ABSTRACT

Title of Dissertation: An Optimization Theoretical Framework for Resource Allocation over Wireless Networks

Zhu Han, Doctor of Philosophy, 2003

Dissertation directed by: Professor K. J. Ray Liu
Department of Electrical and Computer Engineering

With the advancement of wireless technologies, wireless networking has become ubiquitous owing to the great demand of pervasive mobile applications. Some fundamental challenges exist for the next generation wireless network design such as time varying nature of wireless channels, co-channel interferences, provisioning of heterogeneous type of services, etc. So how to overcome these difficulties and improve the system performance have become an important research topic.

Dynamic resource allocation is a general strategy to control the interferences and enhance the performance of wireless networks. The basic idea behind dynamic resource allocation is to utilize the channel more efficiently by sharing the spectrum and reducing interference through optimizing parameters such as the transmitting power, symbol transmission rate, modulation scheme, coding scheme, bandwidth,

etc. Moreover, the network performance can be further improved by introducing diversity, such as multiuser, time, frequency, and space diversity. In addition, cross layer approach for resource allocation can provide advantages such as low overhead, more efficiency, and direct end-to-end QoS provision.

The designers for next generation wireless networks face the common problem of how to optimize the system objective under the user Quality of Service (QoS) constraint. There is a need of unified but general optimization framework for resource allocation to allow taking into account a diverse set of objective functions with various QoS requirements, while considering all kinds of diversity and cross layer approach. We propose an optimization theoretical framework for resource allocation and apply these ideas to different network situations such as:

- Centralized resource allocation with fairness constraint
- Distributed resource allocation using game theory
- OFDMA resource allocation
- Cross layer approach

On the whole, we develop a universal view of the whole wireless networks from multiple dimensions: time, frequency, space, user, and layers. We develop some schemes to fully utilize the resources. The success of the proposed research will significantly improve the way how to design and analyze resource allocation over wireless networks. In addition, the cross-layer optimization nature of the problem provides an innovative insight into vertical integration of wireless networks.

An Optimization Theoretical Framework for Resource Allocation over Wireless Networks

by

Zhu Han

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2003

Advisory Committee:

Professor K. J. Ray Liu, Chairman/Advisor
Professor Steve Tretter
Professor Sennur Ulukus
Professor Gang Qu
Professor Carlos A. Berenstein

DEDICATION

To my wife as well as my parents and sister in China.

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to my advisor, professor K.J. Ray Liu. When I was bewildered in research a few years ago, he helped me find a good direction and placed his confidence in me like a light house in the foggy bay. I wish to thank him for putting so many efforts to help me conduct high quality works especially in how to present. I also value his influences for my vision, attitude, energy, and desire of my professional and personal developments.

I am grateful to my parents for their love and unconditional support. They have made countless efforts to let me have better education chances and make sure that I can explore my future according to my dreams. As a consequence, all my accomplishments are also their accomplishments.

It is said that to do researches and to write a dissertation are similar to draw a picture. The author has to dig into the topic, understand the physical meaning and use mathematical way to describe the findings. During my 23 years of school, I would like to explicitly thank the following people who help me develop the view and enhance my ability of mathematics.

I would like to thank professor Cao, Zhigang and Ma, Zhengxin for their kindly guidance during my undergraduate study in Tsinghua University. I would like to thank professor John Baras for giving me the chance to USA for better education. I would like to thank all the colleagues in ACTERNA (former TTC) for giving me the happy 2.5 years for my first job. I would like to thank Dr. Jane Wang, Dr. Farrokh Rahsid-Farrokhi, Zhu Ji, Andres Kwasinski, Guan-Ming Su, and Masoud

Olfat for cooperations. I would like to thank professor Wu, professor Ulukus, professor Tretter, professor Tits, professor La, and professor Mark for invaluable suggestions on my research. I would also like to thank all the members of CSPL lab and other friends throughout University of Maryland for giving me such happy 6.5 years.

Finally, no happiness of achievement can be completed without sharing with my beloved one, my wife Li Wang, and with her, I am always at the top of the world.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Background

Over the past few decades, wireless communications and networking have witnessed an unprecedented growth, and have become pervasive much sooner than anyone could have imagined [1]. Wireless networks are expected to be the dominant and ubiquitous telecommunication means in the next few decades. The widespread success of cellular and WLAN systems prompts the development of advanced wireless systems to provide other information services beyond voice, such as telecommuting, video conferencing, interactive media, real-time Internet games, etc., at anytime, anywhere. To satisfy growing demands of heterogeneous applications, the future wireless networks are characterized by broadband, high data rate capabilities, integration of services, flexibility, and scalability. Many technical challenges yet remain to achieve these requirements because of the adverse natures of wireless channels.

In this chapter, we give the introduction, motivation, and contribution of our research to overcome these challenges, as well as some basic background knowledge. The organization of this chapter is as follows: First, we present an introduction about nowadays wireless networks and the potential challenges for the next generation wireless network design. Second, the basic knowledge for our research is briefly reviewed. Wireless channel model is briefly discussed. For reliable transmission over such a channel, optimal transceiver design techniques are explained. In order to accommodate multiple users, different multiple access methods are reviewed. With the aim to increase the overall system capacity, frequency should be reused beyond some distance using cellular concept. For enhancing the end-to-end quality of links, different layers of communication protocol should be coordinated together by cross-layer approaches. Finally, we provide the motivations of this dissertation and point out the overall contributions. The organization of the dissertation is given and the contributions of each chapter are presented.

1.1 Introduction

A wireless channel can change rapidly and can be seriously affected by the radio propagation parameters and interferences, thus the topology and link characteristics are dynamically varying in wireless networks. The performance of a wireless network is mainly restrained by the interferences and the time-varying nature of wireless channels. The co-channel interference (CCI) is caused by users sharing the same channel due to the multiple access in wireless networks. Due to the effects such as multipath fading, shadowing, path loss, propagation delay, and noise level, the Signal-to-Interference-Noise-Ratio (SINR) at a receiver output can fluctuate in the order of tens of dBs. Therefore, it is of ample importance to study the fundamental technical issues that have major impacts on the performance of all the wireless systems.

A general strategy to combat these detrimental effects is the dynamic allocation of resources, such as transmitted powers and modulation rates, etc. based on the channel conditions. In power control, transmitted powers are constantly adjusted.

Such a process improves the qualities of weak links. At the same time, it increases CCI during deep fading. In adaptive modulation, the system assigns modulation rates with different constellation sizes and spectral efficiencies to different links, according to their channel conditions. All these resources are interrelated, and there are tradeoffs to allocate them in the interference limited wireless networks.

Moreover, there are other constraints such as fairness, heterogenous QoS provisioning, and practical implementation constraints. Since each user pays the same for his service, it is desirable to have fair resource allocation scheme. In order to provide fair services to all users, we need to define the new fairness concepts. From literature, there exist three popular kinds of fairness: max-min, proportional, and time average. For various applications, the QoS requirements are very different. For example, voice payload is very sensitive for delay, data payload requires low BER, and video payload has burst transmission. There are many practical constraints for wireless system implementation such as maximal transmitted power, minimal throughput, computation capability, implementation cost, etc. So how to optimally allocate the resources under all these constraints has become an important wireless research issue.

Orthogonal Frequency Division Multiplexing Access (OFDMA) is a popular multiple access and signaling scheme for wireless broadband networks. Adaptive modulation techniques in OFDMA provide the potential to vary the number of transmitted bits for a sub-channel, according to instantaneous sub-channel quality, while maintaining an acceptable Bit Error Rate (BER). Resource allocation for OFDMA networks has three major tasks: sub-channel assignment, throughput allocation, and power control.

To enhance the system performance, we explore the multi-dimension diversity.

By using throughput control in MAC layer, we can apply multiuser diversity and time diversity to allocate resources efficiently to different users over time according to their channel conditions. By using OFDM technique, we can apply frequency diversity to fully utilize the limited bandwidth. By using antenna array processing, users from different direction of arrivals have space diversity. All these diversity can be combined together to combat the detrimental effects such as time varying channel, cochannel interference, heterogeneous QoS requirement, etc.

With the advanced signal processing technique, we can further improve the system performance, for example, multiuser detection, space-time processing, etc. All these techniques can be applied in the existing framework.

A critical issue of dynamic resource allocation is the cross-layer optimization over time-varying, heterogeneous environments. Therefore, to support tomorrow's wireless services, it is essential to develop efficient resource management mechanisms that provide an optimal cost-resource-performance tradeoff. Our research considers building a unified optimization framework for dynamic resource allocation to cope with the time-varying channel/traffic conditions, user profiles, and different QoS requirements in various services, with the goal to yield high efficiency under the constraints of minimum infrastructure and service costs. We apply cross layer approaches to the following two cases: multimedia over CDMA networks; Joint power control and blind beamforming.

Basically we have formulate the different resource allocation problem as a constrained optimization problem. The solutions for the problem can be categorized by four basic mathematic tools: analysis, optimal control, game theory, and dynamic programming. We will explain their basic approaches and point out their advantages and disadvantages. Then we will discuss the different problem formulations in details in the following chapters.

1.2 Background

1.2.1 Wireless Channel Model

The mobile radio signal transmitted in a wireless channel experiences attenuation or distortion mainly due to the effects such as path loss, shadowing, and fading. These effects will generally depend on the frequency, location, direction, reflecting coefficients of the surrounding objects, and velocity of the mobile unit. Modelling these effects has been one of the most difficult parts of mobile radio system design. So the statistical models are applied based on measurements. In this subsection, we briefly discuss the three major effects that affect the wireless transmission.

Propagation Loss

Path loss is caused by propagation loss, where the signal is attenuated due to the distance between the transmitter and the receiver. There are many models to depict the statistical behavior of propagation loss.

In the free space, the propagation loss is given by

$$G = \frac{P_r}{P_t} = \frac{G_t G_r \lambda^2}{(4\pi)^2 d^2}, \tag{1.1}$$

where G_t and G_r are the transmitter and receiver antenna gain respectively, d is the path length, and λ is the carrier wavelength.

If there is a ground path in addition of the direct path, the received signal is expressed as [3]:

$$G = \frac{G_t G_r h_t^2 h_r^2}{d^4}, \tag{1.2}$$

where h_t and h_r are transmitter and receiver antenna heights respectively.

In practice, the path loss models derived from measurement results predict the path loss in different environments with a reasonable accuracy. For example, path loss in urban areas using Hata model is given by:

$$L(dB) = 69.55 + 26.6\log(f_c) - 13.82\log(G_t) - \alpha(G_r)$$
$$+(44.9 - 6.55\ log(G_t)\log(d) \tag{1.3}$$

where $\alpha(G_r)$ is a correction factor and a function of type of environment. For small city it is given by

$$\alpha(G_r) = (1.1\ \ \log(f_c) - 0.7)G_r - (1.56\ \ \log(f_c) - 0.8)$$

and for a large city it is replaced by

$$\alpha(G_r) = 3.2\ (\log(11.75)G_r)^2 - 4.97 \quad (f > 300MHz).$$

Shadowing

In addition to path loss, the average received signal power may be affected by shadowing from large obstacles, such as trees, buildings, or mountains. Measurements have shown that the path loss variations at a particular distance due to shadowing effect is a random variable with zero mean log-normal distribution. The shadowing is generally modelled as lognormal distribution[39]. The probability density function (PDF) is given by:

$$\text{PDF}(\rho) = \frac{1}{\sqrt{2\pi}\sigma\rho} exp\{-\frac{(\log\rho - \xi)^2}{2\sigma^2}\}, \rho > 0 \tag{1.4}$$

where ξ is related to the path loss, σ is the shadow standard deviation.

Fading

In wireless channel, reflections from small scatterers generate multiple replicas of the transmitted signal with different delay, phase, and amplitudes at the receiver. The constructive or destructive combination of these multipath signals causes signal strength fluctuation or fading. If the delay spread of the received signal is significantly smaller than the symbol interval, fading causes amplitude fluctuations only. When there is no specular component in the received signal, fading is modelled by a Raleigh distribution:

$$p(r) = \frac{r}{\sigma^2} \exp(-\frac{r^2}{\sigma^2}).$$

When there are scattering components as well as a dominant path, the received signal amplitude has a Ricean distribution:

$$p(r) = \begin{cases} \frac{r}{\sigma^2} e^{-\frac{r^2+A^2}{2\sigma^2}} I_0(\frac{Ar}{\sigma^2}) & A \geq 0, r \geq 0 \\ 0 & r < 0 \end{cases}$$

where I_0 is the Bessel function of first kind and zero-order, and A denotes the peak amplitude of the dominant signal.

If the difference in time of arrival from different paths is larger than a fraction of symbol interval, in addition to fluctuations in amplitude, fading will cause frequency selective distortion as well. Received signal due to multipath signals is given by

$$r(t) = A \sum_{l=1}^{L} \sqrt{\alpha_l} u(t - \tau_l) e^{j(-2\pi f \tau_l)} + n(t),$$

where $n(t)$ is the thermal noise, and τ_l is the delay associated with the l^{th} path.

Random movement of scatters or mobile will cause doppler spread. If the mobile or scatterers are moving with speed v, the doppler shift is given by $f_d = \frac{v}{\lambda}$. If the doppler spread is larger than a fraction of signal bandwidth, fading causes variation in channel response or time-selective fading. The received signal with delay and doppler shift is give by

$$r(t) = A \sum_{l=1}^{L} \sqrt{\alpha_l} u(t - \tau_l) e^{j(2\pi f_d \cos \phi_l t - 2\pi f \tau_l)} + n(t),$$

where ϕ_l is the angle between the path direction and the velocity vector.

1.2.2 Optimal Transceiver Design

In this subsection, we explain some basic facts on how to optimally design the traditional transceiver for peer to peer transmission. We concentrate on the topics like modulation, equalization, channel coding, diversity, and antenna array processing, which are closely related to our research.

Modulation

Modulation is the process of encoding information to form a message source in a manner suitable for transmission. It generally involves translating a base band source signal to a bandpass signal at frequency that is much higher than the baseband frequency. The bandpass signal is called the modulated signal and the baseband source signal is called modulating signal. Modulation may be done by varying the amplitude, phase, or frequency of a high frequency carrier in accordance with the amplitude of the message signal. Demodulation is the process of extracting the baseband message from the carrier so that it may be processed and interpreted by the intended receiver. [1]

For digital modulation technique, the performance of a modulation scheme is often measured in terms of power efficiency and bandwidth efficiency. Power efficiency describes the ability of a modulation technique to transmit digital message at low power levels. The popular power efficient modulations are M-ary orthogonal modulation and M-ary bi-orthogonal modulation. Bandwidth efficiency describes the ability of a modulation scheme to accommodate data within a limited bandwidth. The popular bandwidth efficient modulations are M-ary FSK, M-ary PAM, M-ary PSK, M-ary QAM, MSK, and CPM. In addition to the efficiencies, other factors, such as performance in fading condition, robustness to nonlinear amplifier, and cost of transceiver, also influence the choice of digital communication.

Adaptive modulation is a promising technique to increase the data rate that can be reliably transmitted over fading channels. For this reason some forms of adaptive modulation are being proposed or implemented in many next generation wireless systems. The basic premise of adaptive modulation is a real-time balancing of the link budget in flat fading through adaptive variation of the transmitted power level, symbol transmission rate, constellation size, BER, coding rate/scheme, or any combination of these parameters. Thus, without wasting power or sacrificing BER, adaptive modulation schemes provide a higher average link spectral efficiency (bps/Hz) by taking advantage of fading through adaptation.

Equalization

If the modulation bandwidth exceeds the coherence bandwidth of the wireless channel, iter-symbol-interference (ISI) occurs and modulation pulses are spread in time into adjacent symbols. Equalization in receiver compensates for ISI within time dispersive channels. Equalizer must be adaptive because the wireless channel are varying continuously. The popular adaptive equalizers are maximum-likelihood optimum receiver, linear equalizer, or decision-feedback equalizer.[2] The performance of equalizer directly affects the communication quality.

Channel Coding

Channel coding adds redundant data bits in the transmitted message so that if instantaneous errors occur in the received signal, the receiver can detect the errors or the data still can be recovered. The channel encoder is located between the source encoder where user's digital message sequence is produced and the modulator where the signal is modulated for transmission in the wireless channel.

There are three general types of channel codes: Block codes (Hamming code, Hadamard code, Golay code, cyclic code, BCH code, Reed-Solomon Code, etc.), convolutional codes, and turbo codes. Some techniques exist to combine the channel coding and modulation such as trellis code and bit interleaved coded modulation. Viterbi algorithm is a fast and optimal algorithm to decode convolutional codes.

Diversity

Diversity is a powerful communication technique that provides significant wireless link improvement with little added cost. Diversity exploits the random nature of radio propagation by finding independence within communication system. A simple example can explain the diversity concept: If one radio path undergoes a deep fade, another independent path may have a strong signal, so the transmitted signal can still be correctly received. The popular diversity methods are listed as follows:

- Frequency diversity

  Frequency diversity is implemented by transmitting information on more than one carrier frequency. The rationale is that frequency separated by more than the coherence bandwidth of the channel will be uncorrelated and will not experience the same fades. OFDM modulation technique exploit frequency diversity by providing simultaneous modulation signals with error control coding across a large bandwidth, such that if a particular frequency undergoes a fade, the composite signal from all frequencies will still be demodulated.

- Time diversity

  Time diversity repeatedly transmits information at time spacings that exceed the coherence time of the radio channel, such that multiple repetitions of the signal will be received with independent fading conditions, thereby providing diversity.

Rake receiver for CDMA is a kind of time diversity by exploring the redundancy in the received signals over multipath channel. By demodulating several replicas of the transmitted CDMA signal, where each replica experiences a particular multipath delay, the RAKE receiver is able to align the replicas in time so that a better estimate of the original signal may be formed at the receiver.

Interleaving is a technique to obtain time diversity in digital communication systems without adding any overhead. Interleaving is extremely useful for channel coding because it helps to resist burst errors. Interleaver has two forms: block structure or convolutional structure.

- Space diversity

  Space diversity is very popular diversity technique, due to the fact that the signals received from spatially separated antennas would have essentially uncorrelated envelops for antenna separations of one half wavelength or more. Space diversity reception methods can be classified into four categories: selection diversity, feedback diversity, maximal ratio combining, and equal gain diversity.

- Space-time (space-frequency) diversity

  Multiple-input-multiple-output (MIMO) systems employing multiple transmit and receive antennas will inarguably play a significant role in the development of future broadband wireless communications. By taking diversity of the larger number of propagation paths between the transmit and receive antennas, the detrimental effects of channel fading can be significantly reduced. It has been shown that MIMO systems offer a large potential capacity increase compared to single antenna systems. To exploit this diversity, a considerable number of MIMO modulation and coding methods, known as space-time codes, have been proposed.

- Multiuser diversity

  In multiuser communications, different users have different channel conditions because they are located in different locations and experience different fading. By adaptively assigning resources such as frequency subchannels, we can take advantage of this channel diversity, which is called multiuser diversity. This multiuser diversity stems from channel diversity including independent path loss and fading of users.

Antenna Array Processing

An antenna array processing is a technique for an array of antenna elements connected to a digital signal processor, as shown in Fig. 1.1. Such a configuration dramatically enhances the capacity of a wireless link through a combination of diversity gain, array gain, and interference suppression. Increased capacity translates to higher data rates for a given number of users or more users for a given data rate per user. Multipath paths of propagation are created by reflections and scattering. Also, interference signals are superimposed on the desired signals. Measurements suggest that each path is really a bundle or cluster of paths, resulting from surface roughness or irregularities. The random gain of the bundle is called multipath fading.

The antenna array processing works as follows. Each antenna element "sees" each propagation path differently, enabling the collection of elements to distin-

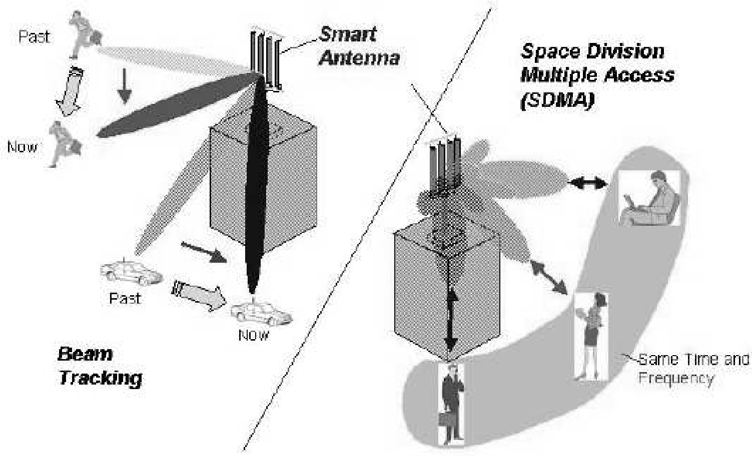


Figure 1.1: Antenna Array Processing

guish individual paths to within a certain resolution. As a consequence, antenna transmitters can encode independent streams of data onto different paths or linear combinations of paths, thereby increasing the data rate, or they can encode data redundantly onto paths that fade independently to protect the receiver from catastrophic signal fades, thereby providing diversity gain. An antenna receiver can decode the data from an antenna transmitter–this is the highest-performing configuration– or it can simply provide array gain or diversity gain to the desired signals transmitted from conventional transmitters and suppress the interference. No manual placement of antennas is required. The antenna array processing electronically adapts to the environment by looking for pilot tones or beacons or by recovering certain characteristics that the transmitted signal is known to have. The antenna array processing can also separate the signals from multiple users who are separated in space (i.e. by angle of arrival) but who use the same radio channel (i.e. center frequency, time-slot, and/or code).

1.2.3 Multiple Access

In this subsection, we will briefly review the most popular multiple access schemes. In the following chapters, we will give different problem formulation for different multiple access schemes and apply different techniques to enhance the system performance.

In wireless communication, it is desirable for each user to transmit and receive simultaneously, which is called duplexing. There are two techniques for duplexing. Frequency division duplexing (FDD) provides two distinct frequency band for transmitting and receiving. Time division duplexing (TDD) uses different time slots for forward and reverse links. There are several pros and cons between FDD and TDD. For FDD, the radio frequency must be carefully designed to reduce the RF cost and handle the different powers of transmit and receive radio signals. For TDD, there are transmission delays and the system is sensitive for propagation delays. So TDD is often applied in cordless phone and fixed wireless networks.

