Resource Allocation in OFDMA Wireless Networks

by

Mehri Mehrjoo

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Orthogonal frequency division multiple access (OFDMA) is becoming a widely deployed mechanism in broadband wireless networks due to its capability to combat the channel impairments and support high data rate. Besides, dealing with small units of spectrum, named sub-carriers, instead of whole spectrum, results in enhanced flexibility and efficiency of the resource allocation for OFDMA networks.

Resource allocation and scheduling in the downlink of OFDMA networks supporting heterogeneous traffic will be considered in this thesis. The purpose of resource allocation is to allocate sub-carriers and power to users to meet their service requirements while maintaining fairness among users and maximizes resource utilization. To achieve these objectives, utility-based resource allocation schemes along with some state-of-the-art resource allocation paradigms such as power control, adaptive modulation and coding, sub-carrier assignment, and scheduling are adopted. On one hand, a utility-based resource allocation scheme improves resource utilization by allocating enough resources based on users' quality of service (QoS) satisfaction. On the other hand, resource allocation based on utilities is not trivial when users demand different traffic types with convex and nonconvex utilities.

The first contribution of the thesis is the proposing of a framework, based on joint physical (PHY) and medium access (MAC) layer optimization, for utility-based resource allocation in OFDMA networks with heterogeneous traffic types. The framework considers the network resources limitations while attempting to improve resources utilization and heterogeneous users' satisfaction of service. The resource allocation problem is formulated by continuous optimization techniques, and an algorithm based on interior point and penalty methods is suggested to solve the problem. The numerical results show that the framework is very efficient in treating the nonconvexity problem and the allocation is accurate comparing with the ones obtained by a genetic search algorithm.

The second contribution of the thesis is the proposing of an opportunistic fair scheduling scheme for OFDMA networks. The contribution is twofold. First, a vector of fair

weights is proposed, which can be used in any scheduling scheme for OFDMA networks to maintain fairness. Second, the fair weights are deployed in an opportunistic scheduling scheme to compensate the unfairness of the scheduling. The proposed scheme efficiently schedules users by exploiting multiuser diversity gain, OFDMA resource allocation flexibility, and utility fair service discipline.

It is expected that the research in the thesis contributes to developing practical schemes with low complexity for the MAC layer of OFDMA networks.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Xuemin (Sherman) Shen for his generous support and supervision, and for the great experience that he shared with me. I learned valuable lessons from his personality and his visions. I will never forget his encouragement when I faced difficulties and felt frustrated during the course of my Ph.D. study. He gave me enough time and freedom to choose the research topic which I am interested in. Without his endless support, my achievements were not possible.

I am also grateful to my Doctoral Committee Members: Professor Sagar Naik, Professor Pin-Han Ho, the internal external examiner, Professor Jun Cai, and also the external examiner, Professor Ben Liang. Their valuable comments and suggestions helped me improve the quality of this dissertation. Their precious time and efforts devoted to this dissertation are highly appreciated.

With great thanks, I want to acknowledge the partial financial support of the Iranian Ministry of Science, Research, and Technology (MSRT), the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the University of Waterloo scholarships and awards for funding this research work. My thanks are also extended to the ECE Department of the University of Waterloo for a nice atmosphere, kind treatment and support.

Special thanks to my friends and colleagues in the broadband communication research (BBCR) group for their fruitful discussions and professional collaborations. My deepest thank goes to Mehrdad Dianati, Naghmeh Mansouri, Somayeh Moazeni, Mohamad Awad, and Ibrahim Alsolami who inspired me by new ideas through constructive discussions.

Finally, and most importantly, I would like to express my deep appreciation to my parents and my husband, Masoud, for all their encouragement, understanding, support, patience, and true love throughout the ups and downs. As always, I thank and praise God by my side.

to my dear Masoud

&

my blossom of life, Samin

Contents

List of Abbreviations **xii**

List of Symbols **xv**

List of Tables **xv**

List of Figures **xviii**

1 Introduction **1**

1.1 Research Motivation 1

1.2 Problem Description 3

1.3 Research Objectives and Contributions 6

1.4 Structure of the Thesis 8

1.5 Bibliographic Notes 9

2 Multi-carrier Transmission Over Wireless Channel **11**

2.1 Radio Channel 12

2.1.1 Wireless Channel Impairments 12

2.1.2 Wireless Channel, Mathematical Model 14

2.1.2.1 Delay Spread 16
2.1.2.2 Coherence Bandwidth 16
2.1.2.3 Doppler Spread 17
2.2 Multi-Carrier OFDM 18
2.2.1 OFDM Transmitter and Receiver 18
2.2.2 PHY Layer Advantages of OFDM 18
2.2.3 MAC Layer Advantages of OFDM 21
2.3 Network Topology and Configuration 23
2.3.1 PHY Layer 25
2.3.2 MAC Layer 25
2.4 Summary 26

3 A Framework for Resource Allocation in OFDMA Networks 29
3.1 Problem Formulation 30
3.1.1 MINLP and NLP Problems 31
3.2 Related Works and Problem Complexity 36
3.2.1 Linear Objective Function 37
3.2.2 Nonlinear Objective Functions and Constraints 38
3.2.2.1 Nonlinear Objectives 38
3.2.2.2 Nonlinear Constraints 39
3.2.3 Problem Complexity 40
3.3 Penalty Function and Interior Point Methods 41
3.3.1 PM/IPM Descriptions 42
3.4 Genetic Algorithm 48

3.4.1 Genetic Algorithm Methodology 49
3.4.2 Genetic Algorithm Implementation 50
3.5 Numerical Results 51
3.5.1 Genetic Algorithm Convergence 53
3.5.2 Interior Point and Penalty Method Performance 54
3.5.3 Resource Utilization Performance 61
3.6 Summary 63

4 Opportunistic Fair Scheduling in OFDMA Networks 67
4.1 Background and Related Works 68
4.2 Opportunistic Fair Scheduling Scheme 69
4.3 Network Model and Problem Formulation 71
4.3.1 OFDMA Resource Allocation Problem 72
4.3.2 Fairness Problem 75
4.4 Solution Algorithms for OFDMA Resource Allocation and Fairness Optimization Problems 77
4.4.1 The Dual Method 78
4.4.2 The Interior Point Method 79
4.5 Complexity of the Proposed Approach 81
4.6 Numerical Results 81
4.6.1 Fixed Users 84
4.6.2 A Fixed and a Mobile User 84
4.6.3 Users With Heterogeneous Rate Requirements 87
4.7 Summary 90

5 Conclusions and Further Works **91**

5.1 Major Research Contributions 91

5.2 Further Works 93

5.3 Final Remarks 97

APPENDICES **98**

A Derivation of $\nabla^2_{rr}\mathcal{L}$, $\nabla_r f(r)$, $\nabla^2_{ww}\mathcal{L}$, and $\nabla_w f(w)$ **99**

A.1 $\nabla^2_{rr}\mathcal{L}$ and $\nabla_r f(r)$ 99

A.2 $\nabla^2_{ww}\mathcal{L}$ and $\nabla_w f(w)$ 101

References **103**

List of Abbreviations

3G	third generation
AMC	adaptive modulation and coding
AWGN	additive white Gaussian noise
BER	bit error rate
BPSK	binary phase shift keying
BS	base station
bps	bit per second
CSI	channel state information
DL	downlink
DSA	dynamic sub-carrier assignment
EWMA	exponential window moving average
FDM	frequency division multiplexing
FFT	fast Fourier transform
GA	genetic algorithm
GPS	generalized processor sharing
IFFT	inverse fast Fourier transform
ISI	intersymbol interference
IPTV	internet protocol television
KKT	Karush-Kuhn-Tucker
LIP	linear integer programming
LOS	line-of-sight

MAC	medium access layer
MINLP	mixed integer nonlinear programming
NLOS	non-line-of-sight
NLP	nonlinear programming
NUM	network utility maximization
OFDM	orthogonal frequency division multiplexing
OFDMA	orthogonal frequency division multiple access
PHY	physical layer
PMP	point-to-multipoint
PM/IPM	penalty method/interior point method
Pr	problem
QoS	quality of service
RF	radio frequency
RMS	root mean square
SNR	signal to noise ratio
TDD	time division duplexing
TDM	time division multiplexing
UL	uplink
UMTS	universal mobile telecommunication system
UWB	ultra wide band
WiFi	wireless fidelity
WiMax	wireless interoperability for microwave access
WLAN	wireless local area networks
WMAN	wireless metropolitan area networks
WWAN	wireless wide area networks

List of Symbols

A	the Jacobian matrix of $C(r)$
α_{ijn}	channel gain of user i on sub-carrier j of OFDMA symbol n
α_{ij}	channel gain of user i on sub-carrier j of each allocation interval
α_s^{max}	length of movement for vector s
α_z^{max}	length of movement for vector z
B	network bandwidth
b	vector of movements for vectors r, s, and z, denoted by $b = [b_r, b_s, b_z]^T$
$C(r)$	vector of inequality constraints in PM/IPM
c	a solution for sub-carrier assignment $[c_{11}, c_{12}, \ldots, c_{1K}, \ldots, c_{M1}, \ldots, c_{MK}]^T$
c_{ij}	a binary variable representing assignment of sub-carrier j to user i
e	$e = (1, 1, ..., 1)^T$
$\mathcal{F}$	a general objective function
g	iteration number in PM/IPM
h	utility set index in $\mathcal{U}$
i	user index belongs to $\mathcal{M} := \{1, 2, ..., M\}$
j	sub-carrier index belongs to $\mathcal{K} := \{1, 2, ..., K\}$
K	total number of sub-carriers in the network
k	convexity controlling factor of utility functions
κ	long distance fading exponent
$\mathcal{L}$	Lagrangian function

L	penalty parameter in PM/IPM
$l1$	minimum required rate in utility functions
$l2$	maximum rate in utility functions
M	total number of users in the network
μ	KKT perturbing variable
N	total number of OFDM symbols in each scheduling interval
$\mathcal{N}_{itr}$	maximum number of iterations
$\mathcal{N}_{fit}$	maximum number of iterations that *elite*'s fitness value does not change
n	symbol index belongs to $\mathcal{N} := \{1, 2, ..., N\}$
P_{BS}	the BS total power budget
$\mathcal{P}_0$	an initial population
$\mathcal{P}_n$	n^{th} population
p	a solution for power allocation, $[p_{11}, p_{12}, \ldots, p_{1K}, \ldots, p_{M1}, \ldots, p_{MK}]^T$
p_{ij}	allocated power to user i on sub-carrier j of each allocation interval
p_{ijn}	required power by user i on sub-carrier j of OFDMA symbol n to transmit r_{ijn}
p_{cross}	probability of crossover
p_{mut}	probability of mutation
$\tilde{R}_i$	total transmitted rate to user i over simulation intervals
R_i	average transmitted rate to user i over T_c scheduling interval
R^i_{min}	minimum service rate requirement of user i
r	a solution for rate allocation, $[r_{11}, r_{12}, \ldots, r_{1K}, \ldots, r_{M1}, \ldots, r_{MK}]^T$
r_{ij}	allocated rate to user i on sub-carrier j of each allocation interval
r_{ijn}	achievable rate by user i on sub-carrier j of OFDMA symbol n
S	a diagonal matrix with diagonal elements given by vector s
s	the vector of slack variables in PM/IPM
σ	a constant for updating KKT perturbing variable
T_c	time constant of the lowpass filter for computing users' transmission rates averages
τ	a constant value in PM/IPM for updating the movement directions

$\mathcal{U}$	a bounded set of M users' feasible utilities subsets, u_h
U_i	utility function of user i
u_h	$u_h = \{u_{h1}, u_{h2}, \ldots, u_{hM}\}$
u_{hi}	is the utility of user i in utility subset h
v	iteration number in GA algorithm
W	population size
W_i	user i's fair weight
x_j^y	a $1 \times M$ allocation vector of x_{ij}^ys in GA algorithm
x_{ij}^y	allocated power to user i on sub-carrier j in GA algorithm
y	a chromosome, a $K \times M$ vector equals $[x_1^y \cdots x_j^y \cdots x_K^y]$
Z	a diagonal matrix with diagonal elements given by vector z
z	a vector containing $(2M+1)K$ Lagrange multipliers

List of Tables

3.1 Notations Descriptions 32
3.2 Simulation Parameters 54
3.3 Users's Allocated Rates on Each Sub-carrier 62

4.1 List of Symbols 73
4.2 Aggregate Utilities of the Scheduling Schemes 88

List of Figures

2.1 Multipath channel 13

2.2 Tapped delay model for multipath channel 15

2.3 An OFDM transceiver structure 19

2.4 Time and frequency illustration of OFDM-sub-carriers (a) two OFDM sub-carriers modulated by binary phase shift keying (BPSK) are illustrated over two OFDM symbols duration, (b) nine OFDM sub-carriers are illustrated in frequency domain 20

2.5 Network platform is the DL of a PMP infrastructure where spectrum and power are allocated to users with heterogeneous service requirements. . 24

2.6 OFDM symbols and sub-carriers in a MAC frame 27

3.1 The population and chromosomes representations 50

3.2 Power allocation distribution on sub-carriers 55

3.3 Convergence of fitness value 56

3.4 Convergence speed comparison of GA and PM/IPM 57

3.5 Performance comparison of GA and PM/IPM 58

3.6 Aggregate penalty term constraints' deviations in PM/IPM 59

3.7 Utility allocation comparison of GA and PM/IPM 60

3.8 Concave and nonconcave utilities corresponding to each group of users 63

3.9 Allocated rate and corresponding utilities to each group of users 64

3.10 Utilization performance of utility-based and greedy schemes 65

4.1 Scheduling instances and sub-carriers allocation illustrations in a MAC frame 70

4.2 Architecture of the proposed scheduler 70

4.3 Comparison between equal rate and equal utility allocation 75

4.4 Simulated scenarios: (a) fixed users, (b) a fixed user and a mobile user, (c) users with heterogeneous rate requirements 83

4.5 Overall network throughput for scenario (a) 85

4.6 Fairness index for scenario (a) 85

4.7 User1 and User2's throughput at different positions of the second scenario 86

4.8 Fairness performance of the second scenario 87

4.9 Utility values of users 1 to 8 for opportunistic and opportunistic fair scheduling schemes versus time 89

Chapter 1

Introduction

1.1 Research Motivation

Recently, the world has witnessed rapidly growing of wireless technology and increasing demand for wireless communication services [1]. Accordingly, the new standards for the next generation wireless networks, such as, IEEE 802.16 [2–5] for wireless metropolitan area networks (WMAN), IEEE 802.11 [6] for wireless local area networks (WLAN), or universal mobile telecommunication system (UMTS) for third generation (3G) wireless networks[6], appear with the trend of providing heterogeneous services over broadband channel. However, successful deployment of the standards faces a number of challenges, e.g., scarce spectrum, complex time-varying wireless channel, and providing quality of service requirements of heterogeneous traffic types or service requirements.

Despite the limited unlicensed radio frequency (specifically below 11 GHz), it is used exhaustively due to the advantages of fast rollout and low administrative/regulatory costs. Besides, current technological barriers of using high frequency bands, that need line-of-sight (LOS) transmission, fade the motivation of developing the applications that use those bands. On the other hand, non-line-of-sight (NLOS) transmission on the unlicensed band suffer from multipath propagation especially in urban areas. Accordingly, wireless transmission techniques that promote spectrum usage efficiency and enable high data rate transmission over multipath radio, such as, frequency division multiplexing (OFDM) have found widespread deployment in current wireless transmission technologies [7–11].

On the other hand, a natural challenge of wireless channel is reducing signal strength, but strong restrictions are taking effect on increasing transmitted signal strength. Limiting the power consumption is one of the requirements of having a green world. In addition, the technological constraints of battery products for electronic mobile devices pose a restriction on available power. In addition to power limitation, wireless channel is highly time varying, which results in different power requirement for each transmission instance. Also, it demands sophisticated power allocation schemes that adaptively allocate limited power and take advantage of users' diversity for power allocation.

Irrespective of wireless medium challenges for traffic transmission, wireless applications, such as, cell phones, are becoming more popular and new applications, such as, mobile computing, and video on demand are promising in the near future. Each of these applications demands its own service requirements and sophisticated service management that should be fair to all users. To come up with a solution for heterogeneous traffic types transmission on wireless channel, researchers have to put a lot of efforts on proposing some resource allocation schemes that consider the aforementioned challenges. In other words, a resource allocation scheme is needed to consider technical issues of transmission technologies and wireless access mechanisms while allocating resources to meet the heterogeneous service requirements. Due to the large diversity of telecommunication networks topologies, constraints, and objectives, many resource allocation schemes have been proposed so far for legacy wireless networks [12]. However, these schemes need renovation and/or redesign due to the advent of new transmission technologies and network applications.

Multicarrier OFDM transmission is a developing aspect and multiservice provisioning is a promising objective in recent wireless networks. OFDMA, the multiple access mechanism based on multicarrier OFDM, results in a flexible resource allocation [10,13] in the sense that instead of allocating whole resources, such as total bandwidth, to only one user at a time, some portion of it can be allocated to each user. The flexibility of

OFDMA resource allocation can be deployed to compensate for wireless channel impairment [14], provide QoS [15, 16], and maintain fairness [17]. These issues have been studied separately in the existing literature, but fairness, QoS, and resource utilization enhancement should be considered simultaneously for efficient resource allocation in practice. The commercial growth of the networks with multicarrier transmission and heterogeneous traffic types strongly depends on proposing efficient resource allocation schemes that consider the aforementioned issues.

1.2 Problem Description

Transmission over wireless medium is the first and most fundamental challenge that the service providers face in a broadband communication network. The medium is impaired by many factors, such as, obstacles, noise, interference, and intersymbol interference (ISI). Obstacles shadow the signal path or cause scattering and diffraction, which result in multipath propagation. Noise weakens the transmitted signal strength, and interferences distort the signal. Basically, the degradations are unpredictable and time-varying. Besides, they become more severe when the signal bandwidth increases. Accordingly, elaborated methods are needed to mitigate channel impairments in broadband networks.

One of the most effective techniques to increase the spectral efficiency and combat the wireless channel impairments in wireless networks is OFDM. The fundamental feature of OFDM is it converts single carrier transmission to multi-carrier transmission, which is advantageous from the PHY and MAC layers points of view. In PHY, OFDM sub-carriers have overlap (it is possible because of their orthogonality) which increases spectral efficiency. In MAC, using OFDMA sub-carriers can improve spectral efficiency in two ways. First, given channel state information (CSI) of sub-carriers, a transmission can be scheduled over sub-carriers that have good status, which results in less effort for retransmission of corrupted signals transmitted on weak sub-carriers. Second, as CSI of sub-carriers for different users are usually independent and uncorrelated, a sub-carrier which is not in good status for a user may be in a good status for another user. An optimal frequency usage will be achieved upon optimal sub-carrier assignment to users.

The performance of OFDMA depends on sub-carrier assignment to users as well as power allocation to each sub-carrier. Therefore, the joint sub-carrier and power allocation problem, denoted by OFDMA resource allocation problem, is formulated as an optimization problem whose solution is an optimal allocation. In an OFDMA resource allocation problem, resources are allocated to users in a way to achieve an objective while satisfying some constraints. Maximizing aggregate users' rates or minimizing total required transmission power are examples of the objective functions. The constraints are imposed by some network limitations or service requirements, such as, maximum available power source or users' minimum rate requirements.

A constraint, related to PHY and MAC implementation of OFDMA, is to allocate a sub-carrier to only one user at a time. In other words, a sub-carrier band cannot be shared by several users simultaneously. Appeared in an OFDMA optimization problem, this constraint causes the feasible region of the problem, i.e., the set of allocations that satisfy all constraints, becomes discrete. An optimization problem with discrete feasible region is categorized among nonconvex optimization problems[1].

In addition to the nonconvexity of the feasible region, the objective function of an OFDMA optimization problem may contribute to the nonconvexity of the problem. Utility-based OFDMA resource allocation problems are among this category of nonconvex problems. Utility function, shortly utility hereafter, is usually a function of rate that shows a user's satisfaction of received service [18]. Some utilities are designed in the literature to achieve a specific objective, but, in this thesis, application layer utilities, i.e., those utilities that represent users' perception of QoS at the application layer are considered. In a utility-based resource allocation scheme, resource are allocated according to users' requirements as long as the allocation is effective in utility increment. For

[1] An optimization problem is nonconvex if either the feasible region or the objective function be nonconvex. Moreover, a function f is *convex* if the domain of f, D_f, is a convex set, i.e., $(1-t)x + ty \in D_f$ for every $x, y \in D_f$ and $t \in [0, 1]$, and $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ for every $x, y \in D_f$ and $0 \leq \theta \leq 1$.

example, a *step* utility function of rate represents that the user expects a threshold rate, allocating less rate is not useful at all, and allocating more rate is wasteful. Due to the advantage of resource utilization enhancement, some utility-based resource allocation schemes, namely utility maximization problems, have been proposed in the literature recently. In a utility maximization problem the effort is on maximizing aggregate users' utilities. As long as users' utilities are concave[2], the utility maximization problem is a convex problem. For many concave utilities, the utility-based resource allocation problem is a convex problem which can be solved using special methods for convex optimization [19]. Therefore, most works in the literature have considered only concave utilities. On the contrary, in case of utility maximization for heterogeneous traffic, some of the utilities, such as voice and video, are nonconcave. Then, the utility-based resource allocation in a multiservice network will not be a convex optimization problem any more. Nonconvexity of the objective function, when combined with the nonconvexity of the feasible region, contributes to difficulty of solving the utility-based OFDMA resource allocation problem.

Unlike convex problems, which there exist several algorithms to solve them up to the optimum solution efficiently, there is no suggested algorithm for nonconvex problems that guarantees an optimum solution. Accordingly, nonconvex problems are usually solved for a local (near optimal) solution by either heuristic search algorithms or nonlinear programming (NLP) solver algorithms. When the feasible region is small and discrete, a search algorithm may find the optimum solution in limited time, but when the feasible region expands, the solution time grows exponentially, and search algorithms become inefficient. Similar to the search algorithms, NLP solver algorithms will result in local solutions. However, they are usually faster than search algorithms. More importantly, the closeness of solutions to optimal depends on the solver algorithm which is used. Precisely, the accuracy of the solution obtained by an NLP solver algorithm depends on the way that it treats the nonconvexity of the problem.

[2]A function f is *concave* if the domain of f, D_f, is a convex set, i.e., $(1-t)x + ty \in D_f$ for every $x, y \in D_f$ and $t \in [0,1]$, and $f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta)f(y)$ for every $x, y \in D_f$ and $0 \leq \theta \leq 1$.

Despite the difficulties of utility-based OFDMA resource allocation problems, they can be applied in many different network scenarios. Most of present wireless access technologies for ultra wide band (UWB), WLAN, WMAN, and cellular networks deploy OFDMA and aim a heterogeneous service provisioning. When applied to these scenarios, utility-based OFDMA resource allocation schemes can efficiently allocate resources to qualify usres' satisfaction and improve resource utilization. In this thesis, utility-based OFDMA resource allocation schemes in the context of heterogeneous service provisioning in the downlink of IEEE 802.16 WMAN is investigated. The attempt is to specify challenging aspects of the problem and suggest a practical and accurate solution algorithm.

1.3 Research Objectives and Contributions

The main objectives of this research are to develop a framework for resource allocation that provides satisfactory QoS and fairness for heterogeneous traffic types in the downlink of point-to-multipoint (PMP) OFDMA networks, while improving network resource utilization. The framework guarantees the users' minimum rate requirements, maintains fairness among users, and enhances resource utilization simultaneously. In order to realize these objectives, the research work is conducted in three stages as follows.

In the first stage, the OFDMA resource allocation problem that guarantees users' minimum rate requirements is formulated based on continuous optimization techniques [20]. The problem of OFDMA resource allocation is usually presented by mixed integer nonlinear programming (MINLP) techniques in the literature [21–30]. However, our proposed optimization problem is an NLP problem, which does not contain integer variables. The NLP problem uses the information of OFDMA sub-carrier status to allocate power or rate to sub-carriers. This information are obtained through a feedback channel via underlying PHY and assumed to be constant in a limited interval. The proposed framework allows any objective function of users' rate be considered in the optimization problem. Indeed, the framework can serve applications with either linear/nonlinear or concave/nonconcave utilities that require a minimum rate for functionality. We investigate the performance of the utility-based OFDMA resource allocation scheme using an iterative search algorithm. We implement a heuristic search algorithm based on genetic algorithm (GA) for the NLP problem. The results of the iterative search algorithm are used as a benchmark in the next stages of the research, where an analytical algorithm is proposed to solve the problem.

In the second stage, inspired by continuous optimization approach used for the OFDM resource allocation problem representation, an algorithm based on a combination of a penalty and an interior point method (PM/IPM) is suggested to solve the NLP problem. Mainly, the approach takes advantage of an interior point method which can be successfully applied to nonlinear programming problems [31]. The success of interior point methods in solving a nonconvex or nonlinear problem strongly depends on how nonconvexity of the problem is treated. We apply a penalty function method to deal with nonconvexity problem. Before applying the interior point method, the nonconvexity of the feasible region is removed by a penalty function method. More precisely, nonconvex constraints are moved to the objective function by a coefficient penalty. Then the interior point method is applied to solve the new problem with convex feasible region. The solutions obtained by PM/IPM are compared with near optimal solutions obtained by GA in terms of speed and efficiency of the algorithms. The proposed PM/IPM is very comprehensive in the sense that users can have heterogeneous rate requirements and the objective function of the resource allocation scheme can be nonconvex.

In the third stage, an opportunistic fair scheduling scheme is proposed for heterogeneous traffic in the downlink of broadband OFDMA networks. In this scheme, users are scheduled for service based on three factors: a) the achievable data rate at the instant of scheduling, b) the average data rate that had been experienced by each user during an observation time window preceding the scheduling instant , c) the assigned fair weight to each user. The scheme uses a fair service discipline to allocate resources to users based on the instantaneous CSI of sub-carriers. The fair weights compensate the unfairness of the opportunistic scheduling and can be adjusted dynamically according to users' average channel status and fairness criterion. More precisely, the fair weights are computed based on the proposed framework for OFDMA resource allocation and can be used in most of the scheduling schemes in OFDMA networks. The scheduling scheme achieves a flexible trade-off between fairness and throughput. Radio resource utilization is enhanced by using adaptive modulation and coding. The scheduler runs a bit loading algorithm, which is embedded in the sub-carrier assignment algorithm, to determine allocated rate to sub-carriers while scheduling users.

1.4 Structure of the Thesis

The wireless channel impairment, happening from various random phenomena in the signal propagation paths, should be well understood and taken into consideration while designing broadband wireless networks. In chapter 2, a brief characterization of the radio channel including small scale and large scale fading is presented. Then, the OFDM and OFDMA design issues to cope with the large and rapid variations in received signal strength and provide a reliable transmission are explained. The OFDMA resource allocation is investigated in the context of a centrally controlled OFDMA broadband network in this thesis. In addition, the required knowledge of PHY and MAC relevant to the resource allocation problem formulation is described, such as the relation among transmission rate, power, channel gain and bit error probability.

The problem formulation for resource allocation in centrally controlled OFDMA networks is presented in chapter 3. First, basic assumptions and constraints of OFDMA and the network are introduced. The OFDMA resource allocation problem is represented by a MINLP first. Then, an equivalent NLP problem for the MINLP one is proposed, which is followed by a discussion about OFDMA resource allocation problem complexity. An iterative search algorithm based on GA and analytical algorithm based on PM/IPM is suggested to solve the NLP problem. Numerical results for scenarios with convex and nonconvex objective functions are conducted to evaluate utility-based resource allocation schemes and verify the accuracy of solutions achieved by PM/IPM.

