm will behave on a particular set of problem parameters. Therefore, we need to perform CPU measurements to see which algorithm is actually faster and for how much. First, we choose the parameters, or a range of them, and then vary one to study the CPU time taken by the algorithms.

One of the practical systems the algorithms in this dissertation can be applied to, is the single cell LTE-Advanced system we consider in chapter 5. For such system the permissible parameter values are (cf. table 5.3): the number of subchannels N ranges from 6 to 550 [3]; the number of antennas M belongs to the set $\{2, 4, 8\}$; the number of users K is preferably large but it is limited to keep the signaling overhead to manageable levels. In our numerical

3. For LTE systems, a subchannel or *Resource Block* comprises 12 contiguous subcarriers which are allocated jointly in a block.

evaluations we use a number of total users between 16 and 64, a number of RT users with minimum rate requirements between 1 and 32 and a fixed number of antennas $M = 4$. We vary the number of subchannels from 10 to 500 and make the rate constraints active by setting the rate constraint to 10% more of the maximum throughput rate. Finally, we repeat the experiment 100 times and measure the elapsed time. Elapsed time as measured by the Matlab function *toc* measures the time it takes the Matlab software to execute and it is an estimate of the CPU time.

Figure 4.10 illustrates the average elapsed time taken by each algorithm for the parameters listed under the figure's title, algorithm 9's performance is shown in circle markers and algorithm 10's one in cross markers. Algorithm 9 has larger computational complexity bound in table 4.5, however for a number of subcarriers lower than 80 in figure 4.10, it actually runs faster than algorithm 10. This is because when the number of subcarriers is low, the power allocation operations per subcarrier are less than those needed to invert a $M \times M$ subchannel matrix. Then, it is faster to perform rate-constrained power allocation than giving up on a subcarrier and trying to find a new one whose channel we will have to invert. Let's define NB as the number of subcarriers where the two curves in figure 4.10 intersect. For a number of subcarriers lower than NB, algorithm 9 is a better choice, it supports higher minimum rates through rate-constrained power allocation and it is faster. However, its elapsed time grows fast with the number of subcarriers, in contrast with algorithm 10 whose elapsed time is almost linear. Therefore, for a number of subcarriers higher than NB algorithm 10 is a better choice.

The number of subcarriers NB for which algorithm 9 is faster than algorithm 10 depends on the number of users among other parameters. As the number of users K increases, it becomes easier for algorithm 9 to find SDMA sets to satisfy the minimum rate requirements. This is illustrated in figures 4.11 and 4.12 for the same parameters as in figure 4.10, but for a number of users equal to 32 and 64 respectively. For $K = 32$ in figure 4.11, the elapsed time approximately doubles the elapsed time in figure 4.10 for both algorithms, i.e. it increases in the same proportion as the increase in the number of users. The number of subcarriers NB where both curves intersect moves from 80 to 270 indicating that the range in which algorithm 9 is faster than algorithm 10 increases. This is also true for $K = 64$ in figure 4.12, where the range in which algorithm 9 is faster than algorithm 10 covers the whole range of practical values. These results show that multiuser diversity improves the CPU performance of algorithm 9.

On the other hand, increasing the number of RT users has a positive effect on the performance of algorithm 10. Figure 4.13 illustrates this when the number of RT users increases from 1 in figure 4.11 to 16. The maximum elapsed time decreases from 0.8 to 0.6 approxi-

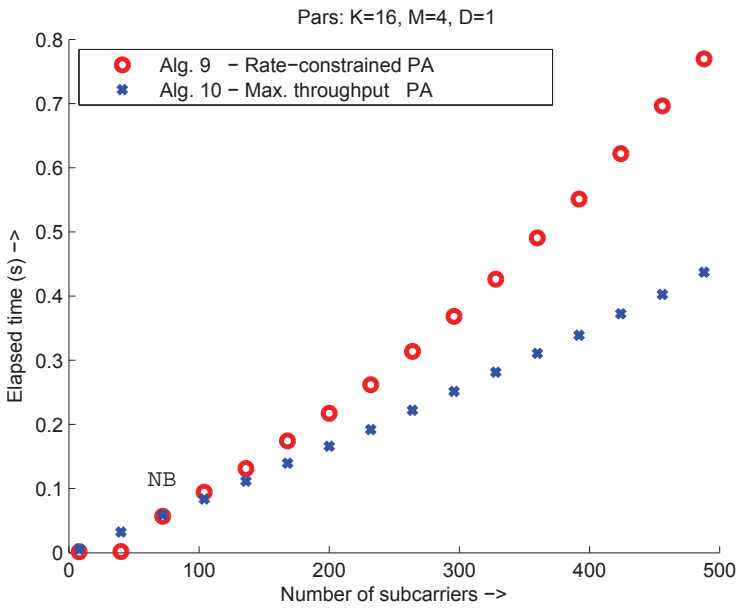


Figure 4.10 Elapsed time of proposed algorithms vs. N for K=16

mately and the shape of algorithm 9's elapsed time also changes, reducing NB from 270 to 170.

The results in this section show that when consider speed only, there is no uniform better choice between algorithm 9 and 10. The quickest algorithm choice depends on the value of the problem parameters. The speed difference is not large, for the parameters used the quickest algorithm is roughly twice faster than the other one. For this reason, the minimum rates range and the closeness to the optimal are more significant selection criteria, which definitely favors algorithm 9.

4.7 Chapter Conclusion

The work collected in this chapter is the result of the search for an efficient method to solve the ZF-constrained optimization problem (3.1–3.7) supporting minimum rate requirements. We found ways to reduce the computational complexity of the two main tasks necessary to solve the problem: power allocation and subcarrier assignment. For the power allocation, we exploited the fact that finding feasible points satisfying the rate constraints with inequality is quicker than finding exact solution points. Thus, we devised a method that satisfies the rate constraints with very few iterations. On the other hand, we looked for points that satisfy the constraint exactly for the power constraint because power cannot be exceeded without creating excessive interference in neighboring cells and there is a quick method to find the