For multiple users' communication, multiple access schemes are developed to share simultaneously the limited bandwidth of radio spectrum. Frequency division multiple access (FDMA), time division multiple access (TDMA), frequency hopped multiple access (FHMA), and code division multiple access (CDMA) are major access techniques [1]. These multiple access techniques have been widely used in current wireless communication systems such as GSM, IS-95, CT2, and DECT.

By using the antenna signal processing technique, space divsion multiple access (SDMA) separates users' signals in different direction of arrivals(DOA). With SDMA, multiple users with different DOA are able to communicate at the same time using the same channel. In addition, the antenna can collect transmitting powers from multipath components, combine them in an optimal manner, suppress interferences from other users, and improve the received SINR. Consequently, less power is required.

In random access protocols, the channels are utilized by users attempting to access a single channel in an uncoordinated manner. Consequently, the transmissions are due to collisions by multiple users. Many packet radio (PR) access techniques are developed to handle the collisions. PR is very easy to implement, but has low spectral efficiency and may have delays. Some of the available PR access techniques are Aloha, carrier sense multiple access (CSMA), carrier sense multiple access with collision detection (CSMA/CD), data sense multiple access (DAMA), and packet reservation multiple access (PRMA).

Orthogonal frequency division multiplexing (OFDM) protocol such as IEEE 802.11 is one of the prime modern schemes for broadband wireless networks, because of its advantages over frequency selective channel and inter-symbol-interferences caused by multipath propagations. In multi-user scenario, the available techniques are OFDM-TDMA, OFDM-CDMA, and frequency division multiple access (OFDMA). In OFDMA system, each user occupies a subset of subcarriers and each carrier is assigned exclusively to only one user at any time, so that there are no intra-cell interferences.

1.2.4 Cellular Concept

In the multi-access techniques mentioned in the previous subsection, because each channel is used by only one user at each time, there is no cochannel interference (CCI). However in order to achieve high capacity with limited radio spectrum while at the same time covering very large areas, we need to introduce channel reuse. Channel reuse will cause CCI and we will discuss how to allocate the

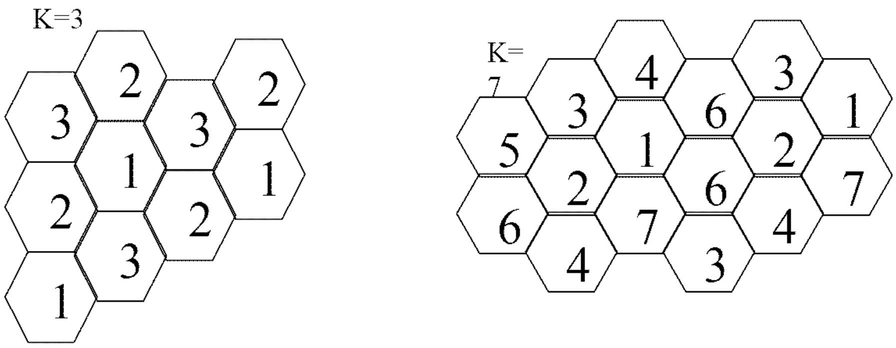


Figure 1.2: Example of Different Reuse Value

wirless resources to minimize CCI. In this subsection, first we introduce the cellular concept. Then we discuss the channel reuse and assignment. Finally, we discuss the handoff.

The cellular techniques offer very high capacity in the limited available spectrum by applying many low power transmitters, which provides coverage to a small portion of the service area. In a cellular system, a large coverage area is broken into many small geographic areas called cells. Each cell is assigned with a small proportion of the total channels, and the adjacent cells are assigned with different groups of channels. The same group of channels can be reused in the cells that are enough far away so that the transmitted powers are attenuated enough and the interferences between cells are minimized. The cellular wireless networks provide a method to use limited spectrums to serve large number of users by reusing the channels throughout the coverage regions.

To mitigate the cochannel interferences, total number of channels are grouped in R_u groups and neighboring cells are assigned with different group of channels. For symmetric cell plans,

$$R_u = (i+j)^2 - ij, \; i,j = 0,1,\ldots. \tag{1.5}$$

Possible value are $R_u = 1, 3, 4, 7, 9, 12, 13....$ In AMPS, $R_u = 7$. In GSM, $R_u = 4$ or 3. In USDC and PDC, $R_u = 7$ or 4. In Fig. 1.2,we show the cell plans with different R_u.

Channels are assigned to different cells to efficiently utilize the spectrum by fixed or dynamic policies. In a fixed assignment, each cell is allocated a certain set of channels and each cell handles channel allocation independently of other cells, which is simple for implementable and fits a network with spatially uniform traffic density. In a dynamic channel assignment, the network will allocate a channel to a cell at call setup. The minimum allowable distance between cochannel cells and traffic density is considered in order to minimize the probability of blocking.

Handoff occurs when a mobile leaves the coverage area of a cell and enters the coverage area of another cell. In channelized wireless system, different radio channels is assigned during a handoff, which is called hard handoff. In CDMA system such as IS-95, the assigned channel to user is not changed, but a different base station is selected for communication. This kind of handoff is called soft handoff.

1.2.5 Cross Layer Approaches

Traditional communication systems are designed in layers. According to OSI reference model, the communication system can be divided into seven layers from top to bottom: Application, Presentation, Session, Transport, Network, Data Link, and Physical Layers. Each layer implements a specific purpose and optimizes its own goal. Obviously it will not be optimal from the whole system point of view. Moreover, the system has to pay the communication overhead between layers. In the communication systems nowadays, the bandwidth becomes more and more limited and precious. So it is necessary and emerging to consider the optimization problem across layers. In this subsection, we briefly discuss three kinds of cross layer designs that are of great interests in recent literature.

Data Link Layer and Physical Layer

The main task of the data link layer is to transform a raw transmission facility into a line that appears free of undetected transmission errors to the network layer. It also considers the flow control and error handling. Within this layer, one sublayer called medium access control (MAC) controls access to the shared channel.

The concern of physical layer is to transmit raw bits over a communication channel. The design issues largely deal with mechanical, electrical, and timing interfaces, and the physical transmission medium, which lies below the physical layer.

Because wireless channels are shared for different users, one user's transmission power is the interference for other users. Moreover in order to fully utilize the multiuser diversity, different users' rates should be controlled in such a way to optimize the overall system performance. So how to consider the resource allocation such as power control and rate adaptation between date link layer and physical layer is essential for wireless communication design. Most of our research works are concentrated on this type of cross layer design.

Application Layer, MAC, and Physical Layer

The application layer contains a variety of protocols that are commonly needed by users. The most popular applications payloads for wireless networks are voice, video, and data. For voice payload, the concern is subjective perception which is affected by the transmission delay and source encoder rate. So MAC layer and physical layer controls are important means to guarantee the recovered voice packet qualities. For video transmission, the transmission is very bursty because of different frames and different video contents. The variable rate transmission over the lower layers can substantially improve the system performances. For data transmission, the reliability reception of data streams is the most important design issue. So powerful channel coding or ARQ is necessary for this type of application. In our research, we apply joint source channel coding with power control for voice transmission over CDMA networks.

Network Layer and Physical Layer in Ad-hoc Networks

The network layer controls the operation of the subnet. A key design issue is determining how packets are routed from source to destination. A wireless ad hoc network consists of a collection of wireless nodes without a fixed infrastructure. Each node in the network serves as a router that forwards packets for other nodes. Each flow from the source to the destination traverses multiple hops of wireless links. Compared with wireline networks where flows contend only at the router with other simultaneous flows through the same router, the unique characteristics of multi-hop wireless networks show that, data stream flows also compete for shared channel bandwidth if they are within the transmission ranges of each other. This presents the problem of designing an appropriate topology aware resource allocation algorithm. so that contending multi-hop flows share the scarce channel capacity, while the total system performance is optimized.

1.3 Motivations and Contributions

Dynamic resource allocation is a general strategy to control the interferences and enhance the performance of wireless networks [42, 25, 22, 31]. The basic idea behind dynamic resource allocation is to utilize the channel more efficiently by sharing the spectrum and reducing interference through optimizing parameters such as the transmitting power, symbol transmission rate, modulation scheme, coding scheme, bandwidth, or combinations of these parameters. Moreover the network performance can be further improved by introducing more diversity and cross-layer considerations. Many interference management techniques have been explored including power control, rate adaptation, dynamic channel allocation, beamforming, multiuser detection, and so on [42, 50, 54, 51, 7]. Joint methods, such as adaptive rate and power control [43, 54, 25], joint power control and beamforming [47, 95, 97], joint power control and multiuser detection for CDMA [86], joint power control, multiuser detection and beamforming [50, 6], channel allocation with modulation and power control , and joint base station, power and channel allocation [77] are also proposed to cancel/suppress the interference.

In the literature, a widely used objective is to minimize the total transmitting power or to maximize the overall system throughput in the network while the SINR targets are achieved for all users [42, 86, 50]. The solution for this objective can be obtained by a matrix inversion in centralized (non-iterative) schemes [43] which require the full knowledge of the entire network, for instance link gains and noise levels; while in distributed (iterative) schemes [42, 43, 44, 47] only local measurements are required thus more suited to a network with limited information available to the users. In addition, various schemes have been proposed to maximize the minimum SINR, to maximize throughput, to maximize the total capacity, or to maximize the expected sum of data rates under energy and delay constraints in [54, 22, 46] and the references therein. Stochastic approximation based power control algorithms have also been studied in some research both in cases of single user receivers and multiuser receivers [86].

As we can see, most of the works tried to optimize some objectives under certain Quality of Service (QoS) constraints in wireless networks. There is a need of unified but general optimization framework for resource allocation to allow taking into account a diverse set of objective functions with various QoS requirements. Also, the QoS provisioning and support remains essential technical challenges in wireless environments. Previous works generally adopt simple QoS parameters, for instance, the fixed targeted SINR is routinely used to characterize the QoS requirements. The transmission strategies in the previous research were designed based on the current channel conditions, and did not consider the time diversity of both short-term and long-term perspectives. Such kind of optimization approach may exhibit disadvantages over a long term period of time for the time diversity is not factored into consideration. Most of the existing schemes mentioned above do not adaptively adjust according to the users' QoS satisfaction levels. In addition, there is a need to summarize the possible solutions for these resource allocation problems. So all these facts give us the motivation for our research.

When the number of users is small in a wireless network, the resource allocation problem can be solved by control optimization theory, where the problem is viewed as a constrained optimization problem. When the number of users are large, resource allocation problem of a wireless network is analogous to that of the human society. In the proposed research, we shall employ the commonly accepted principles of economy analysis, particularly game theory and mechanism design theory, to tackle the problem by motivating self-interested users to adopt a social behavior by sharing resources efficiently and thus to improve the overall system quality. We will develop a unified optimization framework for dynamic resource allocation and provide some solutions for some specific system scenarios.

The significance of this dissertation is primarily targeted at developing a unified optimization framework with different approaches for dynamic resource allocation. The proposed research is interdisciplinary in that it combines concepts in signal processing, economics, decision theory, optimization, information theory, communications, and networking to address the issues in questions. The cross-layer optimization nature of the problem provides an innovative new inside into vertical integration of wireless networks. The goal is to significantly improve and advance the models to design and analyze resource allocation over wireless networks, especially in linking successful optimization control and economy models to the engineering problems.

1.4 Organization of This Dissertation

In this dissertation, we will propose a unified optimization framework that address the wireless communication resource allocation problem and enhance the system performances. The organization of this dissertation is given by:

In Chapter 2, we give the basic mathematical background. A universal view of resource allocation is developed. Because of the channel dynamics, the feasible range of resource allocation is varying. This dynamics gives us difficulty to find the optimal allocation within the feasible range. On the other hand, this dynamics also gives us opportunities to explore the multiple dimension diversity. There are many practical constraints for implementation and the optimization goals are assorted. Most researches in literature concentrate on "sliced view", i.e., the optimization is performed for one goal and under some constraints. In this dissertation, we will formulate the general resource allocation problem from a "universal" view. In order to solve such a problem in an easy way, four different types of mathematical solutions are explained and compared. They are analysis approximation, nonlinear/linear/convex programming, game theory, and dynamic programming.

In Chapter 3, we explain our proposed centralized resource allocation schemes. We give the introduction and motivations of our research for the wireless systems where centralized control is applicable. Then we discuss the fairness issue. Three different criteria for fair resource allocation are explained and compared. Then we discuss how to provide heterogeneous QoS. For different applications, the requirements are quite different. For example, voice packet cannot suffer delay, data packet cannot suffer BER, and video packet transmits in burst. We model the QoS measure for delay sensitive application. In order to combat the time varying channel and cochannel interference, we explore the time diversity and multiuser diversity. In addition, we also explore the space diversity using antenna array processing. We formulate three different problems in three different scenario. In one case, we apply micro-economy concepts such as credit system, user autonomy, and resource awareness for users' efficient resource allocation. From the simulation results, the proposed schemes can satisfy the delay constraint, allocate resource fairly to all users, and have comparable performance to that of the greedy approach where fairness and delay constraint are not considered.

In Chapter 4, we present our distributed resource allocation scheme using game theory. Introduction and motivations are given first. Then brief introduction about game theory is introduced. The challenges for game theory approach are explained. The utility function for each user has to depict some physical meaning. Because of the nonlinearity of the system, the game is hard to be balanced in the desired Nash equilibrium with a high system performance. We present a noncooperative game approach to motivate individual users to adopt a social behavior and enhance the system performance by sharing resources. A performance upper bound is also developed. From the simulation results, the proposed game theory approach can achieve the desired Nash equilibrium and has the similar performance to that of the performance upper bound. Finally, we compare the centralized and distributed resource allocation approaches. Pro and con of the two approaches are compared and analyzed. Possible hybrid system is proposed.

In Chapter 5, we further extend our research for frequency diversity using orthogonal frequency division multiple access (OFDMA) to provide high throughput, combat frequency selective fading, and provide flexible resource allocation. The resource allocation problem is how to assign subchannel, how to allocate bits to each subchannel, and how to control transmitted power. Existing work solve these problems by waterfilling, integer programming, or iterative waterfilling. However the complexity is high, the efficiency is low, and only centralized solution is available. In this dissertation, to overcome the disadvantages of previous schemes, we present three methods for resource allocation in OFDMA networks by cooperative game, non-cooperative game, and subspace method for single cell with multiple users, multi-cell with one user per cell, and multi-cell with multiple users per cell, respectively. In cooperative game, we provide a fair and simple solution. The complexity is only $O(N \log N)$, compared with traditional scheme with $O(N^4)$, where N is the number of subchannel. In the noncooperative game, behaviors of Nash equilibriums are analyzed and a game rule is developed for users to share the subchannel. The unqualified user will be kicked out from using some subchannel, such that the other users can share the subchannel more efficiently and the overall system performance can be improved. In the subspace method, two initialization algorithms are constructed and one iterative subspace improvement algorithm is provided to solve the proposed very complex problem.

In Chapter 6, we present two examples for cross layer approaches. The first one is the multimedia transmission over CDMA. We model and formulate the problem for multimedia over MAC and physical layers. A protocol is constructed for embedded voice coder, adaptive channel coding, adaptive processing gain, and adaptive power, such that the distortion is smoothly and predictably controlled. We develop a fast algorithm to minimize the overall system distortion, under the maximal transmitted power and distortion constraints. From the simulation results, the proposed scheme can increase the number of users and reduce the required transmitted power fundamentally. The second cross layer approach is joint power control and blind beamforming. The objectives are to eliminate additional overheads for measurement and provide a scheme that is robust for estimation errors. The proposed scheme uses a local information from a blind beamforming algorithm and updates the transmitted power in a distributed manner. A Cramer-Rao lower bound is also developed to compared the performance. From the simulation results, the proposed scheme can achieve a large range of BER for the whole networks without requiring training sequence.

In Chapter 7, we draw a conclusion to show that our works explore the multiuser, time, frequency, and space diversity and formulate the problem more accurately and efficiently in a cross layer approach. Then we give some possible future work: effective bandwidth and capacity, video transmission, dynamic programming over HMM model, dynamic reinforcement learning for cooperation in multiuser system, repeated game approach, utility and pricing for multimedia transmission, and Ad Hoc networks with limited resources.

Chapter 2

Generalized Optimization Framework and Mathematics Theoretical Background for Solutions

2.1 General Resource Allocation Formulations

The development in the filed of wireless communications has been nothing short of astonishing in the past decades. We now are witnessing the transition between the mobile telephone era and the era of wireless computing. With the breakthrough advances of digital signal processing high data rate, many of the technical problems associated with the adverse and changing propagation conditions in mobile radio communication have been solved. Multi-megabit data rates to portable mobile terminals are no longer science fiction, but reality. As the engineer seems to have the upper hand in this struggle against nature, very much of the development efforts are concentrated on the social struggle for scarce resources. One of the most technical challenges that limit achieving these requirements is interferences due to the bandwidth limitation and reuse of the bandwidth, which becomes bottleneck of nowadays wireless communications. Dynamic resource allocation is a general strategy to control the interferences and enhance the performance of wireless networks. The basic idea behind dynamic resource allocation is to utilize the channel more efficiently by sharing the spectrum and reducing interference through optimizing parameters such as the transmitting power, symbol transmission rate, modulation scheme, coding scheme, bandwidth, or combinations of these parameters.

In traditional resource allocation, the resources are managed within each layer. For example, in physical layer, the adaptive transmitted power and adaptive modulation are applied to increase the spectrum efficiency and reduce the co-channel interferences. While in application layer, the multimedia encoder is designed to have highest compression rate with small distortion. This kind of layered approaches can be easily implemented and each layer has its own concentration. But the resulting resource management might not be optimal. Because the wireless communications become more and more crowded, there are more and more demands for efficiency of resource allocation. This motivates the cross layers approach. For example, the source coder can get information from the physical layer about the current channel condition. If the channel is good, the source coder can generate more bit stream which will result in higher quality. Otherwise, the source coder will generate the minimal bit stream reduce the burden of physical layer.

In order to optimize the network performance, we need to know what are the available resources, i.e., what are the parameters for optimization. We define the parameter sets as Θ. For different layers, we list possible resources as:

1. **Physical Layer**

   Transmitted power, rate (source rate, channel rate, symbol rate), base station, antenna weight vector ...

2. **MAC Layer**

   The buffer size, waiting time, arrival rate, service rate ...

3. **Network Layer**

   The route from the source to destination.

4. **Application Layer**

   Source coding rate for voice or video encoders.

With these parameters, the fundamental problem for wireless resource allocation is how to efficiently allocate them across layers and to different users so that the network performance is optimized. For the network performance, there are many criteria. For example optimization goal can be overall throughput, overall transmitted power, average distortion, maximum outage rate, overall QoS, etc. or multi-purpose. We represent optimization goal as Σ. For different wireless networks and different situations, Σ can have very different representations. These optimization goals can have sum, product, or other format and are functions of the resources. They can be linear, convex, or nonlinear at all. Sometime the goal itself can be implicit or multiple purposes as well. The most important thing is to define the goal function that can represent the real network performances.

In real implementation, there are many practical constraints. For example, the mobile unit can only generated limited transmitted power. We define the constraints sets as Φ. The typical constraints are maximal power constraint, minimal or maximal rate constraint, minimal distortion, maximal delay time, and other practical constraints. These constraints are functions of the possible resources. These functions might have nonlinear and nonconvex properties.

In the wireless resource allocation, the key problem is how to allocation the limited resources to optimize the system performance under some practical constraints. So the overall fundamental problem can be formulated as:

$$\max_{\Theta} \Sigma \qquad (2.1)$$
$$s.t. \;\; \Phi.$$

For dynamic system, channel conditions are kept changing. Under this condition, the feasible range of the solution that satisfies the constraints is also varying. For the traditional wireless networks, the system is designed to accommodate the worst case situation. One example is shown in Fig. 2.1, where a two-user case is illustrated with the axes representing the throughput for each user. The feasible ranges for time 1 and time 2 are very different. For the traditional system without dynamic resource allocation, the optimal resource allocation point is shown at point A in order to let the system feasible for all times. If we explore diversity for the dynamic system, we can apply allocation point B at time 1 and allocation point C at time 2. Obviously, the resulting solution is much better than that of the traditional scheme. This is because we take consideration of both time diversity and multiuser diversity. The challenge for dynamic resource allocation is how to find the feasible and optimal resource allocation point dynamically for different times in a simple and implementable way. In addition, we can also apply frequency, space, and route diversity to explore the dynamics of the system.

Basically we have formulate the resource allocation as a constrained optimization problem and explain the dynamics. In order to solve the problem, we will

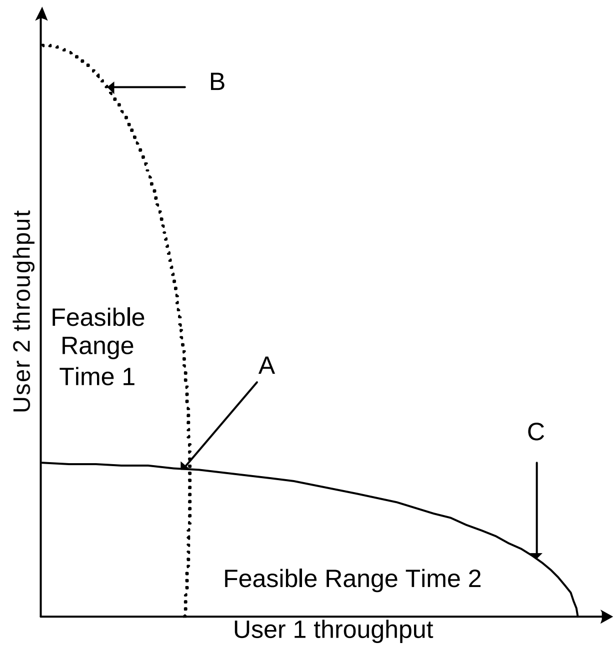


Figure 2.1: Illustrative Example on Dynamics

discuss four basic mathematic tools to solve the fundamental problem. We will explain their basic approaches and point out their advantages and disadvantages. Then we will discuss the different problem formulations in details in the following chapters.

