CROSS-LAYER RESOURCE ALLOCATION FOR MULTI-USER COMMUNICATION SYSTEMS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL
ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Kibeom Seong
March 2008

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(John M. Cioffi) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Arogyaswami Paulraj)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Ravi Narasimhan)

Approved for the University Committee on Graduate Studies.

Abstract

Satisfaction of different QoS demands for various broadband services in wireless networks requires that multi-user packet scheduling intelligently use both channel state information (CSI) and queue state information (QSI). A combination of queue-channel-aware scheduling and power/rate allocation on each transmit dimension is known as "cross-layer resource allocation". This thesis investigates various aspects of cross-layer resource allocation to illustrate its important role in multi-user communication systems. Of particular interest are downlink and uplink wireless systems that use orthogonal frequency division multiplexing (OFDM) modulation and multi-input multi-output (MIMO) transmission with multiple antennas.

There are four major contributions in this thesis: First, queue-proportional scheduling (QPS) is presented and is shown to exhibit superior throughput, delay, and fairness properties. QPS provides a capability that can arbitrarily scale each user's average queueing delay relative to others, which makes QPS suitable for networks driven by heterogeneous traffic. Second, geometric programming (GP) is applied for cross-layer resource allocation in multi-user OFDM systems with CSI, where GP formulations lead to numerical efficiency and strong scalability. Third, efficient power/rate optimization algorithms are developed by using Lagrange dual decomposition for multi-user MIMO-OFDMA (Orthogonal Frequency Division Multiple Access) systems with CSI. Finally, cross-layer resource allocation in multi-user MIMO-OFDMA systems with channel distribution information (CDI) is addressed. It is shown that outage rate region for scheduling can be efficiently characterized by using a Gaussian approximation of mutual information along with a successive feasibility check method. This efficient approach is directly applicable to finding power/rate allocation for QPS as well. Stochastic simulations on a variety of situations demonstrate that QPS outperforms other well-known scheduling policies in terms of throughput, average queueing delay, and delay fairness among the users.

Acknowledgments

I have been extremely fortunate to meet and interact with such outstanding individuals at Stanford. Without the invaluable lessons and support from them, I wouldn't be able to complete any part of this thesis. It is my great pleasure to express my heartfelt appreciations to them.

Foremost, I would like to thank my principal adviser, Prof. John M. Cioffi. I have tremendously benefited from his amazing technical insight and guidance throughout my PhD years. His excellent communication classes taught me the fundamentals of multi-user resource allocation that laid foundations of this dissertation. When I was losing confidence in my research direction, his warm encouragement and continuous support lifted my sprit and motivated me to run again, which eventually led me to this final stage. Prof Cioffi is my best role model as a supervisor and a researcher. I strongly desire to grow into a professional like him, with passion, intelligence, diligence, and good personality.

My special thanks go to my associate adviser, Prof. Arogyaswami Paulraj. I was honored to start my first year at Stanford with his generous help and guidance. His sincere advices deeply motivated me from the very beginning so that I was successfully able to adapt to the new life at Stanford. Moreover, his invaluable class on the space-time wireless communications taught me the fundamentals of the MIMO systems that are widely used in this thesis.

I'd like to extend my appreciations to my reading committee member, Prof. Ravi Narasimhan in University of California, Santa Cruz. He has held regular weekly meetings on wireless topics for our group, and I was very fortunate to join that meeting from the beginning. Many results in this thesis came out during our informative discussions, and his shrewd comments were immensely helpful. I would also like to thank Prof. John Gill for willingly serving as a chair of my oral exam committee.

I am deeply grateful to my master's degree adviser, Prof. Kwang Bok Lee in Seoul National University. For the past 10 years, I asked for his precious advices whenever I had to make some critical decisions on my career path. With his encouragement, I have been able to stay focused to make full use of my potential. I wouldn't be able to come this far without his guidance and support.

I also wish to acknowledge the Stanford Graduate Fellowship program for providing me with the full scholarship for three years and the Samsung Lee Kun Hee Scholarship program for providing a partial financial support.

I was very fortunate to have the opportunities to collaborate with many brilliant and fun people at Stanford. I would like to thank Chiang-yu Chen, Chan-Soo Hwang, Sumanth Jagannathan, Youngjae Kim, Bin Lee, Wooyul Lee, Mehdi Mohseni, Tsuyoshi Tamaki, David Yu, and Rui Zhang for their co-authorship and collaboration on conference or journal papers. It was truly enjoyable to work with them, and I learned a lot from our innumerable discussions.

I would like to thank the former and current members in Prof. Cioffi's research group: Rajiv Agarwal, Mark Brady, Jungsub Byun, Chiang-yu Chen, Sunghyun Cho, Jiwoong Choi, Won-Joon Choi, Seong Taek Chung, Amal Ekbal, Gustavo Fraidenraich, Chan-Soo Hwang, Sumanth Jagannathan, Edward Woongjun Jang, Joonsuk Kim, Youhan Kim, Youngjae Kim, Ryoulhee Kwak, Hyukjoon Kwon, Bin Lee, Inkyu Lee, Jungwon Lee, Wooyul Lee, Vinay Majjigi, Moshe Malkin, Tadashi Minowa, Mehdi Mohseni, Jisung Oh, Sung Hwan Ong, Vahbod Pourahmad, Wonjong Rhee, Kee-Bong Song, Tsuyoshi Tamaki, Dimitris-Alexandros Toubakaris, Shu-ping Yeh, David Yu, and Rui Zhang. I am blessed to know these many extraordinary people and have an opportunity to share innovative thoughts and opinions together. It has been my great honor to be a member of this distinguished group. Also, my special thanks go to our administrative assistants, Joice DeBolt and Pat Oshiro, for their outstanding administrative support. I will miss their kind care and cheerful greetings.

I would like to thank my Korean friends who have made my life at Stanford so memorable: Hoyon Hwang, Jungsoon Jang, Sangoh Jeong, Jaedon Kim, Taeksoo Kim, Young-Han Kim, Jongho Lee, Jongmin Lee, Taesup Moon, Jeonghun Noh, Taegon Park, Seungbum Rim, Wonjoo Suh, Kyoungho Woo, Taesang Yoo, Sungroh Yoon, and many others. I would also like to extend my gratitude to the Fields family (Davis, Mary, Davy and Jim) for letting me be a part of their family since my arrival in the U.S.

Last, but most importantly, I would like to express my sincere appreciations to my family members. I would like to thank my parents and my brother who have always motivated me to pursue my dreams. I am deeply indebted to them for their unconditional love and care. I would also like to thank my parents-in-law and brother-in-law for their constant encouragement. I wish to express my deepest gratitude and love to my wife, Hyunjin, for her endless love that has made this dissertation possible. She is the reason for my blissful life at Stanford. My thanks also go to my son, Yunjune, for his adorable smile which means the World to his daddy. I dedicate this dissertation to my family.

Kibeom Seong
March 2008

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The evolution of wireless communication systems has been dramatic in the past decade. The main application of traditional wireless cellular systems has been voice communications, where one of major design goals is seamless conveyance of stable voice signals to many users with low delay. Along with the rapid evolution of Internet, the need of data communications has been significantly increasing, and the wireless systems such as wireless local area network (WLAN) have been developed to support packet data services such as web-browsing and file transfer protocol (FTP). Future wireless networks, as illustrated in Fig. 1.1, will be driven by a variety of ubiquitous broadband services such as portable telephony, mobile Internet, Voice over IP (VoIP), and IPTV that require different Quality of Service (QoS). Hence, a key design issue is satisfaction of different service classes' QoS requirements. For example, if an operator supports video, voice, and streaming, three different service classes can be defined by the operator containing different sets of QoS parameters such as throughput and delay. Table 1.1 presents an example for a pre-configured QoS requirement of these service classes. In order to meet these distinct QoS demands, as well as to guarantee fairness among the users, development of intelligent multi-user packet schedulers is essential.

A recent breakthrough in wireless communications is the application of link adaptation to reduce the performance loss from channel fading. Channel fading is inevitable in a wireless environment and is caused by signal scattering of a number of

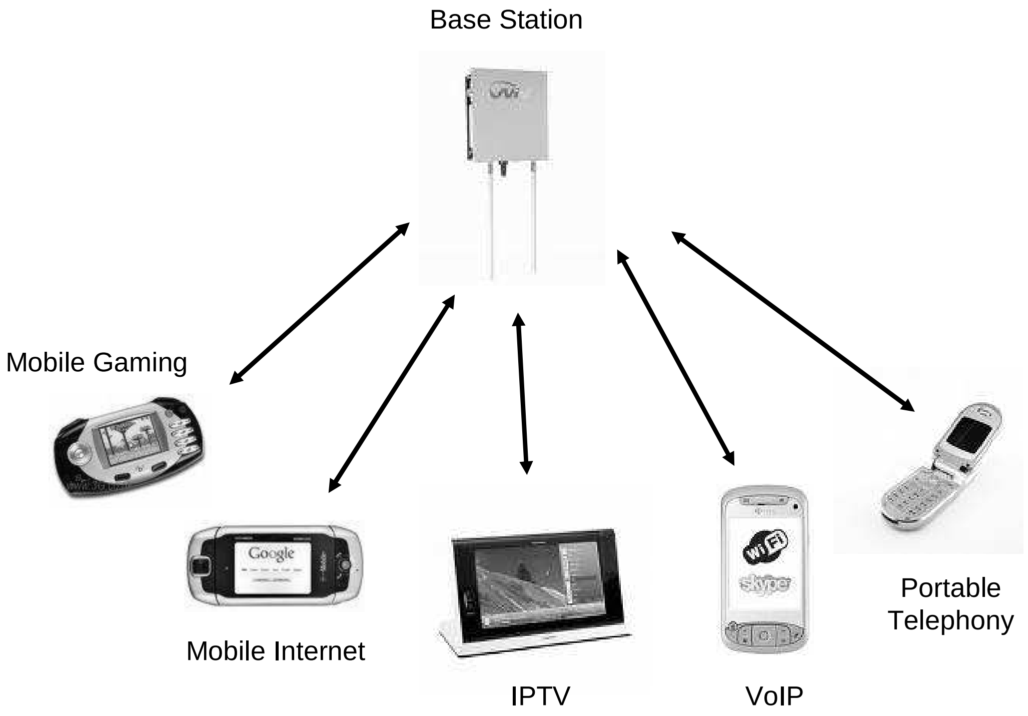


Figure 1.1: Wireless network driven by various ubiquitous broadband services

objects in the channel [25]. Traditionally, fading is considered as a detrimental effect and the design focus was on circumventing this effect through the use of diversity techniques. However, if channel state information (CSI) such as channel gains is available at transmitter, fading can be exploited by adapting the transmit power and data-rate allocation according to the channel conditions [26]. The receiver's CSI can be delivered to the transmitter via a reliable feedback link, or if TDD (Time Division Duplex) mode is utilized, channel reciprocity* can be assumed and the transmitter can equate downlink CSI with the estimated uplink CSI. By opportunistically using the channel fluctuation, achievable throughput can be significantly increased [27, 72].

*Channel reciprocity means that both downlink and uplink channels are identical. This property can hold in the TDD mode if the channel coherence time is much longer than the half-duplex cycle and the interference levels at both transmitter and receiver are assumed to be the same.

The link-adaptation technique was first utilized in a stationary channel for DSL (Digital Subscriber Line) systems, where it is called *loading* [64]. Link adaptation can be extended to wireless systems by carefully considering more rapid channel variations. The multi-user scheduling issues in wireless systems have become more interesting with the link-adaptation technique, which provides better capability to meet a variety of QoS requirements.

Table 1.1: Example for QoS requirement

	Video	Voice	Streaming
Min Reserved Rate	500 kbps	12.2 kbps	50 kbps
Tolerated Delay Jitter	50 ms	20 ms	80 ms
Max Traffic Burst	2 Mbps		150 kbps
Max Sustained Rate	1 Mbps		100 kbps

Realization of link adaptation for multi-user communication systems requires optimal allocation of communication resources, such as the transmit power and data rate. The architecture of many communication networks falls into one of two categories: the broadcast channel (BC) and the multiple-access channel (MAC) [20, 19]. Examples of BC and MAC are the downlink and uplink of cellular networks, respectively. In the uplink MAC, a number of mobile terminals send independent information to the base-station (BS), and in the downlink BC, the BS broadcasts messages, which are often independent, to all the mobile terminals. With perfect CSI at both the BS and each mobile terminal (MT), each user's transmit power and rate in the downlink and uplink can be determined based on the capacity regions of BC and MAC, respectively. Much research work has focused on this information-theoretic approach to resource allocation. However, because these approaches ignore the randomness in packet arrivals and queueing, they have difficulty in guaranteeing each user's throughput or delay requirement. Therefore, satisfaction of each user's QoS requirement needs to consider queue state information (QSI), such as the current queue backlog size, in addition to CSI when the scheduler selects each user's operating data rate. Once service rates are determined by a scheduler in the medium access control layer (MACL), the physical layer (PHYL) determines the corresponding power and rate allocation in each transmit dimension. Because of this interplay between MACL and PHYL, such a combination of queue-channel-aware scheduling and power/rate allocation is known as *cross-layer resource allocation*. This cross-layer approach to resource allocation to account for queueing parameters has been recently studied in [5, 45, 3, 73, 65] and the references therein.

Reference [45] defines *network capacity region* as a set of all packet arrival rate vectors for which it is possible to keep every queue length finite. For bursty input traffic, it is generally difficult to estimate the packet-arrival rates. Thus, resource allocation solely based on CSI is unable to update rate allocation properly according to the dynamics of the input traffic. As a result, even for a packet arrival rate vector within the network capacity region, some users' queue backlogs may become unacceptably large, causing long queueing delay as well as buffer overflow. Certain schedulers assume infinite queue backlog and simply consider CSI only. On the other hand in the packet switching systems, the channel can be described by the circuit connection, so the concept of CSI is irrelevant. The above two schedulers are called 'channel-aware scheduler' and 'queue-aware-scheduler', respectively, to distinguish them from the queue-channel-aware scheduler[†]. The network capacity region may not be achievable with channel-aware scheduler that ignores queueing dynamics. In addition, each user's queueing delay, which is an important QoS parameter, is uncontrollable without considering queue sizes in scheduling. Therefore, it is essential to design an intelligent scheduler that considers both CSI and QSI. It has been shown that the entire network capacity region can be achieved by using this approach in the fading BC and MAC [45, 67, 77].

With increasing demand for high data-rate services, Orthogonal Frequency Division Multiplexing (OFDM) has drawn much attention as a promising technique. OFDM has been widely applied to a variety of telecommunication systems such as WLAN, DVB (Digital Video Broadcasting), and to DSL systems in DMT (Discrete Multi-Tone) form. By converting a frequency selective fading channel into parallel frequency-flat fading subchannels, OFDM achieves high spectral efficiency, as well as

[†] In this thesis, 'scheduling' often implies queue-channel-aware scheduling for simplicity.

lower equalization complexity, for channels with large delay spread [18]. With perfect CSI at both base station and mobile terminals, as the number of tones goes to infinity, OFDM with link adaptation is shown to achieve the capacity of Gaussian BC and MAC with inter-symbol interference (ISI), or with frequency selective fading. OFDM inherently provides a variety of advantages such as FFT (Fast Fourier Transform) realization, simple channel equalization, no ISI, and high degree of design flexibility. Its application to multi-user systems, multi-user OFDM has been considered as a strong candidate for the platform of the future wireless systems. In multi-user OFDM systems, multiple users are allowed to transmit simultaneously from the intelligent tone assignment to each user, which has triggered much work on the power/rate allocation issues in this system.

Achievable throughput can be further boosted by applying MIMO (Multiple Input Multiple Output) techniques to OFDM systems where the transmitter and receiver are equipped with multiple antennas. In a rich scattering environment with the antenna separation in the order of a wave length, each component of channel matrix undergoes independent and identically distributed (i.i.d.) fading. Thus, it is possible to enhance the performance by exploiting spatial diversity effects through the application of space-time channel coding [1, 66]. If CSI is available at the transmitter, the channel capacity, which is shown to be proportional to the number of transmit antennas, can be achieved by optimizing the transmit power across the transmit antennas, after precoding at the transmitter and postprocessing at the receiver based on SVD (Singular Value Decomposition). Thus, MIMO-OFDM, the combination of MIMO with OFDM modulation, has been considered as a key feature for the future wireless systems.

Scheduling in multi-user MIMO-OFDM systems is much more complex than that in TDMA (Time Division Multiple Access) systems where only one MT is allowed to communicate with the BS in each time slot. Because of its simplicity, TDMA is one of the most widely deployed transmission methods in wireless communication. Assuming a single-cell environment, there are no multi-user interactions in each time slot with TDMA, so the achievable rate region can be simply defined from the maximum achievable rate of each user. Therefore, once each user's maximum achievable rate is calculated, the scheduler at the BS can select which user communicates by considering both the finite set of operation rates and QSI to meet delay, throughput, and fairness criteria. Then, the power and rate on each transmit dimension are optimized for the scheduled user. On the other hand, in multi-user MIMO-OFDM systems, multiple users with different QoS demands can transmit at the same time; thus, characterization of the achievable rate region is much more complicated. It is also infeasible to generate the finite number of operation points in PHYL and pass them over to MACL. Thus, the queue-channel-aware scheduling and allocation of power/rate on each tone and transmit antenna must be performed in a combined way.

This thesis comprehensively addresses various aspects of cross-layer resource allocation in multi-user communication systems. The focus is on downlink and uplink OFDM/MIMO-OFDM with or without CSI at the transmitter. Cross-layer resource allocation shows significant improvement in throughput, delay, and fairness properties by intelligently utilizing both CSI and QSI. In particular, *queue-proportional scheduling* (QPS) is presented, and its throughput, delay, and fairness properties are shown to be superior to other well-known scheduling policies. In order to apply QPS and other schedulers to downlink and uplink OFDM/MIMO-OFDM systems with and without CSI, this thesis develops a variety of efficient power/rate optimization algorithms. The rest of this chapter elaborates on practical motivations of this work as well as overviews, and the main contributions of each chapter are outlined.

1.1 Motivation

With the increasing demands on multimedia services from applications such as video streaming and IPTV, a variety of ubiquitous real-time and non-real-time broadband services need to be simultaneously supported in the future wireless networks, where each service demands a different QoS. Non-real-time best effort traffic such as web-browsing and FTP service targets maximization of throughput while tolerating some degree of packet delay. On the other hand, real-time traffic such as potable telephony has a more strict delay constraint with much lower data rate. In order to satisfy these diversified QoS demands, design of intelligent multi-user packet schedulers becomes a key issue in wireless systems. In particular, schedulers must carefully use both CSI and QSI to achieve better throughput, delay, and fairness properties.

TDMA has been widely applied in wireless systems where only one user transmits at each time slot. However, many promising wireless systems under development such as mobile WiMAX (Worldwide Interoperability for Microwave Access) [34] and 3GPP LTE (The 3rd Generation Partnership Project Long Term Evolution) [69] adopt OFDMA (Orthogonal Frequency Division Multiple Access). These applications' OFDM modulation allows simultaneous transmission by multiple users where each tone can be occupied by only one user. In addition, mobile WiMAX and 3GPP LTE systems incorporate MIMO transmission technique with multiple transmit and receive antennas for further enhancement of spectral efficiency. This MIMO-OFDM platform is a strong candidate for future wireless systems.

By simultaneously supporting multiple users with different classes of traffic, higher throughput as well as effective controllability of queueing delay can be achieved with smart scheduling policies. Hence, scheduler design becomes more important and challenging in multi-user MIMO-OFDM systems. With simultaneous multi-user transmission, the scheduling operation is more intertwined with resource allocation; thus, it is desirable to combine them. Optimal power/rate allocation in multi-user systems based on OFDM modulation and MIMO transmission is an active and non-trivial research area. For various cases, optimal resource allocation is an open problem. Even for the cases with known optimal solutions, the numerical complexity is often very high, which makes the optimal solutions intractable with practical system settings.

The typical procedures of cross-layer resource allocation in downlink and uplink MIMO-OFDM systems are illustrated in Fig. 1.2 and Fig. 1.3, respectively. In the downlink case (Fig. 1.2), when the channel is slowly varying, the BS is able to secure each user's downlink CSI via a feedback channel, or by utilizing channel reciprocity in case of a TDD mode. Also, the output queues for each service are located at the BS; thus, perfect CQI at the BS is available for the downlink. Then, cross-layer resource allocation is performed based on both CSI and QSI at the BS to control the power and rate allocation on each user's transmit dimension. Finally, the BS delivers information on power/rate allocation to each corresponding user via control channels,

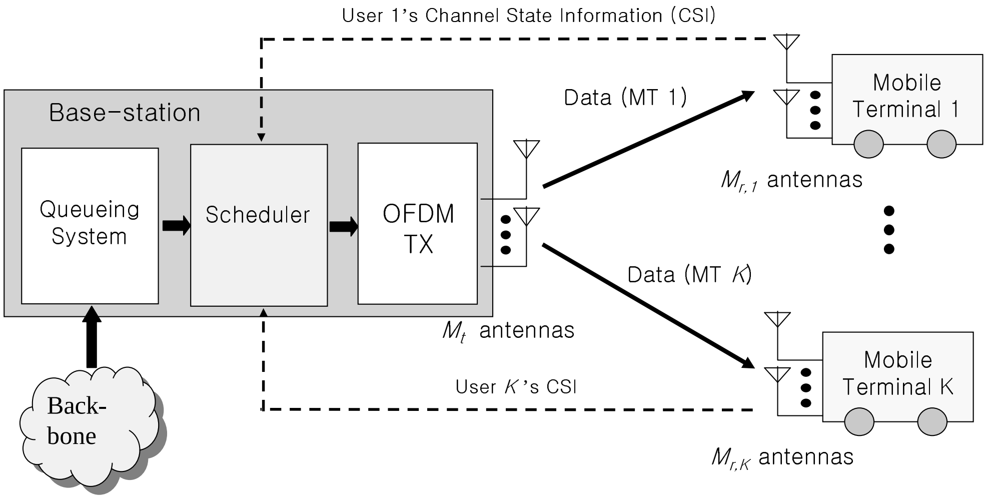


Figure 1.2: Downlink MIMO-OFDM systems

possibly at the same time with data packet delivery.

In Fig. 1.3's uplink with a slowly varying channel condition, the BS obtains uplink CSI from channel estimation. Since the queues are distributed at each MT in the uplink, each MT's QSI needs to be delivered to BS via a reliable link. Then, by using both CSI and QSI at the BS, cross-layer resource allocation calculates the transmit power/rate allocation, which is then delivered to each corresponding user via control channels. Each MT updates its power/rate allocation based on this information and transmits a new packet accordingly. However, in a highly mobile environment with fast channel variation, both the downlink and uplink can experience discrepancies between the CSI at the BS used for scheduling and the channel conditions when actual data transmission happens. Thus, the value of CSI would be limited, which may significantly degrade the system performance.

This thesis is motivated by the challenges and limitations mentioned above. The

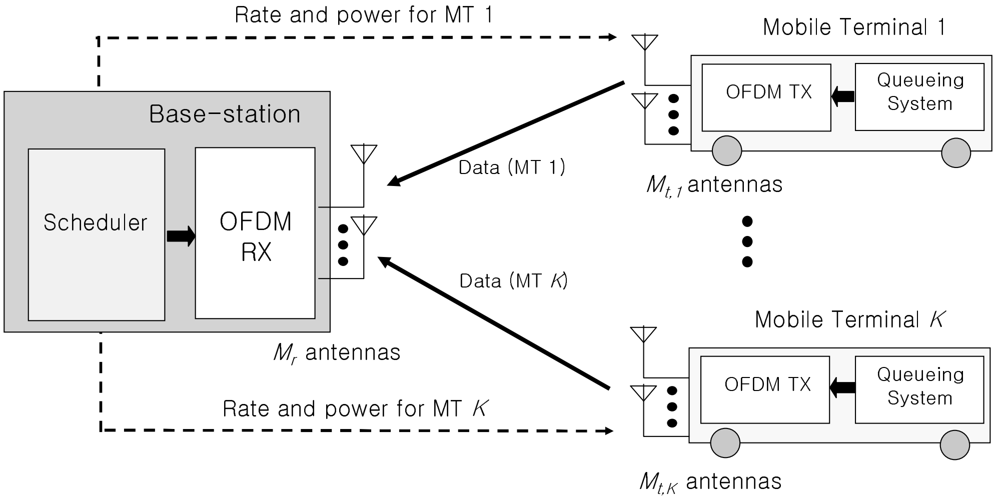


Figure 1.3: Uplink MIMO-OFDM systems

main objectives of this thesis can be summarized as follows:

- Corroboration of the advantages of cross-layer resource allocation.

- Design of intelligent schedulers suitable for meeting heterogeneous QoS requirement.

- Development of efficient power/rate optimization algorithms for cross-layer resource allocation in multi-user systems with and without CSIT (CSI at the Transmitter) that are based on OFDM modulation and MIMO transmission.

1.2 Overview of Thesis

This thesis is composed of seven chapters to cover the objectives listed in the previous section. This section summarizes each chapter with its key contributions emphasized.

Chapter 2 introduces the fundamentals of queueing systems and schedulers. The basics of queueing theory are essential to understand schedulers' throughput and delay performance. Considering a simple queueing system with Poisson packet arrivals and exponentially distributed service time, average queue length is analytically derived, and its relation to average queueing delay is addressed. Also, the stability analysis of queueing systems is provided, which is crucial in characterizing achievable throughput of schedulers. In addition, Chapter 2 illustrates models of the multi-user queueing systems and schedulers that are used throughout this thesis. Basic concepts frequent in the scheduling context such as a *network capacity region* and *throughput optimality* are explained with simple illustrations. Finally, various well-known schedulers including maximum weight matching scheduling (MWMS) are introduced, where they are categorized into channel-aware, queue-aware, and queue-channel-aware schedulers.

With all the background built in Chapter 1 and 2, Chapter 3 presents another scheduling policy called queue-proportional scheduling (QPS). The application of QPS to stationary and time-varying channels is illustrated, and throughput, delay, and fairness properties of QPS are analyzed in this chapter. It is discovered that the direction of average queue length vector under QPS always converges to that of arrival rate vector. Using this finding in addition to the queueing theories introduced in Section 2.1, QPS is proved to be throughput optimal. It is also shown that QPS has the capability of arbitrarily scaling the ratio of each user's average queueing delay relative to others. This unique delay-fairness property is desirable for satisfying diversified QoS requirement in a network. Finally, stochastic simulation results with Poisson packet arrivals and exponentially distributed packet lengths are presented for stationary and fading Gaussian BCs. QPS and various scheduling policies introduced in Section 2.2 are considered in the simulation. The results demonstrate that queue-channel-aware scheduling policies such as QPS and MWMS have much better throughput and delay properties than other types of schedulers, which corroborates the cross-layer approach to resource allocation. In addition, QPS is shown to outperform MWMS in terms of average queueing delay and to provide a more desirable delay-fairness property for QoS satisfaction. Part of the work in Chapter 3 is presented in [56, 54, 55].

Cross-layer resource allocation is defined as the combination of queue-channel-aware scheduling and power/rate allocation in each transmit dimension. Therefore, given schedulers, an essential part of cross-layer resource allocation develops efficient algorithms for power/rate allocation to support the scheduled rate tuple. Chapter 4-6 present a variety of efficient power/rate optimization techniques for multi-user broadcast and multiple-access systems based on OFDM modulation and MIMO transmission. As well as perfect CSIT situations (Chapter 4, 5), cross-layer resource allocation with no CSIT is addressed (Chapter 6).

First, Chapter 4 introduces a powerful optimization tool, *geometric programming* (GP) to cross-layer resource allocation for the OFDM BC and MAC with CSIT. GP is a special form of convex optimization problems for which very efficient solvers have been developed [10]. In an OFDM Gaussian BC with sum-power constraint, a degraded broadcast channel is formed at each tone where its capacity can be achieved by applying the superposition coding at the transmitter and successive interference cancellation at the receivers [24]. Also, the optimal encoding and decoding order is unique, given each tone's channel signal-to-noise ratio (SNR) [20]. By converting capacity equations with careful consideration of optimal orderings, it is shown that major resource allocation problems in the OFDM Gaussian BC can be formulated via GP. Also, the duality relation between the MAC and BC is used to extend the GP formulations in the BC to the MAC. The extension to fading channels is straightforward. With any additional rate constraints of linear form, the GP structure is still maintained in the derived equations, which makes GP a convenient tool for satisfying various throughput QoS demands in the network. After the introduction to GP, Chapter 4 derives GP formulations of QPS as well as of some other scheduling polices. Also, the results of stochastic simulations performed by solving the obtained GP equations are presented. Numerical efficiency and strong scalability of GP make GP suitable for cross-layer resource allocation in multi-user OFDM systems with CSIT driven by heterogeneous QoS requirement. Part of the work in Chapter 4 is presented in [56, 54, 58].

Chapter 5 is devoted to cross-layer resource allocation in downlink and uplink MIMO-OFDMA systems with CSIT. In MIMO-OFDMA, it is assumed that each tone is taken by at most one user; thus, a one-to-one MIMO channel is formed at each tone where its capacity is well-known [68]. However, the optimal allocation of power and rate in MIMO-OFDMA requires optimally assigning tones to each user, which is a non-convex combinatorial problem with the exponential complexity in the number of tones. If the number of tones goes to infinity, the infinite frequency dimensions allow arbitrary frequency sharing within the small bandwidth. Thus, the original problem effectively becomes convex, which results in zero duality gap [80]. With zero duality gap, it is possible to apply Lagrange dual decomposition in efficient solution of the optimization problem. However, with only tens or hundreds of tones in practical systems, this argument is inapplicable and zero duality gap may not be guaranteed, which complicates development of efficient optimal algorithms. Chapter 5 shows that in MIMO-OFDMA BCs and MACs with CSIT, the duality gap vanishes with only tens of tones for weighted sum-rate maximization (WSRmax) and weighted sum-power minimization (WSPmin) problems. From this observation, Lagrange dual decomposition is applied to develop efficient algorithms for optimal allocation of power/rate on each tone and at each transmit antenna. Using derived algorithms, the optimal achievable rate and power regions of MIMO-OFDMA BCs and MACs with CSIT are calculated with the polynomial complexity. Part of the work related to Chapter 5 is presented in [53, 12].

If the channel variation is fast over time, the instantaneous CSI sent to the transmitter via feedback channel becomes unreliable. In a fast-varying mobile environment, scheduling can be performed by utilizing long-term channel statistics. Chapter 6 addresses the cross-layer resource allocation in uplink and downlink MIMO-OFDMA systems with CDIT (Channel Distribution Information at the Transmitter). Under CDIT only, schedulers are unable to update transmission rates according to the instantaneous channel mutual information, which may result in *packet outages*. Hence, schedulers can select a rate tuple from *outage rate region*, which is defined as the set of maximum achievable rates while satisfying each user's specified outage probability constraint [40, 29, 41]. However, characterizing outage rate region of MIMO-OFDMA with CDIT involves very complicated numerical integration; thus, scheduling based on the exact outage rate region becomes intractable.

Chapter 6 shows that the mutual information of the MIMO-OFDMA BC and MAC can be well approximated using a Gaussian distribution. Also, reliable approximations for the mean and variance of the mutual information appear, which can be used to characterize the approximate Gaussian distribution. Based on the Gaussian approximation, a *successive feasibility check* (SFC) efficiently characterizes the entire outage rate region of MIMO-OFDMA BC and MAC with the linear complexity in the number of users and tones. Also, the power/rate allocation under QPS can be efficiently computed by directly applying this approach, i.e. the Gaussian approximation of mutual information in conjunction with a SFC. On the other hand, other gradient-type scheduling polices based on WSRmax, such as MWMS, exhibit the exponential complexity in the number of users for resource allocation. Chapter 6 presents stochastic simulations for the MIMO-OFDMA BC and MAC with CDIT that are performed by using the outage rate region characterized by the Gaussian approximation. The results corroborate superior throughput, delay, and fairness properties of QPS over other scheduling polices. In addition to these fundamental advantages, QPS provides high numerical efficiency that enables QPS to be a preferable scheduler for use in cross-layer resource allocation when only CDIT is available. Part of the work related to Chapter 6 is presented in [57].

Finally, Chapter 7 summarizes the key points in this thesis. Cross-layer approach to resource allocation is essential in multi-user communication systems with heterogeneous QoS requirement. Considering its superior throughput, delay and fairness properties as well as numerical efficiency, QPS is suitable for future wireless networks driven by various ubiquitous broadband services.

1.3 Notations and Abbreviations

In this thesis, bold face letters are used to denote vectors and matrices where matrices are always denoted by upper case letters. $\mathbb{R}^n$ denotes the set of real n-vectors and

$\mathbb{R}^n_+$ denotes the set of nonnegative real n-vectors. Given two column vectors $\mathbf{x}$ and $\mathbf{y}$ of length n, $\sum_{i=1}^n x_i y_i$ is expressed as an inner product $\mathbf{x} \cdot \mathbf{y}$. The curled inequality symbol $\succeq$ (and its strict form $\succ$) denotes the generalized inequality [10]. It denotes the component-wise inequality between vectors: $\mathbf{x} \succeq \mathbf{y}$ means $x_i \geq y_i$, $i = 1, 2, \cdots, n$. A column vector with all entries being 1 is denoted as $\mathbf{1}$; the length of $\mathbf{1}$ will be clear from context. $\mathbb{E}_x$ denotes expectation over the random variable x.

For a square matrix $\mathbf{S}$, $|\mathbf{S}|$, $\mathbf{S}^{-1}$ and $\mathbf{Tr}(\mathbf{S})$ denote its determinant, inverse matrix, and trace, respectively. For any general matrix $\mathbf{M}$, $\mathbf{M}^H$ is its conjugate transpose. $\mathbf{I}$ and $\mathbf{0}$ represent the identity matrix and the matrix with all zero elements, respectively. The Gaussian distribution of a vector with the mean vector $\mathbf{x}$ and the covariance matrix $\mathbf{\Sigma}$ is denoted by $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$. $\mathbb{C}^{x \times y}$ is the space of $x \times y$ matrices with complex entries. $\mathbb{R}^n$ denotes the set of real n-vectors and $\mathbb{R}^n_+$ denotes the set of nonnegative real n-vectors. For a square matrix $\mathbf{S}$, $\mathbf{S} \succeq \mathbf{0}$ means $\mathbf{S}$ is positive semidefinite. $\mathbf{1}\{\cdot\}$ is the indicator function, which takes the value of 1 if the argument is true, and zero otherwise.

The abbreviations used in this thesis are summarized in the following table:

Table 1.2: Summary of abbreviations

BC	Broadcast Channel
BCHPR	Best Channel Highest Possible Rate
BS	Base Station
CDI	Channel Distribution Information
CDIT	Channel Distribution Information at the Transmitter
CSI	Channel State Information
CSIT	Channel State Information at the Transmitter
CTMC	Continuous Time Markov Chain
DSL	Digital Subscriber Line
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FTP	File Transfer Protocol

GP	Geometric Programming
ISI	Inter-Symbol Interference
IWF	Iterative Water-Filling
LQHPR	Longest Queue Highest Possible Rate
LTE	Long Term Evolution
MAC	Multiple Access Channel
MACL	Medium Access Control Layer
MDT	Minimum Draining Time
MIMO	Multiple Input Multiple Output
MT	Mobile Terminal
MWMS	Maximum Weight Matching Scheduling
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PFS	Proportional Fair Scheduling
PHYL	Physical Layer
PRmax	Proportional Rate maximization
QoS	Quality of Service
QPS	Queue Proportional Scheduling
QSI	Queue State Information
SDMA	Spatial Division Multiple Access
SISO	Single Input Single Output
SNR	Signal-to-Noise Ratio
SRM	Sum Rate Maximization
SVD	Singular Value Decomposition
TDD	Time Division Duplex
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
VoIP	Voice over IP
WiMAX	Worldwide Interoperability for Microwave Access

WLAN	Wireless Local Area Network
WSP	Weighted Sum Power
WSPmin	Weighted Sum Power minimization
WSRmax	Weighted Sum Rate maximization
ZMCSCG	Zero Mean Circularly Symmetric Complex Gaussian
3GPP	The 3rd Generation Partnership Project

Chapter 2

Queueing Systems and Schedulers

Performance of scheduling operation can be characterized by the properties such as achievable throughput and queueing delay. With random packet arrivals, these properties are all related to the queueing behavior, which motivates full understanding of queueing systems. The first section of this chapter presents fundamentals of queueing theory. The most well-known queueing systems with Poisson packet arrivals and exponentially distributed service time, called $M/M/1$ queue, is introduced and its average queue length is mathematically derived. In addition, Little's law is explained, which is one of the most powerful and general queueing theories that relates average queue length to average queueing delay. Under some scheduling policy, a rate tuple is declared *achievable* if the scheduler can keep the queue backlog size finite. Therefore, in order to derive achievable throughput under certain schedulers, it is necessary to find the stability condition of queueing systems. In this regard, the first section introduces Lyapunov analysis and shows its application to finding the stability condition of $M/M/1$ queue.

The second section elaborates on the models of multi-user queueing system and scheduler that are used throughout this thesis. Also, the basic concepts required for characterizing schedulers, such as the *network capacity region* and *throughput optimality*, are defined in this section. The last section introduces a variety of well-known schedulers categorized into three types: channel-aware, queue-aware, queue-channel-aware schedulers.

2.1 Queueing Theory Basics

This section introduces an $M/M/1$ queue, Little's theorem for evaluating average queueing delay, and a Lyapunov analysis technique for proving stability of queueing systems [49].

2.1.1 M/M/1 Queue

A queueing system is often described by the notation of $A/S/s/k$. A stands for the arrival process such as Poisson, geometric, and deterministic, and S stands for the service distribution such as exponential, geometric, and deterministic. s denotes the number of servers and k stands for the buffer size where $k = \infty$ when k is absent. In addition, full characterization of the queueing system behavior requires description of the service discipline.

$M/M/1$ queue is a continuous-time queueing system that is widely applied in modeling queueing systems. In the $M/M/1$ queue, the arrivals occur according to a rate λ Poisson process where the number of arrivals within a time interval has a Poisson distribution.

$$P\{N(t_1, t_2) = k\} = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \tag{2.1}$$

where $N(t_1, t_2)$ denotes the number of arrivals in an interval (t_1, t_2) and $t = t_2 - t_1$. Let X_1 denote the time of the first arrival. Further, for $n \geq 1$, let X_n denote the time between the $(n-1)$st and the nth arrival. The sequence $\{X_n, n \geq 1\}$ is called the *sequence of interarrival times*. The event $\{X_1 > t\}$ takes place if and only if no arrivals occur in the interval $[0, t]$. Thus,

$$P\{X_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}. \tag{2.2}$$

Also, the distribution of X_2 conditioned on X_1 is

$$P\{X_2 > t | X_1 = s\} \quad = \quad P\{\text{No arrivals in } (s, s+t] | X_1 = s\}$$

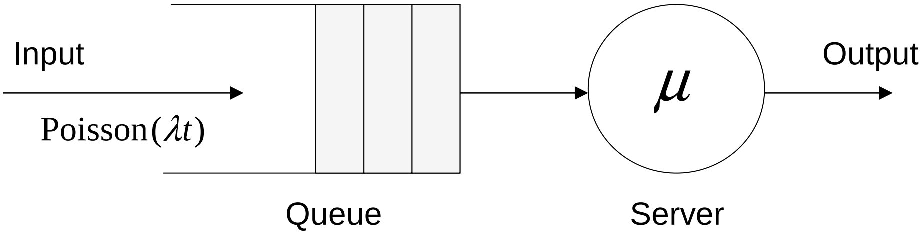


Figure 2.1: M/M/1 queue

$$
\begin{aligned}
&= P\{\text{No arrivals in } (s, s+t]\} \\
&= e^{-\lambda t}. \qquad (2.3)
\end{aligned}
$$

Therefore, from the above two equations, interarrival times are i.i.d. exponentially distributed with the mean of $1/\lambda$. Service is also based on the Poisson process in $M/M/1$ queue; thus, service times are i.i.d. rate μ exponentials, and independent of arrivals: $P(S > t) = e^{-\mu t}$. The M/M/1 queue model is illustrated in Fig. 2.1.

Let $Q(t)$ denote the number of information units in the queue at time t. Without loss of generality, $Q(t)$ is assumed to have nonnegative integer values. For $0 \le t_1 < t_2 \le t_3$, the queue-size process $Q(t)$ has the following Markovian property:

$$P\{Q(t_3) = k|Q(t_1) = i, Q(t_2) = j\} = P\{Q(t_3) = k|Q(t_2) = j\}. \qquad (2.4)$$

Thus, $Q(t)$ forms a continuous-time Markov chain (CTMC), where the past has no influence on the future if the present is specified [52]. The queueing behavior can be fully characterized once the transition rate from state i to j, which is the state transition probability divided by the average time in state i, is obtained.

$$q_{i,j} = \frac{p_{i,j}}{\mathbb{E}[T_i]}, \tag{2.5}$$

where $q_{i,j}$ is the transition rate from state i to j, $p_{i,j}$ denotes the state transition probability, and the average time in state i is denoted by $\mathbb{E}[T_i]$.

For CTMCs, $p_{i,i} = 0$; thus, $p_{0,1} = 1$. When $i \geq 1$, $p_{i,i+1}$ is equal to the probability that an arrival occurs before the service. Therefore, the following equations can be derived:

$$\begin{aligned} p_{i,i+1} &= \int_0^\infty P[A < t, S \in (t, t+dt)]\, dt = \int_0^\infty (1 - e^{-\lambda t})\mu e^{-\mu t}\, dt = \frac{\lambda}{\lambda + \mu}, \qquad (2.6) \\ p_{i,i-1} &= 1 - p_{i,i+1} = \frac{\mu}{\lambda + \mu}. \end{aligned}$$

For $i = 0$, the queue state changes only if there is an arrival. Thus, the average time in state 0 is given by

$$\mathbb{E}[T_0] = \mathbb{E}[\text{arrival time}] = \int_0^\infty P[A > t]\, dt = \frac{1}{\lambda}. \tag{2.7}$$

For $i \geq 1$, the state change is triggered by an arrival or a departure. Therefore, the average time in each state is

$$\begin{aligned} \mathbb{E}[T_i] &= \mathbb{E}[\min\{\text{arrival time, service time}\}] = \int_0^\infty P[\min\{A, S\} > t]\, dt \qquad (2.8) \\ &= \int_0^\infty P[A > t]P[S > t]\, dt = \int_0^\infty e^{-(\lambda+\mu)t}\, dt = \frac{1}{\lambda + \mu}. \end{aligned}$$

Finally, from (2.5)-(2.8), the state-transition rate is

$$\begin{aligned} q_{0,1} &= \lambda, \qquad (2.9) \\ q_{i,i+1} &= \lambda, \quad \text{for} \quad i \geq 1, \\ q_{i,i-1} &= \mu, \quad \text{for} \quad i \geq 1. \end{aligned}$$

Denote the equilibrium queue-length distribution by $\pi_i = \lim_{t\to+\infty} P(Q(t) = i)$, where $\sum_{i=0}^{\infty} \pi_i = 1$. Since the rate at which the process enters and leaves state j is equal, the following global-balance relations must be satisfied.

$$\pi_j \sum_{i=0}^{\infty} q_{j,i} = \sum_{i=0}^{\infty} \pi_i q_{i,j}. \tag{2.10}$$

Thus, $\pi_j \lambda = \pi_{j+1}\mu$ for $j = 0, 1, 2, \cdots$, which results in

$$\pi_j = \left(\frac{\lambda}{\mu}\right)^j \pi_0, \quad \text{for } j = 0, 1, 2, \cdots . \tag{2.11}$$

If the arrival rate is greater than the service rate, i.e. $\lambda > \mu$, then $\pi_j \to \infty$ and the queueing system becomes unstable. If $\lambda = \mu$, $\pi_j = \pi_0$ for every j. Since $0 \leq \pi_j \leq 1$ and $\sum_{j=0}^{\infty} \pi_j = 1$, there is no equilibrium and the queueing system is critically stable. On the other hand, when $\lambda < \mu$, $\pi_0 = (1 - \lambda/\mu)$ and $\pi_j = (1 - \lambda/\mu)(\lambda/\mu)^j$. Since π_j is exponentially decreasing, the queue backlog remains finite; thus, the queueing system is stable.

Define $\rho = \lambda/\mu$ as the traffic intensity. For $\rho < 1$, $\pi_j = (1 - \rho)\rho^j$; thus, the expected queue length,

$$\mathbb{E}[Q] = \Sigma_j j \pi_j = \frac{\rho}{(1 - \rho)}. \tag{2.12}$$

2.1.2 Little's Law

In queueing theory, Little's law states that the average queue size is equal to the average arrival rate multiplied by the average waiting time in the queueing system. This statement is quite general in that it is valid for any probability distributions on arrivals and services as long as the system operates in a first-come-first-served manner.

Suppose that a stable queue is empty at time 0. Denote the number of arrivals in $[0, t]$ by $A(t)$, and let $\lambda = \mathbb{E}[A(1)]$ be the arrival rate and D_i be the delay of the ith packet such that $D_i = W_i + S_i$ where W_i and S_i are waiting time and service time of the ith packet, respectively. Fig. 2.2 illustrates the progression of queue size with time t.

From Fig. 2.2, the following equation can be derived.

$$\frac{\sum_{t=0}^{T} Q(t)}{T} = \frac{\sum_{k=1}^{A(T)} D_k}{T} = \frac{A(T)}{T} \frac{\sum_{k=1}^{A(T)} D_k}{A(T)}. \tag{2.13}$$

By letting $T \rightarrow \infty$, the equation for Little's law is obtained,

$$\mathbb{E}[Q] = \lambda \mathbb{E}[D]. \tag{2.14}$$

For example, the average length of the $M/M/1$ queue is shown to be $\mathbb{E}[Q] = \rho/(1-\rho)$ in (2.12). Hence, by Little's theorem, the average packet delay in the $M/M/1$ queue becomes

$$\mathbb{E}[D] = \frac{\mathbb{E}[Q]}{\lambda} = \frac{1}{\mu(1-\rho)} = \frac{1}{(\mu-\lambda)}. \tag{2.15}$$

In addition, the average waiting time is given by

$$\mathbb{E}[W] = \mathbb{E}[D] - \mathbb{E}[S] = \frac{\lambda}{\mu(\mu-\lambda)}. \tag{2.16}$$

2.1.3 Stability and Lyapunov Analysis

This subsection addresses the stability of $M/M/1$ queue and presents its stability proof based on Lyapunov analysis. Let X_n denote a discrete-time Markov chain on the countable state space $\mathcal{S}$. Define a positive valued function $L : \mathcal{S} \rightarrow \mathbb{R}^+$, which is also called Lyapunov function. If $\lim_{n\rightarrow\infty} L(X_n)$ is finite with probability one, X_n is declared to have *weak stability.* On the other hand, *strong stability* of X_n is equivalent to that $\lim_{n\rightarrow\infty} \mathbb{E}[L(X_n)]$ is finite. Strong stability implies weak stability, but the reverse is not always true.