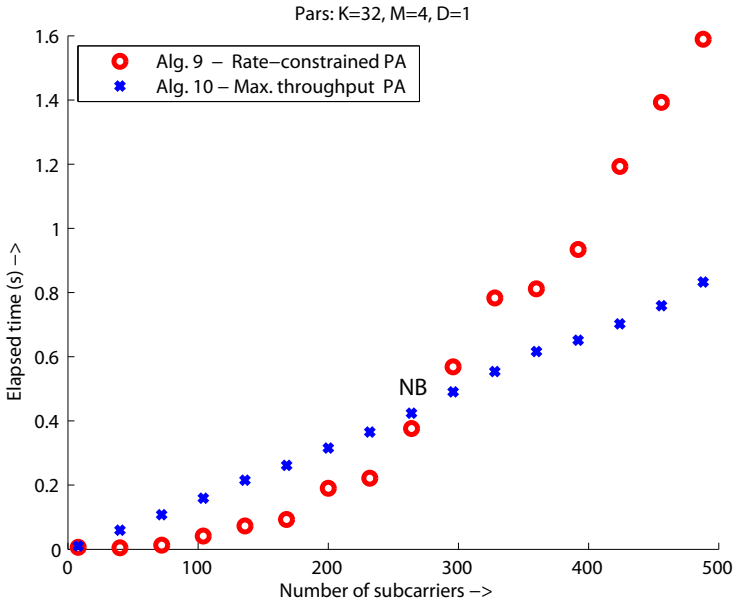


Figure 4.11 a) Elapsed time of proposed algorithms vs. N for K=32

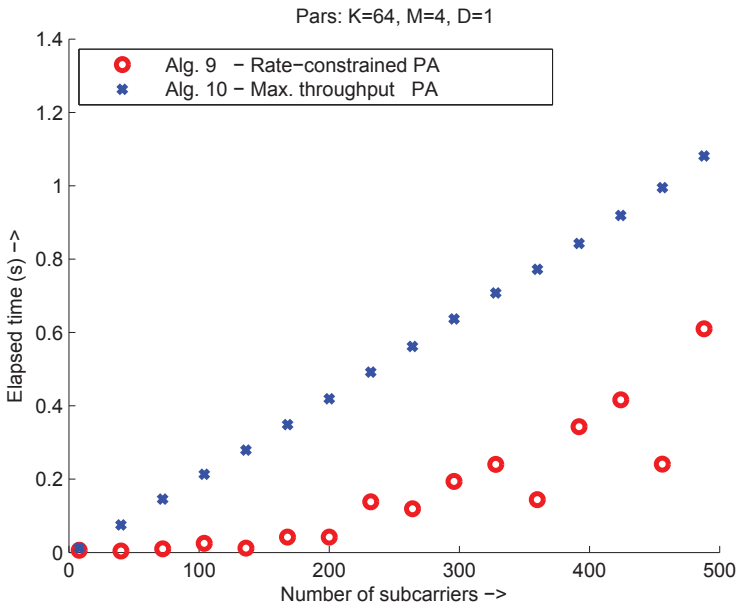


Figure 4.12 Elapsed time of proposed algorithms vs. N for K=64

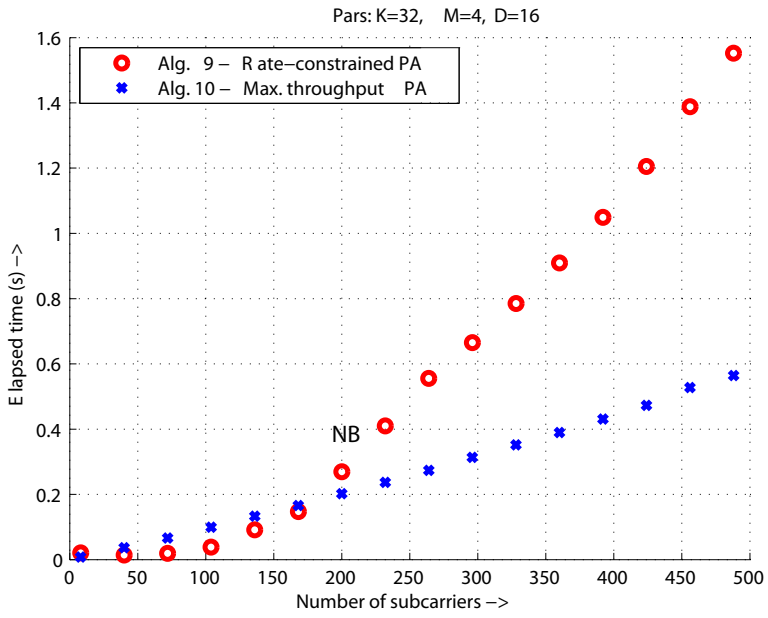


Figure 4.13 Elapsed time of proposed algorithms vs. N for K=32, D=16

exact point. In contrast, if rates are overly assigned the rate requirements in the next frame would be reduced. We found through numerical evaluations that the number of iterations required for the modified power water-filling method is very low because of the user selection process previously performed that selects users with good channel conditions.

For the subcarrier assignment task, we made use of the SUS algorithm that has been extensively researched in the literature. We adapted it for the minimum rate constraints problem by prioritizing the users search. The combination of the proposed algorithms for user selection and power allocation constitutes the proposed heuristic method. We showed through numerical evaluations that it has a performance not far from the optimal solution and that it increases the range of supported minimum rates when compared with other approaches. This is an important result because, in a system with RT users, it is more important to satisfy the rate constraints of the users in need than increasing the rates of the nRT users. Compared with the method proposed in [22], our method does not follow the upper bound as closely but it increases the range of the minimum rates supported. This and the fact that we have reduced the complexity by several orders of magnitude are the main advantages of the proposed method.

CHAPTER 5

APPLICATION TO LTE-ADVANCED SYSTEMS

5.1 Introduction

In this chapter, we study the application of the algorithms proposed in this dissertation to a single-cell system using the current LTE-Advanced technology. LTE is an evolving standard, its most recent version, LTE Advanced — Release 10 [3] — is aimed to fulfil the requirements set by the International Telecommunication Union (ITU) for International Mobile Telecommunications (IMT)-Advanced:

- Increase downlink peak data rate up to 3 Gbps and uplink up to 1.5 Gbps
- Obtain higher spectral efficiency, from a maximum of 16 bps/Hz in release 8 to 30 bps/Hz in release 10
- Increase the number of simultaneously active subscribers
- Improve the performance at cell edges.

The main functionalities introduced in release 10 of LTE-Advanced to meet these requirements are: Carrier Aggregation (CA), enhanced use of multi-antenna techniques and support for relay nodes. Using CA, the available bandwidth can be augmented up to 100 MHz by grouping different parts of the spectrum. On the other hand, MIMO techniques are used to increase the spectral efficiency through transmission of two or more different data streams on two or more different antennas using the same resources in both frequency and time. Finally, the introduction of relay nodes (low power base stations) provides enhanced coverage and capacity at cell edges and can also be used to connect to remote areas without a wired connection.

Algorithms 9 and 10 proposed in chapter 4 play a role in the improvement of the multi-antenna techniques. Particularly, they provide an efficient way to perform Resource Allocation (RA) in an LTE-Advanced system with multiple antennas at the Base Station (BS) and User Equipment (UE) terminals with a single antenna. However, the algorithm's design was done for a general MISO-OFDMA system, thus the design needs to be adapted to the LTE system architecture. To illustrate where the RA algorithm fits in the LTE architecture, we present a general description in the following.

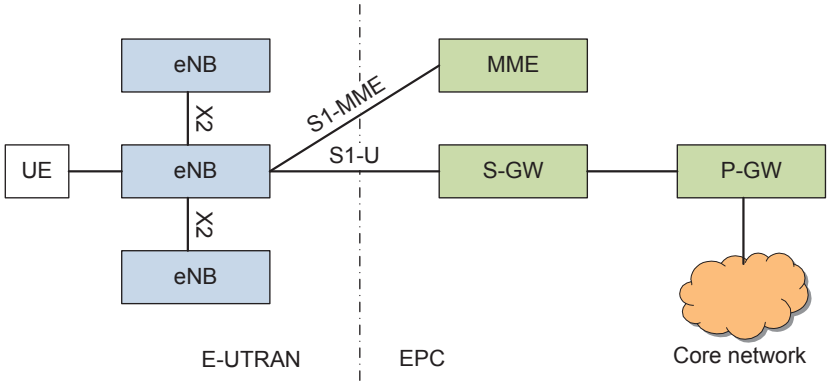


Figure 5.1 LTE General Architecture

5.1.1 LTE General Architecture

Figure 5.1 shows a diagram of the LTE architecture. This consists of two main parts: on the left side of the diagram, the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN), and the Evolved Packet Core (EPC) on the right side. The E-UTRAN consists of the enhanced NodeBs (eNB) or base stations. The EPC interconnects the E-UTRAN to the core network. As illustrated in figure 5.1, the EPC network main elements are [46]:

Mobility Management Entity (MME): The MME is the control node that processes the signaling between the UE and the core network. The main functions supported by the MME can be classified in two: functions related to bearer management including the establishment, maintenance and release of the bearers, and functions related to connection management which includes the establishment of the connection and security between the network and UE.

Serving Gateway (S-GW): All user IP packets are transferred through the serving gateway, which serves as the local mobility anchor for the data bearers when the UE moves between eNBs. It also retains the information about the bearers when the UE is in the idle state and temporarily buffers downlink data while the MME initiates paging of the UE to re-establish the bearers.