2.2 Analysis Solution

For general constrained optimization problem, if the optimization goal and constraint functions are linear or convex or have some nice forms, we can apply the methods such as Lagrange multiplier or convex optimization algorithm to have a nice analysis solution. In order to have clean analytic results for the resource allocation, the approximation and simplification for the optimization goal and constraints are the key factors that affects the performance of analysis solutions.

The Lagrange function for (2.1) can be written as:

$$J = \Sigma + \lambda\Phi. \tag{2.2}$$

where λ is Lagrange multiplier. We differentiate J over Θ and find the solution for λ. By using the constraint functions Φ can can find the optimal solution for (2.1). The difficulty is that λ might not be solvable and even with λ, the optimal solution may be hard to obtain from Φ.

If the optimization goal Σ and the constraint functions Φ in (2.1) are convex functions, we can apply convex optimization methods to solve the problem numerically with great efficient. The convex optimization methods have extensive and useful theory and can be applied to many engineering problems. The convex optimization methods are tractable in theory and practice: there exist algorithms such that

- computation time small, grows gracefully with problem size
- global solutions attained
- non heuristic stopping criteria; provable lower bounds
- handle non-differentiable as well as smooth problems

The popular convex optimization methods include linear optimization, quadratic optimization, geometric programming, vector optimization, dual methods, gradient method, steepest descent method, Newton method, barrier method, interior point method, and cutting plane method.

The challenge for analysis solution is the difficulty to approximate the optimization goal and the constraint functions to a nice and handleable form. The approximations need to be accurate under some conditions. For example, the capacity function can be written as:

$$C = W \log(1 + \Gamma) \tag{2.3}$$

where W is the bandwidth and Γ is the signal to noise ratio. Obviously C is a non-convex and nonlinear function of Γ and has the S-type shape. In order to make it easy to handle, we can assume $\Gamma \gg 0$ such that the above capacity function can be approximated by:

$$C \approx W \log(\Gamma), \tag{2.4}$$

which is a nice concave function.

However in reality, we may not be so lucky to have the good approximations. So this reason limits the usage of the analysis approach. In literature, only simple problems, with small number of users, small number of optimization parameters and simple channel models, can be solved by this type of mathematical solutions.

On the whole, the advantage of this approach is the clean solution. The resource allocation can be calculated fast and directly. However the performance is highly related to how good the approximation and simplification to the reality.

2.3 Optimal Control Solution

Since the problem defined in (2.1) is a constrained optimization problem. It is nature to use the methods such as linear programming, nonlinear programming, or integer programming to solve the problem.

Many major developments are achieved in optimal control theory in the last ten years. First is the merging of linear and nonlinear programming algorithms throughput the user of interior point methods. The second development is the increased emphasis on large scale problems and the associated algorithms that take advantage of problem structure as well as parallel hardware. The third development is the extensive use of iterative unconstrained optimization to solve the difficult least squares problems arising in the training of neural networks. All these developments are extremely useful for resource allocation.

The advantages of such kind of solutions are obvious. In reality, the optimization problems are often nonlinear and nonconvex. The optimal control methods fit this kind of problem very well. For example, if the second order differentials are available for the goal and constraints, we can use the Newton algorithm with Barrier method to solve the problem efficient.

There are some disadvantages of this kind of solution. First because of nonlinearity and nonconvexity, there exits many local optima. Careful or multiple initializations are needed. For the worst case, simulated annealing has to be applied for global optima. Second full knowledge of channel conditions is needed to do the optimizations, which increases the burden of channel estimations and the associated overhead. Moreover the complexity usually increases fast with the increasing of the number of users. So this kind of solutions are very complex and only fit centralized control with small number of users.

2.4 Game Theory Solution

In multi-access wireless networks, since an individual mobile user does not have the knowledge of other users' conditions and cannot cooperate with each other, they act selfishly to maximize their own performances in a distributed fashion. Such a fact motivates us to adopt the game theory [56].

The resource allocation can be modelled as a non-cooperative game that deals largely with how rational and intelligent individuals interact with each other in an effort to achieve their own goals. In the resource allocation game, each mobile user is self-interested and trying to maximize his utility function, where the utility function represents the user's performance and controls the outcomes of the game. So the goal of this kind of solutions is to define the meaningful utility function such that the system can be balanced in the desired social optimal equilibrium.

For noncooperative game, because of each user's greediness, the Nash equilibrium of the game usually turns out to be not optimal. If the users play multi-stage games and the users overall payoff is a weighted average of the payoffs in each stage, we can apply the repeated game theory, which is the best understood class of dynamic games. The repeated game can let users cooperative together to have better Nash equilibrium. The rationale is: even though each user could do better in the short run by defecting instead of cooperating , for a patient user, this short-run gain is outweighed by the prospect of unrelenting future "punishment" from other users. The difficulty to model cooperative game is how to model the punishment.

If there exist limited communications between users, we can apply cooperative game to improve the system performance. The cooperative game is defined as: A cooperative game is a structure in which the players have the option of planning as a group in advance of choosing their actions. The famous results include Nash bargaining solution, coalition analysis, core concept, and Shapley function.

The biggest advantage of this kind of solution is that it can be implemented in a distribute manner with large number of users. This kind of solution is very similar to economy or social problem. Everybody is selfish and the society needs to design the game rule for each individual to improve the social good.

The disadvantage of this kind of solutions is the difficulty to design the utility function. First users' QoS is a hard parameter to describe. Second, even we can describe QoS by a utility function, when the different users compete with each other, this utility function may not produce the desired Nash equilibrium. Some techniques such as pricing and repeated game are applied to improve the equilibrium. In pricing, the system provides prices for the resources and users have to pay the price to get the resources. By doing this, the system can control the outcome of the competition. The prices are determined by the system for the social good or determined by the "demand and request" rule.

2.5 Dynamic Programming Solution

The above three methods only consider the optimization at one time. In reality, some of the applications need to do optimization over different time. Naturally, the dynamic programming technique can be applied. The dynamic programming method makes the optimal decisions based on the distributions of the channels or the sources.

Basic structure of dynamic programming is briefly explained as follows: suppose a discrete time system

$$x_{k+1} = f_k(x_k, u_k, w_k), \quad k = 0, 1, \ldots, N-1, \tag{2.5}$$

where k is the discrete time, x_k is the state which summarizes past information that is relevant for future optimization, u_k is control where decision is selected at time k from a given set, w_k is the random parameter or noise with probability distribution P_k, and N is horizon or number of times control is applied. Define policies $\pi = \{\mu_0, \ldots \mu_{N-1}\}$, where μ_k maps state x_k into control $u_k = \mu_k(x_k)$ and is such that $\mu_k(x_k) \in U_k(x_k)$ for all x_k. We want to select the optimal policy π^* such that the expected cost of π starting at x_0 is minimized:

$$J_{\pi^*}(x_0) = \min_{\pi} J_\pi(x_0) = E\{g_n(x_n) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k)\} \tag{2.6}$$

The dynamic programming considers the optimization over times.

Scheduling is an extreme case of dynamic programming. For scheduling, only one user can transmit each time, which fits the situation such as single cell CDMA systems. If the optimization goal is for each time only, the optimal solution is that only the user with the best channel response transmits. However this will introduce unfairness and long delays. There is a tradeoff for fairness (delay) and system performance. Many scheduling methods are developed to reconcile the tradeoffs.

The advantages of this solution are the optimization over time. The user might not be optimized at a specific time. But his sacrifice for performance will increase the overall system performance and will be compensated back in the future, which explores time diversity.

The disadvantages of this solution lies in two factor. First the distributions are hard to obtain, especially in multiuser cases. Second, the computation complexity is extremely high. So this solution usually is useful when the system model is very simple and there are only few users.

2.6 Comparison of Different Solutions

In the previous sections, we briefly review the existing techniques to solve the resource allocation problem. We explain their basic approaches as well as the pro and con of the solutions. There are many other techniques that can combine with

Table 2.1: Pro and Con of Different Approaches

Methods	Pro	Con
Analysis	easy implementation	approx. far from the reality
Optimal Program	noncovex nonlinear problem	centralized implementation only
Game Theory	distributed implementation	non-optimal equilibrium
Dynamic Program	optimization over time	complex, simple model only

the optimization frame works. We will explain in details in the following chapters.

On the whole, the advantages and disadvantages of the above four methods are listed in Table 2.1. Different solutions may fit different problem formulations. For the research works, it is usually to combine the different techniques to have a simple solution with good performances.

Chapter 3

Centralized Resource Allocation with Time Average Fairness

In Chapter 3 of the dissertation, we will discuss some centralized resource allocation schemes with fairness constraint. These schemes fit the wireless networks where centralized control is implementable. The mathematical tools we used here are analysis approximation, optimal control, and dynamic control. Compared with traditional resource allocation, our proposed schemes can improve the system performance, while maintaining the fairness for all users.

This chapter is organized as follows: First we give the introduction and motivations for our research. Then, we explain the fairness issues and give some popular definitions of fairness. In the rest of this chapter, we list three works for the centralized resource allocation: First one explores the adaptive modulation, second one explores the space-time diversity, and last one constructs a credit system and uses economy approaches.

3.1 Motivations

With the development of wireless networks, the number of mobile users becomes more and more large and the network topology becomes more and more complicated. As a result, the distributed resource allocation becomes more and more popular, because of the great reduction of system cost and efficiency to save the overheads for centralized control. However, on the other hand, for each user, the applications become more and more heterogenous. Image, Video, and data are integrated besides voice in nowadays system. As a result, the optimization problem becomes more and more complicated, because of the goals and constraints for different layers are all different and complicated. The centralized resource allocation has its own advantages to cope with more difficult problems, due to its abundant mathematical theories and more available information. So there is a tradeoff between the centralized and distributed resource allocation and each of them will fit different network scenarios.

The bottle neck for centralized resource allocation lies in two factors. First, the communication overhead might be unacceptable if the network topology is too distributed. For example, for multicell case, it is hard to get the channel estimation from one cell to the other. Second, the optimization parameters will grow too fast with the number of users is increasing, which will increase the system cost greatly. So centralized system fits the wireless network like micro cell scenario where there is a centralized control node, base station, and the number of users are small.

The other advantage for centralized resource allocation is that it can deal with much more complicated problem. The first reason is that there might be very powerful computing ability in base station, where the very complicated optimization problem can be solved. For the cross layer optimization, where voice, video, data, and routing are considered, the optimization problem usually has ugly nonlinear nonconvex form. This computation task can be implemented in base station and the results are assigned to the mobile users to save their computation cost. The second reason is that the information such as channel conditions can be obtained from base station and be applied for the optimization. Consequently, the optimization results are more accurate, converge more fast, and more robust.

The motivation for us to explore the centralized resource allocation lies in two factors: First, we want to explore the fairness to different users, which will impose another constraint for resource allocation and might be related to the concepts of dynamic programming. Second, we want to connect MAC layer and physical layer optimization. We will jointly consider the adaptive modulation, power control, and antenna diversity so as to fully explore the space-time diversity and increase the system performance. We also apply some economy ideas for this type of resource allocation.

3.2 Fairness

Before we dig into our proposed resource allocation schemes, we will briefly review the current scheme where the fairness is not considered. Then we review three popular fairness concepts.

In most of traditional networks, the optimization can be classified to two categories. In the first category of greedy scheme, the system optimization goal is to optimize the system performance such as to maximize the overall throughput, or minimize the overall transmitted power, without considering the fairness between users or each user has the minimal constraint. For this kind of resource allocation, the resources are allocated to each user for their minimal requirement, then the rest is greedily allocated to the users with the best conditions, which is extremely unfair. Since each user pays the same for his service, it is desirable to have fair resource allocation scheme. On straight-forward way is to let each user have the same quality of service, which is the second category of strict fair scheme. However the performance of such a scheme is very low compared to the greedy scheme, because it doesn't consider the time diversity and multiuser diversity. In order to provide fair services to all users, we need to define the new fairness concepts. From literature, there exist three popular kinds of fairness: max-min, proportional, and time average.

max-min fairness

Max-min fairness is a very popular fairness principle, which has been advocated for a long time of resource allocation. The objective of max-min fairness is to maximize, under the practical constraints, the minimum performance of each user can obtain. Max-min fairness basically relies on the following principle: In the domain of feasible resource allocation, one user's (user 1) performance cannot be increased without decreasing some other user's (user 2) performance such that user 1's performance is better than user 2's. The compactness and convexity of the feasible region imply that such a max-min solution exists and is unique. However the max-min fairness criterion gives an absolute priority to the user with bad conditions, which in turn will reduce the system performance.

proportional fairness

An alternative fairness criterion which favors the users with bad conditions less emphatically, is proportionally fair [80].

Definition 3.2.1 *A feasible resource allocation vector $\lambda_{\mathbf{s}}$ for user s is proportionally fair, if and only if for any other feasible resource allocation vector λ'_s, the sum* *of relative change is not positive, i.e.,*

$$\sum_s \frac{\lambda_s' - \lambda_s}{\lambda_s} \leq 0. \tag{3.1}$$

The physical meaning of proportional fairness is that an increase in the allocation of network resources for one user must be compensated by corresponding decreases in the allocations of one or more other users. Another interpretation is that a resource allocation is fair if it is in proportion to the users willingness to pay.

time average fairness

The previous two fairness concept only consider the fair allocation for each time. Since users experience different channels for different times, we can further explore the time diversity by defining the time average fairness. The principle of time average fairness is that the user will get the same time average performance. This fairness also depends on how patient the users can wait for the channel to become better for transmission. For example, for the voice transmission, delay is very strict, so fairness for the users of this type of service is within a short term. While for the data transmission where the delay can be suffered, fairness can be relaxed for a long period of time. The mechanism to maintain such a fairness is that the user will demand more in the future if he cannot get his desired QoS now.

3.3 Joint Power Control and Adaptive Modulation

In multi-access wireless communication systems, power control and adaptive modulation are two important means to increase spectral efficiencies, combat time-varying fading channels, and reduce co-channel interferences. In our approach, the overall uplink transmitted power is minimized under the constraints that there is no reduction in overall network throughput and each user achieves the desired time-average throughput. Adaptive M-QAM modulations with two kinds of antenna diversity are considered. Each user can select a range of modulation rates, according to his channel condition and transmission history. Two subproblems are considered for the development of suboptimal low complexity adaptive algorithms. First at the user level, the following needs to be determined: the range of modulation rates that each user can accept at a specific time to ensure fairness. Then each time at the system level, within the acceptable ranges, the system finds out what throughput allocation for different users requires the lowest overall transmitted power. The scheme can be interpreted as "water filling" each user's throughput in time domain and allocating network throughput to different users at each time. From the simulation results, the proposed scheme reduces the overall transmitted power up to 7dB and increases average spectral efficiency up to 1.2 bit/s/Hz, compared with the previous known power control schemes.

The organization of this section is as follows: First, we give the motivation and sketch of the proposed scheme. Then, a network system model is presented with antenna diversities. Approximations of MQAM are presented. The optimization problem is formulated. The problem is heuristically divided into two sub-problems. Several adaptive algorithms are developed. A power and throughput management system is constructed. The proposed algorithms are evaluated by simulation study.

Motivation and Sketch

Much work has been done for resource allocation, such as power control and adaptive modulation in multi-access wireless channels. In [4, 5, 43, 44], resource allocation has been extensively studied. In [47, 95, 97, 6], beamforming, multiuser detection, and power control have been combined for cellular wireless communication systems. In [7, 65, 8, 9, 10, 11, 12], adaptive modulation techniques have been proposed to enhance the spectrum efficiency for wireless channels. The performance approximation and robustness for estimation errors have been investigated. In [66, 13], adaptive coding provides another way for transmission rate control. In [48, 14, 15], the authors have explored resource allocation problems from the channel capacity point of view. In [16, 54, 49, 17, 18, 19, 53, 22], many adaptive algorithms are constructed to adaptively control the transmitted power and rate to optimize the system performance. In [59], the authors present an "opportunistic" transmission scheduling policy for a single cell TDMA/FDMA system that exploits time-varying channels and maximizes the system performance stochastically. In [57], game theory is introduced in the power control problem. Each user competes with other users for limited resources, and the system is balanced in some equilibriums.

In traditional power control, each link's transmitted power is selected so that its received SINR is larger than or equal to a fixed and predefined targeted SINR threshold, required to maintain its link quality, while the system minimizes the overall transmitted power of all links. However, a link with a bad channel condition requires too much transmitted power and therefore causes unnecessary CCI to other links. This is a major issue that will be addressed in our approach. In adaptive modulation, each link can select a range of different modulation rates; consequently, a range of targeted SINR thresholds can be applied. A joint adaptive power and rate allocation scheme is developed by using M-QAM adaptive modulation with antenna diversity. The optimization goal is to minimize the overall transmitted power under some constraints: The overall network throughput is not reduced; the time-average throughput of each user is maintained as a constant that is determined by the service for which the user pays. In order to solve the problem, the problem is heuristically divided into two sub-problems. First, users determine the ranges of throughput that they can accept at different times and report these ranges back to the system. An algorithm is developed to ensure fairness at the user level. Second, the system determines what is the optimal throughput allocation to different users at each time within the acceptable throughput ranges provided by the users. Three adaptive algorithms are developed to solve this sub-problem at the system level. The whole scheme can be interpreted as "water filling" each user's throughput in time domain and allocating network throughput to different users each time, according to their channel conditions. From the simulation results, the proposed scheme reduces the overall transmitted power up to 7dB and increases the average spectral efficiency up to 1.2 bit/s/Hz, compared with the previous scheme in [47].

System Model and Problem Formulation

System Model

K co-channel links exist in distinct cells, such as in TDMA or FDMA networks. Each link consists of a mobile unit and its assigned base station. Coherent detection is assumed to be possible so that it is sufficient to model this multiuser system by an equivalent baseband model. Antenna arrays with P elements are used only at the base stations. Each link is affected by the multipath fading, with the propagation delay far less than one symbol duration. The maximum number of paths is L. For the uplink case, the signal at the p^{th} antenna array element of the i^{th} base station can be expressed as:

$$x_i^p(t) = \sum_{k=1}^{K} \sum_{l=1}^{L} \sqrt{\rho_{ki} G_{ki} P_k} \alpha_{ki}^{pl} g_k(t - \tau_{ki}) s_k(t - \tau_{ki}) + n_i^p(t) \tag{3.2}$$

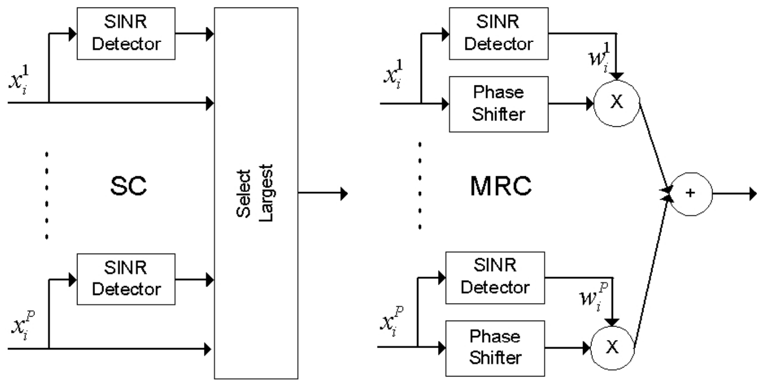


Figure 3.1: Selective Combining and Maximum Ratio Combining

where ρ_{ki} and G_{ki} are the log-normal shadow fading and the path loss from the k^{th} user to the i^{th} base station, respectively, α_{ki}^{pl} is the l^{th} path fading loss from the k^{th} user to the i^{th} base station's p^{th} antenna, P_k is the transmitted power, $g_k(t)$ is the shaping function, $s_k(t)$ is the message symbol, $n_i^p(t)$ is the i^{th} base station's thermal noise at the p^{th} antenna, and τ_{ki} is the channel propagation delay. Here $\tau_{ii} = 0, \forall\, i$ (the delay from the mobile to its assigned BS), and $\tau_{ki}, k \neq i$ (the delay from the mobile to other cell's BS) is uniformly distributed within one symbol duration. The channels change slowly and are stable over a frame with hundreds of symbols. The impulse response from the k^{th} mobile to the p^{th} element of the i^{th} base station is defined as: $h_{ki}^p = \sum_{l=1}^{L} \alpha_{ki}^{pl} r_{ki}^p$, where r_{ki}^p includes the effects of the transmitter, receiver filter, and shaping function $g_k(t - \tau_{ki})$. We define $n_i^p(n)$ as the sampled noise.

Because of the channel distortions, CCI, and thermal noises, the average receivers' SINR can be very low most of the times. Under this condition, in order to satisfy the desired BER, only low modulation rate or even no transmission can be selected. Antenna Diversity is an important means to increase the average receiver's SINR. Consequently, MQAM can be applied with different modulation rates for the desired BER. The antenna outputs can be combined by Maximal Ratio Combining (MRC) or Selective Combining (SC) [3], as shown in Fig. 3.1. MRC diversity requires that the individual signals from each branch be compensated in phase and weighted by the square roots of their SINRs, and then be summed coherently. If perfect knowledge of the branch amplitudes and phases is assumed, when the noise is spatially white, MRC is the optimal diversity-combining scheme and provides the maximum capacity improvement. The disadvantage of MRC is that it requires all knowledge of the branch parameters. SC combiner only chooses the branch with the highest SINR. SC is simpler than MRC but yields suboptimal performance. By using the antenna diversity, the i^{th} base station's combiner output can be written as $\mathbf{w}_i^H \mathbf{x}_i$, where $\mathbf{x}_i = [x_i^1 \dots x_i^P]^T$, and $\mathbf{w}_i$ is a $P \times 1$ combiner weight vector given by:

$$\begin{cases} \text{For MRC:} \;\; |[\mathbf{w}_i]_j| = \sqrt{\Gamma_i^j}, \\ \text{For SC:} \quad [\mathbf{w}_i]_j = \begin{cases} 1, \; j^{th} \text{ antenna has the largest SINR;} \\ 0, \; \text{otherwise.} \end{cases} \end{cases}$$

where Γ_i^p is the received SINR at the p^{th} antenna element that can be calculated from (3.2). The i^{th} base station's combiner output SINR is given by [19]:

$$\Gamma_i = \frac{P_i \rho_{ii} G_{ii} \|\mathbf{w}_i^H \mathbf{h}_{ii}\|^2}{\sum_{k \neq i} P_k \rho_{ki} G_{ki} \|\mathbf{w}_i^H \mathbf{h}_{ki}\|^2 + \mathbf{w}_i^H \mathbf{N}_i \mathbf{w}_i} \tag{3.3}$$

where $\mathbf{h}_{ki} = [h_{ki}^1, \dots, h_{ki}^P]^T$, $\mathbf{N}_i = E\{\mathbf{n}_i \mathbf{n}_i^H\}$, and $\mathbf{n}_i = [n_i^1 \dots n_i^P]^T$.