Assume that X_n is aperiodic and irreducible, and from any state, X_n can only transition to a finitely many states. Also, $L(X_{n+1})$ is assumed to be finite if $L(X_n)$

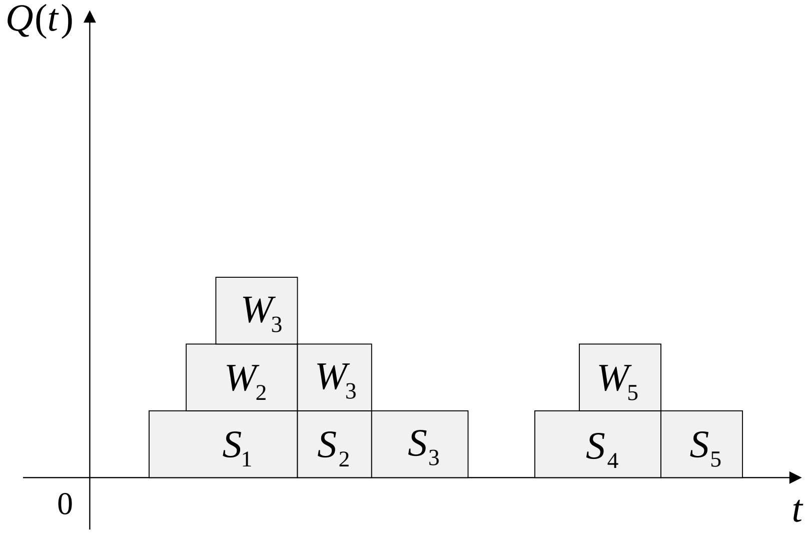


Figure 2.2: Progression of queue backlog size over time

is finite. Define the set $C = \{x \in \mathcal{S} : L(x) \leq B\}$ such that $L(x)$ is bounded in C. Then, the strong stability of X_n can be described by the following theorem.

Theorem 1. *If there exists a monotonically increasing positive function f such that $\mathbb{E}[L(X_{n+1}) - L(X_n)|L(X_n) \notin C] \leq -\epsilon f(L(X_n))$, $\limsup_{n\to\infty} \mathbb{E}[f(L(X_n))] < \infty$.*

Therefore, the stability of X_n is proved if a Lyapunov function L that satisfies the above condition can be found. Consider $M/M/1$ queue and apply Lyapunov analysis to prove its stability. Since the queue size is constant during the period when there is no packet arrival or departure, time can be discretized to when either an arrival or a potential service occurs. Then, the queue size is denoted by $Q(n)$ for $n = 0, 1, \cdots$, which is a Markov chain. With the arrival rate λ and service rate μ, the probability of $Q(n+1) = Q(n) + 1$ is $\lambda/(\lambda + \mu)$, and that of $Q(n+1) = Q(n) - 1$ when $Q(n) > 0$ is $\mu/(\lambda + \mu)$.

Assume the Lyapunov function $L(n)$ is equal to the queue size $Q(n)$. Then, for $Q(n) > 0$, the expected value of conditional discrete derivative of the Lyapunov function can be derived as follows.

$$\begin{aligned}\mathbb{E}[L(n+1)-L(n)|L(n)] &= \mathbb{E}[Q(n+1)-Q(n)|Q(n)] \qquad (2.17)\\ &= \mathbb{E}[Q(n+1)|Q(n)]-Q(n)\\ &= \mathbb{E}[\frac{\lambda}{\lambda+\mu}\left(Q(n)+1\right)+\frac{\mu}{\lambda+\mu}\left(Q(n)-1\right)|Q(n)]-Q(n)\\ &= \left(\frac{\lambda-\mu}{\lambda+\mu}\right)<0 \quad \text{if} \;\; \lambda<\mu.\end{aligned}$$

Hence, from Theorem 1, the $M/M/1$ queue is stable as long as the arrival rate is lower than the service rate. In general, multiple Lyapunov functions can be used to prove the stability of the same queueing system. If $Q^2(n)$ is chosen as the Lyapunov function, then for $Q(n)>0$,

$$\begin{aligned}&\mathbb{E}[L(n+1)-L(n)|L(n)]=\mathbb{E}[Q^2(n+1)-Q^2(n)|Q^2(n)] \qquad (2.18)\\ &=\mathbb{E}[\frac{\lambda}{\lambda+\mu}\left(Q(n)+1\right)^2+\frac{\mu}{\lambda+\mu}\left(Q(n)-1\right)^2-Q^2(n)|Q^2(n)]\\ &=Q^2(n)\left(\frac{\lambda}{\lambda+\mu}+\frac{\mu}{\lambda+\mu}-1\right)+2Q(n)\left(\frac{\lambda-\mu}{\lambda+\mu}\right)+1\\ &=2\sqrt{L(n)}\left(\frac{\lambda-\mu}{\lambda+\mu}\right)+1<0 \quad \text{if} \;\; \lambda<\mu \;\; \text{and} \;\; L(n)>\left(\frac{\lambda+\mu}{2(\mu-\lambda)}\right)^2.\end{aligned}$$

Thus, for $L(n)$ is large enough, $\mathbb{E}[L(n+1)-L(n)|L(n)]$ is negative if $\lambda<\mu$, which means the $M/M/1$ queue is stable.

2.2 Multi-User Queueing System and Scheduler Model

This section describes the models of the multi-user queueing system and the packet scheduler considered throughout this thesis. Also, the important concept of network capacity region and throughput optimality is introduced. Figure 2.3 presents the queueing system and scheduler model for the downlink. The transmitter has K output queues and the packets destined to receiver k enter queue k and wait until they are served. $\{A_k(t), k = 1, \cdots, K\}$ denotes the arrival process, and user k's average bit arrival rate is denoted by λ_k. The queue-state vector at time t is $\mathbf{Q}(t) = [Q_1(t)\ Q_2(t)\ \cdots\ Q_K(t)]^T$ where $Q_k(t)$ denotes the number of bits waiting to be sent to user k. In a very slow fading, instantaneous CSI for each downlink is obtainable at the transmitter by using a reliable CSI feedback link in FDD (Frequency Division Duplex) transmission mode or by utilizing the channel reciprocity in case of TDD transmission. However, if the channel variation is fast, the instantaneous CSIT becomes unreliable and only the long-term channel statistics may be available at transmitter. Based on both CSI and QSI, the scheduler determines the rate allocation on each user where user k's rate is denoted by $\mathbf{R}_k(t)$. In the uplink case, the queueing system is distributed over the users and QSI needs to be reported to the BS via a feedback link. Also, the BS can obtain uplink CSI from the channel estimation. Other than these changes, the scheduling operation is basically the same as that for the downlink.

In this thesis, the detailed assumptions on the above model are as follows: K output queues are assumed to have infinite capacity, and K data sources generate packets according to independent Poisson arrival processes $\{A_i(t), i = 1, \cdots, K\}$, which are stationary counting processes with $\lim_{t\to\infty} A_i(t)/t = a_i < \infty$, and $\mathrm{var}(A_i(t + T) - A_i(t)) < \infty$ for $T < \infty$. Packet lengths in bits $\{B_i\}$ are i.i.d. exponentially distributed with $\mathbb{E}[B_i] = \gamma_i < \infty$ and $\mathbb{E}[{B_i}^2] < \infty$. We assume packet lengths are independent of packet arrival processes; thus, user i's average arrival rate in bits is given by $\lambda_i = a_i\gamma_i$. Packets from source i enter queue i and wait until they are served to receiver i. The scheduling period is denoted by T_s. A time interval $[lT_s, (l + 1)T_s)$ where $l = 0, 1, 2, \cdots$ is denoted by the *time slot* l. At time t, the fading state is represented as $\mathbf{n}(t) = [n_1(t)\ n_2(t)\ \cdots\ n_K(t)]^T$, and the allocated rate vector at time t is represented as $\mathbf{R}(\mathbf{n}(t), \mathbf{Q}(t)) = [R_1(\mathbf{n}(t), \mathbf{Q}(t))\ \cdots\ R_K(\mathbf{n}(t), \mathbf{Q}(t))]^T$, which is determined by the scheduler based on both fading and queue states. For simplicity, $\mathbf{R}(t)$ and $\mathbf{R}\left(\mathbf{n}(t), \mathbf{Q}(t)\right)$ are interchangeably used.

This thesis considers a *quasi-static* fading channel where the channel condition remains stationary within a time slot, and it changes over time slots based on independent and identically distributed (i.i.d.) fading statistics. In addition, it is assumed

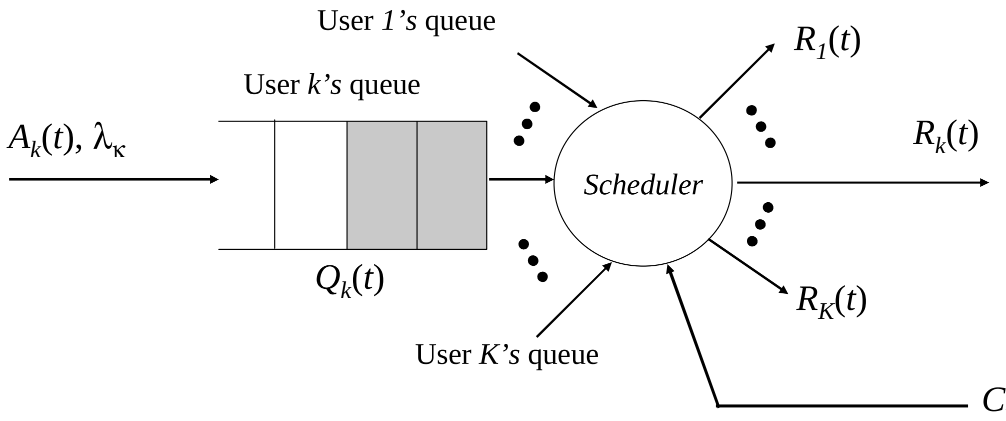


Figure 2.3: Queueing system and scheduler for the downlink

that the rate allocation is determined at the beginning of each time slot and remains unchanged until a new time slot begins. Thus, $T_s\mathbf{R}(lT_s)$ for $l = 0, 1, 2, \cdots$ is equivalent to a vector denoting the number of bits supported by each user over the time slot l. If $R_i(lT_s) > 0$, after the time slot l, a new packet is created for user i with the payload size of $T_sR_i(lT_s)$ and modulated for transmission. Define $Z_i(t)$ as the number of arrived bits at user i's queue over the time interval $[t, t+T_s)$. Then, after a scheduling period, user i's queue-state vector is equal to $Q_i(t+T_s) = \max\{Q_i(t) - T_sR_i(t), 0\} + Z_i(t)$. In this thesis, each scheduling policy has an explicit constraint of $T_s\mathbf{R}(t) \preceq \mathbf{Q}(t)$; thus, $\max\{\cdot, 0\}$ operation can be simply removed. Without loss of generality, $T_s = 1$ is assumed throughout this thesis. Thus, a time interval $[t, t+1)$ with $t = 0, 1, 2, \cdots$ is denoted by the time slot t, and $\mathbf{R}(t)$ for $t = 0, 1, 2, \cdots$ becomes a vector denoting the number of bits supported by each user in the time slot t.

The stability definition of queueing systems given in [45] is adopted in this thesis.

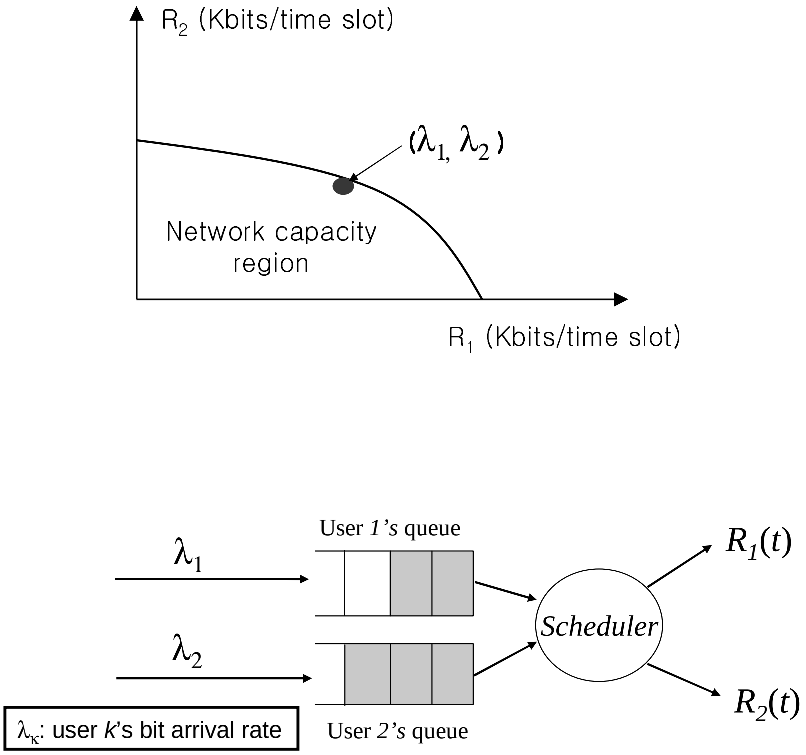


Figure 2.4: Network capacity region for two users

Thus, with the overflow function defined by $g(M) = \limsup_{t\to\infty} \frac{1}{t}\int_0^t 1[Q_i(\tau) > M]\, d\tau$, queue i is said to be stable if $g(M) \to 0$ as $M \to \infty$. An arrival rate vector $\boldsymbol{\lambda}$ is stabilizable if there exists a feasible power-and-rate-allocation policy that keeps all queues stable. A set of stabilizable arrival rate vectors forms the network capacity region [45], and a scheduling method that achieves the entire network capacity region is called *throughput optimal*.

Figure 2.4 illustrates the network capacity region for two user example. A bit arrival rate vector with the unit of bits per time slot, $[\lambda_1\ \lambda_2]^T$ is declared to be within the network capacity region if there exists some scheduler that is able to keep each queue's backlog size finite. Assuming that perfect CSIT is available and the total transmit power is constrained in the downlink*, the *instantaneous capacity region* is

*For the uplink, the individual power is constrained defined on each time slot. If an infinite-length Gaussian codeword is assumed, the entire capacity region can be achieved by applying capacity-achieving transmission and reception schemes with the optimal power and rate allocation [61, 20]. *Ergodic channel capacity region* with the sum-power constraint is defined as the instantaneous capacity region averaged over all the fading states [7, 39]. Thus, with the assumption of perfect CSIT and infinite-length Gaussian codewords, any arrival rate vectors within the ergodic capacity region are stabilizable. On the other hand, the queue backlog size eventually grows infinite for any arrival rate vectors outside the ergodic capacity region. Therefore, the network capacity region becomes equivalent to the ergodic capacity region for this case. In practical systems, the packet duration is limited to the finite time-slot duration, which results in the finite codeword length; thus, the deviation of achievable rate from the channel capacity is inevitable. This deviation from the capacity can be addressed by using the *gap* parameter that properly scales down the received signal-to-noise ration (SNR) in capacity equations [17]. With imperfect CSIT or no CSIT, the network capacity region becomes completely different from the ergodic capacity region.

2.3 Type of Packet Schedulers

This thesis divides schedulers into three types: channel-aware, queue-aware, and queue-channel-aware schedulers. In this section, a variety of well-known schedulers are introduced for each category.

2.3.1 Channel-Aware Scheduler

Channel-aware schedulers consider CSI in performing scheduling and tend to allocate higher data rate on the user with a better channel condition. From this opportunistic transmission, higher system throughput can be achieved by the multi-user diversity effects [37, 72]. QSI is only considered in such a way that the allocated rate is kept no more than the current queue backlog size. Thus, the user with a larger queue backlog size is not guaranteed to be assigned a higher data, which may result in the instability of queueing system even when the arrival rate vector is within the network capacity region. Three schedulers in this category are introduced in the following subsections.

Best Channel Highest Possible Rate Scheduling

Under the Best Channel Highest Possible Rate (BCHPR) scheduling policy, a user with the better channel condition takes higher priority in resource allocation. Also, user i is served only if some transmit power remains after clearing queue backlogs of users with higher priorities than user i. When the queue backlog size is large for every user, BCHPR operates as a type of TDM (Time Division Multiplexing) scheduler that allocates the full power to the user with the best condition. BCHPR is not a throughput-optimal scheduling policy since the queue size for the user with bad channel conditions can grow infinitely even for the arrival rate vector within the network capacity region. This algorithm is mathematically equivalent to allocating a data-rate vector that minimizes the l_1-norm distance from the current queue state vector. The l_1-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$. At time slot t, the BCHPR policy for the downlink supports the rate vector $\mathbf{R}_{BCHPR}(t)$ that is a solution of the following optimization problem.

$$\min \|\mathbf{Q}(t) - \mathbf{r}\|_1 \quad \text{subject to } \mathbf{r} \in C\left(\mathbf{h}(t), P\right). \tag{2.19}$$

where $C\left(\mathbf{h}(t), P\right)$ denotes the capacity region of a broadcast channel when the channel gain vector is $\mathbf{h}(t)$, and where the sum transmit power is constrained to P.

Sum-Rate Maximization Scheduling

The Sum-Rate Maximization (SRM) scheduling policy allocates a data rate vector such that the sum rate of each time slot is maximized. Under SRM, it is possible to support multiple users at the same time slot. SRM supports the rate vector $\mathbf{R}_{SRM}(t)$ at time slot t, which is a solution of the optimization problem given below.

$$\max \sum_{k=1}^{K} r_k \quad \text{subject to } \mathbf{r} \in C\left(\mathbf{h}(t), P\right) \text{ and } \mathbf{r} \preceq \mathbf{Q}(t). \tag{2.20}$$

Fig. 2.5 demonstrates the set of stabilizable arrival rate vectors under SRM for a two-user fading BC. By definition of SRM, the expected rate vector supported by SRM is a single boundary point of the network capacity region. Therefore, the arrival rate vectors are stabilizable only if they are inside the shaded region B in Fig. 2.5. For any arrival rate vectors outside the region B, which includes the region A within the network capacity region, the queueing system becomes unstable. This illustration shows the necessity of dynamic rate scheduling based on both CSI and QSI.

Proportional Fair Scheduling

In the practical scenario where each user has non-symmetric fading statistics, BCHPR and SRM are unable to guarantee or control the fairness among the users in terms of average throughput. For example, the users closer to the BS with a better average SNR may enjoy higher throughput than others, which is uncontrollable under the strategy. Therefore, for the case that the performance metric is defined as the average throughput over certain time horizon, BCHPR and SRM are unable to satisfy the condition.

Proportional Fair Scheduling (PFS) has been designed to meet the requirements on average throughput over the delay time scale in addition to utilizing multi-user diversity effects [72]. PFS converts SRM into the weighted sum-rate maximization problem where user k's weight at time slot t, $w_k(t)$, is defined as the inverse of user k's average throughput, $T_k(t)$, in a past window of length t_c. At time slot t, PFS allocates the rate vector $\mathbf{R}_{PFS}(t)$ which solves the next optimization problem.

$$\max \sum_{k=1}^{K} \frac{r_k}{T_k(t)} \quad \text{subject to } \mathbf{r} \in C\left(\mathbf{h}(t), P\right) \text{ and } \mathbf{r} \preceq \mathbf{Q}(t). \tag{2.21}$$

$T_k(t)$ can be updated by using an exponentially weighted low-pass filter

$$T_k(t+1) = \begin{cases} \left(1 - \frac{1}{t_c}\right) T_k(t) + \frac{1}{t_c} R_{k,PFS}(t), & \text{if user } k \text{ is served at time slot } t \\ \left(1 - \frac{1}{t_c}\right) T_k(t) & \text{otherwise.} \end{cases} \tag{2.22}$$

The parameter t_c is related to the latency time scale of the application. While still

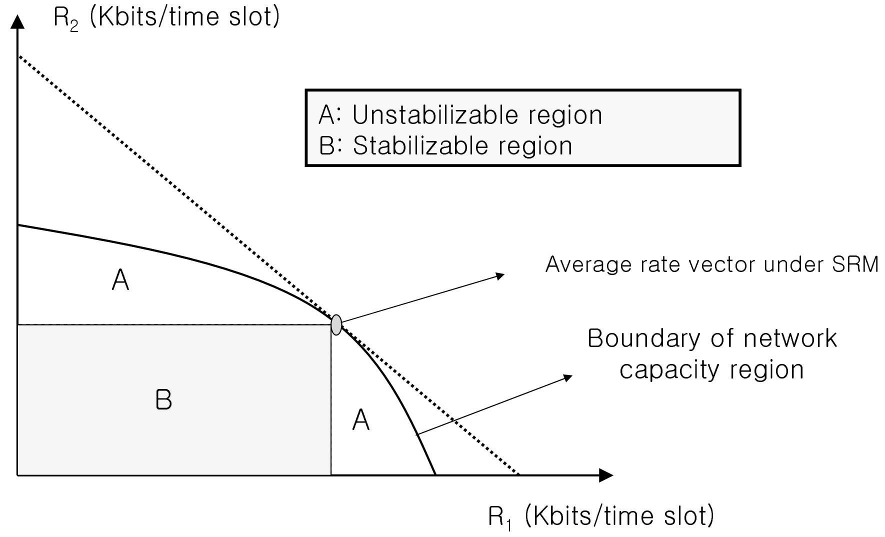


Figure 2.5: Stabilizable arrival rate vectors under SRM for a two-user fading BC

extracting the multi-user diversity benefit, PFS can also guarantee the proportional fairness of average throughput. Nonetheless, PFS cannot guarantee throughput optimality nor provide the controllability of queueing delay since QSI is not properly considered in this scheduling policy.

2.3.2 Queue-Aware Scheduler

Contrary to channel-aware scheduler, queue-aware scheduler mostly considers QSI such that the user with larger queue backlog is guaranteed to have higher rate allocation. The consideration of CSI is merely for deciding the rate amount for the selected users, and the better fading channel condition is never opportunistically exploited to achieve the multi-user diversity effects. One good example is the Longest Queue Highest Possible Rate (LQHPR) policy which schedules a data-rate vector such that the longest queue length is minimized. LQHPR scheduling is equivalent to selecting a rate vector minimizing the l_∞-norm distance from the current queue-state vector. The l_∞-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as $\|\mathbf{x}\|_\infty = \max\{|x_1|, \cdots, |x_n|\}$. Hence, at time slot t, the LQHPR policy assigns the rate vector $\mathbf{R}_{LQHPR}(t)$ that is a solution of the following optimization problem.

$$\min \|\mathbf{Q}(t) - \mathbf{r}\|_\infty \quad \text{subject to } \mathbf{r} \in C\left(\mathbf{h}(t), P\right). \tag{2.23}$$

Inherently, LQHPR tries to equalize every queue's backlog size, but the absence of multi-user diversity effects results in much smaller achievable rate region; thus, LQHPR is not a throughput optimal scheduling policy.

2.3.3 Queue-Channel-Aware Scheduler

Queue-channel-aware scheduler is the combination of channel-aware and queue-aware schedulers. By intelligently considering both CSI and QSI in scheduling, throughput optimality can be achieved as well as the individual queueing delay can be controlled to satisfy QoS requirement. This subsection introduces two well-known scheduling policies in this category: maximum weight matching scheduling and exponential rule.

Maximum Weight Matching Scheduling

Maximum weight matching scheduling (MWMS) maximizes the inner product of the queue-state vector and the achievable rate vector [67, 42]. At time slot t, MWMS assigns the following data-rate vector

$$\begin{aligned} \mathbf{R}_{MWMS}(t) &= \arg\max_{\mathbf{r}} \mathbf{Q}'(t)^T \mathbf{r} \\ &\text{such that } \mathbf{r} \in C\left(\mathbf{h}(t), P\right), \end{aligned} \tag{2.24}$$

where $\mathbf{Q}'(t) = [\beta_1 Q_1(t) \ \cdots \ \beta_K Q_K(t)]^T$. β_i is the user i's priority weight which is set to 1 for all users if everyone has the same priority. This algorithm tends to allocate higher data rate to the user with longer backlog or better channel conditions. By jointly considering queue and channel states, this MWMS policy is proved to achieve throughput optimality in the fading broadcast and multiple access channels [45, 5]. The proof of throughput optimality for MWMS is provided in Appendix by using Lyapunov stability analysis introduced in Section 2.1.

Recent applications of MWMS can be also found in OFDM downlink systems [63] and MIMO downlink systems [73], [65]. Sometimes, a scheduling policy is called *delay optimal* if it minimizes average queueing delay over all K users, which is defined as $\lim_{t\to\infty} \mathbb{E}[\frac{1}{K}\sum_{i=1}^{K} Q_i(t)]$ [77]. For the fading MAC, [77] shows that MWMS indeed minimizes the average queueing delay over all users if symmetric channels and equal packet arrival rates are assumed. This property is a consequence of the polymatroidal structure of the MAC capacity region [71]. However, there are no such structural properties in the fading BC capacity region so that even with symmetry assumptions, MWMS cannot guarantee the minimum average queueing delay.

Exponential Rule

Exponential (EXP) rule is another queue-channel-aware scheduling policy introduced in [2] whose throughput optimality is analytically proved in [60]. Both MWMS and EXP rule solve the weighted sum-rate maximization problem. While the weight vector is linearly proportional to the queue-state vector under MWMS, it is a exponential function of the queue-state vector under EXP rule. In [59], the simulation study shows that EXP rule provides better packet delays and guaranteed throughput compared to SRM or PFS.

At time slot t, EXP rule assigns the following data-rate vector [†]

$$\mathbf{R}_{EXP}(t) = \arg\max_{\mathbf{r}} \sum_{k=1}^{K} \gamma_i r_i \exp\left(\frac{a_i Q_i(t)}{\beta + [\overline{Q}(t)]^{\eta}}\right)$$
$$\text{such that } \mathbf{r} \in C\left(\mathbf{h}(t), P\right), \tag{2.25}$$

where $\overline{Q}(t) = (1/K)\Sigma_k a_i Q_i(t)$, and the positive constants $\gamma_1, \cdots, \gamma_K, a_1, \cdots, a_K$, β, and $\eta \in (0, 1)$ are fixed. With EXP rule, the right selection of these listed positive constants is crucial in the satisfaction of QoS requirement.

2.4 Summary

This chapter provides the fundamentals of queueing systems and schedulers that are prerequisite for understanding queueing behavior and scheduling performance. First, Little's theorem and Lyapunov analysis are introduced. They are widely used for calculating average queueing delay and for proving stability of the queueing system, respectively. Also, the models of multi-user queueing systems and schedulers are presented, and two important concepts required to understand scheduling performance, network capacity region and throughput optimality, are elaborated. Multi-user packet schedulers can be categorized into three types: channel-aware, queue-aware, and queue-channel-aware scheduling policies. This chapter introduces various well-known schedulers in each category.

[†] TDM constraints in original papers are relaxed in this formulation so that multiple users are allowed to simultaneously transmit.

Chapter 3

Queue Proportional Scheduling

This chapter presents another throughput-optimal scheduling policy called queue-proportional scheduling (QPS), which has more desirable delay and fairness properties than MWMS. Given the current queue state, QPS allocates a data-rate vector such that the expected rate vector averaged over all fading states is proportional to the current queue-state vector and is on the boundary of network capacity region. Reference [38] introduced the minimum draining time (MDT) policy, which was shown to be throughput optimal and shown to minimize the draining time of the queue backlogs in a fluid model with no further arrivals. Our work was performed independent of [38], and QPS has the properties of the MDT policy. We present another approach for proving the throughput optimality of QPS, which is different from [38]. Also, using the new proof, QPS is shown to have the capability of arbitrarily scaling the ratio of each user's average queueing delay. This fairness property of QPS is desirable for satisfying different *Quality of Service* (QoS) requirement of each user. The queueing delay for Poisson packet arrivals and exponentially distributed packet lengths is simulated under various scheduling policies. Numerical results corroborate the throughput optimality of QPS and indicate that QPS provides significantly smaller average queueing delay than MWMS. Moreover, it is observed that with the QPS policy, the fairness in terms of average queueing delay can be guaranteed for any arrival rate vectors.

3.1 Definition

In the stationary broadcast channels (BC), QPS assigns a maximum data-rate vector that is proportional to the current queue-state vector. Assuming that every user has equal priority and independent Gaussian noise with unit variance is added at each receiver, the formulation of QPS is given by[*]

$$\begin{aligned} \mathbf{R}_{QPS}(t) &= \mathbf{Q}(t)(\max x) \\ \text{subject to } \ \mathbf{Q}(t)x \in C(\mathbf{h}, P) &\ \text{ and } \ x \leq 1. \end{aligned} \tag{3.1}$$

where $C(\mathbf{h}, P)$ is the BC capacity region with the channel gain vector $\mathbf{h} = [h_1, \cdots, h_K]^T$ and total transmit power P. At time slot t, $\mathbf{R}_{QPS}(t)$ is the rate vector scheduled by QPS, and the queue-state vector in bits is denoted by $\mathbf{Q}(t) = [Q_1(t)\ Q_2(t)\ \cdots\ Q_K(t)]^T$. The application of QPS to Gaussian BC is addressed in [21, 55]. By using the degradedness of a Gaussian BC, the next chapter shows that (3.1) can be converted into geometric programming (GP), which is a special form of convex optimization problems with very efficient interior-point methods. In the stationary multiple-access channel (MAC), the formulation of QPS is the same as (3.1) except that the BC capacity region, $C(\mathbf{h}, P)$ is replaced by MAC capacity region denoted by $C(P_1, \cdots, P_K)$ where P_i is user i's maximum transmit power.

For the queue-state vector $\mathbf{Q}(t)$, Figure 3.1 illustrates two distinct rate vectors supported by MWMS and QPS. The two-user Gaussian BC is considered where user 1's average signal-to-noise ratio (SNR) is 19dB and user 2's average SNR is 13dB. Since both bandwidth and scheduling period are assumed 1, bps/Hz is equivalent to bits/slot. In other words, the rate region in Figure 3.1 shows how many bits can be supported in each time slot. The given queue-state vector satisfies $Q_2(t) = 0.5Q_1(t)$, which results in $R_{QPS}(t) = [4.1 \quad 2.05]^T$ and $R_{MWMS}(t) = [6.34 \quad 0]^T$. From Fig. 3.1, it can be anticipated that as the queue state changes, MWMS will exhibit more fluctuation in the supported rate vector compared to QPS. According to queueing

[*] Scheduling period and bandwidth are assumed to be 1. Thus, the capacity unit, bps/Hz becomes equivalent to bits per time slot. In general, $C(\mathbf{h}, P)$ needs to be defined as the scaled version of capacity region with the scaling factor $T_s W$.

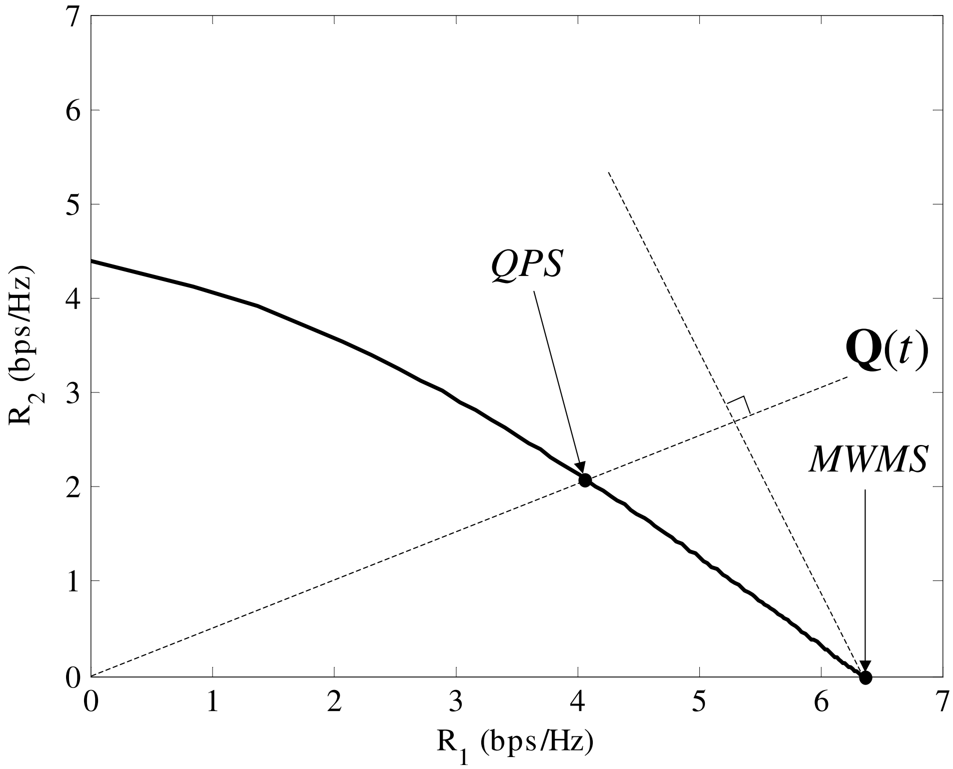


Figure 3.1: Capacity region of two user Gaussian BC, and rate vectors of QPS and MWMS when the queue-state vector is $\mathbf{Q}(t)$ (User 1's SNR=19dB and user 2's SNR=13dB).

theory, lower variance in service rate or arrival rate results in smaller queueing delay [4]. Therefore, QPS can be expected to have smaller average queueing delay than MWMS, upon which the next sections will elaborate.

On the other hand, in a time-varying BC, the QPS algorithm allocates the following data-rate vector at time slot t.

$$\begin{aligned} &\mathbf{R}_{QPS}(t) \in C\left(\mathbf{h}(t), P\right) \quad \text{such that} \\ &\mathbb{E}_{\mathbf{h}(t)}\left[\mathbf{R}_{QPS}(t)\right] = \mathbf{Q}'(t)\left(\max_{\mathbf{Q}'(t)x \in C_{erg}(P)} x\right). \end{aligned} \tag{3.2}$$

where x is a scalar. $\mathbf{Q}'(t) = [\beta_1 Q_1(t) \ \cdots \ \beta_K Q_K(t)]^T$. β_i is the user i's priority weight which is set to 1 for all users if everyone has the same priority. It is assumed

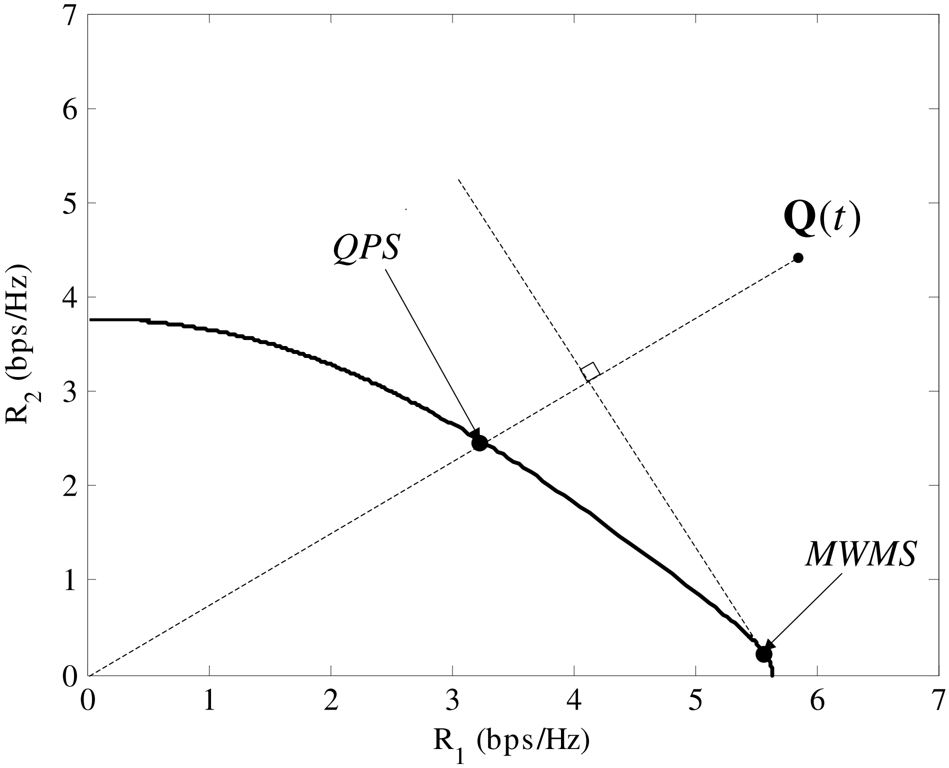


Figure 3.2: Ergodic capacity region of a two-user Rayleigh fading BC, and expected rate vectors of QPS and MWMS when the queue-state vector is $\mathbf{Q}(t)$ ($P = 2$, user 1's average SNR=13dB and user 2's average SNR=7dB).

in (4.13) that independent Gaussian noise with unit variance is added at each receiver. $C(\mathbf{h}(t), P)$ is the instantaneous BC capacity region at time t with the channel gain vector $\mathbf{h}(t) = [h_1(t), \cdots, h_K(t)]^T$ and total transmit power P. Also, $C_{erg}(P)$ denotes the ergodic BC capacity region with the total transmit power constrained to P. Chapter 4 details the analytic expressions for $C(\mathbf{h}(t), P)$ and $C_{erg}(P)$.

Assuming equal priority on each user, $\mathbf{Q}'(t) = \mathbf{Q}(t)$. Then, the average rate vector under the QPS policy, $\mathbb{E}_{\mathbf{h}(t)}[\mathbf{R}_{QPS}(t)]$ is proportional to the queue-state vector and also lies on the boundary surface of the ergodic capacity region. As shown in [70], each boundary point of $C_{erg}(P)$ in a fading BC is a solution to the optimization problem $\max_{\mathbf{r}} \boldsymbol{\mu} \cdot \mathbf{r}$ where $\mathbf{r} \in C_{erg}(P)$ for some $\boldsymbol{\mu} \in \mathbb{R}_+^K$. When such $\boldsymbol{\mu}$ is given, $\mathbf{R}_{QPS}(t)$ is a solution to the optimization problem $\max_{\mathbf{r}} \boldsymbol{\mu} \cdot \mathbf{r}$ where $\mathbf{r} \in C\left(\mathbf{h}(t), P\right)$ for any fading state $\mathbf{h}(t)$. Therefore, the data-rate vector assigned by QPS at time slot t can be expressed as

$$\begin{aligned} \mathbf{R}_{QPS}(t) &= \arg\max_{\mathbf{r}} \boldsymbol{\mu}^T \mathbf{r} \\ \text{such that } \mathbf{r} &\in C\left(\mathbf{h}(t), P\right). \end{aligned} \tag{3.3}$$

Under the QPS policy, $\boldsymbol{\mu}$ is determined based on the current queue-state vector as well as the ergodic capacity region of a fading BC. On the other hand, as shown in (2.24), MWMS only considers the queue-state vector in deriving the weight vector.

Figure 3.2 illustrates two distinct expected rate vectors supported by MWMS and QPS for the queue-state vector $\mathbf{Q}(t)$. A two-user Rayleigh fading BC is considered where $P = 2$, user 1's average signal-to-noise ratio (SNR) is 13dB and user 2's average SNR is 7dB. Each user's average SNR is defined as the average received SNR when the total transmit power is allocated to that user. Since $W = T_s = 1$ is assumed, bps/Hz is equivalent to bits/scheduling period. Thus, the ergodic capacity region in Fig. 3.2 represents the set of vectors denoting each user's expected number of bits served in one scheduling period. Also, note that with $W = T_s = 1$, the network capacity region is the same as the ergodic capacity region. From Fig. 3.2, as the queue state changes, QPS is expected to exhibit smaller variations in the average rate vector compared to MWMS, which may result in a smaller average queueing delay as demonstrated in Section 3.3.

3.2 QPS Properties

This section proves QPS to be throughput optimal in a fading BC[†]. Further, this proof extends to show that QPS has the capability of arbitrarily scaling the ratio of each user's average queueing delay. This unique scaling/fairness property of QPS is desirable for satisfying different QoS requirement of each user.

[†] Every proof in this section is directly applicable to any time-varying BC or MAC with the convex achievable rate region.

3.2.1 Throughput Optimality of QPS

The next theorem shows the convergence property of the expected queue-state vector under QPS, which is crucial in showing throughput optimality and fairness properties.

Theorem 2. *Under the QPS policy in a fading BC, as $t \to \infty$, the expected queue-state vector conditioned on any initial queue state, converges to a vector proportional to the arrival rate vector.*

Proof. Let $\mathbf{q_0} \in \mathbb{R}_+^K$ be the initial queue-state vector, and denote the bit arrival rate vector by $\boldsymbol{\lambda} = [\lambda_1 \ \cdots \ \lambda_K]^T$ where $\lambda_1 > 0$. Consider time slot t when some queues have backlogs, and let $\mathbf{Q}(t)$ be equal to $\mathbf{q_t} = [q_{t,1} \ q_{t,2} \ \cdots \ q_{t,K}]^T$. Without loss of generality, assume $q_{t,1} > 0$. Then, $\mathbf{q_t}$ can be represented as $\mathbf{q_t} = w(t)[\lambda_1, \ \lambda_2 + \Delta\lambda_2, \ \cdots, \ \lambda_K + \Delta\lambda_K]^T$ where $w(t) = q_{t,1}/\lambda_1$ and $\Delta\lambda_i \in \mathbb{R}$ such that $w(t)(\lambda_i + \Delta\lambda_i) = q_{t,i}$ for $i = 2, \cdots, K$. The expectation of $\mathbf{Q}(t+1)$ conditioned on $\mathbf{Q}(t) = \mathbf{q_t}$ becomes

$$\mathbb{E}\left[\mathbf{Q}(t+1)|\, \mathbf{Q}(t) = \mathbf{q_t}\right] = \mathbf{q_t} + \boldsymbol{\lambda} - \mathbb{E}\left[\mathbf{R}_{QPS}(t)|\, \mathbf{Q}(t) = \mathbf{q_t}\right]. \tag{3.4}$$

Under QPS, $\mathbb{E}\left[\mathbf{R}_{QPS}(t)|\, \mathbf{Q}(t) = \mathbf{q_t}\right] = r(t)\left(\mathbf{q_t}/w(t)\right)$ where $r(t)$ equals $\max x$ subject to $x\left(\mathbf{q_t}/w(t)\right) \in C_{erg}(P)$. (3.4) can be converted into the following form.

$$\begin{aligned}\mathbb{E}\left[\mathbf{Q}(t+1)|\, \mathbf{Q}(t) = \mathbf{q_t}\right] = (w(t) - r(t) + 1) \times \\ [\lambda_1, \ \lambda_2 + \gamma(t)\Delta\lambda_2, \ \cdots, \lambda_K + \gamma(t)\Delta\lambda_K]^T,\end{aligned} \tag{3.5}$$

where $\gamma(t) = 1 - 1/(w(t) - r(t) + 1)$. If $\mathbf{q_t} \in C_{erg}(P)$, then $w(t) = r(t)$; hence, $\gamma(t) = 0$ and $\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(t) = \mathbf{q_t}] = \boldsymbol{\lambda}$. Otherwise, $w(t) > r(t)$ and $\gamma(t)$ is strictly less than 1. Let the angle between $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ and $\mathbf{x} \in \mathbb{R}_+^K$ be denoted by $\theta_{\boldsymbol{\lambda}}(\mathbf{x})$ that is

$$\theta_{\boldsymbol{\lambda}}(\mathbf{x}) = \cos^{-1}\left(\frac{\boldsymbol{\lambda}^T\mathbf{x}}{\|\boldsymbol{\lambda}\|_2\|\mathbf{x}\|_2}\right), \quad 0 \le \theta_{\boldsymbol{\lambda}}(\mathbf{x}) \le \frac{\pi}{2}. \tag{3.6}$$

Since $\gamma(t) < 1$, $\theta_{\boldsymbol{\lambda}}(\mathbf{q_t}) \ge \theta_{\boldsymbol{\lambda}}(\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(t) = \mathbf{q_t}])$. This chapter assumes i.i.d. block fading and Poisson packet arrivals. Therefore, each user's queue state is the

1st-order Markov process, which allows the following relation to hold from Chapman-Kolmogorov equations [52].

$$\begin{gathered}\mathbb{E}\left[\mathbf{Q}(t+1)\right|\mathbf{Q}(0)=\mathbf{q_0}]=\\ \mathbb{E}\left[\mathbb{E}\left[\mathbf{Q}(t+1)\right|\mathbf{Q}(t)]\left|\mathbf{Q}(0)=\mathbf{q_0}\right]\quad\text{for}\quad t=1,2,\cdots.\end{gathered}\tag{3.7}$$

Since $\theta_{\boldsymbol{\lambda}}(\mathbf{Q}(t)) \geq \theta_{\boldsymbol{\lambda}}(\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(t)])$, the right-hand side (RHS) of (3.7) has a direction closer to $\boldsymbol{\lambda}$ than $\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0)=\mathbf{q_0}]$. Consequently, the following relation is obtained.

$$\begin{aligned}\theta_{\boldsymbol{\lambda}}\left(\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0)=\mathbf{q_0}]\right) \geq \theta_{\boldsymbol{\lambda}}\left(\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(0)=\mathbf{q_0}]\right)\\ \text{for}\quad t=1,2,\cdots.\end{aligned}\tag{3.8}$$

Define an infinite sequence $\theta_t = \theta_{\boldsymbol{\lambda}}(\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0)=\mathbf{q_0}])$ for $t=1,2,\cdots$. Since θ_t is monotonically decreasing and nonnegative, θ_t is a converging sequence. In the RHS of (3.7), $\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(t)] = \mathbf{Q}(t)+\boldsymbol{\lambda}-\mathbb{E}[\mathbf{R}_{QPS}(t)|\mathbf{Q}(t)] = (1-c)\mathbf{Q}(t)+\boldsymbol{\lambda}$ where $c=\max r$ such that $r\mathbf{Q}(t)\in C_{erg}(P)$. Therefore, (3.7) can be expressed as

$$\begin{gathered}\mathbb{E}\left[\mathbf{Q}(t+1)\right|\mathbf{Q}(0)=\mathbf{q_0}]=\\ (1-c)\mathbb{E}\left[\mathbf{Q}(t)\left|\mathbf{Q}(0)=\mathbf{q_0}\right]+\boldsymbol{\lambda},\quad\text{for}\quad t=1,2,\cdots.\end{gathered}\tag{3.9}$$

By the convergence property, as $t\to\infty$, the angle between $\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0)=\mathbf{q_0}]$ and $\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(0)=\mathbf{q_0}]$ becomes zero. Therefore, to satisfy the equality in (3.9) when $t\to\infty$, the direction of these two vectors should converge to that of $\boldsymbol{\lambda}$. As a result, it can be concluded that $\lim_{t\to\infty}\theta_t=0$, which completes the proof of the theorem. □

Based on Theorem 2, the throughput optimality of QPS can be proved by using Lyapunov stability analysis technique introduced in Chapter 2.1.3.