P-GW: The Packet Data Network (PDN) Gateway is responsible for IP address allocation for the UE, QoS enforcement and flow-based charging. It is also responsible for the filtering of downlink user IP packets into the different QoS-based bearers. For example, the P-GW performs QoS enforcement for guaranteed bit rate bearers.

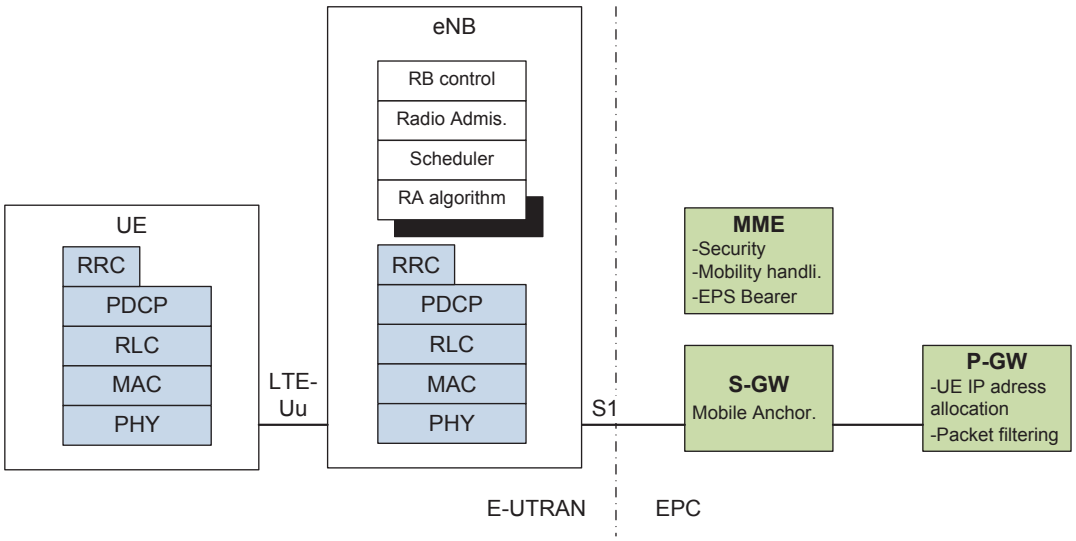


Figure 5.2 LTE Protocol stack

5.1.2 The eNB functions and the role of RA algorithms

The access network of LTE, E-UTRAN, simply consists of a network of eNBs as illustrated in figure 5.1. The eNBs are interconnected to the MME by means of the S1-MME interface and to the S-GW by the S1-U interface. The eNBs connect to the UE by the LTE-Uu air interface, and are interconnected with each other by the X2 interface to facilitate hand-over and interference cancellation. The eNBs have control on the resources for downlink and uplink transmission of the users attached to them.

Figure 5.2 illustrates the protocol layer stack to communicate with the UE. The RA algorithm resides in the eNB's protocol layer stack, below the scheduler and above the Packet Data Convergence Protocol (PDCP) layer. The scheduler receives QoS requests and Channel State Information (CSI) from the UE and determines which RT users to serve at each TS and with how much rate. The RA algorithm finds the resources — beamforming vectors, power and subcarriers — to provide the required minimum rates to the selected RT users. The QoS functions of the Radio Resource Control (RRC) layer interact with the scheduler and RA algorithm. In the previous chapters, we formulated and solved the RA problem without making any assumption on the system parameter values, like the number of antennas, subchannels, etc. In practice, these are restricted to the values agreed upon in LTE-Advanced. After briefly introducing some elements of the LTE-Advanced technology in section 5.2, we describe the permissible values of these parameters in section 5.3.

An assumption made in previous chapters is that of perfect CSI; we have assumed that the Base Station (BS) has access to the channel information of all users at all times and with

infinite precision. This is, of course, impractical in a real system because it would require a large amount of uplink bandwidth to feedback the CSI from the users to the BS. In contrast, the LTE-Advanced design philosophy has been to use the minimum uplink bandwidth to feedback CSI. For this purpose, the CSI information is transmitted implicitly and not explicitly in the current standard.

In order to effectively reduce interuser interference, Zero Forcing (ZF) techniques in Spatial Division Multiple Access (SDMA) systems require an explicit and accurate representation of the channel vectors, which is currently not specified in the standard. In section 5.4, we assume that such explicit channel representation is available at the BS and outline a general procedure that invokes the proposed RA algorithms to perform downlink transmission in a single-cell LTE-Advanced system.

Another assumption made in this dissertation is that of continuous rates. The rate expression in (2.3) corresponds to Shannon channel capacity, which is a continuous concave function. In LTE-Advanced systems the rates are discrete, corresponding to specific modulation schemes and coding rates which are selected according to the current SNR [47]. The envelope of the stepwise discrete rates is a concave function; we have assumed that the quantization step is small, but in reality there is power waste since increasing the power does not give higher rates until the following level is achieved in the stepwise function. For the SISO case, in [48] the discrete rate RA problem was formulated as a integer program, an optimal solution is found but the computational complexity of this method is not compared with the continuous case. In [49], a heuristic efficient solution is proposed for the discrete rate case.

In previous work, LTE-Advanced parameters have been used to evaluate the performance of RA algorithms. For instance, in [50] Chung et al. proposed an RA scheme for LTE downlink transmission which assigns priorities to the users by means of a fuzzy inference system. It heuristically solves the rate maximization problem subject to per subchannel power, delay and minimum rate constraints. The fairness index and the packet drop ratio is compared against other schemes. However, these works do not consider realistic CSI feedback mechanisms.

In [51], MIMO single user (SU) and multiuser (MU) LTE-Advanced transmission modes are compared, where it is found that MU schemes can largely improve throughput but they are more sensitive to CSI accuracy. An scheme that dynamically switches between SU and MU is then proposed.

The performance of several CSI feedback mechanism have been studied in [52] and references therein; the application of quantized feedback to LTE-Advanced has been evaluated in 3GPP recommendations [53] and [54]. The results reported in the literature show that there is a trade-off between improving downlink performance and increasing uplink overhead

for CSI feedback. Currently, LTE-Advanced only specifies an implicit CSI feedback which is not sufficient for interference cancelling techniques. However, proprietary solutions can be implemented sending explicit CSI as data packets in the uplink direction.

In [29], RA for an LTE system is considered supporting both RT and nRT traffic. SDMA beamforming is not considered because only LTE Rel. 8 transmission modes are used. A comparison between different heuristic methods is performed but no comparison against a near-optimal solution is provided.

5.1.3 Chapter Objective

In this chapter, we describe how the algorithms proposed in this dissertation can be adapted to LTE-Advanced. We describe in detail how each problem parameter can be mapped to LTE-Advanced specifications and present some numerical results in subsection 5.3.4. We present a general procedure in section 5.4, where we assume that explicit CSI feedback is available to the BS and we do not specify the method nor the CSI resolution.

5.2 Downlink Transmission Mechanisms in LTE-Advanced

LTE systems use time-slotted OFDMA in their physical layer. Time-slotted OFDMA systems partition the time-frequency space forming a grid of Resource Elements (RE). Each RE consists of 1 subcarrier and 1 OFDM symbol. In LTE systems, a Resource Block (RB) is a collection of 12 contiguous subcarriers in a 0.5 ms time slot in the time domain, corresponding to 7 ODFM symbols for the case of normal cycle prefixing. Figure 5.3 illustrates this time-frequency grid. An RB-pair is the minimum amount of resources that can be allocated for transmission; it consists of 2 time-contiguous RB, i.e., a 1 ms subframe.