In adaptive modulation, the transmitters and receivers can adaptively select the modulation rates, i.e. throughput, according to the channel conditions. It has been shown that adaptive modulation can greatly increase the spectral efficiency of wireless communications [65, 54]. In our approach, adaptive MQAM modulation is applied. It has been shown that BER of square MQAM with Gray bit mapping

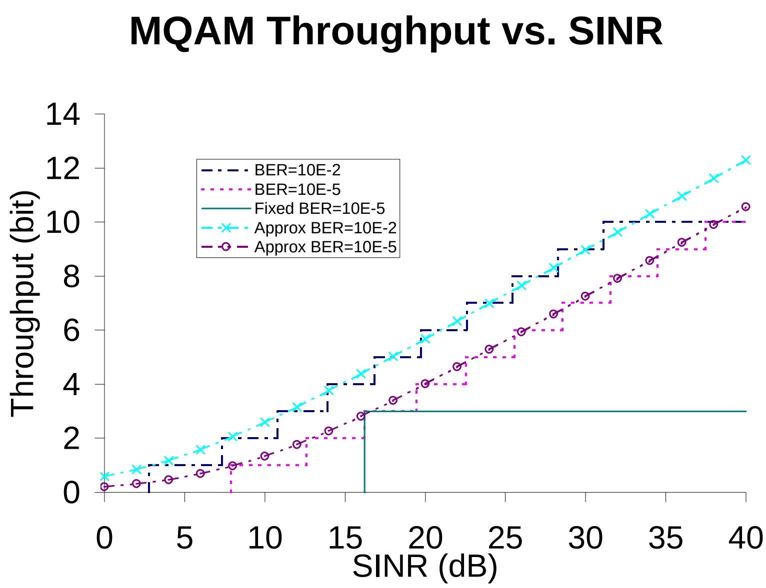


Figure 3.2: BER Approximation and BER Standard Formula for MQAM

as a function of received SINR Γ and constellation size M is approximately given by [3]:

$$\mathrm{BER}(\Gamma, M) \approx \frac{2}{\log_2 M}\left(1 - \frac{1}{\sqrt{M}}\right) erfc\left(\sqrt{1.5\frac{\Gamma}{M-1}}\right) \tag{3.4}$$

where $erfc$ is the complementary error function. This approximation is tight when the SINR Γ is high.

Now the relation between SINR and throughput will be shown. In the i^{th} cell, the i^{th} link between the mobile and its assigned base station uses the modulation with constellation size M_i. Without loss of generality, each user is assumed to have the unit bandwidth. The i^{th} link has throughput $T_i = \log_2(M_i)$. For $\mathrm{BER} = 10^{-2}$ and $\mathrm{BER} = 10^{-5}$, the required SINRs of different constellation sizes are shown in Fig. 3.2. One can see that for the traditional power control with fixed modulation (8-QAM), the receiver must have SINR greater than a specific threshold to have any throughput that satisfies $\mathrm{BER} = 10^{-5}$. In our approach, each user can select a range of different modulation rates. Consequently, the targeted receiver's SINR can be chosen within a range.

It is hard for (3.4) to be inverted and differentiated. In [65, 54], the authors have introduced BER approximations for different modulation rates as:

$$\mathrm{BER}_i \approx c_1 e^{-c_2 \frac{\Gamma_i}{2^{c_3 T_i} - 1}} \tag{3.5}$$

where $c_1 \approx 0.2$, $c_2 \approx 1.5$, and $c_3 \approx 1$. This approximation is tight when the SINR is high. Rearranging (3.5) for a specific BER, the i^{th} link's throughput is given by:

$$T_i = c_4^i \log_2 \left(1 + c_5^i \Gamma_i\right) \text{ bit/s} \tag{3.6}$$

where $c_4^i = 1/c_3^i$ and $c_5^i = -c_2^i / \ln(\mathrm{BER}_i / c_1^i)$. In Fig. 3.2, the approximation is compared with the expression in (3.4) at BER $= 10^{-2}$ and BER $= 10^{-5}$, respectively. It is shown that (3.6) is a good approximation for throughput vs. SINR for a fixed BER.

In reality, the channel estimation errors can affect the performance of adaptive modulation. In our approach, the perfect channel estimation is assumed and it is used in many literature works. Many analysis for the effects of channel estimation errors on adaptive modulation can be found in [8, 10, 11].

Traditional Power Control

In traditional power control problem [47], the SINR of each user is maintained greater than or equal to some threshold γ_i that can provide the adequate link quality. The problem is given by:

$$\min_{\gamma_i} \sum_{i=1}^{K} P_i(\mathbf{\Gamma}) \tag{3.7}$$

$$\text{subject to } \Gamma_i \geq \gamma_i, \ \forall\, i$$

where $\mathbf{\Gamma} = [\Gamma_1 \ldots \Gamma_K]^T$. In this kind of power control, a fixed and predefined targeted SINR threshold γ_i for the desired modulation rate and BER is assigned to each user. Then the transmitted powers are updated to ensure users' targeted SINRs without considering their channel conditions. The system works perfectly in low SINR areas. When the targeted SINRs become high enough, the overall transmitted power will start to increase rapidly. If the targeted SINRs are larger than some specific values, there are no feasible solutions, i.e., the receivers cannot get enough SINR levels, no matter how large the transmitted powers are. One of the underlying reasons for such a problem is that the users with bad channel responses require too many transmitted powers; thus they introduce unnecessarily high CCI to others. Therefore, having a fixed targeted SINR threshold is not an optimal power control approach.

Optimization Problem

All links are assumed to apply MQAM with throughput T_i within a range $[T_i^{min}, T_i^{max}]$, according to their channel conditions, while the overall network throughput $T = \sum_{i=1}^{K} T_i$ is maintained greater than or equal to a constant R. R is equal to the sum of the fixed targeted throughput in the previous scheme [47] in (3.7). R should be selected such that the system is always feasible. If R is too large, it is likely that the overall network throughput will be larger than the overall system capacity, as a result there will be no solution. Each time, the links with bad channel conditions sacrifice their throughput, i.e., they use lower SINR thresholds, which reduce the unnecessary CCI. The links with good channel conditions use higher SINR thresholds, i.e., more bits per symbol are selected, which increases the network throughput. For each link, the time-average throughput is a constant to ensure fairness, and the throughput is "water filled" at different times. For the whole system at any specific time, the overall network throughput is allocated to different links, according to their channel conditions so as to minimize the overall transmitted power. The value of R is also equal to the sum of all users' time average throughput, so that the sum of users' time average throughput and the overall network throughput can coincide. This problem can be summarized as:

$$\min_{T_i, P_i} \sum_{i=1}^{K} P_i \tag{3.8}$$

$$\text{subject to} \begin{cases} \text{Feasibility: } (\mathbf{I} - \mathbf{DF})\mathbf{P} \geq \mathbf{u}, \\ \text{Network Performance: } T \geq R, \\ \text{Throughput Range: } T_i^{min} \leq T_i \leq T_i^{max}, \\ \text{Fairness: } \lim_{N\to\infty} \frac{\sum_{n=1}^{N} T_i(n)}{N} = const_i. \end{cases}$$

where $R = \sum_{i=1}^{K} E(T_i(n))$. Only one type of users is assumed, so $const_i = const_j$, $\forall i, j$. The feasibility constraint $(\mathbf{I}-\mathbf{DF})\mathbf{P} \geq \mathbf{u}$ is the matrix expression for the equalities $\Gamma_i \geq \gamma_i$, $\forall i$ [47], where $\mathbf{u} = [u_1, \ldots, u_K]^T$, $u_i = \gamma_i \mathbf{w}_i^H N_i \mathbf{w}_i / (\rho_{ii} G_{ii} \|\mathbf{w}_i^H \mathbf{h}_{ii}\|^2)$, $\mathbf{P} = [P_1, \ldots, P_K]^T$, $\mathbf{D} = diag\{\gamma_1, \ldots, \gamma_K\}$, and

$$[\mathbf{F}_{ij}] = \begin{cases} 0 & \text{if } j = i, \\ \frac{\rho_{ji} G_{ji} \|\mathbf{w}_i^H \mathbf{h}_{ji}\|^2}{\rho_{ii} G_{ii} \|\mathbf{w}_i^H \mathbf{h}_{ii}\|^2} & \text{if } j \neq i. \end{cases}$$

In the problem defined above, the complexity lies in the optimization over time and grows rapidly with the number of users. In the next part, algorithms are developed to reduce the complexity and distribute the computing efforts to both the system level and the user level.

Problem Partition and Adaptive Algorithms

Problem Partition

The difficulties to solve (3.8) lie in the feasibility and fairness constraints. First, in the feasibility constraint, if the users' transmitted powers are fixed, the targeted

SINR γ_i is linearly constrained. On the other hand, if γ_i is fixed, the constraint is linear for $\mathbf{P}$. However, if both SINRs and powers are considered, it is a Bilinear Matrix Inequality (BMI) problem [68]. The BMI problem is non-convex and non-linear. Only limited tools are available in the literature to find the solutions[68]. Second, in the fairness constraint, the throughput is considered at the different times. It is very difficult to solve the problem by traditional dynamic programming, because the distributions of the received SINRs and transmitted powers are extremely hard to model and calculate. Therefore the problem defined in (3.8) is too difficult to find an analytically optimal solution. A heuristic way is needed to obtain a suboptimal solution with relatively good performances.

If the fairness constraint is not considered, the problem in (3.8) is a pure constrained optimization problem. With the consideration of fairness, the motivation to solve the problem comes from jointly considering the throughput ranges and fairness constraints. First, the users report the ranges of throughput that they can accept. Then the system decides how to allocate the throughput to each user each time, according to these ranges. The acceptable throughput ranges are modified by the users' transmission history. Each time, some users may have more throughput, while others have less. Then the users with more throughput will become less aggressive about transmitting and will request smaller throughput ranges in the near future, and vice versa. From the above idea, the optimization problem in (3.8) is divided into two sub-problems:

1. At the user level, in order to ensure fairness, the users trace their histories of throughput and report the ranges of throughput that they can accept to the system at current time.

2. At the system level, for the whole network each time, the system determines

the optimal throughput allocation to different users, and this allocation requires the lowest overall transmitted power.

Therefore, the overall transmitted power is minimized each time, and fairness is guaranteed. However, the optimal solution for (3.8) is not guaranteed to be achieved. But, from the simulation results, the significant performance improvements over the traditional system [47] will be shown.

An illustrative example for two users is shown in Fig. 3.3. The two axes represent the two users' desired SINRs that are related to their throughput. The provided ranges are the required SINRs for the throughput ranges that the users provide, and these ranges are also restricted by the feasibility constraint. On the dashed line, the overall network throughput $T = T_1 + T_2$ is a constant. At the system level, the goal is to find what is the optimal point each time that requires the minimum overall transmitted power, within the range (shown as the polyhedra) and under the overall throughput constraint $T \geq R$. At the user level, the problem is how to change the throughput ranges over different times to ensure fairness. For example, if user 1 is assigned to have small throughput now, he will be more aggressive about transmitting his data in the future. Consequently, the throughput range will move to the right within the practical range, and user 1 has to be assigned the higher throughput in the future.

Adaptive Algorithm for Throughput Range at the User Level

In this part, the first sub-problem will be solved. An adaptive algorithm is developed at the user level to report the acceptable throughput ranges back to system, so as to ensure fairness. The key idea is to adapt the throughput ranges with joint consideration of the fairness constraint. Instead of having a fixed throughput range $[T_i^{min}, T_i^{max}]$ for each link, the throughput ranges are adaptively changed by taking

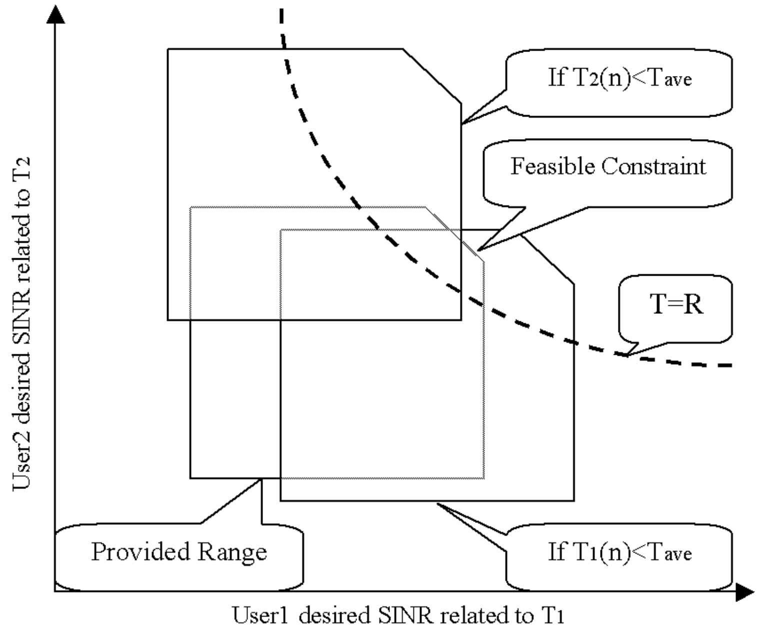


Figure 3.3: Two Users Example for Problem Partition

into account the links' throughput histories. Assume the i^{th} link can select throughput $T_i^{min}(n) \leq T_i(n) \leq T_i^{max}(n)$ at time n, and the average desired throughput for the i^{th} link is T_i^{ave}. Each time, $T_i^{min}(n+1)$ and $T_i^{max}(n+1)$ are modified by the current $T_i(n)$. When $T_i(n)$ is smaller than T_i^{ave}, $T_i^{min}(n+1)$ and $T_i^{max}(n+1)$ are increased so that there is a higher probability that the future throughput $T_i(n+1)$ is larger than T_i^{ave}. When $T_i(n)$ is larger than T_i^{ave}, $T_i^{min}(n+1)$ and $T_i^{max}(n+1)$ are decreased so that there is a higher probability that $T_i(n+1)$ is smaller than T_i^{ave}. $T_i^{min}(n+1)$ and $T_i^{max}(n+1)$ are bounded by $\widehat{T}_i^{min}$ and $\widehat{T}_i^{max}$, which are the practical minimum and maximum throughput boundaries that the i^{th} link can select, respectively. Their values are fixed and predefined by the system. In order to track the history of T_i, $T_i^{mid}(n) = T_i^{mid}(n-1) + \beta(T_i(n) - T_i^{ave})$, $0 < \beta < 1$, where β is a constant that depends on how much delay the user can suffer. If the delay constraint is tight, β should be selected as a relatively larger number, so that the throughput range will move quickly to compensate the user's throughput loss at a specific time. If the user can suffer longer delay, β could be selected as a relatively smaller number, so that the user can wait until the channel becomes better to be compensated back. The selected value of β is also affected by how rapidly the channels change. If the channels change slowly, a smaller β is preferred, so that the user can wait; otherwise, a larger β is selected. Each time, the throughput window is updated by:

$$\begin{cases} T_i^{mid}(n) = T_i^{mid}(n-1) + \beta(T_i(n) - T_i^{ave}), \\ T_i^{min}(n+1) = \min(\max(T_i^{ave} - T_i^{mid}(n) + \widehat{T}_i^{min}, \widehat{T}_i^{min}), \widehat{T}_i^{max}), \\ T_i^{max}(n+1) = \max(\min(\widehat{T}_i^{max} - T_i^{mid}(n) + T_i^{ave}, \widehat{T}_i^{max}), \widehat{T}_i^{min}). \end{cases} \tag{3.9}$$

The above throughput window may move to the opposite direction of the channel changing trend. When the channel is bad, the user selects less throughput. But in the next time, the user has to select a larger throughput because the throughput window moves to a higher throughput area, even if the channel is still bad. With the consideration of the channel changes, a scheme is developed so that the throughput window follows the channel changing trend. This problem can be categorized as a dynamic programming problem given by:

$$[T_i^{min}(n+1), T_i^{max}(n+1)] = f_n([T_i^{min}(n), T_i^{max}(n)], v_n, T_i(n)), \quad n = 0, 1, \ldots, N-1 \tag{3.10}$$

where f_n is a function to select the throughput window at time n, and v_n is the control policy that has a different impact on the outcomes of f_n. The problem in (3.10) is extremely difficult to solve, but an intuitive idea can be applied to find a much simpler solution. Because β may not be an integer, the throughput window developed in (3.9) may not be discrete. If the i^{th} user's assigned throughput at the current time n is smaller than the median of all the users' assigned throughput in the adjacent cells, this means that the i^{th} user is possibly still under the bad channel condition. The lower throughput window is assigned to follow the channel condition, by using the floor of the original throughput window. Here the floor is a function that finds the maximum integer immediately less than the real value. On the other hand, if the i^{th} user's throughput is larger than the median of the users' throughput among the adjacent cells, the higher throughput window is assigned to follow the channel condition by using the ceiling of the original throughput window. Here the ceiling is a function that finds the minimum integer immediately greater than the real value.

In addition, when a user is trapped in a bad channel for a long time, instead of assigning him with a very high throughput range, the algorithm should be able to assign this user with lower throughput. By doing so, the user will not cause too much CCI to others, and the system performance can be improved. The history of $T_i^{min}(n)$ is tracked. If the user detects Z consecutive $T_i^{min}(n)$ equal to $\widehat{T}_i^{max}$, the user will report the acceptable throughput range as $[\widehat{T}_i^{min}, T_i^{max}(n)]$, instead of $[T_i^{min}(n), T_i^{max}(n)]$. Consequently, the system is able to assign the minimal throughput to the user. The throughput ranges are updated by users to BS every power update interval. Because the ranges are discrete and limited by the hardware, the associated overheads to report these ranges are small. In a real system, this information is coded by a powerful error control code to ensure that it comes through without errors. In each iteration, users' throughput windows are updated by:

Adaptive Algorithm for Each User's Throughput Window

1. **Initialization:** $T_i^{min}(0) = \widehat{T}_i^{min}, \;\; T_i^{max}(0) = \widehat{T}_i^{max}, \;\; T_i^{mid}(0) = T_i^{ave}$

2. **Iteration:**

$T_i^{mid}(n) = T_i^{mid}(n-1) + \beta(T_i(n) - T_i^{ave});$

if $T_i(n) > \text{median}(T_j(n)),\ j \in$ all adjacent CCI cells,

$T_i^{min}(n+1) = floor(\min(\max(T_i^{ave} - T_i^{mid}(n) + \widehat{T}_i^{min}, \widehat{T}_i^{min}), \widehat{T}_i^{max}));$

$T_i^{max}(n+1) = floor(\max(\min(\widehat{T}_i^{max} - T_i^{mid}(n) + T_i^{ave}, \widehat{T}_i^{max}), \widehat{T}_i^{min}));$

else

$T_i^{min}(n+1) = ceiling(\min(\max(T_i^{ave} - T_i^{mid}(n) + \widehat{T}_i^{min}, \widehat{T}_i^{min}), \widehat{T}_i^{max}));$

$T_i^{max}(n+1) = ceiling(\max(\min(\widehat{T}_i^{max} - T_i^{mid}(n) + T_i^{ave}, \widehat{T}_i^{max}), \widehat{T}_i^{min}));$

3. **Feedback the Acceptable Throughput Ranges to BS:**
   if $T_i^{min}(n+1) = T_i^{min}(n) = \ldots = T_i^{min}(n-Z+1) = \widehat{T}_i^{max}$,

   report $[\widehat{T}_i^{min}, T_i^{max}(n)]$; else, report $[T_i^{min}(n), T_i^{max}(n)]$;

If a user is never trapped in the bad channel for a long period of time, when $T_i(n)$ is continuously less than T_i^{ave} for some time, $T_i^{min}(n)$ is increased to T_i^{ave}. Then the next $T_i(n+1)$ has to select the throughput equal to or greater than T_i^{ave}; consequently, $T_i^{mid}(n)$ stops increasing. The same analysis can be applied to $T_i^{max}(n)$. Since $T_i^{min}(n)$ and $T_i^{max}(n)$ are bounded and are linearly modified by $T_i^{mid}(n)$, $T_i^{mid}(n)$ is also bounded. If $T_i^{mid}(n)$ is rearranged and summed over the different times,

$$\frac{\sum_{n=1}^{N} T_i(n)}{N} = T_i^{ave} + \frac{(T_i^{mid}(N) - T_i^{ave})}{\beta N}. \tag{3.11}$$

The second term on the right hand side decreases to zero as $N \to \infty$. So $\lim_{N\to\infty} \frac{\sum_{n=1}^{N} T_i(n)}{N} = T_i^{ave}$, i.e., the system is fair, so that each user's time-average throughput is a constant.

If a user is trapped in the bad channel for a long period of time and detects Z consecutive $T_i^{min}(n)$ equal to $\widehat{T}_i^{max}$, the user will report the acceptable throughput range as $[\widehat{T}_i^{min}, T_i^{max}(n)]$. Under this condition, $T_i^{mid}(n)$ will not be bounded. If the channel becomes better in the future and the system assigns more throughput to this user, $T_i^{mid}(n)$ will be increased. Consequently, $T_i^{max}(n)$ will be less than $\widehat{T}_i^{max}$, and the second term on the right hand side of (3.11) will approach to zero, asymptotically. If a user is trapped in the bad channel indefinitely, $T_i^{mid}(n)$ will go to negative infinity, and fairness constraint cannot be satisfied. In practice, this situation seldom happens. If it does happen, there is no practical meaning to guarantee fairness for this user because this user will cause too much CCI that will reduce the system performance a lot.