An opportunistic fair scheduling scheme is proposed in 4 for scheduling heterogeneous traffic in the downlink of OFDMA networks. The scheduler takes advantage of independent channel variation across users to improve the network performance through multiuser diversity. Also, to guarantee fairness, a weighted fairness scheme based on users' average channel gain and required fairness criterion is proposed. In this chapter, first, some opportunistic fair schemes proposed in the literature for multicarrier networks are surveyed. Then, the optimization problems correspondent to the scheduling scheme and the fairness scheme are derived, and separate algorithms appropriate for solving each problem is suggested. Finally, numerical results are conducted to evaluate the performance of the scheduling scheme and illustrate its adaptivity to users' CSI.

The contribution of this thesis is summarized in chapter 5. In addition, the future research directions relevant to the works in this thesis are discussed. Also, final remarks of the thesis are given at the end of this chapter.

1.5 Bibliographic Notes

Most of the research work reported in this dissertation have appeared in peer reviewed papers [12, 20, 32–35] or will be published in [36–42]. The concepts discussed in chapter 2 appeared in [12, 33, 35, 40–42]. The work of chapter 3 can be found in [20, 34, 36, 38, 40, 41]. The material of chapter 4 can be found in [32, 33, 35, 37, 39, 42].

Chapter 2

Multi-carrier Transmission Over Wireless Channel

The emerging technology extends the transmission rate and range of wireless communication beyond the limits of existing technologies while allowing for heterogeneous traffic transmission. To achieve all these goals, qualified protocols should efficiently utilize the spectrum and overcome the deficits of wireless channel simultaneous to maintaining a satisfactory level of service for users with heterogeneous traffic types.

Most current wireless standards support OFDM and OFDMA which, respectively, are robust technique for transmission and flexible mechanism for resource allocation on wireless channel. The OFDM air interface mitigates multipath and interference effects, which are some main challenges of wireless communication. The OFDMA mechanism is very flexible in allocating resources due to its capability of providing fine granularity in accessing the spectrum. As we take advantage of these specifications through this research work, we explain them briefly in this chapter.

First, wireless channel specifications and transmission challenges in broadband networks are explained. Then, we will describe how OFDM can combat the channel impairments and how flexibility and granularity of OFDM can be incorporated in a resource allocation scheme to improve network performance. We consider a general centralized network topology throughout this thesis, which is introduced in the last section of this chapter.

2.1 Radio Channel

2.1.1 Wireless Channel Impairments

Propagation over wireless channel weakens, delays, and deteriorates transmitted signals randomly. Expansion of wireless networks over urban areas necessitates NLOS transmission, where a transmitted signal passes several obstructions on its way to a wireless receiver. When a signal is propagated in NLOS conditions, random phenomenas, such as, reflection, refraction, diffraction, absorption, or scattering deteriorate the signal and result in multiple reception of the signal with different delays and strength. The wireless channel impairments can be categorized as the following phenomena and effects:

- Noise: Additive white Gaussian noise (AWGN) is the main impairment in any communication channel. AWGN has a constant spectral density, so it affects broadband signals more than narrow-band signals. As AWGN is additive, it can be formulated by simple and tractable mathematical models.

- Shadowing: Large obstacles in the propagation path, such as buildings and moving objects, shadow the signal transmission. Although, radio waves propagate around such blockages via diffraction but the power loss drops severely. Shadowing phenomenon causes slow variations of a transmitted signal with respect to the signal duration, so shadowing is sometimes referred to slow fading in the literature.

- Pathloss: A signal power decays in the communication path as the distance increases. Pathloss depends on the environment of traversing signals and is inversely proportional to square carrier frequencies. Broadband signals experience significant pathloss. In addition, pathloss is worse in NLOS than the line-of-sight (LOS)

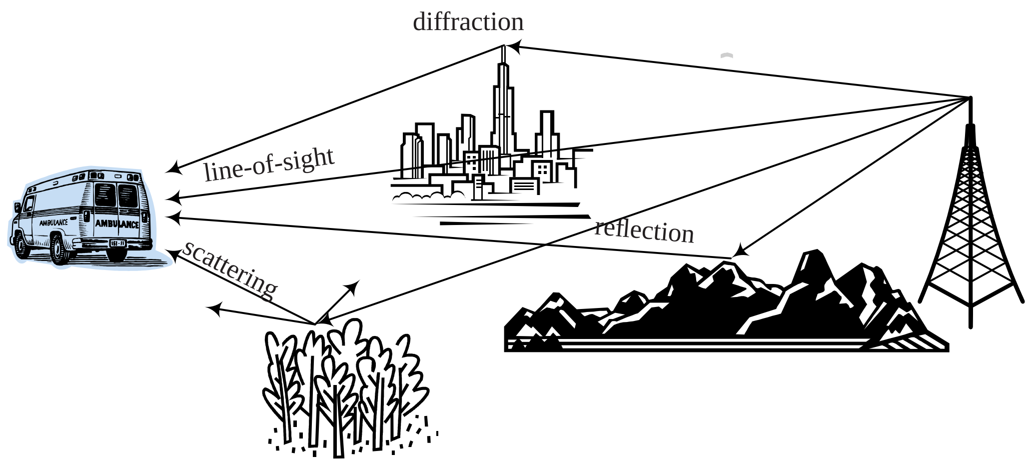

Fig. 2.1: Multipath channel

transmission. Pathloss is a large-scale fading type because its effects are dominant in extended geographical networks.

- Multipath Fading: Large variations in received signal envelope occurred by propagating the transmitted signal via diffraction, scattering, and reflection, as shown in Fig. 2.1, is characterized as multipath fading. The variation of the amplitude of the received signal affected by multipath fading may be very large even over very small distances or small durations. Multipath propagation causes frequency selective fading and intersymbol interference (ISI). The frequency selectivity results from destructive interference of transmitted signal with itself due to multipath reflections. A frequency selective fading channel cause deep fading in some frequency components of the transmitted signal. The locations of the deep fades may change because the interference pattern changes with reflectors movement or changes.

  ISI is due to the signal propagation through different paths and concurrent receptions of different transmitted signals. In a NLOS environment, time dispersion of a multiple propagated signal causes it arrives at the receiver during the next symbol period reception. ISI is a big concern for broadband signal transmission, because the symbol length is short in time and a small delay cause ISI. Traditionally, ISI is overcome by equalization, but it is computationally hard when number of transmitted signals increases.

- Doppler Shift: Time selectivity which is occurred due to relative motion between a transmitter and receiver causes carrier frequency dispersion called Doppler shift. Doppler shift phenomenon depends on movement speed and carrier frequency. Doppler shift reduces SNR and can make carrier recovery and synchronization more difficult for broadband signals. Doppler shift is a main concern for OFDM-based networks, since it can corrupt the orthogonality of the OFDM sub-carriers named intercarrier interference (ICI).

- Interference: It is the conflict resulted when two or more users transmit on the same frequency band. Frequency reuse, which allows users share available bandwidth and improve spectrum utilization, may cause signals from different users to interfere with each other. Interference limits the capacity and coverage of wireless networks.

Typically, the broader is the signal, the worse is the wireless channel impacts. Broadband wireless networks need to be designed to cope with these large and rapid variations in received signal strength. There is no unique solution to all these impairments. However, OFDM is a popular choice for mitigating most of these deficits, because it exploits wireless channel fluctuations and multichannel transmission flexibility for efficient transmission of broadband signals. We will explain, later in this chapter, how OFDM will reduce some of these impairments. For this purpose, we first formulate some of aforementioned channel effects in the followings.

2.1.2 Wireless Channel, Mathematical Model

The communication channel can be modeled as a linear time variant system [43]. Due to multipath propagation and Doppler effect, the channel impulse response to $\delta(\tau)$, Dirac

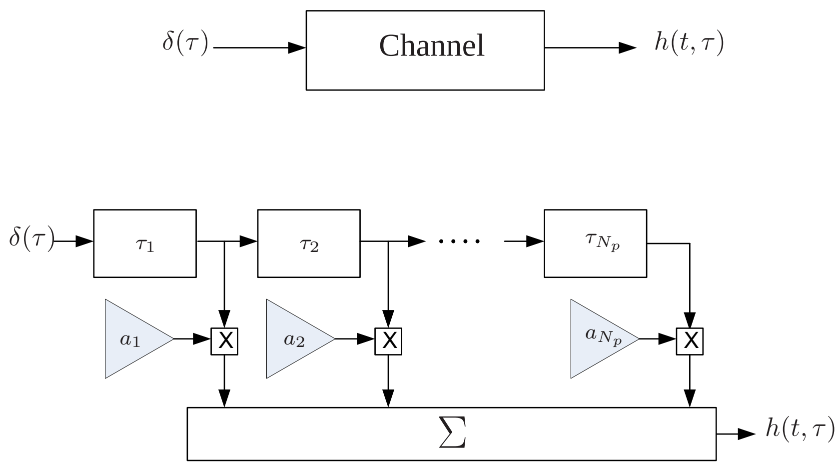

Fig. 2.2: Tapped delay model for multipath channel

impulse function transmitted at the moment τ, is the superpose of the reflected $\delta(\tau)$s:

$$h(\tau,t) \quad = \quad \sum_{p=0}^{N_p-1} a_p(t)e^{j(2\pi f_{D,p}t+\varphi_p)}\delta\left(\tau-\tau_p(t)\right). \tag{2.1}$$

a_p, $f_{D,p}$, φ_p and τ_p refer to the complex-valued amplitude, doppler frequency, phase, and delay of path p among N_p multipath. A systematic representation of $h(\tau,t)$ is a tapped delay line as shown in Fig. 2.2 [44, 45], where the output of each delay block τ_p, is a tap consisting of multiple propagated signal with close delays to τ_p. In practice the number of taps that can be distinguished is very large. Therefore, only those taps with a delay greater than the inverse of the input signal bandwidth, are considered in the receiver detectors [43].

Fading channel effects depend on some channel characteristics such as delay spread, coherence frequency, and Doppler spread and some signal characteristics such as bandwidth and duration time. In the following, we explain these characteristics and their relevant bounds that limit some fading effects, such as, ISI, frequency selectivity, and ICI.

2.1.2.1 Delay Spread

The delay dispersion of channel, identified as root mean square (RMS) delay spread, determines the severity of ISI and frequency selective fading. RMS delay spread τ_{RMS} depends on the channel stationary impulse response $h(t)$

$$h(t) = \sum_{p=0}^{N_p-1} a_p \delta\left(t - \tau_p\right) \tag{2.2}$$

and channel mean delay $\bar{\tau}$:

$$\bar{\tau} = \frac{\int_0^\infty \left|h(t)\right|^2 t dt}{\int_0^\infty \left|h(t)\right|^2 dt} = \frac{\sum_{p=0}^{N_p-1} \left|a_p\right|^2 \tau_p}{\sum_{p=0}^{N_p-1} \left|a_p\right|^2} \tag{2.3}$$

as follows [46]:

$$\tau_{RMS} = \frac{\int_0^\infty \left|h(t)\left(t-\bar{\tau}\right)\right|^2 dt}{\int_0^\infty \left|h(t)\right|^2 dt} = \frac{\sum_{p=0}^{N_p-1} \left|a_p\left(\tau_p-\bar{\tau}\right)\right|^2}{\sum_{p=0}^{N_p-1} \left|a_p\right|^2}. \tag{2.4}$$

RMS delay spread is the standard deviation value of the delay of reflections, weighted proportional to the energy in the reflected signals. To avoid ISI, the symbol duration T_s should be much larger than τ_{RMS} [47].

2.1.2.2 Coherence Bandwidth

In a frequency selective fading channel, the frequency components of a transmitted signal are distorted differently. To avoid frequency selectivity, the signal bandwidth should be smaller than the channel coherence bandwidth B_c, which is the frequency band that the channel is frequency flat fading. Coherence bandwidth is a measure of the channel frequency dispersion, i.e., the extent between two different frequencies f_1 and f_2 where the channel fading is correlated. Accordingly, the fading effect for two tones located apart farther than B_c is uncorrelated.

The correlation can be measured by the channel frequency response autocorrelation function as [48]:

$$R\left(\Delta f\right) = E\{H(f,0)H^*(f-\Delta f,0)\}, \tag{2.5}$$

where $(\cdot)^*$ denotes the complex conjugate, and $H(f,t)$ is the channel time-variant transfer function [49]:

$$H(f,t) \quad = \quad \sum_{p=0}^{N_p-1} a_p(t)e^{j(2\pi(f_{D,p}t - f\tau_p(t)) + \varphi_p)}. \tag{2.6}$$

The coherence bandwidth B_c measures the spectral width of $|R\left(\Delta f\right)|$ over which the channel is considered frequency flat.

2.1.2.3 Doppler Spread

Similar to delay dispersion that causes channel frequency selectivity, frequency dispersion results in channel time selectivity. Doppler spread or frequency dispersion, describing the time varying nature of the channel, occurs by relative mobility of the transmitter and the receiver or the movement of objects in the environment. When a carrier frequency f_c is transmitted on a channel with Doppler frequency f_d, the received signal spectrum is spread over $f_c - f_d$ to $f_c + f_d$. This phenomenon is known as Doppler spread, which cause varying phase shift of the received signal. Such channel has a very short coherence time, i.e., the channel transfer functions variation with time is faster than the ones of the transmitted signal. The time correlation function [14, 48]

$$R\left(\Delta t\right) \quad = \quad E\{H(0,t)H^*(0,t-\Delta t)\} \tag{2.7}$$

quantifies the time varying nature of the channel. From $R\left(\Delta t\right)$, the channel coherence time T_c can be obtained, and it is defined as the time duration over which the channel is essentially flat [49]. If the signal duration T_s is greater than the coherence time of the channel, then the channel will change during the transmission of the baseband message, thus causing distortion at the receiver[47].

Coherence time Tc is the time domain dual of Doppler spread:

$$T_c \approx \frac{1}{f_d^{max}}. \tag{2.8}$$

f_d^{max} is the maximum Doppler spread. If the signal bandwidth is much greater than f_d^{max} the effects of Doppler spread are negligible at the receiver.

2.2 Multi-Carrier OFDM

2.2.1 OFDM Transmitter and Receiver

OFDM is an old technology with a history that goes back to 60s [50, 51]. While OFDM concept is simple, it took a while to find a widespread application in modern telecommunication networks due to implementation issues. Deploying inverse fast Fourier transform (IFFT)/fast Fourier transform (FFT) removed the requirement for a large number of sinusoidal generators in OFDM transmitters and separate filters for sub-carriers in OFDM receivers, which accelerated OFDM emerging in todays market.

The key concept in OFDM is to split a wide band signal into several orthogonal narrow band signals for transmission. In other words, instead of transmitting a volume of bits over a short time duration and a wide frequency band, it is transmitted over a long time duration and several narrow frequency bands. For this purpose, a baseband high data rate stream is divided into K parallel low data rate streams X_l, $l = 0, \cdots, K-1$, in an OFDM transmitter as shown in Fig. 2.3. X_ls are modulated with orthogonal sub-carriers by IFFT and a guard interval greater than the multipath time-spreading is added between the OFDM symbols to eliminate ISI. A cyclic extension of the OFDM symbol, i.e., a copy of the OFDM symbol in the interval $-T_g \leq t \leq 0$, named cyclic prefix, is inserted in the guard interval $T_s - T_g \leq t \leq T_g$, where T_s is the OFDM symbol time. OFDM symbols are modulated by a carrier frequency after passing a parallel to serial converter. At the receiver, the reverse action are taken place to reproduce the baseband high data rate stream. In the receiver side, channel estimation information are obtained and fed back to the transmitter for adaptive transmission schemes, such as adaptive modulation, channel coding, and power allocation.

2.2.2 PHY Layer Advantages of OFDM

Using large number of slow rate streams, which are carried by narrow band sub-carriers, increases robustness against frequency selective fading and immunity against impulsive

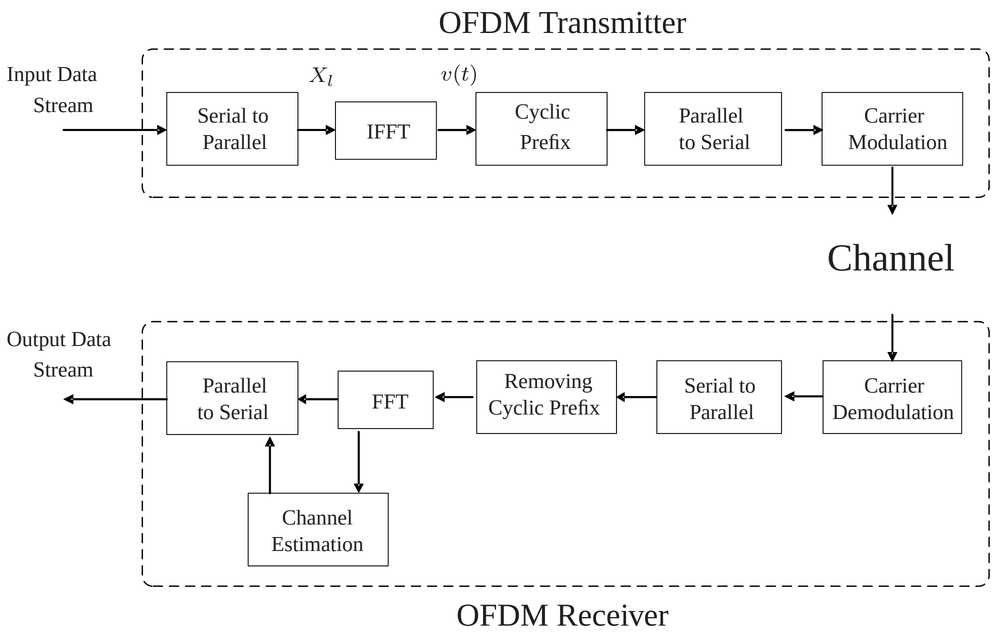

Fig. 2.3: An OFDM transceiver structure

noise. As sub-carriers bandwidth are narrow, the fading that they experience is flat. Also, due to enlarging symbols duration in time domain, OFDM symbol duration is much larger than multipath delay dispersion, which eliminates ISI. Eliminating ISI removes the requirements for equalization and reduces the complexity of an OFDM receiver. Orthogonality means sub-carriers are independent and each one can be adaptively coded and modulated. With orthogonal sub-carriers, there is no need for guard band between sub-carriers, to avoid ICI, because the peak of one sub-carrier occurs when other sub-carriers are at zero as shown in Fig. 2.4. Orthogonality allows the sub-carriers to overlap and save some bandwidth, so OFDM increases spectral efficiency in comparison to frequency division multiplexing (FDM). Cyclic prefix restore the orthogonality of sub-carriers at the receiver.

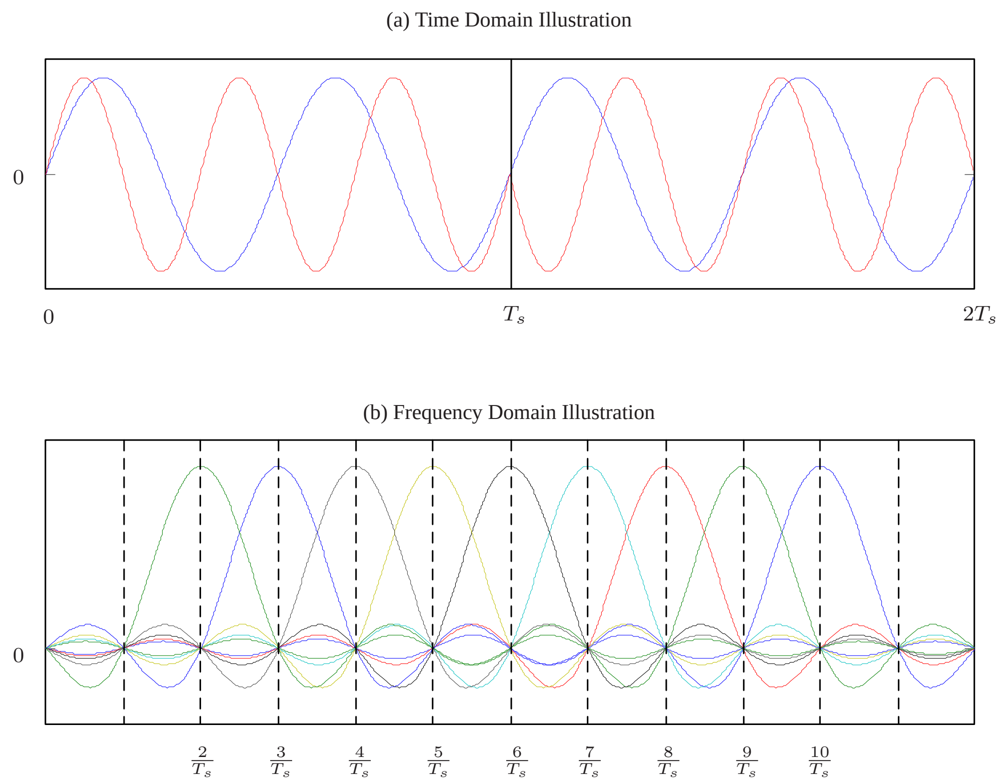

Fig. 2.4: Time and frequency illustration of OFDM-sub-carriers (a) two OFDM sub-carriers modulated by binary phase shift keying (BPSK) are illustrated over two OFDM symbols duration, (b) nine OFDM sub-carriers are illustrated in frequency domain

2.2.3 MAC Layer Advantages of OFDM

Originally, OFDM was proposed as a digital modulation or multiplexing technique, where all sub-carriers in an OFDM symbol carried only a user's data. However, OFDM can be used as a multi-user transmission technique when subsets of sub-carriers in an OFDM symbol are assigned to different users' transmission [14,52]. Multiuser transmission is possible because of the orthogonality of OFDM sub-carriers. Multiuser OFDM, denoted as OFDMA, is superior to traditional multiple access mechanisms such as TDMA and CDMA in terms of ability to exploit multiuser diversity [53]. OFDMA superiority in multiuser diversity gain stems from the fact that sub-carriers, which are the basic units of physical resources, i.e., time and frequency, are small. The fine granularity of resources units increases the flexibility of a resource allocation scheme.

Given a block of OFDMA symbols, the number of both symbols and sub-carriers can be dynamically assigned to each user. Dynamic sub-carrier assignment (DSA) achieves multiuser diversity gain. The multiuser diversity gain arises from the fact that the utilization of given resources varies from one user to another. A sub-carrier may be in deep fading for one user. Allocating this particular sub-carrier to the user with higher channel gain permits higher transmission rate. To achieve multiuser diversity gain, a scheduler at MAC is required to schedule users in appropriate frequency and symbols of an OFDMA block.

Another techniques that enhances the resource allocation schemes in MAC is adaptive modulation and coding (AMC) technique. AMC allows different modulation and coding to be used for the transmission on each sub-carrier. If some sub-carriers suffer from interference or attenuation, they can be allocated lower number of bits or they may not be used for transmission. On the contrary, sub-carriers with high channel gain are modulated by a higher order modulation and carry more bits per sub-carrier. The main objective of adaptive modulation and coding is to compensate for radio channel instability. It has been shown that adaptive modulation can effectively improve the bit error rate (BER) performance on radio channel which had suffered from shadowing and fading.

DSA and AMC are deployed at the transmitter when the fading channel is flat over a block of OFDMA symbols and a perfect CSI is available at the transmitter. Under these assumptions, the normalized transmission rate (bits/sec/Hz) on sub-carrier j is given by [54]:

$$r_j = \log_2\left(1 + p_j \frac{\alpha_j}{N_0}\right), \tag{2.9}$$

where p_j, α_j, and N_0 are, respectively, the allocated power to sub-carrier j, the channel gain of sub-carrier j, and AWGN spectral density. The Shannon capacity in equation (2.9) is an upper bound that asymptotically approaches the transmission rate over wireless channel. In practice, this upper bound is not achieved in networks because of using modulation and coding rates, which allow a specific number of bits is modulated and coded in each sub-carrier. Basically, given CSI, a proper modulation and coding rates can be chosen for the upcoming transmission so that the user bit rate can be maximized. An appropriate modulation and coding rate can be chosen from a lookup table. Also, for some M-ary modulation, such as M-QAM and M-PSK, where M represents the modulation level, approximate equations for obtaining M based on CSI and required bit error probability, P_b, exist. The approximations of the M-QAM and M-PSK P_b are, respectively, given by [55]:

$$P_b \approx \frac{4}{log_2 M} Q\left(\sqrt{\frac{3p_j \frac{\alpha_j}{N_0} log_2 M}{M-1}}\right) \tag{2.10}$$

$$P_b \approx \frac{2}{log_2 M} Q\left(\sqrt{2p_j \frac{\alpha_j}{N_0} log_2 M} \sin\left(\frac{\pi}{M}\right)\right). \tag{2.11}$$

In [56–58] equation(s) (2.10) and/or (2.11) are inverted to obtain the constellation size and power adaptation for a specific P_b. However, the $Q(\cdot)$ function cannot be easily inverted in practice, because numerical inversions are necessary [55]. Alternatively, the exact approximation can be written in a form that is easy to invert [59–62]. Because both modulation schemes are special cases of the M-ary modulation techniques [63], equations (2.10) and (2.11) can be written as

$$P_b \approx c_1 exp\left[\frac{-c_2 p_j \frac{\alpha_j}{N_0}}{2^{c_3 r_j} - c_4}\right], \tag{2.12}$$

where $r_j = log_2 M$ and $c_1 = 0.2$, $c_2 = 1.5$, $c_3 = 1$ and $c_4 = 1$ for M-QAM and $c_1 = 0.05$, $c_2 = 6$, $c_3 = 1.9$ and $c_4 = 1$ for M-QPSK [55]. Constants for different bounds can be found in [64]. By assuming "$=$" instead of "$\approx$" in (2.12) and solving for M, we obtain:

$$M = \sqrt[c_3]{\left(\frac{c_2}{-\ln(\frac{P_b}{c_1})} P_j \frac{\alpha_j}{N_0} + c_4\right)} \tag{2.13}$$

The adaptive modulation transmission rate as a function of P_b can be obtained by substituting (2.13) in $r_j = \log_2 M$:

$$r_j = \frac{1}{c_3} \log_2 \left(c_4 + \frac{c_2}{-\ln(\frac{P_b}{c_1})} P_j \frac{\alpha_j}{N_0}\right). \tag{2.14}$$

Note that the transmission rates equations (2.14) and (2.9) are similar. Thus, a resource allocation scheme that maximizes one of them maximizes the other [65]. This result broadens the applicability of resource allocation schemes to networks that adopt different modulation schemes.

2.3 Network Topology and Configuration

The network topology considered in this thesis is a PMP infrastructure, as shown in Fig. 2.5, which consists of a base-station (BS) and several users located in one hop neighborhood from the BS. The uplink (UL) channel, the transmission from users to the BS, is shared by all users, i.e., UL is a multiple access channel. On the other hand, the downlink (DL) channel, the transmission from the BS to users, is a broadcast channel. We consider resource allocation and scheduling on broadcast channel, which is part of the BS operations in this network.