The LTE physical-layer specification allows for a carrier to consist of a number of resource blocks in the frequency domain ranging from a minimum of six resource blocks up to a maximum of 110 resource blocks. This corresponds to an overall transmission bandwidth ranging from roughly 1 MHz up to 20 MHz with very fine granularity and thus allows for a very high degree of LTE bandwidth flexibility. However, LTE radio-frequency requirements are only specified for a limited set of transmission bandwidths, corresponding to a limited set of possible values for the number of resource blocks within a carrier [55]. In addition, LTE-Advanced supports carrier aggregation to increase the system bandwidth.

User data is included in a Downlink Shared Channel (DLSCH) message and written into the assigned RB for downlink transmission. There are several Transmission Modes (TM) in LTE-Advanced depending on the number of transmit antenna and transmission method used. Table 5.1 lists the possible modes. We focus on TM modes 8 and 9 which support SDMA.

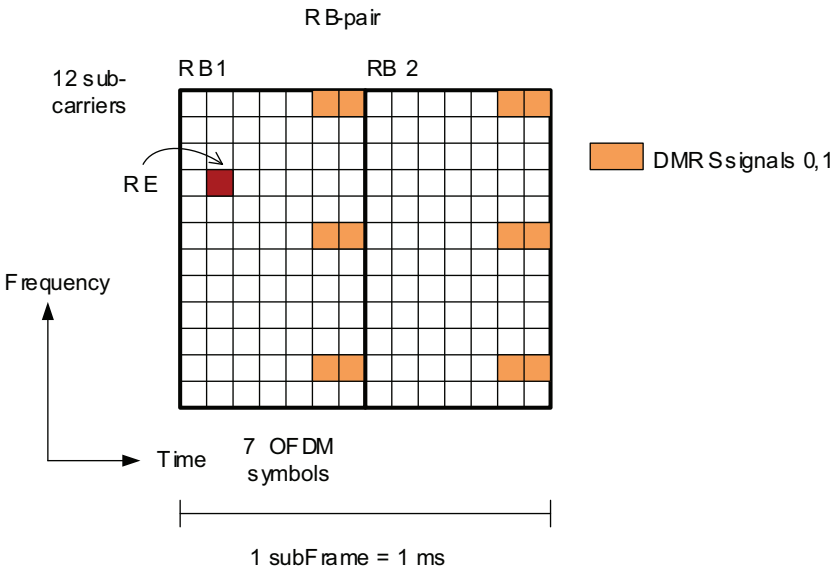


Figure 5.3 LTE time-frequency grid

Table 5.1 LTE-Advanced Transmission modes

TM	Description
1	Single-antenna transmission
2	Transmit diversity
3	Open-loop Codebook-Based Precoding (CBP)
4	Closed-loop CBP
5	Multi-user-MIMO version of transmission mode 4
6	Special case of closed-loop CBP limited to single layer transmission
7	Release-8 non-CBP supporting only single-layer transmission
8	Release-9 non-CBP supporting up to two layers
9	Release-10 non-CBP supporting up to eight layers

Antenna mapping consists of processing the modulation symbols corresponding to the one or two transport blocks and writing the result to different antenna ports. The antenna mapping can be configured in different ways corresponding to different multi-antenna transmission schemes, including transmit diversity, beam-forming, and spatial multiplexing. The input to the antenna mapping thus consists of the modulation symbols (QPSK, 16QAM, 64QAM) corresponding to transport blocks. The symbols of each antenna port are subsequently applied to the OFDM modulator; that is, they are mapped to the basic OFDM time-frequency grid corresponding to that antenna port.

For TM 8 and 9, linear precoding is applied by multiplying the input of the antenna mapping by a beamforming matrix. For the terminal user receivers to properly detect the signal, they need reference signals that have been multiplied by the same precoding matrix,

so that the reference signals are representative of the effective channels. These are named Demodulation Reference Signals (DMRS) and written in separate RE elements in the RB assigned to each user terminal. Figure 5.3 illustrates the location of the DMRS RE, when 8 signals are required (8 antenna ports), the signals are located every 4 subcarriers along the frequency axis, and every subframe along the time axis.

5.3 Parameters Correspondence with LTE-Advanced Systems

The input parameters of algorithms 9 and 10 (presented in chapter 4) to solve problem (3.1–3.7) can be categorized in three groups as listed below

LTE system problem parameters:

M Number of antennas at the BS.

N Number of subchannels available.

$\check{P}$ Total power available at the base station for transmitting over all subchannels.

Scheduling problem parameters:

K Number of users in the cell.

$\check{d}_k$ Minimum rate requirement for user k.

c_k Weight of the user rates in the objective function. These could be computed by the scheduler to implement prioritization or fairness.

CSI feedback problem parameters:

$\mathbf{h}_{k,n}$ the M-component row vector representing the channel between the M antennas at the BS and the receive antenna at user k for each subcarrier n.

In the sequel we describe how these parameters are mapped from the RA problem formulated in chapter 2 to an LTE-Advanced system. The CSI feedback parameter is studied separately in section 5.4.

5.3.1 LTE system problem parameters

Number of transmit antennas M

In LTE-Advanced systems the number of transmit antennas at the BS belongs to the set $\{1, 2, 4, 8\}$. We are interested in the case where the number of transmit antennas is 4 or 8 because these are the configurations where SDMA techniques can be used (i.e. transmission modes 8 and 9). We also restrict ourselves to the case of single antenna terminals.

For ZF beamforming we require one DMRS reference signal per transmitted layer to write the beamforming vector. LTE-Advanced release 10 offers up to 8 DMRS reference signals per

Table 5.2 LTE-Advanced bandwidth and number of resource blocks

Number of resource blocks - N	Bandwidth (KHz)
No carrier aggregation	
Min. 6 Max. 110	1080 19800
With carrier aggregation	
Min. $2 \times 6 = 12$ Max. $5 \times 110 = 550$	2160 99000

cell. Since these DMRS signals are transmitted only on the RBs assigned to the terminal, each terminal can have its own set of DMRS reference signals. Therefore, the first parameter M in subsection 5.3 belongs to the set $\{4, 8\}$ for ZF operation.

Number of subchannels N

In LTE, resource allocation is performed at the RB-pair level. We assign all REs in a RB block to a particular user. Each RB-pair comprises 12 subcarriers (c.f. subsection 5.2). Therefore, the term *subchannel* employed in previous chapters corresponds to the grouping of 12 subcarriers in an RB-pair. The number of RBs available depends on the assigned bandwidth and it ranges from 6 to 110 without considering carrier aggregation. Carrier aggregation consists of grouping several parts of the spectrum located in one or several bands, each spectrum part can have different bandwidth allowing more flexibility for operators to allocate the spectrum and increase the system bandwidth. A maximum of 5 component carriers can be used in the same band or across bands. Table 5.2 lists examples of different bandwidths and number of component carriers. The problem parameter N equals the number of RB available. Thus, it ranges from 6 to 550.

Total power constraint $\check{P}$

In the problem formulation (3.1–3.7) we considered a total power constraint on the beamforming vectors Eq. (3.2) for each time subframe (14 OFDM symbols). It is assumed that the data symbols energy is normalized, thus the transmitted signal energy depends only on the beamforming vectors. Translating this total power constraint to the OFDM time-frequency grid of figure 5.3 means that the sum energy of all RBs assigned to users for transmission, across subcarriers and across layers should translate into a transmitted signal power equal to $\check{P}$. Assume the mapping from the signal level to the actual transmitted signal power is given by a factor f, the same factor must be applied across all layers to preserve the power constraint.

Problem (3.1–3.7) can also be formulated using per antenna power constraints instead of total power constraint [37]. This would require a different solution method, but it would

Table 5.3 LTE system problem parameters

Problem parameter	LTE system permissible values
Number of transmit antennas M	$\{4, 8\}$
Number of subchannels N	$[6 - 550]$
Total power constraint $\check{P}$	System dependant

provide the advantage of reducing the dynamic range of the transmitted signals which eases the power amplifier design. Notice that algorithm 10 in chapter 4 performs power allocation considering power constraints per subchannel instead of total power constraint. This means that the sum energy of all RBs assigned to users for transmission, across layers but *not* across subcarriers, must be equal to $\check{P}/N$.