Adaptive Algorithms for Throughput Allocation at the System Level

In this part, the second sub-problem will be solved, and three adaptive algorithms will be developed at the system level to allocate throughput to different users each time to generate the minimum overall transmitted power. The first one is a full search algorithm that can guarantee to find the optimal solution each time, but the complexity is very high. The second one is a fast search algorithm that analyzes which users contribute more to the overall transmitted power. The last one is an adaptive algorithm by assuming that the throughput is continuous and approximated by (3.6). Then the throughput allocation result is projected to the closest discrete value that satisfies all the constraints.

Full Search Algorithm

Because there is only limited number of discrete throughput T_i that each user can select, and there are only limited number of users, a full search method can be applied to find the optimal throughput allocation. The users provide the acceptable throughput ranges to the system. The system calculates the overall transmitted powers of all combinations of T_i by the iterative algorithm under the constraints in (3.8). The throughput allocation that generates the lowest overall transmitted power is selected. The adaptive algorithm can find the optimal solution each time, but it has very high complexity. The complexity is increased exponentially with the number of users, which is not acceptable in practice. It can be used as a performance bound. The full search adaptive algorithm is given by:

Full Search Adaptive Algorithm for Throughput Allocation

1. **Adaptive Modulation:**
   search all possible $T_i(n)$ for every user subject to the constraints.
   find the combination of $T_i(n)$ that minimizes $\sum_{i=1}^{K} P_i$
   calculated by the iteration.
2. **Iteration:**
   - Initialization: $P_1, \ldots, P_K =$ any positive feasible values
   - Antenna Diversity: $\mathbf{w}_i = \arg\max_{\mathbf{w}_i} \Gamma_i$
   - Power Allocation Update Iteration:
     $\gamma_i =$ required SINR for $T_i(n)$ and desired BER;
     $\mathbf{D} = diag(\gamma_1, \ldots, \gamma_K)$; $\mathbf{P} = \mathbf{DFP} + \mathbf{u}$.
3. **Throughput Range Update:**
   Update $T_i^{mid}(n)$, $T_i^{min}(n)$, and $T_i^{max}(n)$.

Fast Search Algorithm

In order to reduce complexity, a fast search algorithm is developed. The system needs to find out which users contribute more to the overall transmitted power. The gradient of overall transmitted power to each user's targeted SINR is derived. If the users with larger gradients can sacrifice their SINRs a little bit, the overall transmitted power will be reduced significantly.

In the Perron-Frobenius theorem [67], if the spectrum radius of $\mathbf{DF}$, $\rho(\mathbf{DF})$, i.e., the maximum absolute eigenvalue, is less than 1, the minimum overall transmitted power is achieved when $\Gamma_i = \gamma_i$, $\forall$ i, and P_{sum} can be written as:

$$P_{sum} = \sum_{i=1}^{K} P_i(T_i(n)) = \mathbf{1}^T(\mathbf{I} - \mathbf{DF})^{-1}\mathbf{u} \tag{3.12}$$

where $\mathbf{1} = [1 \ldots 1]^T$. Define $\mathbf{Q} = [\mathbf{I} - \mathbf{DF}]^{-1}$. If $(\mathbf{DF}) \in R^{K \times K}$ and $\rho(\mathbf{DF}) < 1$, then $\mathbf{Q} = \sum_{j=0}^{\infty} (\mathbf{DF})^j$. Since $\mathbf{D} = diag(\gamma_1, \ldots, \gamma_K)$, and $[\mathbf{F}]_{ij} > 0, \forall i, j$, if $\gamma_j,\ j = 1 \ldots K,\ j \neq i$ is fixed, every component in $\mathbf{Q}$ is a function of $(\gamma_i)^j, j = 1 \ldots \infty$ with non-negative coefficients. In vector $\mathbf{u}$, every u_i has the non-negative coefficients as well. So $P_{sum} = \sum_{i=1}^{K} P_i = \mathbf{1}^T\mathbf{Q}^{-1}\mathbf{u}$ is also a function of $(\gamma_i)^j$ with non-negative coefficients. The only situation where the coefficients are zeros is when the antenna diversity uses a null for the desired mobile user. This hardly happens in practice. Since $\gamma_i > 0, \forall i$, when the other $\gamma_j,\ j = 1 \ldots K,\ j \neq i$ is fixed, P_{sum} is a convex and increasing function of γ_i. From (3.6), γ_i is an increasing and convex function of throughput T_i. So P_{sum} is also an increasing and convex function of T_i, when the other $T_j,\ j = 1 \ldots K,\ j \neq i$ is fixed. Consequently, the overall transmitted power is minimized when the network throughput constraint is equal, i.e., $T = R$. This is because any T_i can be reduced to have smaller overall transmitted power, if $T > R$.

Now the gradients of overall transmitted power can be deduced. If the derivatives are taken with respect to γ_i at both sides of (3.12), the i^{th} element of gradient $\mathbf{g} = [g_1 \ldots g_K]^T$ is given by [16]:

$$g_i = \frac{\partial P_{sum}}{\partial \gamma_i} = \frac{c_i P_i}{\Gamma_i} \tag{3.13}$$

where $c_i = \mathbf{1}^T(\mathbf{I} - \mathbf{DF})^{-1}\mathbf{v}_i$, and $[\mathbf{v}_i]_j = 1$, if $j = i$; $[\mathbf{v}_i]_j = 0$, otherwise. The value of c_i reflects how severe the CCI is. When the CCI is large, c_i tells how much the i^{th} user causes the CCI to other users. When the CCI is small, $c_i \approx c_j,\ \forall i, j$. Since the adaptive algorithm only needs the direction of the gradients and does not need the amplitudes, the value of c_i can be ignored, i.e., $c_i = 1, \forall i$, when the CCI is small. Equation (3.13) is very significant in that it provides a very simple way to find the gradients. In this case, SINRs can be measured at each base station's antenna diversity output, and the feedback channels can be used to get the mobile transmitted power values to calculate the gradients. Consequently, the complexity can be reduced greatly.

With the gradients, a greedy algorithm is developed. First, because the network throughput constraint $T = R$ is non-linear, the overall transmitted power P_{sum} is no longer a convex function of γ_i under this constraint. The gradients of different users are compared. If a user with a larger gradient selects lower throughput, i.e., he requires a lower targeted SINR threshold, the overall transmitted power is greatly reduced. So first the throughput that generates the lowest overall transmitted power is decided for the user with the highest gradient, subject to the constraints. When the throughput of the user with the largest gradient is changed, the throughput of the other users is modified in the order from the lower gradient to higher gradient to compensate the network throughput constraint $T = R$. By doing this, more throughput is allocated to the users with small gradients, and less throughput is assigned to the users with large gradients; Consequently, the overall transmitted power will be reduced significantly. Note that the throughput of the user with the largest gradient may not end up with the lowest throughput $T_i^{min}(n)$, because the increase of the sum of other user's powers may be larger than the decrease of this user's power. In the next iteration, the throughput of the user with the largest gradient is fixed, and the system finds the optimal throughput for the user with the second highest gradient, and so on until we find the throughput of the last user in the row. Because every user only searches for a fixed amount of throughput range and reordering is needed, if the gradient is calculated by (3.13) and $c_i \approx c_j, \forall\ i, j$ for simplicity, this sub-optimal algorithm has the complexity of only $O(K^2 \log_2 K)$. If the CCI is severe and $c_i \neq c_j$, then the complexity is $O(K^3 \log_2 K)$. If the user index is rearranged from the largest gradient to the lowest, i.e., $g_1 \geq g_2 \geq \ldots \geq g_K$, and any non-feasible solution has $P_{sum} = \infty$, the sub-optimal adaptive iterative algorithm is summarized by:

Fast Search Adaptive Algorithm for Throughput Allocation

1. **Initialization:**
   $T_1(0) = T_1^{ave}, \ldots, T_K(0) = T_1^{ave},$

   $P_1, \ldots, P_K =$ any feasible positive const.

2. **Adaptive Modulation**
   for $i = 1$ to K

   for $T_i = T_i^{min}(n)$ to $T_i^{max}(n)$

   1.Modify from $T_K(n)$ to $T_{i+1}(n)$

   to satisfy the constraint $T = R$ (exhaust T_K first.)

   2.Run iteration

   •Antenna Diversity: $\mathbf{w}_i = \arg\max_{\mathbf{w}_i} \Gamma_i$

   •Power Allocation Update Iteration:

   $\gamma_i =$ required SINR for T_i and desired BER,

   $\mathbf{D} = diag(\gamma_1 \ldots \gamma_K), \quad \mathbf{P} = \mathbf{DFP} + \mathbf{u}.$

   3.Find T_i that generates the lowest power for the i^{th} user.

   end

   end

3. **Throughput Range Update:** Update $T_i^{mid}(n)$, $T_i^{min}(n)$, and $T_i^{max}(n)$.

The algorithm is suboptimal because the optimal throughput for one user may not be optimal for all users. The algorithm may stop at some local minimum points or the boundary points. From the simulation results that will be shown later, the sub-optimal algorithm has relatively good performance.

Projected Gradient Algorithm

As what have been stated, the feasible constraint in (3.8) is a BMI constraint. Here the approximation of throughput in (3.6) is used, and a projected gradient algorithm [68] is developed to change each user's targeted SINR to find the minimal overall transmitted power. The throughput allocation results are probably not integers. The results are projected to the nearest discrete throughput allocation that satisfies the constraints. Then the above two steps are employed again, until the discrete throughput allocations are the same in two consecutive iterations.

First, the projected gradient method will be developed. The throughput is now supposed to be continuous and has the value $\widetilde{T}_i$. If each user's targeted SINR is changed by the gradient in (3.13), the overall throughput constraint $\sum_{i=1}^{K} \widetilde{T}_i = R$ cannot be satisfied. The gradient needs to be modified such that the overall throughput constraint holds. The plane that is tangent to the curve $\sum_{i=1}^{K} \widetilde{T}_i = R$ at point $[\gamma_1, \ldots, \gamma_K]$ needs to be found, where $\gamma_i = (2^{\widetilde{T}_i/c_4^i} - 1)/c_5^i$. Without loss of generality, this plane can be moved to the origin. The plane can be expressed as: $\sum_{i=1}^{K} k_i x_i = 0$, where $k_i = c_4^i c_5^i/(1 + c_5^i \gamma_i)$. The modified gradient is given by $\mathbf{q} = [q_1 \ldots q_K]^T$. By the definition of projection, vector $\mathbf{q}$ satisfies equation $\|\mathbf{g} - \mathbf{q}\|^2 = \min_{\forall \mathbf{q} \in plane} \|\mathbf{g} - \mathbf{q}\|^2$. The right hand side needs to be minimized to get the optimal vector, i.e., the projection $\mathbf{q}$.

The best throughput allocation $\widetilde{T}_i$ is obtained from the above projected gradient algorithm. $\widetilde{T}_i$ needs to be projected to a discrete value. The projection problem can be written as:

$$\min_{T_i} \|\widetilde{T}_i - T_i\|^2 \tag{3.14}$$

$$\text{subject to } \sum_{i=1}^{K} T_i = R,\ T_i^{min}(n) \leq T_i \leq T_i^{max}(n),\ \text{ and } (\mathbf{I} - \mathbf{DF})\mathbf{P} \geq \mathbf{u}.$$

where $\widetilde{T}_i$ is projected to the discrete value with the constraint $\sum_{i=1}^{K} T_i = R$. However the discrete throughput projection may not be feasible or not in the ranges. If so, the second closest point needs to be found to see if it satisfies all the constraints. The search is continued until a feasible solution is found. The projected gradient algorithm is given by:

Projected Gradient Algorithm

1. **Initialization:**
   $T_1(0) = T_1^{ave}, \ldots, T_K(0) = T_K^{ave}$,
   $P_1, \ldots, P_K =$ any feasible positive const.
2. **Iteration:** Stop when T_i is stable.
   - Antenna Diversity: MRC or SC
   - Adaptive Threshold Allocation

     do

     {SINR Range:$\gamma_i^{min} = (2^{T_i^{min}/c_4^i} - 1)/c_5^i;\ \gamma_i^{max} = (2^{T_i^{max}/c_4^i} - 1)/c_5^i$

     Projected Gradient:

     $\mathbf{g} = \bigtriangledown P_{sum};\ \mathbf{q} = \text{projection}(\mathbf{g});\ \gamma_i = \gamma_i - \mu \cdot q_i\ \forall\ i;$

     Within Range:if $(\gamma_i > \gamma_i^{max})\ \ \gamma_i = \gamma_i^{max};$ if $(\gamma_i < \gamma_i^{min})\ \ \gamma_i = \gamma_i^{min}$}

     while (γ_i not stable, not boundary)

- Adaptive Modulation: Select $\widetilde{T}_i = c_4^i \log_2(1 + c_5^i \gamma_i)$.
- Throughput Projection:

  Project $\widetilde{T}_i$ to the nearest T_i that satisfies the constraints.
- Power Update Iteration:

  $\gamma_i = (2^{T_i/c_4^i} - 1)/c_5^i$; $\mathbf{D} = \mathrm{diag}(\gamma_1, \gamma_2 ... \gamma_K)$; $\mathbf{P} = \mathbf{DFP} + \mathbf{u}$.

3. **Throughput Range Update:** Update $T_i^{mid}(n)$, $T_i^{min}(n)$, and $T_i^{max}(n)$.

In the algorithm, μ is a small constant, whose value decides the rate of convergence and the variance of the final results. Whether or not γ_i is stable is decided by comparing the maximum difference of γ_i in two consecutive steps. The algorithm has complexity of $O(K^3)$. However, because two iterations are needed each time, the complexity is higher than that of the fast search algorithm but still much lower than that of the full search algorithm, when the number of users is large. From the simulation results, it will be shown that the projected gradient algorithm can find the optimal solution each time.

The algorithm starts from any feasible rate and power allocation. In each iteration, the gradient of the overall transmitted power is calculated and projected on a plane where the network performance constraint is satisfied. This modified gradient is at least pointing in the direction where the overall transmitted power is increasing. The algorithm modifies the SINR allocation at the opposite direction of this modified gradient, so that the new overall transmitted power is less than or equal to that of the old iteration. When the algorithm finds the SINR allocation solution, this SINR allocation must be feasible, and the transmitted powers are updated by fixing the targeted SINRs. This power update iteration converges to a unique solution [5, 47].

Joint Power Control and Throughput Management System

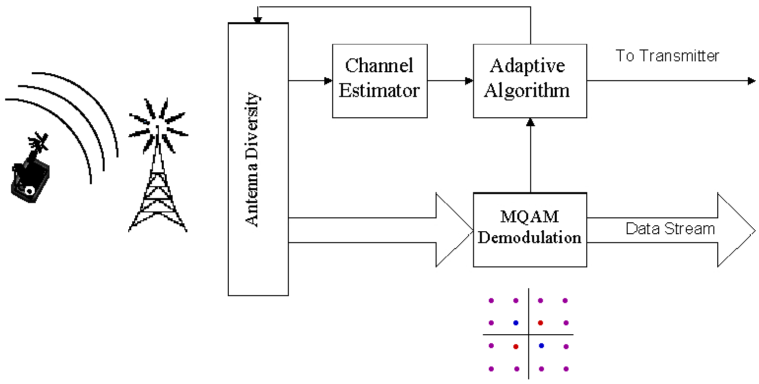


Figure 3.4: Power Control and Throughput Management System

With the adaptive algorithms, a joint power control and throughput management system is constructed in Fig. 3.4. Because of users' multipath fading, shadowing, and random locations in their respective cells, the channel conditions are varying. Therefore, accurate techniques for "real time" estimations of channel conditions are essential [8, 11, 3]. The fluctuations of channels are assumed to be tracked perfectly by the base stations. This information is sent back to the mobile users via an error-free feedback channel. The time delay in this feedback channel is also assumed to be negligible, compared to the speed of channel and CCI variations. All these assumptions are reasonable in slowly varying channels.

The way the system works and the distribution of computing efforts are shown as follows: At the user level, the users compute and provide the system with their acceptable throughput ranges, according to their transmission histories and current channel conditions. At the system level, where the base stations have much stronger computing power, the adaptive algorithm module gets the estimation of users' channel responses from the channel estimation module. Then power control and modulation rates are computed. The power control and best throughput allocation information is sent back to the mobile users. Then the mobile users, accordingly, adapt their transmission rates and powers.

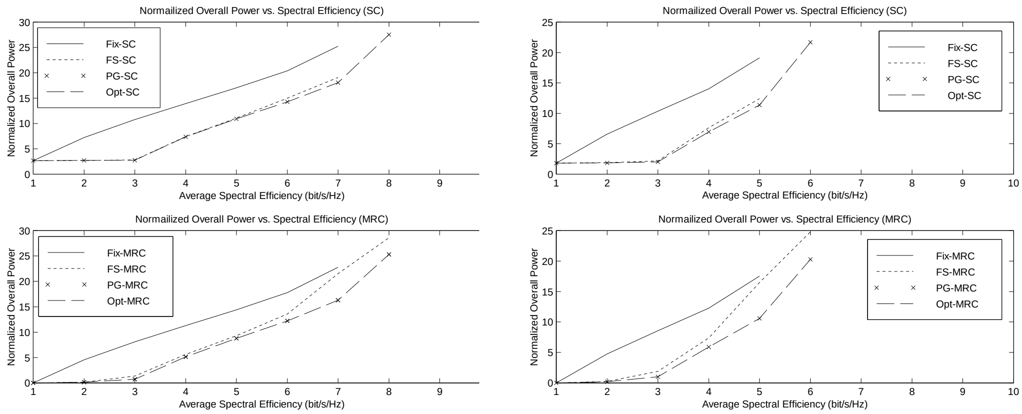
For the mobile device with battery power supply, the maximum transmitted power is limited. In the optimization problem, the maximum power constraint can also be considered. In the proposed approaches, this constraint can be easily implemented by the full search and fast search algorithms. The algorithms are modified such that only the throughput allocation that satisfies the maximum power constraint will be selected. However, in the projected gradient method, the maximum power constraint will impose another very complex and non-linear constraint in the proposed adaptive algorithm.

Simulation Results

In order to evaluate the performances of the proposed algorithms, a network with hexagonal cells is simulated. The radius of each cell is 1000m. Two adjacent cells do not share the same channel, i.e., the reuse factor is 7. There are 50 cells in the networks. One base station is placed at the center of each cell. In each cell, one user is placed randomly with a uniform distribution. In the simulation, the fading is considered as complex Gaussian distributions with three multipaths of equal powers. The fading is independent between two resource allocation intervals. Each base station has a P-element antenna array. Noise power is 10^{-3}. $Z = 50$. 3dB log-normal distribution is considered. In our approach, the two slopes path loss model [20] is applied to obtain the average received power as a function of distance. We select the basic path loss exponent as 2, the additional path loss exponent as 2, the base station antenna height as 50m, the mobile antenna height as 2m, and the carrier frequency as 900-MHz. In Table 3.1, the overall transmitted power of our proposed system is shown with respect to different number of antennas. The value is normalized with the case where only single antenna is applied.

Table 3.1: Normalized Transmitted Power with Respect to No. of Antennas

No. of antennas	2	3	4	5	6	7
MRC(BER=10^{-2})	0.2722	0.1271	0.0904	0.0598	0.0468	0.0412
SC(BER=10^{-2})	0.3746	0.2095	0.1677	0.1260	0.1083	0.1065
MRC(BER=10^{-5})	0.1519	0.0787	0.0572	0.0463	0.0349	0.0264
SC(BER=10^{-5})	0.1958	0.1248	0.0873	0.0797	0.0716	0.0655



(a) BER=10^{-2}

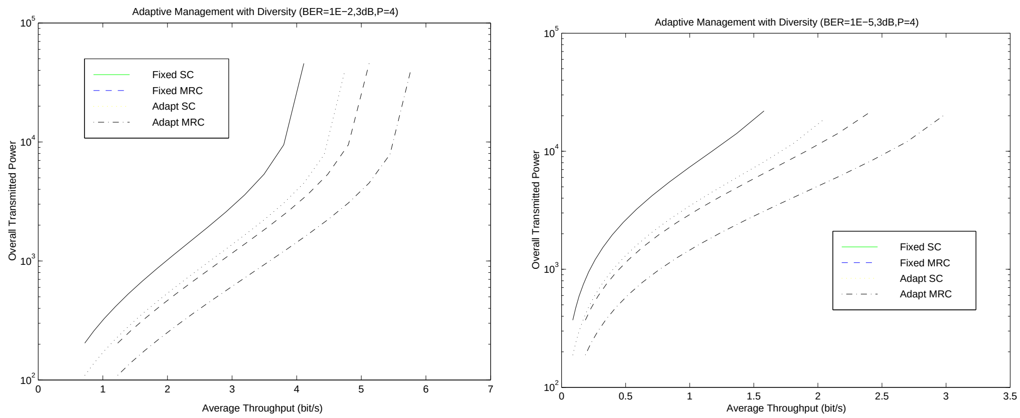
(b) BER=10^{-5}

Figure 3.5: Normalized Power (dB) vs. Throughput

Two different BER requirements (BER $= 10^{-2}$ and BER $= 10^{-5}$) are shown respectively . The overall transmitted power can have a reduction of about 75% to 95% for MRC compared to the single antenna case. The performance of SC is consistently worse than that of MRC. Since SC can apply non-coherent modulation, the complexity is much smaller. When the desired BER is decreased, SC performs closer to MRC. With the number of antennas P increasing, from simulations, the decrease of powers saturates around $P = 4$. Therefore, $P = 4$ is chosen for rest of simulations.