This thesis considers a centralized resource allocation scheme, where the BS allocates OFDM sub-carriers and power to users based on CSI. Users estimate CSI and report it to the BS on each MAC frame. It is assumed that the estimation error is negligible and CSI remains constant during the next frame duration [66]. The BS determines sub-carrier assignments and power allocations based on CSI and broadcast an allocation vector on a

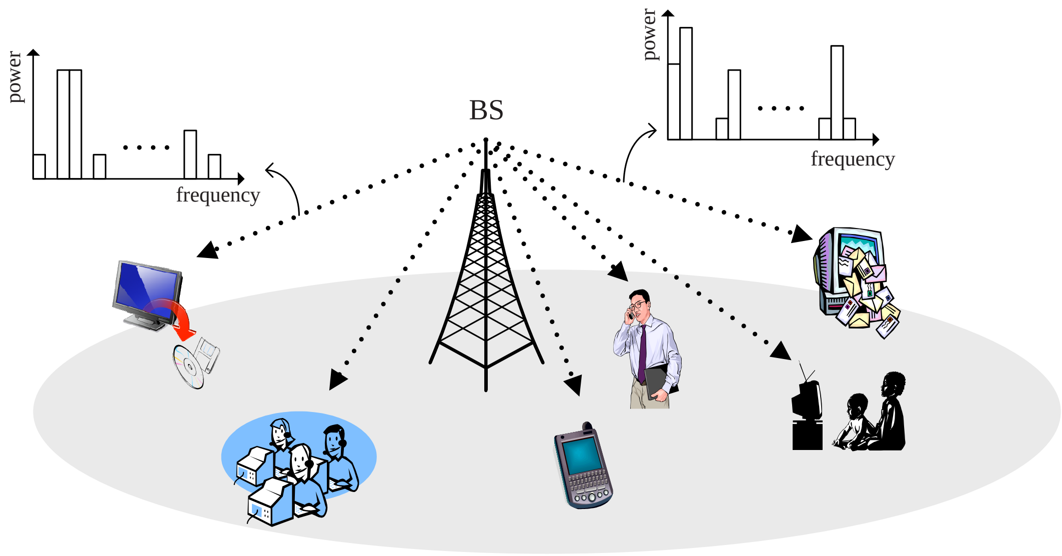

Fig. 2.5: Network platform is the DL of a PMP infrastructure where spectrum and power are allocated to users with heterogeneous service requirements.

signaling channel at the beginning of each MAC frame. In the following, some of PHY and MAC specification related to resource allocation problem formulation are reviewed.

2.3.1 PHY Layer

The PHY layer is responsible for raw bit transmission. We assume a single physical channel shared among all users and, hence, the channel access is controlled by a MAC protocol. The radio technology used in the physical channel can be any widely deployed one, such as WiFi or WiMAX. All users are equipped with identical communication devices and are capable of performing all the required networking functions and services.

For simplicity, ideal wireless channel without transmission error is assumed unless otherwise is mentioned. CSI is basic to achieving efficient resource allocation. The information is estimated at the receiver and fed back to the transmitter. As the characteristics of slow fading channel are different from fast fading channel for OFDM networks, different estimation algorithms should be used for each case [10]. Estimation algorithms take advantage of the correlation between time [67] or frequency [68] instances of channel to estimate the channel. As CSI in OFDM networks is presented in both time and frequency domain, a channel estimation algorithm for OFDM networks should consider both time and frequency domain characteristics. As the time correlation between symbols of a fast fading channel decreases with time faster than a slow fading channel, fast fading channel estimation is more complicated. We assume the channel estimation is taking effect through pilot assisted methods, i.e, the complex envelope of the fading channel is estimated using pilot symbols [69, 70]. As these methods give the channel estimation for pilot sub-carriers, the channel estimation of the other sub-carriers can be derived by interpolation.

2.3.2 MAC Layer

Radio resource allocation is part of the MAC sub-layer tasks in the current layered network architecture. MAC functionality in controlling access to shared resources will improve if it can acquire time-varying information of resources, mainly CSI, and users' requirements from PHY and upper layers, respectively. Adaptive resource allocation schemes deploy the provided information to smartly allocate the air links to users based on users' QoS requirements and channel quality. Therefore, cross-layer design and optimization across PHY and MAC are suggested for wireless resource allocation and scheduling schemes [71–73].

Using a cross-layer design between PHY and MAC, users' CSI is known at the beginning of each transmission frame. Upon receiving a feedback channel estimation, the BS makes new decisions for allocation of shared resources and informs users of the new assignment. The period of resource allocation fetching depends on the speed of variation of the fading channel. Adaptive resource management techniques are successfully applied to slow varying fading channel, such as in fixed or nomadic applications where the channel is static or quasi-static.

The BS broadcasts information based on OFDMA in the DL. Users' backlogged traffic, buffered in separate queues at the BS, are transmitted on assigned sub-carrier and allocated power determined by the resource allocation scheme. UL and DL subframes are interleaving in a time division duplexing (TDD) manner in a MAC frame as shown in Fig. 2.6. A MAC frame consists of frame header, DL and UL subfames, and guard bands. The frame header is used for synchronizing users with the BS and carrying users' profiles, e.g., the code rate or the sub-carrier allocated to each user. All MAC frames are assumed to have the same fixed length, which can be easily achieved in practice by commonly used link layer functions, such as fragmentation or concatenation of the upper layer packets.

2.4 Summary

In this chapter, the fading channel characteristics were explained, and the mathematical model of wireless fading channel was presented. Then, based on the channel model, we described how OFDM can improve communication over fading channel. An overview of

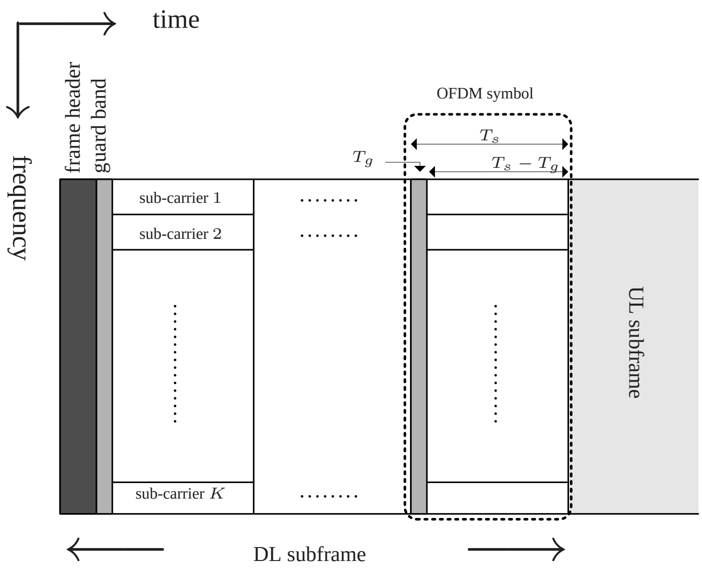

Fig. 2.6: OFDM symbols and sub-carriers in a MAC frame

the OFDM and OFDMA transceivers structures along with an explanation of their operations were presented. In addition, the required knowledge of PHY and MAC relevant to the resource allocation problem formulation was described, such as the relation among transmission rate, power, channel gain, bit error probability, and infrastructure used.

Chapter 3

A Framework for Resource Allocation in OFDMA Networks

Resource allocation is a very broad topic in telecommunication field due to the extended scope of targets, e.g., diverse service provisioning, different infrastructure accommodation, or mobility support. In this chapter, we consider resource allocation in multicarrier OFDMA networks when users have heterogeneous rate requirements. We investigate how the flexibility and granularity of OFDMA can be incorporated in a resource allocation scheme to improve network performance and resource utilization.

We formulate the joint optimization problem of sub-carrier assignment and power allocation in OFDMA networks as an MINLP problem first. A major challenge in solving the optimization problem is non-convexity caused by the combinatorial nature of sub-carrier assignment problem and/or non-convex objective functions. To avoid combinatorial optimization, we formulate the resource allocation as a nonlinear programming (NLP) with continuous variables. The problem formulation follows by a discussion about the complexity and performance of the proposed schemes exist in the literature. We suggest an approach based on PM/IPM to solve the NLP problem. Using a two-step implementation, first, the penalty method is applied to convert the non-convex feasible region to a convex one. Then, the interior point method is deployed to solve the new problem which is non-convex only in the objective function. To evaluate the performance of PM/IPM, we implement a genetic algorithm that achieves near optimal solutions of the problem by iterative searching. Numerical results are presented at the end of the chapter to demonstrate that PM/IPM can solve the problem within limited time while the solutions are close to the ones obtained by the genetic algorithm. In additions, the sensitivity of PM/IPM to users' channel gains and the effects of utility-based resource allocation are investigated.

3.1 Problem Formulation

In wireless OFDMA networks, sub-carrier assignment to users and power allocation to sub-carriers, referred as OFDMA resource allocation, affect the network performance significantly. In practice, to assign sub-carriers and allocate power efficiently[1], an OFDMA resource allocation is presented as an optimization problem whose objective function and constraints are determined based on users' requirements and network specifications. Depending on the definition of the objective functions, different utilization performance are expected. Resource allocation algorithms available in the literature focus on two main objectives: either data rate maximization or power minimization. Using a general objective function of rate, we present an optimization problem for sub-carrier assignment and power allocation constrained by the BS maximum power and users' minimum rate requirement. The problem formulation of power minimization is not discussed here. Interested readers are referred to [58] and [22].

The restrictions imposed by OFDMA networks specifications and users' requirements determine the feasible region, i.e., the set of feasible allocation that satisfy all constraints. Due to the exclusive sub-carrier assignment of OFDMA, the feasible region is discrete and consequently nonconvex. The objective function of the problem depends on users' demand and networks service providers' goals, which usually is a nonlinear

[1] An efficient resource allocation is the one that allocates as much resource as is needed by a user as long as resource is available.

function in practice. MINLP techniques are used when a discrete network structure and continuous parameters are simultaneously formulated [74]. Accordingly, most proposed schemes for the OFDMA resource allocation are based on MINLP. We review some of these schemes in section 3.2.

Following the work in the literature, first, we present an MINLP problem for the OFDMA resource allocation. The feasible region of the MINLP problem contains integer variables representing sub-carriers assigned to users and continuous variables representing power allocated to sub-carriers. Then, we prove that the set of constraints including the integer variables, in the MINLP problem, can be substituted by a set of nonlinear constraints with continuous variables. Accordingly, we present an NLP problem that unifies sub-carrier assignment and power allocation in a rate (or power) allocation problem. For more readability of formulas, the network parameters used in the optimization problems are given in Table 3.1.

3.1.1 MINLP and NLP Problems

We consider a network platform shown in Fig. 2.5, which consists of the BS and several users located in one hop neighborhood from the BS in a PMP infrastructure. The BS assigns sub-carriers to users and allocates a fraction of the BS total power, P_{BS}, to each user in each resource allocation interval. A solution of the resource allocation problem is denoted by a rate allocation vector r or a power allocation vector p as below:

$$r = [r_{11}, r_{12}, \ldots, r_{1K}, \ldots, r_{M1}, \ldots, r_{MK}]^T \tag{3.1}$$

$$p = [p_{11}, p_{12}, \ldots, p_{1K}, \ldots, p_{M1}, \ldots, p_{MK}]^T. \tag{3.2}$$

Similarly, a sub-carrier assignment vector is denoted by c, where

$$c = [c_{11}, c_{12}, \ldots, c_{1K}, \ldots, c_{M1}, \ldots, c_{MK}]^T \tag{3.3}$$

and c_{ij} is

$$c_{ij} = \begin{cases} 1 & \text{if sub-carrier } j \text{ is assigned to user } i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Table 3.1: Notations Descriptions

Notation	**Description**
M	total number of users in the network
K	total number of sub-carriers in the network
i	user index belongs to $\mathcal{M} := \{1, 2, ..., M\}$
j	sub-carrier index belongs to $\mathcal{K} := \{1, 2, ..., K\}$
α_{ij}	channel gain of user i on sub-carrier j
p_{ij}	allocated power to user i on sub-carrier j
r_{ij}	allocated rate to user i on sub-carrier j
R^i_{min}	minimum service rate requirement of user i
B	network bandwidth
P_{BS}	BS total power budget

Every user can use several sub-carriers, but each sub-carrier can be assigned to at most one user. Mathematically, this restriction is given by

$$\sum_{i=1}^{M} c_{ij} \leq 1 \quad \forall j \in \mathcal{K}. \tag{3.5}$$

If sub-carrier j has not been assigned to user i, then allocated power to user i on sub-carrier j must be zero. Therefore, for every user $i \in \mathcal{M}$ and every sub-carrier $j \in \mathcal{K}$, we must have the following condition:

$$\text{if } c_{ij} = 0 \ \text{ then } \ p_{ij} = 0. \tag{3.6}$$

We include this restriction in the optimization problem through the following constraint:

$$p_{ij} \leq P_{BS} c_{ij} \ \ \forall i \in \mathcal{M}, \ \forall j \in \mathcal{K}. \tag{3.7}$$

Note that, if $c_{ij} = 0$, (3.7) implies $p_{ij} \leq 0$ that along with the non-negativity constraint $p_{ij} \geq 0$ yields $p_{ij} = 0$ and satisfies (3.6). When $c_{ij} = 1$, (3.7) is reduced to the redundant constraint $p_{ij} \leq P_{BS}$, because of the existence of the following constraint, which assures total allocated power to the sub-carriers in each time slot is limited to P_{BS}:

$$\sum_{i=1}^{M}\sum_{j=1}^{K} c_{ij}p_{ij} \leq P_{BS}. \tag{3.8}$$

As (3.7) includes (3.6), variables c_{ij}'s can be removed from (3.8) as follows:

$$\sum_{i=1}^{M}\sum_{j=1}^{K} p_{ij} \leq P_{BS}. \tag{3.9}$$

If noise spectral density equals to one and rate adaptation is assumed to be continuous [47], the approximate transmission rate for user i on sub-carrier j, r_{ij}, is given by:

$$r_{ij} = \frac{B}{K}\log_2\left(1 + \alpha_{ij}p_{ij}\right). \tag{3.10}$$

Moreover, quality of service (QoS) requirements are projected on the objective function and constraints of the optimization problem. R^i_{min}, the minimum service rate requirement of user i with rate r_i is guaranteed through the following constraint:

$$r_i = \sum_{j=1}^{K} r_{ij} \geq R^i_{min} \quad \forall i \in \mathcal{M}.$$

Also, QoS requirements of users, in terms of rate, can be taken into account through users' utilities, which represent users' satisfaction of allocated rate. However, to present a general optimization problem that unifies most of the existing problems for OFDMA resource allocation, general objective function $\mathcal{F}(r)$, is used in this subsection. $\mathcal{F}(\mathbf{r})$ can be substitute by any function of rate, such as, sum of users' weighted rate, $\sum \omega_i r_i$, or sum of users' utilities, $\sum u_i(r_i)$, where ω_i and u_i are the assigned weight and utility to user i. The optimization problem Pr_1, which is an MINLP problem, is resulted:

$$Pr_1: \quad \max_{c,p} \quad \mathcal{F}(r) \tag{3.11}$$
$$\text{s.t} \quad r_{ij} = \frac{B}{K}\log_2(1+\alpha_{ij}p_{ij}) \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K}, \tag{3.12}$$
$$r_i = \sum_{j=1}^{K} r_{ij} \geq R^i_{min} \ \forall i \in \mathcal{M}, \tag{3.13}$$
$$\sum_{i=1}^{M}\sum_{j=1}^{K} p_{ij} \leq P_{BS}, \tag{3.14}$$
$$\sum_{i=1}^{M} c_{ij} \leq 1 \ \forall j \in \mathcal{K}, \tag{3.15}$$
$$0 \leq p_{ij} \leq P_{BS}c_{ij} \ \ \forall i \in \mathcal{M},\ \forall j \in \mathcal{K}, \tag{3.16}$$
$$c_{ij} \in \{0,1\} \ \ \forall i \in \mathcal{M},\ \forall j \in \mathcal{K}. \tag{3.17}$$

We eliminate integer variables c_{ij}'s and formulate the problem as a continuous nonlinear one-stage programming problem Pr_2:

$$Pr_2: \quad \max_{p} \quad \mathcal{F}(r) \tag{3.18}$$
$$\text{s.t} \quad r_{ij} = \frac{B}{K}\log_2(1+\alpha_{ij}p_{ij}) \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K}, \tag{3.19}$$
$$r_i = \sum_{j=1}^{K} r_{ij} \geq R^i_{min} \forall i \in \mathcal{M}, \tag{3.20}$$
$$\sum_{i=1}^{M}\sum_{j=1}^{K} p_{ij} \leq P_{BS}, \tag{3.21}$$
$$p_{\hat{i}j}p_{ij} = 0 \ \ \forall j \in \mathcal{K}, \forall i \in \mathcal{M} \setminus \{\hat{i}\}, \tag{3.22}$$
$$0 \leq p_{ij},\ \forall i \in \mathcal{M},\ \forall j \in \mathcal{K}. \tag{3.23}$$

Proposition 3.1.1 *There is a one-to-one correspondence between the set of feasible solutions of Pr_1 and the set of feasible solutions of Pr_2.*

We prove it by showing that from each feasible solution of Pr_2, a feasible solution of Pr_1 is obtained and vice versa.

Let p^* be a feasible solution of Pr_2. For every $i \in \mathcal{M}$ and $j \in \mathcal{K}$, define c^*_{ij} as follows:

$$c^*_{ij} = \begin{cases} 1 & \text{if } p^*_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.24}$$

Clearly p^* and c^* satisfy (3.12), (3.13), (3.14), (3.16), and (3.17). We claim that this solution also satisfies (3.15). If this is not true, there exists some $j \in \mathcal{K}$ so that $\sum_{i=1}^{M} c^*_{ij} \geq 2$. This implies that there are at least two i_1 and i_2 such that $c^*_{i_1 j} = c^*_{i_2 j} = 1$. However, the derivation of c^*_{ij} from p^*_{ij} in (3.24) yields $p^*_{i_1 j} > 0$ and $p^*_{i_2 j} > 0$. Hence $p^*_{i_1 j} p^*_{i_2 j} > 0$ which is in contradiction to the fact that p^* satisfies (3.22). So p^*, c^* must also satisfy (3.15).

Next, assume that (p^*, c^*) is a feasible solution of Pr_1. Thus p^* satisfies (3.19), (3.20), (3.21), (3.23). If p^* does not satisfy (3.22), then there must be $\bar{i}, \check{i} \in \mathcal{M}$ and $\bar{j} \in \mathcal{K}$ such that $p^*_{\bar{i}\bar{j}} p^*_{\check{i}\bar{j}} > 0$ or equivalently $p^*_{\bar{i}\bar{j}} > 0$ and $p^*_{\check{i}\bar{j}} > 0$ for some $\check{j}$. Constraint (3.16) implies that $c^*_{\bar{i}\bar{j}} = 1$ and $c^*_{\check{i}\bar{j}} = 1$. Thus $\sum_{i=1}^{M} c^*_{i\bar{j}} \geq c^*_{\bar{i}\bar{j}} + c^*_{\check{i}\bar{j}} \geq 2$, which is in contradiction to the assumption that $(p^*,\ c^*)$ satisfies (3.15). Thus p^* also satisfies (3.22) and therefore, is a feasible solution of Pr_2. For every feasible solution of Pr_1 and associated feasible solution of Pr_2, the rate allocation vectors are identical. Thus, $Proposition\ 2.1$ implies there is a one-to-one correspondence between the set of optimal solutions of Pr_1 and Pr_2; As a result, they have the same optimal value.

Problem Pr_2 can be written only in terms of allocated rate r_{ij}, if an equivalent constraint of r_{ij} replaces constraint (3.22). It can be shown that the following constraints are equivalent to (3.22):

$$\begin{aligned}
&(a) \quad r_{\hat{i}j} r_{ij} = 0 \ \ \forall j \in \mathcal{K}, \forall i \in \mathcal{M} \setminus \{\hat{i}\}, \\
&(b) \quad r_{\hat{i}j} + r_{ij} = \max\{r_{\hat{i}j}, r_{ij}\} \ \ \forall j \in \mathcal{K}, \forall i \in \mathcal{M} \setminus \{\hat{i}\}, \\
&(c) \quad |r_{\hat{i}j} - r_{ij}| = r_{\hat{i}j} + r_{ij} \ \forall j \in \mathcal{K}, \forall i \in \mathcal{M} \setminus \{\hat{i}\}, \\
&(d) \quad (r_{\hat{i}j} - r_{ij})^2 = (r_{\hat{i}j} + r_{ij})^2 \ \forall j \in \mathcal{K}, \forall i \in \mathcal{M} \setminus \{\hat{i}\}.
\end{aligned}$$

We use (a) in the rest of the chapter, because they are differentiable and have a simple representation. Thus, Pr_2 can be restated as follows:

$$Pr_3: \quad \max_{r} \quad \mathcal{F}(r) \tag{3.25}$$

$$\text{s.t} \quad \sum_{j=1}^{K} r_{ij} \geq R^{i}_{min} \, \forall i \in \mathcal{M}, \tag{3.26}$$

$$\sum_{i=1}^{M} \sum_{j=1}^{K} \frac{1}{\alpha_{ij}} (2^{\frac{r_{ij}K}{B}} - 1) \leq P_{BS} \ \ \forall i \in \mathcal{M}, \ \forall j \in \mathcal{K}, \tag{3.27}$$

$$r_{\hat{i}j} r_{ij} = 0 \ \ \forall i \in \mathcal{M} \setminus \{\hat{i}\} \, \forall j \in \mathcal{K}, \tag{3.28}$$

$$0 \leq r_{ij}, \ \forall i \in \mathcal{M}, \ \forall j \in \mathcal{K}. \tag{3.29}$$

As the objective function is continuous over the range of r and the feasible region of Pr_3 is closed and bounded, the *extreme value theorem* (*Weierstrass Theorem*) [75] implies that Problem (Pr_3) has global optimal solution(s):

Theorem 3.1.1 (extreme value theorem) *Let f be a continuous real-valued function whose domain, D_f, is bounded and closed. Then there exist x_1 and x_2 in D_f such that:*

$$f(x_1) \leq f(x) \leq f(x_2) \ \ \forall x \in D_f.$$

Although *Weierstrass Theorem* guarantees that the global optimal solution exists, finding such a global solution for a general continuous objective function is hard, i.e., there is no polynomial time algorithm for obtaining the global optimal solution.

3.2 Related Works and Problem Complexity

In general, objective function $\mathcal{F}$ is a function of users' rates. The choice of $\mathcal{F}$ along with the set of constraints affect both computational complexity of Pr_3 and the network performance. The following discussion will provide an insight into the problem in terms of achievable performance and complexity for different objective functions and constraints.

3.2.1 Linear Objective Function

Common linear objective functions, used in the OFDMA resource allocation problems, are $\mathcal{F}(r) = \sum_i r_i$ and $\mathcal{F}(r_i) = \sum_i \omega_i r_i$. The former, known as bit rate maximization problem, maximizes total users' data rate and the later, known as weighted rate maximization problem, maximizes aggregate users' rate multiplied by a vector of weights, ω_i's, subject to a given power budget. Bit rate maximization problem, is the most common objective function deployed in [11, 23–29, 62]. [23] and [24] consider joint sub-carrier and power allocation with power constraint as an MINLP problem. [25] formulates the problem by allowing a sub-carrier to be shared by multiple users. The optimization problem is decoupled into two subproblems, sub-carrier assignment to users and power allocation to sub-carriers, and a two-step algorithm is proposed for solution. In the first step, a sub-carrier is assigned to only one user who has the best channel gain on that sub-carrier. In the second step, the amount of transmit power to be allocated to each sub-carrier is determined by water-filling scheme [76] to maximize overall data rate. To reduce computational complexity of water-filling, equal power allocation scheme may be adopted. It has been shown that water-filling and equal power allocation schemes have only marginal performance difference [77]. Accordingly, a suboptimal solution in [26] allocates uniform power to sub-carriers. Given the channel gain and the fixed power allocation (P_{BS}/K), sub-carriers' rates (r_{ij}'s) are known. The problem is converted into a linear integer programming (LIP) problem with integer variables c_{ij}'s. Then a reduced computational complexity algorithm is deployed to solve LIP by, first, assigning sub-carriers to maximize total users' data rates, irrespective of users' minimum required data rate constraints, and, second, adjusting sub-carriers assignment to satisfy users' minimum required rate constraints.

A geometric programming (GP), a special form of convex optimization, has been proposed in [78] for weighted rate maximization or weighted power minimization. There exist several algorithms to solve GP efficiently and optimally. However, GP is not applicable in some OFDMA resource allocation problems because converting or approximating objectives and constraints to be compatible with GP [79] is challenging.

3.2.2 Nonlinear Objective Functions and Constraints

To providing QoS and fairness or maximizing resource utilization, some OFDMA resource allocation schemes have been proposed that use nonlinear objective functions or add a set of nonlinear constraints in the optimization problem.

3.2.2.1 Nonlinear Objectives

The objective functions can be chosen properly to achieve some specific objects.

Max-min fairness solution is addressed in [29] by maximizing the minimum users' data rates, i.e., $\max \min r_i$. A convex feasible region is obtained for the problem by relaxing the constraint of exclusively allocating one sub-carrier to only one user. Assuming equal amount of power is allocated to each sub-carrier, [29] proposes an algorithm to assign sub-carriers to users.

Rate proportional fairness schemes have been proposed in [32, 80]. A set of rates that maximizes aggregate logarithms of users' data rates is rate proportional fair. This set of rates is chosen as a fair weight allocation set and is deployed in a scheduling scheme that determines users' transmissions order according to users' channel gains and fair weights.

An appropriate form of the objective function in networks with heterogeneous traffic is to maximize users' aggregate utility functions. Assuming concave or linear utility functions, [30, 81] investigate the utility-based resource allocation in OFDMA networks for both discrete and continuous adaptive rate. The optimization problem is decomposed into two problems: DSA and APA. The DSA problem is represented as a uniform power allocation problem, and the APA problem is represented as a fixed sub-carrier assignment problem. Different approaches are proposed for solving DSA, APA, and joint DSA/APA problems. DSA is relaxed to a nonlinear integer (binary) problem, and a sorting search algorithm is proposed for sub-carrier assignment. When all utility functions are linear or sub-carriers bandwidth is small enough to be considered infinitesimal (rate region is concave), sorting search algorithm gives optimal solutions. Otherwise, the solution is suboptimal, and sorting search algorithm only reduces the computation complexity. A sequential-linear-approximation water-filling algorithm is proposed to solve the APA continuous rate adaptation. The relaxed nonlinear concave problem is approached by a series of linear optimization problems derived by a sequential-linear-approximation algorithm named Frank-Wolfe method [82]. For APA with discrete rate adaptation, a greedy algorithm is deployed to allocate bits and the corresponding power. In each bit loading iteration, the greedy algorithm allocates power to some sub-carriers that maximize the utility argument per power. Assuming concave utility functions, the greedy algorithm results in optimal bit loading and power allocation. Finally, a joint DSA and APA solution is proposed for the original problem. For continuous rate adaptation, combinations of iterative sub-carrier assignment, power allocation, and the updates of marginal utilities are deployed. A new sub-carrier assignment is derived based on the sub-gradient of concave utility functions; the corresponding power allocation is determined by linear approximation of the objective function; the algorithm stops when the marginal utility function is negligible. For discrete rate adaptation, a combination of sorting-search DSA and the greedy APA algorithm is deployed.

3.2.2.2 Nonlinear Constraints

A set of constraints can be added to the problem to force a notion of fairness or QoS.

To resolve unfair rate allocation of bit rate maximization problems and balance between capacity and fairness, [27] formulates the problem by adding a set of nonlinear constraints which assures proportional users' data rates. The primal solution of the constrained fairness problem is computationally complex to be obtained, so a low-complexity suboptimal algorithm that separates sub-carrier assignment and power allocation is proposed. The decoupled allocation algorithm, first assigns sub-carriers assuming uniform power allocation. Then, an optimal power allocation algorithm maximizes total capacity while maintaining proportional fairness.