Theorem 3. *In a fading BC, the QPS policy is throughput optimal.*

Proof. With $W=T_s=1$, the network capacity region is equivalent to $C_{erg}(P)$. Thus, we need to show that for any $\boldsymbol{\lambda}\in\text{int }C_{erg}(P)$ where int $\mathcal{S}$ denotes the interior of a set $\mathcal{S}$, the queue lengths for all users can be kept finite. First, choose the Lyapunov function $L\left(\mathbf{Q}(t)\right) = \sum_{i=1}^{K} Q_i(t)$. The evolution of $L\left(\mathbf{Q}(t)\right)$ in one scheduling interval is $L\left(\mathbf{Q}(t+1)\right) = \sum_{i=1}^{K} Q_i(t+1) = \sum_{i=1}^{K} \left(Q_i(t) + Z_i(t) - R_i(t)\right)$. Conditioned on $\mathbf{Q}(t) = \mathbf{q_t}$, the expected drift of the Lyapunov function is

$$\mathbb{E}\left[L\left(\mathbf{Q}(t+1)\right) - L\left(\mathbf{Q}(t)\right) \middle| \mathbf{Q}(t) = \mathbf{q_t}\right] = \sum_{i=1}^{K} \left(\lambda_i - \mathbb{E}\left[R_i(t) \middle| \mathbf{Q}(t) = \mathbf{q_t}\right]\right). \quad (3.10)$$

To prove the throughput optimality of QPS, it is required to show that as queue lengths grow sufficiently large, (3.10) becomes strictly negative for any $\boldsymbol{\lambda} \in \text{int } C_{erg}(P)$ [74]. By Theorem 2, in the stationary regime, $\mathbb{E}\left[\mathbf{Q}(t)\right] = w(t)\boldsymbol{\lambda}$ for some $w(t) \geq 0$. Thus, $\mathbf{Q}(t)$ can be represented as $\mathbf{Q}(t) = \mathbb{E}\left[\mathbf{Q}(t)\right] + \mathbf{N}(t) = w(t)\boldsymbol{\lambda} + \mathbf{N}(t)$ where $\mathbf{N}(t) = [N_1(t) \ \cdots \ N_K(t)]^T$ and $\mathbb{E}\left[N_i(t)\right] = 0$ for $i = 1, \cdots, K$. As $w(t)$ increases, $\mathbf{Q}(t) = w(t)(\boldsymbol{\lambda} + \mathbf{N}(t)/w(t)) \rightarrow w(t)\boldsymbol{\lambda}$ with probability 1, which results in $\mathbb{E}\left[\mathbf{R}_{QPS}(t) \middle| \mathbf{Q}(t) = \mathbf{q_t}\right] \rightarrow \mathbb{E}\left[\mathbf{R}_{QPS}(t) \middle| \mathbf{Q}(t) = w(t)\boldsymbol{\lambda}\right]$ with probability 1.

$\mathbb{E}[\mathbf{R}_{QPS}(t)|\mathbf{Q}(t) = w(t)\boldsymbol{\lambda}] = \boldsymbol{\lambda}(\max r)$ such that $\boldsymbol{\lambda} r \in C_{erg}(P)$. If $\boldsymbol{\lambda} \in \text{int } C_{erg}(P)$, then $\max r > 1$. Thus, when $\|\mathbf{q_t}\|_\infty$ grows sufficiently large, the Lyapunov drift in (3.10) becomes strictly negative for any $\boldsymbol{\lambda} \in \text{int } C_{erg}(P)$. □

3.2.2 Fairness and Delay Properties of QPS

This subsection shows that for any set of arrival rates, QPS can arbitrarily scale the ratio of each user's average queueing delay. Also, it is shown that without new packet arrivals, QPS minimizes the expected time to empty all the backlogs. First, the next theorem shows that QPS has a capability of guaranteeing fairness among users in terms of average queueing delay.

Theorem 4. *In a fading BC under the QPS policy, as $t \rightarrow \infty$, each user's average queueing delay becomes equalized.*

Proof. From Theorem 2, the average queue-state vector becomes proportional to the arrival rate vector as $t \rightarrow \infty$. By Little's theorem introduced in Chapter 2.1.2, the average queue length is the same as a product of the arrival rate and average queueing delay [6]. Therefore, with QPS policy, each user's average queueing delay is equalized after the convergence. □

In general, QPS can satisfy a different QoS for each user in terms of average queueing delay. This property is shown in the following corollary to Theorem 4.

Corollary 1. *Let $\boldsymbol{\beta}$ denote the priority vector on average queueing delay. For example, $\beta_1 = 2\beta_2$ means that the average delay of user 1 should be half of user 2's average delay. This priority can be satisfied with the QPS policy by replacing $\mathbf{Q}(t)$ with the modified queue-state vector $\mathbf{Q}'(t) = [\beta_1 Q_1(t)\ \beta_2 Q_2(t)\ \cdots\ \beta_K Q_K(t)]^T$.*

Proof. From Theorem 2, the average of a modified queue state vector $\mathbf{Q}'(t)$ converges to $\boldsymbol{\lambda} x$ for some $x \in \mathbb{R}_+$. Thus, user i's average queue length converges to $(\lambda_i x)/\beta_i$, and by Little's theorem, user i's average queueing delay becomes x/β_i. □

One reasonable way of choosing the priority vector $\boldsymbol{\beta}$ is to find a vector proportional to each user's maximum achievable average rate when no other users transmit. The next theorem shows that without new packet arrivals, QPS minimizes the expected time to empty all the queue backlogs.

Theorem 5. *Let the initial queue-state vector be $\mathbf{Q}(0) = \mathbf{q_0} \in \mathbb{R}_+^K$, and assume that there are no more packet arrivals after $t = 0$. Then, in a fading BC, the QPS policy presuming the constant queue-state vector of $\mathbf{q_0}$ for all $t \geq 0$ minimizes the expected time until all the queue backlogs are cleared.*

Proof. Let $\mathbb{E}[T_X]$ denote the expected time until a scheduling algorithm X empties all the queue backlogs $\mathbf{q_0}$. The total supported data vector is $\mathbf{q_0}$. Thus, given $\mathbb{E}[T_X]$, the average data vector allocated per each scheduling period can be expressed as $\mathbb{E}[\mathbf{R}_X] = \frac{\mathbf{q_0}}{\mathbb{E}[T_X]}$. Since $C_{erg}(P)$ is convex, $\mathbb{E}[\mathbf{R}_X] \in C_{erg}(P)$ is always satisfied. Therefore, $\mathbb{E}[T_X]$ is minimized by assigning $\mathbf{R}_{opt}(\mathbf{h}(t), \mathbf{Q}(t)) \in C(\mathbf{h}(t), P)$ at time slot t such that

$$\mathbb{E}_{\mathbf{h}(t)}\left[\mathbf{R}_{opt}(\mathbf{h}(t), \mathbf{Q}(t))\right] = \mathbf{q_0}\left(\max_{\mathbf{q_0} r \in C_{erg}(P)} r\right). \tag{3.11}$$

From the definition of QPS, it can be easily seen that $\mathbf{R}_{opt}\left(\mathbf{h}(t), \mathbf{Q}(t)\right)$ is equal to $\mathbf{R}_{QPS}\left(\mathbf{h}(t), \mathbf{q_0}\right)$, which completes the proof of the theorem. □

In actual systems with random packet arrivals, the property in Theorem 5 can be approximated by replacing $\mathbf{q_0}$ with the current queue-state vector $\mathbf{Q}(t)$. Therefore, at each scheduling time, QPS tries to minimize the expected time to empty current queue backlogs. This property of QPS results in low average queueing delay, which will be demonstrated to be much smaller than MWMS in the next section.

3.3 Numerical Results

This section presents simulation results with Poisson packet arrivals and exponentially distributed packet lengths for both the stationary Gaussian BC and the fading Gaussian BC, in order to demonstrate stability, delay, and fairness properties of the QPS algorithm. In the simulation, average packet length for each user, scheduling period, and signal bandwidth are all equal to 1. Also, the average queue length over all users is defined as $\lim_{t\to\infty} \mathbb{E}[\frac{1}{K}\sum_{i=1}^{K} Q_i(t)]$. Stochastic simulation results in Gaussian BC are presented in Fig. 3.3-3.5. In these simulations, noise power is assumed to be 0.1. In Fig. 3.3 and Fig. 3.4, the average queue length is evaluated for different values of λ_1 with two users and ten users, respectively. Four scheduling algorithms are compared in both figures: QPS, MWMS, Longest Queue Highest Possible Rate (LQHPR) and Best Channel Highest Possible Rate (BCHPR). As explained in Chapter 2.3, LQHPR allocates a data-rate vector such that the longest queue length is minimized. Under the BCHPR policy, a user with the better channel condition takes higher priority in resource allocation, user i is served only if some transmit power remains after clearing queue backlogs of users with higher priorities than user i.

For the two user case in Fig. 3.3, a Gaussian BC channel presented in Fig. 3.1 is considered where the power constraint $P = 2$ and the channel gain vector is $\mathbf{h} = [2\ 1]^T$; thus, user 1's SNR=19dB, and user 2's SNR=13dB. Also, the bit-arrival rate vector satisfies $\boldsymbol{\lambda} = \lambda_1[1\ 0.5]^T$. From Fig. 3.1, $\boldsymbol{\lambda} \in \text{int}\, C(\mathbf{h}, P)$ if and only if $\lambda_1 < 4.1$. Fig. 3.3 demonstrates that the average queue length of QPS is about 30% smaller than

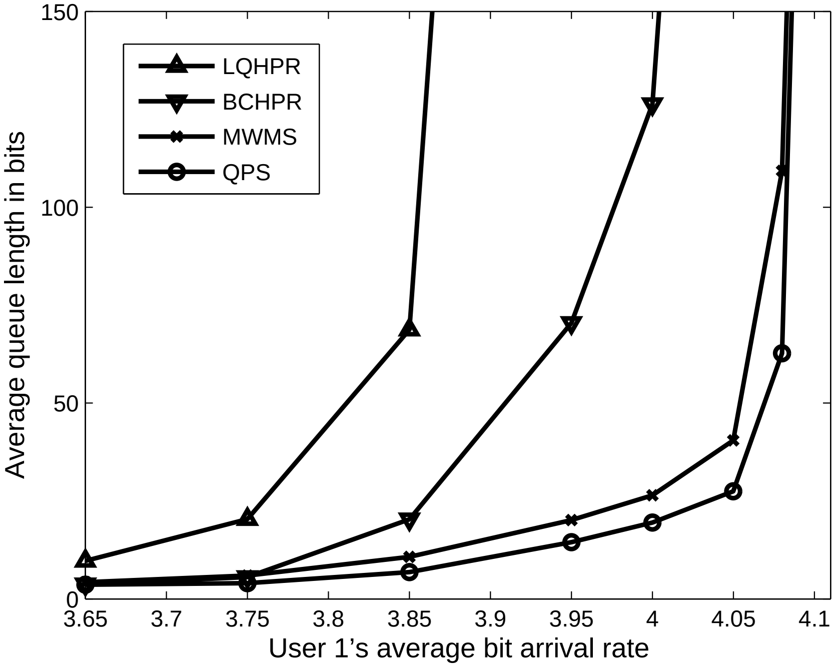


Figure 3.3: Average queue length vs user 1's bit arrival rate under LQHPR, BCHPR, MWMS and QPS (2 users, user 1's SNR=13dB and user 2's SNR=7dB, $\lambda_2 = 0.5\lambda_1$).

that of MWMS for any $\lambda_1 < 4.1$. Since MWMS is a throughput optimal policy, this observation corroborates the throughput optimality of QPS. LQHPR and BCHPR, which are not throughput optimal, have much longer average queue lengths than MWMS. Simulation results with 10 users are shown in Fig. 3.4. The total transmit power is $P = 10$ and user i's channel gain $h_i = 2 - 0.1(i-1)$ and $\lambda_i = \lambda_1(0.9)^{i-1}$ for $i = 1, \cdots, 10$. QPS provides about 40-50% smaller average queue length than MWMS, which is a more prominent difference than in the two user case. BCHPR is also observed to have around 20% smaller average queue length than MWMS at small λ_1. However, as λ_1 approaches to the boundary of a network capacity region, the average queue length of BCHPR grows faster than MWMS.

The fairness properties of QPS, MWMS and BCHPR with 10 users are illustrated in Fig. 3.5. The simulation environment is identical with Fig. 3.4 and $\lambda_1 = 1.32$ is considered. Fig. 3.5 shows the arrival rate vector as well as each user's average

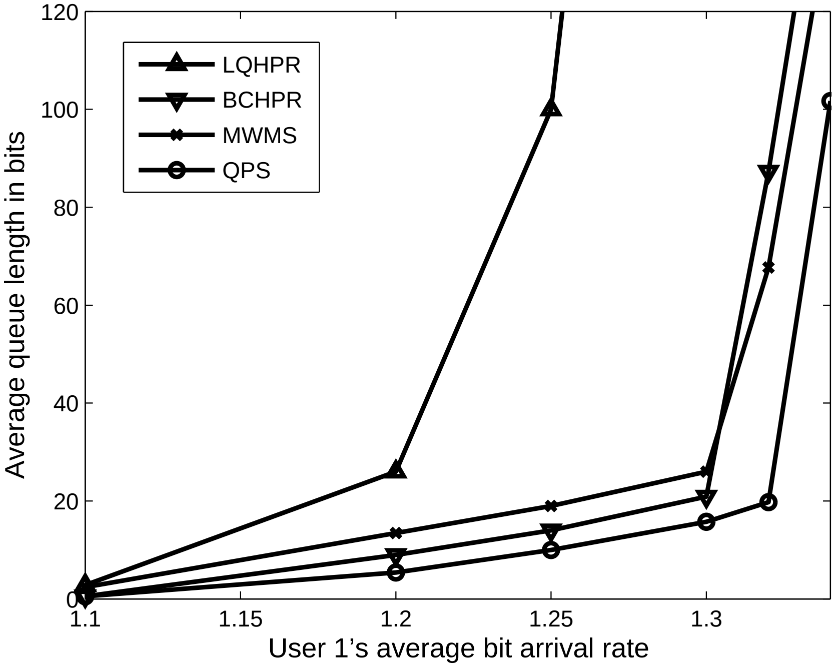


Figure 3.4: Average queue length vs user 1's bit arrival rate under LQHPR, BCHPR, MWMS and QPS (10 users, user i's channel gain $h_i = 2-0.1(i-1)$ and $\lambda_i = \lambda_1(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

queueing delay in slots for above three scheduling policies. It is observed that fairness among users is not satisfied under the BCHPR, which results in intolerably long average queue length for users with worse channel conditions. In this simulation result, MWMS tends to equalize each user's average queue length. Since each user has a different arrival rate, by Little's theorem, MWMS shows smaller average queueing delay for the user with higher arrival rate. On the other hand, the average queue length of QPS is shown to be almost proportional to the arrival rate vector so that each user's average queueing delay is equalized. Therefore, under the QPS policy, fairness among users is guaranteed in terms of average queueing delay.

Fig. 3.6-3.8 provides the simulation results in the fading BC. For the two user case in Fig. 3.6, the Rayleigh fading BC channel presented in Fig. 3.2 is considered where the total power constraint $P = 2$, user 1's average SNR=13dB, and user 2's

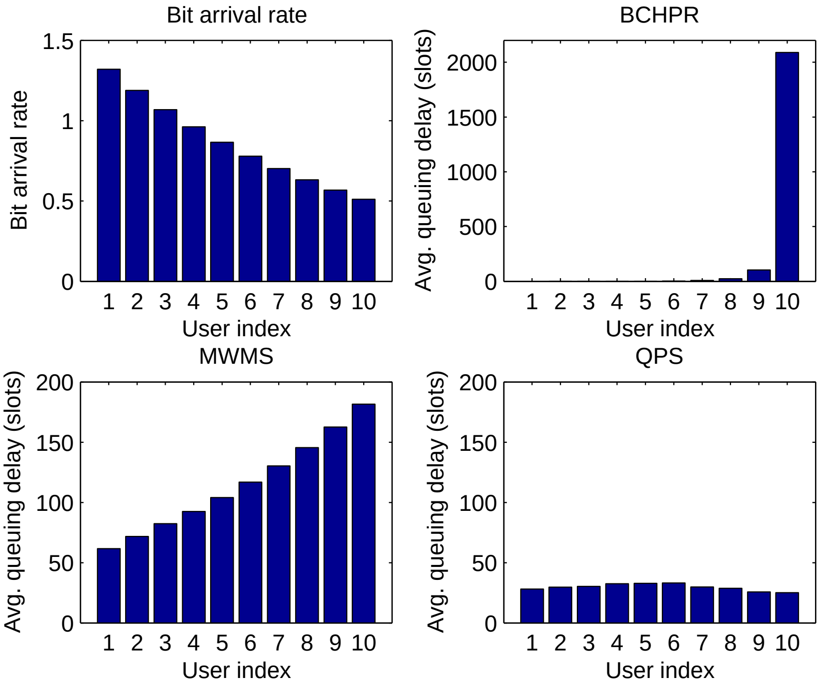


Figure 3.5: Each user's bit arrival rate and each user's average queueing delay under BCHPR, MWMS and QPS (10 users, user i's channel gain $h_i = 2 - 0.1(i-1)$ and $\lambda_i = 1.32(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

average SNR=7dB. Also, the bit arrival rate of user 2 is assumed to be the half of user 1's. Thus, the bit arrival rate vector can be represented as $\boldsymbol{\lambda} = \lambda_1[1\ 0.5]^T$. From Fig. 3.2, $\boldsymbol{\lambda} \in \text{int } C_{erg}(P)$ if and only if $\lambda_1 < 3.9$.

In Fig. 3.6 and Fig. 3.7, average queue lengths are evaluated for different values of λ_1 when $K = 2$ and $K = 10$, respectively. In both figures, QPS, MWMS, BCHPR, and LQHPR are compared. Fig. 3.6 shows that the average queue length under QPS is about 30% smaller than that of MWMS for any $\lambda_1 < 3.9$. Since MWMS is a throughput optimal policy, this observation corroborates the throughput optimality of QPS in a fading BC. LQHPR and BCHPR, which are not throughput optimal, have about 12% and 5% throughput loss, respectively. Simulation results with 10 users are presented in Fig. 3.7. $P = 10$ and user i's average SNR is equal to $20-(i-1)$ (dB) for $i = 1, \cdots, 10$. Also, the bit arrival rate is identical for all users. QPS provides about

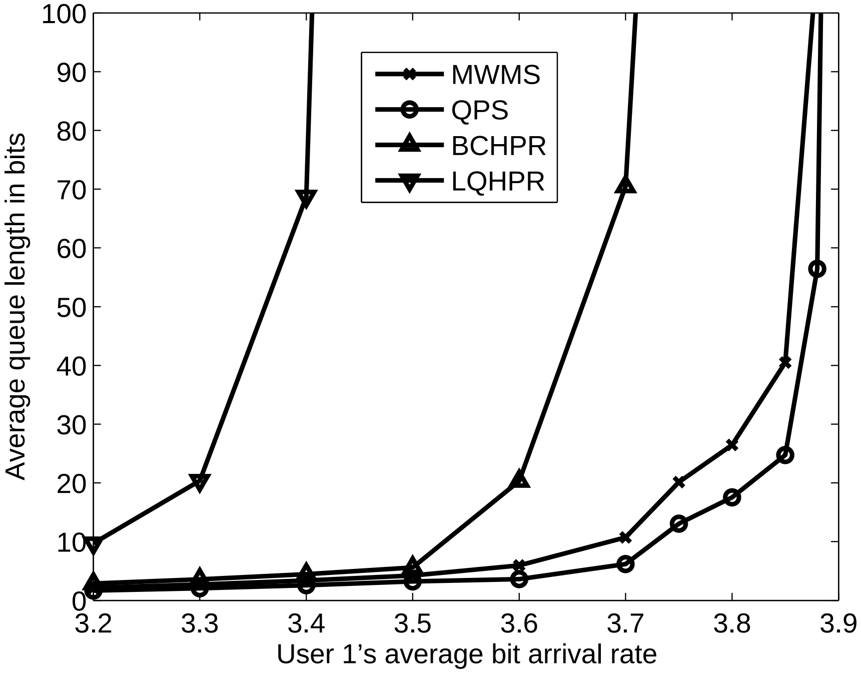


Figure 3.6: Average queue length vs user 1's bit arrival rate under five scheduling policies ($K = P = 2$, $M = 10$, user 1's average SNR=13dB, and user 2's average SNR=7dB, $\lambda_2 = 0.5\lambda_1$).

a 40-50% reduction in average queue length compared to MWMS, a larger difference than in the two user case. The throughput loss of LQHPR and BCHPR is around 30% and 10%, respectively, which is also much greater than in Fig. 3.6.

Delay fairness properties for the above four scheduling policies are illustrated in Fig. 3.8 where $K = 10$, $P = 10$, user i's average SNR is $20 - 0.5(i - 1)$ (dB), and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$. Fig. 3.8 demonstrates each user's average queueing delay in slots. It is observed that fairness among the users is not satisfied under BCHPR, which provides intolerably long average queueing delay for the users with worse channel conditions. Also, in Fig. 3.8, MWMS tends to provide smaller average queueing delay for the users with higher bit arrival rates. On the other hand, QPS guarantees fairness by equalizing every user's average queueing delay. Moreover, from Corollary 1 in Chapter 3.2, the ratio of each user's average queueing delay is

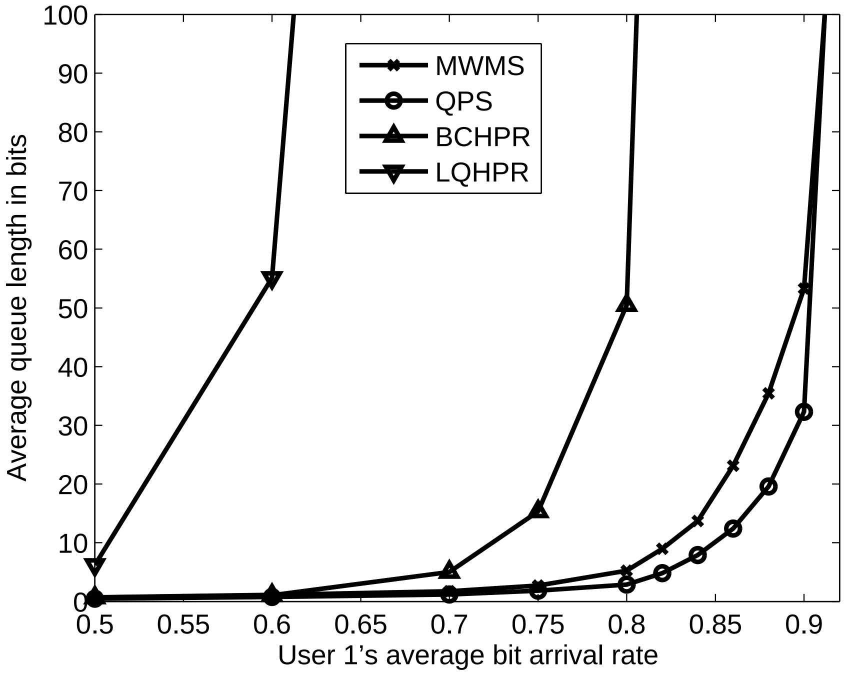


Figure 3.7: Average queue length vs user 1's bit arrival rate under five scheduling policies ($K = P = 10$, $M = 10$, user i's average SNR (dB) $= 20-(i-1)$, and $\lambda_i = \lambda_1$ for $i = 1, \cdots, 10$).

arbitrarily scalable by applying a modified queue-state vector to the QPS algorithm.

3.4 Summary

This chapter presents queue-proportional scheduling (QPS) and investigates its interesting properties. The throughput optimality of QPS is proved and it is shown that QPS can arbitrarily scale each user's average queueing delay relative to others, which is essential in satisfying each user's different QoS demand. Stochastic simulation results in both stationary and fading broadcast channels demonstrate the advantages of QPS over other schedulers such as maximum weight matching scheduling (MWMS), in terms of throughput, delay, and fairness. QPS is a promising queue-channel-aware scheduling policy for use in cross-layer resource allocation.

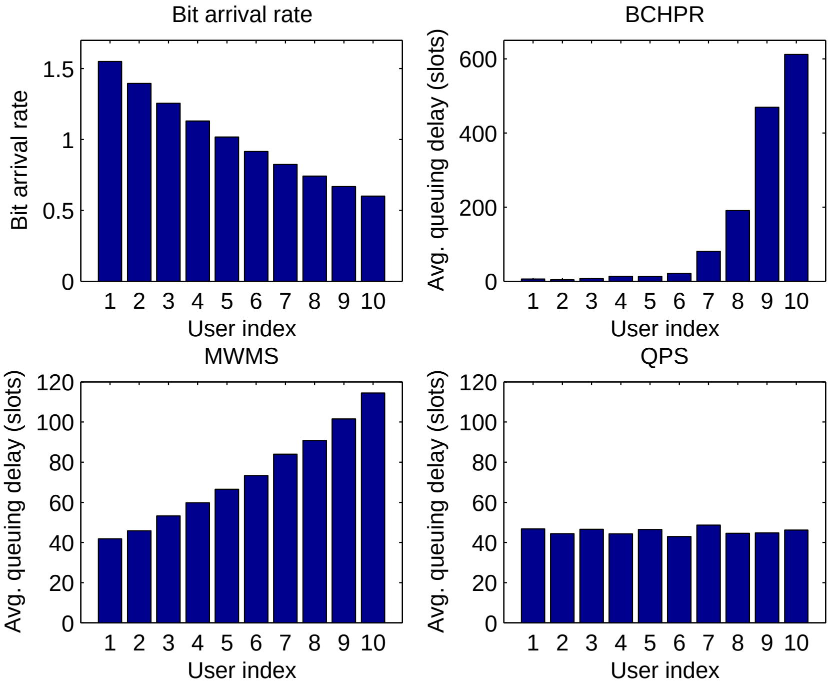


Figure 3.8: Each user's average queueing delay under four scheduling policies ($K = P = 10$, $M = 10$, user i's average SNR (dB) $= 20 - 0.5(i-1)$, and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

Chapter 4

Application of Geometric Programming

This chapter introduces yet another powerful tool, geometric programming (GP), into the family of numerical algorithms for cross-layer resource allocation problems in OFDM MAC and BC with CSIT. GP is a special case in convex optimization for which very efficient interior-point methods have been developed [10]. Reference [15] presents many interesting applications of GP in communication systems.

To achieve the channel capacity, superposition coding and successive decoding at the base-station can be used on each OFDM tone for downlink and uplink systems, respectively [32, 71]. With the application of such techniques, OFDM systems can dynamically allocate power and rate to each tone such that various QoS requirement of each user is satisfied. If each user's target data rate is fixed, minimization of transmit power reduces inter-cell interference levels in both up and down links as well as extends the battery life of each mobile terminal in the uplink.

Much progress has been made on resource allocation for the scalar Gaussian MAC and BC with ISI, where each user and the base-station are equipped with a single antenna. Cheng and Verdu [13] characterized the capacity region of Gaussian MAC with ISI and showed that the optimal input power spectral densities can be viewed as a generalization of the single-user water-filling spectrum. However, the lack of efficient numerical algorithms triggered much research for efficient resource allocation by using the inherent structure of Gaussian MAC. A breakthrough was made by Tse and Hanly [71], where polymatroid structure was used to characterize the capacity region of fading MAC, and marginal utility functions were introduced to develop algorithms that have strong greedy flavors. These results can be directly extended to Gaussian MAC and BC with ISI [70].

Recently, [81] proposed an efficient algorithm applicable to sum-rate maximization in the Gaussian OFDM MAC by using iterative water-filling (IWF) technique, which was first introduced for power control in interference channels [79]. The application of IWF has been further extended to the sum-power minimization problem in the Gaussian OFDM MAC by [46]. However, for general weighted sum-rate maximization or weighted sum-power minimization problems in the Gaussian OFDM MAC and BC, finding numerical algorithms with lower complexity still remains non-trivial. Also, because of the increasing demand in multi-media services such as video and audio streaming, real-time and non real-time traffic often coexist in the network. Thus, the constraints of resource allocation problems become more complicated, which requires the development of new algorithms.

This chapter first shows that, in the fading broadcast channel (BC) with CSIT, many schedulers including QPS and MWMS can be formulated via GP by using the degradedness of BC. This formulation simply extends to SISO-OFDM systems. A derived GP formulation allows simulation of the queueing delay for Poisson packet arrivals and exponentially distributed packet lengths under various scheduling policies. Numerical results corroborate the throughput optimality of QPS and indicate that QPS provides significantly smaller average queueing delay than MWMS. Moreover, QPS guarantees fairness in terms of average queueing delay for any arrival rate vectors. Compared to other schedulers, QPS has more variables and constraints, which may increase the complexity. This chapter also presents a scheme to simplify QPS by approximating the ergodic BC capacity region to a hypersphere. This method achieves the complexity of QPS comparable to other policies with a very small increase in the queueing delay.

With more generalization, GP formulation completely characterizes the achievable rate region as well as the achievable power region for both the OFDM broadcast and multiple access channels with CSIT. The GP perspective of multi-user OFDM resource allocation provides numerical efficiency as well as strong scalability for any additional constraints of GP form. First, the next subsection presents brief introduction of GP.

4.1 Geometric Programming

GP is a class of nonlinear optimization problems with a special form. Although GP in the standard form is a non-convex optimization problem, simple change in variables can convert it to a convex optimization problem that provides strong numerical efficiency [10]. A variety of recent algorithms allow efficient and reliable solution of GP [9]. GP applications can be found in many communication fields such as information theory, coding and signal processing, resource allocation, and queueing theory [15].

GP uses monomial and posynomial functions. A monomial function has the form of $h_j(\mathbf{x}) = c_j x_1^{a_{j,1}} x_2^{a_{j,2}} \cdots x_n^{a_{j,n}}$, where $\mathbf{x} \succ 0$, $c_j > 0$ and $a_{j,l} \in \mathbb{R}$. A posynomial is a sum of monomials $f_i(\mathbf{x}) = \sum_k c_{ik} x_1^{a_{ik,1}} x_2^{a_{ik,2}} \cdots x_n^{a_{ik,n}}$. Then, GP is

$$\begin{array}{rlll} \text{minimize} & f_0(\mathbf{x}) & & \\ \text{subject to} & f_i(\mathbf{x}) & \leq & 1 \\ & h_j(\mathbf{x}) & = & 1, \end{array} \tag{4.1}$$

where f_0 and f_i are posynomials and h_j are monomials. Although this is not a convex optimization problem, a change of variables: $y_i = \log x_i$ and $b_{ik} = \log c_{ik}$ converts it into a convex form:

$$\begin{array}{rlll} \text{minimize} & p_0(\mathbf{y}) & = & \log \Sigma_k \exp(\mathbf{a}_{0k}^T \mathbf{y} + b_{0k}) \\ \text{subject to} & p_i(\mathbf{y}) & = & \log \Sigma_k \exp(\mathbf{a}_{ik}^T \mathbf{y} + b_{ik}) \leq 0 \\ & q_j(\mathbf{y}) & = & \mathbf{a}_j^T \mathbf{y} + b_j = 0. \end{array} \tag{4.2}$$

The solution of this problem can be readily found by using well-developed efficient GP algorithms [9, 15].

As an example of GP formulation, consider the problem

$$\begin{array}{rl} \text{maximize} & x^2/y \\ \text{subject to} & 1 \le x \le 5 \\ & x^3 + 2z^2/y \le \sqrt{z} \\ & x/z = y^2, \end{array} \tag{4.3}$$

where $x, y, z > 0$. This problem can be readily turned into GP in the standard form as follows:

$$\begin{array}{rl} \text{minimize} & x^{-2}y \\ \text{subject to} & x^{-1} \le 1 \\ & \frac{1}{5}x \le 1 \\ & x^3 z^{-1/2} + 2y^{-1}z^{3/2} \le 1 \\ & xy^{-2}z^{-1} = 1. \end{array} \tag{4.4}$$

This standard GP formulation can be transformed into a convex optimization problem in $\tilde{x} = \log x$, $\tilde{y} = \log y$, and $\tilde{z} = \log z$:

$$\begin{array}{rl} \text{minimize} & -2\tilde{x} + \tilde{y} \\ \text{subject to} & -\tilde{x} \le 0 \\ & \tilde{x} - \log 5 \le 0 \\ & \log\left(\exp(3\tilde{x} - 1/2\tilde{z}) + \exp(-\tilde{y} + 3/2\tilde{z} + \log 2)\right) \le 0 \\ & \tilde{x} - 2\tilde{y} - \tilde{z} = 0. \end{array} \tag{4.5}$$

4.2 QPS via GP in a Fading BC with CSIT

This section presents GP formulations of QPS as well as other scheduling methods such as MWMS, BCHPR and LQHPR for the fading BC with CSIT. First, the next subsection describes the model of a fading BC.

4.2.1 Fading Broadcast Channel Model

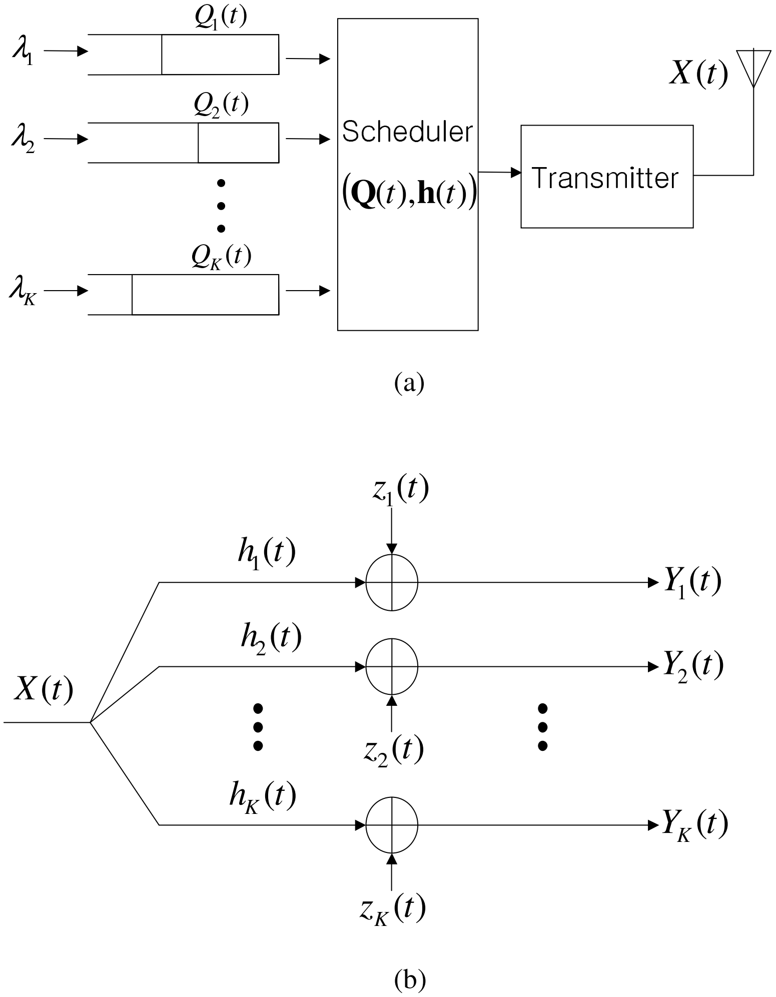
A block fading channel is assumed where the fading state is constant over one scheduling period and each scheduling period undergoes independent and identically distributed (i.i.d.) fading. Also, both the transmitter and receivers are assumed to have perfect knowledge of CSI. The capacity region of a Gaussian BC can be achieved by using superposition coding at the transmitter in conjunction with successive interference cancellation at each receiver [20]. With this optimal scheme, one user can remove the interference caused by other users' messages encoded earlier. Consider a Gaussian BC with a single transmitter sending independent messages to K users over two-sided bandwidth $2W$. It is assumed that the transmitter has a peak power constraint of P on each transmission. At time t, the received signal of user i is

$$Y_i(t) = h_i(t)X(t) + z_i(t), \quad i = 1, \cdots, K, \tag{4.6}$$

where the transmitted signal $X(t)$ is composed of K independent messages, the complex channel gain of user i is denoted by $h_i(t)$, and $z_i(t)$'s are i.i.d. zero-mean Gaussian band-limited noises with power N_0W. The models of fading broadcast channels and queueing systems that are used in this section are summarized in Fig. 4.1. As in [39], the channel gain can be combined with the noise component by defining an effective noise $\tilde{z}_i(t) = z_i(t)/h_i(t)$. Then, the equivalent received signal is given by

$$\tilde{Y}_i(t) = X(t) + \tilde{z}_i(t), \quad i = 1, \cdots, K, \tag{4.7}$$

where the power of $\tilde{z}_i(t)$ conditioned on the channel gain is defined as $n_i(t) = N_0W/|h_i(t)|^2$. Without loss of generality, $W = 1$ is assumed for simplicity. The effective noise power $\mathbf{n} = [n_1 \; n_2 \; \cdots \; n_K]^T$ is used to denote a fading state. The ergodic capacity region of a fading BC is the set of all long-term average rate vectors achievable in a fading BC with arbitrarily small probability of error. A power control policy $\mathcal{P}$ over all possible fading states is defined as a function that maps from any fading state $\mathbf{n}$ to each user's transmit power $P_i(\mathbf{n})$. Let Ω denote the set of all power policies satisfying the sum-power constraint P, which is given by



(a)

(b)

Figure 4.1: (a) Block diagram of the queueing system and scheduler. (b) Fading broadcast channel models.

$$\Omega = \left\{ \mathcal{P} : \sum_{i=1}^{K} P_i(\mathbf{n}) \leq P, \text{ for all } \mathbf{n} \right\}. \tag{4.8}$$

For each fading state, the channel is a degraded Gaussian BC where the capacity region is achieved by later encoding a message of the user with smaller effective noise power. With this optimal ordering, user i's capacity for a fading state $\mathbf{n}$ is

$$R_i\left(\mathbf{P}(\mathbf{n})\right) = \log_2 \left(1 + \frac{P_i(\mathbf{n})}{n_i + \sum_{k=1}^{K} P_k(\mathbf{n}) 1\left[n_i > n_k\right]} \right) \tag{4.9}$$

where $\mathbf{P}(\mathbf{n}) = [P_1(\mathbf{n}) \ P_2(\mathbf{n}) \ \cdots \ P_K(\mathbf{n})]^T$ and $1[\cdot]$ is the indicator function, which equals 1 if its argument is satisfied; 0 otherwise. Then, the capacity region of a Gaussian BC for the fading state $\mathbf{n}$ and transmit power P is

$$\begin{aligned} C(\mathbf{n}, P) = \{R_i : R_i \leq R_i\left(\mathbf{P}(\mathbf{n})\right), \ i = 1, 2, \cdots, K, \\ \text{where } \sum_i P_i(\mathbf{n}) = P\}. \end{aligned} \tag{4.10}$$

Let $C_{BC}(\mathcal{P})$ denote the set of achievable rates averaged over all fading states for a power policy $\mathcal{P}$

$$C_{BC}(\mathcal{P}) = \left\{R_i : R_i \leq \mathbb{E}_{\mathbf{n}}\left[R_i\left(\mathbf{P}(\mathbf{n})\right)\right], \ i = 1, \cdots, K\right\}. \tag{4.11}$$

With the sum-power constraint P and perfect CSI at the transmitter and receivers, the ergodic capacity region of a fading BC is given by [39]

$$C_{erg}(P) = \bigcup_{\mathcal{P} \in \Omega} C_{BC}(\mathcal{P}) \tag{4.12}$$

where the region $C_{erg}(P)$ is convex.

4.2.2 GP Formulation of QPS

The QPS algorithm presented in Chapter 3 allocates the following data-rate vector at time slot t:

$$\begin{aligned} \mathbf{R}_{QPS}\left(\mathbf{n}(t), \mathbf{Q}(t)\right) &\in C\left(\mathbf{n}(t), P\right) \quad \text{such that} \\ \mathbb{E}_{\mathbf{n}(t)}\left[\mathbf{R}_{QPS}\left(\mathbf{n}(t), \mathbf{Q}(t)\right)\right] &= \mathbf{Q}'(t)\left(\max_{\mathbf{Q}'(t)x \in C_{erg}(P)} x\right), \end{aligned} \tag{4.13}$$

where x is a scalar. Assuming equal priority on each user, $\mathbf{Q}'(t) = \mathbf{Q}(t)$. Then, the average rate vector under the QPS policy, $\mathbb{E}_{\mathbf{n}(t)}[\mathbf{R}_{QPS}(\mathbf{n}(t), \mathbf{Q}(t))]$ is proportional to the queue-state vector and also lies on the boundary surface of the ergodic capacity region. As shown in [70], each boundary point of $C_{erg}(P)$ in a fading BC is a solution to the optimization problem $\max_{\mathbf{r}} \boldsymbol{\mu} \cdot \mathbf{r}$ where $\mathbf{r} \in C_{erg}(P)$ for some $\boldsymbol{\mu} \in \mathbb{R}_+^K$. When such $\boldsymbol{\mu}$ is given, $\mathbf{R}_{QPS}(\mathbf{n}(t), \mathbf{Q}(t))$ is a solution to the optimization problem $\max_{\mathbf{r}} \boldsymbol{\mu} \cdot \mathbf{r}$ where $\mathbf{r} \in C\left(\mathbf{n}(t), P\right)$ for any fading state $\mathbf{n}(t)$. Therefore, the data-rate vector assigned by QPS at time slot t can be expressed as

$$\begin{aligned} \mathbf{R}_{QPS}\left(\mathbf{n}(t), \mathbf{Q}(t)\right) = \arg\max_{\mathbf{r}} &\, \boldsymbol{\mu}^T \mathbf{r} \\ \text{such that } &\mathbf{r} \in C\left(\mathbf{n}(t), P\right). \end{aligned} \tag{4.14}$$

Under the QPS policy, $\boldsymbol{\mu}$ is determined based on the current queue-state vector as well as on the ergodic capacity region of the fading BC. By contrast, as shown in (2.24), MWMS only considers the queue-state vector in deriving the weight vector.

By utilizing the degradedness of the BC for each fading state, the rate allocation of QPS can be formulated via GP. Assume that the M most recent fading states are sampled, which are denoted by $\left\{\mathbf{n}^{(1)}, \cdots, \mathbf{n}^{(M)}\right\}$. To reduce the correlation among samples, the sampling period needs to be extended in consideration of fading coherence time. The sampling period is simply assumed equal to one scheduling period because of i.i.d. block fading over each scheduling time. Without loss of generality, $\mathbf{n}^{(M)}$ is assumed to denote the current fading state $\mathbf{n}(t)$. Then, consider a family of M parallel Gaussian broadcast channels, such that in the mth component channel, user i has effective noise variance $n_i^{(m)}$, rate $R_i^{(m)}$, and power $P_i^{(m)}$. Each BC has a power constraint of P. At time slot t, QPS allocates the data-rate vector $\mathbf{R}_{QPS}(\mathbf{n}^{(M)}, \mathbf{Q}(t))$ that is a solution of the following optimization problem.

$$\frac{1}{M}\sum_{m=1}^{M}\mathbf{R}_{QPS}\left(\mathbf{n}^{(m)},\mathbf{Q}(t)\right)=\mathbf{Q}(t)\left(\max_{\mathbf{Q}(t)x\in C_{erg}(P)} x\right)$$
$$\mathbf{R}_{QPS}\left(\mathbf{n}^{(m)},\mathbf{Q}(t)\right)\in C\left(\mathbf{n}^{(m)},P\right)\quad\text{for all } m. \tag{4.15}$$

From (4.9) and (4.10), the capacity region of the mth Gaussian BC is given by

$$C\left(\mathbf{n}^{(m)},P\right)=\{R^{(m)}_{\pi_m(i)}:R^{(m)}_{\pi_m(i)}\leq \log_2\left(1+\frac{\alpha^{(m)}_{\pi_m(i)}P}{n^{(m)}_{\pi_m(i)}+\sum_{j<i}\alpha^{(m)}_{\pi_m(j)}P}\right), i=1,2,\cdots,K \text{ where } \sum_i \alpha^{(m)}_{\pi_m(i)}=1\}, \tag{4.16}$$

where $\pi_m(\cdot)$ is the permutation such that $n^{(m)}_{\pi_m(1)}<n^{(m)}_{\pi_m(2)}<\cdots<n^{(m)}_{\pi_m(K)}$, and $\alpha^{(m)}_{\pi_m(i)}$ is the fraction of the total transmit power used for user $\pi_m(i)$'s signal in the mth Gaussian BC. When $\mathbf{R}^{(m)}$ is on the boundary of the capacity region, solving the $\alpha^{(m)}_{\pi_m(i)}$'s in terms of the rate vector $\mathbf{R}^{(m)}$ yields the following equations.

$$\sum_{i=1}^{l}\alpha^{(m)}_{\pi_m(i)}P=\sum_{i=1}^{l}\left(n^{(m)}_{\pi_m(i)}-n^{(m)}_{\pi_m(i-1)}\right)\times\exp\left(\log 2\sum_{j=i}^{l}R^{(m)}_{\pi_m(j)}\right)-n^{(m)}_{\pi_m(l)},\quad l=1,\cdots,K, \tag{4.17}$$

where $n^{(m)}_{\pi_m(0)}\equiv 0$. As shown in [39], (4.16) is equivalent to

$$C(\mathbf{n}^{(m)},P)=\{R^{(m)}_{\pi_m(i)}:\sum_{i=1}^{K}\left(n^{(m)}_{\pi_m(i)}-n^{(m)}_{\pi_m(i-1)}\right)\times\exp\left(\log 2\sum_{j=i}^{K}R^{(m)}_{\pi_m(j)}\right)-n^{(m)}_{\pi_m(K)}\leq P \text{ and } R^{(m)}_{\pi_m(i)}\geq 0,\ i=1,2,\cdots,K\}. \tag{4.18}$$

Using this relation, (4.15) can be converted into

$$\begin{aligned}
\text{minimize} \quad & \log\left(\exp(-x)\right) \\
\text{subject to} \quad & \log\left(\exp\left(-R_i^{(m)}\right)\right) \leq 0, \quad \forall\ i \text{ and } m \\
& \log\left(\exp\left(-Q_i(t)\right)\exp\left(R_i^{(M)}\right)\right) \leq 0, \quad \forall\ i \\
& \log\sum_{i=1}^{K}\left(\frac{n_{\pi_m(i)}^{(m)} - n_{\pi_m(i-1)}^{(m)}}{P + n_{\pi_m(K)}^{(m)}}\right) \\
& \times\exp\left(\log 2\sum_{j=i}^{K} R_{\pi_m(j)}^{(m)}\right) \leq 0, \quad \forall\ m \\
& \mathbf{Q}(t)x - \frac{1}{M}\sum_{m=1}^{M}\mathbf{R}^{(m)} = 0,
\end{aligned} \tag{4.19}$$

where the second set of constraints is added to avoid allocating redundant power to some users with short queue lengths. If the optimization variable is defined as $\mathbf{y} = [x\ (\mathbf{R}^{(1)})^T\ \cdots\ (\mathbf{R}^{(M)})^T]^T \in \mathbb{R}_+^{(KM+1)}$, (4.19) is a standard geometric program with the globally optimal solution $\mathbf{y}^* = [x^*\ (\mathbf{R}^{*(1)})^T\ \cdots\ (\mathbf{R}^{*(M)})^T]^T$. Then, the data-rate vector supported under the QPS policy is $\mathbf{R}_{QPS}(\mathbf{n}^{(M)}, \mathbf{Q}(t)) = \mathbf{R}^{*(M)}$, and the corresponding power allocation can be obtained by solving (4.33) for $m = M$. This GP formulation of QPS can be extended to OFDM systems, as discussed in the next subsection.