In Fig.3.5 (a) and (b), the normalized overall transmitted power as a function of average spectral efficiency (bit/s/Hz) is compared for the fixed scheme [47], fast search scheme, projected gradient algorithm, and optimal full search scheme with MRC and SC diversity at BER $= 10^{-2}$ and BER $= 10^{-5}$, respectively. We normalize the power with MRC scheme when the spectral efficiency equals to 1. Each user is assumed to have the same desired time-average throughput $T_1^{ave} = \ldots = T_K^{ave}$. Define window size $= (T_i^{max} - T_i^{ave}) = (T_i^{ave} - T_i^{min}) = 2$ bit/s, $\forall i$. Each user is assumed to have unit bandwidth. From the simulation results, the projected gradient algorithm can find the optimal solution obtained by the full search algorithm. Because there is only one allocation scheme available when the average spectral efficiency is equal to one, all the algorithms perform the same. When the average spectral efficiency increases, the proposed algorithms greatly reduce the overall transmitted power and increase the maximum achievable throughput. The suboptimal fast search algorithm has the performance between those of the fixed scheme and optimal scheme. The results show that the proposed scheme can reduce the overall transmitted power by about 7dB, when the average spectral efficiency is larger than 2. The proposed scheme also increases the maximum spectral efficiency by about 1 bit/s/Hz. In the lower spectral efficiency range, the suboptimal fast search algorithm has almost the same performance as that of the optimal solution. If the MRC diversity is employed, it reduces about 3dB to 4dB more transmitted powers than those of SC diversity. The SC diversity and proposed sub-optimal algorithm have a lower complexity.

In order to further study the projected gradient method, the throughput is assumed to be continuous. In Fig. 3.6 (a) and (b), the MQAM performances are compared with MRC and SC diversity at BER $= 10^{-2}$ and BER $= 10^{-5}$,

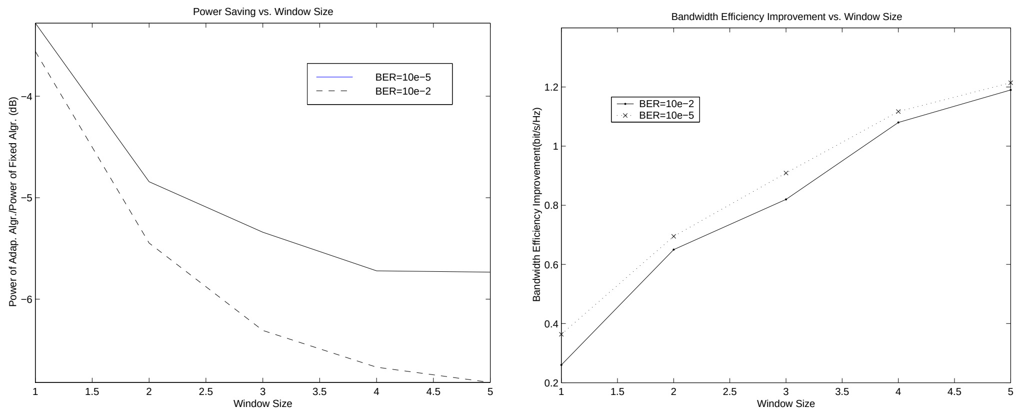


(a) BER=10^{-2} (b) BER=10^{-5}

Figure 3.6: MQAM Performance with Continuous Throughput Assumption

respectively. The simulations are conducted from time 1 to 1000. From the results, it is shown that the adaptive algorithms can improve the average spectral efficiency by 0.9 bit/s/Hz, and decrease the overall transmitted power by 40% less than those of the fixed schemes. The MRC scheme again has better performance than SC scheme. The overall transmitted power of MRC is 40% less than that of SC. The maximum achievable spectral efficiency of MRC is about 0.7 bit/s/Hz to 0.9 bit/s/Hz higher than that of SC. However, this improvement is decreasing as BER is getting smaller.

In Fig. 3.7 (a) and (b), the average power saving and average spectral efficiency gain are shown as the functions of window size. The overall transmitted power can be reduced up to 7dB, and the spectral efficiency can be increased up to 1.2 bit/s/Hz. The power stops decreasing and spectral efficiency increasing speed is reduced, as the window size is growing. This is because of the time-average throughput constraint for each user. The user that gets better throughput at this time must pay back in the future. So there is no need to have a very large window



(a) Power Saving (b) Spectral Efficiency Gain

Figure 3.7: Effects of Window Size

size. Only a limited number of modulation rates are necessary; consequently, the system complexity can be simple.

3.4 Link Quality and Power Management with Space-Time Diversity

In multi-access wireless networks, dynamic allocation of resource such as link qualities and transmitted powers is an important means to combat time varying fading environments and co-channel interferences (CCI). In most prior works, every link's quality is maintained by having a fixed signal-to-interference-noise-ratio (SINR) requirement. We discover that such a constraint is too strong and can degrade the performance of the whole wireless networks, because a user with a bad channel response requires too much transmitted power and therefore causes unnecessary CCI to other users. In our approach, we alleviate this constraint and explore the time diversity. For each user, the time average link quality is maintained as a constant to ensure fairness. For the whole system, we want to minimize the overall transmitted power. In order to solve this problem, each user provides the system with a SINR range that is acceptable, according to his channel conditions and transmission history. Then the system allocates the resources according to these ranges, channel conditions, and other practical constraints. Each time, some users may sacrifice their performances to reduce the overall network transmitted power. These users' temporary sacrifices will improve the system performance and will be paid back in long term. The scheme can be conceived as "water filling" the wireless network resources to different users at different times. In addition, by combining our proposed scheme with beamforming, we can have one more degree of freedom to combat CCI's in different directions of arrivals and different channel conditions over time.

The organization of section is as follows: First, we give the motivation and sketch of the proposed scheme. Then, we present the system model. We first explain the traditional power control problem and point out its shortcomings. We develop adaptive algorithms to reduce the overall transmitted power by alleviating the fixed link quality constraint and exploring time diversity. We discuss downlink cases and point out the differences from the uplink cases. Joint beamforming and our proposed scheme is presented. We have simulation studies.

Motivation and Sketch

Resource allocation for the wireless networks has been extensively studied in the literature. In [94, 21, 22, 43, 44, 23], classical power control algorithms are presented, and their convergence is proved. In [54, 24, 25, 16, 18, 55, 117, 22], the authors study combining rate adaption and power control to optimize the system performance. In [47, 95, 26, 27], beamforming, power control, multiuser detection, and base station assignment are combined for cellular wireless communication system. In [48, 28], the problem of optimal resource allocation is considered from the information theoretic point of view. Throughput capacity and delay-limited capacity are extensively studied. In [49, 30], dynamic programming is considered for integrating link adaptation and power control to improve the overall throughput. In [31, 32], game theory is introduced to power control problem. The utility functions are designed for users to compete resources with each other. The system is balanced in some equilibrium.

In the traditional power control, the overall transmitted power is minimized, while each user modifies his transmitted power, so that his received SINR is larger than or equal to a fixed and predefined targeted SINR threshold required maintaining his link quality. However a user with a bad channel response will transmit a very high power, therefore he can cause unnecessary CCI to other users. As a result, the overall network performance is degraded. In our approach, by alleviating the fixed link quality constraint and exploring the time diversity, we develop adaptive joint link quality and power management schemes with fairness constraint for both uplink and downlink. The schemes encourage some users to sacrifice their performances in a short period, with the incentive that the overall network transmitted power can be reduced and the users' temporary sacrifices will be paid back in a long term.

In the proposed schemes, each user provides the system with a SINR range that he can accept each time. Then the system employs adaptive algorithms to assign different users their targeted SINR's, according to their acceptable SINR ranges, channel conditions, and other practical constraints. Different users may have different assigned SINR's each time, while each user's time average SINR is maintained as a constant to ensure fairness for the link quality that the user has paid for. In order to ensure fairness, users adjust their acceptable SINR ranges, according to their channel conditions and transmission histories. If a user has a smaller assigned SINR now, he will provide the system with a higher acceptable SINR range in the future, such that the system has to assign a higher SINR to him. The scheme can be conceived as "water filling" the wireless network resources to different users at different times, according to their channel conditions. Moreover, the joint consideration of the proposed scheme and beamforming can be interpreted as to combat CCI's in different DOA's and different channel conditions over time. As will be shown in the simulation results, the proposed schemes reduce up to 60% of the overall transmitted power, increase the maximal achievable SINR by up to 6dB compared with the previous work [47, 95], thus the schemes greatly increase the network performance.

System and Channel Models

Consider M co-channel links that may exist in distinct cells of multicell networks. Each link consists of a mobile user and his assigned base station. Assume coherent detection is possible so that it is sufficient to model this system by an equivalent baseband model. Each link is affected by propagation loss, shadowing fading, and multipath Rayleigh fading. For uplink, the output signal at the i^{th} base station can be expressed as:

$$x_i(t) = \sum_{m=1}^{M} \sum_{l=1}^{L} \sqrt{P_m G_{mi} \rho_{mi}} \alpha_{mi}^l g_m(t - \tau_{mi}^l) s_m(t - \tau_{mi}^l) + n_i(t) \tag{3.15}$$

where L is the maximal number of multipath, P_m is the m^{th} user's transmitted power, G_{mi} and ρ_{mi} are the path loss and the log-normal shadow fading from the m^{th} user to the i^{th} base station, respectively, α_{mi}^l is the Rayleigh fading for the l^{th} path, $g_m(t)$ is the shaping function, $s_m(t)$ is the mobile's message symbol, τ_{mi}^l is the transmission delay, and $n_i(t)$ is the thermal noise. We assume that the channels change slowly and are stable over a frame with hundreds of symbols. We also assume the multipath delay is far less than one symbol duration, i.e. $\tau_{ii}^l \approx 0,\ \forall i, l$ (the delay from the mobile user to his assigned BS), and the delay from the user to any other cell $\tau_{mi}^l, m \neq i$ is uniformly distributed in $[0, T]$, where T is the sample duration.

Define the impulse response from the m^{th} mobile user to the i^{th} base station by:

$$h_{mi} = \sum_{l=1}^{L} \alpha_{mi}^l r_{mi} \tag{3.16}$$

where r_{mi} includes the effects of transmission delay, transmitter filter, receiver filter, and shaping function. Then we can express the sampled received signal at time k as:

$$x_i(k) = \sum_{m=1}^{M} h_{mi} \sqrt{P_m G_{mi} \rho_{mi}} s_m(k) + n_i(k) \tag{3.17}$$

where $n_i(k)$ is the sampled thermal noise. The i^{th} user's SINR can be written as:

$$\Gamma_i = \frac{P_i \rho_{ii} G_{ii} |h_{ii}|^2}{\sum_{m \neq i} P_m \rho_{mi} G_{mi} |h_{mi}|^2 + N_i} \tag{3.18}$$

where $N_i = E(|n_i|^2)$.

Now we discuss the downlink cases. One issue that complicates the downlink problem is the possible lack of direct measurements of downlink channel responses at the base stations, especially for frequency-division-duplex (FDD) systems. The other issue is the lack of efficient downlink algorithms, even though the downlink channel responses are available. To obtain the optimal power control involves a complicated multi-variable optimization. In our approach, we use the virtual uplink power control technique [95], which just involves simple computations. The received signal at the m^{th} mobile receiver is given by:

$$\widetilde{y}_m(t) = \sum_{i=1}^{M}\sum_{l=1}^{L}\sqrt{\widetilde{P}_i\widetilde{G}_{im}\widetilde{\rho}_{im}}\widetilde{\alpha}_{im}^l\widetilde{g}_{im}(t-\widetilde{\tau}_{im}^l)\widetilde{s}_i(t-\widetilde{\tau}_{im}^l)+\widetilde{n}_m(t) \tag{3.19}$$

where $\widetilde{s}_i$ is the message signal transmitted from the i^{th} base station to its associated mobile user, $\widetilde{n}_m$ is the thermal noise at the m^{th} mobile user, $\widetilde{P}_i$ is the signal power, and $\widetilde{G}_{im}$, $\widetilde{\rho}_{im}$, $\widetilde{\alpha}_{im}^l$, $\widetilde{g}_{im}$, and $\widetilde{\tau}_{im}^l$ have the same definitions as those of the uplink cases. The impulse response from the i^{th} base station to the m^{th} mobile user is defined as:

$$\widetilde{h}_{im} = \sum_{l=1}^{L}\widetilde{\alpha}_{im}^l\widetilde{r}_{im} \tag{3.20}$$

where $\widetilde{r}_{im}$ includes the effects of receiver matched filter, shaping function, and transmitter filter. Then the sampled received signal vector is given by:

$$\widetilde{y}_m(k) = \sum_{i=1}^{M}\widetilde{h}_{im}\sqrt{\widetilde{P}_i\widetilde{\rho}_{im}\widetilde{G}_{im}}\widetilde{s}_i(k)+\widetilde{n}_m(k). \tag{3.21}$$

The SINR at the m^{th} mobile receiver can be expressed as:

$$\widetilde{\Gamma}_m = \frac{\widetilde{P}_m\widetilde{\rho}_{mm}\widetilde{G}_{mm}|\widetilde{h}_{mm}|^2}{\sum_{i\neq m}\widetilde{P}_i\widetilde{\rho}_{im}\widetilde{G}_{im}|\widetilde{h}_{im}|^2+\widetilde{N}_m} \tag{3.22}$$

where $\widetilde{N}_m$ is the thermal noise power at the m^{th} mobile user.

Joint Adaptive Link Quality and Power Management

In this part, we will first review the traditional power control problem and indicate the disadvantages of this kind of approaches. Then we give the reformulated problems for both uplink and downlink cases. Adaptive algorithms are developed to solve the problems.

Traditional Power Control

In the traditional uplink power control, the transmitted power of each mobile user is selected, so that each user has $\Gamma_i \geq \gamma_i$, for $i = 1, ..., M$, while the overall transmitted power used by all mobile users is minimized. Here γ_i is a fixed and predefined targeted SINR threshold to maintain the required link quality. Given that the path gains and the transmitted powers are non-negative, the matrix version of the traditional power control formulation with the fixed link quality is given by:

$$\min_{P_i} \sum_{i=1}^{M} P_i \tag{3.23}$$
$$\text{subject to } (\mathbf{I}-\mathbf{DF})\mathbf{P} \geq \mathbf{u}$$

where $\mathbf{u} = [u_1, \ldots, u_M]^T$ with $u_i = \gamma_i N_i / \rho_{ii} G_{ii}$, $\mathbf{P} = [P_1, \ldots, P_M]^T$, $\mathbf{D} = \text{diag}\{\gamma_1, \ldots, \gamma_M\}$, and

$$[\mathbf{F}]_{ij} = \begin{cases} 0 & \text{if } j = i, \\ \frac{\rho_{ji} G_{ji} |h_{ji}|^2}{\rho_{ii} G_{ii} |h_{ii}|^2} & \text{if } j \neq i. \end{cases} \tag{3.24}$$

If the spectral radius of $\mathbf{DF}$, $\rho(\mathbf{DF})$, i.e. the maximal eigenvalue of $\mathbf{DF}$, is inside the unit circle, the system has feasible solutions, i.e., there exists a positive power allocation that $\Gamma_i \geq \gamma_i$, for $i = 1, ..., M$. By Perron-Frobenius theorem [67] , the optimum power vector for this problem is $\widehat{\mathbf{P}} = (\mathbf{I}-\mathbf{DF})^{-1}\mathbf{u}$. The optimal solution of the power vector is achieved when the equations of the constraint are held, i.e. $\Gamma_i = \gamma_i, \forall i$. It has been shown that this is a NP hard problem[91]. Many adaptive algorithms [94, 43, 44, 47] have been developed to decrease the system complexity by updating the transmitted powers in a distributed manner.

In the traditional power control scheme mentioned above, each user adjusts his transmitted power to maintain the fixed and predefined SINR thresholds. When these targeted SINR thresholds are small and CCI's are minor, the system works perfectly well. However, when the targeted SINR thresholds become large, each user transmits a higher power and causes higher CCI to other users. The overall transmitted power will start to increase rapidly. If the targeted SINR thresholds are larger than some specific levels, CCI's will be so large that no feasible solutions exist, i.e. no matter how large the transmitted powers are, the receivers cannot get enough SINR levels. The reason for such a problem is that the user with the bad channel condition transmits too much power, thus introduces unnecessarily high CCI to other users. Consequently the overall system performance is reduced. Therefore, having the fixed and predefined targeted SINR thresholds constraint as the problem defined in (3.24) is not a good approach for wireless resource allocations.

Proposed Approach for Uplink

In our approach, we alleviate the constant SINR constraint by allowing users to have the time-varying SINR thresholds, according to their channel conditions. We assume the i^{th} user can accept the instantaneous SINR threshold within a range from $\gamma_i^{\min}$ to $\gamma_i^{\max}$, according to his channel condition, while the overall network link quality is kept higher than or equal to a value for adequate overall network performances. Each time, the users with bad channel conditions sacrifice their SINR's (because such levels of SINR's may not improve anything for these users), and are assigned with lower SINR thresholds. At the same time, the users with good channel conditions get higher SINR's. Consequently they have better link qualities. For each user, the time average link quality is kept as a constant to ensure fairness that the user has paid for. Each time, some users may sacrifice their performances to reduce the overall network transmitted power. These users' temporary sacrifices will be paid back in a long term. The scheme can be conceived as "water filling" wireless network resources in the time domain and to the different users, according to users' channel conditions. The user's link quality can have different definitions for different scenarios. For example, for adaptive modulation systems, the throughput and BER can be approximated by simple exponential expressions in [54, 25]. For adaptive coding systems and multimedia transmissions, the coding performance and distortion can also be approximated as functions of γ_i [66]. In our approach, we define the link quality as γ_i directly, which fits the situations such as power limited communications [43]. For the other link quality functions, the schemes described in the rest of section can be easily extended in the similar way. The matrix version of this problem formulation can be expressed as:

$$\min_{P_i,\gamma_i} \sum_{i=1}^{M} P_i \tag{3.25}$$

$$\text{subject to } \begin{cases} (\mathbf{I}-\mathbf{DF})\mathbf{P} \geq \mathbf{u}, \\ \sum_{i=1}^{M} \gamma_i \geq \psi, \\ \gamma_i^{min} \leq \gamma_i \leq \gamma_i^{max}, \\ E(\gamma_i) = \gamma_i^{ave}, \end{cases}$$

where γ_i^{ave} is the time average i^{th} link's quality, and ψ is the network overall link quality that our system needs to guarantee, which is at least as large as that of the traditional power control in (3.23) and is also the sum of time average throughput, i.e., $\psi = \sum_{i=1}^{M} \gamma_i^{ave}$.

It is worthy to emphasize that the inequality $(\mathbf{I}-\mathbf{DF})\mathbf{P} \geq \mathbf{u}$ is a bilinear matrix inequality (BMI) [68]. If we fix the powers, the targeted SINR's are linearly constrained; if we fix targeted SINR's, the powers are linearly constrained. However, if both are considered, it is a BMI problem. In the previous works [94, 47], each user's targeted SINR is the same, thus the inequality constraint is linear. While in the proposed scheme, different user can select different γ_i, so the constraint is not linear any more. A BMI problem is non-convex and can have multiple local optima.

The time-diversity fairness constraint $E(\gamma_i) = \gamma_i^{ave}$ in (3.25) involves optimiza-

Table 3.2: *Adaptive Algorithm for Moving Acceptable SINR Range*

`Initial:`	$\gamma_i^{min}(0) = \widehat{\gamma}_i^{min}$; $\gamma_i^{max}(0) = \widehat{\gamma}_i^{max}$; $\gamma_i^{mid} = \gamma_i^{ave}$.
`Iteration:`	$\gamma_i^{mid}(n) = \gamma_i^{mid}(n-1) + \beta(\gamma_i(n) - \gamma_i^{ave})$; $\gamma_i^{min}(n+1) = \min(\max(\gamma_i^{ave} - \gamma_i^{mid}(n) + \widehat{\gamma}_i^{min}, \widehat{\gamma}_i^{min}), \widehat{\gamma}_i^{max})$; $\gamma_i^{max}(n+1) = \max(\min(\widehat{\gamma}_i^{max} - \gamma_i^{mid}(n) + \gamma_i^{ave}, \widehat{\gamma}_i^{max}), \widehat{\gamma}_i^{min})$.

tions at different times. The difficulties to solve it analytically by the techniques such as dynamic programming lie in how to represent the channel models with CCI's and the computational complexity with large number of users. In our approach, we develop a moving SINR window algorithm and a projected gradient algorithm to heuristically solve (3.25). The basic idea is to first change the acceptable SINR ranges, according to the transmission histories and channel conditions, so that the fairness constraint is satisfied. Then within these SINR ranges, a projected gradient algorithm finds the allocation that produces the minimal overall transmitted power.

Instead of having fixed γ_i^{min} and γ_i^{max}, we assume that the i^{th} user can select SINR level $\gamma_i^{min}(n) \leq \gamma_i(n) \leq \gamma_i^{max}(n)$ at time n and the targeted time average SINR is γ_i^{ave}. Each time, $\gamma_i^{min}(n+1)$ and $\gamma_i^{max}(n+1)$ are modified by current $\gamma_i(n)$. When $\gamma_i(n)$ is smaller than γ_i^{ave}, $\gamma_i^{min}(n+1)$ and $\gamma_i^{max}(n+1)$ are increased, so that there is a higher probability that $\gamma_i(n+1)$ is larger than γ_i^{ave}; else $\gamma_i^{min}(n+1)$ and $\gamma_i^{max}(n+1)$ are decreased. $\gamma_i^{min}(n+1)$ and $\gamma_i^{max}(n+1)$ are bounded by $\widehat{\gamma}_i^{min}$ and $\widehat{\gamma}_i^{max}$, which are the minimal and maximal SINR's that are fixed and predefined by the system. In order to track the history of γ_i, we define

$$\gamma_i^{mid}(n) = \gamma_i^{mid}(n-1) + \beta(\gamma_i(n) - \gamma_i^{ave}), \quad 0 < \beta < 1, \tag{3.26}$$

where β is a delay sensitive factor. If a user's payload is a voice traffic and cannot suffer much delay, β should select a relatively larger number, such that the link quality will be compensated quickly. If a user's payload is a data traffic and can suffer some delay, β can select a relatively small number, so that the user can wait until the channel becomes better to transmit. Each time, $\gamma_i^{min}(n)$, $\gamma_i^{max}(n)$, and $\gamma_i^{mid}(n)$ are updated by each user in Table 1.