An alternative way of fair allocation of resources is combining a fair scheduling algorithm with resource allocation techniques. [28] develops a resource management scheme by integrating DSA and generalized processor sharing (GPS) scheduling to maximize network throughput subject to the constraints on the total transmit power, user's SNR requirement, and GPS fair scheduling. A fixed modulation level has been considered for all sub-carriers. At the first step of the algorithm, the number of sub-carriers allocated to users are determined with a modified GPS scheduling based on users' required rate and fairness constraint. At the second step, an algorithm is deployed to determine the set of required number of sub-carriers of each user, derived in the step one. As a user with a higher SNR requirement consumes more power, sub-carriers with the largest channel gain are assigned to users with the highest SNR requirement, as long as the total transmission power for each user does not exceed the total transmission power constraint. Also, the principle of generalized processor sharing is deployed as a constraint of the optimization problem in [62] to allocate sub-carriers fairly among users.

Furthermore, an associated set of constraints to a specific QoS characteristic can be include to guarantee users' required QoS, e.g., [11] provides users' minimum rate requirements, and [28] guarantees tolerable signal to noise ratio of users' receivers by including corresponding rate and signal to noise ratio constraints to the optimization problem.

3.2.3 Problem Complexity

Note that when either a set of constraints is added to the problem or a nonlinear objective function is deployed, the problem remains nonconvex. The complexity of the problem is caused by nonconvexity of the feasible region and/or non-concavity of the objective function. The sets of c_{ij}'s and p_{ij}'s in the MINLP problem, Pr_1, as well as p_{ij}'s and r_{ij}'s in the NLP problem, Pr_2 and Pr_3 are nonconvex. The nonconvexity arises from the fact that a sub-carrier should be allocated exclusively to one user. For example, consider two feasible allocation power vectors $p = [1, 0, 0, 0, 1, 1]$ and $\hat{p} = [0, 1, 1, 1, 0, 0]$ in a simple network which consists of two users and three sub-carriers. For $\alpha \in (0, 1)$, the convex combination of p and $\hat{p}$, which is

$$\alpha p + (1-\alpha)\hat{p} = [\alpha, (1-\alpha), (1-\alpha), (1-\alpha), \alpha, \alpha], \tag{3.30}$$

does not belong to the feasible region, and the definition of convex feasible region is not held. An optimization problem whose objective function is non-concave (in a maximization problem) and its feasible region is nonconvex, is categorized among nonconvex optimization problems, which are difficult to be solved for a global optimum.

In general, nonconvex optimization problems are NP-hard [83], and there is no polynomial time algorithm to find their global optimum has been found yet. Therefore, OFDMA resource allocation problems can be solved for a local optimal solution by exhaustive search algorithms. Search algorithms span almost the entire feasible region of the problem to find the highest local maximum (or lowest local minimum). As they do not stop searching when they find a local optimum, it is expected that the algorithms achieve near optimal solutions when searching time approaches infinity. However, the long response time of search algorithms limits their usage and is a barrier in developing elaborated OFDMA resource allocation schemes, while OFDMA is emerging in broadband wireless networks, and the OFDMA resource allocation arises in many contexts. This motivates us to investigate continuous optimization approaches, rather than discrete methods, that can treat the nonconvexity of the OFDMA resource allocation problem. To the best of our knowledge, using continuous optimization approaches for the OFDMA resource allocation problem has not been addressed in the literature yet.

3.3 Penalty Function and Interior Point Methods

We propose an interior point based approach to solve the OFDMA resource allocation problem. We were motivated by the increasing trend toward improving the interior point theory and methods and applying them on new problems. Specifically, it is highly expected that interior point methods will be helpful in solving MINLP problems [84] and are successful in solving continuous nonlinear problems, particularly with convex feasible regions [31, 85]. We apply the proposed method to solve Pr_3 which contains continuous variables only, i.e., is an NLP problem. The success of interior point methods in solving a nonconvex nonlinear problems strongly depends on how nonconvexity of the problem is treated. Although the proposed formulation for problem Pr_3 is continuous, the feasible region of the problem is nonconvex yet. So, we use a penalty function method to remove the nonconvexity of the feasible region. More precisely, nonconvex constraints are moved to the objective function by a coefficient penalty. We apply the proposed penalty function method combined with an interior point method to solve the NLP problem for the OFDMA resource allocation problem.

3.3.1 PM/IPM Descriptions

In Pr_3, all constraints except (3.28) are convex. We add this set of constraints to the objective function as a penalty term, which is negative when one of the constraints in (3.28) is violated, and zero otherwise. After adding the penalty term to the objective function, the new objective function becomes:

$$P_L \max_r \quad f(r) = \mathcal{F}(r) - \frac{L}{2} \sum_{\hat{i}=1}^{M} \sum_{i=1, i \neq \hat{i}}^{M} \sum_{j=1}^{K} r_{\hat{i}j} r_{ij}, \tag{3.31}$$

where positive constant L is the penalty parameter. The new objective function along with the constraints of Pr_3 form the following problem:

$$Pr_4 : \quad \max_r \quad f(r) \tag{3.32}$$

$$\text{s.t} \quad C(r) \geq 0, \tag{3.33}$$

where $C(r)$ is the vector of inequality constraints (3.26), (3.27) and (3.29), which is represented as follows:

$$C(r) = \begin{pmatrix} \sum_{j=1}^{K} r_{1j} - R_{min}^{1} \\ \vdots \\ \sum_{j=1}^{K} r_{Mj} - R_{min}^{M} \\ -\sum_{i=1}^{M}\sum_{j=1}^{K} \frac{1}{\alpha_{ij}}(2^{\frac{Kr_{ij}}{B}} - 1) + P_{BS} \\ r_{11} \\ \vdots \\ r_{MK} \end{pmatrix}. \tag{3.34}$$

Instead of solving Pr_3, we solve Pr_4 whose feasible region is convex. However, an optimal solution of Pr_4 with a positive L will not be an optimal solution of Pr_3, unless the (positive) penalty term is zero. By making L larger, we penalize constraint violations more severely, thereby forcing the minimizer of the penalty function to be smaller. We formally prove this statement in the following proposition:

Proposition 3.3.1 *The value of penalty term $\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K} r_{\hat{i}j}r_{ij}$ at an optimal solution of Problem P_L decreases, as L increases.*

Let L_1 and L_2 be two penalty parameters so that $L_1 \leq L_2$. Denote optimal solutions of Problems P_{L_1} and P_{L_2}, with r_1 and r_2, respectively. Since r_1 is an optimal solution associated with parameter L_1, the value of the objective function of P_{L_1} at r_1 is larger than the value of the objective function of P_{L_2} at r_2, so

$$\mathcal{F}(r_2) - \frac{L_1}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij} \leq \mathcal{F}(r_1) - \frac{L_1}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij}, \tag{3.35}$$

and consequently

$$\frac{L_1}{2}\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij}\right) \leq \mathcal{F}(r_1) - \mathcal{F}(r_2). \tag{3.36}$$

Similarly, since r_2 is an optimal solution of P_{L_2}, the value of the objective function of P_{L_2} at r_2 is greater than its value at r_1. Hence

$$\mathcal{F}(r_1) - \frac{L_2}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij} \leq \mathcal{F}(r_2) - \frac{L_2}{2}\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij}, \quad (3.37)$$

and consequently

$$\frac{L_2}{2}\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij}\right) \geq \mathcal{F}(r_1) - \mathcal{F}(r_2). \quad (3.38)$$

Inequalities (3.36) and (3.38) imply that

$$\begin{aligned}\frac{L_2}{2}&\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij}\right) \geq \mathcal{F}(r_1) - \mathcal{F}(r_2)\\ &\geq \frac{L_1}{2}\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij}\right). \end{aligned} \quad (3.39)$$

Hence

$$\left(\frac{L_2}{2} - \frac{L_1}{2}\right)\left(\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij} - \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij}\right) \geq 0. \quad (3.40)$$

Using the assumption that $L_1 \leq L_2$, we have

$$\sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_1)_{\hat{i}j}(r_1)_{ij} \geq \sum_{\hat{i}=1}^{M}\sum_{i=1,i\neq\hat{i}}^{M}\sum_{j=1}^{K}(r_2)_{\hat{i}j}(r_2)_{ij}, \quad (3.41)$$

which completes the proof.

Therefore, the larger L is, the more penalized the constraint violations of penalty term is, and the smaller the penalty term will be. Indeed, it is shown in Theorem 17.1 of [31] that for a large enough choice of L, global optimal solution(s) of Pr_4 is (are) optimal solution(s) of Pr_3. However, the maximization of $f(r)$ in P_L becomes more difficult to perform as L becomes large [31]. In this thesis, we find an appropriate value for L through a simple search method.

Even though the objective function of Pr_4 is a non-concave nonlinear function, but its feasible region is convex. Convexity of the feasible region motivates us to use some interior point methods to solve Pr_4.

Before applying the interior point method, we first convert the inequality constraints in $C(r)$ to equality constraints by associating a positive slack variable to each constraint. Denote the $(2M+1)K$ vector of slack variables with s. Hence, Pr_4 is converted to the following minimization problem:

$$Pr_5: \quad \min_{r} \quad -f(r) \tag{3.42}$$

$$\text{s.t} \quad C(r) - s = 0, \tag{3.43}$$

$$s \geq 0. \tag{3.44}$$

A necessary condition for a feasible solution of Pr_5 to be optimal is to satisfy the following conditions, called Karush-Kuhn-Tucker (KKT) conditions:

$$\nabla f(r) - A^T(r)z = 0, \tag{3.45}$$

$$C(r) - s = 0, \tag{3.46}$$

$$Sz = 0, \tag{3.47}$$

$$s \geq 0, \;\; z \geq 0\,. \tag{3.48}$$

In the aforementioned KKT conditions, S is a diagonal matrix with diagonal elements given by vector s, and vector z contains $(2M+1)K$ Lagrange multipliers used in the definition of the Lagrangian function of Pr_5:

$$\mathcal{L}\left(r,s,z\right) = f(r) - z^T\left(C(r) - s\right). \tag{3.49}$$

The matrix A in (3.45) is the Jacobian matrix of $C(r)$ represented by:

$$A = \begin{pmatrix} & \Theta & \\ \frac{-K\ln(2)2^{\frac{Kr_{11}}{B}}}{B\alpha_{11}} & \cdots & \frac{-K\ln(2)2^{\frac{Kr_{MK}}{B}}}{B\alpha_{MK}} \\ & I & \end{pmatrix}, \tag{3.50}$$

where I is an identity matrix of dimension $MK \times MK$, and Θ is the following $M \times MK$ matrix:

$$\Theta = \begin{pmatrix} 1_{(1,K)} & 0_{(1,K)} & \dots & 0_{(1,K)} \\ 0_{(1,K)} & 1_{(1,K)} & \dots & 0_{(1,K)} \\ \vdots & \vdots & & \vdots \\ 0_{(1,K)} & 0_{(1,K)} & \dots & 1_{(1,K)} \end{pmatrix}, \tag{3.51}$$

where $1_{(1,K)}$ and $0_{(1,K)}$ are K vectors of 1 and 0, respectively.

To find an approximation for a local optimum of a nonlinear problem, interior point methods solve a series of perturbed KKT conditions in which only the right-hand-side in equation (3.47) is replaced by a vector μe:

$$\nabla f(r) - A^T(r)z = 0, \tag{3.52}$$

$$C(r) - s = 0, \tag{3.53}$$

$$Sz = \mu e, \tag{3.54}$$

$$s \geq 0, \quad z \geq 0\ , \tag{3.55}$$

with $e = (1, 1, ..., 1)^T$ and $\mu > 0$. Interior point methods start with an initial interior point in the feasible region that satisfies perturbed KKT conditions for some μ and proceeds to find another interior point that satisfies perturbed KKT conditions (3.52)-(3.55) for a smaller value of μ. As the method proceeds, μ is decreased, and consequently the solution of the perturbed KKT conditions approaches the solution of the KKT conditions, in which $\mu = 0$. It is expected that after several iterations the solution will converge to a point that satisfies the KKT conditions of the problem [31].

In each iteration of the interior point method, directions and lengths of movements are updated based on the first and second order gradients of the objective function and constraints. The vector of movement directions for variables r, s, and z, denoted by $b = [b_r, b_s, b_z]^T$, is computed by solving the following linear system of equations:

$$\begin{pmatrix} \nabla^2_{rr}\mathcal{L} & 0 & -A^T(r) \\ 0 & Z & S \\ A(r) & -I & 0 \end{pmatrix} \begin{pmatrix} b_r \\ b_s \\ b_z \end{pmatrix} = \begin{pmatrix} \nabla_r u(r) - A^T(r)z \\ Sz - \mu e \\ C(r) - s \end{pmatrix},$$

Here, Z denotes the diagonal matrix whose diagonal elements are given by vector z. As matrices $\nabla^2_{rr}\mathcal{L}$ and $\nabla_r f(r)$ depend on the objective function chosen for the problem, we provide their descriptions in Appendix, section A.1, for a chosen objective function.

After obtaining movement directions, the length of movement in each direction, step length, denoted by α_s^{max} and α_z^{max}, are specified as below:

$$\alpha_s^{max} = \max\left\{\alpha \in (0,1] : s + \alpha b_s \geq (1-\tau)\, s\right\}, \tag{3.56}$$

$$\alpha_z^{max} = \max\left\{\alpha \in (0,1] : z + \alpha b_z \geq (1-\tau)\, z\right\}, \tag{3.57}$$

where $\tau \in (0,1)$. A large value of τ close to one, e.g., $\tau = 0.995$, is usually chosen to avoid s and z approaching zero too quickly. Now, the new interior point, slack variables, and Lagrange multipliers, (r^+, s^+, z^+), are determined with the information of movement directions and step lengths accordingly:

$$r^+ = r + \alpha_s^{max} b_r, \tag{3.58}$$

$$s^+ = s + \alpha_s^{max} b_s, \tag{3.59}$$

$$z^+ = z + \alpha_z^{max} b_z. \tag{3.60}$$

For the next iteration, μ is updated to a smaller value, say $\mu^+ < \mu$. There are several strategies to choose μ^+. Among them we use a linear method to update μ:

$$\mu^+ = \sigma\mu \qquad \sigma \in (0,1). \tag{3.61}$$

Since $\sigma < 1$, μ approaches zero over several iterations. However, choosing a very small σ or a very large σ will cause faster or slower convergence, respectively. Although fast convergence is always desired, it may cause some parameters, such as s and z, to approach zero too quickly, which reduces the performance of the method, e.g., the offered solution may be infeasible or far from optimality.

The interior point method is terminated when a stopping criterion is achieved. In this work, the initial value of $\mu_0 = 1$ has been chosen, and when μ approaches a very small value or the change in allocated rate vector, r, is negligible, the method stops. Algorithm 1 presents a summary of the interior point method implementation steps used in our simulation.

Algorithm 1 The solution algorithm for Pr_5

Input: $M, K, P_{BS}, B, \alpha, U_i,$ *initial_r*, s_0, μ_0, τ, σ

Output: r

Setting up and initialization:

Choose *initial_r* and compute $s_0 > 0$.

Choose $\mu_0 > 0$ and compute $z_0 > 0$ accordingly.

Set parameters $\tau \in (0, 1)$ and $\sigma \in (0, 1)$.

Set $k = 0$ and $Exit\_flag = 0$.

Interior point method main loop:

while $Exit\_flag == 0$ **do**

  Solve (3.56) to obtain movement direction $b = (b_r, b_s, b_z)$.

  Compute α_s^{max}, and α_z^{max} using (3.56) and (3.57).

  Compute $(r^{k+1}, s^{k+1}, z^{k+1})$ using (3.58) to (3.60).

  Set $\mu^{k+1} \leftarrow \mu^k$ and $k \leftarrow k + 1$.

  Compute $Exit\_flag$.

end while

return r.

3.4 Genetic Algorithm

In our simulation, we use GA as an intelligent search algorithm to find near-optimal solutions. GA is a randomized adaptive search method that processes a large number of search points at each iteration, then generates a new set of feasible points based on characteristics of the old search points. GA deploys a randomization search technique that avoids searching process being stopped when a local optimum is attained and continues searching the feasible region for a better local optimum [86]. Also, adaptive search based on the previous search points limits computational complexity, i.e., the computational burden does not necessarily increase with an increase in dimensions of search region [87].

3.4.1 Genetic Algorithm Methodology

In GA context, feasible solutions of a problem are represented by a data structure named (chromosome), and a fitness function is defined to evaluate feasible solutions. The algorithm begins with forming an initial population (first generation) of random feasible solutions. Then, the initial population is improved toward the optimal solution by generating a new population from the current chromosomes through several iterations. The evolution is in favor of chromosomes with better fitness values, because they are more likely to be inherited to the next generation. The new population is generated in each iteration through the following operators:

- **Selection:** The operator chooses better chromosomes of current generation to form a population of parent chromosomes. The larger is the fitness value of a chromosome, the higher is the probability of it being selected as a parent.

- **Crossover:** The operator generates new chromosomes (children) from parents chosen by selection operator. A crossover between a pair of parents is performed by selecting a point on the chromosomes of the two parents and swapping the chromosomes beyond that point.

- **Mutation:** The operator probabilistically changes an arbitrary element of a chromosome to a new value. Mutation avoids the algorithm stopping in a local optimum by generating new chromosomes which may have a better fitness value than

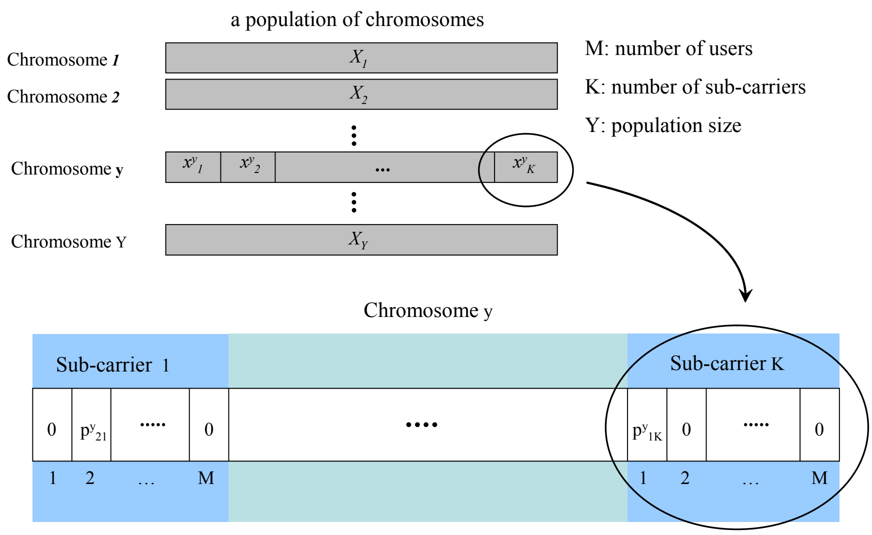

Fig. 3.1: The population and chromosomes representations

the ones of the chromosome of the current local optimum.

3.4.2 Genetic Algorithm Implementation

The specifics of chromosomes and fitness function as well as operators implementation depend on the problem to be solved. A $K \times M$ vector is chosen for the chromosome in our implementation, where K and M are the numbers of sub-carriers and users, respectively. Chromosome y of the population is a vector $[x_1^y \cdots x_j^y \cdots x_K^y]$ of x_j^y, where $j \in \mathcal{K}$ represents a sub-carrier index, as shown in Fig. 3.1. x_j^y is a $1 \times M$ allocation vector of a continuous value x_{ij}^y, where $i \in \mathcal{M}$ is a user's index, that shows allocated power to user i on sub-carrier j, p_{ij}^y. Each x_j^y contains only one non-zero element, x_{ij}^y, due to the constraint of exclusive sub-carrier assignment to a user.

An initial population, $\mathcal{P}_0$, of N chromosomes is formed by allocating a random user to each sub-carrier of each chromosome. The minimum required power, that satisfies user' minimum required rate, is assigned to the users that are allocated to sub-carriers in initial population. Each chromosome is a feasible solution, so it should satisfy all the constraints of the problem. If a chromosome does not satisfy the problem constraints, the procedure of chromosome generation will be repeated. The fitness function is the objective function of the optimization problem. Selection operator is a fitness proportionate selection, also known as roulette-wheel selection, that selects individuals with a probability proportional to their fitness values. This selection operator gives a chance to weak solutions (low fitness values) to be selected, hoping that those weak solutions will result in some good solutions (high fitness value) in crossover operation. Using a uniform distribution, p_{cross}, a point j from $\{M, \cdots, (K-1)M\}$ is chosen for crossover operation. In other words, crossover is performed over sub-carriers. Mutation operation chooses a mutating element from $\{1, \cdots, KM\}$ with a uniform distribution, p_{mut}. Actually, the mutating element indicates a new user i for sub-carrier j, so allocated power to the previous user of sub-carrier j is altered to zero, and a random power is allocated to the mutating element. Crossover and mutation are repeated if new generated chromosomes do not satisfy the problem constraints. Once a new population $\mathcal{P}_n$ is generated through selection and crossover and mutation, it replaces the old one. However, as the chromosome with the best fitness value, referred to as *elite*, may be lost in selection, crossover, and mutation operators, an elitism operation is performed before substituting $\mathcal{P}_{v-1}$ with $\mathcal{P}_v$. Elitism operation substitutes the corresponding chromosome to the least fitness value of $\mathcal{P}_v$ with *elite*. GA stops after $\mathcal{N}_{itr}$ iterations or when there is no increment in *elite*'s fitness value for $\mathcal{N}_{fit}$. Numerical parameters of GA are listed in Table 3.2 and the pseudo code of the solution is outlined in Algorithm 2.

3.5 Numerical Results

In this section, the convergence of GA is investigated in 3.5.1, which then will be used as a benchmark to evaluate the performance of PM/IPM in terms of optimality and sensitivity to network parameters in 3.5.2. As the focus of the resource allocation in this thesis is on utility maximization problems, in 3.5.3, we demonstrate how the resource utilization performance is enhanced in utility-based resource allocation problem.

In our simulation, traffic arriving at the BS is first buffered in separate infinite queues

Algorithm 2 GA implementation for the problem

Input: $M, K, \mathcal{N}_{itr}, \mathcal{N}_{fit}, p_{\mathit{cross}}, p_{\mathit{mut}}, P_{BS}, B, \alpha, \mathcal{F}$

Output: p_{ij}

Setting up and initialization:

Generate initial population,$\mathcal{P}_0$.

Find *elite*$_0$.

$v = 1$ and $Exit\_flag == 0$.

Genetic algorithm main loop:

while $Exit\_flag == 0$ **do**

  Perform *selection* using roulette wheel sampling scheme.

  for $y = 1 : W$ **do**

    while constraints $(3.20) to (3.23)$ are not held **do**

      crossover with probability p_{cross}.

    end while

    while constraints $(3.20) to (3.23)$ are not held **do**

      Mutation with probability p_{mut}.

    end while

    Find *elite*$_n$.

    $\mathcal{P}_{v+1} = \mathcal{P}_v$.

    Replace the worst chromosome with $elite_{v-1}$.

    $Exit\_flag =$ **Check_termin_conditions**.

    $v = v + 1$

  end for

end while

return p_{ij}.

dedicated to each user, then, is forwarded to users on the down-link path using assigned sub-carriers and allocated power. We assume the objective function is aggregate utility maximization. In its simplest form, the utility function of user i may be a linear function of its rate, $U_i = r_i$, or an exponential function of rate such as $U_i = 1 - exp(\frac{-r_i}{b})$, where b defines the curvature of the utility function. However, for the worst case, we allow utility functions to be non-concave and nonlinear. There are two sets of users with concave and convex utility functions expressed by equation (3.62) [88]. r_i denotes allocated rate to user i, $l1$ and $l2$ are thresholds, and k controls the shape of the utility function. The function is concave for $k < 1$ and convex for $k > 1$. $k = 0.7$ and $k = 2$ have been chosen for concave and convex utility functions, respectively. The fading channel is frequency selective Rayleigh fading. Sub-carriers are divided between two groups of sub-carriers with good average channel gain and sub-carriers with weak average channel gain. Other simulation parameters are listed in Table 3.2.

$$U_i\left(r\right) = \begin{cases} 0 & r \leq l_1, \\ sin^k\left(\frac{\pi}{2}\frac{r_i - l_1}{l_2 - l_1}\right) & l_1 < r \leq l_2, \\ 1 & r > l_2. \end{cases} \tag{3.62}$$

3.5.1 Genetic Algorithm Convergence

To evaluate convergence performance of GA, a scenario consisting of 4 users with concave utility functions is considered. It is assumed that average channel gains are 1 and 0.3 on the first and the second half of the sub-carriers, respectively, for all users. In the first iteration of GA, sub-carriers are assigned to users exclusively and randomly; This assignment of sub-carriers is irrespective of users' channel gain on sub-carriers. Then, the required power to achieve a minimum rate requirement of each user is allocated uniformly to sub-carriers assigned to each user. It is expected that more power is allocated to the sub-carriers with better average channel gain as iterations proceed, to gain higher rate and utility.

Fig. 3.2 depicts the distribution of allocated power to the sub-carriers in the first and

Table 3.2: Simulation Parameters

Parameter	Value
maximum power budget of the BS	20 Watt
total bandwidth	2400 Hz
number of sub-carriers	24
number of users	4
minimum required rate of users with convex utility	100 bit/symbol
minimum required rate of users with concave utility	1 bit/symbol
number of iterations	30000
crossover probability	0.75
mutation probability	0.1
initial population,	200

the last iteration of GA. A comparison between the two distributions illustrates that GA evolves toward allocating more power to the good status sub-carriers and less power to the bad status (weak) sub-carriers, i.e., evolution of the algorithm toward maximizing the objective function by utilizing the resources efficiently. To show the speed of convergence, the best fitness value, the best users' total utility of a chromosomes, in each iteration is illustrated in

Fig. 3.3. The curve is monotonically increasing due to elitism technique, i.e., the best individual of current population is transfered to the next population, so the best fitness value never drops. As expected, there is a noticeable trade off between optimality and short solution time.