4.2.3 Extension to OFDM Systems

In a fading BC with inter-symbol interference (ISI), the ISI can be completely removed by exploiting OFDM techniques with sufficient number of tones, i.e. the frequency response can be made flat within each tone [18]. OFDM systems will have K users and L tones. On each tone, the channel is equivalent to a fading BC without ISI, which becomes a degraded Gaussian BC for the fixed fading state. Therefore, by extending the results from Section 4.2, QPS for OFDM systems in a fading BC can be also converted into GP. At tone l, M sampled fading state vectors are denoted by

$\{\mathbf{n}^{(l,1)}, \cdots, \mathbf{n}^{(l,M)}\}$ where $\mathbf{n}^{(l,m)} = [n_1^{(l,m)} \ \cdots \ n_K^{(l,m)}]^T$. For the mth sampled fading state, $n_i^{(l,m)}$, $R_i^{(l,m)}$, and $P_i^{(l,m)}$ denote the effective noise variance, rate, and power on user i's tone l, respectively. Without loss of generality, the Mth sample is assumed to denote the current fading state. Also, a total-power constraint of P is imposed on each transmission of OFDM symbols. Define $\pi_{l,m}(\cdot)$ as the permutation such that $n^{(l,m)}_{\pi_{l,m}(1)} < n^{(l,m)}_{\pi_{l,m}(2)} < \cdots < n^{(l,m)}_{\pi_{l,m}(K)}$. By carefully applying above updates to (4.18) and (4.19), QPS in OFDM systems can be converted into the following GP:

$$\begin{aligned}
\text{minimize} \quad & \log\left(\exp(-x)\right) \\
\text{subject to} \quad & \log\left(\exp\left(-R_i^{(l,m)}\right)\right) \leq 0, \quad \forall \ i, \ l, \text{ and } m \\
& \log\left(\exp\left(-Q_i(t)\right)\exp\left(\sum_{l=1}^{L} R_i^{(l,M)}\right)\right) \leq 0, \quad \forall \ i \\
& \log \sum_{l=1}^{L}\sum_{i=1}^{K}\left(\frac{n^{(l,m)}_{\pi_{l,m}(i)} - n^{(l,m)}_{\pi_{l,m}(i-1)}}{P + \sum_{s=1}^{L} n^{(s,m)}_{\pi_{s,m}(K)}}\right) \\
& \times \exp\left(\log 2 \sum_{j=i}^{K} R^{(l,m)}_{\pi_{l,m}(j)}\right) \leq 0, \quad \forall \ m \\
& \mathbf{Q}(t)x - \frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L}\mathbf{R}^{(l,m)} = 0,
\end{aligned} \tag{4.20}$$

where $n^{(l,m)}_{\pi_{l,m}(0)} \equiv 0$. Denote the optimization variable by $\mathbf{y} = [x \ (\mathbf{R}^{(1,1)})^T \ \cdots \ (\mathbf{R}^{(L,M)})^T]^T \in \mathbb{R}_+^{(KLM+1)}$. Then, (4.20) is a standard geometric program with the globally optimal solution $\mathbf{y}^* = [x^* \ (\mathbf{R}^{*(1,1)})^T \ \cdots \ (\mathbf{R}^{*(L,M)})^T]^T$. Consequent rate allocation on tone l under the QPS policy is $\mathbf{R}_{QPS}^{(l)}(\mathbf{n}^{(l,M)}, \mathbf{Q}(t)) = \mathbf{R}^{*(l,M)}$ for $l = 1, \cdots, L$, and the corresponding power allocation can be obtained by applying (4.33) on each tone with $m = M$.

4.2.4 Other Scheduling Policies via GP

This subsection provides GP formulations of three other scheduling methods in a fading BC: Maximum Weight Matching Scheduling (MWMS), Best Channel Highest

Possible Rate (BCHPR), and Longest Queue Highest Possible Rate (LQHPR).

Maximum Weight Matching Scheduling via GP

At time slot t, the rate allocation under MWMS can be found by solving (2.24), which is the weighted sum-rate maximization problem over $C(\mathbf{n}(t), P)$ considering the queue state vector $\mathbf{Q}(t)$ as the weight vector. Using (4.18), MWMS can be formulated as the following GP:

$$\begin{aligned} \text{minimize} \quad & \log\left(\exp(-\mathbf{Q}(t)^T\mathbf{r})\right) \\ \text{subject to} \quad & \log\left(\exp(-r_i)\right) \le 0, \quad \forall\ i \\ & \log\left(\exp\left(-Q_i(t)\right)\exp\left(r_i\right)\right) \le 0, \quad \forall\ i \\ & \log\sum_{i=1}^{K}\left(\frac{n_{\pi(i)}(t) - n_{\pi(i-1)}(t)}{P + n_{\pi(K)}(t)}\right) \\ & \times\exp\left(\log 2\sum_{j=i}^{K} r_{\pi(j)}\right) \le 0, \end{aligned} \tag{4.21}$$

where $\pi(\cdot)$ is the permutation such that $n_{\pi(1)}(t) < n_{\pi(2)}(t) < \cdots < n_{\pi(K)}(t)$. Let $\mathbf{r}^*$ be the solution of (4.21), then $\mathbf{R}_{MWMS}\left(\mathbf{n}(t), \mathbf{Q}(t)\right) = \mathbf{r}^*$.

Best Channel Highest Possible Rate via GP

Under the BCHPR policy, a user with the better channel condition takes higher priority in resource allocation. Also, user i is served only if some transmit power remains after clearing queue backlogs of users with higher priorities than user i. This algorithm is equivalent to allocating a data-rate vector that minimizes the l_1-norm distance from the current queue-state vector. The l_1-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$. At time slot t, the BCHPR policy supports the rate vector $\mathbf{R}_{BCHPR}\left(\mathbf{n}(t), \mathbf{Q}(t)\right)$ that is a solution of the following optimization problem.

$$\min\|\mathbf{Q}(t) - \mathbf{r}\|_1 \quad \text{subject to } \mathbf{r} \in C\left(\mathbf{n}(t), P\right). \tag{4.22}$$

With the constraint of $\mathbf{r} \preceq \mathbf{Q}(t)$, the solution of the above problem is unaffected by $\sum_{i=1}^{K} Q_i(t)$. After removing this summation from the objective, (4.22) can be converted into the following GP.

$$\begin{aligned} \text{minimize} \quad & \log\left(\exp\left(-\mathbf{1}^T\mathbf{r}\right)\right) \\ \text{subject to} \quad & \log\left(\exp(-r_i)\right) \leq 0, \quad \forall\ i \\ & \log\left(\exp\left(-Q_i(t)\right)\exp\left(r_i\right)\right) \leq 0, \quad \forall\ i \\ & \log\sum_{i=1}^{K}\left(\frac{n_{\pi(i)}(t) - n_{\pi(i-1)}(t)}{P + n_{\pi(K)}(t)}\right) \\ & \times \exp\left(\log 2 \sum_{j=i}^{K} r_{\pi(j)}\right) \leq 0. \end{aligned} \tag{4.23}$$

Let $\mathbf{r}^*$ be the solution of (4.23), then $\mathbf{R}_{BCHPR}\left(\mathbf{n}(t), \mathbf{Q}(t)\right) = \mathbf{r}^*$. When $\mathbf{Q}(t) \succeq \mathbf{r}$ for any $\mathbf{r} \in C\left(\mathbf{n}(t), P\right)$, the BCHPR policy solely depends on channel conditions. At each scheduling time, it allocates total power to the single user with the best channel condition, which is a sum-rate maximizing scheme in a fading BC [72].

Longest Queue Highest Possible Rate via GP

LQHPR schedules a data-rate vector to minimize the longest queue length, which is equivalent to selecting a rate vector that minimizes the l_∞-norm distance from the current queue-state vector. The l_∞-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as $\|\mathbf{x}\|_\infty = \max\{|x_1|, \cdots, |x_n|\}$. Hence, at time slot t, the LQHPR policy assigns the rate vector $\mathbf{R}_{LQHPR}\left(\mathbf{n}(t), \mathbf{Q}(t)\right)$ that is a solution of the following optimization problem.

$$\min \|\mathbf{Q}(t) - \mathbf{r}\|_\infty \quad \text{subject to } \mathbf{r} \in C\left(\mathbf{n}(t), P\right). \tag{4.24}$$

Let x denote the upper bound on $\|\mathbf{Q}(t) - \mathbf{r}\|_\infty$ such that $-x\mathbf{1} \prec \mathbf{Q}(t) - \mathbf{r} \prec x\mathbf{1}$. Then, the above equation can be represented as

$$\text{minimize} \qquad \log\left(\exp\left(x\right)\right)$$

$$\begin{aligned}
\text{subject to} \quad & \log\left(\exp(-r_i)\right) \leq 0, \quad \forall\, i \\
& \log\left(\exp\left(-Q_i(t)\right)\exp(-x + r_i)\right) \leq 0, \quad \forall\, i \\
& \log\left(\exp\left(Q_i(t)\right)\exp(-x - r_i)\right) \leq 0, \quad \forall\, i \\
& \log \sum_{i=1}^{K} \left(\frac{n_{\pi(i)}(t) - n_{\pi(i-1)}(t)}{P + n_{\pi(K)}(t)} \right) \\
& \times \exp\left(\log 2 \sum_{j=i}^{K} r_{\pi(j)} \right) \leq 0. \qquad (4.25)
\end{aligned}$$

Define the optimization variable as $\mathbf{y} = [x \ \ \mathbf{r}^T]^T$ then, (4.25) is a standard geometric program with the globally optimal point $\mathbf{y}^* = [x^* \ \ \mathbf{r}^{*T}]^T$. The data-rate vector supported under LQHPR is $\mathbf{R}_{LQHPR}\left(\mathbf{n}(t), \mathbf{Q}(t)\right) = \mathbf{r}^*$.

4.2.5 Hypersphere Approximation of the Ergodic Capacity Region of a Fading BC

At each scheduling time, QPS solves (4.19) which has $KM + 1$ optimization variables and $KM + 2K + M$ constraints*. In order to capture the fading statistics, QPS requires the number of sampled fading states, $M \gg 1$. Even though GP can be efficiently solved and the constraint matrix of (4.19) is sparse, $M \gg 1$ implies that the computational complexity of QPS can be higher than other scheduling polices such as MWMS, which has K variables and $2K + 1$ constraints. The expected rate vector under QPS is a boundary point of the ergodic capacity region that is proportional to the current queue-state vector. The rate allocation satisfying this condition can be obtained by solving (4.14) with a proper weight vector $\boldsymbol{\mu}$ determined from the current queue-state vector and ergodic capacity region. With the QPS policy, $\boldsymbol{\mu}$ is a normal vector of the tangent plane, which is drawn at the boundary point of $C_{erg}(P)$ supported by QPS. Thus, if the boundary surface of $C_{erg}(P)$ can be characterized with a simple function, finding $\boldsymbol{\mu}$ becomes much easier, and the computational complexity of QPS becomes comparable to other scheduling policies.

*In OFDM systems with L tones, QPS solves (4.20) that has $KML + 1$ optimization variables and $KML + 2K + M$ constraints.

This section proposes a simple method to approximate the boundary surface of $C_{erg}(P)$ by a hypersphere. By allowing a small increase in the average queueing delay, this hypersphere-approximation method solves the complexity issue of QPS. First, $K+1$ boundary points on $C_{erg}(P)$ are sampled to characterize the K-dimensional hypersphere. K points correspond to each user's average rate when total transmit power is allocated to that user. They are equivalent to the intercept of each user's rate axis with $C_{erg}(P)$. The remaining point is the maximum average sum-rate vector achieved by transmitting only to the best user at each scheduling period. The next lemma provides the uniqueness of K-dimensional hypersphere constructed by using these $K+1$ rate vectors.

Lemma 1. *In a fading BC with K users, there exists a unique K-dimensional hypersphere characterized with each user's maximum average rate vector and the maximum average sum-rate vector.*

Proof. Let user i's maximum average rate vector be denoted by $\mathbf{x}_i = a_i\mathbf{e_i} \in \mathbb{R}^K_+$ where $\mathbf{e_i}$ is a unit vector whose ith element is 1 and the others are 0's. Also, denote the maximum average sum-rate vector by $\mathbf{x}_s \in \mathbb{R}^K_+$. In a fading BC, the sum rate is maximized by allocating full power to the best user. When excluding the trivial case where the best user is always identical, $\mathbf{x}_s$ exists outside the $K-1$ dimensional hyperplane that passes through $\mathbf{x}_i$'s for $i = 1, \cdots, K$. The center of the K-dimensional hypersphere is denoted by $\mathbf{x}_c \in \mathbb{R}^K$. Then, $\|\mathbf{x}_c - \mathbf{x}_s\|_2 = \|\mathbf{x}_c - \mathbf{x}_i\|_2$ for $i = 1, \cdots, K$. Therefore, the following linear equation is obtained.

$$A\mathbf{x}_c = \mathbf{b} \quad \text{where} \quad A = \begin{bmatrix} 2(\mathbf{x}_1 - \mathbf{x}_s)^T \\ 2(\mathbf{x}_2 - \mathbf{x}_s)^T \\ \vdots \\ 2(\mathbf{x}_K - \mathbf{x}_s)^T \end{bmatrix} \in \mathbb{R}^{K\times K}$$

$$\text{and} \quad \mathbf{b} = \begin{bmatrix} \mathbf{x}_1^T\mathbf{x}_1 - \mathbf{x}_s^T\mathbf{x}_s \\ \mathbf{x}_2^T\mathbf{x}_2 - \mathbf{x}_s^T\mathbf{x}_s \\ \vdots \\ \mathbf{x}_K^T\mathbf{x}_K - \mathbf{x}_s^T\mathbf{x}_s \end{bmatrix} \in \mathbb{R}^{K\times 1}. \tag{4.26}$$

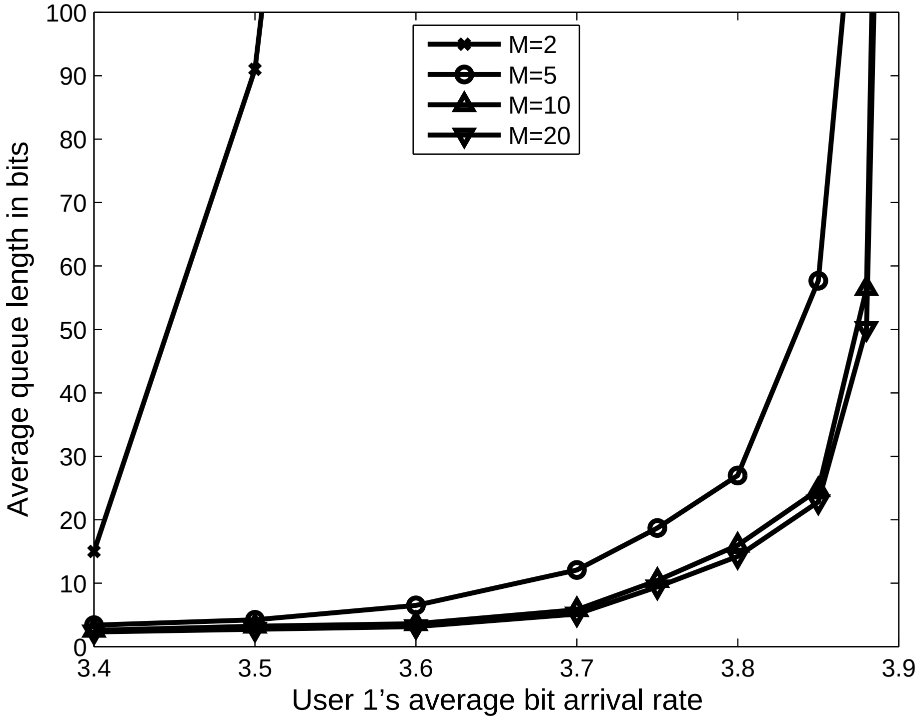


Figure 4.2: Average queue length under QPS vs user 1's bit arrival rate for M=2, 5, 10, and 20 ($K = P = 2$, user 1's average SNR=13dB, and user 2's average SNR=7dB, $\lambda_2 = 0.5\lambda_1$).

A is nonsingular since every row of A is independent of the other rows. Thus, $\mathbf{x}_c$ has a unique solution, which is $\mathbf{x}_c = A^{-1}\mathbf{b}$. □

Let $\mathbf{x_b}$ denote a boundary point of the hypersphere that is proportional to the current queue-state vector. Then, $\mathbf{x_b} = k\mathbf{Q}(t)$, where $k \geq 0$. The value of k can be found by solving $\|\mathbf{x}_c - \mathbf{x}_1\|_2 = \|\mathbf{x}_c - k\mathbf{Q}(t)\|_2$. If the weight vector for QPS acquired from the hypersphere approximation is denoted by $\boldsymbol{\mu'}$, then $\boldsymbol{\mu'} = \mathbf{x_b} - \mathbf{x_c}$.

4.2.6 Numerical Results and Discussion

By using the GP formulations derived in the previous subsections, this subsection presents simulation results with Poisson packet arrivals and exponentially distributed packet lengths to demonstrate stability, delay, and fairness properties of the QPS algorithm. In the simulations, the average packet length for each user, the scheduling period, and the signal bandwidth are all equal to 1. Also, the average queue length over all users is defined as $\lim_{t\to\infty}\mathbb{E}[\frac{1}{K}\sum_{i=1}^{K}Q_i(t)]$. First, Fig. 4.2 demonstrates the effect of the number of sampled fading states, M on the average queue length under QPS. A Rayleigh fading BC presented in Fig. 3.2 is considered where $P = 2$, user 1's average SNR=13dB, and user 2's average SNR=7dB. Also, the bit-arrival rate of user 2 is assumed to be half that of user 1's. Thus, the bit-arrival rate vector can be represented as $\boldsymbol{\lambda} = \lambda_1[1\ 0.5]^T$. From Fig. 3.2, $\boldsymbol{\lambda} \in \text{int}\ C_{erg}(P)$ if and only if $\lambda_1 < 3.9$. The average queue lengths are for different values of λ_1 when M=2, 5, 10, and 20. Fig. 4.2 shows that as M increases, larger throughput and smaller average queue length can be achieved with QPS. About 10% throughput loss occurs with $M = 2$, compared to the maximum achievable throughput. However, this loss quickly vanishes with larger M, which becomes much less than 1% for $M = 5$. Also, it is shown that for $M > 10$, the additional decrease in average queue length is quite small, which suggests that about 10 independent fading samples are sufficient in using QPS.

Fig. 4.3 and Fig. 4.4 present average queue lengths for different values of λ_1 when $K = 2$ and $K = 10$, respectively. Both figures use $M = 10$ and compare five scheduling algorithms: QPS, QPS with the hypersphere approximation, MWMS, BCHPR and LQHPR. For the two user case in Fig. 4.3, the channel and input traffic conditions are assumed to be the same as in Fig. 4.2. Fig. 4.3 shows that the average queue length of QPS is about 30% smaller than that of MWMS for any $\lambda_1 < 3.9$. Since MWMS is a throughput optimal policy, this observation corroborates the throughput optimality of QPS. LQHPR and BCHPR, which are not throughput optimal, have about 12% and 5% throughput loss, respectively. QPS using the hypersphere approximation of $C_{erg}(P)$ slightly increases the average queue length compared to QPS. However, its average queue length is still much smaller than MWMS. Simulation results with 10 users are presented in Fig. 4.4. $P = 10$ and user i's average SNR is equal to $20-(i-1)$ (dB) for $i = 1, \cdots, 10$. Also, the bit arrival rate is identical for all users. QPS is observed to provide about a 40-50% reduction in average queue length

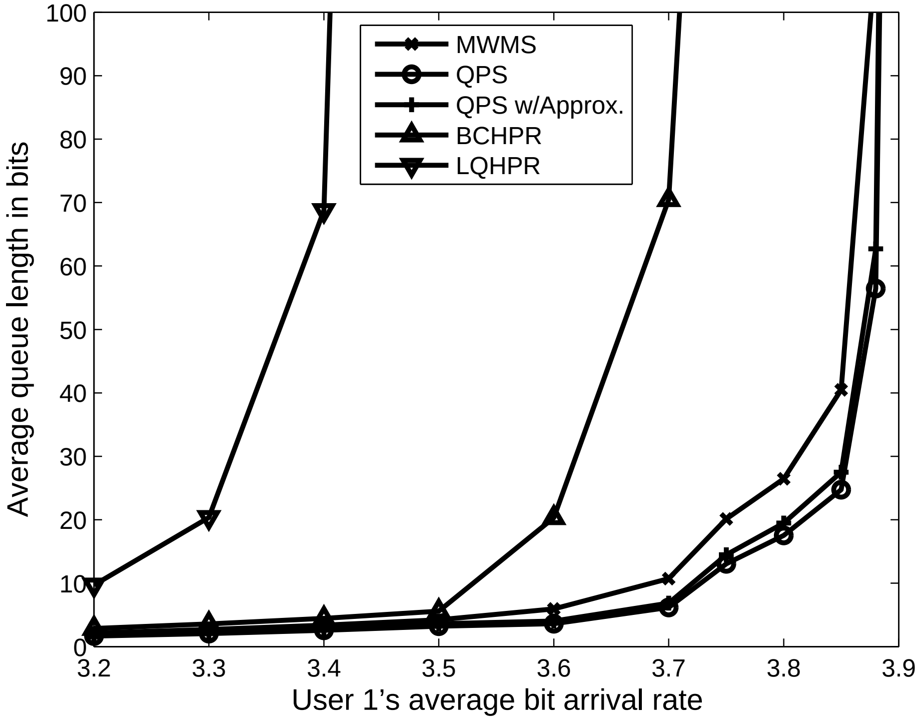


Figure 4.3: Average queue length vs user 1's bit arrival rate under five scheduling policies ($K = P = 2$, $M = 10$, user 1's average SNR=13dB, and user 2's average SNR=7dB, $\lambda_2 = 0.5\lambda_1$).

compared to MWMS, a larger difference than in the two user case. The throughput loss of LQHPR and BCHPR is around 30% and 10%, respectively, which is also much greater than in Fig. 4.3. Accuracy of the hypersphere approximation is somewhat lower than in the two user case, but this method still gives about a 30% decrease in the average queue length compared to MWMS.

The fairness properties of QPS, QPS with the hypersphere approximation, MWMS and BCHPR with 10 users are illustrated in Fig. 4.5 and Fig. 4.6. $P = 10$, $M = 10$, user i's average SNR is equal to $20 - 0.5(i - 1)$ (dB), and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$. First, Fig. 4.5 demonstrates each user's average queue length in bits for the above four scheduling policies. It is observed that fairness among users is not satisfied under the BCHPR, which provides intolerably long average queueing

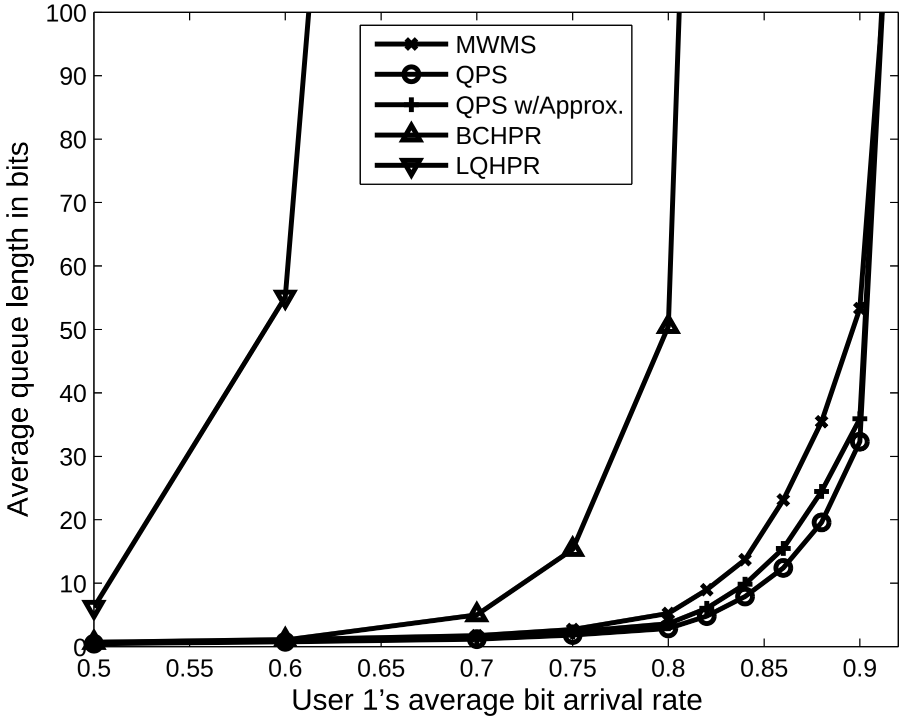


Figure 4.4: Average queue length vs user 1's bit arrival rate under five scheduling policies ($K = P = 10$, $M = 10$, user i's average SNR (dB) $= 20 - (i - 1)$, and $\lambda_i = \lambda_1$ for $i = 1, \cdots, 10$).

delay for users with worse channel conditions. MWMS approximately equalizes every user's average queue length. Since each user has a different arrival rate, by Little's theorem, MWMS provides smaller average queueing delay for the user with higher bit arrival rate. On the other hand, each user's average queue length under QPS is shown to be proportional to the bit arrival rate vector so that average queueing delay of every user is equalized. Therefore, under the QPS policy, fairness among users is guaranteed in terms of average queueing delay. QPS with the hypersphere approximation also shows a similar tendency with QPS, but some deviation from the arrival rate vector is observed because of the approximation error. Fig. 4.6 presents each user's average queueing delay in slots, which indicates that QPS equalizes every user's average queueing delay.

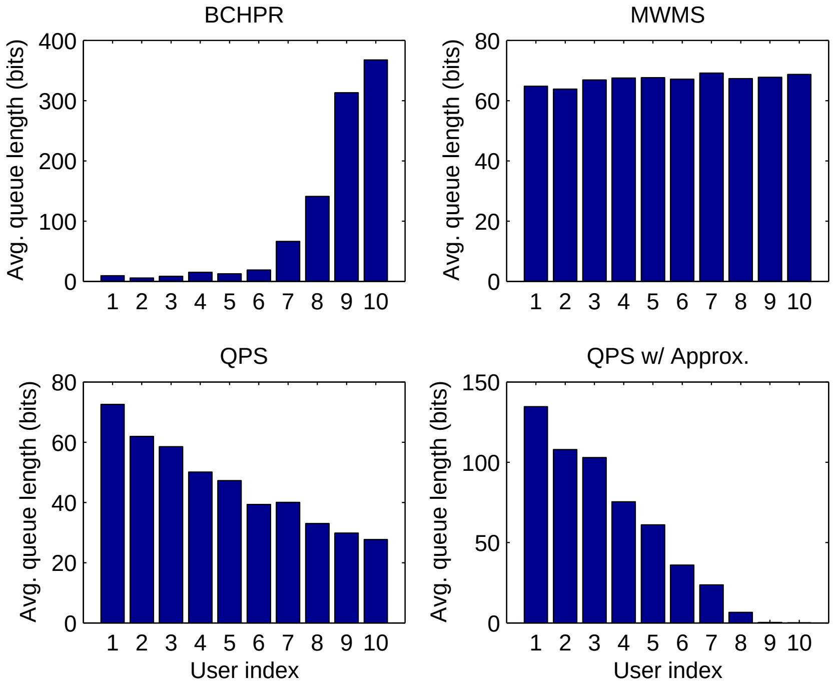


Figure 4.5: Each user's average queue length under four scheduling policies ($K = P = 10$, $M = 10$, user i's average SNR (dB) $= 20 - 0.5(i-1)$, and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

4.3 GP for OFDM BC and MAC with CSIT

This section further extends the GP formulation to various resource allocation problems for the OFDM MAC and BC with CSIT. The primary focus is on the following three major resource allocation problems in OFDM MAC and BC: weighted sum-rate maximization (WSRmax), weighted sum-power minimization (WSPmin) and proportional-rate maximization (PRmax). These problems are essential in performing cross-layer resource allocation to guarantee each user's QoS satisfaction. With the appropriate choice of the weight vector, WSRmax can be used to support any gradient-based scheduling policies such as MWMS and PFS. For services requesting constant-rate, WSPmin is useful in minimizing the inter-cell interference as well as maximizing the battery power of mobile terminals while allowing different priorities on each user. PRmax can be used for guaranteeing proportional fairness among users

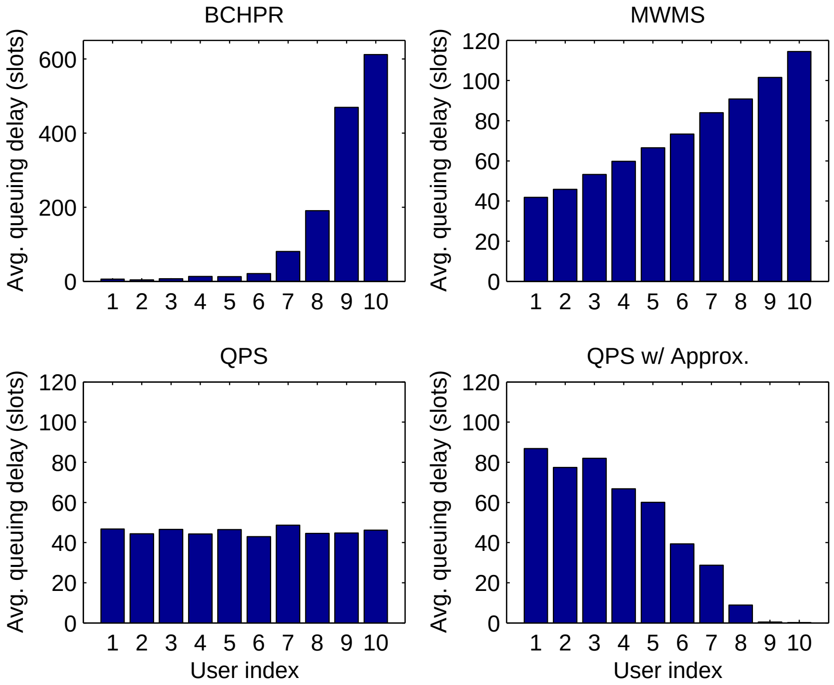


Figure 4.6: Each user's average queueing delay under four scheduling policies ($K = P = 10$, $M = 10$, user i's average SNR (dB) $= 20 - 0.5(i-1)$, and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

or to perform QPS that provides desirable delay and fairness properties as shown in Chapter 3. By using the degradedness of the BC on each tone, as well as duality relation between MAC and BC [35], this section shows that all these resource allocation problems in the OFDM MAC and BC can be formulated as GP problems. This GP perspective of multi-user OFDM resource allocation problems provides numerical efficiency as well as strong scalability for any additional constraints of GP form.

4.3.1 System Model and Problem Formulation

In this subsection, OFDM downlink and uplink system models are described, and the WSRmax, WSPmin, and PRmax problems are mathematically formulated. Considering a transmission system with K users and N tones where the base-station (BS) and each user are equipped with a single antenna, it is assumed that inter-symbol interference (ISI) is completely removed by exploiting OFDM techniques, i.e. the frequency response is flat within each tone. In the downlink case, total transmit power is constrained to P_{tot}, and in the uplink case, each user's individual power is constrained to P_i where i is the user index.

On user k's tone n, the channel gain is denoted by $H_k(n)$, and a zero-mean independent and identically distributed (i.i.d.) Gaussian noise with variance $\sigma_k^2(n)$ is added at the receiver. For the uplink case, $\sigma_k(n)$ is replaced with $\sigma(n)$ since the BS is the only receiver. The channel SNR for user k's tone n is defined as $g_k(n) = |H_k(n)|^2/\sigma_k^2(n)$, and $r_k(n)$ and $p_k(n)$ denote rate and power allocation on user k's tone n. Perfect CSI is assumed at both the BS and each user, which enables dynamic allocation of power and rate on each tone according to channel conditions. Multiple users are allowed to share each tone, and the base-station performs superposition coding in the downlink and successive decoding in the uplink. Fig. 4.7 summarizes OFDM BC and MAC models. Formulations of each resource allocation problem in OFDM BC and MAC are presented in the next two subsections.

Resource Allocation Problems for OFDM BC

In the downlink, the base station encodes multi-user messages using superposition coding with a proper encoding order. Also, each receiver performs successive decoding with a decoding order identical to the encoding order. It can be assumed that the ordering is the same on every tone, which is shown to be sufficient for achieving the overall capacity region [44]. Let $\pi(\cdot)$ denote the message-encoding order at the base-station, where $\pi(i) < \pi(j)$ means that user i's message is encoded earlier than user j's message. With superposition coding, one user can remove the interference caused by other users' messages encoded earlier. Therefore, the rate for user k's tone n is represented as

$$r_k(n) = \frac{1}{2}\log_2\left(1 + \frac{p_k(n)g_k(n)}{1 + g_k(n)\sum_{i:\pi(i)>\pi(k)} p_i(n)}\right). \tag{4.27}$$

First, WSRmax problem can be formulated as follows.

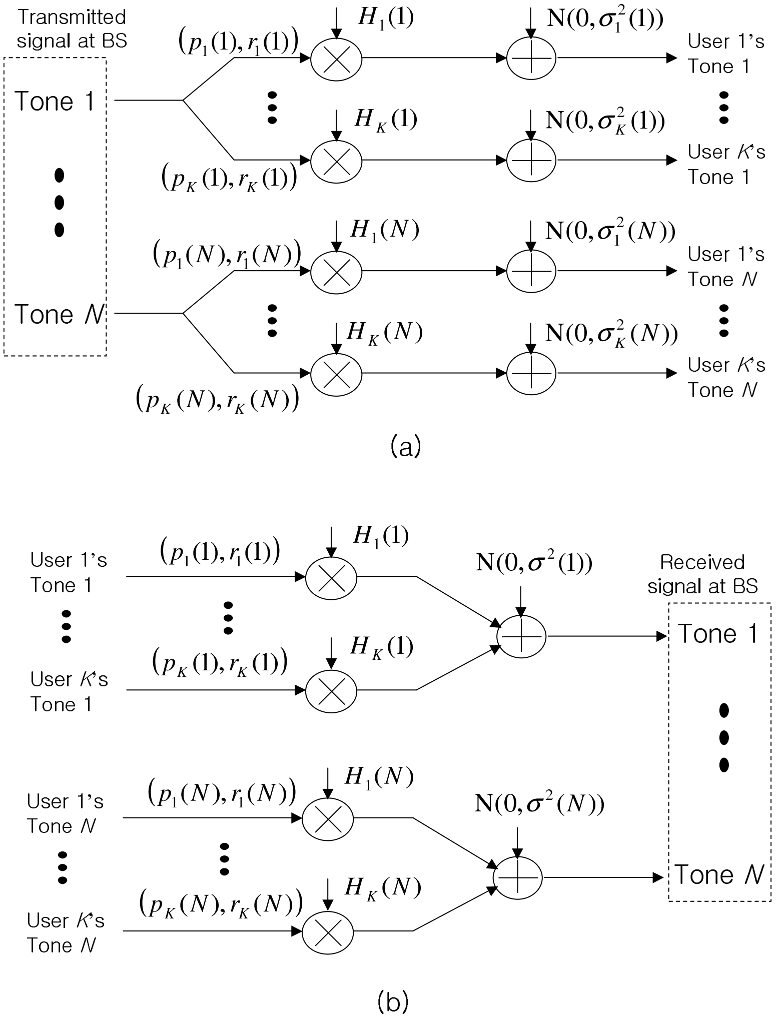


Figure 4.7: (a) OFDM BC model. (b) OFDM MAC model.

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{K} \mu_k \sum_{n=1}^{N} r_k(n) \\ \text{subject to} \quad & \sum_{k=1}^{K} \sum_{n=1}^{N} p_k(n) \leq P_{tot} \\ & p_k(n) \geq 0 \ \ \forall k \ \ \text{and} \ \ \forall n, \end{aligned} \tag{4.28}$$

where $\mu_k \geq 0$ is the weight on rates assigned to user k. Under the total power constraint, this problem finds the optimal power and rate allocation maximizing the weighted sum rate. The boundary surface of achievable rate region in BC or MAC can be traced by solving WSRmax for all possible weight vectors. WSRmax becomes equivalent to MWMS if the weight vector is replaced by the current queue-state vector. A dual version of WSRmax is WSPmin, which finds the rate and power allocation minimizing the weighted sum power with minimum rate constraints on each user. In the downlink, transmit power comes from a single source at the base station. Thus, sum-power minimization (SPmin) problem is of particular interest in BC, which is formulated as

$$\begin{aligned} \text{minimize} \quad & \sum_{k=1}^{K} \sum_{n=1}^{N} p_k(n) \\ \text{subject to} \quad & \sum_{n=1}^{N} r_k(n) \geq R_k \ \ \forall k \\ & p_k(n) \geq 0 \ \ \forall k \ \ \text{and} \ \ \forall n, \end{aligned} \tag{4.29}$$

where R_k is user k's minimum rate constraint. The third problem is PRmax that maximizes the sum rate while maintaining a preset ratio of each user's data rate:

$$\text{maximize} \quad \sum_{k=1}^{K} \sum_{n=1}^{N} r_k(n)$$

$$\begin{aligned}\text{subject to}\quad & \sum_{k=1}^{K}\sum_{n=1}^{N} p_k(n) \le P_{tot} \\ & p_k(n) \ge 0 \;\; \forall k \;\; \text{and} \;\; \forall n \\ & \sum_{n=1}^{N} r_1(n) : \cdots : \sum_{n=1}^{N} r_K(n) = \gamma_1 : \cdots : \gamma_K,\end{aligned} \tag{4.30}$$

where $\{\gamma_i\}_{i=1}^K$ is a set of non-negative values that defines the proportional fairness among users. Under a total power constraint, the boundary point of achievable rate region that satisfies the given proportional fairness is found by solving PRmax. If γ_k is replaced with the current queue length of user k for $k = 1, \cdots, K$, PRmax is equivalent to QPS for the stationary channel, and its extension to the time-varying channels is straightforward from Section 4.2.

Resource Allocation Problems for OFDM MAC

In the uplink case, the base station performs successive decoding with interference cancellation, in which each user's message is successively decoded and subtracted from the received signal. As in the downlink, the same ordering can be assumed over the tones without reducing achievable rates. Let $\pi(\cdot)$ denote the decoding order at the base-station, where $\pi(i) < \pi(j)$ means that user i's message is decoded earlier than user j's message. Then, the rate for user k's tone n is represented as

$$r_k(n) = \frac{1}{2}\log_2\left(1 + \frac{p_k(n)g_k(n)}{1 + \sum_{i:\pi(i)>\pi(k)} p_i(n)g_i(n)}\right). \tag{4.31}$$

Using this definition of $r_k(n)$, formulations of WSRmax and PRmax in the MAC are the same as those in the BC except for the power constraint. In the BC, the total power constraint is considered, but each user has an individual power constraint in the MAC. Thus, the total power constraint, $\sum_{k=1}^K \sum_{n=1}^N p_k(n) \le P_{tot}$, is replaced with individual power constraints, $\sum_{n=1}^N p_k(n) \le P_k$ for all k in WSRmax and PRmax for the MAC.

Compared with SPmin in the BC, WSPmin in the MAC includes the weight on each user's power in the objective. Therefore, $\sum_{k=1}^K \sum_{n=1}^N p_k(n)$ in (4.29) is replaced with $\sum_{k=1}^{K} \lambda_k \sum_{n=1}^{N} p_k(n)$ where $\lambda_k \geq 0$ is the weight on power assigned to user k. Other than this change in the objective, all the constraints are identical in both cases.

4.3.2 Optimal Resource Allocation via GP

In this subsection, WSRmax, WSPmin and PRmax problems for the OFDM BC and MAC are formulated as GP problems. The GP formulation of OFDM resource-allocation problems is closely related to the message encoding and decoding order. According to [44], the optimal ordering for various OFDM resource-allocation problems is identical across every tone, which implies that $K!$ possible orderings exist regardless of the number of tones. In the downlink, each tone's channel forms a degraded BC where the largest rate region is achieved by encoding the user with a higher channel SNR later [20]. The next subsection shows that after determining tone-dependent optimal orderings on every tone, WSRmax, SPmin, and PRmax in the BC can be converted into GP problems. The optimal power/rate allocation obtained by solving GP must conform to the optimal ordering that is one of the $K!$ tone-independent orderings. From the duality relation between the BC and its dual MAC, WSRmax, WSPmin, and PRmax in the MAC can also be solved via GP.

GP Formulations for OFDM BC

In the OFDM BC, the achievable rate region of tone n can be represented as

$$C_{BC}\left(\mathbf{m}(n), \sum_{k=1}^{K} p_k(n)\right) = \{r_{\pi_n(i)}(n) : r_{\pi_n(i)}(n) \leq$$
$$\frac{1}{2}\log\left(1 + \frac{p_{\pi_n(i)}(n)}{m_{\pi_n(i)}(n) + \sum_{j<i} p_{\pi_n(j)}(n)}\right), i = 1, \cdots, K\}, \tag{4.32}$$

where the effective noise variance of user k's tone n, $m_k(n) = 1/g_k(n)$, $\mathbf{m}(n) = [m_1(n), \cdots, m_K(n)]^T$, and $\pi_n(\cdot)$ is the permutation at tone n such that $m_{\pi_n(1)}(n) < m_{\pi_n(2)}(n) < \cdots < m_{\pi_n(K)}(n)$. That is, $\pi_n(\cdot)$ is in order of decreasing channel SNRs on tone n, which is reverse to the encoding order providing the largest rate region. When $\mathbf{r}(n) = [r_1(n), \cdots, r_K(n)]^T$ is on the boundary of the capacity region, solving

$p_{\pi_n(i)}(n)$'s in terms of the rate vector $\mathbf{r}(n)$ yields the following equations:

$$\begin{aligned}\sum_{i=1}^{l} p_{\pi_n(i)}(n) = \sum_{i=1}^{l} \big(m_{\pi_n(i)}(n) - m_{\pi_n(i-1)}(n)\big) \\ \times \exp\left(2\ln 2 \sum_{j=i}^{l} r_{\pi_n(j)}(n)\right) - m_{\pi_n(l)}(n),\ l = 1, \cdots, K,\end{aligned} \tag{4.33}$$

where $m_{\pi_n(0)}(n) \equiv 0$. As shown in [39], (4.32) becomes equivalent to

$$\begin{aligned} C_{BC}\left(\mathbf{m}(n), \sum_{k=1}^{K} p_k(n)\right) = \{r_{\pi_n(i)}(n) : \\ \sum_{i=1}^{K} \big(m_{\pi_n(i)}(n) - m_{\pi_n(i-1)}(n)\big) \exp\left(2\ln 2 \sum_{j=i}^{K} r_{\pi_n(j)}(n)\right) \\ \leq \sum_{k=1}^{K} p_k(n) + m_{\pi_n(K)}(n),\ r_i(n) \geq 0,\ i = 1, \cdots, K\}. \end{aligned} \tag{4.34}$$

From the above relations, the WSRmax problem given in (4.28) can be converted into the following GP:

$$\begin{aligned} \text{minimize} \quad & \log\exp\left(-\sum_{k=1}^{K} \mu_k \sum_{n=1}^{N} r_k(n)\right) \\ \text{subject to} \quad & \log\exp\left(-r_k(n)\right) \leq 0, \quad \forall\ k, n \\ & \log \sum_{n=1}^{N}\sum_{k=1}^{K} \left(\frac{m_{\pi_n(k)}(n) - m_{\pi_n(k-1)}(n)}{P_{tot} + \sum_{l=1}^{N} m_{\pi_l(K)}(l)}\right) \\ & \times \exp\left(2\ln 2 \sum_{i=k}^{K} r_{\pi_n(i)}(n)\right) \leq 0, \end{aligned} \tag{4.35}$$

where the optimization variables are $r_k(n)$'s. Once the optimal rate allocation is found, the corresponding power allocation can be derived using (4.33).