When $\gamma_i(n)$ is continuously less than γ_i^{ave} for some time, $\gamma_i^{min}(n)$ is increased to γ_i^{ave}. Then the next $\gamma_i(n+1) \geq \gamma_i^{ave}$, consequently, $\gamma_i^{mid}(n)$ stops increasing. The same analysis can be applied to $\gamma_i^{max}(n)$. Since $\gamma_i^{min}(n)$ and $\gamma_i^{max}(n)$ are bounded and they are linearly modified by $\gamma_i^{mid}(n)$, $\gamma_i^{mid}(n)$ is also bounded. Rearrange $\gamma_i^{mid}(n)$ in (3.26) and average over times, we have

$$\frac{\sum_{n=1}^{N} \gamma_i(n)}{N} = \gamma_i^{ave} + \frac{(\gamma_i^{mid}(N) - \gamma_i^{ave})}{\beta N}. \tag{3.27}$$

Since $\gamma_i^{mid}(N)$ is bounded, the second term on the right hand side decreases to zero as $N \to \infty$. So we prove that $\lim_{N\to\infty} \frac{\sum_{n=1}^{N} \gamma_i(n)}{N} = \gamma_i^{ave}$, i.e. the proposed algorithm guarantees fairness.

Now we can construct the adaptive algorithm to adjust each user's targeted SINR threshold to reduce the overall transmitted power. We need to find out which users cause larger CCI's and contribute more to the overall transmitted power. If these users can sacrifice their targeted SINR's a little bit, the overall transmitted power will be reduced significantly. P_{sum} can be written as $P_{sum} = \mathbf{1}^T(\mathbf{I}-\mathbf{DF})^{-1}\mathbf{u}$. From [53], we know P_{sum} is a convex and increasing function of γ_i, when the other $\gamma_j,\ j = 1 \ldots M,\ j \neq i$ are fixed. Take derivatives of γ_i of P_{sum}, then we have the

i^{th} element of gradient $\mathbf{g} = [g_1 \dots g_M]^T$ of the overall uplink transmitted power P_{sum} as:

$$\begin{aligned} g_i &= \frac{\partial P_{sum}}{\partial \gamma_i} \\ &= \mathbf{1}^T[(\mathbf{I}-\mathbf{DF})^{-1}\frac{\partial \mathbf{u}}{\partial \gamma_i} - (\mathbf{I}-\mathbf{DF})^{-1}\frac{\partial(\mathbf{I}-\mathbf{DF})}{\partial \gamma_i}(\mathbf{I}-\mathbf{DF})^{-1}\mathbf{u}] \\ &= \mathbf{1}^T(\mathbf{I}-\mathbf{DF})^{-1}[\widehat{\mathbf{D}}_i\mathbf{FP} + \hat{\mathbf{u}}_i] \end{aligned} \tag{3.28}$$

where $\widehat{\mathbf{D}}_i$ is a $M \times M$ matrix, $\hat{\mathbf{u}}_i$ is a $M \times 1$ vector, and

$$[\widehat{\mathbf{D}}_i]_{jk} = \begin{cases} 1 & \text{if } i = j = k, \\ 0 & \text{otherwise.} \end{cases}$$

$$[\hat{\mathbf{u}}_i]_j = \begin{cases} N_i/(\rho_{ii}G_{ii}|h_{ii}|^2) & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

Reorder (3.28), we have

$$g_i = \frac{c_i\left(N_i + \sum_{j\neq i} P_j\rho_{ji}G_{ji}|h_{ji}|^2\right)}{\rho_{ii}G_{ii}|h_{ii}|^2} = \frac{c_iP_i}{\Gamma_i} \tag{3.29}$$

where Γ_i is the SINR detected at the base station's antenna output for the i^{th} user, $c_i = \mathbf{1}^T(\mathbf{I}-\mathbf{DF})^{-1}\mathbf{v}_i$, and $[v_i]_j = 1$, if $j = i$; $[v_i]_j = 0$, otherwise. c_i reflects severeness of CCI's and tells which user causes more CCI to other users. When CCI's are small, $c_i \approx c_j$, $\forall i, j$. Since we only care the direction of the gradient and do not care the amplitude, we can ignore the value of c_i when CCI's are small. By using this gradient, we know how to reduce the overall transmitted power.

Since each user can have his targeted SINR threshold in a range from $\gamma_i^{min}(n)$ to $\gamma_i^{max}(n)$, $\forall\ i$, at time n, if we do not have any more constraint, every user will have $\gamma_i^{min}(n)$ as his targeted SINR threshold, so that the transmitted powers are minimized. However, the network performance is consequently degraded. So we assume the overall link quality of the network $\sum_{i=1}^{M}\gamma_i(n) \geq \psi$. Because we optimize the overall transmitted power P_{sum} which is an increasing function of γ_i [53], the optimal solution will occur when $\sum_{i=1}^{M} \gamma_i(n) = \psi$. If we change each user's targeted SINR according to (3.29), the constraint $\sum_{i=1}^{M} \gamma_i(n) = \psi$ will not hold. We have to modify the gradient by projecting the gradient onto the plane where the constraint holds. Define the modified gradient as $\mathbf{q} = [q_1 \ldots q_M]^T$. By the definition of a projection, vector $\mathbf{q}$ is the vector closest to $\mathbf{g} = [g_1 \ldots g_M]^T$ in the space $[\Omega : \sum_{i=1}^{M} \gamma_i(n) = 0]$, i.e., $\mathbf{q} = \arg\min_{\forall \mathbf{x} \in \Omega} \|\mathbf{g} - \mathbf{x}\|^2$, where $\mathbf{x}$ is a vector in Ω. Suppose $\mathbf{x} \in \Omega$. The distance between $\mathbf{q}$ and $\mathbf{x}$ is given by:

$$\|\mathbf{g} - \mathbf{x}\|^2 = \sum_{i=1}^{M-1} (x_i - g_i)^2 + (-\sum_{i=1}^{M-1} x_i - g_M)^2. \quad (3.30)$$

Take derivatives with respect to each argument, set the derivatives to zeros, write the equations in matrix form, we can get the optimal projection $\mathbf{q}$ of $\mathbf{g}$ onto Ω, where $q_m = -\sum_{i=1}^{M-1} q_i$ and

$$\begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_{M-1} \end{bmatrix} = \frac{1}{M} \begin{bmatrix} M-1 & -1 & \ldots & -1 \\ -1 & M-1 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ldots & M-1 \end{bmatrix} \begin{bmatrix} g_1 - g_M \\ g_2 - g_M \\ \vdots \\ g_{M-1} - g_M \end{bmatrix}. \quad (3.31)$$

Now we can construct an adaptive algorithm to move along the projected gradient $\mathbf{q}$ to reduce the overall transmitted power, as summarized in Table 2. We initialize the algorithm with $\gamma_1(0) = \gamma_1^{ave}, \ldots, \gamma_M(0) = \gamma_M^{ave}$. The initialization is assumed to be feasible. μ is a small constant, whose value decides the rate of convergence and the variance of the final result. The convergence criteria for the adaptive algorithm can be implemented, according to the Karush-Kuhn-Tucker(KKT) conditions [38]. For the specific problem in (3.25), the KKT conditions are given by the following theorem:

Table 3.3: *Adaptive Algorithm for Uplink*

`Initialization:`	$\gamma_1(0) = \gamma_1^{ave}, \ldots, \gamma_M(0) = \gamma_M^{ave}$.
`Adaptive Threshold Allocation`	`do` $\{$ $\mathtt{g} = \bigtriangledown P_{sum}$; $\mathtt{q} = \mathtt{projection}(\mathtt{g})$; $\gamma_i(n) = \gamma_i(n) - \mu \mathtt{q}_i \ \forall \ i$; `if` $(\gamma_i(n) > \gamma_i^{max}(n))$, $\gamma_i(n) = \gamma_i^{max}(n)$; `if` $(\gamma_i(n) < \gamma_i^{min}(n))$, $\gamma_i(n) = \gamma_i^{min}(n).\}$ `while` $(\gamma_i(n)$ `is not convergent.`$)$
`Power Update Iteration:`	$\mathtt{D} = \mathtt{diag}(\gamma_1(n), \gamma_2(n), \ldots, \gamma_M(n))$; $\mathtt{P} = \mathtt{DFP} + \mathtt{u}$.
`SINR Range Update:`	`Update` $\gamma_i^{mid}(n), \gamma_i^{min}(n)$, `and` $\gamma_i^{max}(n)$.

Theorem 3.4.1 *The convergence criteria of the proposed algorithm in Table 2 is: when γ_i hits the boundary, the projected gradient q_i will point inside the range; else $q_i = 0$, i.e.,*

$$\begin{cases} q_i \geq 0 & \textit{if } \gamma_i = \gamma_i^{min}(n), \\ q_i \leq 0 & \textit{if } \gamma_i = \gamma_i^{max}(n), \\ q_i = 0 & \textit{otherwise.} \end{cases} \tag{3.32}$$

Under such conditions, the algorithm cannot further decrease the overall transmitted power and falls into a local minimum.

Proof: Each time, we know the acceptable SINR ranges for different users and the fact that the optima occur when $\psi = \sum_{i=1}^{M} \gamma_i$. We can rewrite the optimization problem in (3.25) as:

$$\min_{\gamma_i} \ f = \mathbf{1}^T (\mathbf{I} - \mathbf{DF})^{-1} \mathbf{u} \tag{3.33}$$

$$\text{subject to} \begin{cases} h \ = \psi - \sum_{i=1}^{M} \gamma_i = 0, \\ g_1^i = \ \gamma_i - \gamma_i^{max}(n) \leq 0, \ \forall i, \\ g_2^i = \ \gamma_i^{min} - \gamma_i \leq 0, \ \forall i. \end{cases}$$

Write the Lagrange multiplier for this constrained optimization problem as:

$$L(\gamma_i, \lambda, \mu_1^i, \mu_2^i) = f + \lambda h + \sum_{i=1}^{M} u_1^i g_1^i + \sum_{i=1}^{M} u_2^i g_2^i. \tag{3.34}$$

where λ, μ_1^i, and μ_2^i, $\forall i$ are the Lagrange coefficients. Assume the local minimum occurs at $[\gamma_1^* \dots \gamma_M^*]^T$. The KKT conditions are that there exit u_1^i, u_2^i, and λ, such that the following conditions hold at γ_i^*, $\forall i$.

$$\begin{aligned} \bigtriangledown f + \bigtriangledown h^T \lambda + \textstyle\sum_{i=1}^{M} \bigtriangledown {g_1^i}^T u_1^i + \sum_{i=1}^{M} \bigtriangledown {g_2^i}^T u_2^i = \mathbf{0}&, \\ \forall i, \ {u_1^i}^T g_1^i = 0, \ {u_2^i}^T g_2^i = 0&, \\ \forall i, \ u_1^i \geq 0, \ u_2^i \geq 0&, \end{aligned} \tag{3.35}$$

where $\bigtriangledown g_1^i = 1$ and $\bigtriangledown g_2^i = -1$, $\forall i$. we have $\mathbf{q} = \bigtriangledown f + \bigtriangledown h^T \lambda$ at point γ_i^*, $\forall i$. When $\gamma_i^{max}(n) > \gamma_i^* > \gamma_i^{min}(n)$, we select $u_1^i = 0$ and $u_2^i = 0$. Under this condition, q_i needs to be zero. If $\gamma_i^{max}(n) = \gamma_i^*$, we select $u_2^i = 0$. Because $q_i \leq 0$ and $\bigtriangledown g_1^i = \mathbf{1}$, we can have $u_1^i \geq 0$. If $\gamma_i^{min}(n) = \gamma_i^*$, we select $u_1^i = 0$. Because $q_i \geq 0$ and $\bigtriangledown g_2^i = -\mathbf{1}$, we can have $u_2^i \geq 0$. So we prove that the equations in (3.32) satisfies the KKT conditions in (3.35). □

The power update step in Table 2 can be implemented in a distributed iteration manner as in [94], which only needs local channel information. In each update, the targeted SINR's are calculated at the base station, and then the powers are updated, according to the targeted SINR's in the distributed iterations [94, 43, 44, 47]. The power update equation in the algorithm in Table 2 has been proved [47] to fit the *standard function* [94]. The power update step converges to a unique solution, when the targeted SINR's are feasible. In the proposed algorithm, the targeted SINR's are selected, so that the overall transmitted power is reduced. Starting from any feasible initialization, γ_i is always within the feasible range $|\rho(\mathbf{DF})| < 1$. So the power update step converges.

Proposed Approach for Downlink

Similar to the uplink cases, we develop the proposed link quality and power management algorithm for the downlink cases. Define $\widetilde{P}_i$ as the downlink transmitted power. The optimization problem is

$$\min_{\widetilde{P}_i,\gamma_i} \sum_{i=1}^{M} \widetilde{P}_i \tag{3.36}$$

$$\text{subject to} \quad \begin{cases} (\mathbf{I} - \mathbf{D}\widetilde{\mathbf{F}})\widetilde{\mathbf{P}} \geq \tilde{\mathbf{u}}, \\ \sum_{i=1}^{M} \gamma_i \geq \psi, \\ \gamma_i^{min} \leq \gamma_i \leq \gamma_i^{max}, \\ E(\gamma_i) = \gamma_i^{ave}, \end{cases}$$

where $\widetilde{\mathbf{P}} = [\widetilde{P}_1 \ldots \widetilde{P}_M]^T$, $D = \text{diag}\{\gamma_1 \ldots \gamma_M\}$, $\tilde{\mathbf{u}} = [\tilde{u}_1 \ldots \tilde{u}_M]^T$, and

$$\tilde{u}_i = \frac{\gamma_i \widetilde{N}_i}{\tilde{\rho}_{ii}\widetilde{G}_{ii}|\tilde{h}_{ii}|^2}, \tag{3.37}$$

and

$$[\widetilde{\mathbf{F}}]_{ji} = \begin{cases} 0 & \text{if } j = i, \\ \frac{\tilde{\rho}_{ij}\widetilde{G}_{ij}|\tilde{h}_{ij}|^2}{\tilde{\rho}_{jj}\widetilde{G}_{jj}|\tilde{h}_{jj}|^2} & \text{if } j \neq i. \end{cases} \tag{3.38}$$

Similar to the uplink cases, the overall transmitted power $\widetilde{P}_{sum} = \sum_{i=1}^{M} \widetilde{P}_i$ is a convex and increasing function of γ_i, if $\gamma_j,\ j \neq i,\ j = 1 \ldots M$, is fixed [53]. By using the similar deductions of the overall transmitted power as those in the uplink cases, the m^{th} gradient element $\tilde{\mathbf{g}}_m$ of the overall downlink transmitted power is given by:

$$\tilde{g}_m = \frac{\tilde{c}_m \widetilde{P}_m}{\widetilde{\Gamma}_m} \tag{3.39}$$

where $\widetilde{\Gamma}_m$ is the SINR detected at the m^{th} mobile user and $\tilde{c}_m = \mathbf{1}^T(\mathbf{I} - \mathbf{D}\widetilde{\mathbf{F}})^{-1}\mathbf{v}_m$.

Table 3.4: *Adaptive Algorithm for Downlink*

`Initialization:`	$\gamma_1(0) = \gamma_1^{ave}, \ldots, \gamma_M(0) = \gamma_M^{ave}$.
`Adaptive Threshold Allocation`	`do` $\{$ $\widetilde{\mathtt{g}} = \bigtriangledown \widetilde{P}_{sum}$; $\widetilde{\mathtt{q}} = \mathtt{projection}(\widetilde{\mathtt{g}})$; $\gamma_i(n) = \gamma_i(n) - \mu \widetilde{\mathtt{q}}_i, \ \forall \ i$; `if` $(\gamma_i(n) > \gamma_i^{max}(n))$, $\gamma_i(n) = \gamma_i^{max}(n)$; `if` $(\gamma_i(n) < \gamma_i^{min}(n))$, $\gamma_i(n) = \gamma_i^{min}(n).\}$ `while` $(\gamma_i(n)$ `is not convergent.`$)$
`Iteration:`	
`Virtual Uplink Power Update:`	$\mathtt{P} = \mathtt{D}\widetilde{\mathtt{F}}^T \mathtt{P} + \mathtt{u}$.
`Downlink Power Update:`	$\widetilde{\mathtt{P}} = \mathtt{D}\widetilde{\mathtt{F}}\widetilde{\mathtt{P}} + \widetilde{\mathtt{u}}$.
`SINR Range Update:`	`Update`$\gamma_i^{mid}(n), \gamma_i^{min}(n),$ `and` $\gamma_i^{max}(n)$.

For the discussion of downlink in this part, we still assume the SINR as the link quality index and the overall network link quality is greater than or equal to ψ each time. We can use (3.31) to get the projection of the gradient, such that $\sum_{i=1}^{M} \gamma_i(n) = \psi$ holds. For each user, we use the same moving SINR window algorithm to ensure fairness as that for the uplink cases.

If the uplink and downlink are reciprocal, such as time divided duplex (TDD) systems, we can use uplink channel responses as downlink channel responses and construct a virtual uplink [95], whose channel responses are similar to those of the downlink. Then we find the powers and targeted SINR's at the base stations of the virtual uplink. Finally, we use the same powers and targeted SINR's for the real downlink. In order to update the transmitted power, we use the algorithm in [95]: downlink SINR is measured in each mobile user; knowing his previous transmitted power and targeted SINR, the mobile user uses a feedback channel to update the transmitted power from the base station. The algorithm is summarized in Table 3.

Joint Consideration with Beamforming

The antenna array processing techniques such as beamforming can efficiently improve the received SINR's and system performances [47, 95, 27]. The antenna arrays point their beams towards the directions of the desired signals while trying to null the CCI's. In this part, we jointly consider the proposed schemes in the previous part with beamforming, and explain why such joint schemes are superior to the traditional joint power control and beamforming schemes [47, 95].

We consider a system with antenna arrays at the base stations only. There are P elements for each antenna array. For uplink, the sampled received signal vector $\mathbf{x}_i(k)$ can be expressed as:

$$\mathbf{x}_i(k) = \sum_{m=1}^{M} \mathbf{h}_{mi} \sqrt{P_m \rho_{mi} G_{mi}} s_m(k) + \mathbf{n}_i(k) \tag{3.40}$$

where $\mathbf{h}_{mi} = [h_{mi}^1, \ldots, h_{mi}^P]^T$, $h_{mi}^p = \sum_{l=1}^{L} \alpha_{mi}^l a_{mi}^p(\theta_l) r_{mi}^p$, $a_{mi}^p(\theta_l)$ is the p^{th} antenna element response to the signal from the direction θ_l, and $\mathbf{n}_i(k)$ is the sampled thermal noise vector.

With adaptive beamforming, the output of each antenna array element is combined together with beamforming weight vector $\mathbf{w}_i$. The aim is to adjust the weight vector to achieve the maximal SINR at the output of the combiner. If the channel response from the desired user is known, the minimal variance distortion response (MVDR) solution to this problem can be used to minimize the total interferences at the output of beamformer, while the gain for the desired user is kept as a constant [34]. For uplink, the MVDR problem can be defined as:

$$\min_{\mathbf{w}_i} \|\mathbf{w}_i^H \mathbf{x}_i\|^2$$

$$\text{subject to } \|\mathbf{w}_i^H \mathbf{h}_{ii}\|^2 = 1, \; i = 1, ..., M.$$

Define the correlation matrix as $\mathbf{\Phi}_i = E[\mathbf{x}_i \mathbf{x}_i^H]$. The optimal weight vector is given by:

$$\widehat{\mathbf{w}}_i = \frac{\mathbf{\Phi}_i^{-1} \mathbf{h}_{ii}}{\mathbf{h}_{ii}^H \mathbf{\Phi}_i^{-1} \mathbf{h}_{ii}}. \tag{3.41}$$

Assuming the transmitted signals from different sources are uncorrelated and zero mean, and the additive noise is spatially and temporally white, we can write the i^{th} user's power at the beamformer output of the i^{th} base station as:

$$E[\|\mathbf{w}_i^H \mathbf{x}_i\|^2] = P_i \rho_{ii} G_{ii} + \sum_{m \neq i} P_m \rho_{mi} G_{mi} \|\mathbf{w}_i^H \mathbf{h}_{mi}\|^2 + \mathbf{w}_i^H \mathbf{N}_i \mathbf{w}_i \tag{3.42}$$

where $\mathbf{N}_i$ is the noise correlation matrix. The effective SINR at the i^{th} base station's beamformer output is given by:

$$\Gamma_i = \frac{P_i \rho_{ii} G_{ii}}{\sum_{m \neq i} P_m \rho_{mi} G_{mi} \|\mathbf{w}_i^H \mathbf{h}_{mi}\|^2 + \mathbf{w}_i^H \mathbf{N}_i \mathbf{w}_i}. \tag{3.43}$$

For the downlink case, the complexity of beamforming may increase because the calculations for the beamformer weight vectors need the knowledge of downlink channel responses for the whole network. This requires channel measurements at the mobile users and feedback mechanisms to send the information to the base stations, which will cost too much overhead and reduce the capacity. In order to calculate the downlink beamforming weight vectors, we can only measure the channel response from the base station to its assigned mobile user. We try to maximize the received power at the desired mobile user with a fixed norm downlink beamforming vector as:

$$\widetilde{\mathbf{w}}_m = \arg\max \|\widetilde{\mathbf{w}}^H \widetilde{\mathbf{h}}_{mm}\|^2 \tag{3.44}$$

$$\text{subject to } \|\widetilde{\mathbf{w}}\|^2 = 1$$

where $\widetilde{\mathbf{w}}_m$ is the downlink beamforming weight vector for the m^{th} user, and $\widetilde{\mathbf{h}}_{mm} = [\widetilde{h}_{mm}^1, \ldots, \widetilde{h}_{mm}^P]^T$ with $\widetilde{h}_{mm}^p$ as the channel response from the p^{th} antenna.