3.5.2 Interior Point and Penalty Method Performance

We evaluate the performance of PM/IPM in terms of optimality, solution time, and sensitivity of solution to users' channel gain variations on sub-carriers. The results achieved

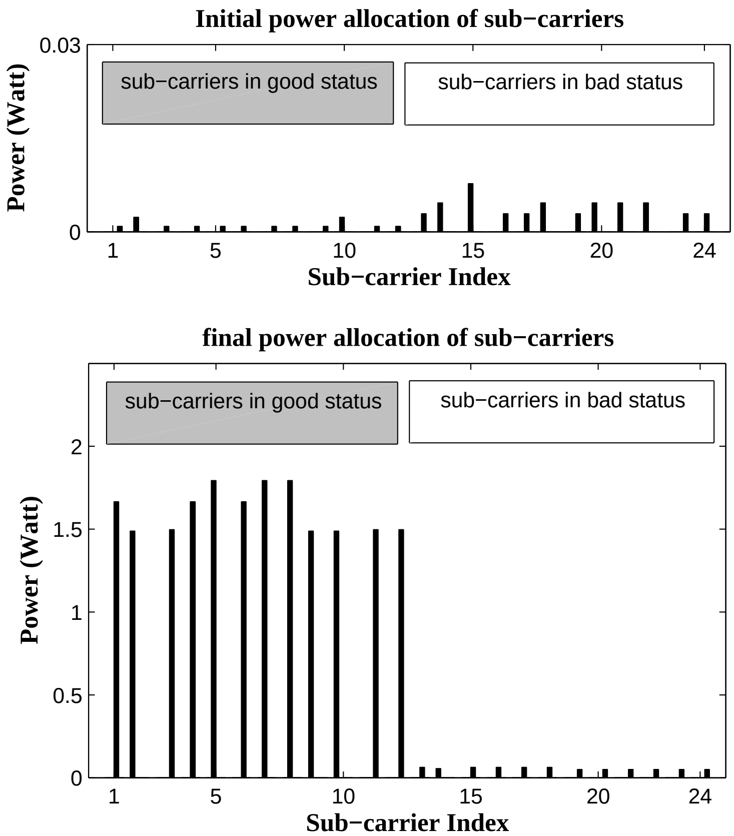

Fig. 3.2: Power allocation distribution on sub-carriers

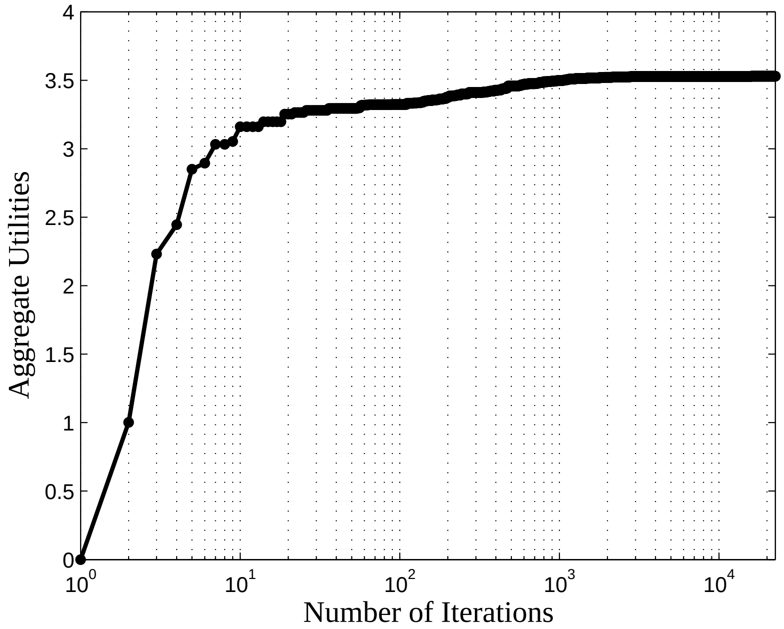

Fig. 3.3: Convergence of fitness value

by GA is used as a benchmark. A network of 4 users with convex utility functions but diverse channel gain on sub-carriers is considered. We use small number of users because GA results are intractable for large number of users. Average channel gain on sub-carriers is higher for users 1 and 3 than users 2 and 4.

Fig. 3.4 shows the convergence speed of GA and PM/IPM over time. The iterations of GA and PM/IPM stop when the improvement in rate allocation vector is less than $1e-13$. GA has a very slow convergence speed, although it starts from an initial allocation with better aggregate utilities than the ones of PM/IPM. In comparison, PM/IPM converges very fast while its maximum achievable aggregate utilities and convergence time depend on the value of σ. The smaller is σ, the faster is the method, and the less accurate is the result. The data tips on the diagram show the time and aggregate utilities with x and y, respectively. It can be seen that, at $29.61\ sec$, PM/IPM with $\sigma = 0.95$ obtains the same aggregate utilities as the one of GA, i.e., 3.609, which is obtained in about $8916\ sec$. When σ increases beyond 0.99, PM/IPM has no further improvement in achievable ag-

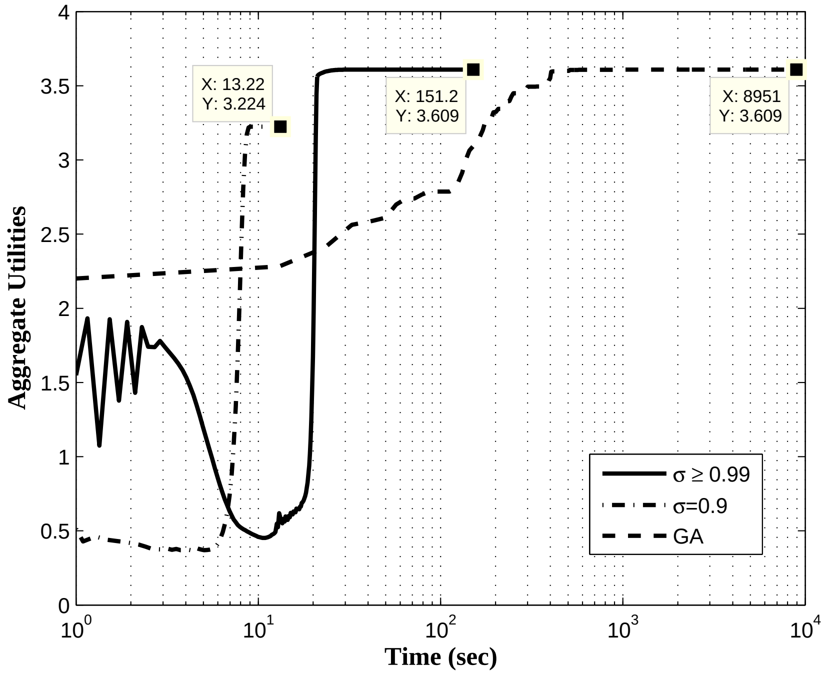

Fig. 3.4: Convergence speed comparison of GA and PM/IPM

gregate utilities or convergence speed. Fig. 3.5 compares total utility achieved by GA over iterations and PM/IPM over time, respectively. GA has a fast convergence for the first $1e + 3$ iterations, but it slows down beyond that, so it can reach to the optimum in an infinite time. On the other hand, in $t = 29\ sec$, PM/IPM with $\sigma = 0.95$ obtains the same aggregate utility as the ones of GA, which is obtained in about $5e + 3\ sec$. When σ increases beyond 0.99, PM/IPM has no more improvement in achievable aggregate utilities or convergence speed.

The convergence of PM/IPM is determined by the aggregate utilities and constraints' violations in the penalty term. For PM/IPM convergence, aggregate utilities should be maximized subject to the fact that constraints' violations are negligible or close to zero. Fig. 3.6 illustrates aggregate constraints' deviations (from zero), for two different values of σ, when PM/IPM iterations proceed over time. The negligible aggregate deviations at convergence points, especially for $\sigma = 0.99$, ensures the rate allocation satisfies the

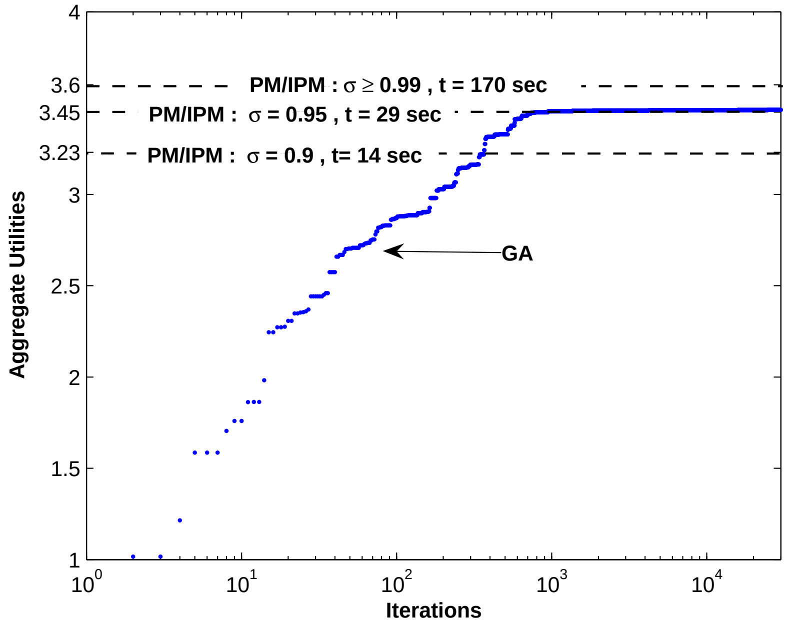

Fig. 3.5: Performance comparison of GA and PM/IPM

exclusive sub-carrier allocation. Besides, a comparison between Fig. 3.4 and Fig. 3.6 shows aggregate constraints' deviations and aggregate utilities convergence happen simultaneously, which satisfies the convergence requirements of the problem.

Moreover, a comparison between rate allocation of GA and PM/IPM, shown in Fig. 3.7, demonstrates the performance of PM/IPM in recognizing diverse channel status and its capability in allocating resources. Let all users have the same channel status, except that average channel gain on sub-carriers is higher for users 1 and 3 than those of users 2 and 4. Therefore, more resources should be allocated to the users with better average channel quality to gain user diversity and maximize aggregate utilities. The numeric tables in Fig. 3.7 represent that both GA and PM/IPM allocate more rate to users 1 and 3 than users 2 and 4. Also, it can be seen that PM/IPM allocates equal rate to the users with the same average channel quality on sub-carriers.

Table 3.3 presents rate allocation and exclusive sub-carrier assignment by PM/IPM,

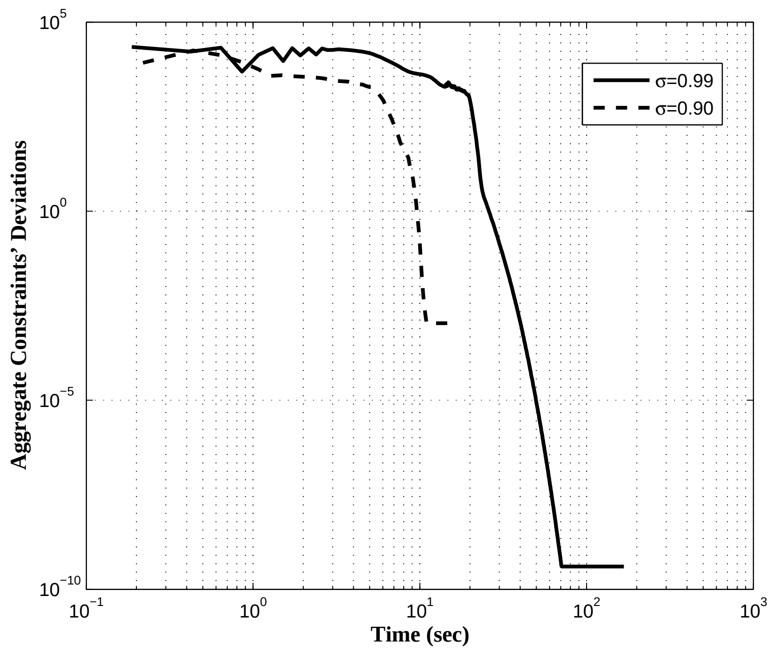

Fig. 3.6: Aggregate penalty term constraints' deviations in PM/IPM

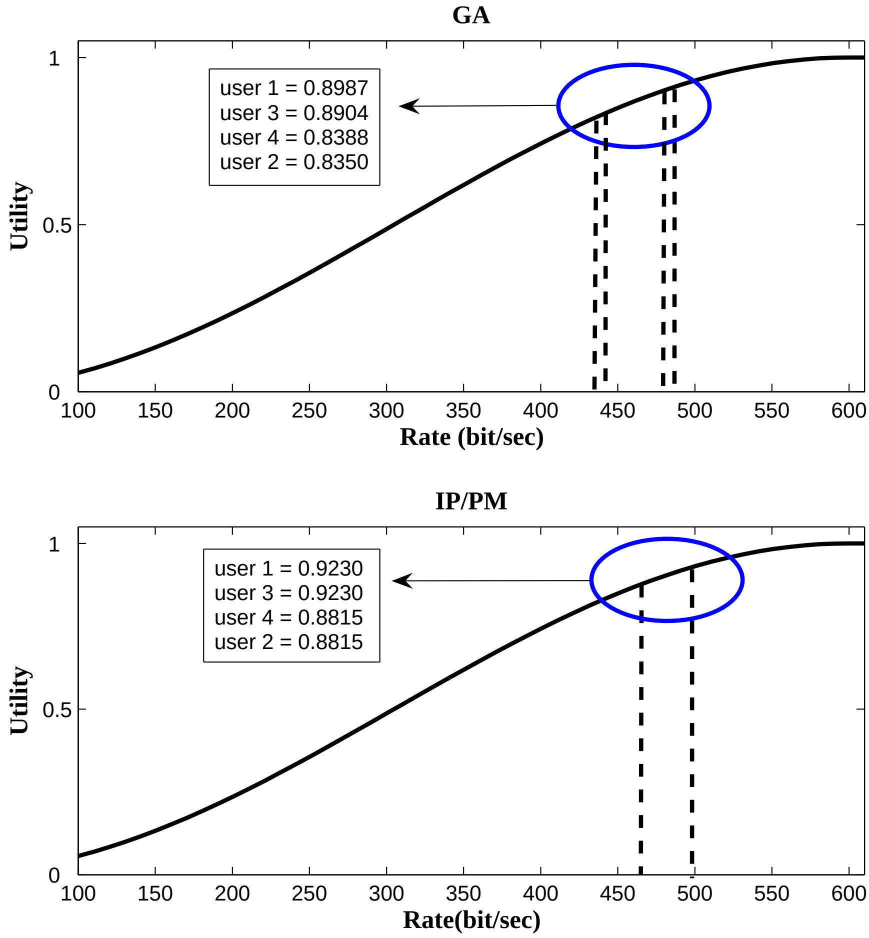

Fig. 3.7: Utility allocation comparison of GA and PM/IPM

the vectors of allocated rate to sub-carriers, $n = 1, \ldots, 24$, of users 1 to 4, r_1 to r_4, along with the corresponding channel gains of the users on the sub-carriers, α_1 to α_4. The gray rows of the table represent the assigned sub-carriers to users, and the sub-carriers on white rows are unassigned. The result confirms the success of PM/IPM in exclusive sub-carrier assignment since no sub-carrier has been assigned to two users. In addition, a sub-carrier is assigned to a user that has the best channel gain on that sub-carrier, which results in a solution closer to the optimum. In numerical results given in Table 3.3, all users achieve a utility equal to one, so some sub-carriers are not needed to be assigned to any user.

3.5.3 Resource Utilization Performance

The numerical analysis is conducted in this section to show that how considering users' utilities and application level QoS requirement in a resource allocation problem can improve the efficiency of network utilization and users' satisfaction. We divide the users into two groups of users with concave and nonconcave utilities shown in Fig. 3.8. The average channel gains equal to 0.8 for the first half of the sub-carriers and 0.2 for the second half. Fig. 3.9 demonstrates the allocated rate to the two groups of users along with their corresponding utilities. In our scenario, the network resources are tight, so all users cannot achieve utility equal to unity at the same time. Upon this circumstances, the resources are allocated to users with nonconcave utility first. The allocated rate shows that users with nonconcave utility require less rate to achieve $\%100$ satisfasction. The rest of resources are allocated to users with concave utility. As the number of users increases, less rate is allocated to each user. Therefore, utility degradation is worse for users with nonconcave utility than users with concave utility. To demonstrate the effectiveness of resource utilization, defined as the total users' utilities, we compare the resource utilization performance of a greedy scheme with the ones of a utility-based scheme. The greedy scheme allocates resources evenly between the two groups of users. Fig. 3.10 shows that utility-based resource allocation utilizes the network resources more effective than the greedy scheme. With a small increase in allocated rate to users with nonconcave utility

Table 3.3: Users's Allocated Rates on Each Sub-carrier

n	α_1	r_1	α_2	r_2	α_3	r_3	α_4	r_4
1	0.50	37	0.10	0	0.02	0	0.30	0
2	1.30	0	1.04	0	0.59	0	0.40	0
3	0.11	0	1.04	0	6.13	470	0.75	0
4	0.11	0	0.41	0	0.27	0	2.13	221
5	0.29	0	1.97	0	3.48	0	0.98	0
6	3.34	0	1.97	0	1.04	0	1.52	0
7	0.49	0	0.79	0	0.44	0	4.20	0
8	0.52	0	2.25	309	0.83	0	0.02	0
9	1.94	0	1.99	290	0.43	0	0.06	0
10	1.03	0	0.25	0	0.99	0	0.21	0
11	1.33	0	0.20	0	2.22	0	0.95	0
12	1.27	0	0.44	0	1.10	0	0.82	0
13	0.01	0	0.10	0	0.34	0	0.85	43
14	0.62	0	2.89	0	1.58	0	0.30	0
15	1.16	0	3.69	0	1.66	0	2.51	0
16	0.28	0	0.93	0	7.21	128	1.41	0
17	0.47	0	0.42	0	0.37	0	3.16	0
18	3.79	0	0.59	0	1.03	0	1.75	0
19	3.24	0	0.37	0	0.06	0	4.34	334
20	2.37	260	0.06	0	0.42	0	0.65	0
21	1.98	0	2.68	0	2.11	0	0.40	0
22	0.31	0	0.33	0	0.34	0	1.34	0
23	0.83	0	1.18	0	0.33	0	0.46	0
24	3.63	301	0.15	0	1.27	0	2.21	0

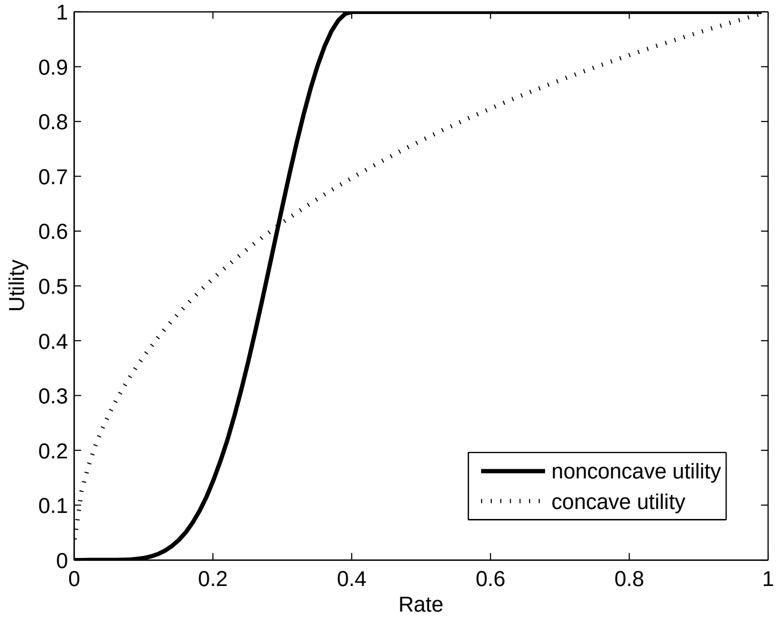

Fig. 3.8: Concave and nonconcave utilities corresponding to each group of users

and a small decrease in allocated rate to users with concave utility, users' satisfaction level of service (utilities) for users with nonconcave utility increases. However, utility degradation of users with concave utility is negligible. Overall, the sum of users' utilities increases with utility-based resource allocation which takes advantage of the diversity of the application level QoS requirement of users.

3.6 Summary

The non-convexity of OFDMA resource allocation optimization problem has been studied in this chapter. A framework for the resource allocation has been developed and a novel approach based on a penalty function method and an interior point method (PM/IPM) has been applied to solve the optimization problem. Numerical results have demonstrated that the proposed approach performs well in achieving near optimal solu-

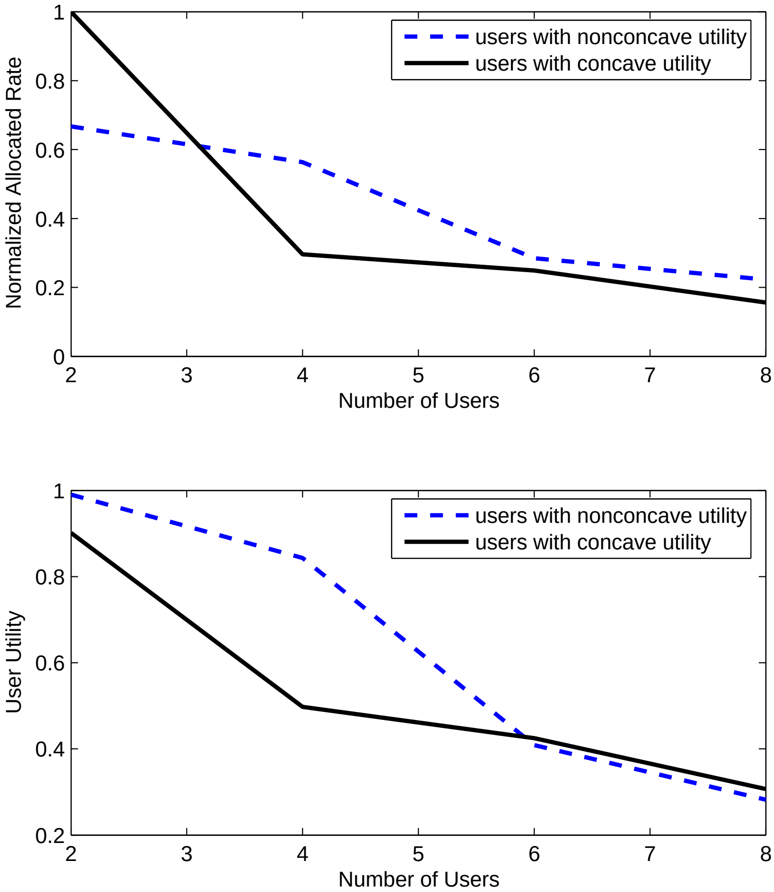

Fig. 3.9: Allocated rate and corresponding utilities to each group of users

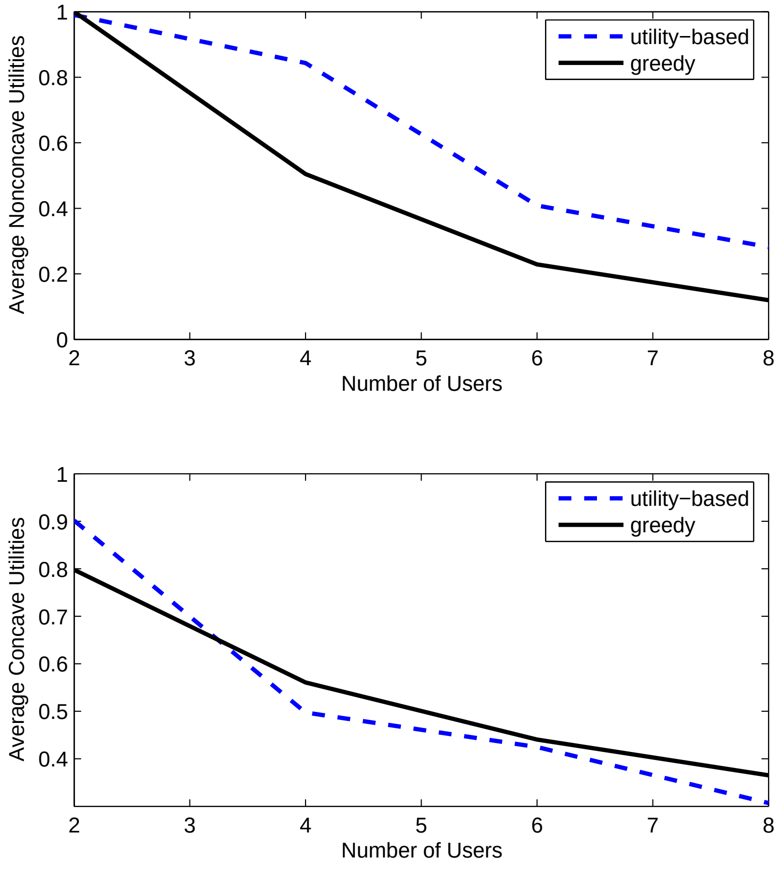

Fig. 3.10: Utilization performance of utility-based and greedy schemes

tions while satisfies the non-convex (sub-carrier assignment) constraints.

Chapter 4

Opportunistic Fair Scheduling in OFDMA Networks

Scheduling the transmissions in a telecommunication medium features a resource allocation scheme. Inspired by the framework proposed for the OFDMA resource allocation in the previous chapter, we propose an opportunistic fair scheduling scheme for OFDMA broadcast wireless channel where users have heterogeneous rate requirements. The proposed scheme jointly considers multiuser diversity gain, OFDMA resource allocation flexibility, and utility fair service discipline. Fairness among users is maintained by deploying a utility-based fair scheme that computes a set of fair weights and assigns them to users. In each scheduling interval, the resource allocated to each user is proportional to its assigned weighting factor and its channel quality on OFDM sub-carriers.

The proposed scheduler is designed with a modular structure, consisting of OFDMA Resource Allocation Module and Fairness Module. We present two separate optimization programming problems representing OFDMA Resource Allocation Module and Fairness Module to reduce the complexity, and we suggest fast algorithms to solve the problems. We present simulation results to demonstrate the performance of the proposed scheduling scheme in terms of throughput and fairness in a wireless network where users can be either fixed or mobile with heterogeneous rate requirements.

4.1 Background and Related Works

Opportunistic scheduling, which allocates resources to users with the best channel quality in each scheduling interval, is a throughput-optimal scheme for wireless networks with fading channel [89]. Opportunistic scheduling improves throughput and channel utilization especially when it exploits OFDMA, which provides more flexibility in resource allocation by dividing a broadband channel into several narrow band channels. An opportunistic scheduler in the DL needs to allocate the resources, i.e., the base station's sub-carriers and power, to users that have the best channel gain on some sub-carriers in each scheduling interval. Therefore, an OFDMA resource allocation module is needed in any opportunistic scheduler for OFDMA networks.

Despite throughput and channel utilization enhancement, severe unfairness occurs by opportunistic scheduling when averages of channel quality of users differ significantly. For example, the scheduler may not provide fair service to a user that has been shaded by neighborhood buildings in an urban area, because the channel quality of that user is always less than other users in the neighborhood. Hence, a variant of opportunistic scheduling scheme that maintains a level of fairness to unfortunate users, namely, opportunistic fair scheduling, is needed in practical networks.

Recently, some opportunistic fair scheduling schemes for multi-carrier transmission techniques have been appeared in the literature. In [90], a throughput maximization problem with deterministic and probabilistic fairness constraints for code division multiple access (CDMA) networks is proposed. To reduce complexity, the scheduling problem is decoupled into two separate tractable optimization problems: a scheduling problem that maximizes total system throughput and a fairness problem that controls and/or updates long-term fairness constraints. The proposed approach is appropriate for CDMA networks. Downlink opportunistic scheduling for OFDMA networks is considered in [91], where the scheduling is constrained by users' quality of service and fairness requirements. The utility-based fairness in [91] aims at maximizing the total network utility while guaranteeing minimum utility for individual users. Proportional fair scheduling for OFDMA networks is considered in [92]. This paper proposes clustering sub-carriers into sub-bands in order to reduce feedback overhead and complexity of the scheduling scheme. Whereas, the current literature consider various techniques to combat the complexity of multi-carrier opportunistic fair scheme for scheduling homogeneous traffic, the challenges of scheduling heterogeneous traffic with opportunistic scheduling schemes have not been addressed yet. Our work unifies many of the results found in the literature while proposing a utility proportional fair approach for multiservice OFDMA networks, i.e., OFDMA networks with heterogeneous traffic.

4.2 Opportunistic Fair Scheduling Scheme

The proposed opportunistic fair scheduling scheme jointly considers a utility-based fair resource allocation scheme and an OFDMA resource allocation scheme to allocate resources and schedule transmissions in the downlink. In each scheduling interval, depicted in Fig 4.1, the scheduling scheme selects a subset for transmission, assigns sub-carriers to selected users, and determines the transmission power and the coding and modulation scheme of each sub-carrier. All these allocations and assignments are determined by OFDMA Resource Allocation Module involved in the scheduler architecture shown in Fig. 4.2. Also, the architecture contains Fairness Module which performs in parallel with OFDMA Resource Allocation Module. Fairness Module includes Fair Weight and Transmission History blocks, as shown in Fig. 4.2. Considering the availability of CSI, a_{ij} of sub-carrier j for user i, the Fair Weight block generates a set of fair weights W_i's, associated to users $i = 1 \cdots M$, based on a utility-based fairness scheme. Then, the weights along with a set of average transmitted rate to users, R_i's, are used in OFDMA Resource Allocation Module to allocate the resources fairly.