As a summary, table 5.3 lists the permissible values for the LTE system problem parameters.

5.3.2 Scheduling Parameters

The packet scheduler operates in the eNB in a layer above the PDCP layer as illustrated in figure 5.2 (c.f. subsection 5.1.2). It schedules users according to the the traffic conditions, queues state and user priorities. Problem parameters $K, \check{d}_k, c_k$ are selected by the scheduler and passed to the RA algorithm.

In general, the number of users K is limited by the amount of CSI feedback data the system is able to support. Ideally, one would like to receive explicit CSI from *all* terminals to select the best users for each subchannel as this would allow us to fully exploit multiuser diversity. Another possibility is to balance multiuser diversity and the amount of feedback data by receiving *implicit* feedback data from all terminals, which requires less bandwidth. Once a set of users has been identified as having good channel conditions, *explicit* feedback data can be requested from these terminals only. In this way, the number of supported users K can be increased with less impact on the amount of feedback data needed. RT users would have to send explicit feedback data at all times.

The number of users scheduled per subcarrier is limited by the number of antennas M, but the total number of users with minimum rate requirements D is limited by the number of subchannels N, i.e. $D \leq N$. On the other hand, given minimum required rates per frame for each RT user, the scheduler distributes these rate requirements among subframes determining the $\check{d}_k$ for each subframe. The distribution of the minimum rate requirements is part of the scheduler design and is not addressed here.

The user weight factors $\{c_k\}$ are used to implement priorities among users and are determined by the scheduler as done in [18, 19].

5.3.3 Mapping Algorithms 9 and 10 Output to LTE Parameters

The output of Algorithms 9 and 10 in chapter 4 are the feasible values of the optimization variables $\boldsymbol{\alpha_{k,n}}$ and $\mathbf{w}_{k,n}$ which would be used by the LTE-Advanced system to make the RA decisions.

Subchannel assignment $\boldsymbol{\alpha_{k,n}}$: Determines the users assigned to each subchannel n. For efficient signaling of the RA from the base station to the terminal, different RA types are supported in LTE. In RA type 2, for instance, a starting resource block and an allocated number of resource blocks are signaled to the terminal. In order to save signaling bits on the Physical Downlink Control Channel (PDCCH), these two parameters are not explicitly signaled. Instead, a Resource Indication Value (RIV) is derived, which is signaled in the downlink control information.

Beamforming vectors $\mathbf{w}_{k,n}$: The data destined to selected users is multiplied by the beamforming vectors and written in the corresponding RBs of the time-frequency grid. The precoding vector $\mathbf{w}_{k,n}$ is also applied to the pilot signals and written to the DMRS signal in the RB assigned to user k. This non codebook-based precoding mechanism is explained in subsection 5.4.3.

5.3.4 RA Results Using LTE Parameters

Figures 5.4 and 5.5 illustrate the weighted sum rate achieved and the CPU elapsed time for the LTE parameters listed in table 5.4. The number of RT users varies from 1 to 12 and the minimum rates $\check{d}_k$ are set to 10% over the rates obtained when solving the problem without rate constraints. This way, we ensure that the rate constraints of the RT users are active.

In this numerical evaluation we assume perfect CSI feedback and use algorithm 1 to obtain the upper dual bound and the heuristic algorithms 9 and 10 to perform practical RA. We compare these results against Papoutsis' method (c.f. section 4.3.2) and find that the heuristic algorithms 9 and 10 achieve a sum rate lower than the one achieved by Papoutsis' method. However, they support a wider range of RT users. In figure 5.4 Papoutsis' method supports up to 8 RT users, for a higher number of RT users Papoutsis' method is not able to find feasible points. In contrast, the heuristic algorithm 9 supports up to 11 RT users. Other configurations of LTE parameters show the same trend: Papoutsis' method achieves a sum rate closer to the upper bound than the proposed heuristics, but it supports a narrower

Table 5.4 LTE-Advanced simulation parameters

K	Number of users	16
$\check{P}$	Total power constraint	20
N	Number of subcarriers	16
M	Number of antennas	4
$\{c_k\}$	users weight	1

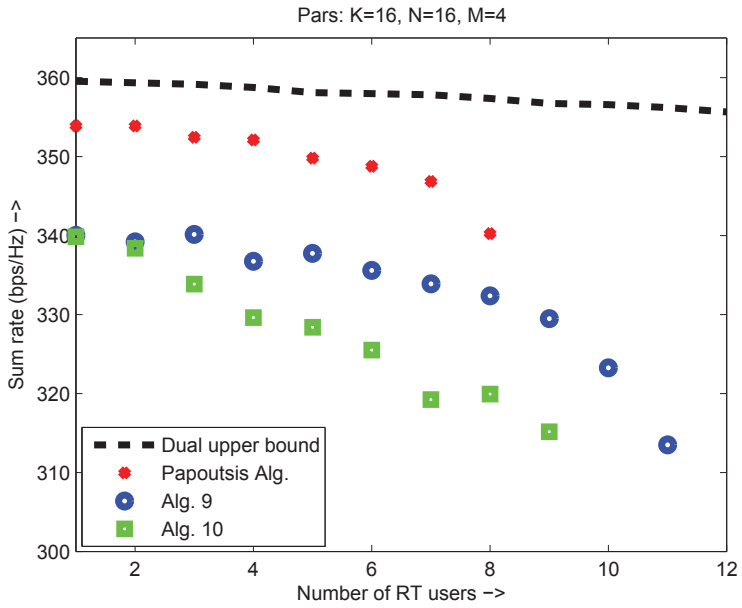


Figure 5.4 Algorithms performance vs. the number of RT users

range of RT users. In addition, the proposed heuristics are faster than Papoutsis' method as shown in figure 5.5 for the same LTE parameters. Papoutsis' method and algorithm 10 elapsed time grows linearly with the number of RT users, but Papoutsis' method has a much larger slope. On the other hand, algorithm 9 elapsed time grows non-linearly. For a number of RT users lower than 10, algorithm 9 is faster than algorithm 10, and for a higher number of RT users is slower.

5.4 CSI Feedback in LTE-Advanced Systems

As mentioned above, the design philosophy in LTE has been to reduce the amount of feedback data and concurrently to provide a sufficient representation of the channel state. In this section, after briefly presenting codebook-based precoding which uses a low overhead implicit CSI feedback, we present a general procedure for downlink ZF beamforming in an

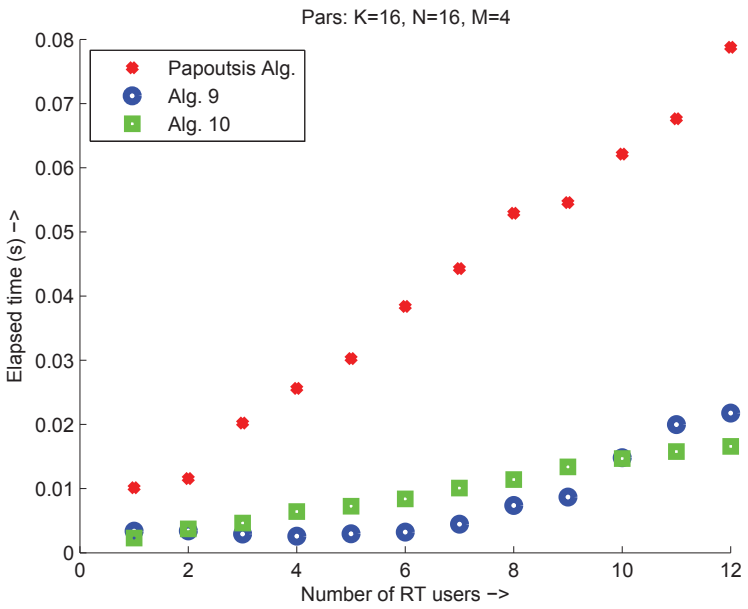


Figure 5.5 Elapsed time vs. the number of RT users

LTE-Advanced system using explicit CSI feedback. This procedure invokes the RA heuristic algorithms proposed in chapter 4.