The SPmin in (4.29) can also be formulated via GP since the optimal ordering on each tone is the one achieving the largest rate region.

$$\begin{aligned}
\text{minimize} \quad & \log \sum_{n=1}^{N} \sum_{k=1}^{K} \left(m_{\pi_n(k)}(n) - m_{\pi_n(k-1)}(n) \right) \\
& \times \exp \left(2 \ln 2 \sum_{i=k}^{K} r_{\pi_n(i)}(n) \right) \\
\text{subject to} \quad & \log \exp\left(-r_k(n)\right) \leq 0, \quad \forall\ k, n \\
& \log \left(\exp(R_k) \exp(-\sum_{n=1}^{N} r_k(n)) \right) \leq 0 \ \ \forall\ k.
\end{aligned} \tag{4.36}$$

Similarly, PRmax in (4.30) can be converted into the following GP form:

$$\begin{aligned}
\text{minimize} \quad & \log \exp \left(- \sum_{k=1}^{K} \sum_{n=1}^{N} r_k(n) \right) \\
\text{subject to} \quad & \log \exp\left(-r_k(n)\right) \leq 0, \ \ \forall\ k, n \\
& \log \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\frac{m_{\pi_n(k)}(n) - m_{\pi_n(k-1)}(n)}{P_{tot} + \sum_{l=1}^{N} m_{\pi_l(K)}(l)} \right) \\
& \times \exp \left(2 \ln 2 \sum_{i=k}^{K} r_{\pi_n(i)}(n) \right) \leq 0 \\
& x\boldsymbol{\gamma} - \sum_{n=1}^{N} \mathbf{r}(n) = 0,
\end{aligned} \tag{4.37}$$

where $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_K]^T$. Define the optimization variable as $\mathbf{y} = [x\, \mathbf{r}(1)^T \ \cdots \ \mathbf{r}(N)^T]^T \in \mathbb{R}^{(KN+1)\times 1}$. With the optimal solution $\mathbf{y}^* = [x^*\ \mathbf{r}^*(1)^T \ \cdots \ \mathbf{r}^*(N)^T]^T$, the vector $x^*\boldsymbol{\gamma}$ denotes each user's allocated data rate.

GP Formulations for OFDM MAC

By using duality relationship between BC and MAC, the results obtained for the downlink can be extended for GP formulations of WSRmax, WSPmin and PRmax in the uplink. Given a BC, its dual MAC has the channel SNRs and a total power constraint that are the same as in the original BC. [35] showed that any points in the BC capacity region can be also achieved in its dual MAC if the decoding order of the dual MAC receiver is reversed with respect to the encoding order in the BC transmitter. Since the total power for both channels is identical, the rate allocation that minimizes sum power in the MAC can be solved via GP of the dual BC by using (4.36). From the above argument on ordering, the decoding order on tone n in the MAC is equal to the permutation $\pi_n(\cdot)$ defined in the previous subsection. Once the optimal rate allocation for SPmin is obtained by solving GP, the corresponding power allocation in the MAC can be determined from the following equation.

$$p_{\pi_n(k)}(n) = \frac{(2^{2r_{\pi_n(k)}(n)} - 1) \cdot 2^{2\sum_{i=k+1}^{K} r_{\pi_n(i)}(n)}}{g_{\pi_n(k)}(n)}, \ \forall \ k, n, \tag{4.38}$$

where $r_{\pi_n(K+1)}(n) \equiv 0$ for all n. This equation is derived by applying the tone-dependent optimal ordering to (4.31).

In the uplink case, each user has different power source so that the WSPmin problem is more useful than SPmin. With general non-equal weights, the tone-dependent optimal ordering can be different from that defined in SPmin. However, by utilizing channel scaling method, the optimal ordering for WSPmin in the MAC can be easily determined, and this problem becomes solvable via GP as well. The scaled power is $p'_k(n) = \lambda_k p_k(n)$, where λ_k is the weight on user k's power. Then, close observation of (4.31) reveals that if the channel SNR is also scaled such that $g'_k(n) = g_k(n)/\lambda_k$, the mutual information in terms of scaled powers and channel SNRs remains the same as that before scaling [13]. Therefore, WSPmin in the MAC converts into SPmin in terms of $p'_k(n)$ and $g'_k(n)$, which is solved via GP.

GP formulation of WSRmax in the MAC is not straightforward compared to other problems so far. The optimal tone-independent ordering is automatically determined from the given weight vector, but this ordering doesn't guarantee the feasibility of GP formulations because of the individual power constraints. By employing Lagrange dual decomposition, it is shown that WSRmax can be solved via iterative GP. First, WSRmax in the MAC converts into the minimization problem by multiplying -1 and then taking the exponential of the objective. The Lagrangian of this problem is defined over domain $\mathcal{D}$ as

$$\begin{aligned}\mathcal{L}(\{p_k(n)\},\{r_k(n)\},\boldsymbol{\lambda}) = \exp&\left(-\sum_{k=1}^{K}\mu_k\sum_{n=1}^{N}r_k(n)\right)\\ &+\sum_{k=1}^{K}\lambda_k\left(\sum_{n=1}^{N}p_k(n)-P_k\right),\end{aligned} \tag{4.39}$$

where $\boldsymbol{\lambda} \succeq 0$ and the domain $\mathcal{D}$ is defined as the set of all non-negative $p_k(n)$'s for all k and n. Then, the Lagrange dual function is represented as

$$f(\boldsymbol{\lambda}) = \min_{\{p_k(n)\},\{r_k(n)\}\in\mathcal{D}} \mathcal{L}(\{p_k(n)\},\{r_k(n)\},\boldsymbol{\lambda}). \tag{4.40}$$

For a fixed $\boldsymbol{\lambda}$, the minimization problem in (4.40) can be formulated via GP as the following: First, the scaled power is $p'_k(n) = \lambda_k p_k(n)$ and the scaled channel SNR is $g'_k(n) = g_k(n)/\lambda_k$. Then, in terms of $p'_k(n)$ and $g'_k(n)$, the minimization of Lagrangian in (4.39) is equivalent to maximizing the weighted sum rate and minimizing the sum power simultaneously. In the dual BC, the optimal encoding order on each tone for WSRmax and SPmin is equal to the order of increasing scaled channel SNR. From this reasoning, (4.40) can be converted into GP as follows.

$$\begin{aligned}\text{minimize}\quad & \log\Bigg(\exp\left(-\sum_{k=1}^{K}\mu_k\sum_{n=1}^{N}r_k(n)\right)\\ & +\sum_{n=1}^{N}\sum_{k=1}^{K}\left(m'_{\pi'_n(k)}(n)-m'_{\pi'_n(k-1)}(n)\right)\\ & \times\exp\left(2\ln 2\sum_{i=k}^{K}r_{\pi'_n(i)}(n)\right)\Bigg)\\ \text{subject to}\quad & \log\exp\left(-r_k(n)\right)\le 0,\quad \forall\ k,n,\end{aligned} \tag{4.41}$$

where $m'_k(n) = 1/g'_k(n)$, and $\pi'_n(\cdot)$ is the permutation at tone n such that $m'_{\pi'_n(1)}(n) < m'_{\pi'_n(2)}(n) < \cdots < m'_{\pi'_n(K)}(n)$, or $\pi'_n(\cdot)$ is in order of decreasing scaled channel SNRs on tone n. With the optimal rate and power allocation obtained by this GP, $f(\boldsymbol{\lambda})$ can be derived from (4.39).

Finally, the dual optimal solution is obtained by maximizing $f(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda} \succeq 0$. The original WSRmax in the MAC is a convex optimization problem with zero duality gap; thus, the dual optimal objective always equals the primal optimal objective [10]. This maximization can be done by iterating the following steps until each user's power converges to individual power constraint: find $f(\boldsymbol{\lambda})$ via GP for a fixed $\boldsymbol{\lambda}$, and update $\boldsymbol{\lambda}$ to the direction of increasing $f(\boldsymbol{\lambda})$. $\boldsymbol{\lambda}$ can be efficiently updated by using an ellipsoid method, a type of sub-gradient search methods for non-differentiable functions. This method converges in $\mathcal{O}(n^2)$ iterations where n is the number of variables [10]. The details of the ellipsoid method are provided in Appendix C. A sub-gradient for $f(\boldsymbol{\lambda})$ required in the ellipsoid method is $d_k = \sum_{n=1}^{N} p_k^*(n) - P_k$ for all k, where $\{p_k^*(n)\}$ optimizes the minimization problem in the definition of $f(\boldsymbol{\lambda})$.

PRmax in the MAC can be also converted into iterative GP following similar steps as in WSRmax. For PRmax, the weighted sum rate is replaced with the sum rate, and the constraint on proportional fairness is added when the Lagrangian is minimized. Considering these changes, GP formulation of PRmax in the MAC becomes

$$\begin{array}{ll} \text{minimize} & \log\left(\exp\left(-\sum_{k=1}^{K}\sum_{n=1}^{N} r_k(n)\right)\right. \\ & +\sum_{n=1}^{N}\sum_{k=1}^{K}\left(m'_{\pi'_n(k)}(n) - m'_{\pi'_n(k-1)}(n)\right) \\ & \left.\times \exp\left(2\ln 2\sum_{i=k}^{K} r_{\pi'_n(i)}(n)\right)\right) \\ \text{subject to} & \log\exp\left(-r_k(n)\right) \le 0, \quad \forall\ k, n \\ & x\boldsymbol{\gamma} - \sum_{n=1}^{N}\mathbf{r}(n) = 0, \end{array} \tag{4.42}$$

where the optimization variable is the same as in (4.37).

4.3.3 Numerical Results and Discussion

This subsection provides some simulation results generated using GP formulations for multi-user OFDM resource-allocation problems. Fig. 4.8 presents two achievable

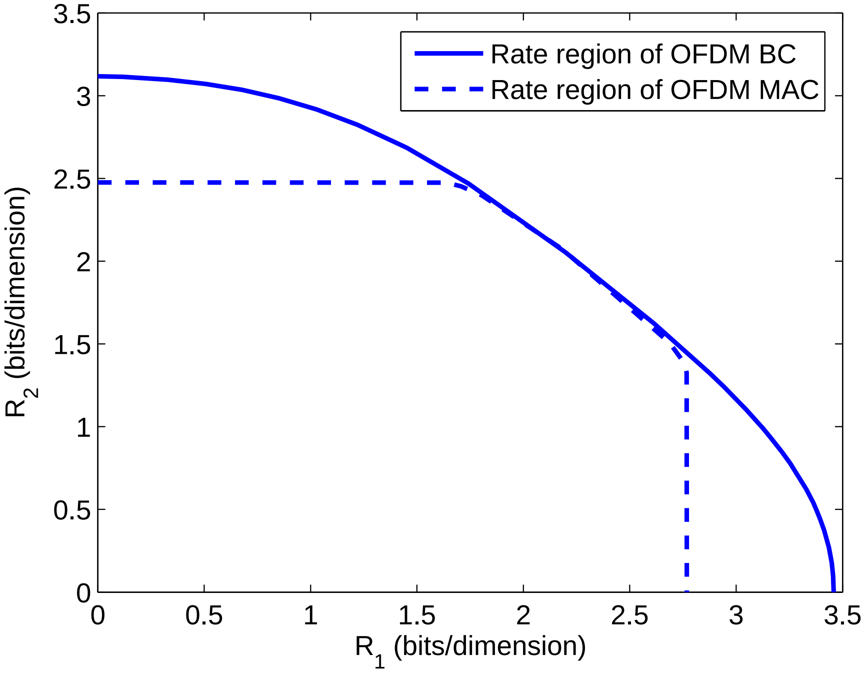


Figure 4.8: Rate regions of OFDM BC and MAC ($N = 64$, $K = 2$, $P_{tot} = NK = 128$ in BC, $P_1 = P_2 = \frac{P_{tot}}{2} = 64$ in MAC. Each user's average channel SNR per tone = 10 dB)

rate regions of OFDM BC and MAC where $N = 64$, $K = 2$, $P_{tot} = NK = 128$ in BC, $P_1 = P_2 = \frac{P_{tot}}{2} = 64$ in MAC. Channel SNRs are assumed to be i.i.d. exponentially distributed with each tone's average SNR of 10 dB. The same set of channel SNRs are used for both BC and MAC. In Fig. 4.8, boundary points of rate regions are obtained by solving WSRmax via GP for all possible weight vectors. Since $P_1 + P_2 = P_{tot}$ and since both OFDM BC and MAC have the same channel SNRs, the duality relation holds between these two channels. Therefore, both rate regions always share at least one boundary point, which can be observed in Fig. 4.8.

Fig. 4.9 illustrates the power region for the same OFDM MAC as in Fig. 4.8, with the target rate vector of $\mathbf{R} = [2.05 \;\; 2.19]^T$ bits per dimension. Boundary points of power region are characterized by solving WSPmin via GP for all possible weight vectors. The given target rate vector is a boundary point shared by both OFDM BC

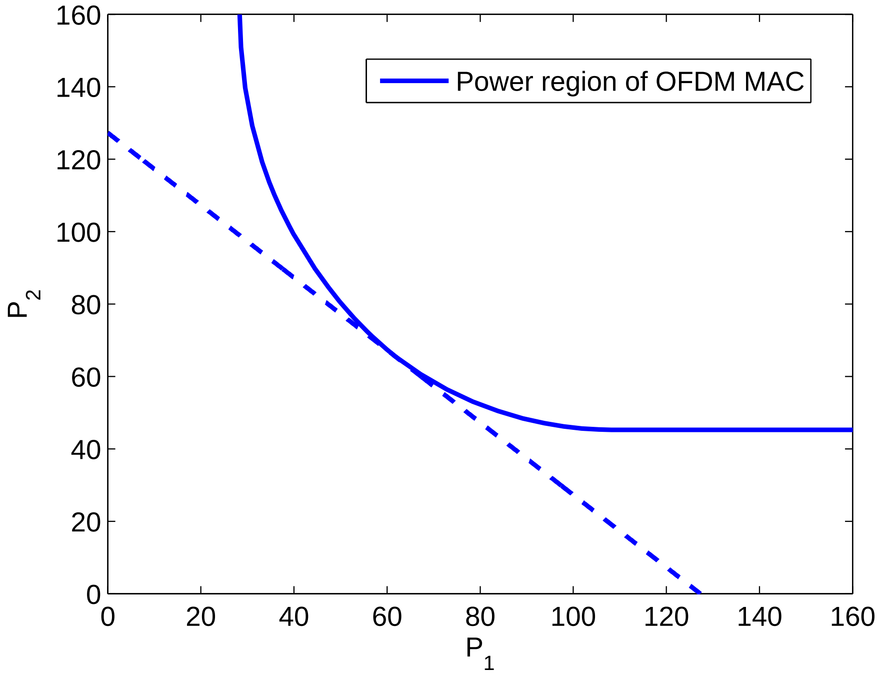


Figure 4.9: Power region of OFDM MAC ($N = 64$, $K = 2$, target rates $\mathbf{R} = [2.05 \;\; 2.19]^T$(bits/dim), and the channel SNRs are the same as in Fig. 4.8.)

and MAC in Fig. 4.8. Thus, as can be seen in Fig. 4.9, the minimum sum power required to support these target rates is equal to the total power used in Fig. 4.8.

4.4 Summary

In fading broadcast channels, the geometric programming (GP) formulation of QPS, which is also applicable to OFDM systems, is presented. GP is a special form of convex optimization problems with well-developed efficient algorithms. Stochastic simulations performed by solving formulated GP problems demonstrate that QPS provides significantly smaller average queueing delay compared to other scheduling policies such as MWMS for any arrival rate vector within the network capacity region, and it exhibits more desirable fairness property.

Furthermore, three major resource allocation problems in both downlink and uplink OFDM systems are formulated via GP: weighted sum-rate maximization, weighted sum-power minimization, and proportional-rate maximization. Without violating GP structure, a variety of rate constraints can be added, which is essential in performing cross-layer resource allocation to satisfy each user's various QoS requirement. In multi-user OFDM systems, GP emerges as a powerful tool that provides high numerical efficiency as well as strong scalability.

Chapter 5

Lagrange Dual Decomposition for MIMO-OFDMA

One popular realization of multi-user OFDM systems is called orthogonal frequency division multiple access (OFDMA), which assigns each subchannel or tone to at most one user [31]-[14]. When a single cell environment is assumed, there is no multiuser interference at each tone owing to this FDMA (Frequency Division Multiple Access) constraint inherent in OFDMA systems. The spectral efficiency can be further increased by employing multiple antennas at base stations and terminals in rich-scattering environments [68, 22]. Such multiple-input multiple-output (MIMO) systems enable a dramatic increase in capacity known as spatial multiplexing gain. This chapter addresses MIMO-OFDMA systems where each tone is occupied by at most one user, and the assigned user occupies the MIMO channel formed at the corresponding tone as a one-to-one communication link.

If the instantaneous CSI is available at the transmitter via a reliable feedback link, the transmitter of MIMO-OFDMA systems can dynamically allocate power and rate on each tone and each transmit antenna to satisfy each user's QoS demand, which is essential in multi-user communication systems. For a single-user, single-tone MIMO systems with CSIT, the MIMO channel capacity is achieved by multiplying precoding and post-processing matrices at transmitter and receiver, respectively, based on

SVD (Singular Value Decomposition) of channel matrix. The effective MIMO channel becomes equivalent to orthogonal parallel channels where each channel gain takes the singular value. Thus, the capacity is achieved by waterfilling the transmit power across these parallel effective channels [68, 50]. On the other hand, MIMO-OFDMA systems have FDMA constraints that make the optimal tone assignment a combinatorial problem with the exponential complexity in the number of tones [31]. For the single-input single-output (SISO) OFDMA systems, much previous work has considered convex relaxation methods by introducing time-sharing or frequency-sharing variables for efficient suboptimal solution [51, 14]. However, this approach employs a different system model from the original OFDMA system. Thus, it eventually requires a heuristic approximation that might lead to a significant suboptimality in some cases.

On the other hand, Yu and Lui [80] showed that in multi-carrier applications, even though the original resource allocation problems are non-convex, the duality gap becomes zero as the number of tones goes to infinity. Therefore, with a very large number of tones, Lagrange dual-decomposition methods can be used to find the optimal solutions accurately. This argument is based on the fact that if the optimal value of an optimization problem is a concave (or convex) function of the constraint vector, the duality gap is zero regardless of convexity of the original problem. With infinite dimensions, arbitrary time-sharing or frequency-sharing become feasible in resource allocation, which enables this condition to be satisfied. In this chapter, motivated by these results, downlink and uplink MIMO-OFDMA resource allocation problems are solved in the dual domain by using Lagrange dual decomposition, and efficient algorithms are developed for the following two major problems: weighted sum-rate maximization (WSRmax) and weighted sum-power minimization (WSPmin). The existing duality gap is actually evaluated, and the results show that with the practical number of tones, the optimal objective is virtually concave in terms of the constraint vector, which validates the proposed dual approach.

The organization of this chapter is as follows: Section 5.1 presents the system model and problem formulation. In Section 5.2, general theory on the duality gap of non-convex optimizations is introduced, and duality gap of MIMO-OFDMA problems is closely investigated in Section 5.3. Section 5.4 presents efficient resource allocation algorithms for downlink and uplink MIMO-OFDMA systems. Finally, numerical results are discussed in Section 5.5 and Section 5.6 summarizes this chapter.

5.1 System Model and Problem Formulation

This section describes downlink and uplink MIMO-OFDMA system models and formulates two resource allocation problems.

5.1.1 Downlink MIMO-OFDMA Systems

First, consider a downlink MIMO-OFDMA system with K users and N tones where the base station (BS) is equipped with t transmit antennas and K mobile terminals with $r_1, \cdots, r_K$ receive antennas, respectively. It is assumed that the inter-symbol interference (ISI) is completely removed by the cyclic prefix in OFDM techniques, i.e. the frequency response is flat within each tone. The total transmit power is constrained to P_{tot}. At user k's tone n, a MIMO channel is formed, which is given by

$$\mathbf{y}_k(n) = \mathbf{H}_k(n)\mathbf{x}_k(n) + \mathbf{z}_k(n), \tag{5.1}$$

where $\mathbf{y}_k(n) \in \mathbb{C}^{r_k \times 1}$, $\mathbf{H}_k(n) \in \mathbb{C}^{r_k \times t}$, and $\mathbf{x}_k(n) \in \mathbb{C}^{t \times 1}$ denote, respectively, the received signal vector, the channel matrix, and the transmitted signal vector at user k's tone n. $\mathbf{z}_k(n) \in \mathbb{C}^{r_k \times 1}$ is a vector of independent zero-mean complex Gaussian noise entries with variance 1/2 per real component at user k's receiver, i.e. $\mathbf{z}_k(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The MIMO matrix channel $\mathbf{H}_k(n)$ is assumed to be perfectly known to the transmitter and user k's receiver. Let S_i denote the set of tones allocated to user i. Because of the FDMA constraint in MIMO-OFDMA systems, each tone is allowed to be used by at most one user; hence, $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^{K} S_i \subseteq \{1, 2, \cdots, N\}$. Also, let the covariance matrix of the transmitted signal of user k's tone n be denoted by $\mathbf{S}_k(n) = \mathbb{E}[\mathbf{x}_k(n)\mathbf{x}_k(n)^H]$. Then, the total power constraint can be expressed as the following:

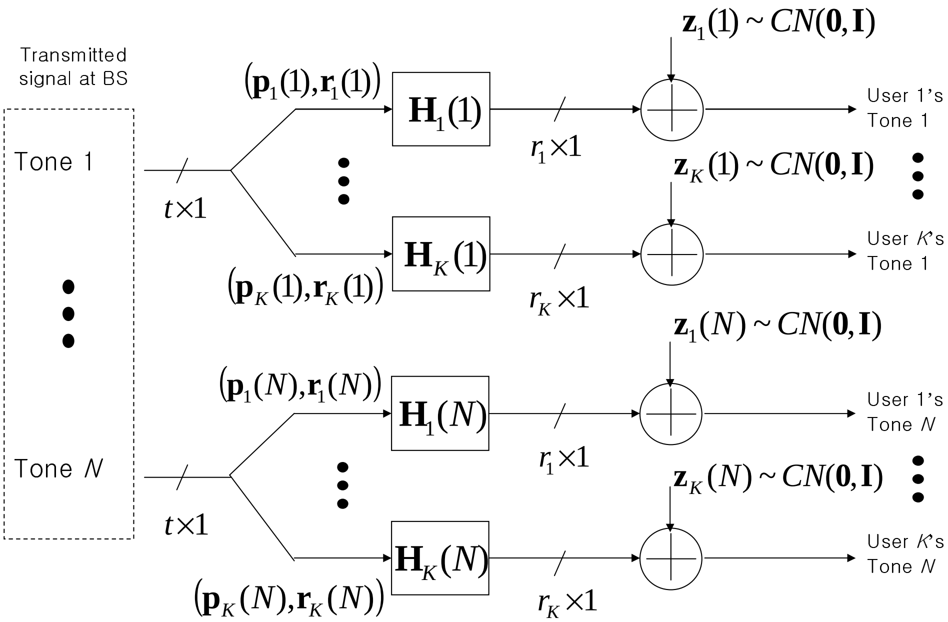


Figure 5.1: Block diagram of MIMO-OFDMA BC

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbf{Tr}\left(\mathbf{S}_k(n)\right)=\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{l=1}^{t}p_{k,l}(n)\leq P_{tot}, \tag{5.2}$$

where $p_{k,l}(n)$ denotes user k's power allocation on tone n and transmit antenna l. Given $\mathbf{S}_k(n)$, let $r_{k,l}(n)$ denote user k's rate allocation on tone n and transmit antenna l such that $\sum_{l=1}^{t} r_{k,l}(n)$ is equal to or less than the mutual information $\mathcal{I}(\mathbf{x}_k(n);\mathbf{y}_k(n))$ over the choice of the distribution of $\mathbf{x}_k(n)$. The block diagram of MIMO-OFDMA BC is illustrated in Fig. 5.1.

As shown in [68] and [50], the maximization of this MIMO mutual information can be achieved by using singular value decomposition (SVD). By applying SVD, the channel matrix $\mathbf{H}_k(n)$ can be written as

$$\mathbf{H}_k(n) = \mathbf{U}_k(n)\mathbf{\Sigma}_k(n)\mathbf{V}_k(n)^H, \tag{5.3}$$

where $\mathbf{U}_k(n)$ and $\mathbf{V}_k(n)$ are unitary matrices with dimension $r_k \times r_k$ and $t \times t$, respectively, and $\mathbf{\Sigma}_k(n)$ is a non-negative diagonal matrix with dimension $r_k \times t$. The diagonal entries of $\mathbf{\Sigma}_k(n)$ are $\sigma_{k,1}(n)^{1/2}, \ldots, \sigma_{k,L_k}(n)^{1/2}$, the non-negative square roots of the eigenvalues of $\mathbf{H}_k(n)\mathbf{H}_k(n)^H$, where $L_k = \min(r_k, t)$. Let $\tilde{\mathbf{y}}_k(n) = \mathbf{U}_k(n)^H \mathbf{y}_k(n)$, $\tilde{\mathbf{x}}_k(n) = \mathbf{V}_k(n)^H \mathbf{x}_k(n)$ and $\tilde{\mathbf{z}}_k(n) = \mathbf{U}_k(n)^H \mathbf{z}_k(n)$. Then, the original channel is equivalent to the following channel.

$$\tilde{\mathbf{y}}_k(n) = \mathbf{\Sigma}_k(n)\tilde{\mathbf{x}}_k(n) + \tilde{\mathbf{z}}_k(n), \tag{5.4}$$

where $\tilde{\mathbf{z}}_k(n)$ has the same distribution as $\mathbf{z}_k(n)$ and $\mathbb{E}[\tilde{\mathbf{x}}_k(n)\tilde{\mathbf{x}}_k(n)^H] = \mathbb{E}[\mathbf{x}_k(n)\mathbf{x}_k(n)^H]$. Thus, the one-to-one MIMO channel is decomposed to L_k independent subchannels:

$$\tilde{y}_{k,l}(n) = \sigma_{k,l}(n)^{1/2}\tilde{x}_{k,l}(n) + \tilde{z}_{k,l}(n), \ 1 \leq l \leq L_k. \tag{5.5}$$

Consequently, the rate allocation $r_{k,l}(n)$ satisfies the following equality

$$\sum_{l=1}^{t} r_{k,l}(n) = \sum_{l=1}^{L_k} \log_2(1 + p_{k,l}(n)\sigma_{k,l}(n)). \tag{5.6}$$

If $\mathbf{Tr}\left(\mathbf{S}_k(n)\right)$ is assumed to be constrained to $P_{sum,k}(n)$, the maximum value of $\sum_{l=1}^{t} r_{k,l}(n)$ is

$$\max_{p_{k,l}(n)\geq 0, \sum_{l=1}^{t} p_{k,l}(n) \leq P_{sum,k}(n)} \sum_{l=1}^{t} r_{k,l}(n) = \sum_{l=1}^{L_k} [\log_2(\mu_{k,l}(n)\sigma_{k,l}(n))]^+, \tag{5.7}$$

where $(x)^+$ denotes $\max\{x, 0\}$, the optimal power allocation $p_{k,l}^*(n) = (\mu_k(n) - \sigma_{k,l}(n)^{-1})^+$, and $\mu_k(n)$ satisfies $\sum_{l=1}^{L_k}(\mu_k(n) - \sigma_{k,l}(n)^{-1})^+ = P_{sum,k}(n)$. In other words, when $\mathbf{Tr}\left(\mathbf{S}_k(n)\right)$ is given, the optimal power allocation on each transmit antenna of user k's tone n is water-filling over the channel eigenvalues.

Utilizing the above results, the WSRmax problem in downlink MIMO-OFDMA systems can be formulated as

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{K} \mu_k \sum_{n \in S_k} \sum_{l=1}^{t} r_{k,l}(n) \\ \text{subject to} \quad & \sum_{k=1}^{K} \sum_{n \in S_k} \sum_{l=1}^{t} p_{k,l}(n) \leq P_{tot}, \\ & S_i \cap S_j = \emptyset \;\; \forall i \neq j, \\ & \bigcup_{k=1}^{K} S_k \subseteq \{1, 2, \cdots, N\}, \\ & p_{k,l}(n) \geq 0 \;\; \forall k, n, \text{and } l, \end{aligned} \tag{5.8}$$

where the relation between the rate and power allocation is as defined in (5.13), and $\mu_k \geq 0$ is the weight assigned to user k. Given the weight vector and the channel matrices, the solution of this problem finds the power allocation that maximizes the weighted sum rate with total power constraint. The boundary of the achievable rate region can be traced by solving this problem for all possible weight vectors $\boldsymbol{\mu}$. In general, (5.8) is not a convex optimization problem since it finds the optimal set of tones for each user, which is a combinatorial problem whose complexity increases exponentially with N. This argument also holds for the following WSPmin problem which is a dual problem of WSRmax:

$$\begin{aligned} \text{minimize} \quad & \sum_{k=1}^{K} \lambda_k \sum_{n \in S_k} \sum_{l=1}^{t} p_{k,l}(n) \\ \text{subject to} \quad & \sum_{n \in S_k} \sum_{l=1}^{t} r_{k,l}(n) \geq R_k \;\; \forall k, \\ & S_i \cap S_j = \emptyset \;\; \forall i \neq j, \\ & \bigcup_{k=1}^{K} S_k \subseteq \{1, 2, \cdots, N\}, \\ & p_{k,l}(n) \geq 0 \;\; \forall k, n, \text{and } l, \end{aligned} \tag{5.9}$$

where $\lambda_k \geq 0$ is the weight assigned to user k. Given the weight vector and the channel matrices, this problem finds a power distribution that minimizes the weighted sum power with minimum rate constraint on each user. In the downlink, the inter-cell interference can be reduced by solving the WSPmin problem. This problem is of particular interest in the uplink case since the battery life of the mobile terminal is critical. Because of the FDMA nature of MIMO-OFDMA systems, the optimal solution of WSPmin in the broadcast channel (BC) is equivalent to that in its dual multiple access channel (MAC) where the role of transmitter and receivers in the BC is reversed.

5.1.2 Uplink MIMO-OFDMA Systems

This subsection presents the system model and problem formulation for uplink MIMO-OFDMA systems. The major differences from downlink MIMO-OFDMA systems are that each mobile terminal has its own power constraint P_k. Consider an uplink MIMO-OFDMA system with K users and N tones where the base station (BS) is equipped with r receive antennas and K mobile terminals with $t_1, \cdots, t_K$ transmit antennas, respectively. At user k's tone n, a MIMO channel is formed, which is represented as

$$\mathbf{y}_k(n) = \mathbf{H}_k(n)\mathbf{x}_k(n) + \mathbf{z}_k(n), \tag{5.10}$$

where $\mathbf{y}_k(n) \in \mathbb{C}^{r\times 1}$, $\mathbf{H}_k(n) \in \mathbb{C}^{r\times t_k}$, and $\mathbf{x}_k(n) \in \mathbb{C}^{t_k\times 1}$ denote, respectively, the received signal vector, the channel matrix, and the transmitted signal vector at user k's tone n. $\mathbf{z}_k(n) \in \mathbb{C}^{r\times 1}$ is a vector of independent zero-mean complex Gaussian noise entries with variance 1/2 per real component at the BS, i.e. $\mathbf{z}_k(n) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The channel matrix $\mathbf{H}_k(n)$ is assumed to be perfectly known to the receiver and corresponding user's transmitter. Then, user k's power constraint can be given as

$$\sum_{n=1}^{N} \mathbf{Tr}\left(\mathbf{S}_k(n)\right) = \sum_{n=1}^{N}\sum_{l=1}^{t_k} p_{k,l}(n) \leq P_k. \tag{5.11}$$

The block diagram of MIMO-OFDMA MAC is illustrated in Fig. 5.2.

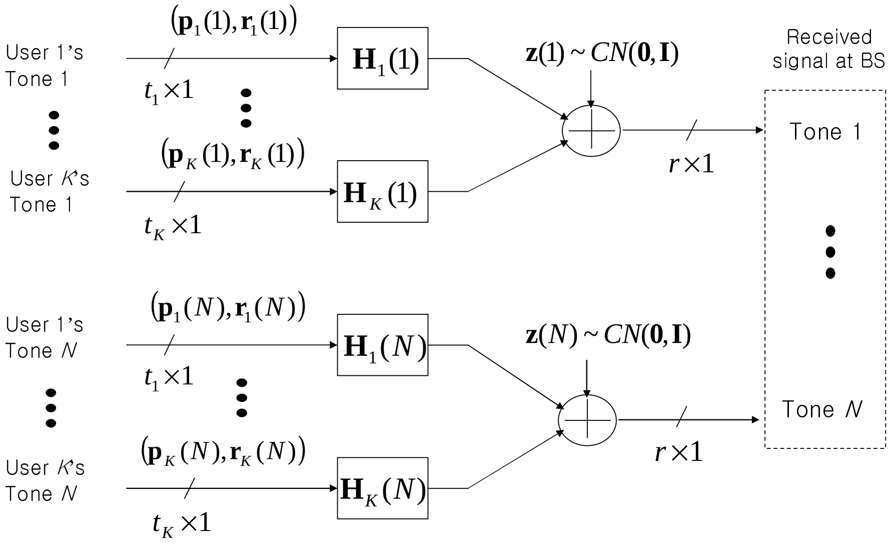


Figure 5.2: Block diagram of MIMO-OFDMA MAC

By using SVD as in the downlink case, the uplink one-to-one MIMO channel is decomposed to M_k independent subchannels as

$$\tilde{y}_{k,l}(n) = \sigma_{k,l}(n)^{1/2}\tilde{x}_{k,l}(n) + \tilde{z}_{k,l}(n),\ 1 \leq l \leq M_k, \tag{5.12}$$

where $M_k = \min(t_k, r)$ and other symbols denote the same quantities as in the downlink case with only changes in dimension: $L_k \to M_k$, $t \to t_k$, and $r_k \to r$. Consequently, the rate allocation $r_{k,l}(n)$ satisfies the following equality

$$\sum_{l=1}^{t_k} r_{k,l}(n) = \sum_{l=1}^{M_k} \log_2(1 + p_{k,l}(n)\sigma_{k,l}(n)). \tag{5.13}$$

If $\mathbf{Tr}\,(\mathbf{S}_k(n))$ is assumed to be constrained to a certain value, the optimal power allocation on each transmit antenna of user k's tone n is water-filling over the channel eigenvalues as shown in the previous subsection.

Formulation of WSRmax in the MAC is the same as in the BC except for power constraints. Total power constraint, $\sum_{k=1}^{K}\sum_{n\in S_k}\sum_{l=1}^{t} p_{k,l}(n) \leq P_{tot}$, is considered in the BC. On the other hand, in the MAC, each user has an individual power constraint, $\sum_{n\in S_k}\sum_{l=1}^{t_k} p_{k,l}(n) \leq P_k$ for all k. As mentioned in the previous subsection, the optimal solution of WSPmin in the MAC is equivalent to that in its dual BC where the roles of transmitters and receiver in the MAC are reversed. Therefore, by solving (5.9) with the number of transmit and user k's receive antenna, r and t_k, respectively, the power and rate allocation for WSPmin can be obtained in the uplink MIMO-OFDMA systems.

The next section shows that the duality gap for each of the aforementioned non-convex problems is virtually negligible with realistic number of tones, which makes it possible to develop efficient algorithms by using Lagrange dual decomposition.

5.2 General Theory on Duality Gap

This section introduces some conditions under which the duality gap is zero for general non-convex optimization problems in multi-tone systems. With N tones and K users, the optimization problem has the following general form.

$$\begin{aligned} \text{maximize} \quad & \sum_{n=1}^{N} f_n(\mathbf{x}_n) \\ \text{subject to} \quad & \sum_{n=1}^{N} \mathbf{h}_n(\mathbf{x}_n) \preceq \mathbf{P}, \end{aligned} \tag{5.14}$$

where $\mathbf{x}_n \in \mathbb{R}^K$ are vectors of optimization variables; $f_n(\cdot)$ are $\mathbb{R}^K \rightarrow \mathbb{R}$ functions, which are not necessarily concave; and $\mathbf{h}_n(\cdot)$ are $\mathbb{R}^K \rightarrow \mathbb{R}^L$ functions that are not necessarily convex. Constant $\mathbf{P}$ is an L-vector of constraints. The Lagrangian of (5.14) is defined as

$$\mathcal{L}(\{\mathbf{x}_n\}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} f_n(\mathbf{x}_n) + \boldsymbol{\lambda}^T \left(\mathbf{P} - \sum_{n=1}^{N} \mathbf{h}_n(\mathbf{x}_n) \right), \tag{5.15}$$

where $\boldsymbol{\lambda}$ is a vector of Lagrange dual variables. The dual objective $g(\boldsymbol{\lambda})$ is defined as an unconstrained maximization of the Lagrangian such that $g(\boldsymbol{\lambda}) = \max_{\{\mathbf{x}_n\}} \mathcal{L}(\{\mathbf{x}_n\}, \boldsymbol{\lambda})$. Then, the dual optimization problem becomes

$$\begin{aligned} \text{minimize} \quad & g(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \succeq 0. \end{aligned} \tag{5.16}$$

From duality theory, $g^* \geq f^*$ where f^* and g^* are primal and dual optimal values, respectively. The duality gap d^* is defined as $d^* = g^* - f^*$. When $f_n(\mathbf{x}_n)$'s are concave and $\mathbf{h}_n(\mathbf{x}_n)$'s are convex, (5.15) is a convex optimization problem, which guarantees zero duality gap. Zero duality gap implies that the globally optimal solution can be obtained by using Lagrange dual decomposition. More fundamentals of Lagrange dual decomposition and duality gap are provided in Appendix B. Though the above optimization problem in (5.15) is non-convex, duality gap is zero if either of the following two conditions is satisfied [80, 16].

Theorem 6. *If $\mathbf{x}_n^*(\boldsymbol{\lambda}) = \arg\max_{\mathbf{x}_n} \mathcal{L}(\{\mathbf{x}_n\}, \boldsymbol{\lambda})$, as a function of $\boldsymbol{\lambda}$, is continuous at $\boldsymbol{\lambda}^*$, the duality gap equals zero.*

Theorem 7. *Concavity of the optimal $\Sigma_n f_n$ in $\mathbf{P}$ implies zero duality gap.*

The condition in Theorem 6 is sufficient for that in Theorem 7 but the converse is not always true. Recently, [80] shows that in non-convex multi-carrier optimization problems with the general form of (5.15), the concavity condition in Theorem 7 is always satisfied when the number of tones goes to infinity. However, existing duality gap for a problem with practical number of tones cannot be estimated from this argument. In the next section, duality gap of MIMO-OFDMA resource allocation problems is closely investigated.

5.3 Duality Gap of Non-convex Optimizations

According to [80], more dimensions in resource-allocation problems for multi-carrier systems result in higher likelihood of satisfying the time-sharing condition that guarantees zero duality gap. The extension from SISO-OFDMA to MIMO-OFDMA introduces spatial multiplexing capability, which is effectively similar to adding more parallel channels. Thus, the duality gap for MIMO-OFDMA is generally smaller than that for SISO-OFDMA. Also, from the dual relation between the BC and MAC [35], the solutions of WSRmax and WSPmin in the BC are directly linked to those in the MAC*. This section analyzes the duality gap for the WSRmax and WSPmin problems in downlink SISO-OFDMA systems.

First, consider the downlink WSRmax problem given in (5.8). For any fixed subchannel assignment, $\{S_k\}$, the optimal solution of this problem can be obtained by multi-level water-filling [31] that is given as follows.

$$\begin{aligned} p_{k,l}(n) &= \begin{cases} \left(\mu_k Y - \frac{1}{\sigma_{k,l}(n)}\right)^+ & \text{if } n \in S_k, \\ 0 & \text{if } n \notin S_k. \end{cases} \\ Y &= \frac{P_{tot} + \sum_{k=1}^{K}\sum_{n\in S_k}\sum_{\{l:p_{k,l}(n)>0\}}^{L_k}\frac{1}{\sigma_{k,l}(n)}}{\sum_{k=1}^{K}\sum_{n\in S_k}\sum_{\{l:p_{k,l}(n)>0\}}^{L_k}\mu_k}. \end{aligned} \tag{5.17}$$

Finding the optimal subchannel assignment requires K^N searches. Hence, in terms of K and N, the overall optimization requires $\mathcal{O}(NK^N)$ operations†, which is exponentially complex. In general, the optimal subchannel allocation can change as total power varies, which may destroy the concavity of the optimal objective function in terms of total power. A simple example for this argument is illustrated for downlink SISO-OFDMA systems in Fig. 5.3 when $N = 2$, $K = 2$, $\boldsymbol{\mu} = [1 \;\; 2]^T$, and each user's channel SNR vectors are $[10 \;\; 160]^T$ and $[160 \;\; 10]^T$. The maximum weighted sum rate

*For MIMO-OFDMA BC and MAC, WSPmin is basically an identical problem.

†The complexity order for performing SVD on an $r \times t$ channel matrix is $\mathcal{O}(\min(rt^2, tr^2))$. Thus, the overall complexity order becomes $\mathcal{O}(\min(rt^2, tr^2)NK^N)$. This SVD complexity in terms of the number of transmit and receive antennas is ignored throughout this chapter since it is commonly applied to every case.

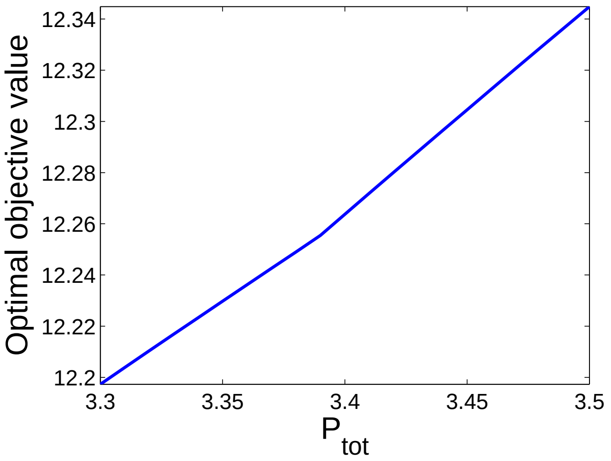


Figure 5.3: Maximum weighted sum rate in SISO-OFDMA BC versus P_{tot} ($K = 2$, $N = 2$, $\boldsymbol{\mu} = [1 \;\; 2]^T$ and channel SNR vectors are $[10 \;\; 160]^T$ and $[160 \;\; 10]^T$)

is plotted for $P_{tot} = 3.3 \sim 3.5$. At $P_{tot} = 3.39$, the optimal subchannel assignment changes from $S_1 = \{2\}, S_2 = \{1\}$ to $S_1 = \emptyset, S_2 = \{1, 2\}$. Since each user's water-level is different, a discrete change in the slope occurs at the transition point, which breaks down the concavity at $P_{tot} = 3.39$. With the same subchannel allocation, the optimal weighted sum rate is concave in total power. However, whenever the optimal set of tones changes, a sudden jump in the slope appears, which might make the curve non-concave with that total power. As the number of tones grows, changes in the optimal subchannel allocation occur more frequently when the sum power varies.

On the other hand, the degree of the discrete slope change tends to decrease with more tones since the bandwidth affected by each set change becomes narrower. Thus, the overall curve is expected to be more concave as the number of tones increases. Fig. 5.4 illustrates the maximum weighted sum rate versus total power in the SISO-OFDMA BC when $N = 8$, $K = 2$, $\boldsymbol{\mu} = [1 \;\; 2]^T$, and each user's channel SNR vectors are $\frac{1}{\sigma^2}[1^2 \; 2^2 \; \cdots \; N^2]^T$ and $\frac{1}{\sigma^2}[N^2 \; (N-1)^2 \; \cdots \; 1^2]^T$. σ^2 denotes the noise power at each tone. As P_{tot} sweeps from 0 to 16, changes in the optimal subchannel allocation occur at least five times on each of three plots in Fig. 5.4. However, discrete slope changes are almost undetectable in this figure. Hence, in practical downlink MIMO-OFDMA systems with more than a hundred tones, the duality gap of WSRmax in

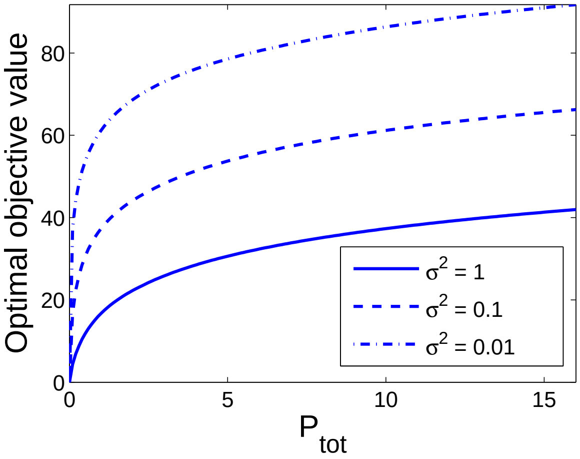


Figure 5.4: Maximum weighted sum rate in SISO-OFDMA BC versus P_{tot} ($K = 2$, $N = 8$, $\boldsymbol{\mu} = [1 \;\; 2]^T$ and channel SNR vectors are $\frac{1}{\sigma^2}[1^2 \; 2^2 \; \cdots \; N^2]^T$ and $\frac{1}{\sigma^2}[N^2 \; (N-1)^2 \; \cdots \; 1^2]^T$)

the MIMO-OFDMA BC is expected to be virtually zero, and the optimal solution can be derived in the dual domain.