The OFDMA resource allocation block determines users' achievable rates based on CSI at each scheduling interval, and the fair weight block computes the set of fair weights based on the averages CSI. The weights do not change during the communication interval, unless average CSI of sub-carriers for a user changes or transmission to a user is

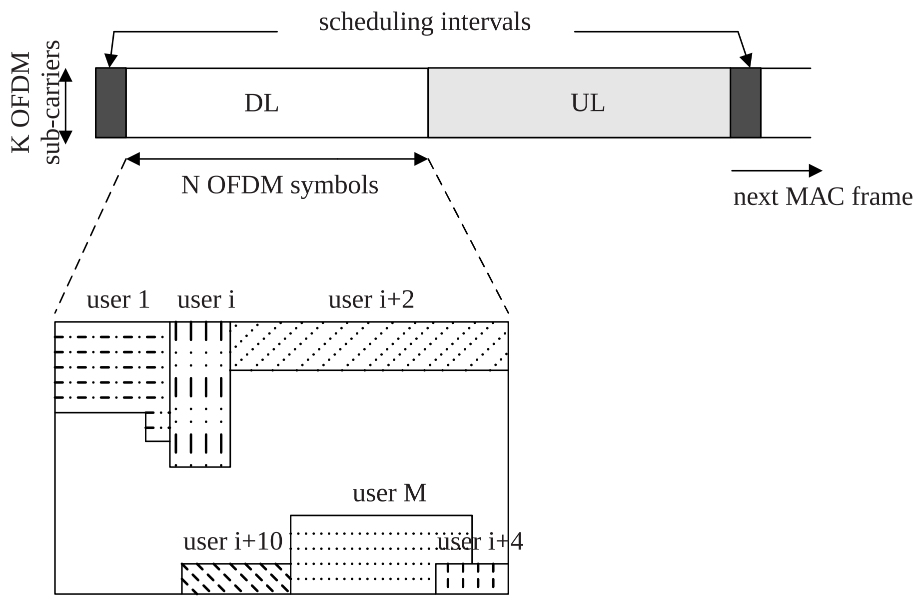

Fig. 4.1: Scheduling instances and sub-carriers allocation illustrations in a MAC frame

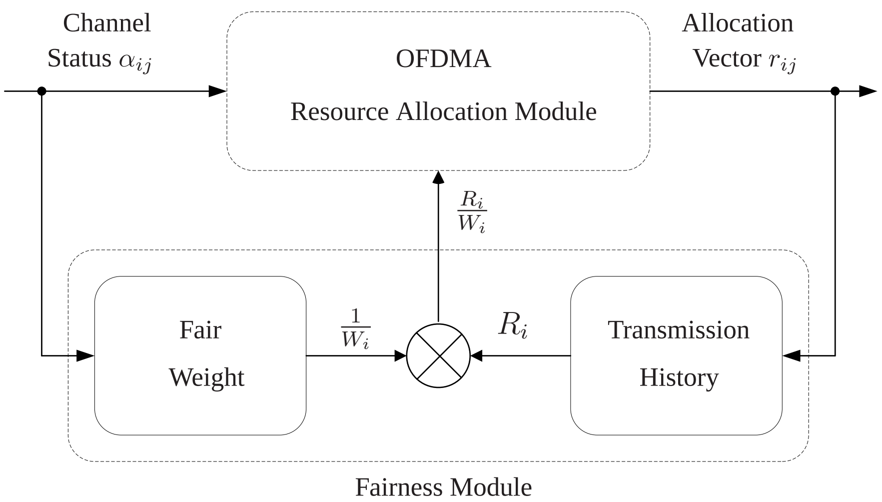

Fig. 4.2: Architecture of the proposed scheduler

terminated. W_i's are calculated based on a fairness criterion, such as max-min or proportional fairness, as will be explained later. The scheduler attempts to make R_i's as close as possible to W_i's to maintain fairness. This is achieved by maintaining the following equalities:

$$\frac{R_1}{W_1} = \frac{R_2}{W_2} = \cdots = \frac{R_M}{W_M}. \tag{4.1}$$

R_i is updated at the beginning of each scheduling interval by an exponentially weighted moving average (EWMA) technique. EWMA puts more emphasis on recent data and less emphasis on older data by applying weighting factors, which decrease exponentially as data grows older. This technique is advantageous in the sense that the fairness scheme attempts to compensate for unfairness of recent allocations as soon as possible. Equation (4.2) gives the EWMA of transmitted rate to user i at the beginning of scheduling interval m:

$$R_i(m) = (1 - \frac{1}{T_c})R_i(m-1) + (\frac{1}{T_c})r_i(m-1), \tag{4.2}$$

where r_i is the transmitted rate to user i, and T_c is a constant that determines smoothness of the exponentially decreasing weighting factors. A large T_c results in smoother decaying of the weighting factors and considering larger number of scheduling intervals in averaging. Accordingly, if allocation of rates has been unfair in the past scheduling intervals, it is more probable that the scheduler compensates for that in the next scheduling intervals.

4.3 Network Model and Problem Formulation

We present separate mathematical optimization programming problems for OFDMA Resource Allocation Module and Fairness Module. The OFDMA resource allocation, described in subsection 4.3.1, is an optimization problem whose objective function represents the scheduler objectives, and its constraints are determined based on OFDMA network specifications. Similarly, we present an optimization problem that considers users' heterogeneous rate requirements and average CSI to compute proportional fair weights in subsection 4.3.2.

4.3.1 OFDMA Resource Allocation Problem

Our network consists of a BS and several users located in one hop neighborhood from the BS. Users' backlogged traffic, buffered in separate queues at the BS, is scheduled at the beginning of each down-link interval consisting of N OFDM symbols. The BS assigns OFDM sub-carriers to users and allocates a fraction of its power, P_{BS}, to each sub-carrier of any OFDM symbol at each scheduling interval, located at the beginning of each down-link interval, as shown in Fig. 4.1. Table 4.1 tabulates symbols representing various network parameters.

Without loss of generality, we assume that noise spectral density and sub-carriers bandwidth are equal to one. Then, allocated rate to user i on sub-carrier j of OFDM symbol n, r_{ijn}, is

$$r_{ijn} = \log_2\left(1 + \alpha_{ijn} p_{ijn}\right). \tag{4.3}$$

Total allocated power to the sub-carriers of each OFDM symbol is limited by P_{BS}, i.e.,

$$\sum_{i=1}^{M}\sum_{j=1}^{K} p_{ijn} \leq P_{BS} \;\; \forall n \in \mathcal{N}. \tag{4.4}$$

Implementation of OFDM requires exclusive allocation of a sub-carrier to a single user. This constraint is mathematically represented by

$$r_{\hat{i}jn} \cdot r_{ijn} = 0 \;\; \forall \hat{i} \in \mathcal{M}, i \neq \hat{i}, \; \forall j \in \mathcal{K}, \; \forall n \in \mathcal{N}. \tag{4.5}$$

Constraint (4.5) implies that if sub-carrier j is assigned to user $\hat{i}$, i.e., $r_{\hat{i}jn} \neq 0$, allocated rate to every other user on sub-carrier j of OFDM symbol n must be zero.

To balance the achievable transmission rate and fairness, the opportunistic fair scheduler allocates sub-carrier j of OFDM symbol n to user i that has the maximum $r_{ijn}/(R_i/W_i)$.

Table 4.1: List of Symbols

Symbol	**Description**
M	number of users in the network
K	number of OFDM sub-carriers
N	number of OFDM symbols in the down-link interval
i	user index belongs to $\mathcal{M} := \{1, 2, ..., M\}$
j	sub-carrier index belongs to $\mathcal{K} := \{1, 2, ..., K\}$
n	symbol index belongs to $\mathcal{N} := \{1, 2, ..., N\}$
R_i	average transmitted rate to user i
W_i	fair weight of user i
R^i_{min}	minimum service rate requirement of the i^{th} user
P_{BS}	the BS total power budget
α_{ijn}	channel gain of user i on sub-carrier j of OFDM symbol n
p_{ijn}	required power by user i on sub-carrier j of OFDM symbol n to transmit r_{ijn}
r_{ijn}	achievable rate by user i on sub-carrier j of OFDM symbol n

The probability of assigning sub-carrier j to user i increases when the achievable transmission rate of user i on sub-carrier j is high or average transmitted rate to user i is smaller than its fair weight. The objective can mathematically be written as

$$\max \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{i=1}^{M}\left(\frac{r_{ijn}}{\frac{R_i}{W_i}}\right). \tag{4.6}$$

The objective function (4.6) along with constraints (4.3), (4.4), (4.5) model the mathematical optimization problem (Pr_6) of the opportunistic fair scheduling scheme.

$$\begin{aligned} Pr_6: \quad & \max_{r_{ijn}} \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{i=1}^{M}\left(\frac{r_{ijn}}{\frac{R_i}{W_i}}\right) && (4.7)\\ \text{s.t} \quad & \sum_{i=1}^{M}\sum_{j=1}^{K}\frac{2^{r_{ijn}}-1}{\alpha_{ijn}} \leq P_{BS}\ \forall n \in \mathcal{N}, && (4.8)\\ & r_{\hat{i}jn}\cdot r_{ijn} = 0\ \ \forall \hat{i} \in \mathcal{M}, i \neq \hat{i},\ \forall j \in \mathcal{K},\ \forall n \in \mathcal{N}, && (4.9)\\ & r_{ijn} \geq 0\ \ \forall i \in \mathcal{M},\ \forall j \in \mathcal{K},\ \forall n \in \mathcal{N}. && (4.10)\end{aligned}$$

The optimal solution of Pr_6 allocates rate to users on all sub-carriers for each OFDM symbol in a scheduling interval that achieves maximum throughput subject to the fairness criterion defined by (4.1). In practice, providing CSI of each sub-carrier over all symbols of each scheduling interval results in large messaging overhead on the reverse feedback channel. Besides, because of the correlation among CSI of a sub-carrier over consecutive symbols, the CSI of each sub-carrier is assumed to remain constant for all symbols over a scheduling interval. Accordingly, index n representing symbols of each scheduling interval can be dropped, and Pr_6 can be simplified to problem Pr_7:

$$\begin{aligned} Pr_7: \quad & \max_{r_{ij}} \sum_{j=1}^{K}\sum_{i=1}^{M}\left(\frac{r_{ij}}{\frac{R_i}{W_i}}\right) && (4.11)\\ \text{s.t} \quad & \sum_{i=1}^{M}\sum_{j=1}^{K}\frac{2^{r_{ij}}-1}{\alpha_{ij}} \leq P_{BS}, && (4.12)\\ & r_{\hat{i}j}\cdot r_{ij} = 0 \quad \forall \hat{i} \in \mathcal{M}, i \neq \hat{i},\ \forall j \in \mathcal{K}, && (4.13)\\ & r_{ij} \geq 0 \quad \forall i \in \mathcal{M},\ \forall j \in \mathcal{K}. && (4.14)\end{aligned}$$

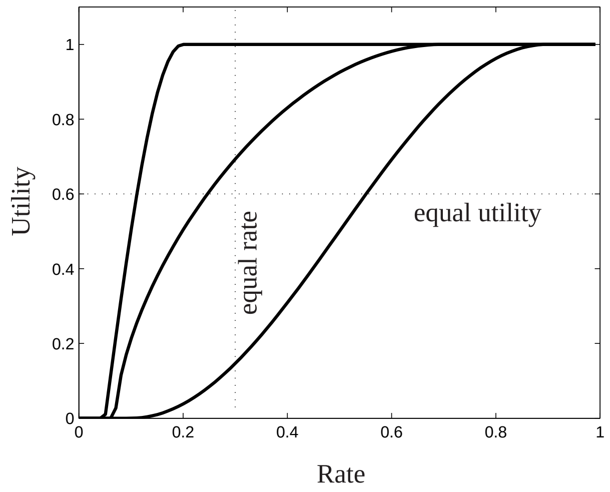

Fig. 4.3: Comparison between equal rate and equal utility allocation

4.3.2 Fairness Problem

This subsection describes how fair weights can be derived based on utility proportional fairness.

Fairness in its simplest form can be defined as equal rate allocation. However, when users have diverse service requirements and channel, an equal rate allocation results in under-utilization of network resources. For example, a user with voice service needs less rate than the ones of a user with a video service to be satisfied. An equal rate allocation to these users may make the first user not to use the extra rate while the second user starves. Fig. 4.3 shows the utilities of three different applications. The dashed line labeled “equal rate” illustrates that equal rate allocation does not provide equal user satisfaction. On the other hand, equal allocation of utilities, which is interpreted as equal users’ satisfaction, utilizes the network resources more efficiently. Thus, we will consider utility fairness instead of rate fairness [80].

In this chapter, the fair weights are determined based on utility proportional fairness where the allocated resources are proportional to users’ demands. Utility proportional fair is advantageous in a network when its users have heterogeneous rate requirements, since no user is ignored because of its high resource requirement. Consider $\mathcal{U} = \{u_h | u_h = \{u_{h1}, u_{h2}, \ldots, u_{hM}\}\}$, a bounded set of M users' feasible utilities subset u_h, where u_{hi} is the utility of user i. Utility proportional fairness is defined [93]:

Definition 4.3.1 ***Utility proportional fairness-*** *A set of utilities u_h is* utility proportional fair *if for any feasible utility set $\hat{u_h}$, the sum of proportional changes in their utilities is non-positive:*

$$\sum_{i=1}^{M} \frac{\hat{u}_{hi}(\hat{r}_i) - u_{hi}(r_i)}{u_{hi}(r_i)} \leq 0. \tag{4.15}$$

A straightforward way to obtain a proportional fair allocation $u_h \in \mathcal{U}$ is to find a set that maximizes $\sum_i \log(u_{hi})$ over the convex set of feasible allocations $\mathcal{U}$ [94, 95].

$$\max_h \ \mathcal{F} = \sum_i \log(u_{hi}) \tag{4.16}$$

We denote the set of r_{ij} that results in the proportional fair set u_h as w_{ij}, which are used in Fairness Module to derive fair weights W_i. The set of w_{ij} is the solution of the optimization problem that maximizes $\sum_i \log(u_{hi})$ subject to the network resources limits. The fairness optimization problem has a power constraint similar to (4.3) and (4.4). However, as we attempt to find a long term fair allocation of resources, the average CSI over time is deployed in the problem, so $\overline{a_{ij}}$ will be used instead of a_{ijn} in (4.3). Also, the exclusive sub-carrier assignment, constraint (4.5), is relaxed, because this problem is solved for fair weights regardless of specific sub-carrier allocation. Also, we add the minimum rate requirement constraint to make sure that the fair allocation qualifies minimum QoS requirements. Accordingly, utility proportional fair weights can be obtained by solving the optimization problem Pr_8.

$$Pr_8: \ \max_{w_{ij}} \ \mathcal{F} \tag{4.17}$$

$$\text{s.t} \quad -\sum_{j=1}^{K} w_{ij} \leq -R^i_{min} \qquad \forall i \in \mathcal{M}, \tag{4.18}$$

$$\sum_{i=1}^{M}\sum_{j=1}^{K} \frac{2^{w_{ij}}-1}{\overline{a_{ij}}} \leq P_{BS}, \tag{4.19}$$

$$w_{ij} \geq 0 \ \ \forall i \in \mathcal{M}, \quad \forall j \in \mathcal{K}. \tag{4.20}$$

The allocation of w_{ij} is a long term fair rate allocation to user i on sub-carrier j. Therefore, the fair weight of user i is inferred as follows:

$$W_i = \sum_{j=1}^{K} w_{ij} \tag{4.21}$$

If the scheduler allocates resources to users such that in a long duration of time the set of aggregate transmitted rates to users is proportional to the set of fair weights, W_i, i.e., the set of equations (4.1) is satisfied, the scheduling scheme is utility proportional fair. The set of W_i is valid until the average channel gains of a user suddenly changes or the transmission to a user is terminated. A practical approach is to periodically update the fair weights.

4.4 Solution Algorithms for OFDMA Resource Allocation and Fairness Optimization Problems

Problem Pr_7 needs to be solved in every scheduling interval, while Pr_8 is solved only when its input parameters are changed. Problems Pr_7 and Pr_8 are nonconvex optimization problems in general and finding their optimal solutions is nontrivial [20]. Problem Pr_7 is nonconvex in feasible region, while Pr_8 is nonconvex because of nonconvex utility functions in the objective function. The efficiency of a method in solving a nonconvex problem strongly depends on how nonconvexity of the problem is treated. Therefore, we apply different approaches to treat the nonconvexity of each problem.

We use a Lagrange dual decomposition method to solve P_2. The method does not guarantee an optimal solution, but it can efficiently obtain near optimal solution(s) with a practical number of sub-carriers [78]. The adaptation of Lagrange dual decomposition method hinges on the results reported in [96] that the duality gap[1] vanishes as the number of sub-carriers increases.

Whereas Lagrange dual decomposition method is applied to solve P_2, an interior point method is applied to solve P_3, because the objective function is sum of users' utilities which can be non-linear functions of users' rates, and interior point methods are shown to be successful in solving non-linear optimization problems efficiently [31].

4.4.1 The Dual Method

If $\mu_i = W_i/R_i$, the objective function of problem Pr_7 is to maximize $\sum_{i=1}^{M}\left(\mu_i \sum_{j=1}^{K} r_{ij}\right)$. Constraints (4.13) and (4.14) form the domain $\mathcal{D}$ that Lagrangian of Pr_7 can be defined over it as

$$\mathcal{L}\left(\{r_{ij}\},\lambda\right) = \sum_{i=1}^{M}\sum_{j=1}^{K}\mu_i r_{ij} - \lambda\left(\frac{2^{r_{ij}}-1}{\alpha_{ij}} - P_{BS}\right), \tag{4.22}$$

where λ is the Lagrange multiplier. The dual problem of Pr_7, is expressed as

$$\min_{\lambda}\max_{\{r_{ij}\}\in\mathcal{D}}\mathcal{L}\left(\{r_{ij}\},\lambda\right). \tag{4.23}$$

The solution of the dual problem gives λ that minimizes the maximum value of $\mathcal{L}$ over the domain $\mathcal{D}$ and determines the set of rate allocations to sub-carriers, r_{ij}, that maximizes $\mathcal{L}$. The optimization problem (4.23) is a minimization problem with one scalar variable λ that can be solved by an iterative algorithm. We use algorithm 3 to solve the problem. In each iteration of algorithm 3, the set of r_{ij} that maximizes $\mathcal{L}$ is determined by solving

[1] The difference between the primal optimal and dual optimal solution

K decomposed problems of rate allocation to sub-carriers. As allocation of sub-carriers to users are independent, the optimization problems (4.24) are solved in parallel to obtain allocated rate to sub-carriers.

$$\max_{\{r_{ij}\}\in\mathcal{D}} \sum_{i=1}^{M} \mu_i r_{ij} - \lambda\left(\frac{2^{r_{ij}}-1}{\alpha_{ij}}\right) \ \forall j = 1\cdots K. \tag{4.24}$$

When adaptive modulation is used, allocated rate to each sub-carrier is determined from a discrete set of rates. Accordingly, the solution of problem (4.24) is determined by searching over the domain $\mathcal{D}$. The search algorithm is performed in real-time, because the size of the domain $\mathcal{D}$ is confined by the number of modulation levels and sub-carriers.

4.4.2 The Interior Point Method

For notational simplicity, a solution of Pr_8 is denoted by a weight allocation vector w:

$$w \ = \ [w_{11}, w_{12}, \ldots, w_{1K}, \ldots, w_{M1}, \ldots, w_{MK}]^T, \tag{4.25}$$

where $w_i = \sum_{j=1}^{K} w_{ij}$ represents allocated weight to user i. We put the inequality constraints in a vector $C(w)$, which is represented as follows:

$$C(w) = \begin{pmatrix} \sum_{j=1}^{K} w_{1j} - R_{min}^{1} \\ \vdots \\ \sum_{j=1}^{K} w_{Mj} - R_{min}^{M} \\ -\sum_{i=1}^{M}\sum_{j=1}^{K}\frac{1}{\alpha_{ij}}(2^{\frac{Kw_{ij}}{B}}-1)+P_{BS} \\ w_{11} \\ \vdots \\ w_{MK} \end{pmatrix}, \tag{4.26}$$

and convert the inequality constraints to equality constraints by associating a positive slack variable to each constraint. Hence, Pr_8 is converted to the following minimization

Algorithm 3 Solution Algorithm for the Dual Problem

Input: $M, K, P_{BS}, \alpha_{ij}, \mu_i, bit\ loading\ set$

Output: r_{ij}

Setting up and initialization:

Set $h = 1$, $\epsilon = 1$, $Exit\_flag = 1$, $\lambda_{h-1} = \lambda_h = 0$.
Solve (4.24) for r_{ij}.
Compute $\Delta p = P_{BS} - p_{ij}$.
if $\Delta p > 0$ **then**
  return r_{ij}.
else
  while $Exit\_flag > 1e-5$ **do**
    if $\Delta p > 0$ **then**
      $\epsilon = 0.99 * \epsilon$.
      $\lambda_h = \lambda_{h-1}$.
      $\Delta p_h = \Delta p_{h-1}$.
    else
      $\lambda_{h-1} = \lambda_h$.
      $\Delta p_{h-1} = \Delta p_h$.
    end if
    $\lambda_h = \lambda_h + |\epsilon * \Delta p|$.
    Solve (4.24) for r_{ij}.
    Update Δp.
    $Exit\_flag = \lambda_h - \lambda_{h-1}$.
    $h = h + 1$.
  end while
end if
return r_{ij}..

problem:

$$Pr_9: \quad \min_{w} \quad -\sum_{i} log(u_i(w)) \tag{4.27}$$

$$\text{s.t} \quad C(w) - s = 0, \tag{4.28}$$

$$s \geq 0. \tag{4.29}$$

The rest of the interior point algorithm is implemented as described in chapter 3, so we avoid repeating it here. The correspondent $\nabla^2_{ww}\mathcal{L}$ and $\nabla_w f(w)$ for Pr_9 can be found in Appendix A.2.

4.5 Complexity of the Proposed Approach

The decomposition of (4.23) into K equations (4.24) reduces the exponential complexity to the linear complexity in K [96]. The solution of (4.24) is obtained by a heuristic search method due to the discreteness of the domain $\mathcal{D}$. The search algorithm is feasible for a practical network, because the size of $\mathcal{D}$ is confined by the number of modulation levels, users, and sub-carriers. When adaptive modulation is used, allocated number of bits to each sub-carrier is a discrete variable that can be chosen from the bit loading vector of the modulation technique [34].

Problem Pr_8 is required to be solved only when the network characteristics, e.g., users' average channel gain or the number of admitted users to the network, change. The scheduling scheme starts with default fair weights, e.g., all equal to one, and updates the fair weights with the ones obtained by solving Pr_8 during the first iterations of the scheduling scheme.

4.6 Numerical Results

Performance of the opportunistic fair scheduling scheme is evaluated in this section. The investigated performance metrics are the overall network throughput and fairness index of the proposed scheme, which are compared with the ones of a pure opportunistic scheduling scheme. We implement a multi-carrier pure opportunistic scheme similar to the opportunistic fair except $\frac{R_i}{W_i} = 1$ for $i = 1 \cdots M$.

To compare the performance in terms of fairness, a fairness metric needs to be defined first. We use Gini fairness index which is an inequality measure of resource sharing that measures deviation from equations (4.1) for each scheduler. Let the total allocated rate to user i over the simulated intervals be symbolized $\tilde{R}_i$. We examine the inequality among the set of proportions $z = \{z_i | z_i = \tilde{R}_i / W_i\}$ by Gini fairness index, I, defined as follows:

$$I = \frac{1}{2M^2\bar{z}} \sum_{x=1}^{M} \sum_{y=1}^{M} |z_x - z_y|, \tag{4.30}$$

where $\bar{z} = \frac{\sum_{i=1}^{M} z_i}{M}$. The Gini fairness index takes values between 0 and 1. A rate allocation is perfectly fair if $I = 0$. A high value of I, close to 1, indicates higher unfairness among the proportions.

The wireless channel is simulated to experience both frequency selective and large-scale fading [47], [28]. Users receive six Rayleigh distributed multipath signals. The real and imaginary components of the received signals to different users are generated from an uncorrelated multidimensional Gaussian distribution with zero mean and an identity covariance matrix. Uncorrelated multi-path components lead to uncorrelated user frequency responses in the frequency domain. Thus, full multiuser diversity can be exploited. The large-scale fading is distance dependent and follows the inverse-power law:

$$|\gamma_{ij}|^2 = D_i^{-\kappa} |\alpha_{ij}|, \tag{4.31}$$

where D_i is the distance between the BS and user i in meters, κ is pathloss exponent, and γ_{ij} is pathloss of user i on sub-carrier j. Numerical values of the wireless channel used in the simulation are: doppler frequency$= 30$ Hz, and $\kappa = 2$.

The network accepts users with nonconcave and concave utility functions, respectively. The users' utility functions are expressed by equation (4.32) [88], where r_i denotes allocated rate to user i, l_{min} and l_{max} are lower and upper rate thresholds, and k controls

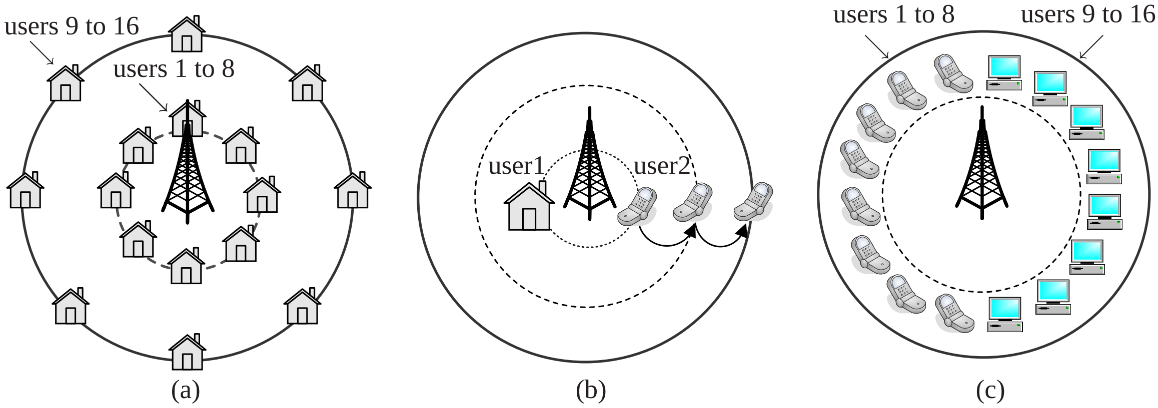

Fig. 4.4: Simulated scenarios: (a) fixed users, (b) a fixed user and a mobile user, (c) users with heterogeneous rate requirements

the shape of the utility function. The function is concave for $k < 1$ and convex for $k > 1$. $k = 2$,$l_{min} = l1 = 10$,$l_{max}l2 = 600$ and $k = 0.7$,$l_{min} = l_3 = 1$,$l_{max} = l_4 = 800$ have been chosen for convex and concave utility functions, respectively.

$$u_i\left(r\right) = \begin{cases} 0 & if r \leq l_{min}, \\ sin^k\left(\frac{\pi}{2}\frac{r_i - l_{min}}{l_{max} - l_{min}}\right) & l_{min} < r \leq l_{max}, \\ 1 & r > l_{max}. \end{cases} \quad (4.32)$$

The simulated network consists of a BS, with total power equal to 20 Watt, located at the center of the cell with 800m radius, that transmits accumulated traffic in its queues to users over 64 sub-carriers. We show the scheduling schemes performance for diverse channel gains and traffic types by considering the three scenarios shown in Fig. 4.4. In the first scenario, Fig. 4.4-(a), users' utilities are concave. Users are fixed, but their channel gains are different due to pathloss and multipath fading. The second scenario, Fig. 4.4-(b), considers a fixed user and a mobile user that has variable pathloss due to the movement, and both users have the same concave utility. The third scenario, Fig. 4.4-(c), consists of users with nonconcave and concave utilities.