5.4.1 Codebook-based Precoding

LTE-Advanced codebook-based precoding provides a good trade-off between the amount of feedback data and the channel state resolution for MIMO single user transmission. Another advantage of codebook-based precoding is that it distributes the computation of the precoding matrices among terminals, alleviating the BS of this task which can overload the BS when the number of terminals increase. In codebook-based precoding, the CSI representation is implicit, i.e., codebook-based precoding does not provide the BS with an explicit representation of the channel vectors $\{\mathbf{h_{k,n}}\}$. Instead, an index of the best precoding matrix, its rank and a Channel Quality Indicator (CQI) is computed by the user terminal and sent back to the BS [3]. This corresponds to TM 4,5 and 6 in table 5.1.

5.4.2 Explicit CSI Feedback Assumption

MIMO multi-user transmission requires finer codebook granularity than that of current implicit CSI-feedback in LTE-Advanced as reported in [51]. This is because a finer CSI representation helps to improve the performance of interference cancellation. In practice, a trade-off between improving downlink transmission performance and increasing uplink over-

head needs to be met. In LTE-Advanced release 10, increasing the feedback resolution was discussed in [56], but no specific technique was proposed; it is expected that this will be part of future releases.

Based on this expectation, the general procedure outlined next assumes explicit CSI at the BS, but we do *not* specify the method to obtain such information. One possibility could be to request CSI explicit information from pre-selected user terminals, which would be sent back to the BS as part of the uplink data channel. To reduce side information, not all users are required to send back explicit CSI, but only those with good channel conditions or high priority. The others can simply send the implicit CSI back to the BS, which is enough to monitor their channel SU quality.

5.4.3 Non Codebook-based Precoding

For transmission mode numbers 8 and 9, LTE-Advanced uses non codebook-based precoding. This allows us to use arbitrary precoding matrices $\mathbf{W}_n$, instead of the predefined codebook matrices used in the other transmission modes, and therefore to implement ZF interference cancellation among the scheduled users. Assuming we have an estimate of the channel matrix $\hat{\mathbf{H}}_\mathbf{n}$ at the BS and that we have computed the user allocated power vector $\mathbf{q}_n$, the ZF precoding matrix is given by (cf. section 4.2, Eq. (4.7))

$$\mathbf{W_n} = \hat{\mathbf{H}}_\mathbf{n}^\dagger \, \mathrm{diag}(\sqrt{\mathbf{q}_n}), \quad \forall n, \tag{5.1}$$

where $\hat{\mathbf{H}}_\mathbf{n}^\dagger$ is the pseudo-inverse of the estimated channel matrix $\hat{\mathbf{H}}_\mathbf{n}$.

Grouping the downlink processed signals [1] corresponding to several users in a $1 \times g_n$ vector $\tilde{\mathbf{s}}_\mathbf{n}$, where g_n is the number of users scheduled on subchannel n, the signal to write into the RBs on each layer after ZF precoding is given by the product $\mathbf{W_n}\tilde{\mathbf{s}}_\mathbf{n}$. This signal is then written to the time-frequency OFDM grid of each port.

In addition, we perform ZF precoding on the pilot DMRS signal $\mathbf{s}_\mathbf{n}^\mathbf{P}$, using the same precoding matrix, obtaining the product $\mathbf{W_n}\mathbf{s}_\mathbf{n}^\mathbf{P}$. After ZF precoding the signal is written to the RB assigned to each user, more specifically it is written to the DMRS RE location illustrated in figure 5.3, cf. subsection 5.2.

At the user terminals, the received data signal is

$$\mathbf{y}_1^{(n)} = \mathbf{H_n}\mathbf{W_n}\tilde{\mathbf{s}}_\mathbf{n} + \mathbf{z}_1 \quad \forall n, \tag{5.2}$$

1. Downlink processing consists of CRC insertion, channel coding, rate matching, PHY hybrid ARQ functionality, bit-level scrambling and Data modulation — see chap. 10 in [57].

and the received pilot signal is

$$\mathbf{y}_2^{(n)} = \mathbf{H_n W_n s_n^P} + \mathbf{z}_2 \quad \forall n, \tag{5.3}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are the AWGN noise vectors at the receivers.

Notice that for each user both signals (5.2) and (5.3) experience the same channel and are precoded by the same matrix. The symbols are decoded independently by each receiver since there is no cooperation among receivers.

To summarize how the proposed algorithms 9 and 10 can be applied to an LTE-Advanced system, we describe the following general procedure to transmit downlink data using explicit CSI feedback

1. Perform Downlink processing on the transport blocks
2. Obtain CSI estimates, $\hat{\mathbf{H}}_\mathbf{n}$ from feedback channel
3. Invoke algorithm 9 or 10 in chapter 4 to perform RA with the input and output parameters specified in subsections 5.3 and 5.3.3.
4. Signal the allocated PRs to each user according to the obtained $\{\alpha_{k,n}\}$
5. Multiply DMRS pilot signals by precoding matrix (5.1) and write to corresponding REs
6. Multiply DL processed signal $\tilde{\mathbf{s}}_\mathbf{n}$ by precoding matrix (5.1) and write to corresponding REs
7. Transmit.

The procedure above is repeated for every subframe. The user terminals detect the user symbols using a single user receiver and the channel estimated by decoding the received DMRS.

5.5 Chapter Conclusion

In this chapter we illustrated how the algorithms proposed in this dissertation can be used in a single-cell LTE-Advanced system. We showed the parameter mapping between the proposed algorithms and the parameters in such system. Using the general outlined procedure, the proposed algorithms can effectively be used to solve the RA problem.

We assumed that an estimate of the CSI is available at the BS and it is used as an input to the proposed algorithms. This, however, needs to be further investigated since the LTE-Advanced design philosophy has been to use the minimum of uplink bandwidth to feedback CSI thus using implicit CSI. Therefore, further research is needed to evaluate the trade-off between the amount of side information and the performance benefits obtained by

feeding back explicit CSI in the uplink data channel. In addition, we need to investigate RA algorithms that use implicit instead of explicit CSI.

CHAPTER 6

CONCLUSION

In this dissertation we have used optimization techniques in the design of RA algorithms for 4G wireless access networks. We designed several algorithms to provide solution points to the RA problem for MISO-OFDMA systems supporting minimum rate requirements. The solution points given by these algorithms differ in their distance to the optimal solution and in the computational complexity to obtain them. The first designed algorithm — Alg. 2 — uses the dual Lagrange method and employs a simple heuristic that searches around the dual optimal solution to obtain a feasible point. This method gives us the best solution point among all the methods proposed in this dissertation. In addition, the dual formulation gives us a relation between the problem feasibility and the minimum rate constraints through the shape of the dual function.

Using numerical evaluations, we compared the solution given by algorithm 2 and a direct method that performs enumeration on the binary variable for small problem sizes. We observed that the difference is small for the cases when algorithm 2 is able to find feasible points. We also observed that the method is not able to find solution points when the minimum rate requirements are close to the feasibility boundary. For larger problem sizes we could not use the direct method, thus we relied on other method to obtain limits on the duality gap: we computed the difference between the upper bound given by Alg. 1 and the feasible point obtained by algorithm 2. We found that the duality gap is within 3.5 % for the cases where algorithm 2 is able to find solution points and thus, within the 3.5 % tolerance, we can use the dual upper bound as a benchmark for more efficient heuristic methods.