The optimal solution for the WSPmin problem in (5.9) can be obtained by the following steps: First, choose a subchannel assignment, and for each user, distribute enough power over its assigned tones in a water-filling fashion to satisfy its rate constraint R_k. User k's power distribution for the given tone assignment, S_k, can be formulated as follows:

$$\begin{aligned} p_{k,l}(n) &= \begin{cases} \left(M_k - \frac{1}{\sigma_{k,l}(n)}\right)^+ & \text{if } n \in S_k, \\ 0 & \text{if } n \notin S_k. \end{cases} \\ M_k &= \left(\frac{2^{R_k}}{\prod_{n\in S_k}\prod_{\{l:p_{k,l}(n)>0\}}^{L_k} \sigma_{k,l}(n)}\right)^{1/\left(\sum_{n\in S_k}\sum_{\{l:p_{k,l}(n)>0\}}^{L_k} 1\right)}. \end{aligned} \tag{5.18}$$

Second, after iterating the first step for all K^N possible set selections, pick one of them that minimizes the weighted sum power (WSP). Therefore, WSPmin also requires $\mathcal{O}(NK^N)$ operations. If the optimal WSP is a convex function of the constraint vector $\mathbf{R}$, the duality gap will be zero from the condition in Theorem 7. Similar to WSRmax

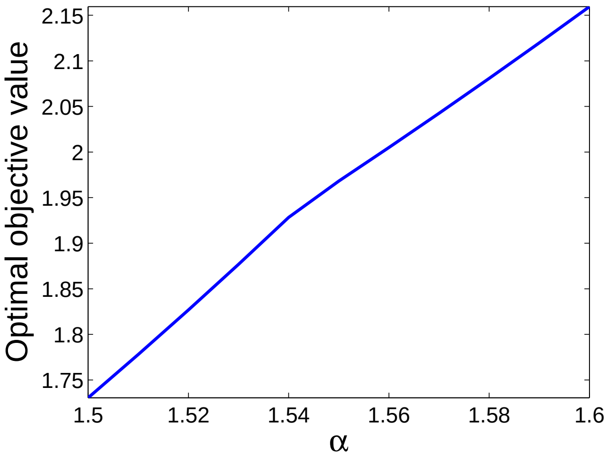


Figure 5.5: Minimum weighted sum power in SISO-OFDMA BC versus rate constraints ($K = 2$, $N = 2$, $\boldsymbol{\lambda} = [1 \;\; 2]^T, \mathbf{R} = \alpha[2 \;\; 1]^T$ (bits/symbol) and channel SNR vectors are $[40 \;\; 160]^T$ and $[10 \;\; 90]^T$)

case, Fig. 5.5 shows that the convexity of the minimum WSP for SISO-OFDMA BC may no longer hold when the optimal subchannel assignment changes. In this figure, $N = 2$, $K = 2$, $\boldsymbol{\lambda} = [1 \;\; 2]^T$, the rate constraint vector is $\mathbf{R} = \alpha[2 \;\; 1]^T$ (bits/symbol), and user 1 and 2's channel SNR vectors are $[40 \;\; 160]^T$ and $[10 \;\; 90]^T$. When α varies from 1.5 to 1.6, the optimal subchannel allocation changes from $S_1 = \{1\}, S_2 = \{2\}$ to $S_1 = \{2\}, S_2 = \{1\}$. This change causes sudden jump in the slope of the curve, which results in non-convexity at this transition point of $\alpha = 1.54$. However, from the same argument provided in WSRmax case, the amount of slope change decreases when the number of tones rises as demonstrated in Fig. 5.6 where $N = 8$, $K = 2$, $\boldsymbol{\lambda} = [1 \;\; 2]^T, \mathbf{R} = \alpha[2 \;\; 1]^T$ (bits/symbol), and channel SNR vectors are the same as those defined in Fig. 5.4. When α sweeps from 1 to 10, discrete slope changes seem to be negligible in this figure. Therefore, in practice, the WSPmin problem in the downlink and uplink MIMO-OFDMA systems can be solved in the dual domain with much less computational complexity.

Based on the results in this section, Lagrange dual decomposition can be used to derive efficient algorithms for both WSRmax and WSPmin problems as shown in the following section.

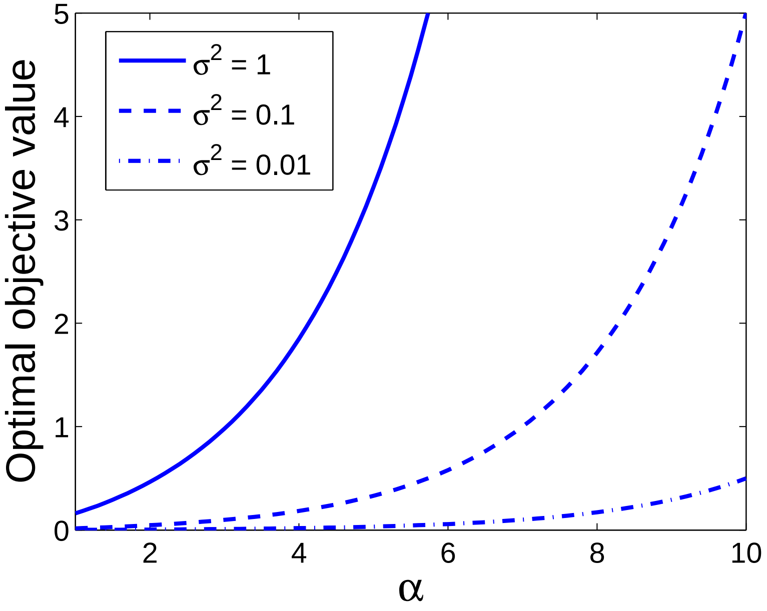


Figure 5.6: Minimum weighted sum power in SISO-OFDMA BC versus rate constraints ($K = 2$, $N = 8$, $\boldsymbol{\lambda} = [1 \ \ 2]^T, \mathbf{R} = \alpha[2 \ \ 1]^T$ (bits/symbol) and channel SNR vectors are the same as those defined in Fig. 5.4.)

5.4 Efficient Resource Allocation Algorithms

In this section, very efficient power and rate allocation algorithms are developed by using Lagrange dual decomposition for solution of the WSRmax and WSPmin problems in downlink and uplink MIMO-OFDMA systems.

5.4.1 Downlink MIMO-OFDMA Systems

The Lagrangian of WSRmax problem in (5.8) is defined over domain $\mathcal{D}$ as

$$\begin{aligned}\mathcal{L}(\{p_{k,l}(n)\},\{r_{k,l}(n)\},\lambda) &= \sum_{k=1}^{K}\mu_k\sum_{n=1}^{N}\sum_{l=1}^{t} r_{k,l}(n)\\ &\quad -\lambda\left(\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{l=1}^{t} p_{k,l}(n)-P_{tot}\right), \qquad (5.19)\end{aligned}$$

where the domain $\mathcal{D}$ is defined as the set of all non-negative $p_{k,l}(n)$'s for $k=1,\cdots,K$, $n=1,\cdots,N$, and $l=1,\cdots,t$ such that for each n, only one user can have positive power allocation from the FDMA constraint. Then, the Lagrange dual function is

$$g(\lambda)=\max_{\{p_{k,l}(n)\}\in\mathcal{D}}\mathcal{L}(\{p_{k,l}(n)\},\{r_{k,l}(n)\},\lambda). \qquad (5.20)$$

Equation (5.19) suggests that the maximization of $\mathcal{L}$ can be decomposed into the following N independent optimization problems

$$g_n'(\lambda)=\max_{\{p_{k,l}(n)\}\in\mathcal{D}}\left\{\sum_{k=1}^{K}\mu_k\sum_{l=1}^{t} r_{k,l}(n)-\lambda\sum_{k=1}^{K}\sum_{l=1}^{t} p_{k,l}(n)\right\}, \qquad (5.21)$$

for $n=1,\cdots,N$. Then, the Lagrange dual function becomes

$$g(\lambda)=\sum_{n=1}^{N} g_n'(\lambda)+\lambda P_{tot}. \qquad (5.22)$$

Assume user k is active on tone n. With a fixed λ, the object of the max operation in (5.21) is a concave function of $p_{k,l}(n)$. By taking the derivative of this object regarding $p_{k,l}(n)$, the next optimality condition is obtained, which maximizes $g'_n(\lambda)$.

$$p_{k,l}(n) = \left(K_k - \frac{1}{\sigma_{k,l}(n)}\right)^+, \quad l = 1, 2, \cdots, L_k, \tag{5.23}$$

where $K_k = \mu_k/(\lambda \log 2)$ and $L_k = \min(t, r_k)$. In case of $L_k < t$, $p_{k,l}(n) = 0$ for $L_k < l \leq t$. Initialize $S_k = \emptyset$ for $k = 1, \cdots, K$. By searching over all K possible user assignments for tone n, $g'_n(\lambda)$ can be obtained as

$$\begin{aligned} g'_n(\lambda) = \max_k \Bigg\{ & \mu_k \sum_{l=1}^{L_k} \log_2 \left(1 + \left(K_k - \frac{1}{\sigma_{k,l}(n)}\right)^+ \sigma_{k,l}(n)\right) \\ & - \lambda \sum_{l=1}^{L_k} \left(K_k - \frac{1}{\sigma_{k,l}(n)}\right)^+ \Bigg\}, \quad n = 1, \cdots, N. \end{aligned} \tag{5.24}$$

For tone n, if user u is associated with the value of $g'_n(\lambda)$ in (5.24), $S_u \cup \{n\} \rightarrow S_u$, and $p_{k,l}(n)$ for $k \neq u$ is set to zero.

Once the above equation is solved for all n, the overall Lagrange dual function $g(\lambda)$ is derived from (5.22). Finally, it is required to find $\lambda^* \geq 0$ that minimizes $g(\lambda)$. The update of λ can be done by using a simple bisection method until the sum power converges [10]. Hence, in terms of the number of users and tones, $\mathcal{O}(NK)$ executions are required to find the optimal solution, which shows the linear complexity of the proposed algorithm in N. If the converged sum power is equal to the total power constraint, the duality gap is zero and this solution is in fact globally optimal. From (5.24), the user selection at tone n can change at some level of λ where a quantum leap may occur in the sum power. Thus, if P_{tot} is within this gap, the sum power cannot converge to P_{tot} by using above bisection method on λ. However, the previous section shows that the duality gap quickly vanishes as the number of tones increases, and the solution obtained in the dual domain becomes a globally optimal solution. Therefore, the subchannel assignment at $\lambda = \lambda^*$ can be assumed to be optimal, and the global optimal solution can be found by doing multi-level water-filling with this set. The algorithm for solution of the WSRmax problem can be summarized as follows:

Algorithm 1: WSRmax in downlink MIMO-OFDMA with CSIT

```
$\lambda_{\min} = 0, \lambda_{\max} = \delta N$ where $\delta$ is sufficiently large
**While** $\lambda_{\max} - \lambda_{\min} > \epsilon_1$
    $\lambda = (\lambda_{\max} + \lambda_{\min})/2$
    Find $\{p_{k,l}(n)\}$ and $\{S_k\}$ by solving (5.23) and (5.24).
    **If** $\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{l=1}^{t} p_{k,l}(n) \leq P_{tot}$, then
        $\lambda_{\max} = \lambda$
    **Else**, $\lambda_{\min} = \lambda$
    **End If**
**End While**
**If** $P_{tot} - \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{l=1}^{t} p_{k,l}(n) > \epsilon_2$, then
    With the obtained $\{S_k\}$, perform multi-level water-filling by (5.17).
**End If**
**Return** $\{p_{k,l}(n)\}$ and $\{S_k\}$
```

Similarly, the WSPmin problem can be also solved by using dual decomposition. The Lagrangian of WSPmin problem in (5.9) is defined over domain $\mathcal{D}$ as

$$\begin{aligned}\mathcal{L}(\{p_{k,l}(n)\}, \{r_{k,l}(n)\}, \boldsymbol{\mu}) = &\sum_{k=1}^{K} \lambda_k \sum_{n=1}^{N} \sum_{l=1}^{t} p_{k,l}(n) \\ &- \sum_{k=1}^{K} \mu_k \left(\sum_{n=1}^{N} \sum_{l=1}^{t} r_{k,l}(n) - R_k \right).\end{aligned} \tag{5.25}$$

Then, the Lagrange dual function is represented as

$$g(\boldsymbol{\mu}) = \min_{\{p_{k,l}(n)\} \in \mathcal{D}} \mathcal{L}(\{p_{k,l}(n)\}, \{r_{k,l}(n)\}, \boldsymbol{\mu}). \tag{5.26}$$

From (5.25), the minimization of $\mathcal{L}$ can be decomposed into N independent optimization problems as follows

$$g'_n(\boldsymbol{\mu}) = \min_{\{p_{k,l}(n)\}\in\mathcal{D}} \left\{ \sum_{k=1}^{K} \lambda_k \sum_{l=1}^{t} p_{k,l}(n) - \sum_{k=1}^{K} \mu_k \sum_{l=1}^{t} r_{k,l}(n) \right\}, \tag{5.27}$$

for $n = 1, \cdots, N$. Thus, the Lagrange dual function is

$$g(\boldsymbol{\mu}) = \sum_{n=1}^{N} g'_n(\boldsymbol{\mu}) + \sum_{k=1}^{K} \mu_k R_k. \tag{5.28}$$

With a fixed $\boldsymbol{\mu}$, the object of min operation in (5.27) is a convex function of $p_{k,l}(n)$. Hence, taking the derivative of this object regarding $p_{k,l}(n)$ results in the following condition, which minimizes $g'_n(\boldsymbol{\mu})$:

$$p_{k,l}(n) = \left(M_k - \frac{1}{\sigma_{k,l}(n)} \right)^+, \quad l = 1, 2, \cdots, L_k, \tag{5.29}$$

where $M_k = \mu_k / (\lambda_k \log 2)$. In case of $L_k < t$, $p_{k,l}(n) = 0$ for $L_k < l \leq t$. Initialize $S_k = \emptyset$ for $k = 1, \cdots, K$. By searching over all K possible user assignments for tone n, $g'_n(\boldsymbol{\mu})$ is obtained as

$$\begin{aligned} g'_n(\boldsymbol{\mu}) = \min_k \Bigg\{ & \lambda_k \sum_{l=1}^{L_k} \left(M_k - \frac{1}{\sigma_{k,l}(n)} \right)^+ \\ & -\mu_k \sum_{l=1}^{L_k} \log_2 \left(1 + \left(M_k - \frac{1}{\sigma_{k,l}(n)} \right)^+ \sigma_{k,l}(n) \right) \Bigg\}, \end{aligned} \tag{5.30}$$

for $n = 1, \cdots, N$. At tone n, if user u is associated with the value of $g'_n(\boldsymbol{\mu})$ in (5.30), $S_u \cup \{n\} \rightarrow S_u$, and $p_{k,l}(n)$ for $k \neq u$ is set to zero.

After solving (5.30) for all n, $g(\boldsymbol{\mu})$ is derived from (5.28). Finally, the dual optimal solution is obtained by maximizing $g(\boldsymbol{\mu})$ over non-negative μ_k's. Though $g(\boldsymbol{\mu})$ is concave, a search method based on gradient is infeasible since the dual function is not differentiable. However, the search direction for non-differentiable functions can be found by using subgradient-type methods. Suppose $\boldsymbol{\mu}^*$ maximizes $g(\boldsymbol{\mu})$[‡], a vector

[‡] In case the goal of optimization is to maximize the objective function like WSRmax, suppose

$\mathbf{d}$ is called a subgradient of $g(\boldsymbol{\mu})$ at $\boldsymbol{\mu}$ if and only if $\boldsymbol{\mu}^*$ can not lie in the half-space $\{\boldsymbol{\mu}' : \sum_{k=1}^{K} d_k(\mu'_k - \mu_k) \geq 0\}$. For the WSPmin problem, the subgradient must satisfy $g(\boldsymbol{\mu}') \leq g(\boldsymbol{\mu}) - \sum_{k=1}^{K} d_k(\mu'_k - \mu_k)$ for any $\boldsymbol{\mu}' \succeq 0$. A subgradient of this problem is derived in the following proposition.

Proposition 1. *For the WSPmin problem with a dual objective $g(\boldsymbol{\mu})$ defined in (5.26), the following choice of $\mathbf{d}$ is a subgradient for $g(\boldsymbol{\mu})$:*

$$d_k = \sum_{n=1}^{N} \sum_{l=1}^{t} r^*_{k,l}(n) - R_k \qquad k = 1, \cdots, K, \tag{5.31}$$

where $\{r^*_{k,l}(n)\}$ and $\{p^*_{k,l}(n)\}$ optimize the minimization problem in the definition of $g(\boldsymbol{\mu})$.

Proof. Since $\{r^*_{k,l}(n)\}$ and $\{p^*_{k,l}(n)\}$ are already in $\mathcal{D}$, for any $\boldsymbol{\delta} \succeq 0$,

$$\begin{aligned} g(\boldsymbol{\delta}) &\leq \mathcal{L}(\{p^*_{k,l}(n)\}, \{r^*_{k,l}(n)\}, \boldsymbol{\delta}) \\ &= g(\boldsymbol{\mu}) - \sum_{k=1}^{K} (\delta_k - \mu_k) \left(\sum_{n=1}^{N} \sum_{l=1}^{t} r^*_{k,l}(n) - R_k \right). \end{aligned} \tag{5.32}$$

□

The update of $\boldsymbol{\mu}$ is efficiently performed with the ellipsoid method until every user's rate converges [80, 44]. The ellipsoid method is one efficient sub-gradient search method for updating the dual variables, and it is known to converge in $\mathcal{O}(n^2)$ iterations where n is the number of variables [10]. The details of ellipsoid method appear in Appendix C.

In terms of K and N, the overall optimization needs $\mathcal{O}(K^2)$ runs of optimization problem with the complexity of $\mathcal{O}(NK)$. Hence, $\mathcal{O}(NK^3)$ executions are required to find the optimal solution of WSPmin by using the proposed algorithm. As discussed in WSRmax case, the discontinuity in power allocation can happen at $\boldsymbol{\mu}^*$ for the WSPmin problem as well. In this situation, a solution finds the subchannel assignment at $\boldsymbol{\mu}^*$, and allocates power in a water-filling fashion to satisfy the rate constraint

$\boldsymbol{\mu}^*$ minimizes $g(\boldsymbol{\mu})$.

for each user. By doing so, the global optimal solution is derived from a dual domain in downlink MIMO-OFDMA systems. The algorithm for solution of the WSPmin problem can be summarized as follows:

Algorithm 2: WSPmin in downlink MIMO-OFDMA with CSIT

```
1: Select a large ellipsoid and assign $\boldsymbol{\mu}$ as the center of the initial ellipsoid
2: **While** (the volume of the ellipsoid) $> \epsilon_1$
3:      Find $\{p_{k,l}(n)\}$ and $\{S_k\}$ by solving (5.29) and (5.30).
4:      Update $\boldsymbol{\mu}$ by the ellipsoid method
5: **End While**
6: **If** $\sum_{k=1}^{K} \left| \sum_{n\in S_k} \sum_{\{l:p_{k,l}(n)>0\}}^{L_k} \log_2 \left( \frac{\mu_k \sigma_{k,l}(n)}{\lambda_k \log 2} \right) - R_k \right|^2 > \epsilon_2$, then
7:      With the obtained $\{S_k\}$, perform multi-level water-filling by (5.18).
8: **End If**
9: **Return** $\{p_{k,l}(n)\}$ and $\{S_k\}$
```

5.4.2 Uplink MIMO-OFDMA Systems

The Lagrangian of WSRmax problem for the uplink MIMO-OFDMA systems is defined over domain $\mathcal{D}$ as

$$\begin{aligned}\mathcal{L}(\{p_{k,l}(n)\}, \{r_{k,l}(n)\}, \boldsymbol{\lambda}) = &\sum_{k=1}^{K} \mu_k \sum_{n=1}^{N} \sum_{l=1}^{t_k} r_{k,l}(n) \\ &- \sum_{k=1}^{K} \left(\lambda_k \sum_{n=1}^{N} \sum_{l=1}^{t_k} p_{k,l}(n) - P_k \right), \end{aligned} \tag{5.33}$$

where the domain $\mathcal{D}$ is defined as the set of all non-negative $p_{k,l}(n)$'s for $k = 1, \cdots, K$, $n = 1, \cdots, N$, and $l = 1, \cdots, t_k$ such that for each n, only one user can have positive power allocation from the FDMA constraint. Then, the Lagrange dual function is

$$g(\boldsymbol{\lambda}) = \max_{\{p_{k,l}(n)\}\in\mathcal{D}} \mathcal{L}(\{p_{k,l}(n)\}, \{r_{k,l}(n)\}, \boldsymbol{\lambda}). \tag{5.34}$$

Equation (5.33) suggests that the maximization of $\mathcal{L}$ can be decomposed into the following N independent optimization problems

$$g_n'(\boldsymbol{\lambda}) = \max_{\{p_{k,l}(n)\}\in\mathcal{D}} \left\{ \sum_{k=1}^{K} \mu_k \sum_{l=1}^{t_k} r_{k,l}(n) - \sum_{k=1}^{K} \lambda_k \sum_{l=1}^{t_k} p_{k,l}(n) \right\}, \tag{5.35}$$

for $n = 1, \cdots, N$. Then, the Lagrange dual function becomes

$$g(\boldsymbol{\lambda}) = \sum_{n=1}^{N} g_n'(\boldsymbol{\lambda}) + \sum_{k=1}^{K} \lambda_k P_k. \tag{5.36}$$

Under the constraints that user k is active on tone n and that $\boldsymbol{\lambda}$ is fixed, the objective of the max operation in (5.35) is a concave function of $p_{k,l}(n)$. By taking the derivative of this objective with respect to $p_{k,l}(n)$, the next optimality condition is obtained, which maximizes $g_n'(\boldsymbol{\lambda})$:

$$p_{k,l}(n) = \left(K_{up,k} - \frac{1}{\sigma_{k,l}(n)} \right)^+, \quad l = 1, 2, \cdots, M_k, \tag{5.37}$$

where $K_{up,k} = \mu_k/(\lambda_k \log 2)$ and $M_k = \min(t_k, r)$. In case of $M_k < t_k$, $p_{k,l}(n) = 0$ for $M_k < l \leq t_k$. Initialize $S_k = \emptyset$ for $k = 1, \cdots, K$. By searching over all K possible user assignments for tone n, $g_n'(\boldsymbol{\lambda})$ can be obtained as

$$\begin{aligned} g_n'(\boldsymbol{\lambda}) = \max_k \Bigg\{ & \mu_k \sum_{l=1}^{M_k} \log_2 \left(1 + \left(K_{up,k} - \frac{1}{\sigma_{k,l}(n)} \right)^+ \sigma_{k,l}(n) \right) \\ & -\lambda_k \sum_{l=1}^{M_k} \left(K_{up,k} - \frac{1}{\sigma_{k,l}(n)} \right)^+ \Bigg\}, \quad n = 1, \cdots, N. \end{aligned} \tag{5.38}$$

For tone n, if user u is associated with the value of $g_n'(\boldsymbol{\lambda})$ in (5.38), $S_u \cup \{n\} \rightarrow S_u$, and $p_{k,l}(n)$ for $k \neq u$ is set to zero.

Once the above equation (5.38) is solved for all n, the overall Lagrange dual function $g(\boldsymbol{\lambda})$ is derived from (5.36). Final solution requires determination of $\boldsymbol{\lambda}^* \succeq 0$ that maximizes $g(\boldsymbol{\lambda})$. As in the previous subsection, an efficient update of $\boldsymbol{\lambda}$ uses the ellipsoid method until every user's power converges. A sub-gradient of this problem required for ellipsoid method is provided in the following proposition.

Proposition 2. *For the uplink WSRmax problem with a dual objective $g(\boldsymbol{\lambda})$ defined in (5.34), the following choice of $\mathbf{d}$ is a subgradient for $g(\boldsymbol{\lambda})$:*

$$d_k = P_k - \sum_{n=1}^{N} \sum_{l=1}^{t_k} p^*_{k,l}(n) \qquad k = 1, \cdots, K, \tag{5.39}$$

where $\{r^*_{k,l}(n)\}$ and $\{p^*_{k,l}(n)\}$ optimize the minimization problem in the definition of $g(\boldsymbol{\lambda})$.

Proof. Since $\{r^*_{k,l}(n)\}$ and $\{p^*_{k,l}(n)\}$ are already in $\mathcal{D}$, for any $\boldsymbol{\delta} \succeq 0$,

$$\begin{aligned} g(\boldsymbol{\delta}) &\geq \mathcal{L}(\{p^*_{k,l}(n)\}, \{r^*_{k,l}(n)\}, \boldsymbol{\delta}) \\ &= g(\boldsymbol{\lambda}) + \sum_{k=1}^{K} (\delta_k - \lambda_k) \left(P_k - \sum_{n=1}^{N} \sum_{l=1}^{t_k} p^*_{k,l}(n) \right). \end{aligned} \tag{5.40}$$

□

Thus, suppose $\boldsymbol{\lambda}^*$ minimizes $g(\boldsymbol{\lambda})$, $\boldsymbol{\lambda}^*$ can not lie in the half-space $\{\boldsymbol{\delta} : \sum_{k=1}^{K} d_k(\delta_k - \lambda_k) \geq 0\}$. In terms of K and N, the overall optimization needs $\mathcal{O}(K^2)$ runs of optimization problem with the complexity of $\mathcal{O}(NK)$. Hence, the proposed algorithm requires $\mathcal{O}(NK^3)$ executions to find the optimal solution of WSRmax in the uplink. As discussed in the downlink case, the discontinuity in power allocation can happen at $\boldsymbol{\lambda}^*$ for this problem as well. In this situation, a solution finds the subchannel assignment, $\{S_k\}$, at $\boldsymbol{\lambda}^*$ and allocates power in a water-filling fashion to satisfy each user's power constraint. This multi-level water-filling can be expressed as

$$\begin{aligned} p_{k,l}(n) &= \begin{cases} \left(Y_k - \frac{1}{\sigma_{k,l}(n)} \right)^+ & \text{if } n \in S_k, \\ 0 & \text{if } n \notin S_k. \end{cases} \\ Y_k &= \frac{P_k + \sum_{n \in S_k} \sum_{\{l : p_{k,l}(n) > 0\}}^{L_k} \frac{1}{\sigma_{k,l}(n)}}{\sum_{n \in S_k} \sum_{\{l : p_{k,l}(n) > 0\}}^{L_k} 1}. \end{aligned} \tag{5.41}$$

As a result, the global optimal solution is derived from a dual domain in uplink

MIMO-OFDMA systems. The algorithm for solution of the WSRmax problem in the uplink can be summarized as follows:

Algorithm 3: WSRmax in uplink MIMO-OFDMA with CSIT

```
1: Select a large ellipsoid and assign $\boldsymbol{\lambda}$ as the center of the initial ellipsoid
2: While (the volume of the ellipsoid) $> \epsilon_1$
3:     Find $\{p_{k,l}(n)\}$ and $\{S_k\}$ by solving (5.37) and (5.38).
4:     Update $\boldsymbol{\lambda}$ by the ellipsoid method
5: End While
6: If $\sum_{k=1}^{K} \left| P_k - \sum_{n\in S_k} \sum_{\{l:p_{k,l}(n)>0\}}^{L_k} p_{k,l}(n) \right|^2 > \epsilon_2$, then
7:     With the obtained $\{S_k\}$, perform multi-level water-filling by (5.41).
8: End If
9: Return $\{p_{k,l}(n)\}$ and $\{S_k\}$
```

As noted in 5.1, WSPmin in the MAC is equivalent to that in its dual BC where the only difference is the reversed role between transmitters and receiver.

5.5 Numerical Results and Discussion

This section provides some simulation results generated by using proposed efficient resource allocation algorithms for MIMO-OFDMA BC and MAC. Fig. 5.7 and Fig. 5.8 show achievable rate and power regions of SISO-OFDMA BC with $N = 8$ and $K = 2$. The user 1 and 2's channel SNR vectors are $10[1^2\ 2^2\ \cdots\ N^2]^T$ and $10[N^2\ (N-1)^2\ \cdots\ 1^2]^T$, respectively. Each region is generated by using both optimal exhaustive search and Lagrange dual-decomposition methods. Fig. 5.7 illustrates the achievable rate region when $P_{tot} = NK = 16$. The boundary points are characterized by solving WSRmax for all possible weight vectors. In this figure, the rate region obtained by employing Lagrange dual decomposition is indistinguishable from the optimal rate region, which implies zero duality gap in this case. Fig. 5.8 shows the achievable power

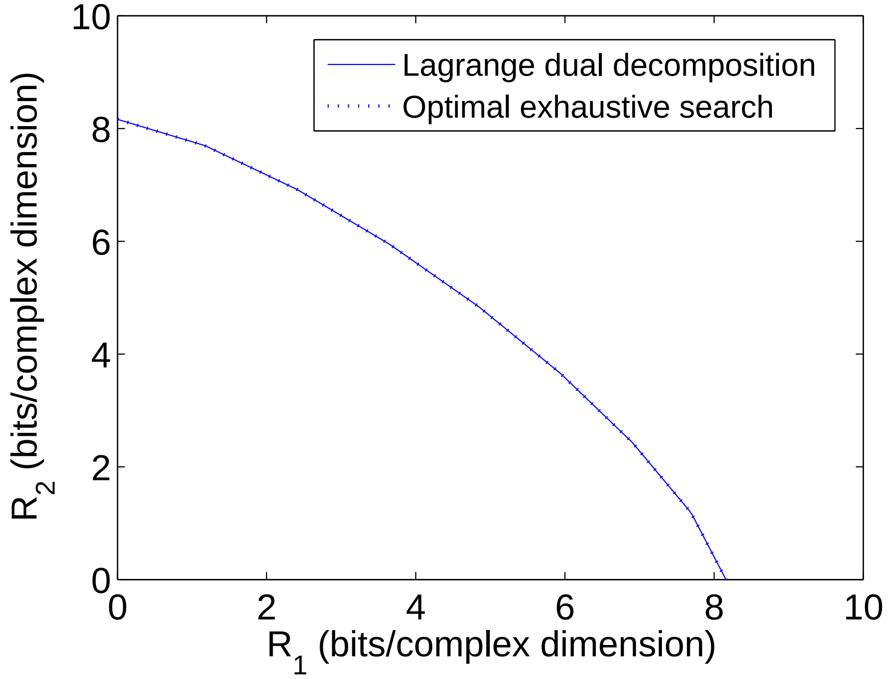


Figure 5.7: Rate region of SISO-OFDMA BC obtained by using Optimal exhaustive search versus Lagrange dual decomposition methods ($N = 8$, $K = 2$, channel SNR vectors are $10[1^2\ 2^2\ \cdots\ N^2]^T$ and $10[N^2\ (N-1)^2\ \cdots\ 1^2]^T$. $P_{tot} = NK = 16$)

region when the target rate vector $\mathbf{R} = [4.84\ \ 4.84]^T$ bits per complex dimension[§]. The boundary points are characterized by solving WSPmin for all possible weight vectors. It can be observed that optimal exhaustive search and Lagrange dual decomposition achieve the identical power region. Since the target rate vector lies on the boundary of rate region in Fig. 5.7, the minimum sum power to achieve this rate vector must equal $P_{tot} = 16$, which can be verified in Fig. 5.8. The results in Fig. 5.7 and Fig. 5.8 suggest that in practical MIMO-OFDMA systems with much more than eight tones, the proposed dual approach can find optimal solutions with the significantly lower computational complexity than the optimal exhaustive search.

Fig. 5.9 presents the rate and power regions for two user MIMO-OFDMA BC and MAC with 1024 tones, two transmit and two receive antennas. These results

[§] Bits/complex dimension is equivalent to bits/tone/antenna

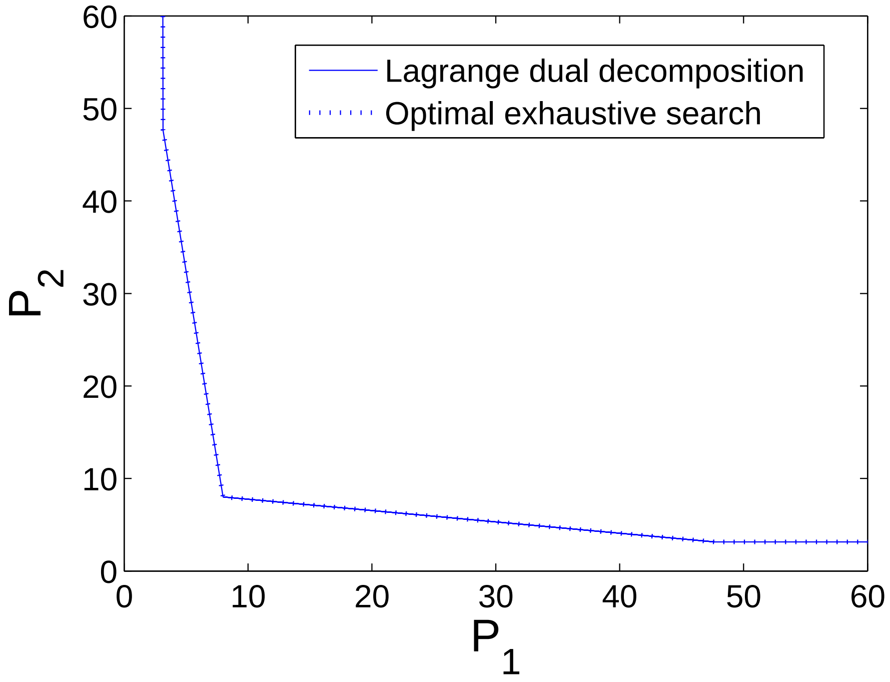
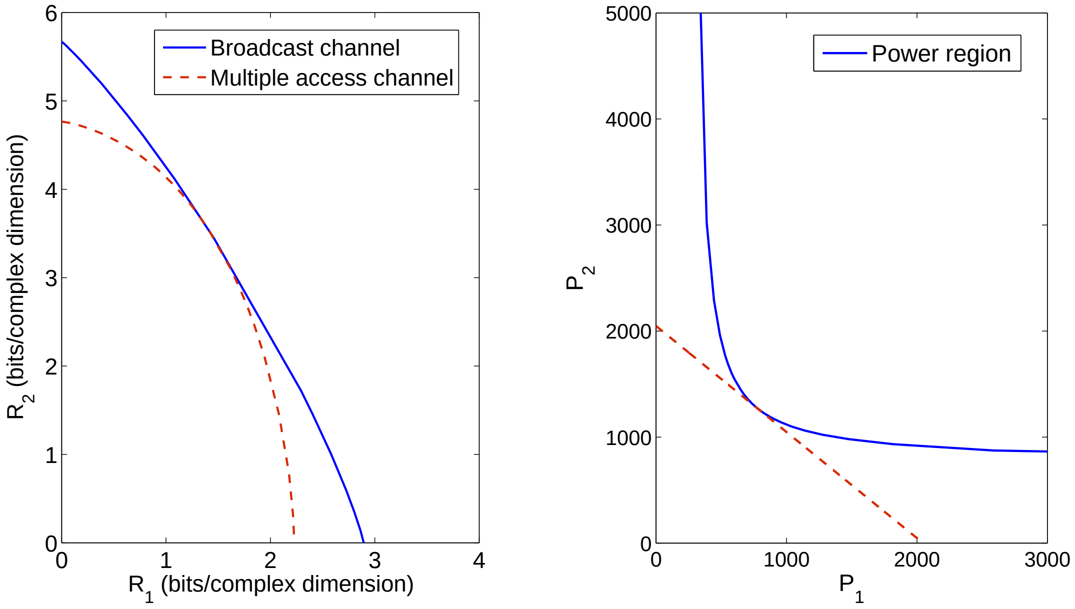


Figure 5.8: Power region of SISO-OFDMA BC obtained by using Optimal exhaustive search versus Lagrange dual decomposition methods ($N = 8$, $K = 2$, channel SNR vectors are the same as those in Fig. 5.7, $\mathbf{R} = [4.84 \;\; 4.84]^T$ bits/complex dimension)

are obtained by applying the proposed efficient algorithms based on Lagrange dual decomposition. It is assumed that the average channel SNR on each tone is 10dB for user 1 and 20dB for user 2. Also, the channel matrix $\mathbf{H}_k(n)$ is assumed to have independent zero-mean complex-Gaussian entries with the same variance and undergo independent fading across the tones. Fig. 5.9(a) shows the rate region of BC and MAC where the total power constraint for BC is $P_{tot} = 2048$, and where the individual power constraint for MAC is $P_1 = P_2 = 1024$. Since $P_1 + P_2 = P_{tot}$, at least one rate tuple must be common in both BC and MAC rate regions, which can be observed in this subfigure. Fig. 5.9(b) presents the power region when the target data rate is $\mathbf{R} = [1.08 \;\; 4.12]^T$ bits/complex dimension. This power region is identical for BC and MAC since it is characterized by solving WSPmin for every possible weight



(a) Rate region ($P_{tot} = 2048, P_1 = P_2 = 1024$)

(b) Power region ($\mathbf{R} = [1.08 \;\; 4.12]^T$)

Figure 5.9: Rate and power regions for MIMO-OFDMA BC and MAC ($N = 1024$, $K = 2$, Two transmit and two receive antennas)

vector. The target rate vector is on the boundary of achievable rate region for MIMO-OFDMA BC in Fig. 5.9(a). Thus, the minimum achievable sum power should be equal to the total power constraint for this MIMO-OFDMA BC, which is indicated by the dashed line in Fig. 5.9(b).

5.6 Summary

In downlink and uplink MIMO-OFDMA systems with CSIT, efficient algorithms are developed for weighted sum-rate maximization and weighted sum-power minimization problems. Though these are originally non-convex problems with the exponential complexity, the duality gap of each problem is shown to quickly vanish as the number of tones grows. From this observation, Lagrange dual decomposition is employed to efficiently solve both problems. Simulation results show that with only eight tones, virtually optimal solutions are obtained by using the proposed dual approach.

Chapter 6

Scheduling in MIMO-OFDMA with No CSIT

Previous chapters assumed that perfect instantaneous CSI is available at the transmitter for use in cross-layer resource allocation. However, if the coherence time of fading channels is not sufficiently long, the current channel state may be quite different from the CSI delivered through the feedback channel, which would limit the value of CSIT. Thus, in a highly mobile environment, instantaneous perfect CSIT becomes infeasible, and the transmitter may only have long-term channel distribution information (CDI) of fading states. In multi-user MIMO systems, if the instantaneous CSI is available at the transmitter via a reliable feedback link, it is possible to apply spatial division multiple access (SDMA) where data streams sent through multiple transmit antennas simultaneously can be destined to multiple users rather than a single user [11, 78, 62]. Without instantaneous CSIT, not only is the application of SDMA quite difficult, but also SDMA's throughput gain might become insignificant.

This chapter considers cross-layer resource allocation in MIMO-OFDMA systems with CDI at the transmitter (CDIT) where only the long-term statistics of fading states are available at the transmitter. MIMO-OFDMA implies that each tone is occupied by at most one user, and the assigned user makes use of the MIMO channel formed at the corresponding tone as a one-to-one communication link. With only CDIT, an outage event inevitably occurs since the transmitter is unable to adapt the data rate to the current channel mutual information. An outage is declared when the mutual information of the instantaneous channel is below the transmission data rate. In this situation, the receiver cannot successfully decode the original data, and a packet error occurs. Thus, the outage probability can be generally considered as the packet-error probability, and in practical systems, the base station selects the maximum transmission rate to achieve the target packet-error rate. If the target outage probability is given for each user, and the transmitter knows the statistics of fading channels, it is possible to characterize the maximum achievable rate region, called the *outage rate region* [40, 29, 41]. All rate tuples within this region can be supported while satisfying the target outage probabilities.

Characterization of the outage rate region for MIMO-OFDMA systems is non-trivial since the statistics of the mutual information lack closed-form expressions and require complicated numerical integrals. In [30, 47, 75], it is shown that the mutual information of MIMO channels can be well approximated as having a Gaussian distribution, an approximation that becomes more accurate as the number of transmit and receive antennas increases. Since the mutual information of MIMO-OFDMA channels takes the form of a summation of the mutual information of MIMO channels on each tone, the Gaussian approximation can be extended to MIMO-OFDMA systems. Based on the Gaussian approximation method, this chapter presents efficient numerical algorithms to characterize the outage rate region of downlink and uplink MIMO-OFDMA systems. Given a rate tuple and outage probabilities, this algorithm checks the feasibility of the given rate vector, and if feasible, finds the tone assignment for each user in an efficient manner. Thus, the data rate of each user can be quickly updated without violating the target outage probabilities.

In addition to characterizing the outage rate region of MIMO-OFDMA BC and MAC, the Gaussian approximation in conjunction with a *successive feasibility check* can be directly used in efficiently finding the rate and power allocation on each tone under queue-proportional scheduling (QPS), which makes it easy to apply QPS to MIMO-OFDMA BC and MAC with CDIT. On the other hand, it is generally difficult to apply other gradient-based scheduling policies* such as maximum weight matching

*Gradient-based scheduling policies require maximization of weighted sum rate.

scheduling (MWMS) to MIMO-OFDMA with CDIT because of the high numerical complexity. Stochastic simulation results demonstrate that QPS provides much more desirable delay and fairness properties compared to other well-known schedulers.

The organization of this chapter is as follows: Section 6.1 describes the overall system models, and Section 6.2 provides a discussion on the mutual information of MIMO-OFDMA channels and its Gaussian approximation. Section 6.3 presents an efficient numerical algorithm for characterizing the outage rate region as well as its application to QPS. Finally, Section 6.4 presents numerical results and a discussion, and concluding remarks appear in Section 6.5.

6.1 System Models

This section presents the models of downlink and uplink MIMO-OFDMA systems as well as queueing systems for use in cross-layer resource allocation.

6.1.1 Downlink MIMO-OFDMA Systems

First, consider a downlink MIMO-OFDMA system with K users and N tones where the base station (BS) is equipped with t transmit antennas and K mobile terminals with $r_1, \cdots, r_K$ receive antennas, respectively. It is assumed that the inter-symbol interference (ISI) is completely removed by the cyclic prefix in OFDM techniques, i.e. the frequency response is flat within each tone. The total transmit power is constrained to P_{tot}. At user k's tone n, a MIMO channel is formed, which is given as

$$\mathbf{y}_k(n) = \mathbf{H}_k(n)\mathbf{x}_k(n) + \mathbf{z}_k(n) = \sqrt{\frac{\rho_k}{t}}\mathbf{H'}_k(n)\mathbf{x}_k(n) + \mathbf{z}_k(n), \tag{6.1}$$

where $\mathbf{y}_k(n) \in \mathbb{C}^{r_k \times 1}$, $\mathbf{H}_k(n) \in \mathbb{C}^{r_k \times t}$, and $\mathbf{x}_k(n) \in \mathbb{C}^{t \times 1}$ denote, respectively, the received signal vector, the channel matrix, and the transmitted signal vector at user k's tone n. $\mathbf{z}_k(n) \in \mathbb{C}^{r_k \times 1}$ is a vector of independent zero-mean complex Gaussian noise entries with variance $1/2$ per real component at user k's receiver. The block diagram of MIMO-OFDMA BC is provided in Fig. 5.1.

The channel matrix, $\mathbf{H}_k(n)$, is assumed to be perfectly known to the receiver and unknown to the transmitter; thus, it is impossible to adapt the transmission strategy in response to the instantaneous channel matrix. ρ_k denotes user k's channel SNR per receive antenna at each tone. $\mathbf{H}_k(n) = \sqrt{\rho_k/t}\mathbf{H'}_k(n)$, and the elements of $\mathbf{H'}_k(n)$ are assumed to be independent zero-mean circularly symmetric complex Gaussian (ZMCSCG) random variables[†] with variance 1/2 per real dimension. The channel SNR ρ_k is only a function of the user index k, and for each user, it is assumed to be the same across every tone. This is not exactly true if the propagation loss increases at higher frequencies. However, the increment in loss depends on the ratio of the carrier frequency to the signal bandwidth; a larger ratio reduces the difference in propagation loss across the tones. In practice, the carrier frequency is much larger than the signal bandwidth[‡], a fact that validates the constant variance assumption over every tone. Across each user's tones, the MIMO channels are correlated depending on the coherence bandwidth.

From the ZMCSCG property of channel matrices, every antenna is assumed to transmit independent Gaussian distributed signals with equal average power as in [30]. Also, it is assumed that each user allocates equal power on assigned tones to that user. Since the channel statistics on the same user's tones are identical, this equal power allocation is optimal for uplink OFDMA systems with CDIT. In addition to the uplink, [33] shows that the equal power allocation over the whole bandwidth is the best strategy for downlink OFDMA systems with CDIT. Thus, when the total transmit power is constrained to P_{tot}, the allocated power on one transmit antenna of each tone is equal to $\frac{P_{tot}}{tN}$, and user k's SNR per receive antenna at each tone is given by $\rho'_k = \rho_k \frac{P_{tot}}{tN}$. Consequently, the mutual information between $\mathbf{x}_k(n)$ and $\mathbf{y}_k(n)$ with a given $\mathbf{H}_k(n)$ is represented as the following [68].

$$\mathcal{I}_k(n) = \log_2 \left| \mathbf{I} + \frac{\rho'_k}{t} \mathbf{H'}_k(n) \mathbf{H'}_k(n)^H \right| \qquad \text{(bits/sec/Hz)}. \tag{6.2}$$

[†] A complex Gaussian random variable $Z = X + jY$ is ZMCSCG if X and Y are independent real Gaussian random variables with zero mean and equal variance [48].

[‡] For example, mobile WiMAX systems have carrier frequency greater than 2 GHz and system bandwidth of 1.25∼20 MHz.

Let S_i denote the set of tones allocated to user i. Because of the FDMA constraint in OFDMA systems, each tone is allowed to be used by at most one user; hence, $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^{K} S_i \subseteq \{1, 2, \cdots, N\}$. Then, the mutual information for user k, i.e. between $\{\mathbf{x}_k(1), \cdots, \mathbf{x}_k(N)\}$ and $\{\mathbf{y}_k(1), \cdots, \mathbf{y}_k(N)\}$ with given $\{\mathbf{H}_k(1), \cdots, \mathbf{H}_k(N)\}$ is

$$\mathcal{I}_k = \frac{1}{N}\sum_{n\in S_k}\mathcal{I}_k(n) = \frac{1}{N}\sum_{n\in S_k}\log_2\left|\mathbf{I} + \frac{\rho'_k}{t}\mathbf{H'}_k(n)\mathbf{H'}_k(n)^H\right| \qquad \text{(bits/sec/Hz)}. \quad (6.3)$$

If the channel matrix is approximately constant during a transmission interval, given a target rate R_k, an outage is declared for user k when $\{\mathcal{I}_k < R_k\}$[§], and the probability of this event is called outage probability of user k. It is difficult to express the distribution of $\mathcal{I}_k$ in a simple closed form, and characterizing the relation between the target rate and the outage probability may require complicated numerical integrals, which are intractable even for the single-user case. The next section shows that the mutual information can be well approximated by a Gaussian distribution, and using this result, efficient algorithms are presented to characterize the outage rate region given each user's outage probability.

The models of queueing system and scheduler in this chapter are basically the same as those described in Chapter 2.2 except for the additional consideration of outage events. The rate vector at time slot t, $\mathbf{R}(t)$ is determined by the scheduler based on the outage rate region and perfect QSI. Because of the outage events, user i's queue-state vector after one scheduling period[¶] is expressed as

$$Q_i(t+1) = \begin{cases} \max\{Q_i(t) - R_i(t), 0\} + Z_i(t) & \text{w.p. } 1-\text{Prob}\{R_i(t) \text{ is in outage}\}; \\ Q_i(t) + Z_i(t) & \text{w.p. Prob}\{R_i(t) \text{ is in outage}\}. \end{cases} \quad (6.4)$$

where w.p. means 'with probability'. In practice, the outage event is detected at the receiver, and notification of the outage to the transmitter may result in some delay effects. For simplicity, those effects are ignored by assuming that the outage is

[§] During each scheduling period, one codeword is assumed to be sent across user k's tones with the transmission rate of R_k bps/Hz

[¶] T_s is assumed to be 1

immediately known at the transmitter.