4.6.1 Fixed Users

In the first scenario, shown in Fig. 4.4-(a), there are 16 users, half of the them are uniformly located on a circle with 50 meters radius, and the other half are located on the cell edge at equal angular distance. As users have diverse channel gains, we investigate the effect of channel diversity on throughput and fairness performance of the scheduling schemes using this scenario.

Fig. 4.5 shows overall throughput versus the number of users for the opportunistic and opportunistic fair scheduling schemes in the first scenario. As the opportunistic scheme assigns a sub-carrier to a user that has the highest channel gain on it, its throughput is the upper bound. The opportunistic fair has lower throughput than opportunistic because in some scheduling intervals it assigns a sub-carrier to a user that lacked service for a long time. Both scheduling schemes exploit multi-user diversity as more users join the inner circle, i.e., when the number of users increases from 2 to 8 in Fig. 4.5. Users 9 to 16 are far from the BS and their channel gains are always much lower than the users located on inner circle, so they do not increase multi-user diversity gain.

Fig. 4.6 shows the Gini fairness index of the first scenario. The fairness index of opportunistic and opportunistic fair increases as the number of users increases. Increasing user diversity has an adverse effect on fairness. However, this effect is moderated in the opportunistic fair scheme at low spatial diversity (i.e., users 1 to 8).

4.6.2 A Fixed and a Mobile User

In the second scenario, a fixed user and a mobile user that moves away from the base station are considered. At first, users 1 and 2 are located close to the BS at the same distance. Then, user 2 moves away from the BS toward the edge of the cell. We investigate the adaptivity of the opportunistic fair scheduling in capturing the network status variations using this scenario.

Fig. 4.7 shows the throughput of user1 and user2 at the three positions for oppor-

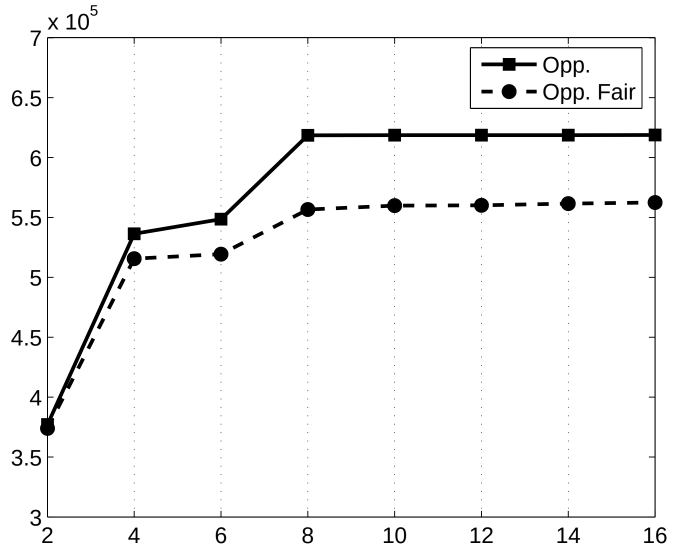

Fig. 4.5: Overall network throughput for scenario (a)

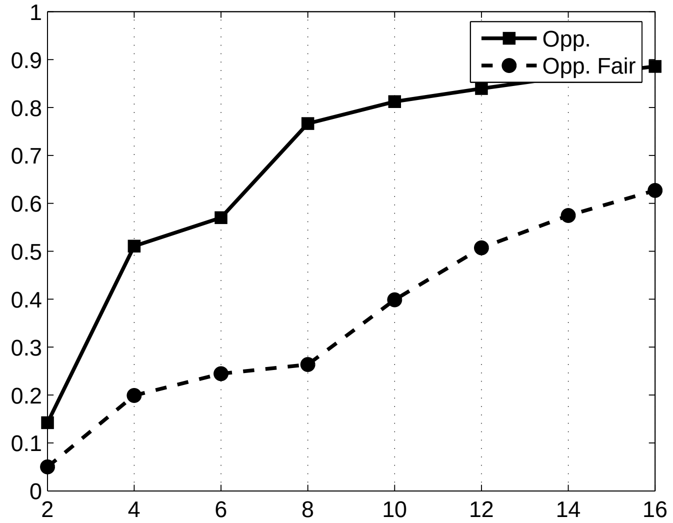

Fig. 4.6: Fairness index for scenario (a)

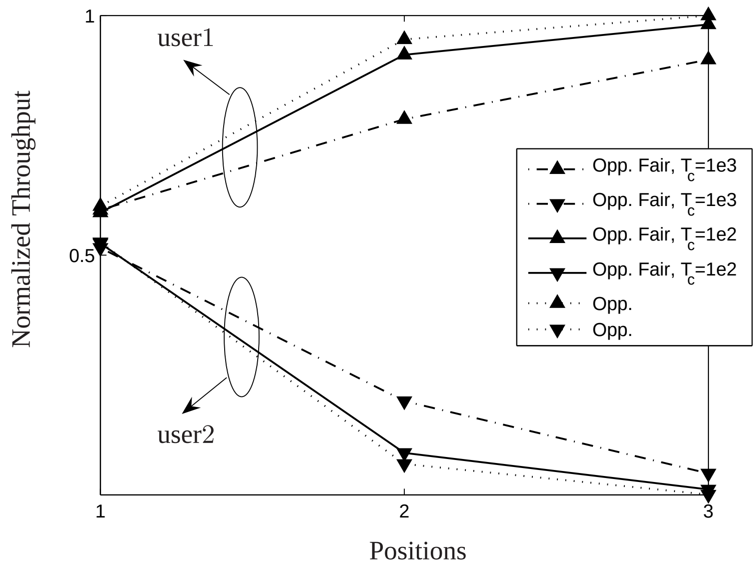

Fig. 4.7: User1 and User2's throughput at different positions of the second scenario

tunistic and opportunistic fair schemes. The throughput of opportunistic fair has been illustrated for two different time constants, T_c, for the lowpass filter of the transmission history. As user2 moves away from the BS and its channel gain drops, the opportunistic scheduling allocates less rate to it and finally ignores it when it is very far. On the other hand, the opportunistic fair scheduling scheme, which intends to allocate proportional rates to the fair weights, allocates more rate to user 2 than the ones of opportunistic allocation. The diagram shows opportunistic fair with smaller T_c is less effective, comparing to the one with larger T_c, in compensating the bad channel gain of user2 as it moves away from the BS. The reason is smaller number of scheduling intervals are considered and compensated for in the fairness scheme when T_c is small. Therefore, the scheduler has shorter time to compensate for the unfairness.

Fig. 4.8 shows the Gini fairness index of the opportunistic and opportunistic fair scheduling with two different T_c in the second scenario. When both users are close to the BS and their channel are almost similar, unfairness of opportunistic scheduling is not observed. However, as user2 moves and its channel degrades, the opportunistic

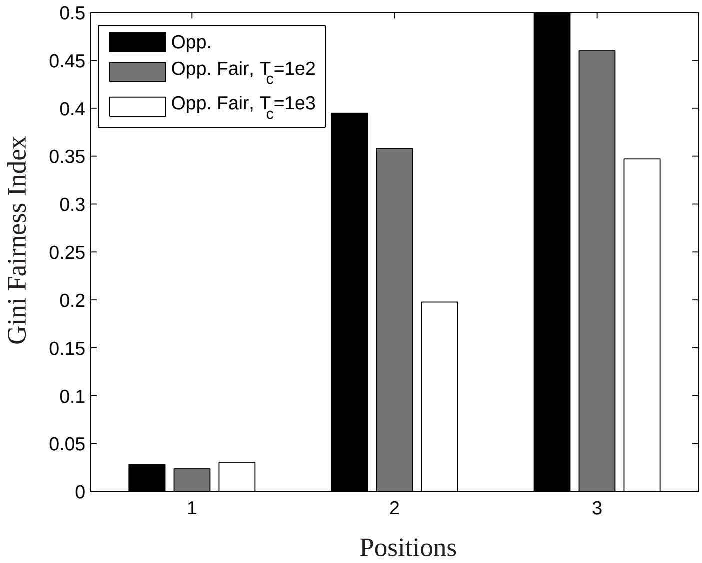

Fig. 4.8: Fairness performance of the second scenario

fair scheme treats it more fairly than the opportunistic scheme, so the fairness index of the opportunistic scheme deteriorates when user2 is at positions 2 and 3. Opportunistic fair scheme with larger T_c outperforms the one with smaller T_c in terms of the fairness performance .

The performance study of the second scenario indicates that the opportunistic fair scheduling can capture the network changes and adapt the fairness scheme accordingly. The adaptivity of the scheme can be adjusted by controlling the transmission history duration, which is one of the components of the fairness module. Furthermore, the trade off between fairness and throughput can be adjusted similarly.

4.6.3 Users With Heterogeneous Rate Requirements

In the third scenario, all 16 users are at the same distance from the BS, on a circle with 50 meters radius, but they are running two different applications with different utility functions. The first group of users, users 1 to 8, are subscribed to a service with a

Table 4.2: Aggregate Utilities of the Scheduling Schemes

Scheduling Scheme	$\sum_{i=1}^{8} U_i(r_i)$	$\sum_{i=9}^{16} U_i(r_i)$	$\sum_{i=1}^{16} U_i(r_i)$
Opportunistic	161.4702	141.0092	302.4793
Opportunistic Fair	306.2989	196.6107	502.9096

nonconcave utility function. The second group of users, users 9 to 16, are subscribed to a service with a concave utility function.

The utility values of users 1 to 8 over 100 samples of the channel, when their traffic is scheduled by opportunistic scheme and opportunistic fair scheme, are represented in Fig. 4.9-a and Fig. 4.9-b, respectively. The figures reveal that, first, opportunistic scheme ignores some users with low channel gains over the simulation interval, such as user 8 in Fig. 4.9-a. This fact causes severe unfairness in service provisioning when the user diversity is high. Second, the rate allocations and hence the utility distributions of users for opportunistic scheme is not as regular as the ones of opportunistic fair scheme. Accordingly, opportunistic scheduling is not effective in service provisioning for the applications that should be scheduled almost regularly.

The data statistics of the simulation, shown in Table 4.2, depicts that the utilization of resources or users' satisfaction of received service, which is represented by the sum of users' utilities, is improved for opportunistic fair scheduling more than that of the opportunistic scheduling scheme. Moreover, the users with convex utilities have a higher aggregate utility than the ones of the users with concave utilities. The reason is the gradient of the convex utility function is higher than the gradient of concave utility function at lower rates in our simulation. Therefore, for the same allocated rate, convex utility value is larger than the concave value.

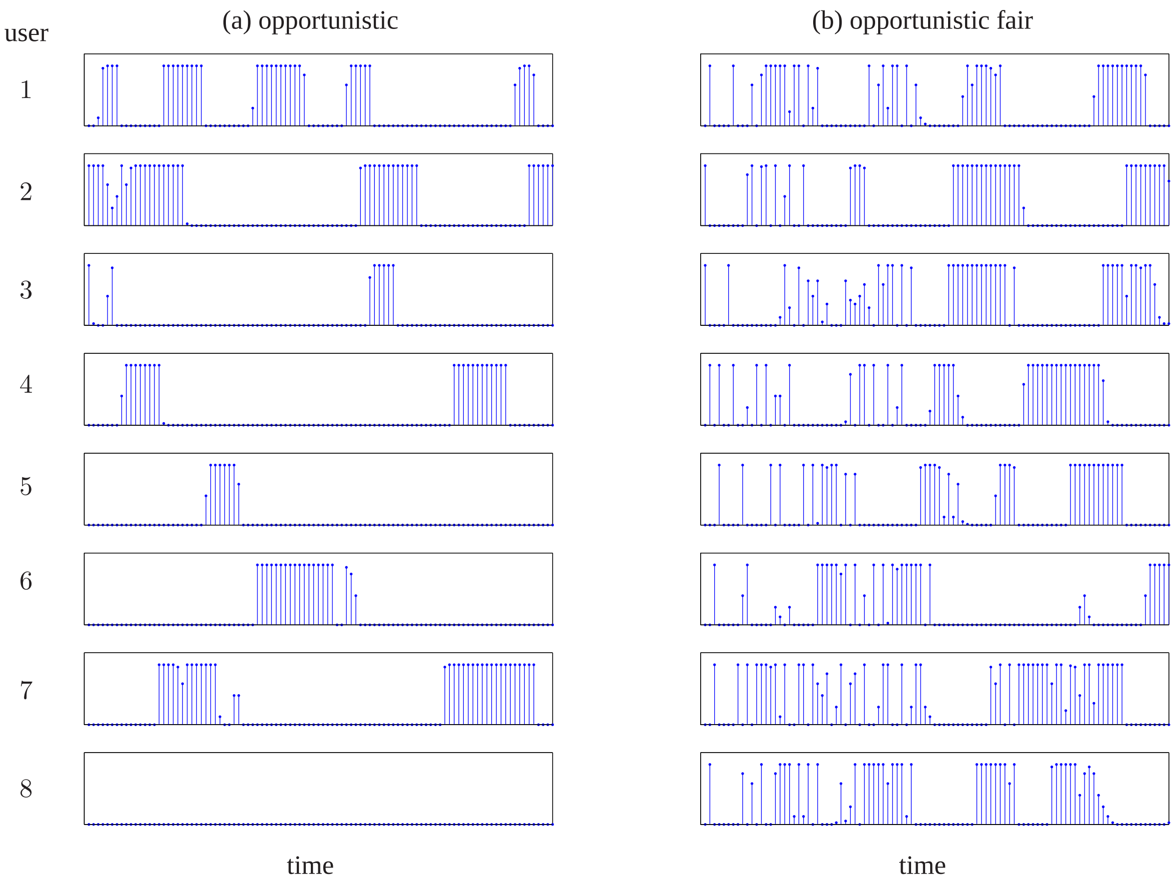

Fig. 4.9: Utility values of users 1 to 8 for opportunistic and opportunistic fair scheduling schemes versus time

4.7 Summary

An opportunistic fair scheduling scheme is proposed for the downlink of OFDMA networks where users have heterogeneous rate requirements. The scheduler takes sub-carriers channel gain and fairness requirements into account to assign sub-carriers to users and allocate rate to each sub-carrier. We consider fairness constraints by adopting the utility proportional fair criteria, computing a set of fair weights associated to users, and allocating the resources according to the fair weights. The proposed scheme is adaptive because the fair weights can be modified dynamically when the network characteristics change due to mobility of users, admitting a new user, or changing the fairness policy of the network service provider. As the fair weights are computed based on a utility-based resource allocation scheme, the resource utilization of the network improves and allocated resources conform the requirement of the users, which are represented by their utilities. Comparing to opportunistic scheduling scheme, the proposed opportunistic fair scheduling scheme provides fairer and smoother service with the cost of the throughput.

Chapter 5

Conclusions and Further Works

The success of wireless networks in supporting a variety of applications and being expanded in a large commercial scale is strongly tied to the performance of corresponded resource allocation schemes. The research in this thesis focuses on resource allocation schemes for OFDMA networks with heterogeneous traffic types which simultaneously provide QoS, maintain fairness, and improve network utilization. Following, we summarize the major research contributions of the thesis, and propose further works.

5.1 Major Research Contributions

- We investigate the OFDMA resource allocation problem and develop a framework for the resource allocation in a very generic form. The framework exploits many aspects of an efficient resource allocation scheme, such as, collaboration between MAC and PHY, and adaptivity to resource changes to improve the resource utilization performance while satisfying heterogeneous users' demands and maintaining fairness among users. Also, the framework captures the essential characteristics of the network and users' requirements, such as, exclusive sub-carrier allocation constraint and users' minimum QoS requirements.

We have followed a new direction in problem formulation. Unlike previous formulations for the OFDMA resource allocation problem, in the literature, which are based on combinatorial optimization techniques, we use continuous optimization techniques for the problem formulation. Given this method of problem formulation, we present an NLP problem that can be solved by continuous optimization algorithms rather than combinatorial ones. Our suggested algorithm is a combination of interior point methods and penalty function methods. The proposed algorithm treats the non-convexity of the problem and exploits the strength of interior point methods in solving NLP problems.

The proposed framework for OFDMA resource allocation can be applied to many centralized networks with multiservice support. More importantly, the application of the framework can be extended to network utility maximization (NUM) problems with either convex or non-convex objective functions. The new problem formulation method sheds some light on the future research about deploying continuous optimization techniques for solving the OFDMA resource allocation problem. Also, the simple and fast algorithms deployed facilitate performance analysis of a variety of OFDMA resource allocation schemes.

- We propose an opportunistic fair scheduling scheme for OFDMA networks with heterogeneous traffic types. In the proposed scheduler, a fairness enforcement technique has been integrated with an opportunistic scheduling scheme to maintain longterm fairness and smooth service delivery. The fairness scheme assigns some fair weights to users which maintain utility proportional fairness among users. The fair weights are determined based on users' average channel gains and utility functions, so fair weights can be assigned for long durations as long as users' average channel gains are static and no user joins/departs the network. On the other hand, when the channel statistics are dynamic, the fair weights can be computed periodically, with a period which is proportional to the rate of variations. The fairness scheme can be adjusted to maintain a measure of fairness tightly or loosely, i.e., the tradeoff between throughput and fairness is adjustable.

Our proposed scheduling scheme design is modular, which helps in separating the tasks of the scheduler between an OFDMA resource allocation module and a fairness module. Besides, as individual modules are less complicated than the combined one, fast and simple algorithms can be used in each module to reduce the complexity of the scheduling. To achieve an optimal resource allocation, we formulate the tasks of each module with an optimization programming problem. We apply dual method algorithms to the OFDMA resource allocation problem, where the objective function is a linear function. When dual method algorithms are applied to the OFDMA resource allocation problem the duality gap is not zero, but it is reduced significantly for practical parameters selection for the problem. More precisely, when the number of sub-carriers grows bigger than the number of users, the duality gap vanishes. Applying duality methods is advantageous in our scheme, because the computational complexity of rate allocation and users' scheduling is reduced. Due to the non-linearity of the fairness module optimization problem, we apply an interior point method combined with a penalty function method. Using interior point and penalty function methods in the fairness optimization problem facilitates the utility fairness implementation; hence, it improves resource utilization.

5.2 Further Works

The proposed schemes for resource allocation and scheduling in this thesis tackle some challenges of the OFDMA resource allocation problem such as non-convexity issues and heterogeneous traffic support. However, there are still many open issues to extend the research and deserve further investigation:

- The research in this thesis investigates conceptual aspects of resource allocation schemes for OFDMA networks. However, fine tuning of the schemes parameters and improving the algorithms convergence speed remain for further research. For example, the width of the exponential window in the moving average technique should be adjusted according to the required tradeoff between throughput and fairness. In addition, the performance of the proposed schemes have been derived for saturated buffers; the effect of traffic model variations on the performance can be studied to specify if adapting or modifying the schemes are required. Some important aspects of PM/IPM that can be discussed include complexity analyses, the initial choice of penalty parameter, strategies for updating the KKT perturbations, and appropriate criteria for terminating inner iterations.

- Scalability is a necessary factor of the algorithms for solving optimization problems of the OFDMA resource allocation. When the number of sub-carrier increases in the network, the algorithms may take longer time, which is not acceptable for real-time applications. Some techniques such as sub-carrier clustering may be taken into account to downsize the allocation variables, which result in loosing some diversity gain [97]. For practical implementation the tradeoff between scalable algorithms and achieving high diversity gain deserves to be recognized and controlled.

- The OFDMA resource allocation problems in this thesis are constrained by the total transmit power, and the utility functions are functions of rate. However, in some applications the objective of the resource allocation is to minimize transmission power [98], or maximize an objective function which is not a function of rate, e.g., maximizing aggregate utilities where the utilities are a function of delay [99]. As the problem formulations are different in such cases, one can specify the applicability of the proposed algorithms in this thesis to those problems and the required modifications of the algorithms if it is needed.

- The results from this thesis and other research [100] indicate that collaborating with PHY layer can significantly improve the performance of resource allocation schemes in wireless OFDMA networks. The performances of proposed resource allocation schemes, which are based on cross-layer design with PHY layer, depend on the accuracy of PHY layer models or measurements.

In this thesis, short-term time variations of channel are exploited in the OFDMA resource allocation scheme, and users' average channel gains are deployed in the fairness scheme. The performance study of the schemes proposed in chapter 3 and 4 is based on the assumption of Rayleigh distribution for the amplitude of the channel gains. The performance analysis can be extended for wireless medium with different fading characteristics, such as, Rician or Nakagami fading channel distributions. Furthermore, some other channel statistical specifications can be deployed in adjusting resource allocation parameters. For example, the average fade duration, which quantifies how long the signal spends below a threshold [101], can be deployed to determine the periods of fair weights calculation.

Perfect CSI is assumed to be available in the BS when channel information are obtained by measurement. However, there exist some uncertainty in achieved CSI due to unreliable feedback channel, which may result in wrong decisions being made by the schemes [102]. It is important to study the effects of imperfect CSI, e.g. estimation error and feedback delay. Also, feedback overhead reduction deserves to be investigated specifically in practical networks with large number of sub-carriers and users. Effective approaches are needed to maintain the diversity gains while reducing the feedback overhead.

- The PHY layer capacity and the resource allocation scheme deployed in the link layer directly affect the available resources for admitting new call requests. While satisfying users' requirements and network constraints, a call admission strategy tries to allow as many user as possible to access the resources simultaneously. Accordingly, an admission control strategy in the network layer is corresponded to stochastic transmissions inherent in channel-aware networks [103]. Designing an admission control strategy that benefits from the efficiency, fairness, and improved resource utilization of the proposed resource allocation schemes in this work is of great importance for commercial implementation of the schemes as well as theoretically extending the work.

- The framework in chapter 3 formulates the rate allocation in centralized OFDMA networks as a NUM problem. A large number of work based on utility maximization approach for network resource allocation has been performed already (look at [104] and the references therein), such as TCP congestion control [105], sharing link capacities among sources, and bandwidth allocation in wireless networks [106]. In particular, the previous research focused on proposing distributed algorithms to solve NUM problems where the utilities were assumed to be concave. However, the demand for transmission heterogeneous traffic types, i.e., real-time and non-real-time, requires non-concave utilities be considered in NUM problems. Then, the proposed distributed algorithms for convex NUM problems may not be tractable for non-convex ones. Inspired by the formulation method and proposed algorithms in this thesis, non-convex NUM problems can be considered in future research.

- Users' minimum rate requirement constraints and utility-based resource allocation satisfy users' essential QoS requirements, but some applications such as video or streaming media need stringent QoS requirements that are not satisfied by this approach. Besides, a higher resource utilization is achieved if more video statistics are properly used in the resource allocation scheme [107]. For commercial video applications, such as video on demand and internet protocol television (IPTV), some traffic characteristics, such as, different importance of encoded video layers, burstiness of video content, and decoding dependency constraints of multimedia can be taken into account for further resource utilization and resource allocation efficiency [108]. However, considering these aspects pose new challenges on resource allocation problem formulation. Nevertheless, the dramatic increase in video demand on wireless broadband networks drives motivations for developing suitable resource allocation schemes, which attain the highest overall video quality given the limited resources while delivering consistent and smooth service and maintaining fairness among users who subscribed the same QoS.

- The resource allocation schemes in the literature can be categorized as centralized or decentralized schemes. The former is corresponded to networks with a PMP infrastructure, but the later can be applied to either PMP (single-hop or multihop) or adhoc networks.

The network infrastructure considered in this thesis is a centralized single-hop, which is a fundamental infrastructure in many networks, e.g., cellular and relay networks. The proposed resource allocation scheme has been specially designed for a stand alone network. However, when some centralized single-hop networks are adjacent, such as in cellular networks, the inter-cell interference should be taken into account. It has been revealed that collaboration of BSs will resolve the inter-cell interference problem and result in better performance [80], providing that effective decentralized schemes with low complexity for the collaboration are suggested. Similarly, the proposed centralized scheme can be extended to multi-hop relay networks. In multihop networks the co-existence of multiple links for transmission causes more complexity, because the active links in each resource allocation intervals should be determined in addition to the rate and sub-carrier allocation to the transmission on each link [109].

Decentralized schemes are of great importance for the resource allocation in the UL of centralized networks or in decentralized networks. The proposed resource allocation scheme in this thesis can be applied in the UL with some slight changes [110]. For example, the BS power constraint is replaced by a per user power constraint. Furthermore, as the nodes in the network, except the BS, are not usually able to monitor all channels used for other nodes transmissions, an effective information exchange mechanism is needed to take full advantage of the channel diversity among different nodes.

5.3 Final Remarks

The coexistence of real-time and non-real-time traffic in future wireless networks is promising. Therefore, resource allocation schemes that support multiple traffic types and maintain fairness simultaneously are demanded.

While considering most important resource allocation paradigms, such as adaptive rate or power allocation, dynamic frequency allocation, and scheduling, this thesis focuses on the OFDMA fine resolution and flexibility in resource allocation. We present a framework which considers the OFDMA network restrictions as well as heterogeneous users' fairness and QoS constraints while attempting to improve the wireless system scarce resources utilization. By taking a different direction from previous works, we introduce a new problem formulation and solution, based on continuous optimization techniques, for the OFDMA resource allocation optimization problem, which produces results with reasonable accuracy in practical time duration.

The new formulation for the OFDMA resource allocation, in this thesis, facilitates applying other continuous optimization approaches that may treat nonconvexity problem more efficient than discrete optimization methods that have been proposed so far in the literature. For instance, scalable and fast-converging continuous methods can be searched for to be applied to this new problem formulation in further works.

Appendix A

Derivation of $\nabla^2_{rr}\mathcal{L}$, $\nabla_r f(r)$, $\nabla^2_{ww}\mathcal{L}$, and $\nabla_w f(w)$

The mathematical representations of $\nabla^2_{rr}\mathcal{L}$ and $\nabla_r f(r)$, required by the interior point algorithms in chapter 3, as well as $\nabla^2_{ww}\mathcal{L}$ and $\nabla_w f(w)$, required in chapter 4, are presented in section A.1 and A.2, respectively.