The computational complexity of algorithm 2 is large because it performs enumeration over all SDMA sets and needs to compute all pseudo-inverse matrices, making the algorithm practical only for off-line processing. To provide practical algorithms that we can implement in real-time systems, we designed two heuristic methods: algorithms 9 and 10. They select an SDMA set for each subchannel and then solve a power allocation problem. The algorithms stop if the rates are feasible, otherwise, they search a new subchannel assignment to provide the RT users in need with more subchannels. The criteria to select SDMA sets for each subchannel is based on the well-known SUS algorithm, but adapted to the minimum rate constraints case. Power allocation is performed using an innovative approximation method that gives near-optimal solution points with much fewer computations. The difference be-

tween algorithms 9 and 10 is that the latter one considers power constraints per subchannel, providing a smaller minimum rates support but higher computational efficiency.

As opposed to algorithm 2, algorithms 9 and 10 can be implemented in a real-time system. We studied the application of these algorithms in the case of a single cell using LTE-Advanced technology, where the BS is equipped with multiple antennas and transmits downlink to single antenna users. We showed that the input and output algorithm parameters can be mapped to LTE-Advanced parameters. This mapping also gives us a practical range of the input values to the algorithm parameters, which is useful for algorithm evaluation. There is, however, a basic assumption in our algorithms that differs from the LTE-Advanced design philosophy. That is, we assume that explicit CSI is available at the BS. In contrast, in LTE-Advanced such information is not available at the BS and has to be inferred from the implicit CSI feedback or estimated by the BS. Specific methods to obtain this CSI are not indicated in the standard.

6.1 Future Work

6.1.1 CSI Feedback Aspect

The current release of LTE-Advanced (release 10) can support SMDA ZF beamforming by using the method explained in chapter 5. This method, however, uses an explicit CSI feedback mechanism that is *not* part of the current standard. Therefore, if one implements this method, the BS has to estimate the CSI information from the transmitted uplink reference signals. This, however, can overload the BS as the number of users increase. It is expected that new releases will enhance the CSI feedback mechanism to provide the resolution needed for SDMA ZF beamforming without excessively increasing the uplink feedback bandwidth. One possible direction is that of CSI compression, where redundancy of the time-correlated CSI information is eliminated by adequate compression techniques. Other possibility is to provide CSI feedback at several resolution levels. In this approach, all users would feedback implicit CSI, as in the current release, which requires low bandwidth but has poor resolution. This information would suffice to pre-select users that are approximately semi-orthogonal to each other and have large vector norms. Then, the pre-selected users would be requested to feedback more precise CSI using adaptive codebooks, and the BS would make the final user selection and ZF beamforming vector computation based on this high resolution CSI. Using a combination of these two levels of CSI resolution, we can find a balance between the required uplink feedback bandwidth and the achieved performance. Evaluating theses schemes, designing the adaptive codebooks and studying the implications of CSI delay on the performance are important parts of required future work.

On the other hand, the RA algorithms can be re-designed so that they do not require explicit CSI as input parameter, but only the precoding matrix and channel quality indexes, PMI and CQI, of adaptive codebooks. In this method, the BS would form the SDMA grouping, by searching users that belong to *compatible* groups in the vector space. The BS would have pre-computed the beamforming vectors based on the highest resolution received from the selected users. This could lead to more computationally efficient algorithms because the vector search space is gradually reduced, and it would eliminate the need of estimating explicit CSI. The effect of discrete rates needs also to be studied using the discrete rates provided in the current standards LTE and WiMAX.

6.1.2 Multi-cell Interference Extension

The problem formulated and solved in this dissertation considers only a single cell. We did not address the interference coming from other cells using the same channel bandwidth. Systems where the neighboring cells share the same channel bandwidth are important because spectrum re-use is necessary in dense areas. In addition, cell-edge users can greatly benefit from inter-cell interference reducing techniques. One avenue of future research is to extend the concepts and methods used in this dissertation to the multi-cell scenario. This scenario has already been considered by several researchers. However, from the work in [25] — reviewed in section 1.3.3, a number of problems remain open: The algorithm convergence was proved only for one of the methods and need to be proved for the others. Another line of research is to consider the case where cell-edge users are served simultaneously by multiple BSs. Such configuration has practical applications in current wireless access networks like LTE.

In [26] (c.f. sec. 1.3.3), downlink coordinated transmission is considered in a multi-cell OFDM system, where the BSs have multiple antennas and the users have single antennas. This is a direct extension of the MISO-OFDM studied in this dissertation to the inter-cell interference coordinated case. An evaluation of algorithm performance shows that joint transmission is very sensitive to synchronization errors, which remains as an open problem together with the analysis of the tradeoff between the amount of needed feedback vs. gained performance, and the method to decide which users to serve using coordinated transmission.

6.1.3 Muti-frame Problem Extension

In this dissertation we used the Lagrange dual method to solve a static optimization for resource allocation per time slot in a OFDM downlink. As part of future work we should consider a multi-temporal extension to this formulation. In this case, we would take into account the more general problem of maximizing the ergodic rates for time-varying fading

channels, taking advantage of temporal diversity. For the SISO case, this has been treated in [5] where an stochastic approximation is used to estimate the channels on-line. Such approach should be investigated for the case with multiple antennas at the transmitter.

REFERENCES

[1] CTIA. (2012, Nov.) Consumer data traffic increased 104 percent according to semi-annual survey. CTIA-The Wireless Association. [Online]. Available: http://files.ctia.org/pdf/CTIA_Survey_MY_2012_Graphics-_final.pdf

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

[3] 3GPP, "3GPP: Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception; Further advancements for E-UTRA physical layer aspects," 3GPP TR V10.5, Tech. Spec.n Group Radio Access Network.

[4] IEEE, "IEEE Standard for Local and metropolitan area networks - part 16: Air Interface for Broadband Wireless Access Systems, 2011," Standard IEEE P802.16m, Institute of Electrical and Electronic Engineers.

[5] X. Wang and G. Giannakis, "Resource allocation for wireless multiuser OFDM networks," *IEEE Trans. in Information Theory*, vol. 57, no. 7, pp. 4359–4372, 2011.

[6] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks part I: Theoretical framework," *IEEE Transactions in Wireless Communications*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[7] C. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[8] I. Kim, I. Park, and Y. Lee, "Use of linear programming for dynamic subcarrier and bit allocation in multiuser OFDM," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 4, pp. 1195–1207, 2006.

[9] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks part II: Algorithm development," *IEEE Transactions in Wireless Communications*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[10] Z. Shen, J. Andrews, and B. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2726–2737, 2005.

[11] D. Bartolomé and A. Pérez-Neira, "Practical implementation of bit loading schemes for multiantenna multiuser wireless OFDM systems," *IEEE Transactions on Communications*, vol. 55, no. 8, pp. 1577–1587, Aug. 2007.

[12] T. F. Maciel and A. Klein, "A resource allocation strategy for SDMA/OFDMA systems," in *Proc. of IST Mobile and Wireless Communications Summit*, Jul. 2007, pp. 1–5.

[13] B. Ozbek and D. L. Ruyet, "Adaptive resource allocation for SDMA-OFDMA systems with genetic algorithm," in *6th International Symposium on Wireless Communication Systems, ISWCS*, Sep. 2009, pp. 483–442.

[14] Y. Tsang and R. Cheng, "Optimal resource allocation in SDMA/multi-input-single-output/OFDM systems under QoS and power constraints," in *Proc. of WCNC*, Mar. 2004, pp. 1595–1600.

[15] P. Chan and R. Cheng, "Capacity maximization for zero-forcing MIMO-OFDMA downlink systems with multiuser diversity," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1880–1889, 2007.

[16] L. Xingmin, T. Hui, S. Qiaoyun, and L. Lihua, "Utility based scheduling for downlink OFDMA/SDMA systems with multimedia traffic," in *Proc. IEEE Wireless Communications and Networking Conference, WCNC*, Mar. 2010, pp. 130–134.