6.1.2 Uplink MIMO-OFDMA Systems

In uplink MIMO-OFDMA systems, the major differences from downlink MIMO-OFDMA systems are that each mobile terminal (MT) has its own power constraint, and the queueing systems are distributed across the users. Thus, as noted in the introduction, a feedback link is required to send each user's queue backlog to the BS. Consider an uplink MIMO-OFDMA system with K users and N tones where the BS is equipped with r receive antennas and K mobile terminals with $t_1, \cdots, t_K$ transmit antennas, respectively. At user k's tone n, the following MIMO channel is formed:

$$\mathbf{y}_k(n) = \mathbf{H}_k(n)\mathbf{x}_k(n) + \mathbf{z}_k(n) = \sqrt{\frac{\rho_k}{t_k}}\mathbf{H'}_k(n)\mathbf{x}_k(n) + \mathbf{z}_k(n), \tag{6.5}$$

where $\mathbf{y}_k(n) \in \mathbb{C}^{r\times 1}$, $\mathbf{H}_k(n) \in \mathbb{C}^{r\times t_k}$, and $\mathbf{x}_k(n) \in \mathbb{C}^{t_k\times 1}$ denote, respectively, the received signal vector, the channel matrix, and the transmitted signal vector at user k's tone n. $\mathbf{z}_k(n) \in \mathbb{C}^{r\times 1}$ is a vector of independent zero-mean complex Gaussian noise entries with variance 1/2 per real component at the BS. As in downlink case, the MIMO matrix channel is assumed to be perfectly known to the receiver and unknown to the transmitter. ρ_k denotes user k's channel SNR per receive antenna at each tone. The block diagram of MIMO-OFDMA MAC is presented in Fig. 5.2.

The entries of $\mathbf{H'}_k(n)$ are ZMCSCG random variables with variance 1/2 per real dimension. In addition, the channel SNR on each tone, ρ_k is only a function of the user index k. Therefore, each mobile terminal transmits an independent Gaussian distributed signal with equal average power on each antenna per allocated tone. If tone n is occupied by user k, the allocated power on one transmit antenna at tone n is equal to $\frac{P_k}{t_k|S_k|}$, where P_k represents user k's power constraint and $|S_k|$ denotes the cardinality of the set S_k, i.e. the number of tones assigned to user k. Therefore, user k's SNR per receive antenna at each tone is given by $\rho'_k = \rho_k \frac{P_k}{t_k|S_k|}$. As a result, the mutual information for user k, i.e. between $\{\mathbf{x}_k(1), \cdots, \mathbf{x}_k(N)\}$ and $\{\mathbf{y}_k(1), \cdots, \mathbf{y}_k(N)\}$ with given $\{\mathbf{H}_k(1), \cdots, \mathbf{H}_k(N)\}$ is represented as

$$\mathcal{I}_k = \frac{1}{N}\sum_{n\in S_k}\mathcal{I}_k(n) = \frac{1}{N}\sum_{n\in S_k}\log_2\left|\mathbf{I}+\frac{\rho'_k}{t_k}\mathbf{H'}_k(n)\mathbf{H'}_k(n)^H\right| \qquad \text{(bits/sec/Hz)}. \quad (6.6)$$

The stochastic properties and models of queueing systems and schedulers are the same as in the downlink case. The only difference is that in the uplink, each MT has one output queue assumed to have infinite capacity. Thus, each MT's QSI needs to be sent to the BS through a reliable feedback channel. The delay in reporting QSI to the BS may cause some degradation in scheduling performance such as the increase of queueing delay. As long as the packet arrival rate is not intensively high, the change in queue states from new packet arrivals remains very slow. Also, previous rate allocations can be tracked at the scheduler, which makes it feasible to compensate the inaccuracy of QSI in case of large feedback delay. For simplicity, this chapter assumes that perfect instantaneous QSI is available at the BS. The next section presents efficient algorithms to characterize the outage rate region given each user's outage probability, based on a Gaussian approximation of the MIMO channel mutual information.

6.2 Gaussian Approximation of Mutual Information

Efficient evaluation of the mutual information is essential in characterizing the outage rate region given each user's outage probability. However, the distribution for the mutual information presented in the previous section takes a very complicated form, and the evaluation of outage probability involves the integration of Laguerre polynomials, a procedure that is numerically complex. Recently, [30, 75] show that the mutual information of MIMO channels is close to a Gaussian distribution. In MIMO-OFDMA systems, the MIMO mutual information is summed over multiple tones as shown in (6.3) and (6.6). The sum of jointly Gaussian random variables also has a Gaussian distribution. In addition, by central limit theorem, the sum of i.i.d.

random variables with arbitrary distribution will have a distribution more like Gaussian. Thus, as the correlation among the tones allocated to user k becomes smaller, the mutual information for user k in MIMO-OFDMA systems can be better approximated by a Gaussian distribution compared to that of single MIMO channels. Fig. 6.1 illustrates the experimental distribution of the mutual information for multi-tone MIMO channels as well as a Gaussian distribution with the same mean and variance when $N = 5$, $t = 2$, and $r = 2$. Each tone is assumed to fade independently. In this figure, it can be observed that the Gaussian distribution shows a very good match to the exact distribution at both low and high SNR ranges.

Therefore, if the mean and variance of the overall channel mutual information is reliably estimated, the outage probability can be characterized by using Gaussian approximation methods. However, it is difficult to exactly estimate the mean and variance of the mutual information of MIMO channels. [36] provides a closed-form analytical expression for approximate mean and variance values of the MIMO mutual information. These are derived by characterizing the asymptotic probability distribution of MIMO mutual information. In spite of the asymptotic nature, [36] shows that the derived variance is quite accurate at all SNR ranges even for a small number of transmit and receive antennas. These results are presented and extended to MIMO-OFDMA systems in the following two propositions.

Proposition 3. *The mean of mutual information of user k's channel in downlink MIMO-OFDMA systems can be approximated as follows:*

$$\begin{aligned}\mathbb{E}[\mathcal{I}_k] &\approx \frac{t|S_k|}{N}[c\log_2(1+\rho'_k-\rho'_k v(c,\rho'_k)) \\ &+ \log_2(1+\rho'_k c-\rho'_k v(c,\rho'_k))-(\log_e 2)v(c,\rho'_k)],\end{aligned} \tag{6.7}$$

where $c = \frac{r_k}{t}$, $\rho'_k = \rho_k \frac{P_{tot}}{tN}$, $|S_k|$ *is the number of tones assigned to user* k, *and*

$$v(c,\rho'_k) = \frac{1}{2}\left[1+c+\frac{1}{\rho'_k}-\sqrt{\left(1+c+\frac{1}{\rho'_k}\right)^2-4c}\right]. \tag{6.8}$$

This equation is also valid for the uplink MIMO-OFDMA systems with the following

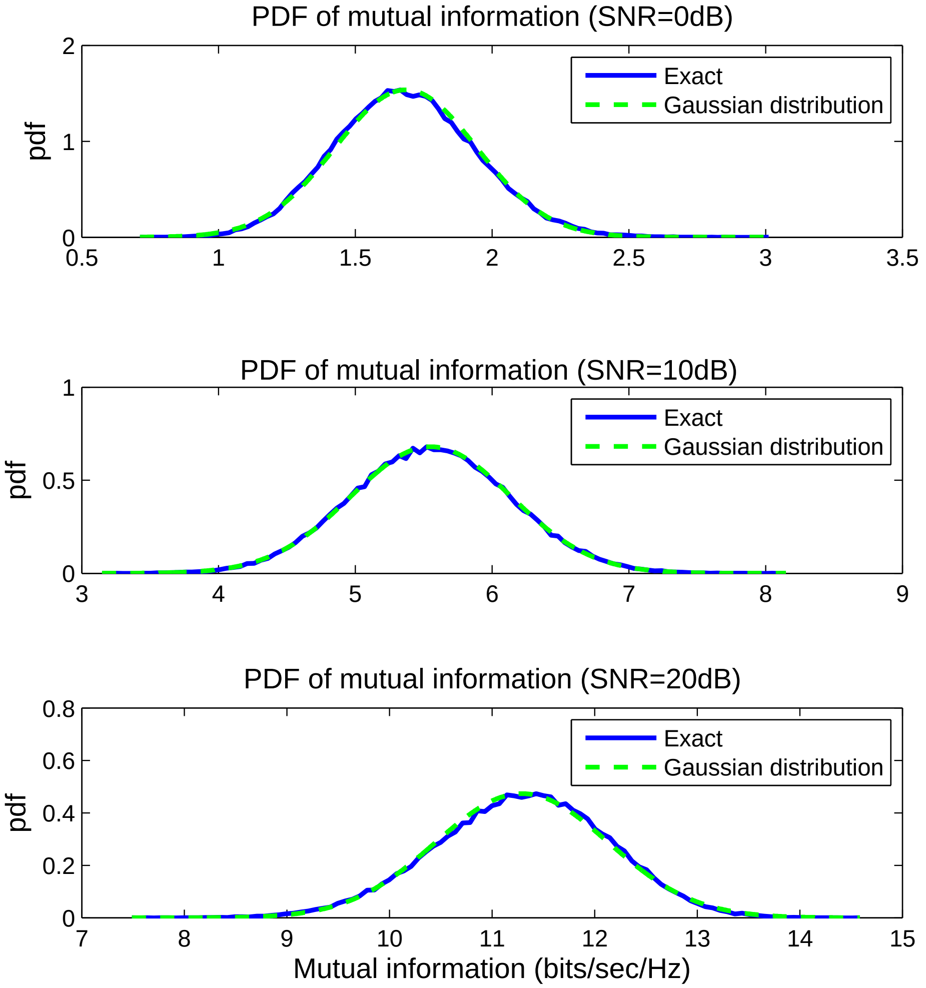


Figure 6.1: Probability distribution of the mutual information for MIMO-OFDMA channels ($N = 5$, $t = r = 2$, and $P_{tot} = tN = 10$)

changes: $t \rightarrow t_k$, $r_k \rightarrow r$, and $\rho'_k = \rho_k \frac{P_k}{t_k|S_k|}$.

Proof. Each tone is assume to have the same distribution and the expectation of the sum of random variables is equal to the sum of the individual expectations. Thus, the approximation of the mean value is simply the right hand side of (11) in [36] multiplied by $|S_k|/N$. □

Proposition 4. *The variance of mutual information for user k's channel in downlink MIMO-OFDMA systems can be approximated as*

$$Var[\mathcal{I}_k] \approx -\frac{(\log_2 e) \sum_{i=1}^{|S_k|} \sum_{j=1}^{|S_k|} \sigma_{i,j}^{(k)}}{N^2} \log_2 \left(1 - \frac{v(c, \rho'_k)^2}{c}\right), \tag{6.9}$$

where the definition of each variable is the same as that in Proposition 3. $\sigma_{i,j}^{(k)}$ denotes the entry at the ith row and the jth column of $\mathbf{\Sigma}^{(k)} \in \mathbb{C}^{|S_k| \times |S_k|}$. $\mathbf{\Sigma}^{(k)}$ is the normalized covariance matrix for mutual information of each tone assigned to user k, which has unit diagonal entries. This equation is also valid for the uplink MIMO-OFDMA systems with the same variable changes as in the previous proposition.

Proof. Let $\mathcal{I}_k = 1/N \sum_{i=1}^{|S_k|} x_i$ where $\mathbf{x} = [x_1 \ \cdots \ x_{|S_k|}]^T$ denotes a vector for the mutual information of each tone assigned to user k which has the mean of $\boldsymbol{\mu} = [\mu_1 \ \cdots \ \mu_{|S_k|}]^T$. Then, the normalized covariance matrix of $\mathbf{x}$ is defined as $\mathbf{\Sigma}^{(k)} = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]/\mathbb{E}[(x_1 - \mu_1)^2] \in \mathbb{C}^{|S_k| \times |S_k|\|}$. Hence, the variance of $\mathcal{I}_k = 1/N \mathbf{1}^T \mathbf{x}$ becomes equal to $\mathbb{E}[(x_1 - \mu_1)^2]/N^2 \mathbf{1}^T \mathbf{\Sigma}^{(k)} \mathbf{1}$. $\mathbf{1}^T \mathbf{\Sigma}^{(k)} \mathbf{1} = \sum_{i=1}^{|S_k|} \sum_{j=1}^{|S_k|} \sigma_{i,j}^{(k)}$ where $\sigma_{i,j}^{(k)}$ denotes the entry at the ith row and the jth column of $\mathbf{\Sigma}^{(k)}$. As a result, the approximation of $Var[\mathcal{I}_k]$ becomes the one in (13) in [36] multiplied by $\sum_{i=1}^{|S_k|} \sum_{j=1}^{|S_k|} \sigma_{i,j}^{(k)}/N^2$. □

If all the tones assigned to user k fade independently, $\sigma_{i,j}^{(k)} = 0$ for $i \neq j$. Thus, $\sum_{i=1}^{|S_k|} \sum_{j=1}^{|S_k|} \sigma_{i,j}^{(k)}$ in (6.9) takes a value of $|S_k|$. When the tones are correlated, some off-diagonal entries of $\mathbf{\Sigma}^{(k)}$ become positive, which results in higher variance of mutual information. The outage rate is defined as the maximum achievable rate to satisfy the target outage probability. With larger variance in mutual information, the outage

[‖] Note that every tone is assumed to have the same fading statistics; thus, the same variance of mutual information.

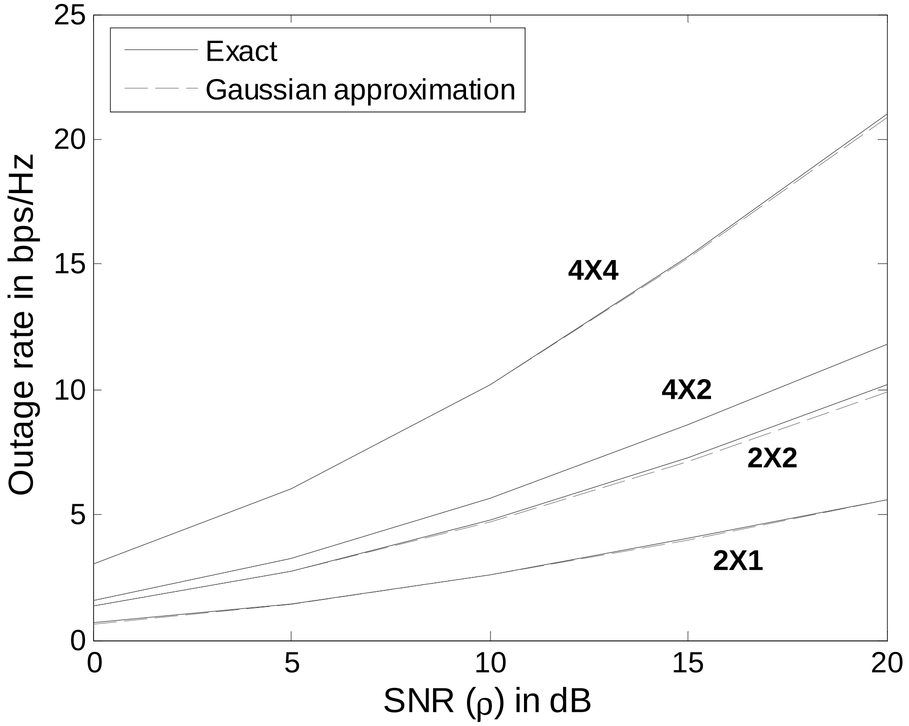


Figure 6.2: Outage rate versus SNR per receive antenna ($N = 5$, $P_{tot} = tN$, and outage probability=10%)

rate becomes lower for the same target outage probability. Therefore, it is desirable to maximize the tone spacing by using distributed tone allocation in order to minimize fading correlation among the adjacent tones.

Fig. 6.2 and Fig. 6.3 illustrate accuracy of the Gaussian approximation by using derived mean and variance values. In these figures, the outage rate in MIMO-OFDMA systems is evaluated for target outage probability of 10% and 1%, respectively, by using the exact distribution of mutual information as well as the approximated Gaussian distribution. $N = 5$ and a variety of antenna configurations are considered: 2×1, 2×2, 4×2, and 4×4 where $m \times n$ denotes m transmit and n receive antennas. It is assumed that each tone undergoes independent fading. Both figures show that the outage rate obtained by using the Gaussian approximation is almost the same as the exact outage rate for any channel SNR ranging from 0 to 20dB, even when there are only two transmit or receive antennas. As the number of antennas increases, the

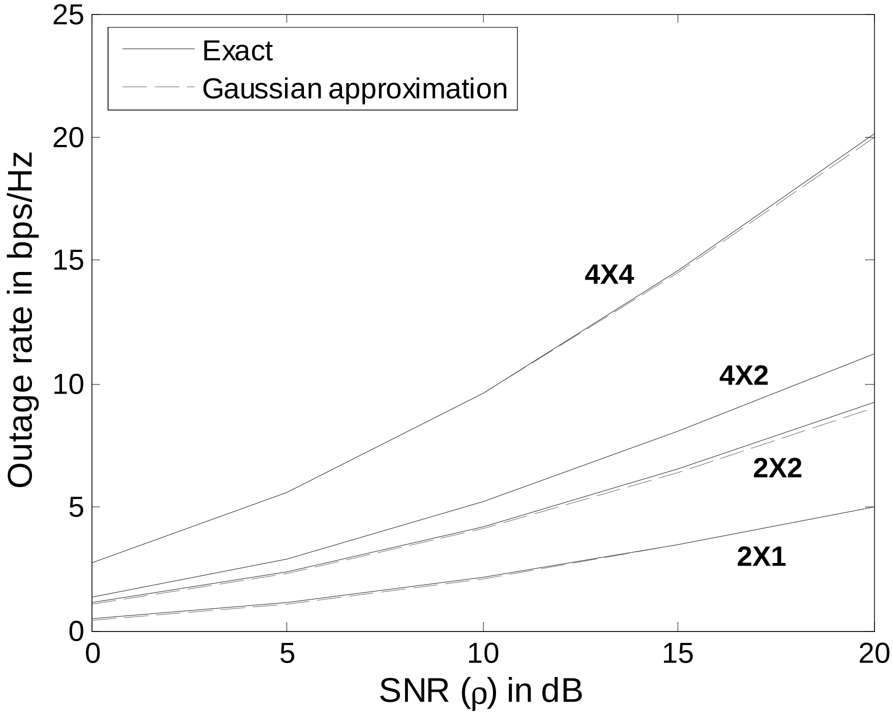


Figure 6.3: Outage rate versus SNR per receive antenna ($N = 5$, $P_{tot} = tN$, and outage probability=1%)

Gaussian approximation becomes more accurate, which will further reduce the gap in outage rates.

6.3 Efficient Algorithms for Outage Rate Region and QPS

Based on the Gaussian approximation derived in the previous section, this section presents efficient algorithms to characterize the outage rate region of the MIMO-OFDMA BC and MAC. In addition, it is shown that these algorithms can be directly used in finding the rate and power allocation for QPS in MIMO-OFDMA systems. By using the Gaussian approximation, the outage probability can be reliably estimated in a very efficient way. Furthermore, given the outage probability of each user, it is possible to check the feasibility of a certain rate tuple, i.e. the rate tuple can be checked if it is within the outage rate region with the given outage probabilities. The outage probability constraint on user k is denoted by o_k. Also, denote $N_k = |S_k|$ where $\sum_{k=1}^{K} N_k = N$. The estimated mean and variance of user k's mutual information are denoted by $N_k a_k$ and $(N_k + \gamma_k) b_k^2$, respectively, where $\gamma_k = \sum_{i \neq j} \sigma_{i,j}^{(k)}$. The quantities a_k and b_k are not related to N_k in the downlink, but they are so related in the uplink since the channel SNR, ρ_k is inversely proportional to N_k. Also, the value of γ_k reflects the fading correlation among the allocated tones, which is dependent on the tone separation. It is assumed that given N_k, N_k tones are distributively selected from the set of available tones such that γ_k is minimized. If user k's target rate is denoted by R_k, with an approximate distribution $\mathcal{I}_k \sim \mathcal{N}(N_k a_k, (N_k + \gamma_k) b_k^2)$, the outage probability is represented as follows.

$$p_k = P\{\mathcal{I}_k < R_k\} = 1 - \frac{1}{2}\mathrm{erfc}\left(\frac{R_k - N_k a_k}{\sqrt{2(N_k + \gamma_k)} b_k}\right), \tag{6.10}$$

where the complementary error function is defined as $\mathrm{erfc}(x) = 2/\sqrt{\pi} \int_x^\infty e^{-t^2}\, dt$. Since $p_k \leq o_k$ from the constraint, for the fixed N_k, the maximum achievable data rate of user k is given by

$$R_{k,\max} = N_k a_k + \sqrt{2(N_k + \gamma_k)} b_k \mathrm{erfc}^{-1}(2 - 2o_k). \tag{6.11}$$

In (6.11), $R_{k,\max}$ always increases with a larger N_k. In multi-user systems, the outage rate region is defined as the set of achievable rate vectors without violating any given constraint on each user's outage probability. The boundary of the outage rate region generally can be characterized by solving either of the following two problems: weighted sum-rate maximization (WSRmax) or proportional-rate maximization (PRmax) problems. First, WSRmax problem can be formulated as

$$\text{maximize} \qquad \sum_{k=1}^{K} w_k R_k$$

$$\begin{aligned}\text{subject to}\quad & R_k = N_k a_k + \sqrt{2(N_k+\gamma_k)}b_k \mathrm{erfc}^{-1}(2-2o_k), \quad k=1,2,\cdots,K \\ & \sum_{k=1}^{K} N_k = N, \text{ and } N_k \text{ is a nonnegative integer,}\end{aligned} \tag{6.12}$$

where $w_k \geq 0$ is the weight assigned to user k such that $\sum_{k=1}^{K} w_k = 1$. The boundary of the outage rate region can be traced by solving this problem for all possible weight vectors. Because of the integer constraints on N_k, (6.12) may require an exhaustive search to find the optimal solution for both uplink and downlink. The number of every possible choice for $\{N_1, \cdots, N_K\}$ is equal to the number of ways obtaining an ordered subset of $K-1$ elements from $\{1, 2, \cdots, N\}$. If each ordered element in the obtained subset is labeled by $\{A_1, A_2, \cdots, A_{K-1}\}$, $N_k = A_k - A_{k-1}$ for all $k = 1, \cdots, K$ where $A_0 = 0$ and $A_K = N$. Thus, the number of all the possible cases for $\{N_1, \cdots, N_K\}$ is ${}_N P_{K-1} = N!/(N-K+1)!$ and $\mathcal{O}(N!/(N-K+1)!) \approx \mathcal{O}(N^K)$ executions are needed to solve (6.12). This exponential complexity makes this problem intractable with the large number of users or tones. Even when the integer constraint on N_k is assumed to be relaxed, this problem is not a convex optimization problem, since the objective for maximization is not concave in the practical situations where the outage probability is less than 50%. Therefore, it is hard to solve (6.12) with the polynomial complexity.

On the other hand, the formulation of PRmax problem is as follows.

$$\begin{aligned}\text{maximize}\quad & x \\ \text{subject to}\quad & w_k x \leq N_k a_k + \sqrt{2(N_k+\gamma_k)}b_k \mathrm{erfc}^{-1}(2-2o_k), \quad k=1,2,\cdots,K \\ & \sum_{k=1}^{K} N_k = N, \text{ and } N_k \text{ is a nonnegative integer.}\end{aligned} \tag{6.13}$$

where $w_k \geq 0$ denotes the weight assigned to user k such that $\sum_{k=1}^{K} w_k = 1$. Solving (6.13) provides a rate tuple on the boundary of the outage rate region that is proportional to the weight vector $\mathbf{w} = [w_1, \cdots, w_K]^T$. Therefore, the boundary can be characterized by solving this problem for all possible weight vectors. This problem can be solved very efficiently by using the following successive feasibility check (SFC) algorithm, which is the combination of bisection search and rate feasibility check.

Algorithm 4: Successive Feasibility Check

```
Main Function
$\lambda_{\min} = 0, \lambda_{\max} = \delta N$ where $\delta$ is sufficiently large
Calculate $a_k$ and $b_k$ for all $k$ by (6.3) for the downlink (or (6.6) for the uplink)
While $\lambda_{\max} - \lambda_{\min} > \epsilon$
    $\lambda = (\lambda_{\max} + \lambda_{\min})/2$
    $\mathbf{R}_o = \mathbf{w}\lambda$ where $\mathbf{w}$ is a given weight vector
    $[N_{o,1}, \cdots, N_{o,K}] =$ Check Feasibility($\mathbf{R}_o$)
    If feasible, then
        $\lambda_{\min} = \lambda$
    Else, $\lambda_{\max} = \lambda$
    End If
End While
Return $\mathbf{R}_o$ and $[N_{o,1}, \cdots, N_{o,K}]$

Function $[N_{o,1}, \cdots, N_{o,K}] =$ Check Feasibility($\mathbf{R}_o$)
$N_{remain} = N$
For $k = 1, \cdots, K$
    For $n = 1, \cdots, N_{remain}$
        $N_k = n$
        Update $a_k$ and $b_k$ using (6.7) and (6.9), respectively
        $R_k = N_k a_k + \sqrt{2(N_k + \gamma_k)} b_k \mathrm{erfc}^{-1}(2 - 2o_k)$
        If $R_k >= R_{o,k}$, then
            $N_{remain} = N_{remain} - n$, $N_{o,k} = n$ and break
        Else If $n = N_{remain}$, then
            Return "Infeasible"
        End If
    End For
End For
Return "Feasible" and $[N_{o,1}, \cdots, N_{o,K}]$
```

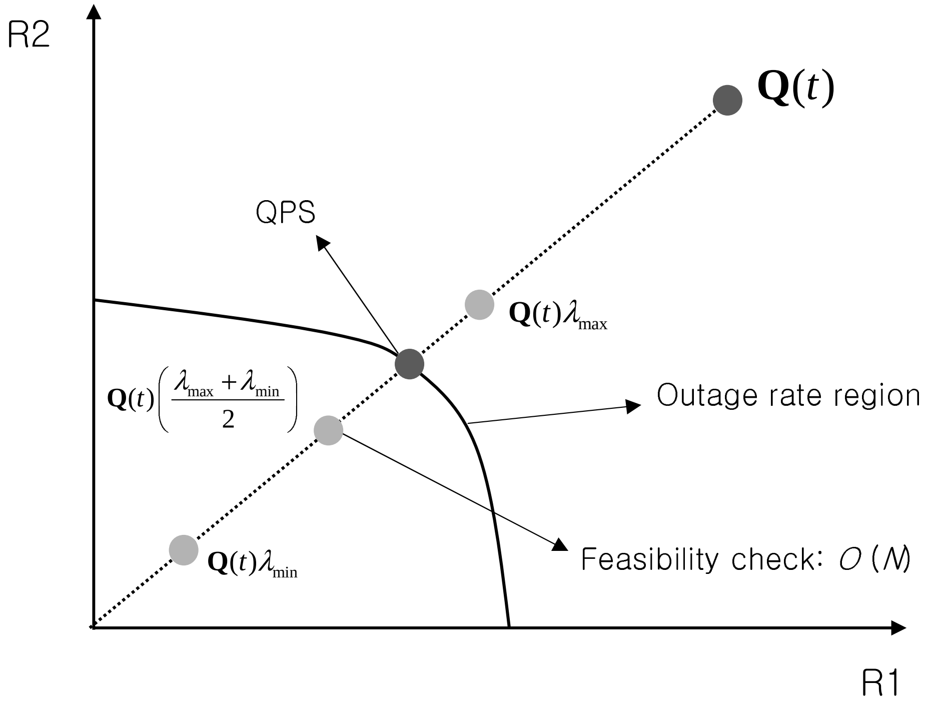

R2
$\mathbf{Q}(t)$
QPS
$\mathbf{Q}(t)\lambda_{\max}$
$\mathbf{Q}(t)\left(\frac{\lambda_{\max}+\lambda_{\min}}{2}\right)$
Outage rate region
Feasibility check: $O(N)$
$\mathbf{Q}(t)\lambda_{\min}$
R1

Figure 6.4: Successive feasibility check for QPS

The parameter δ used in initializing $\lambda_{\max}$ needs to be large enough to guarantee that $\lambda_{\max}\mathbf{w} = N\delta\mathbf{w}$ is outside the outage rate region.

If the weight vector is replaced with the current queue-state vector $\mathbf{Q}(t)$, the PRmax problem becomes equivalent to finding the rate tuple supported by QPS and its corresponding tone and power allocation. Fig. 6.4 illustrates the SFC algorithm for use in applying QPS to two user MIMO-OFDMA systems. First, λ_{min} is set to zero and λ_{max} is initialized with a large value such that the rate tuple, $\mathbf{Q}(t)\lambda_{max}$ is guaranteed to be infeasible. Then, the target rate tuple is chosen to be $\mathbf{R}_o = \mathbf{Q}(t)(\lambda_{max}+\lambda_{min})/2$ and its feasibility is checked as shown in Algorithm 4. If feasible, $\lambda_{min} \leftarrow (\lambda_{max} + \lambda_{min})/2$; otherwise, $\lambda_{max} \leftarrow (\lambda_{max} + \lambda_{min})/2$. This procedure is repeated until the difference between λ_{min} and λ_{max} becomes sufficiently small. As a result, SFC provides the rate tuple $\mathbf{R}_{QPS} = \mathbf{R}_o$ as well as each user's tone and power allocation for QPS.

Each feasibility check requires $\mathcal{O}(NK)$ executions and one-dimensional bisection search only adds constant scaling to the complexity order. Therefore, the overall complexity order of the above SFC algorithm is $\mathcal{O}(NK)$. The characterization of outage rate region as well as application of QPS in MIMO-OFDMA BC and MAC with CDIT can be efficiently achieved by using the SFC algorithm. On the other hand, in order to apply the gradient-type scheduling polices such as MWMS to MIMO-OFDMA with CDIT, WSRmax in (6.12) needs to be solved, which has the exponential complexity in the number of users. Hence, in terms of complexity, QPS has the advantages over the gradient-type schedulers for use in MIMO-OFDMA BC and MAC with CDIT.

With the application of a Gaussian approximation and a SFC, the next section presents a variety of numerical results to compare the performance of different schedulers in both up and down links MIMO-OFDMA systems with CDIT.

6.4 Numerical Results and Discussion

In MIMO-OFDMA systems with CDIT, stochastic simulations with Poisson packet arrivals are performed to evaluate the average queueing delays achieved by different schedulers. The outage rate regions for MIMO-OFDMA BC and MAC are characterized by using the Gaussian approximation in conjunction with the SFC, and four scheduling methods are applied: a fixed rate allocation, MWMS, LQHPR, and QPS. The simulation parameters common to all the simulation results in this section are as follows: 64 tones, two transmit and one receive antennas, 4 μsec OFDM symbol period, tone spacing = 312.5 kHz, scheduling period = 10 msec, each user's target outage probability = 10%, and average packet size = 1 Kbyte. The packet length is exponentially distributed and packet arrivals to each user have i.i.d. Poisson distributions. Over scheduling periods, i.i.d. block fading is assumed, and each tone allocated to the same user is assumed to undergo i.i.d. Rayleigh fading.

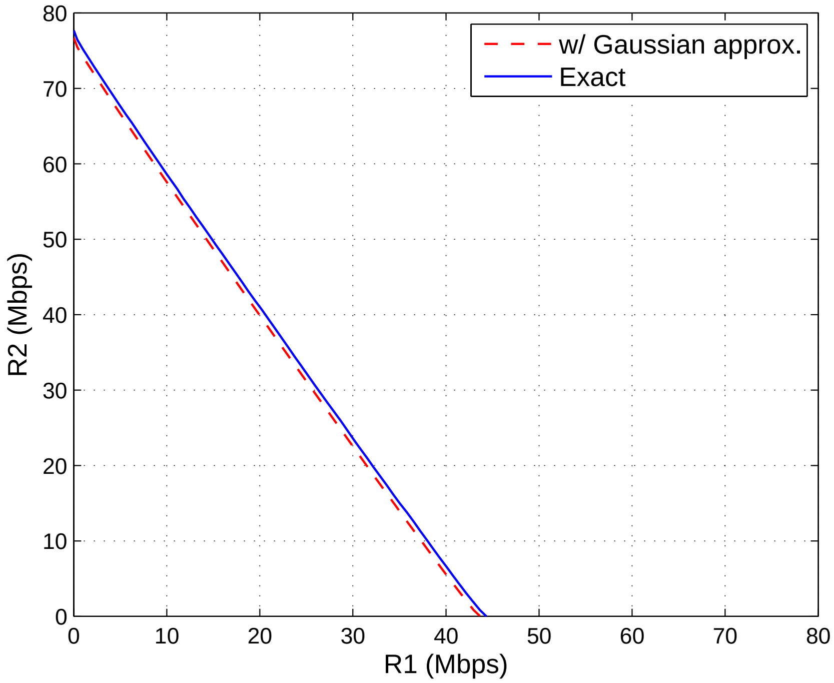


Figure 6.5: Outage rate region of two-user MIMO-OFDMA BC

6.4.1 Downlink MIMO-OFDMA with CDIT

This subsection presents the numerical results for MIMO-OFDMA BC with CDIT. Fig. 6.5 shows the outage rate region of two-user MIMO-OFDMA BC obtained from the exact distribution of mutual information and by employing the approximated Gaussian distribution with SFC. The total transmit power, P_{tot} is assumed to be equal to $N = 64$, and user 1 and 2's channel SNR per receive antenna at each tone are set to 10 and 16 dB, respectively. This figure shows that the outage region derived by using the proposed approach is quite close to the exact one. It is also observed that the outage rate region of MIMO-OFDMA BC takes a triangular shape with some non-convexity, which makes this region slightly smaller than that achieved by TDM transmission. However, if the effect of correlation among the tones is considered, each user's maximum outage rate will be significantly reduced since with every tone assigned to the single user, adjacent tones may exhibit high correlation that increases

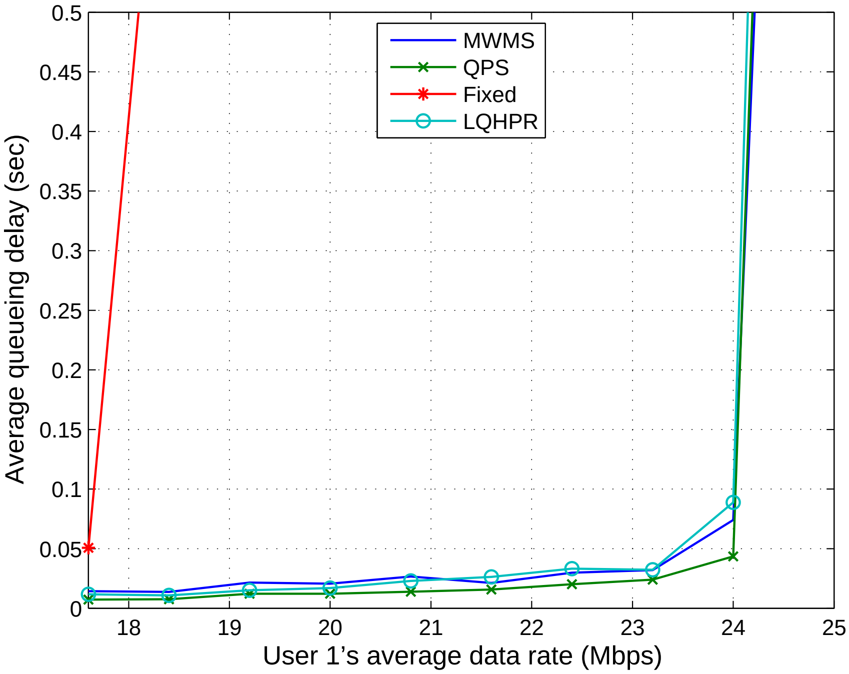


Figure 6.6: Average queueing delay vs User 1's average data rate for MIMO-OFDMA BC with $K = 2$

the variance in channel mutual information. On the other hand, when each user takes significant portion of total bandwidth, the correlation among each user's tones can be kept relatively low by applying the distributed tone allocation. Therefore, the outage rate region of MIMO-OFDMA BC will have more convex shape if the assumption of i.i.d. fading across the tones is removed.

Fig. 6.6 and Fig. 6.7 present the stochastic simulation results for MIMO-OFDMA BC with CDIT. The average queueing delay with Poisson packet arrivals is evaluated for QPS, MWMS, LQHPR, and a fixed operation point. By Little's theorem introduced in Chapter 2.1.2, the average queueing delay over all users can be defined as $\lim_{t\to\infty} \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}[Q_i(t)]/\lambda_i$ where λ_i denotes user i's average bit arrival rate or average data rate. In Fig. 6.6, two-user MIMO-OFDMA BC in Fig. 6.5 is considered. In the simulation, the average queueing delay is evaluated when each user's average data rate increases while satisfying $\lambda_2 = 1.04\lambda_1$. The fixed operation point is set to

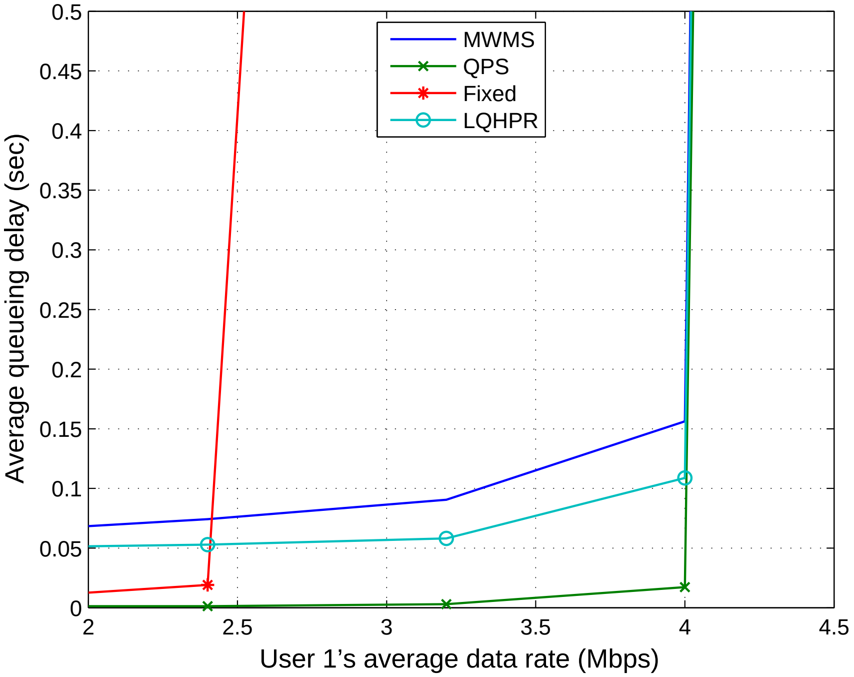


Figure 6.7: Average queueing delay vs User 1's average data rate for MIMO-OFDMA BC with $K = 10$

$[20 \;\; 40]^T$ Mbps, which is on the boundary of outage rate region. Fig. 6.6 shows that with this fixed operation point, user 1's achieved throughput is 17.8 Mbps. Thus, about 10% throughput loss from the target data rate is observed, which is caused by the packet outage events. On the other hand, user 1's achievable throughput is as large as 24 Mbps under QPS, MWMS, and LQHPR. $[27 \;\; 28]^T$ Mbps is the boundary point of outage rate region satisfying the condition, $\lambda_2 = 1.04\lambda_1$. With about 10% throughput loss from this boundary point, 24 Mbps corresponds to user 1's maximum achievable throughput, which corroborates the throughput optimality of QPS and MWMS. One of TDM scheduling policies, LQHPR also happens to achieve this maximum throughput since the outage rate region in Fig. 6.5 is very close to the TDM rate region. In Fig. 6.6, it can also be observed that QPS provides smaller average queueing delay than other schedulers.

In Fig. 6.7, ten-user MIMO-OFDMA BC is considered where $P_{tot} = N = 64$,

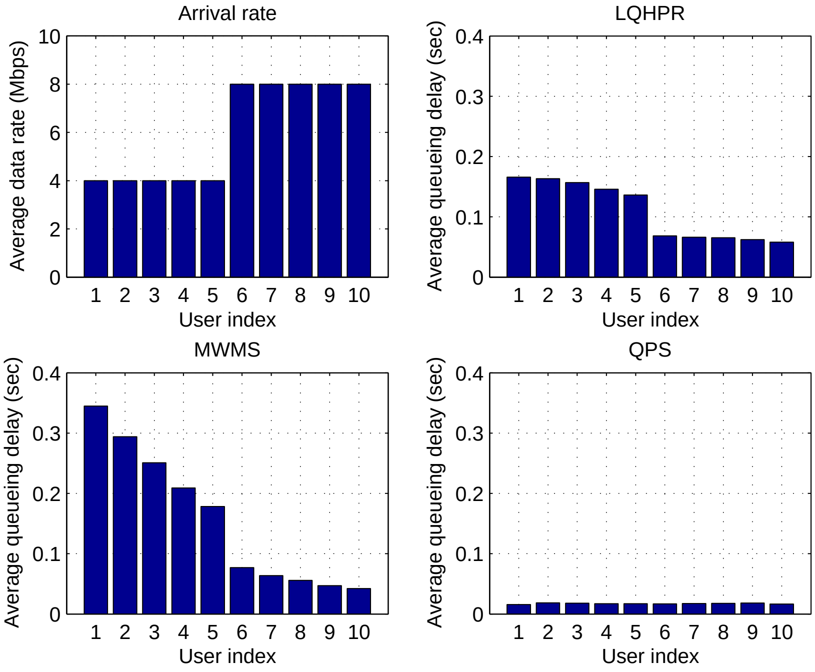


Figure 6.8: Average queueing delay profile for MIMO-OFDMA BC with $K = 10$

and each user's channel SNR per receive antenna at each tone is $[10\ 11\ \cdots\ 19]^T$ dB, respectively. The average queueing delay is evaluated when each user's average data rate increases while satisfying $\lambda_k = \lambda_1$ for $k = 2, \cdots, 5$, and $\lambda_k = 2\lambda_1$ for $k = 6, \cdots, 10$. The fixed operation point is chosen from the boundary of outage rate region such that user 1's target rate is 2.8 Mbps. Fig. 6.7 shows that with this fixed operation point, user 1's achieved throughput is about 2.4 Mbps, which can be boosted up to 4 Mbps by applying QPS, MWMS, or LQHPR. Fig. 6.7 also shows that QPS provides much smaller average queueing delay than MWMS and LQHPR. This is because with the TDM rate region, only QPS supports multiple users at the same time.

Fig. 6.8 presents each user's average queueing delay for the same ten-user MIMO-OFDMA BC as in Fig. 6.7. It can be seen that every user's average queueing delay is perfectly equalized under QPS. Furthermore, as shown in Chapter 3, QPS can arbitrarily scale the ratio of each user's average queueing delay to satisfy QoS

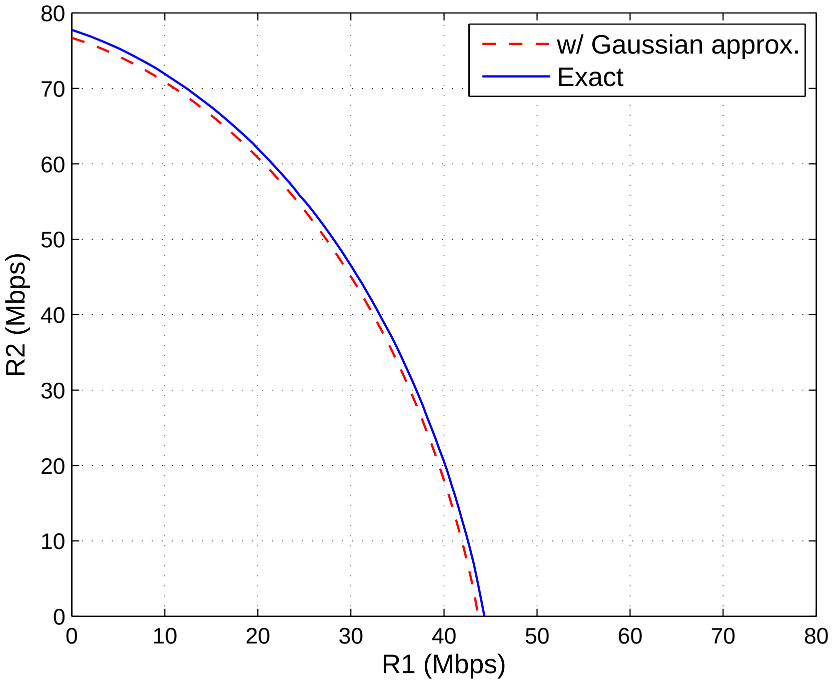


Figure 6.9: Outage rate region of MIMO-OFDMA MAC with $K = 2$

parameters.

6.4.2 Uplink MIMO-OFDMA with CDIT

This subsection presents the numerical results for MIMO-OFDMA MAC with CDIT. Fig. 6.9 provides the outage rate region of two-user MIMO-OFDMA MAC obtained from both the exact distribution of mutual information and approximated Gaussian distribution combined with SFC. Each user's total transmit power, P_k is assumed to be $N = 64$ for $k = 1, 2$, and user 1 and 2's channel SNR per receive antenna at each tone are set to 10 and 16 dB, respectively. As in the downlink case, the proposed approach is shown to provide a very good approximation of the actual outage rate region. The outage rate region presented in Fig. 6.9 is much larger than the outage rate region achieved by TDM transmission. As explained in the previous subsection, the convexity of the outage rate region will be more pronounced if the

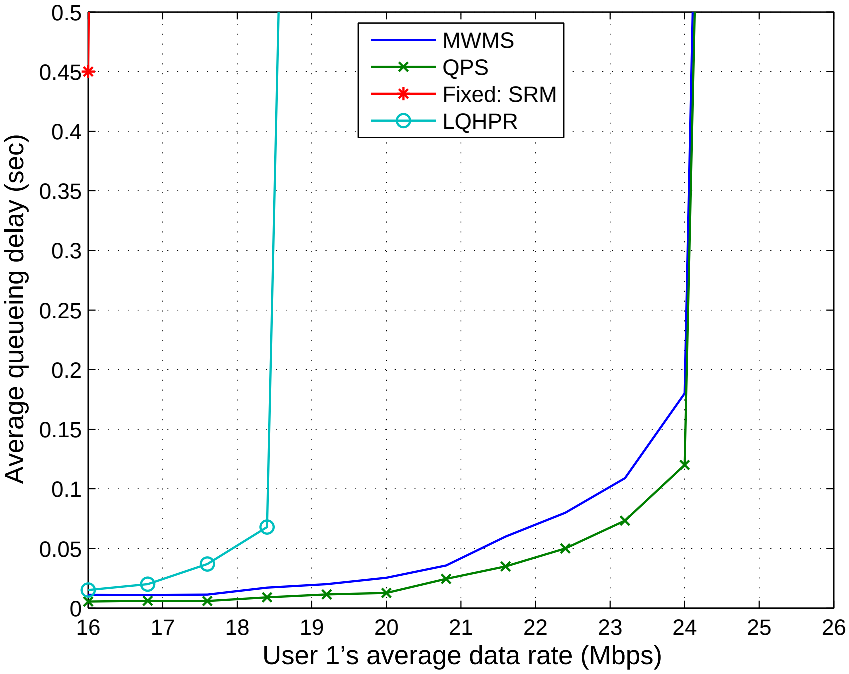


Figure 6.10: Average queueing delay vs User 1's average data rate for MIMO-OFDMA MAC with $K = 2$

effect of correlation among the tones is considered.