A.1 $\nabla^2_{rr}\mathcal{L}$ and $\nabla_r f(r)$

The objective function of Pr_5, based on utility functions (3.62), is represented by:

$$f(r) = \; - \; (U_1(r_1) + \ldots + U_M(r_M)) + \frac{L}{2}(\sum_i \sum_{\hat{i}} (r_{\hat{i}1} r_{i1} + \ldots + r_{\hat{i}K} r_{iK})). \quad \text{(A.1)}$$

Accordingly, $\nabla_r f(r) = \left(\frac{\partial f}{\partial r_{11}}, \ldots, \frac{\partial f}{\partial r_{MK}}\right)^T$ is computed as follows:

$$\nabla_r f(r) = -\begin{pmatrix} \frac{\partial U_1(r_1)}{\partial r_{11}} \\ \vdots \\ \frac{\partial U_1(r_1)}{\partial r_{MK}} \\ \vdots \\ \frac{\partial U_M(r_M)}{\partial r_{M1}} \\ \vdots \\ \frac{\partial U_M(r_M)}{\partial r_{MK}} \end{pmatrix} + L \begin{pmatrix} \sum_i r_{i1} - r_{11} \\ \vdots \\ \sum_i r_{iK} - r_{MK} \\ \vdots \\ \sum_i r_{i1} - r_{M1} \\ \vdots \\ \sum_i r_{iK} - r_{MK} \end{pmatrix}, \tag{A.2}$$

where, for $j = 1, \cdots, K$, and $\theta = \frac{\pi}{2}\frac{r-l_1}{l_2-l_1}$:

$$\frac{\partial U_{\breve{i}}}{\partial r_{ij}} = \begin{cases} \frac{k\pi}{2(l_2-l_1)} sin^{(k-1)}(\theta) cos(\theta) & if \;\; i = \breve{i}, \\ 0 & otherwise. \end{cases} \tag{A.3}$$

To obtain $\nabla^2_{rr}\mathcal{L}$, $\nabla^2_{rr} f(r)$ and $\nabla^2_{rr} C(r)$ are computed first:

$$\nabla^2_{rr} f(r) = \quad - \quad \begin{pmatrix} G(r_1) & 0_{(K,K)} & \ldots & 0_{(K,K)} \\ 0_{(K,K)} & G(r_2) & \ldots & 0_{(K,K)} \\ \vdots & \vdots & & \vdots \\ 0_{(K,K)} & 0_{(K,K)} & \ldots & G_(r_M) \end{pmatrix} \tag{A.4}$$

$$+L \quad \begin{pmatrix} 0_{(K,K)} & I_{(K,K)} & \ldots & I_{(K,K)} \\ I_{(K,K)} & 0_{(K,K)} & \ldots & I_{(K,K)} \\ \vdots & \vdots & & \vdots \\ I_{(K,K)} & I_{(K,K)} & \ldots & 0_{(K,K)} \end{pmatrix}, \tag{A.5}$$

where

$$G(r_i) = \begin{pmatrix} \frac{\partial^2 U_i}{\partial r_{i1}\partial r_{i1}} & \ldots & \frac{\partial^2 U_i}{\partial r_{i1}\partial r_{iK}} \\ \frac{\partial^2 U_i}{\partial r_{i2}\partial r_{i1}} & \ldots & \frac{\partial^2 U_i}{\partial r_{i2}\partial r_{iK}} \\ \vdots & & \vdots \\ \frac{\partial^2 U_i}{\partial r_{iK}\partial r_{i1}} & \ldots & \frac{\partial^2 U_i}{\partial r_{iK}\partial r_{iK}} \end{pmatrix}. \tag{A.6}$$

$0_{(K,K)}$ is a $K \times K$ matrix with all zero entries, and $I_{(K,K)}$ is a $K \times K$ identity matrix. The second partial derivatives of the utility functions required for calculating $G(r_i)$ functions are:

$$\frac{\partial^2 U_i}{\partial r_{i\breve{j}} \partial r_{ij}} = \frac{K\pi^2}{4(l_2 - l_1)^2} \left((k-1) sin^{(k-2)}(\theta) cos^2(\theta) - sin^k(\theta) \right), \tag{A.7}$$

for $\breve{j}$ and $j \in \{1, \cdots, K\}$.

Finally, $\nabla^2_{rr} C(r)$ for calculating $\nabla^2_{rr}\mathcal{L}$ is obtained by:

$$\nabla^2_{rr} C(r) = \left(\frac{Kln(2)}{B}\right)^2 \begin{pmatrix} \frac{2^{\frac{Kr_{11}}{B}}}{\alpha_{11}} & 0 & \dots & 0 \\ 0 & \frac{2^{\frac{Kr_{12}}{B}}}{\alpha_{12}} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{2^{\frac{Kr_{MK}}{B}}}{\alpha_{MK}} \end{pmatrix}. \tag{A.8}$$

A.2 $\nabla^2_{ww}\mathcal{L}$ and $\nabla_w f(w)$

The objective function of Pr_9, based on utility functions (4.32), is represented by:

$$f(w) = \;-\; log(U_1(w_1) - \ldots - log(U_M(w_M)). \tag{A.9}$$

Accordingly, $\nabla_w f(w)$ is computed as follows:

$$\nabla_w f(w) = -\left(\frac{k\pi}{2\left(l_2 - l_1\right)\frac{cos(\theta)}{sin(\theta)}}\right)\begin{pmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\end{pmatrix} - \left(\frac{k\pi}{2\left(l_4 - l_3\right)\frac{cos(\theta)}{sin(\theta)}}\right)\begin{pmatrix}0\\ \vdots\\ 0\\ 1\\ \vdots\\ 1\end{pmatrix} \tag{A.10}$$

$$= -\begin{pmatrix}\frac{k\pi}{2(l_2-l_1)\frac{cos(\theta)}{sin(\theta)}}\\ \vdots\\ \frac{k\pi}{2(l_2-l_1)\frac{cos(\theta)}{sin(\theta)}}\\ \frac{k\pi}{2(l_4-l_3)\frac{cos(\theta)}{sin(\theta)}}\\ \vdots\\ \frac{k\pi}{2(l_4-l_3)\frac{cos(\theta)}{sin(\theta)}}\end{pmatrix}, \tag{A.11}$$

where, $\theta = \frac{\pi}{2}\frac{w_{ij}-l_{min}}{l_{max}-l_{min}}$. Note that utility functions (4.32) are convex for $i = 1, \cdots, \frac{M}{2}$ and concave for $i = \frac{M}{2} + 1, \cdots, M$.

To obtain $\nabla^2_{ww}\mathcal{L}$, $\nabla^2_{ww}f(w)$ and $\nabla^2_{ww}C(w)$ are computed first:

$$\nabla^2_{ww}f(w) = -\begin{pmatrix}G(w_1) & 0_{(K,K)} & \dots & 0_{(K,K)}\\ 0_{(K,K)} & G(w_2) & \dots & 0_{(K,K)}\\ \vdots & \vdots & & \vdots\\ 0_{(K,K)} & 0_{(K,K)} & \dots & G_(w_M)\end{pmatrix}, \tag{A.12}$$

where

$$G(w_i) = \begin{pmatrix}\frac{\partial^2 f_i}{\partial w_{i1}\partial w_{i1}} & \dots & \frac{\partial^2 f_i}{\partial w_{i1}\partial w_{iK}}\\ \frac{\partial^2 f_i}{\partial w_{i2}\partial w_{i1}} & \dots & \frac{\partial^2 f_i}{\partial w_{i2}\partial w_{iK}}\\ \vdots & & \vdots\\ \frac{\partial^2 f_i}{\partial w_{iK}\partial w_{i1}} & \dots & \frac{\partial^2 f_i}{\partial w_{iK}\partial w_{iK}}\end{pmatrix}, \tag{A.13}$$

and $0_{(K,K)}$ is a $K \times K$ matrix with all zero entries. The second partial derivatives of the objective functions required for calculating $G(w_i)$ are:

$$\frac{\partial^2 f_i}{\partial w_{i\breve{j}}\partial w_{ij}} = \frac{K\pi^2}{4(l_{max}-l_{min})^2}\left(\frac{1}{sin^2\left(\theta\right)}\right), \tag{A.14}$$

for $\breve{j}$ and $j \in \{1, \cdots, K\}$.

We need $\nabla^2_{ww}C(w)$ to calculate $\nabla^2_{ww}\mathcal{L}$. As the users' minimum rate requirement R^i_{min} equals l_{min} and w_i's are bounded by l_{min} and l_{max}, first, we rewrite $C(w)$ as follows:

$$C(w) = \begin{pmatrix} \sum_{j=1}^K w_{1j} - \frac{l_1+l_2}{2} \\ \vdots \\ \sum_{j=1}^K w_{\left(\frac{M}{2}\right)j} - \frac{l_1+l_2}{2} \\ \sum_{j=1}^K w_{\left(\frac{M}{2}+1\right)j} - \frac{l_3+l_4}{2} \\ \vdots \\ \sum_{j=1}^K w_{Mj} - \frac{l_3+l_4}{2} \\ -\sum_{i=1}^M \sum_{j=1}^K \frac{1}{\alpha_{ij}}(2^{w_{ij}}-1) + P_{BS} \\ w_{11} \\ \vdots \\ w_{MK} \end{pmatrix}, \tag{A.15}$$

Then, $\nabla^2_{ww}C(w)$ is derived as follows:

$$\nabla^2_{ww}C(w) = (ln(2))^2 \begin{pmatrix} \frac{2^{w_{11}}}{\alpha_{11}} & 0 & \dots & 0 \\ 0 & \frac{2^{w_{12}}}{\alpha_{12}} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{2^{w_{MK}}}{\alpha_{MK}} \end{pmatrix}. \tag{A.16}$$

References

[1] J. Govil, “An Empirical Feasibility Study of 4Gs Key Technologies,” in *Proc. IEEE International Conference on Electro/Information Technology (EIT)*, 2008, pp. 267–270.

[2] IEEE, *IEEE Std 802.16-2001*, IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Std., April 2002.

[3] ——, *IEEE Std 802.16a-2003*, IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems-Amendment 2: Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11 GHz, Std., January 2003.

[4] ——, *IEEE Std 802.16-2004*, IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Std., October 2004.

[5] ——, *IEEE 802.16e-2005*, IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems - Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, Std., February 2006.

[6] *The IEEE 802.11 Handbook, A Designer’s Companion*, 2005.

[7] S. Hwang, J. Um, M. Song, C. Kim, H. Park, and Y. Kim, “Design and Verification of IEEE 802.22 WRAN Physical Layer,” in *Proc. 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, 2008, pp. 1–6.

[8] C. Snow, L. Lampe, and R. Schober, “Performance Analysis and Enhancement of Multiband OFDM for UWB Communications,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 6, pp. 2182–2192, 2007.

[9] Y.-J. Choi, K. B. Lee, and S. Bahk, “All-IP 4G Network Architecture for Efficient Mobility and Resource Management,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 42–46, 2007.

[10] J. Cai, “Transceiver Design in OFDM Wireless Communication Systems Supporting Heterogeneous Packet Transmission,” Ph.D. dissertation, The University of Waterloo, 2004.

[11] M. K. Awad and X. Shen, “OFDMA Based Two-hop Cooperative Relay Network Resources Allocation,” in *Proc. IEEE International Conference on Communications (ICC)*, vol. 2, 2008, pp. 4414–4418.

[12] M. Mehrjoo and X. Shen, “Resource Management Techniques for IEEE 802.16 Wireless Metropolitan Area Networks,” in *Encyclopedia of Wireless and Mobile Communications*. CRC Press - Taylor and Francis, 2007, ch.

[13] S. Kim, X. Wang, and M. Madihian, “Optimal Resource Allocation in Multi-Hop OFDMA Wireless Networks with Cooperative Relay,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1833–1838, 2008.

[14] H. Liu and G. Li, *OFDM-Based Broadband Wireless Networks: Design and Optimization*. Hobken, New Jersey:Wiley-Interscience, 2005.

[15] J. So, “Performance Analysis of VoIP Services in the IEEE 802.16e OFDMA System with Inband Signaling,” *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1876–1886, 2008.

[16] M. Pischella and J. Belfiore, “Resource Allocation for QoS-aware OFDMA Using Distributed Network Coordination,” *IEEE Trans. Veh. Technol.*, accepted for future publication.

[17] G. Li and H. Liu, “Resource Allocation for OFDMA Relay Networks With Fairness Constraints,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 11, pp. 2061–2069, 2006.

[18] M. Dianati, X. Shen, and K. Naik, “A New Fairness Index for Radio Resource Allocation in Wireless Networks,” in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 2, 2005, pp. 712–717.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[20] M. Mehrjoo, S. Moazeni, and X. Shen, “A New Modeling Approach for Utility-Based Resource Allocation in OFDM Networks,” in *Proc. IEEE International Conference on Communications (ICC)*, 2008, pp. 337–342.

[21] C. Mohanram and S. Bhashyam, “Joint Subcarrier and Power Allocation in Channel-Aware Queue-Aware Scheduling for Multiuser OFDM,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 9, pp. 3208–3213, 2007.

[22] Y. Chen and J. Chen, “A Fast Subcarrier, Bit, and Power Allocation Algorithm for Multiuser OFDM-Based Systems,” *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 873–881, 2008.

[23] K. Letaief and Y. Zhang, “Dynamic Multiuser Resource Allocation and Adaptation for Wireless Systems,” *IEEE Wireless Commun. Mag.*, vol. 13, no. 4, pp. 38–47, 2006.

[24] M. Ergen, S. Coleri, and P. Varaiya, "QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems," *IEEE Trans. Broadcast.*, vol. 49, no. 4, pp. 362–370, 2003.

[25] J. Jang and K. Lee, "Transmit Power Adaptation for Multiuser OFDM Systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171–178, 2003.

[26] Y. Zhang and K. Letaief, "Multiuser Adaptive Subcarrier-and-Bit Allocation with Adaptive Cell Selection for OFDM Systems," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1566–1575, 2004.

[27] Z. Shen, "Multiuser Resource Allocation in Multichannel Wireless Communication Systems," Ph.D. dissertation, The University of Texas at Austin, 2006.

[28] J. Cai, X. Shen, and J. Mark, "Downlink Resource Management for Packet Transmission in OFDM Wireless Communication Systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1688–1703, 2005.

[29] W. Rhee and J. Cioffi, "Increase in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation," in *Proc. IEEE Vehicular Technology Conference (VTC)*, 2000, pp. 1085–1089.

[30] G. Song and Y. Li, "Cross-Layer Optimization for OFDM Wireless Networks-Part II: Algorithm Development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, 2005.

[31] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, 2006.

[32] M. Mehrjoo, M. Dianati, X. Shen, and K. Naik, "Opportunistic Fair Scheduling for the Downlink of IEEE 802.16 Wireless Metropolitan Area Networks," in *Proc. of 3rd international conference on Quality of service in heterogeneous wired/wireless networks (QShine)*, 2006, pp. 52–60.

[33] M. Mehrjoo, X. Shen, and K. Naik, "A Joint Channel and Queue-Aware Scheduling for IEEE 802.16 Wireless Metropolitan Area Networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2007, pp. 1549–1553.

[34] M. Mehrjoo, M. K. Awad, and X. Shen, "Resource Allocation for OFDM-Based WiMAX," in *WiMAX network planning and optimization*, Y. Zhang, L. Yang, and J. Ma, Eds. CRC Press - Taylor and Francis, in press, ch.

[35] M. Mehrjoo and X. Shen, "An Efficient Scheduling Scheme for Heterogeneous Traffic in IEEE 802.16 Wireless Metropolitan Area Networks," in *Proc. International Symposium on Telecommunications (IST)*, 2008, pp. 263–267.

[36] M. Mehrjoo, S. Moazeni, and X. Shen, "An Interior Point Penalty Method for Utility Maximization Problems in OFDMA Networks," submitted to IEEE International Conference on Communications (ICC), 2009.

[37] M. Mehrjoo, M. K. Awad, M. Dianati, and X. Shen, "Opportunistic Fair Scheduling for OFDMA Broadcast Channels," submitted to IEEE International Conference on Communications (ICC), 2009.

[38] M. Mehrjoo, S. Moazeni, and X. Shen, "Resource Allocation in OFDMA Networks Based on Interior Point Methods," prepared for submission, 2008.

[39] M. Mehrjoo, M. K. Awad, M. Dianati, and X. Shen, "Opportunistic Fair Scheduling for OFDMA Networks with Heterogeneous Traffic Support," prepared for submission, 2008.

[40] M. Fathi, M. Mehrjoo, and H. Taheri, "A Stability-Based Scheduling Scheme for OFDMA Networks," submitted to IEEE International Conference on Communications (ICC), 2009.

[41] M. K. Awad, M. Veluppillai, M. Mehrjoo, X. Shen, and J. W. Mark, "Downlink Resource Allocation for OFDMA-Based Multiservice Networks with Imper-

fect CSI," submitted to IEEE International Conference on Communications (ICC), 2009.

[42] B. Lin, M. Mehrjoo, P. Ho, L. Xie, and X. Shen, "Capacity Enhancement with Relay Station Placement in Wireless Cooperative Networks," submitted to IEEE Wireless Communications and Networking Conference (WCNC), 2009.

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

[44] F. Capar and F. Jondral, "A Rayleigh Fading Channel Model for Multicarrier Systems: A Tapped Delay Line Model," in *Proc. IEEE 59th Vehicular Technology Conference (VTC)*, 2004, pp. 181–185.

[45] M. K. Awad, K. T. Wong, and Z. Li, "An Integrated Overview of the Open Literature's Empirical Data on the Indoor Radiowave Channel's Delay Properties," *IEEE Trans. Antennas Propag.*, vol. 56, no. 5, pp. 1451–1468, 2008.

[46] P. Vandenameele, L. Perre, and M. Engels, *Space Division Multiple Access for Wireless Local Area Networks*. Kluwer Academic Publishers, 2002.

[47] J. W. Mark and W. Zhuang, *Wireless Communications and Networking*. Upper Saddle River, New Jersey, USA: Pearson Education International, 2003.

[48] S. Haykin and M. Moher, *Modern Wireless Communications*. Pearson/Prentice Hall, 2005.

[49] K. Fazel and S. Kaiser, *Multi-Carrier and Spread Spectrum Systems*. John Wiley & Sons, 2003.

[50] R. W. Chang, "Synthesis of Band-Limited Orthogonal Signals for Multichannel Data Transmission," *Bell syst. tech. J.*, vol. 45, pp. 1775–1796, 1966.

[51] B. R. Saltzberg, "Performance of an Efficient Parallel Data Transmission System," *IEEE Trans. Commun.*, vol. 15, pp. 805–811, 1967.

[52] Z. Abichar, Y. Peng, and J. Chang, "WiMax: The Emergence of Wireless Broadband," *IEEE IT Prof.*, vol. 8, no. 4, pp. 44–48, 2005.

[53] H. Yin and S. Alamouti, "Ofdma: A broadband wireless access technology," in *Proc. IEEE Sarnoff Symposium*, 2006, pp. 1–4.

[54] K. Kumaran and H. Viswanathan, "Joint Power and Bandwidth Allocation in Downlink Transmission," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1008–1016, 2005.

[55] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[56] I. Kim, H. Lee, B. Kim, and Y. Lee, "On the Use of Linear Programming for Dynamic Subchannel and Bit Allocation in Multiuser OFDM," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2001, pp. 3648–3652.

[57] S. Pietrzyk, G. J. M. Janssen, A. N. Unit, P. T. C. Sp, and P. Warsaw, "Radio Resource Allocation for Cellular Networks Based on OFDMA with QoS Guarantees," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2004, pp. 2694–2699.

[58] C. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, 1999.

[59] Y. Chang, F. Chien, and C. Kuo, "Performance Comparison of OFDM-TDMA and OFDMA with Cross-Layer Consideration," in *Proc.IEEE 64th Vehicular Technology Conference (VTC)*, 2006, pp. 1–5.

[60] G. Song and Y. Li, "Adaptive Resource Allocation Based on Utility Optimization in OFDM," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2003, pp. 586–590.

[61] X. Zhang, E. Zhou, R. Zhu, S. Liu, and W. Wang, “Adaptive Multiuser Radio Resource Allocation for OFDMA Systems,” in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2005, pp. 3846–3850.

[62] D. Niyato and E. Hossain, “Adaptive Fair Subcarrier/Rate Allocation in Multirate OFDMA Networks: Radio Link Level Queuing Performance Analysis,” *IEEE Trans. Veh. Technol.*, vol. 55, no. 6, pp. 1897–1907, 2006.

[63] S. Haykin, *Digital communications*. John Wiley & Sons, 1988.

[64] S. Chung and A. Goldsmith, “Degrees of Freedom in Adaptive Modulation: A Unified View,” *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, 2001.

[65] X. Qiu and K. Chawla, “On the Performance of Adaptive Modulation in Cellular Systems,” *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 884 – 895, 1999.

[66] J. Cai, X. Shen, and J. Mark, “Robust Channel Estimation for OFDM Wireless Communication Systems - an H∞ Approach,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 2060–2071, 2004.

[67] H. Minn and V. Bhargava, “An Investigation into Time-Domain Approach for OFDM Channel Estimation,” *IEEE Trans. Broadcast.*, vol. 46, pp. 240–248, 2000.

[68] O. Edfors, M. Sandell, and V. Beek, “OFDM Channel Estimation by Singular Value Decomposition,” *IEEE Trans. Commun.*, vol. 45, pp. 931–939, 1998.

[69] P. Schramm and R. R. Muller, “Multicarrier Signal Detection and Parameter Estimation in Frequency-Selective Rayleigh Fading Channels,” *IEEE Trans. Commun.*, vol. 45, pp. 1560–1563, 1998.

[70] S. Coleri, M. Ergen, A. Puri, and A. Bahai, “Channel Estimation Techniques Based on Pilot Arrangement in OFDM Systems,” *IEEE Trans. Broadcast.*, vol. 48, no. 3, pp. 223–229, 2002.

[71] Y. Cheng, X. Ling, W. Song, L. X. Cai, W. Zhuang, and X. Shen, "A Cross-Layer Approach for WLAN Voice Capacity Planning," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 678–688, 2007.

[72] V. Mahinthan, H. Rutagemwa, J. W. Mark, and X. Shen, "Cross-Layer Performance Study of Cooperative Diversity System Equipped with ARQ," *IEEE Trans. Veh. Technol.*, in press.

[73] P. Pham, S. Perreau, and A. Jayasuriya, "New Cross-Layer Design Approach to adhoc Networks under Rayleigh Fading," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 28–39, 2005.

[74] M. Bussieck and A. Pruessner, "Mixed-Integer Nonlinear Programming," *Society for Industrial and Applied Mathematics/Optimization Activity Group (SIAG/OPT) Newsletter: Views and News*, vol. 14, no. 1, pp. 1–7, 2003.

[75] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1996.

[76] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.

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

[78] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal Resource Allocation for OFDMA Downlink Systems," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2006, pp. 1394–1398.

[79] S. Kandukuri and S. Boyd, "Optimal Power Control in Interference-Limited Fading Wireless Channels with Outage-Probability Specifications," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 46–55, 2002.

[80] M. Dianati, X. Shen, and K. Naik, "Cooperative Fair Scheduling for the Downlink of CDMA Cellular Networks," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1749–1760, 2007.

[81] G. Song and Y. Li, "Cross-layer Optimization for OFDM Wireless Networks-Part I: Theoretical Framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, 2005.

[82] A. Mordecai, *Nonlinear Programming: Analysis and Methods*. 1976: Englewood Cliffs, N.J.: Prentice Hall.

[83] S. A. Vavasis, *Nonlinear optimization: complexity issues*. New York, NY, USA: Oxford University Press, Inc., 1991.

[84] A. Forsgren, P. E. Gill, and M. H. Wright, "Interior Methods for Nonlinear Optimization," *Society for Industrial and Applied Mathematics (SIAM)*, vol. 44, no. 4, pp. 525–597, 2002.

[85] D. F. Shanno and R. J. Vanderbei, "Interior-Point Methods for Non-convex Nonlinear Programming: Orderings and Higher-Order Methods," *Math. Program.*, vol. 87, pp. 303–316, 2000.

[86] D. Whitley, "A Genetic Algorithm Tutorial," *Statistics and Computing*, vol. 4, no. 2, pp. 65–85, 1994.

[87] C. Marrison and R. Stengel, "The Use of Random Search and Genetic Algorithms to Optimize Stochastic Robustness Functions," in *Proc. of American Control Conference*, vol. 2, 1994, pp. 1484–1489.

[88] D. M. Chiu and A. S.-W. Tam, "Fairness of Traffic Controls for Inelastic Flows in the Internet," *Comput. Netw.*, vol. 51, no. 11, pp. 2938–2957, 2007.

[89] R. Knopp and P. A. Humblet, "Information Capacity and Power Control in Single Cell Multiuser Communications," in *Proc. IEEE International Conference on Communications (ICC)*, 1995, pp. 331–335.

[90] Y. Liu and E. Knightly, "Opportunistic Fair Scheduling over Multiple Wireless Channels," in *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, 2003, pp. 1106–1115.

[91] Z. Zhang, Y. He, and E. K. P. Chong, "Opportunistic Scheduling for OFDM Systems with Fairness Constraints," *EURASIP J. Wirel. Commun. Netw.*, vol. 2008, no. 2, pp. 1–12, 2008.

[92] A. Wang, L. Xiao, S. Zhou, X. Xu, and Y. Yao, "Dynamic Resource Management in the Fourth Generation Wireless Systems," in *Proc. International Conference on Communication Technology (ICCT)*, 2003, pp. 1095–1098.

[93] T. Harks, "Utility Proportional Fair Bandwidth Allocation: An Optimization Oriented Approach," in *Proc. of 3rd International Workshop on QoS in Multiservice IP Networks (QoS-IP)*. Springer, 2005, pp. 61–74.

[94] F. Kelly, "Charging and Rate Control for Elastic Traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, 1997.

[95] J.-Y. Boudec, "Rate Adaptation, Congestion Control and Fairness: A Tutorial," [online]. Available:http://ica1www.epfl.ch/PS_files/LEB3132.pdf.

[96] W. Yu and R. Lui, "Dual Methods for Nonconvex Spectrum Optimization of Multicarrier Systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, 2006.

[97] L. Nuaymi, *WiMAX: Technology for Broadband Wireless Access*. John Wiley & Sons, 2007.

[98] Z. Han, Z. Ji, and K. Liu, "Power Minimization for Multi-Cell OFDM Networks Using Distributed Non-Cooperative Game Approach," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2004, pp. 3742–3747.

[99] Y.-J. Choi and S. Bahk, "Delay-Sensitive Packet Scheduling for a Wireless Access Link," *IEEE Trans. Mobile Comput.*, vol. 5, no. 10, pp. 1374–1383, 2006.

[100] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-Layer Design for Wireless Networks,," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 74–80, 2003.

[101] T. S. Rappaport, *Wireless Communications: Principles and Practice*. 2nd edition, Prentice Hall PTR, 2001.

[102] A. Leke and J. Cioffi, "Impact of Imperfect Channel Knowledge on the Performance of Multicarrier Systems," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 1998, pp. 951–955.

[103] G. Song, "Cross-Layer Resource Allocation and Scheduling in Wireless Multicarrier Networks," Ph.D. dissertation, Georgia Institute of Technology, 2005.

[104] Y. Yi and M. Chiang, "Stochastic Network Utility Maximisation-A Tribute to Kellys Paper Published in This Journal a Decade Ago," *EUROPEAN TRANSACTIONS ON TELECOMMUNICATIONS*, vol. 19, no. 4, pp. 421–442, 2008.

[105] R. J. La and V. Anantharam, "Utility-Based Rate Control in the Internet for Elastic Traffic," *IEEE/ACM Trans. Netw.*, vol. 10, no. 2, pp. 272–286, 2002.

[106] N. Lu and J. Bigham, "Utility-Maximization Bandwidth Adaptation for Multi-class Traffic QoS Provisioning in Wireless Networks," in *Proc. 1st ACM International Workshop on Quality of Service & Security in Wireless and Mobile Networks (Q2SWinet)*, 2005, pp. 136–143.

[107] C. E. Luna, L. P. Kondi, and A. K. Katsaggelos, "Maximizing User Utility in Video Streaming Applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 2, pp. 141–148, 2003.

[108] Z. Han and K. J. R. Liu, *Resource Allocation for Wireless Networks, Basics, Techniques, and Applications*. Cambridge University Press, 2008.

[109] T. Ng and W. Yu, "Joint Optimization of Relay Strategies and Resource Allocations in a Cooperative Cellular Network," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 2, pp. 328–339, 2007.

[110] M. Shen, G. Li, and H. Liu, “Design Tradeoffs in OFDMA Traffic Channels,” in *proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 757–760.