[17] D. Perea-Vega, J. Frigon, and A. Girard, "Near-optimal and efficient heuristic algorithms for resource allocation in MISO-OFDM systems," in *IEEE International Conference on Communications ICC*, May 2010, pp. 1–6.

[18] C. Tsai, C. Chang, F. Ren, and C. Yen, "Adaptive radio resource allocation for downlink OFDMA/SDMA systems with multimedia traffic," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1734–1743, 2008.

[19] W. Chung, L. Wang, and C. Chang, "A low-complexity beamforming-based scheduling for downlink OFDMA/SDMA systems with multimedia traffic," in *Proc. of IEEE GLOBECOM*, Nov. 2009, pp. 1–5.

[20] W. Huang, K. Sun, and T. Bo, "A new weighted proportional fair scheduling algorithm for SDMA/OFDMA systems," in *Proc. 3rd Int. Conf. on Communications and Networking in China, ChinaCom*, Aug. 2008, pp. 538–541.

[21] S. K. V. Papoutsis, I. Fraimis, "User selection and resource allocation algorithm with fairness in MISO-OFDMA," *IEEE Communications Letters*, vol. 14, no. 5, pp. 411–413, 2010.

[22] V. Papoutsis and S. Kotsopoulos, "Resource Allocation Algorithm for MISO-OFDMA Systems with QoS Provisioning," in *Proc. ICWMC, The Seventh International Conference on Wireless and Mobile Communications*, Jun. 2011.

[23] V. Tralli, P. Henarejos, and A. Perez-Neira, "A low complexity scheduler for multiuser MIMO-OFDMA systems with heterogeneous traffic," in *Proc. International Conference on Information Networking, ICOIN*, Jan. 2011, pp. 251–256.

[24] X. Wang and G. Giannakis, "Ergodic capacity and average rate-guaranteed scheduling for wireless multiuser OFDM systems," in *International Symposium on Information Theory, ISIT.*, Jul. 2008, pp. 1691–1695.

[25] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allocation in downlink multicell OFDMA networks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 6, pp. 2835–2848, 2009.

[26] E. Björnson, N. Jaldén, and M. Bengtsson, "Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6086–6101, 2011.

[27] D. Perea-Vega, J. Frigon, and A. Girard, "Dual-Based Bounds for Resource Allocation in Zero-forcing OFDMA-SDMA Systems," in *EURASIP Journal on Wireless Communications and Networking — Accepted for publication.*

[28] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[29] J. Brehmer, *Utility Maximization in Nonconvex Wireless Systems.* Springer Berlin Heidelberg, 2012.

[30] L. V. S. Boyd, *Convex Optimization.* Cambridge University Press, 2004.

[31] J. Brehmer and W. Utschick, "Utility Maximization in the Multi-User MISO Downlink with Linear Precoding," in *Proc. of IEEE International Conference on Communications, ICC*, Nov. 2009.

[32] B. K. R. Fourer, D. Gay, *AMPL: A Modeling Language for Mathematical Programming.* Brooks Cole Co., 2002.

[33] B. Murtagh and M. Saunders, "A projected lagrangian algorithm and its implementation for sparse nonlinear constraints," *Mathematical Programming Study*, vol. 16, pp. 84–117, 1982.

[34] M. Bussieck. (2012, Oct.) MINLP solver software. HumboldtUniversitat zu Berlin,. [Online]. Available: http://www.math.hu-berlin.de/~stefan/minlpsoft.pdf

[35] I. Wong and B. Evans, *Resource Allocation In Multiuser Multicarrier Wireless Systems.* Springer, 2008.

[36] D. Bertsekas, *Convex Analysis and Optimization.* Athena Scientific – Belmont, MA, 2003.

[37] A. Wiesel, Y. Eldar, and S. Shamai, "Optimal generalized inverses for zero forcing precoding," in *Proc. of 41st Annual Conf. on Information Sciences and Systems*, 2007.

[38] D. Perez-Palomar and M. Lagunas, "Joint transmit-receive space-time equalization in spatially correlated MIMO channels: a beamforming approach," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 730–743, 2003.

[39] D. Perez-Palomar and J. Rodriguez-Fonollosa, "Practical algorithms for a family of waterfilling solutions," *IEEE Transactions on Signal Procesing*, vol. 53, no. 2, pp. 686–695, 2005.

[40] Matlab. (2012, Oct.) fminbnd documentation. [Online]. Available: http://www.mathworks.com/help/optim/ug/fminbnd.html

[41] T. Yoo and A. Goldsmith, "Sum-rate optimal multi-antenna downlink beamforming strategy based on clique search," in *Proc. IEEE GLOBECOM conference*, vol. 3, Nov. 2005.

[42] J. Wang, D. Love, and M. Zoltowski, "User Selection with Zero-Forcing Beamforming Achieves the Asymptotically Optimal Sum Rate," *IEEE Transactions on Signal Procesing*, vol. 56, no. 8, pp. 3713–3723, 2008.

[43] G. Dimic and N. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, 2005.

[44] J. Kim, S. Park, J. Lee, J. Lee, and H. Jung, "A scheduling Algorithm combined with zero-forcing beamforming for a multiuser MIMO wireless system," in *Proc. Vehicular Technology Conference, VTC*, 2005.

[45] J. Mao, J. Gao, Y. Liu, and G. Xie, "Simplified Semi-Orthogonal User Selection for MU-MIMO Systems with ZFBF," *IEEE Wireless Communications Letters*, vol. 1, no. 1, pp. 42–45, 2012.

[46] S. Sesia, I. Toufik, and M. Baker, *The LTE Network Architecture – A comprehensive tutorial*. Wiley, Feb. 2009.

[47] F. Khan, *LTE for 4G mobile broadband - air interface technologies and performance*. Cambridge University Pres, 2009.

[48] S. Gortzen and A. Schmeink, "Optimality of Dual Methods for Discrete Multiuser Multicarrier Resource Allocation Problems," *IEEE Transcations on Wireless Communications*, vol. 11, no. 10, pp. 3810–3817, 2012.

[49] Z. Ren, S. Chen, W. Ma, and B. Hu, "A feasible downlink scheduling algorithm in OFDMA systems with discrete rate constraints," in *Proc. 19th International Conference on Telecommunications*, Apr. 2012.

[50] W. Chung, C. Chang, and L. Wang, "An Intelligent Priority Resource Allocation Scheme for LTE-A Downlink Systems," *IEEE Wireless Communications Letters*, vol. 1, no. 3, pp. 241–244, 2012.

[51] W. Kuo and W. Liao, "Downlink MIMO in LTE-advanced: SU-MIMO vs. MU-MIMO," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 140–147, 2012.

[52] Y. Huang, L. Yang, M. Bengtsson, and B. Ottersten, "A Multiuser Downlink System Combining Limited Feedback and Channel Correlation Information," in *International Conference on Communications, ICC*, May 2010.

[53] R1-104933, "Performance Evaluation of double Codebook Structure," NTT DoCoMo, 3GPP Contribution to RAN WG1 Meeting 62, Aug. 2010.

[54] R1-094845, "MU-MIMO Performance Comparison of Two Feedback Assumptions: Rel-8 Codebook and Spatial Covariance," Motorola, 3GPP Contribution to RAN WG1 Meeting 59, Nov. 2010.

[55] 3GPP, "3GPP TS 36.104: Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception," 3GPP TR 36.104 V11.1.0, Tech. Spec.n Group Radio Access Network.

[56] R1-111688, "DL MIMO Feedback Enhancements," Qualcomm Inc., 3GPP Contribution to RAN1 Meeting 65, May. 2011.

[57] E. Dahlman, S. Parkvall, and J. Skold, *4G LTE/ LTE-Advanced for mobile Broadband*. Elsevier, 2011.