Fig. 6.10 and Fig. 6.11 present the stochastic simulation results for MIMO-OFDMA MAC with CDIT. In Fig. 6.10, two-user MIMO-OFDMA MAC in Fig. 6.9 is considered. Each user's average data rate increases while maintaining $\lambda_2 = 1.84\lambda_1$. The fixed operation point is set to $[16.6 \ \ 64.8]^T$ Mbps, which is on the boundary of outage rate region. Fig. 6.10 shows that user 1's achieved throughput is slightly less than 16 Mbps with the fixed operation point and about 18.5 Mbps with LQHPR. The throughput can be enhanced up to 24 Mbps by using throughput optimal schedulers, QPS or MWMS. In Fig. 6.10, it can also be observed that QPS provides 30-40% smaller average queueing delay than MWMS.

In Fig. 6.11, ten-user MIMO-OFDMA MAC is considered where $P_k = N = 64$ for $k = 1, \cdots, 10$, and each user's channel SNR per receive antenna at each tone is $[10 \ 11 \ \cdots \ 19]^T$ dB, respectively. Each user's average data rate increases while

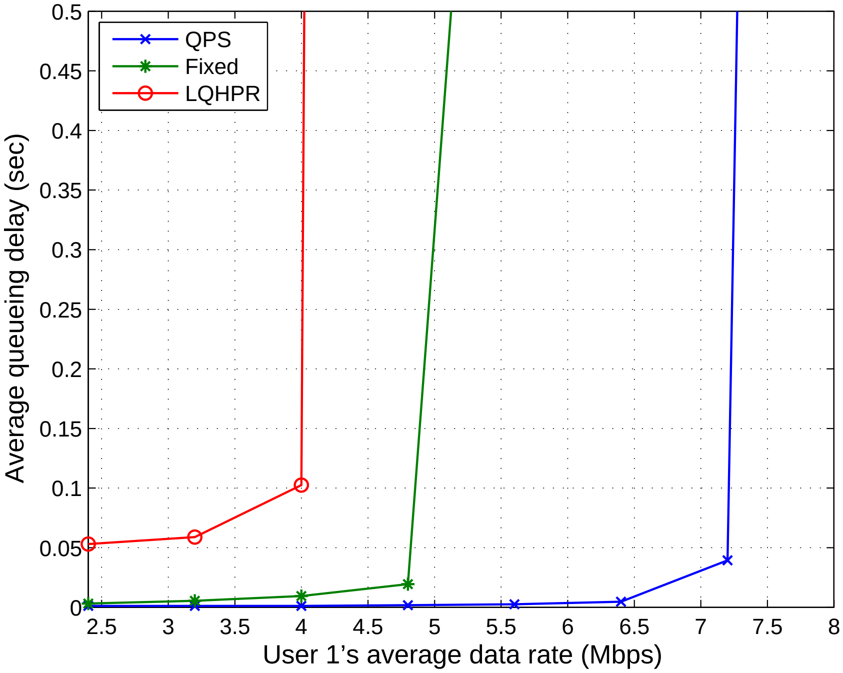


Figure 6.11: Average queueing delay vs User 1's average data rate for MIMO-OFDMA MAC with $K = 10$

satisfying $\lambda_k = \lambda_1$ for $k = 2, \cdots, 5$, and $\lambda_k = 2\lambda_1$ for $k = 6, \cdots, 10$. The fixed operation point is chosen from the boundary of outage rate region such that user 1's target rate is 5.5 Mbps. In Fig. 6.11, user 1's achieved throughput is shown to be 4.8 Mbps with this fixed operation point, 4 Mbps with LQHPR, and 7.2 Mbps with QPS. Hence, around 80% throughput gain is achieved by QPS compared to LQHPR. The numerical results for MWMS are excluded in Fig. 6.11 because of its exponential complexity in the number of users. On the other hand, the rate allocation under QPS can be efficiently searched with the linear complexity by using SFC. Fig. 6.12 provides each user's average queueing delay for the same ten-user MIMO-OFDMA MAC as in Fig. 6.11. It can be observed that every user's average queueing delay is kept much smaller as well as perfectly equalized under QPS.

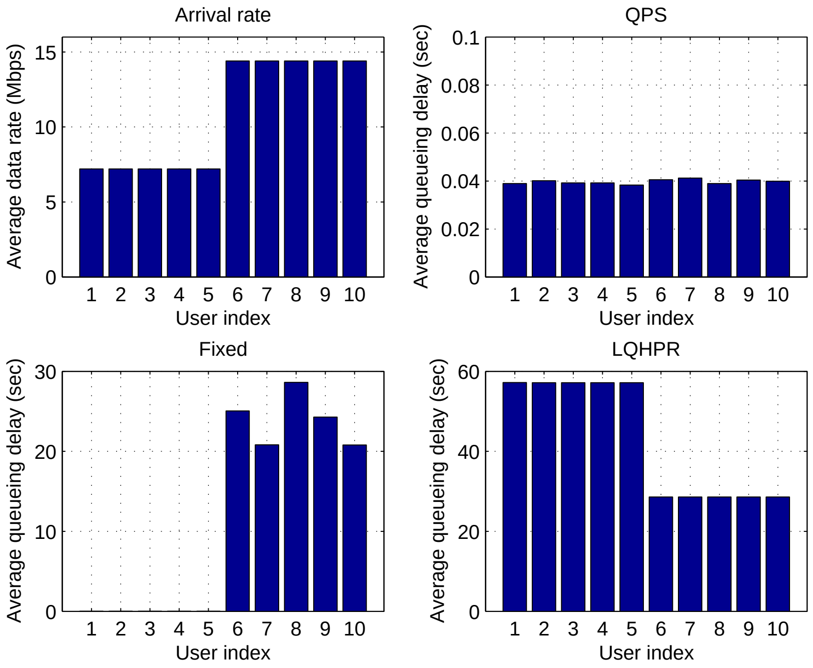


Figure 6.12: Average queueing delay profile for MIMO-OFDMA MAC with $K = 10$

6.5 Summary

When only CDIT is available, characterization of the outage rate region is essential for cross-layer resource allocation. This chapter efficiently characterizes the outage rate regions of the MIMO-OFDMA BC and MAC with CDIT, by using a Gaussian approximation of channel mutual information combined with a successive feasibility check. The proposed approach is also directly applicable to the solution of power/rate optimizations for QPS with the linear complexity in the number of tones and users. On the other hand, power/rate optimizations for gradient-type schedulers such as MWMS require the exponential complexity in the number of users. Stochastic simulations to evaluate average queueing delay are performed by using the proposed method, which demonstrate superior delay and fairness properties of QPS to other scheduling policies. With numerical efficiency as well as good delay and fairness properties, QPS is a promising scheduling policy for use in cross-layer resource allocation for MIMO-OFDMA systems with CDIT.

Chapter 7

Conclusion and Future Work

With the dramatic increase in multimedia services, future wireless networks will have more diversified QoS demands to support a variety of ubiquitous broadband services such as portable telephony, mobile Internet, VoIP, and IPTV. In order to guarantee QoS satisfaction, proper design of dynamic resource allocation will become one of the key issues in the future wireless networks. This thesis illustrates the important role of cross-layer approach to resource allocation in multi-user communication systems.

7.1 Conclusion

In Chapter 3, queue-proportional scheduling (QPS) is suggested as a promising queue-channel-aware scheduler with good throughput, delay, and fairness properties. QPS is shown to be a throughput optimal scheduling policy, and the stochastic simulations demonstrate that QPS achieves smaller average queueing delay than other scheduling policies such as maximum weight matching scheduling (MWMS). Furthermore, QPS is capable of arbitrarily scaling the ratio of each user's average queueing delay, which is essential for satisfying various QoS requirement at the same time.

This thesis also investigates the application of QPS and other schedulers to the multi-user systems based on OFDM modulation as well as MIMO transmission. By using various optimization techniques, efficient power/rate allocation algorithms are developed to employ those schedulers in downlink and uplink SISO-OFDM and

MIMO-OFDMA systems with CSIT. In Chapter 4, geometric programming (GP) is introduced and applied to cross-layer resource allocation for SISO-OFDM BC and MAC with CSIT. With strong numerical efficiency and scalability, GP emerges as a powerful tool for this application.

In MIMO-OFDMA BC and MAC with CSIT, finding the optimal power/rate allocation on each tone is basically a combinatorial problem with the exponential complexity. The analysis in Chapter 5 suggests that in resource allocation problems for MIMO-OFDMA with CSIT, the duality gap vanishes with the practical number of tones. From this observation, efficient algorithms based on Lagrange dual decomposition are derived to solve weighted sum-rate maximization and weighted sum-power minimization problems with the polynomial complexity.

In highly mobile environments, the value of CSI feedback is limited and the base station needs to perform resource allocation only using channel distribution information (CDI). In Chapter 6, cross-layer resource allocation in MIMO-OFDMA BC and MAC with CDIT is investigated. A simple method to accurately characterize outage rate region is presented, based on a Gaussian approximation of mutual information in conjunction with a successive feasibility check. This method is directly applicable to efficiently finding power/rate allocation under QPS with the linear complexity, while supporting other gradient-type schedulers such as MWMS requires the exponential complexity. In multi-user systems with only CDIT, QPS has clear advantages over other schedulers in terms of numerical efficiency in addition to good delay and fairness properties.

This thesis presents stochastic simulation results in a variety of situations by using the proposed efficient cross-layer resource allocation algorithms. In typical wireless networks based on MIMO-OFDM transmission, queue-channel-aware scheduling achieves about 20-50% throughput gain over channel-aware scheduling, and it provides better queueing delay and fairness properties. Particularly, QPS enables precise control of each user's average queueing delay relative to others, which is crucial in satisfying different QoS demands. Cross-layer resource allocation is essential for future wireless networks driven by various ubiquitous broadband services.

7.2 Future Work

Further analytical and experimental study on queueing delay property is recommended to better understand performance of the QPS and other scheduling policies. Stochastic simulations in this thesis mostly present the average queueing delay of each scheduler. However, it is of great interest to characterize each user's packet delay distributions. With specific QoS parameters imposed on each service, the probability of violating delay constraints can be evaluated to provide more realistic comparison of scheduling policies. Further, it is important as well as challenging to define the fundamental limits on achievable queueing delay by multi-user packet schedulers.

Also, efficient power/rate optimization algorithms for cross-layer resource allocation in MIMO-OFDM BC and MAC with CSIT need to be developed. This thesis focuses on the MIMO-OFDMA system without SDMA capability, where its system throughput can be further increased by using SDMA. Though the capacity regions of Gaussian MIMO-OFDM BC and MAC have been completely characterized [76, 43], the complexity of finding the optimal solution is still prohibitively high. In practice, it is required to find near-optimal solutions that can be easily found. Application of zero-forcing beamforming (ZF-BF) [78] or zero-forcing generalized decision feedback equalizer (ZF-GDFE) [12] can be good candidates for suboptimal multi-user MIMO systems. Efficient power/rate optimization algorithms for these suboptimal systems are required for schedulers to support any rate tuple in the achievable rate region.

Chapter 6 assumes that the packet outage is known at the transmitter and retransmission of the failed packet is ignored. New wireless systems such as 3GPP LTE [69] and mobile WiMAX [34] employ an advanced retransmission strategy based on hybrid automatic repeat request (H-ARQ) to achieve robustness and high spectral efficiency [23]. It is an interesting topic to investigate the effects of H-ARQ on scheduling performance and to design intelligent schedulers when H-ARQ is used.

This thesis mainly considers broadcast channels and multiple-access channels where the full coordination at the base-station is feasible. Of great interest is to study cross-layer resource allocation in interference channels or relay channels without full coordination. Interference channels can be encountered in many applications such as multi-cell wireless networks and DSL networks. Also, relay channels can be found in a multi-hop ad hoc wireless network where multiple hops are required for efficient communication between mobile nodes far apart [28]. With no or partial coordination, it becomes more challenging to satisfy each user's different QoS demand and to develop efficient power/rate optimization algorithms.

Appendix A

Throughput Optimality of MWMS

Throughput optimality of maximum weight matching scheduling (MWMS) can be proved by applying the Lyapunov analysis introduced in Chapter 2.1.3. Without loss of generality, such proof assumes that the scheduling period and system bandwidth are equal to 1. A time interval $[t, t+1)$ with $t = 0, 1, 2, \cdots$ is denoted by the *time slot* t. The rate allocation is determined at the beginning of each time slot and remains unchanged until the new time slot begins. $\mathbf{R}(t)$ for $t = 0, 1, 2, \cdots$ is a vector denoting the number of bits supported by each user in the time slot t. At time slot t, MWMS assigns the data-rate vector that maximizes the inner product of the queue-state vector and the achievable rate vector as formulated below.

$$\begin{aligned} \mathbf{R}_{MWMS}(t) = \arg\max_{\mathbf{r}} \mathbf{Q}'(t)^T\mathbf{r} \\ \text{such that } \mathbf{r} \in C(t), \end{aligned} \tag{A.1}$$

where $\mathbf{r} = [r_1\ r_2\ \cdots\ r_K]^T$, $C(t)$ denotes the instantaneous achievable rate region at time t, and $\mathbf{Q}'(t) = [\beta_1 Q_1(t)\ \cdots\ \beta_K Q_K(t)]^T$. β_i is the user i's priority weight which is set to 1 for all users if everyone has the same priority. Define $Z_i(t)$ as the number of arrived bits at user i's queue in the time slot t. Then, after a scheduling period, user i's queue-state vector is equal to $Q_i(t+1) = Q_i(t) - R_i(t) + Z_i(t)$. With the network capacity region denoted by C_{net}, proving throughput optimality of MWMS is equivalent to showing that for any $\boldsymbol{\lambda} \in \text{int}\ C_{net}$, the queue lengths for all users can be kept finite under MWMS.

First, choose the Lyapunov function $L(\mathbf{Q}(t)) = \sum_{i=1}^{K} \beta_i Q_i^2(t)$. The evolution of $L(\mathbf{Q}(t))$ after one scheduling interval is

$$\begin{aligned} L(\mathbf{Q}(t+1)) &= \sum_{i=1}^{K} \beta_i Q_i^2(t+1) = \sum_{i=1}^{K} \beta_i \left(Q_i(t) - R_i(t) + Z_i(t)\right)^2 \qquad \text{(A.2)} \\ &\leq L(\mathbf{Q}(t)) - 2\sum_{i=1}^{K} \beta_i Q_i(t) \left(R_i(t) - Z_i(t)\right) + \sum_{i=1}^{K} \beta_i \left(Z_i^2(t) + R_i^2(t)\right). \end{aligned}$$

Conditioned on $\mathbf{Q}(t) = \mathbf{q_t}$, the expected drift of the Lyapunov function is

$$\begin{aligned} \mathbb{E}\left[L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t)) | \mathbf{Q}(t) = \mathbf{q_t}\right] \leq -2\sum_{i=1}^{K} \beta_i q_{t,i} \left(\mathbb{E}[R_i(t)|\mathbf{Q}(t) = \mathbf{q_t}] - \lambda_i\right) \\ + \sum_{i=1}^{K} \beta_i \left(\mathbb{E}[Z_i^2(t)] + \mathbb{E}[R_i^2(t)|\mathbf{Q}(t) = \mathbf{q_t}]\right). \quad \text{(A.3)} \end{aligned}$$

$\mathbb{E}[Z_i^2(t)]$ and $\mathbb{E}[R_i^2(t)|\mathbf{Q}(t) = \mathbf{q_t}]$ are bounded since the bit arrival process $Z_i(t)$ has a finite mean and variance and the size of network capacity region is finite. Hence, for some positive $\Theta < \infty$,

$$\sum_{i=1}^{K} \beta_i \left(\mathbb{E}[Z_i^2(t)] + \mathbb{E}[R_i^2(t)|\mathbf{Q}(t) = \mathbf{q_t}]\right) \leq \Theta. \qquad \text{(A.4)}$$

From definition of MWMS in (A.1), $\sum_{i=1}^{K} \beta_i q_{t,i} \gamma_i$ is maximized over all the vectors $\boldsymbol{\gamma} = [\gamma_1 \ \cdots \ \gamma_K]^T$ in the network capacity region C_{net}. Therefore, for any arrival rate vector $\boldsymbol{\lambda} \in C_{net}$, the following inequality holds.

$$\sum_{i=1}^{K} \beta_i q_{t,i} \mathbb{E}[R_i(t)|\mathbf{Q}(t) = \mathbf{q_t}] \geq \sum_{i=1}^{K} \beta_i q_{t,i} \lambda_i. \qquad \text{(A.5)}$$

Since $\boldsymbol{\lambda}$ is assumed to be strictly in the interior of the network capacity region, a positive vector $\boldsymbol{\epsilon} = [\epsilon \ \cdots \ \epsilon]^T$ can be added to produce another vector $\boldsymbol{\lambda} + \boldsymbol{\epsilon}$ in the network capacity region. Thus, from (A.5),

$$\sum_{i=1}^{K} \beta_i q_{t,i} \left(\mathbb{E}[R_i(t) | \mathbf{Q}(t) = \mathbf{q_t}] - \lambda_i \right) \geq \epsilon \sum_{i=1}^{K} \beta_i q_{t,i}. \tag{A.6}$$

By applying (A.4) and (A.6) in (A.3), the expected drift of the Lyapunov function can be expressed as

$$\mathbb{E}\left[L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t)) | \mathbf{Q}(t) = \mathbf{q_t}\right] \leq \Theta - 2\epsilon \sum_{i=1}^{K} \beta_i q_{t,i}. \tag{A.7}$$

By choosing any $\alpha > 0$ and defining the compact region, then

$$\Lambda = \left\{ \mathbf{q_t} \in \mathbb{R}^K \mid q_{t,i} \geq 0, \ \sum_{i=1}^{K} \beta_i q_{t,i} \leq \left(\frac{\Theta + \alpha}{2\epsilon} \right) \right\}. \tag{A.8}$$

For $\mathbf{q_t} \notin \Lambda$, the expected drift of the Lyapunov function is less than $-\alpha$; thus, MWMS is throughput optimal.

Appendix B

Lagrange Dual Decomposition

An optimization problem in the standard form can be represented as

$$\begin{array}{rrcll} \text{minimize} & f_0(\mathbf{x}) & & & \\ \text{subject to} & f_i(\mathbf{x}) & \leq & 0, & i = 1, \cdots, m \\ & h_i(\mathbf{x}) & = & 0, & i = 1, \cdots, p, \end{array} \tag{B.1}$$

with variable $x \in \mathbb{R}^n$. The domain $\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom} f_i \cap \bigcap_{i=1}^{p} \mathbf{dom} h_i$ where $\mathbf{dom} f$ denotes the domain of the function f is assumed nonempty and the optimal value of (B.1) is denoted by p^*.

The main idea of Lagrange dual decomposition is to include the constraints in the objective function by using the *Lagrange multipliers* associated with each constraint. The *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ of the problem (B.1) is

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}), \tag{B.2}$$

where λ_i and μ_i are the Lagrange multipliers associated with the ith inequality and the ith equality constraints, respectively. The vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are called the *dual variables.*

Then, the *Lagrange dual function* $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ is defined as the minimum value of the Lagrangian over $\mathbf{x}$.

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{x \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \mu_i h_i(\mathbf{x}) \right). \tag{B.3}$$

When the Lagrangian is unbounded below in $\mathbf{x}$, the dual function takes on the minus infinite value. The dual function is always concave since it is the pointwise infimum of a family of affine functions in $(\boldsymbol{\lambda}, \boldsymbol{\mu})$.

Because of the constraints, for any $\boldsymbol{\lambda} \succeq \mathbf{0}$ and any $\boldsymbol{\mu}$, the following condition holds

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{x \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq f_0(\tilde{\mathbf{x}}), \tag{B.4}$$

where $\tilde{\mathbf{x}}$ denotes any feasible point.

Therefore, $g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq p^*$ for any $\boldsymbol{\lambda} \succeq \mathbf{0}$ and any $\boldsymbol{\mu}$. The *Lagrange dual problem* associated with (B.1) is defined as

$$\begin{aligned} &\text{maximize} \quad g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &\text{subject to} \quad \boldsymbol{\lambda} \succeq \mathbf{0}. \end{aligned} \tag{B.5}$$

The original problem (B.1) is sometimes called the *primal problem.* If $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ are optimal for (B.5), they are called *dual optimal variables.* Regardless of convexity of the primal problem, the above dual problem is always a convex optimization problem since the objective and constraint are convex.

By definition, the optimal objective value of the Lagrange dual problem, d^* is the best lower bound on p^* that can be obtained from the Lagrange dual function. The *optimal duality gap* is defined as $p^* - d^*$, the difference between primal and dual optimal values. If the primal optimization problem is a convex optimization problem where the objective and constraints form convex sets, the zero duality gap is guaranteed; thus, the optimal solution to the original problem can be obtained by solving the Lagrange dual problem. However, for non-convex primal optimization problems, the nonzero duality gap may exist.

The Lagrange dual decomposition can be given a simple geometric interpretation in terms of the set

$$\mathcal{G} = \{(f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}), h_1(\mathbf{x}), \cdots, h_p(\mathbf{x}), f_0(\mathbf{x})) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} | \mathbf{x} \in \mathcal{D}\}, \quad \text{(B.6)}$$

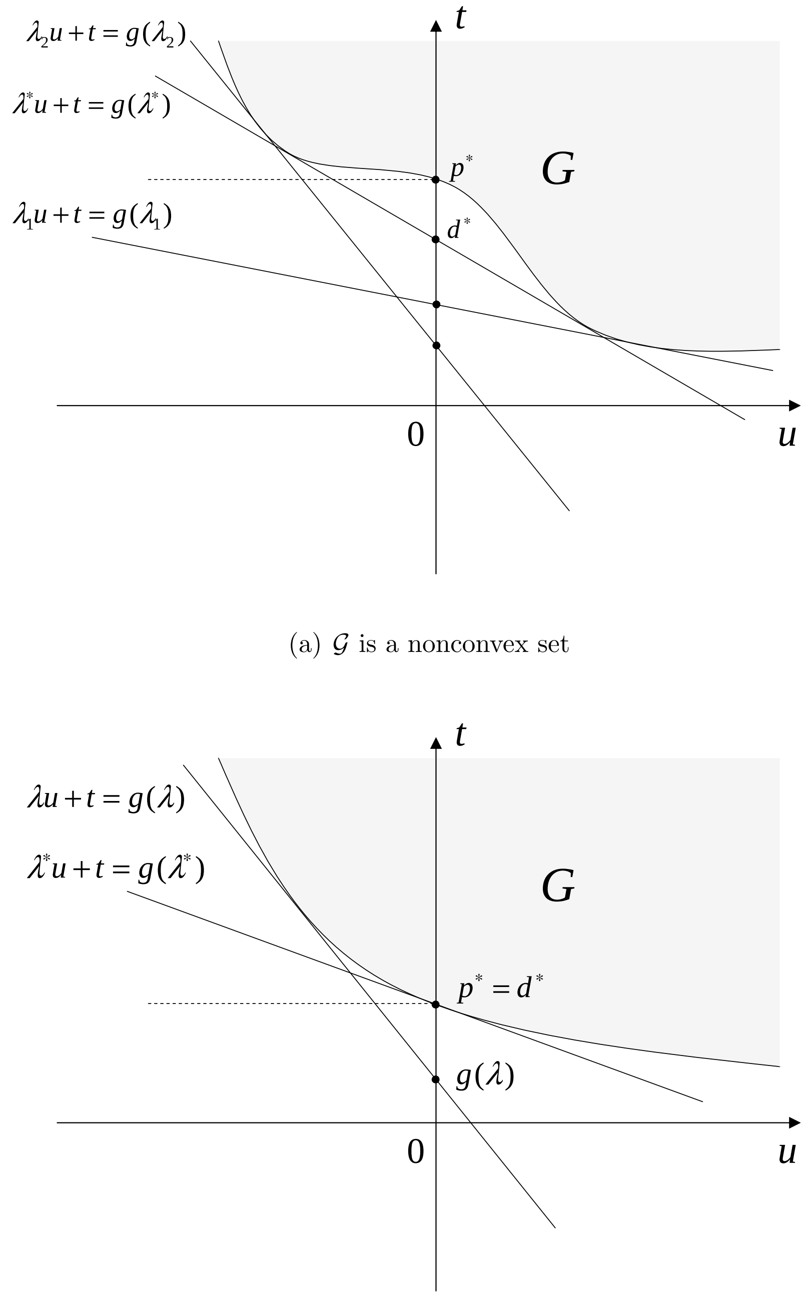
which is the set of values taken on by the constraints and objective functions. Then, p^* can be expressed as

$$p^* = \inf \{t | (\mathbf{u}, \mathbf{v}, t) \in \mathcal{G}, \mathbf{u} \preceq \mathbf{0}, \mathbf{v} = \mathbf{0}\}. \quad \text{(B.7)}$$

The dual function at $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is given by

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf \left\{(\boldsymbol{\lambda}, \boldsymbol{\mu}, 1)^T (\mathbf{u}, \mathbf{v}, t) | (\mathbf{u}, \mathbf{v}, t) \in \mathcal{G}\right\}. \quad \text{(B.8)}$$

The maximization of this dual function over $\boldsymbol{\lambda} \succeq \mathbf{0}$ provides the dual optimal value, d^*. Geometric interpretation of the primal and dual optimal values with only one inequality constraint is illustrated in Fig. B.1, which shows that the duality gap becomes zero when $\mathcal{G}$ is a convex set. It is shown that this statement holds when multiple constraints are present. The convexity of $\mathcal{G}$ does not necessarily imply that the primal problem is a convex optimization problem though the opposite argument is always true. Thus, the convexity of $\mathcal{G}$ is a more general condition to guarantee the zero duality gap.


$\lambda_2 u + t = g(\lambda_2)$
$\lambda^* u + t = g(\lambda^*)$
$\lambda_1 u + t = g(\lambda_1)$
t
p^*
d^*
G
0
u
t
$\lambda u + t = g(\lambda)$
$\lambda^* u + t = g(\lambda^*)$
G
$p^* = d^*$
$g(\lambda)$
0
u

(a) $\mathcal{G}$ is a nonconvex set

(b) $\mathcal{G}$ is a convex set

Figure B.1: Geometric interpretation of dual function with one inequality constraint. Given λ, $\lambda u + t$ is minimized over $\mathcal{G} = \{(f_1(\mathbf{x}), f_0(\mathbf{x}))|\mathbf{x} \in \mathcal{D}\}$, which provides a supporting hyperplane with a slope of $-\lambda$. $g(\lambda)$ is the intersection of this hyperplane with the t-axis.

Appendix C

Dual Update Methods

In order to solve the dual optimization problem, minimize $g(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda} \geq 0$, efficient updates of the dual variable $\boldsymbol{\lambda}$ plays a critical role. All the components of $\boldsymbol{\lambda}$ can be updated simultaneously along some search direction, and because of the convexity of $g(\boldsymbol{\lambda})$, this gradient-type search of optimal dual variable is guaranteed to converge to the global optimum. Unfortunately, the gradient of $g(\boldsymbol{\lambda})$ is unavailable when the dual function is not differentiable. However, it is always feasible to use the sub-gradient in finding a search direction. A vector $\mathbf{d}$ is called a subgradient of $g(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}$, if the following condition is satisfied for all $\boldsymbol{\lambda'}$.

$$g(\boldsymbol{\lambda'}) \geq g(\boldsymbol{\lambda}) + \mathbf{d}^T(\boldsymbol{\lambda'} - \boldsymbol{\lambda}). \tag{C.1}$$

Subgradient is a generalization of gradient, which can be also applied to the non-differentiable functions. $\mathbf{d}$ is a subgradient if the linear function with slope $\mathbf{d}$ passing through $(\boldsymbol{\lambda}, g(\boldsymbol{\lambda}))$ lies entirely below $g(\boldsymbol{\lambda})$.

The updates of dual variables can also be based on the cutting-plane methods. The main idea is to localize the set of candidate $\boldsymbol{\lambda}$'s within some closed and bounded set, and eliminate about the half of the region from the candidate set by evaluating the subgradient of $g(\boldsymbol{\lambda})$ at an properly chosen center of such a region. The iterations continue until the size of the candidate set converges to an optimal $\boldsymbol{\lambda}$. From the definition of the subgradient in (C.1), $g(\boldsymbol{\lambda'}) \geq g(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda'}$ satisfying the following condition.

$$\mathbf{d}^T(\boldsymbol{\lambda'} - \boldsymbol{\lambda}) \geq 0. \tag{C.2}$$

Thus, all $\boldsymbol{\lambda'}$'s in the half-plane defined by (C.2) can be removed at each step. This cutting-plane method is a generalization of the single-variable bisection method to multiple dimensions.

The ellipsoid method is a very efficient cutting-plane dual update method, where the candidate region is defined as the minimized ellipsoid that includes all the candidate $\boldsymbol{\lambda}$'s. The key idea in the ellipsoid method is to localize $\boldsymbol{\lambda}^*$, the optimal solution, in a sequence of ellipsoids $\mathcal{E}^{(k)}$ with vanishing volumes so that $\boldsymbol{\lambda}^{(k)}$, the centers of these ellipsoids, eventually converge to $\boldsymbol{\lambda}^*$ [8].

This iterative algorithm starts with an initial ellipsoid $\mathcal{E}^{(0)}$ that contains $\boldsymbol{\lambda}^*$. At each iteration, $\boldsymbol{\lambda}^{(k)}$ is chosen as the center of ellipsoid $\mathcal{E}^{(k)}$ and a subgradient of $g(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^{(k)}$ denoted by $\mathbf{d}^{(k)}$ is determined. By definition of subgradient, $\mathbf{d}^{(k)}$ satisfies $g(\boldsymbol{\lambda}^{(k)} + \Delta\boldsymbol{\lambda}) \geq g(\boldsymbol{\lambda}^{(k)}) + \Delta\boldsymbol{\lambda}^T\boldsymbol{d}^{(k)}$ for any $\Delta\boldsymbol{\lambda}$, which means that $\boldsymbol{\lambda}^*$ should be in the half-ellipsoid as follows.

$$\mathcal{E}^{(k)} \bigcap \left\{ \boldsymbol{\lambda} : \boldsymbol{d}^{(k)T} \left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(k)} \right) \leq 0 \right\}. \tag{C.3}$$

In the next iteration, $\mathcal{E}^{(k+1)}$ is defined as the minimum volume ellipsoid covering the half-ellipsoid in (C.4). Suppose $\mathbf{A}^{(k)}$ is the matrix describing $\mathcal{E}^{(k)}$ as

$$\mathcal{E}^{(k)} = \left\{ \boldsymbol{\lambda} : \left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(k)} \right)^T \mathbf{A}^{(k)^{-1}} \left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^{(k)} \right) \leq 1 \right\}. \tag{C.4}$$

Given $\mathcal{E}^{(k)}$ and $\mathbf{d}^{(k)}$, $\boldsymbol{\lambda}^{(k+1)}$ and $\mathcal{E}^{(k+1)}$ can be expressed as

$$\begin{aligned} \boldsymbol{\lambda}^{(k+1)} &= \boldsymbol{\lambda}^{(k)} - \frac{1}{n+1}\mathbf{A}^{(k)}\tilde{\mathbf{d}}^{(k)}, & \text{(C.5)} \\ \mathbf{A}^{(k+1)} &= \frac{n^2}{n^2-1}\left(\mathbf{A}^{(k)} - \frac{2}{n+1}\mathbf{A}^{(k)}\tilde{\mathbf{d}}^{(k)}\tilde{\mathbf{d}}^{(k)T}\mathbf{A}^{(k)} \right), & \text{(C.6)} \end{aligned}$$

where $\tilde{\mathbf{d}}^{(k)} = \mathbf{d}^{(k)}/\sqrt{\mathbf{d}^{(k)T}\mathbf{A}^{(k)}\mathbf{d}^{(k)}}$ and n is the dimension of $\boldsymbol{\lambda}$. The volumes of these ellipsoids can be shown to decrease exponentially, i.e. $\mathbf{Vol}(\mathcal{E}^{(k+1)}) < e^{-(1/2n)}\mathbf{Vol}(\mathcal{E}^{(k)})$ and converge in $\mathcal{O}(n^2)$ iterations. The iteration stops when $\sqrt{\mathbf{d}^{(k)T}\mathbf{A}^{(k)}\mathbf{d}^{(k)}} < \epsilon$.

Bibliography

[1] S. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE J. Select. Areas Commun.*, 16:1451–1458, October 1998.

[2] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting. CDMA data QoS scheduling on the forward link with variable channel conditions. *Bell Labs Tech. Report*, April 2000.

[3] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting. Providing Quality of Service over a shared wireless link. *IEEE Commun. Mag.*, pages 150–154, February 2001.

[4] S. Asmussen. *Applied Probability and Queues.* Springer, New York, 2000.

[5] R. Berry and E. Yeh. Cross-layer wireless resource allocation. *IEEE Signal Processing Mag.*, 45:59–68, September 2004.

[6] D. Bertsekas and R. Gallager. *Data Networks.* Prentice Hall, NJ, 1987.

[7] E. Biglieri, J. Proakis, and S. Shamai. Fading channels: information-theoretic and communications aspects. *IEEE Trans. Inform. Theory*, 44:2619–2692, October 1998.

[8] S. Boyd. *Optimization Projects: EE392O Lecture Notes.* Stanford Univ., Stanford, CA, September 2003.

[9] S. Boyd, S. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. *Optimization and Engineering*, 8:67–127, March 2007.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2003.

[11] G. Caire and S. Shamai. On the achievable throughput of a multiantenna Gaussian broadcast channel. *IEEE Trans. Inform. Theory*, 49:1691–1706, July 2003.

[12] C. Y. Chen, K. Seong, R. Zhang, and J. M. Cioffi. Optimized resource allocation for upstream vectored DSL systems with zero-forcing generalized decision feedback equalizer. *IEEE J. Select. Topics Sig. Proc.*, 1:686–699, December 2007.

[13] R. Cheng and S. Verdú. Gaussian multiaccess channels with ISI: Capacity region and multisuer water-filling. *IEEE Trans. Inform. Theory*, 39:773–785, May 1993.

[14] Y. W. Cheong, R. S. Cheng, K. B. Lataief, and R. D. Murch. Multiuser OFDM with adaptive subcarrier, bit, and power allocation. *IEEE J. Select. Areas Commun.*, 17:1747–1758, October 1999.

[15] M. Chiang. Geometric programming for communication systems. *Foundations and Trends in Communications and Information Theory*, 2:1–154, August 2005.

[16] M. Chiang, S. Zhang, and P. Hande. Distributed rate allocation for inelastic flows: Optimization frameworks, optimality conditions, and optimal algorithms. In *Proc. IEEE INFOCOM*, pages 2679–2690, Miami, FL, March 2005.

[17] J. M. Cioffi. *Digital Communications: EE379A Lecture Notes.* Stanford Univ., Stanford, CA, January 2003.

[18] J. M. Cioffi. *Advanced Digital Communication: EE379C Lecture Notes.* Stanford Univ., Stanford, CA, April 2005.

[19] T. Cover. Broadcast channels. *IEEE Trans. Inform. Theory*, 18:2–14, January 1972.

[20] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications, 1991.

[21] A. Eryilmaz, R. Srikant, and J. R. Perkins. Throughput-optimal scheduling for broadcast channels. In *Proc. of SPIE*, pages 70–78, 2001.

[22] G. J. Foschini and M. J. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, 6:311–335, March 1998.

[23] F. Frederiksen and T. Kolding. Performance and modeling of WCDMA/HSDPA transmission/H-ARQ schemes. In *Proc. IEEE Veh. Technol. Conf. (VTC)*, volume 1, pages 472–476, 2002.

[24] R. G. Gallager. Capacity and coding for degraded broadcast channels. *Problemy Peredaci Informaccii*, 10:3–14, September 1974.

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

[26] A. Goldsmith and P. Varaiya. Capacity of fading channels with channel side information. *IEEE Trans. Inform. Theory*, 43:1986–1992, November 1997.

[27] A. J. Goldsmith and S. G. Chua. Variable-rate variable-power MQAM for fading channels. *IEEE Trans. Commun.*, 45:1218–1230, October 1997.

[28] A. J. Goldsmith and S. B. Wicker. Design challenges for energy-constrained ad hoc wireless networks. *IEEE Wireless Commun. Mag.*, 9:8–27, August 2002.

[29] S. V. Hanly and D. N. Tse. Multiaccess fading channels-Part II: Delay-limited capacities. *IEEE Trans. Inform. Theory*, 44:2816–2831, November 1998.

[30] B. M. Hochwald, T. L. Marzetta, and V. Tarokh. Multi-antenna channel hardening and its implications for rate feedback and scheduling. *IEEE Trans. Inform. Theory*, 50:1893–1909, September 2004.

[31] L. M. C. Hoo, B. Halder, J. Tellado, and J. M. Cioffi. Multiuser transmit optimization for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms. *IEEE Trans. Commun.*, 52:922–930, June 2004.

[32] D. Hughes-Hartogs. *The capacity of a degraded spectral Gaussian broadcast channel.* Ph.D. dissertation, Inform. Syst. Lab., Ctr. Syst. Res., Stanford Univ., Stanford, CA, July 1975.

[33] C. Ibars and Y. Bar-Ness. Outage capacities of a multi-carrier WLAN downlink under different resource sharing techniques. In *Proc. IEEE 7th International Symposium on Spread Spectrum Techniques and Applications*, volume 1, pages 144–149, 2002.

[34] *IEEE 802.16e-2005 and IEEE 802.16-2004/Cor1-2005.* IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1, February 2006.

[35] N. Jindal, S. Vishwanath, and A. Goldsmith. On the duality of Gaussian multiple-access and broadcast channels. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, page 500, June 2002.

[36] M. Kamath and B. Hughes. The asymptotic capacity of multiple-antenna Rayleigh-fading channels. *IEEE Trans. Inform. Theory*, 51:4325–4333, December 2005.

[37] R. Knopp and P. A. Humblet. Information capacity and power control in single-cell multiuser communications. In *Proc. IEEE Int. Conf. Commun. (ICC)*, volume 1, pages 331–335, June 1995.

[38] R. Leelahakriengkrai and R. Agrawal. Scheduling in multimedia wireless networks. In *Proc. 17th Int. Teletraffic Congress*, pages 285–298, Salvador da Bahia, Brazil, December 2001.

[39] L. Li and A. Goldsmith. Capacity and optimal resource allocation for fading broadcast channels-Part I: Ergodic capacity. *IEEE Trans. Inform. Theory*, 47:1083–1102, March 2001.

[40] L. Li and A. Goldsmith. Capacity and optimal resource allocation for fading broadcast channels-Part II: Outage capacity. *IEEE Trans. Inform. Theory*, 47:1103–1127, March 2001.

[41] L. Li, N. Jindal, and A. Goldsmith. Outage capacities and optimal power allocation for fading multiple-access channels. *IEEE Trans. Inform. Theory*, 51:1326–1347, April 2005.

[42] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand. Achieving 100% throughput in an input-queued switch. In *Proc. IEEE INFOCOM*, volume 1, pages 296–302, San Francisco, CA, March 1996.

[43] M. Mohseni. *Capacity of Gaussian vector broadcast channels.* Ph.D. dissertation, STAR Lab., Stanford Univ., Stanford, CA, September 2006.

[44] M. Mohseni, R. Zhang, and J. M. Cioffi. Optimized transmission for fading multiple-access and broadcast channels with multiple antennas. *IEEE J. Select. Areas Commun.*, 24:1627– 1639, August 2006.

[45] M. J. Neely, E. Modiano, and C. E. Rohrs. Power allocation and routing in multibeam satellites with time-varying channels. *IEEE/ACM Trans. Networking*, 11:138–152, February 2003.

[46] J. Oh, S. Kim, and J. M. Cioffi. Optimum power allocation and control for OFDM in multiple access channels. In *Proc. 60th IEEE Veh. Tech. Conf. (VTC)*, September 2004.

[47] O. Oyman, R. U. Nabar, H. Bolcskei, and A. J. Paulraj. Characterizing the statistical properties of mutual information in MIMO channels. *IEEE Trans. Signal Processing*, 51:2784–2795, November 2003.

[48] A. Paulraj, R. Nabar, and D. Gore. *Introduction to Space-Time Wireless Communications.* Cambridge University Press, 2003.

[49] B. Prabhakar. *Network Algorithms: EE384M Lecture Notes.* Stanford Univ., Stanford, CA, September 2004.

[50] G. G. Raleigh and J. M. Cioffi. Spatio-temporal coding for wireless communication. *IEEE Trans. Commun.*, 46:357–366, March 1998.

[51] W. Rhee and J. M. Cioffi. Increase in capacity of multiuser OFDM system using dynamic subchannel allocation. In *Proc. IEEE Veh. Technol. Conf. (VTC)*, pages 1085–1089, Tokyo, Japan, May 2000.

[52] S. Ross. *Stochastic Processes.* John Wiley and Sons Inc., 1996.

[53] K. Seong, M. Mohseni, and J. M. Cioffi. Optimal resource allocation for OFDMA downlink systems. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 1394–1398, Seattle, WA, July 2006.

[54] K. Seong, R. Narasimhan, and J. M. Cioffi. Cross-layer resource allocation via geometric programming in fading broadcast channels. In *Proc. IEEE Veh. Technol. Conf. (VTC)*, Melbourne, Australia, May 2006.

[55] K. Seong, R. Narasimhan, and J. M. Cioffi. Queue proportional scheduling in Gaussian broadcast channels. In *Proc. IEEE Int. Conf. Commun. (ICC)*, Istanbul, Turkey, June 2006.

[56] K. Seong, R. Narasimhan, and J. M. Cioffi. Queue proportional scheduling via geometric programming in fading broadcast channels. *IEEE J. Select. Areas Commun.*, 24:1593–1602, August 2006.

[57] K. Seong, R. Narasimhan, and J. M. Cioffi. Scheduling for fading multiple access channels with heterogeneous QoS constraints. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Nice, France, June 2007.

[58] K. Seong, D. Yu, Y. Kim, and J. M. Cioffi. Optimal resource allocation via geometric programming for OFDM broadcast and multiple access channels. In *Proc. IEEE GLOBECOM*, San Francisco, CA, November 2006.

[59] S. Shakkottai and A. Stolyar. Scheduling algorithms for a mixture of real-time and nonreal-time data in HDR. In *Proc. 17th Int. Teletraffic Congress*, pages 793–804, Salvador da Bahia, Brazil, December 2001.

[60] S. Shakkottai and A. Stolyar. Scheduling for multiple flows sharing a time-varying channel: the exponential rule. *Amer. Mathematical Soc. Translations*, Series 2 (a volume in memory of F. Karpelevich), Y. M. Suhov, Ed., vol. 207, 2002.

[61] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, July 1948.

[62] M. Sharif and B. Hassibi. On the capacity of MIMO broadcast channels with partial side information. *IEEE Trans. Inform. Theory*, 51:506–522, February 2005.

[63] G. Song, Y. Li, L. Cimini Jr, and H. Zheng. Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels. In *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Atlanta, Georgia, 2004.

[64] T. Starr, J. M. Cioffi, and P. J. Silverman. *Understanding Digital Subscriber Line Technology.* Prentice-Hall, Englewood Cliffs, NJ, 1999.

[65] C. Swannack, E. Uysal-Biyikoglu, and G. W. Wornell. Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel. In *Proc. 42nd Annual Allerton Conf. Commununications, Control and Computing*, Allerton, IL, October 2004.

[66] V. Tarokh, N. Seshadri, and A. Calderbank. Space-time codes for high data rate wireless communication: performance criterion and code construction. *IEEE Trans. Inform. Theory*, 44:744–765, March 1998.

[67] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automat. Contr.*, 37:1936–1948, December 1992.

[68] E. Telatar. Capacity of multi-antenna Gaussian channels. *European Transactions on Telecommunications*, 10:585–598, November 1999.

[69] *The 3rd Generation Partnership Project, Technical Specification Group Radio Access Network; Long Term Evolution (LTE) Physical Layer-General Description.* 3GPP Std., TS 36.201, v.8.1.0, December 2007.

[70] D. Tse. Optimal power allocation over parallel Gaussian broadcast channels. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, page 27, Ulm, Germany, June 1997.

[71] D. Tse and S. Hanly. Multiaccess fading channels-Part I: polymatroid structure, optimal resource allocation and throughput capacities. *IEEE Trans. Inform. Theory*, 44:2796–2815, November 1998.

[72] P. Viswanath, D. N. C. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Trans. Inform. Theory*, 48:1277–1294, June 2002.

[73] H. Viswanathan and K. Kumaran. Rate scheduling in multiple antenna downlink. In *Proc. 39th Annual Allerton Conf. Commununications, Control and Computing*, pages 747–756, Allerton, IL, October 2001.

[74] J. Walrand. *An Introduction to Queueing Networks.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[75] Z. Wang and G. B. Giannakis. Outage mutual information of space-time MIMO channels. *IEEE Trans. Inform. Theory*, 50:657–662, April 2004.

[76] H. Weingarten, Y. Steinberg, and S. Shamai. The capacity region of the Gaussian MIMO broadcast channel. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, 2004.

[77] E. Yeh and A. Cohen. Throughput and delay optimal resource allocation in multi-access fading channels. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, page 245, Yokohama, Japan, 2003.

[78] T. Yoo and A. Goldsmith. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming. *IEEE J. Select. Areas Commun.*, 24:528–541, March 2006.

[79] W. Yu, G. Ginis, and J. M. Cioffi. Distributed multiuser power control for digital subscriber lines. *IEEE J. Select. Areas Commun.*, 20:1105–1115, June 2002.

[80] W. Yu and R. Lui. Dual methods for non-convex spectrum optimization of multi-carrier systems. *IEEE Trans. Commun.*, 54:1310–1322, July 2006.

[81] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi. Iterative water-filling for Gaussian vector multiple access channels. *IEEE Trans. Inform. Theory*, 50:145–152, January 2004.