It is well known that the beamforming can effectively reduce CCI's in different DOA's. However if the desired users are almost at the same direction, the beamforming is less effective, because the beam pattern cannot distinguish the desired signals from the undesired interferences. Under this condition, some of the co-channel users will cause severe CCI's to the others. In the traditional joint power control and beamforming schemes with the fixed link quality requirement, in order for the system to operate well all the time, the worst case scenario has to be considered to choose the users' link qualities. In our proposed joint schemes,

Table 3.5: *Joint Beamforming and Proposed Resource Allocation*

`Initialization:` $\gamma_i(0) = \gamma_i^{ave}, \ \forall i.$
`Iteration:` `1.Beamforming: find the optimal` $\mathtt{w}_i$, $\forall i$ `2.Adaptive Threshold Allocation: find the targeted SINR` $\gamma_i, \ \forall i.$ `3.Power Update: update powers by` $\mathtt{w}_i$ `and` $\gamma_i, \ \forall i.$
`SINR Range Update: Update` $\gamma_i^{mid}(n), \gamma_i^{min}(n),$ `and` $\gamma_i^{max}(n), \ \forall i.$

each user's time average link quality is maintained instead. When beamforming cannot improve SINR's of some users, these users can sacrifice their temporary link qualities with the incentive that their link qualities can be compensated back, when the DOA's change better and beamformers become more effectively. Consequently, the overall transmitted power can be reduced a lot. It can be interpreted as the system to "water-fill" the users' link qualities, according to the different channel conditions as well as the different DOA's over time. Therefore, our proposed schemes have one more degree of freedom to reduce the overall transmitted power. The proposed joint beamforming and resource allocation scheme is shown in Table 4.

In the rest of this part, we will analyze a two-user example to illustrate the underlying reason for the performance improvements. Consider a network with two users and one base station. For the uplink, the two users' SINR's are given by:

$$\begin{cases} \Gamma_1 = \frac{P_1 \rho_1 G_1}{\|\mathbf{w}_1^H \mathbf{h}_2\|^2 P_2 \rho_2 G_2 + \mathbf{w}_1^H \mathbf{N}_1 \mathbf{w}_1}, \\ \Gamma_2 = \frac{P_2 \rho_2 G_2}{\|\mathbf{w}_2^H \mathbf{h}_1\|^2 P_1 \rho_1 G_1 + \mathbf{w}_2^H \mathbf{N}_2 \mathbf{w}_2}, \end{cases}$$

where P_i, ρ_i, G_i, and $\mathbf{N}_i$ are the power, shadow fading, propagation gain, and thermal noise matrix, respectively, $\mathbf{w}_i$ is the i^{th} user's beamforming weight vector, and $\mathbf{h}_i$ is the fading and array response for the i^{th} user. The overall transmitted power can be written as:

$$P_1 + P_2 = \frac{1}{1 - \Gamma_1 \Gamma_2 \|\mathbf{w}_1^H \mathbf{h}_2\|^2 \|\mathbf{w}_2^H \mathbf{h}_1\|^2} \left(\frac{\Gamma_1 \mathbf{w}_1^H \mathbf{N}_1 \mathbf{w}_1}{\rho_1 G_1} + \frac{\Gamma_2 \mathbf{w}_2^H \mathbf{N}_2 \mathbf{w}_2}{\rho_2 G_2} \right).$$

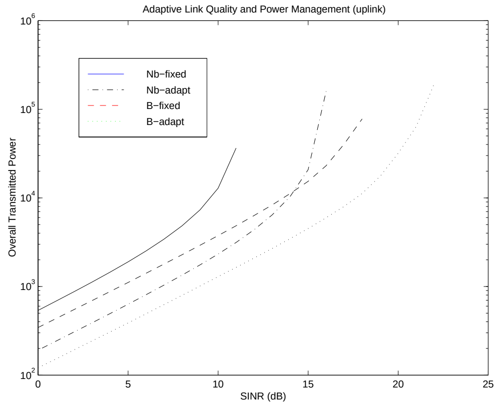
where $\|\mathbf{w}_1^H \mathbf{h}_2\|^2$ and $\|\mathbf{w}_2^H \mathbf{h}_1\|^2$ represent the effects of beamformers to suppress the interferences. In the previous joint power control and beamforming scheme [47], $\Gamma_1 = \Gamma_2$. Under this condition, in order to have a feasible solution of positive power allocation, the following condition must be satisfied at any time:

$$\Gamma_1 = \Gamma_2 < \frac{1}{\sqrt{\|\mathbf{w}_1^H \mathbf{h}_2\|^2 \|\mathbf{w}_2^H \mathbf{h}_1\|^2}}. \tag{3.45}$$

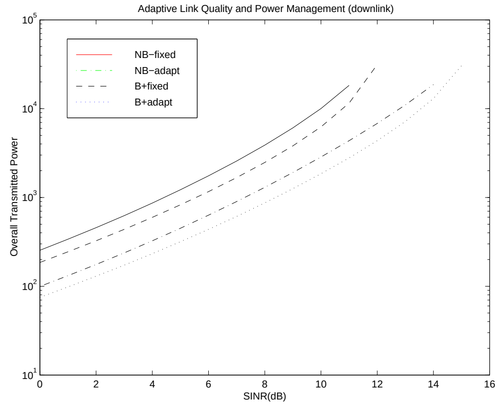
Because the channel responses $\mathbf{h}_1$ and $\mathbf{h}_2$ change randomly, the beamformers cannot be very effective for some channel responses at some time. Consequently, the system has to be designed for the worst case situation, the overall transmitted power cannot be reduced, and the maximal achievable targeted SINR is low. The underlying reason is that there is no freedom to optimize the overall transmitted power by adjusting each user's targeted SINR. In the proposed scheme, $E(\Gamma_1) = E(\Gamma_2)$ over time. When the beamformers cannot reduce the interferences well, i.e., the term $\|\mathbf{w}_1^H \mathbf{h}_2\|^2 \|\mathbf{w}_2^H \mathbf{h}_1\|^2$ is large, our proposed algorithm cleverly reduces the targeted SINR's (the value of $\Gamma_1 \Gamma_2$ will be reduced), so that the overall transmitted power is reduced. The DOA's are frequently changed by the reflections around the moving users. The algorithm waits to increase the targeted SINR's and compensate the previous losses, until the beamformers become more effectively for distinguishing the interfering users. This is the reason why the joint schemes can be used to combat CCI's more efficiently, which will be shown in the simulation results in the next part.

Simulation Results

In order to evaluate the performances of the proposed algorithms, a network with 50 hexagonal cells is simulated. The radius of each cell is 1000m. Two adjacent cells do not share the same channel. One base station is placed at the center of each cell, and one user is located randomly within the cell with the uniform distribution. The uplink and downlink work in TDD. In the simulations, we consider three multipath Rayleigh fadings with equal powers. The delay spread between different paths is far less than one symbol duration. The angle of arrival for each path is a uniform random variable in $[0, 2\pi]$. Each base station has one traditional antenna or four-element antenna arrays. $\beta = 0.1$ and $N_i = 10^{-3}$. The



(a) Uplink



(b) Downlink

Figure 3.8: Overall Transmitted Power as a Function of Average Targeted SINR

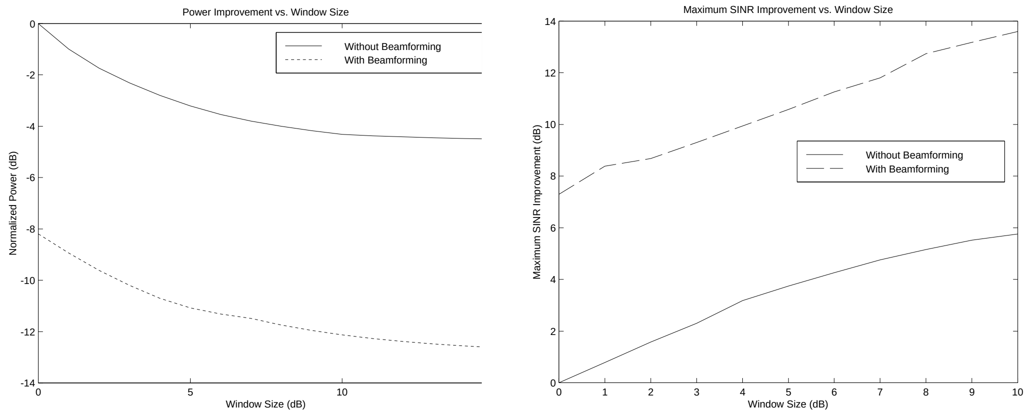
channel fading is stable within each frame and is independent between frames. We have 10000 simulation runs.

Path loss is due to the decay of the intensity of a propagating radio wave. In the simulations, we use the two slope path loss model [20] [33] to obtain the average received power as a function of distance. According to this model, the average path loss is given by:

$$G = \frac{K_0}{r^{b_1}(1 + r\lambda_c/(4h_b h_m))^{b_2}} \tag{3.46}$$

where K_0 is a constant, r is the distance between the mobile user and the base station, $b_1 = 2$ is the basic path loss exponent, $b_2 = 2$ is the additional path loss component, h_b is the base station antenna height, h_m is the mobile antenna height, and λ_c is the wavelength of the carrier frequency. We assume the mobile antenna height is 2m and the base station antenna height is 50m. The carrier frequency is 900-MHz.

In the urban microcell system, the link quality is also affected by the shadowing of the line of sight path from terrain, buildings, and trees. The shadowing is generally modelled as lognormal distribution[39]. The probability density function



(a) Power Improvement (b) SINR Improvement

Figure 3.9: Effects of Window Size

(PDF) is given by:

$$PDF(\rho) = \frac{1}{\sqrt{2\pi}\sigma\rho} exp\{-\frac{(\log\rho - \xi)^2}{2\sigma^2}\}, \rho > 0 \tag{3.47}$$

where ξ is related to the path loss, σ is the shadow standard deviation. In the simulation, for each link, 3dB log-normal shadow fading is considered.

Fig. 3.8 illustrates the overall transmitted power as a function of the average targeted SINR. Fig. 3.8 (a) shows the uplink case. We compare performances of the fixed SINR assignment algorithm [47] and those of the proposed adaptive resource management, with and without beamforming. Here we assume that each user has the same desired time average SINR threshold $\gamma_1^{ave} = \ldots = \gamma_M^{ave}$. For the SINR range, we assume $\widehat{\gamma}_i^{min} = \gamma_i^{min}(0) = \gamma_i^{ave} - \Delta\gamma$ and $\widehat{\gamma}_i^{max} = \gamma_i^{max}(0) = \gamma_i^{ave} + \Delta\gamma$, where $\Delta\gamma$ is defined as window size and $\Delta\gamma$ = 5dB. The solid curve (NB-fixed) shows the algorithm with the fixed SINR assignment and without beamforming [47]. The dash-dot curve (NB-adapt) shows the adaptive link quality and power management without beamforming. The dashed curve (B-fixed) shows the algorithm with the fixed SINR assignment and with beamforming [47]. The dot curve (B-adapt) shows the adaptive link quality and power management with beamforming. The simulation results show that, compared with the fixed SINR assignment algorithm [47], the proposed algorithms significantly reduce the overall transmitted power by 60% and extend the maximal achievable SINR by 6dB by using the adaptive link quality and power management alone. Beamforming can further reduce the overall power by 60% and the maximal achievable SINR is improved by another 7dB. Joint beamforming and our proposed algorithms can further reduce CCI's especially at the higher SINR area, where CCI's become more severe. Fig. 3.8 (b) shows the downlink case. We compare the performances of the adaptive downlink algorithm and those of the fixed SINR assignment [95], with beamforming and without beamforming. Here similar as the uplink, we select $\Delta\gamma = 5$dB. We use the simplifications mentioned in the previous part. From Fig. 3.8 (b), we can see that the adaptive SINR threshold allocation can have 60% reduction of the overall transmitted power, which in turn reduces CCI's and increases the network capacity. Furthermore, the feasible SINR areas are extended by 4dB. The beamforming can further reduce the overall power by 40%. But at the higher SINR area, because of the simplification of downlink beamforming algorithm, the advantage of beamforming is decreasing. From the simulation results, we can see that it is an efficient method to combat the time varying nature of channel and CCI's by dynamically allocating resources.

In Fig. 3.9, we show the effects of window size $\Delta\gamma$ on the performance of the proposed algorithms in uplink. In Fig. 3.9 (a), we normalize the overall transmitted power with the previous scheme [47] and compare that for window sizes. We can see that the proposed algorithm can reduce about 4dB of the overall transmitted power. When the window size increases, the speed of power reduction decreases and power stops decreasing, after window size is greater than some value. This is

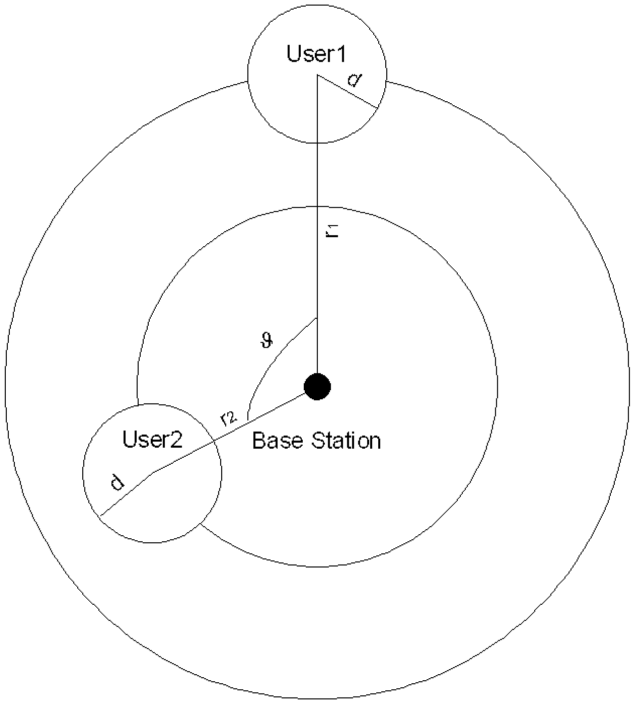
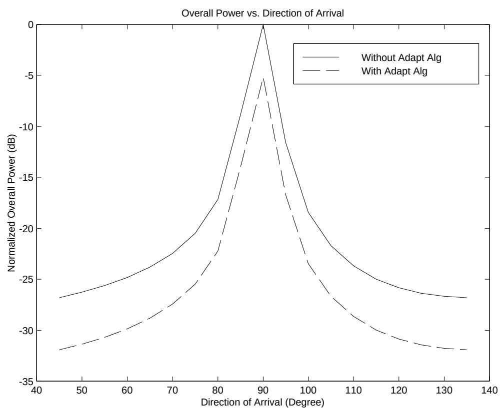
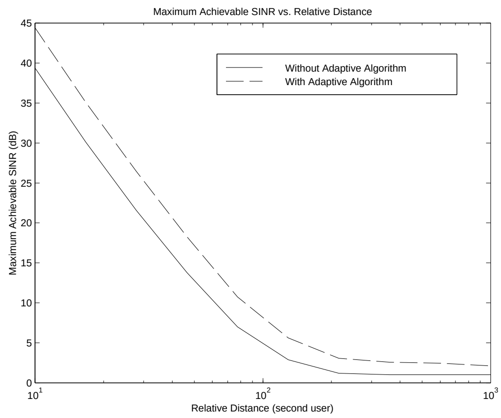


Figure 3.10: Simulation System Setup 2

because the constraint that each user's time average SINR is a constant. A user with a good channel condition now gets a higher SINR. In the future, the user has to pay back and be assigned with a lower SINR. When the proposed algorithm is combined with beamforming, the point where the overall transmitted power stops decreasing moves to a higher $\Delta\gamma$. In Fig. 3.9 (b), we compare the maximal SINR improvement vs. window size. We can see that the proposed algorithm can increase the maximal SINR by up to 6dB. The increasing speed of the maximal achievable SINR is reduced as window size increasing. Here again, joint beamforming and proposed resource allocation algorithm has a better performance.

In order to further show that joint beamforming and proposed resource allocation can combat CCI's in different DOA's and different channel conditions over





(a) Overall Power vs. DOA (b)Max Achievable SINR vs. Relat. Dist.

Figure 3.11: Performance Improvement by Joint Considering Beamforming

time, an uplink network with two mobile users and one base station is setup as shown in Fig. 3.10. The distances between the two mobile users and the base station are r_1 and r_2, respectively. The difference between two users' DOA's is ϑ. The multipath fading is modeled by Jakes model[3]. Three multipath discrete scatterers are uniformly randomly placed on a disk with radius $d = 10$m centered at each mobile user. We select $\Delta\gamma = 5$dB and $P = 4$. The other settings are the same as before.

In Fig. 3.11 (a), we compare the overall transmitted power vs. DOA. Here the first user is located at 90 degree and $r_1 = 1000$m. The second user is located in different DOA and $r_2 = 50$m. We can see that even when DOA's for the two users are almost the same (the second user is located from 85 to 95 degree), the proposed algorithm can still reduce the overall transmitted power by about 5dB. When DOA's are different, the joint beamforming and proposed resource allocation can further reduce the overall transmitted power. In Fig. 3.11 (b), we compare the maximal achievable SINR vs. the relative distance (r_2). Here the first user is located at 90 degree and $r_1 = 1000$m. The second user is located at 90 degree and

r_2 varies from 10m to 1000m. In this situation, both users suffer severe CCI's from each other's transmitted powers. The maximal achievable SINR reduces sharply with increasing of r_2. When r_2 is small, the proposed algorithm can improve the performance by 6dB compared to the fixed SINR assignment algorithm. When r_2 is almost equal to r_1, the proposed algorithm can still improve the performance by about 2dB, which is due to the constantly changing DOA's of the multipath.

3.5 Credit System, User Autonomy, and Resource Awareness

Future wireless networks will support the growing demands of heterogeneous services. Dynamic resource allocation is essential to guarantee quality of service (QoS) and enhance the network performance. We propose a novel resource allocation framework to cope with the time-varying channel conditions, co-channel interferences, and different QoS requirements in various kinds of services. We define a QoS measurement for delay sensitive applications. We introduce a credit system, where users have their autonomy to decide when and how to use their resources, and users can borrow or lend resources from the system. We also develop a simple feedback mechanism to report the system with the users' QoS satisfaction levels and channel conditions. Then the system will adapt its resource allocation strategy according to the users' feedbacks to favor the users with the bad QoS satisfaction levels or the good channels. We develop adaptive algorithms at both the user and system levels. From simulations, the proposed algorithms efficiently allocate the resources to different types of users. The users' delay constraints are satisfied and the links can survive under a long period of bad channels.

The rest of the section is organized as follows: First, we have the motivation and sketch of the proposed scheme. Then, we give the system model and MQAM modulation throughput approximation. We explain our resource allocation framework. We have numerical studies.

Motivation and Sketch

The future wireless systems are expected to provide other information services. Current wireless systems choose single-to-interference-noise ratio (SINR) as the QoS measure for voice communications. The resource allocation problem in the context of voice communications becomes power control problem [42, 43, 44, 51, 57], where the transmitted powers are constantly adjusted to achieves the users' target SINR. It has been shown in [54, 55, 40, 53, 64] that jointly considering power control and adaptive modulation can provide a variable rate and variable power ability to combat with the time varying channel and CCI. In our approach, we concentrate on the resources such as the transmitted powers and throughput of MQAM modulation.

The goal is to develop a framework of dynamic resource allocation with credit system and user autonomy for heterogeneous types of users, based on a QoS measure for delay sensitive applications. We view the problem at two levels: the macro system level and the micro user level. We also develop a feedback mechanism between the two levels. The motivations and how the framework operates are explained as follows:

1. **Micro User Level:** The goal is to let each user have the autonomy to decide when and how to use his resources according to the channel conditions and his application type. A credit system is constructed, where each user can borrow and lend resources from the system to transmit his information

during different periods of times. By doing so, his resources can be "water filled" in time during the transmission, which not only guarantees the QoS, but also ensures the survival of link during the long period of bad channel conditions.

2. **Macro System Level:** The goal is to create an environment to improve the overall network performance under the users' QoS constraints. It receives the feedbacks from the users to adapt the strategy for the environment, so that the user with the bad QoS satisfaction level or good channel condition can be allocated with more resources. Moreover, the system should encourage some users to sacrifice their performance temporarily, so that the overall network performance can be improved. These users may have the incentive to sacrifice in hope for the long-term payback.

3. **Feedback Mechanism:** The goal for feedback mechanism is to provide a simple but efficient way for each user to report his level of QoS satisfaction and channel condition, on which the system will be based to modify the optimization strategies.

System Model and Approximation

For the purpose to illustrate the idea and performance of our proposed framework, we consider a K-user uplink Direct-Sequence CDMA system in a single cell where each user is assigned with a signature sequence and an antenna array of L elements is employed at the base station (BS). For simplicity, we assume a synchronous system with processing gain H. For uplink, over one bit period, the received signal vector of the antenna array at the BS is:

$$\mathbf{y}(t) = \sum_{k=1}^{K} \sqrt{P_k G_k} b_k s_k(t) \mathbf{a}_k + \mathbf{n}_0(t) \tag{3.48}$$

where P_k, b_k, and s_k are the transmit power, bit, and signature of user k, respectively, G_k is the uplink gain from user k to the BS, the spatial signature $\mathbf{a}_k$ is the array response vector of user k, and $\mathbf{n}_0(t)$ represents the white Gaussian noise vector. We apply the chip rate filtering and sample at the chip rate. The sampled output is represented as:

$$\mathbf{Y} = \sum_{k=1}^{K} \sqrt{P_k G_k} b_k \mathbf{s}_k \mathbf{a}_k^T + \mathbf{N}_0 \tag{3.49}$$

where $\mathbf{Y}$ has the size H-by-L, whose l^{th} column represents the outputs of the l^{th} antenna element, $\mathbf{s}_k$ is the signature sequence of user k, and $\mathbf{N}_0$ represents the space and time white noise with zero mean and variance σ^2.

Suppose we apply a two-dimensional temporal-spatial linear filter $\mathbf{X}_i$ to decode the bit b_i in the MMSE sense [40]. The filter $\mathbf{X}_i$ with size H-by-L is:

$$\mathbf{X}_i = \arg \min_{\mathbf{X}_i} E[|tr(\mathbf{X}_i^H \mathbf{Y}) - b_i|^2] \tag{3.50}$$

where $tr(\cdot)$ is the trace operation. The i^{th} user's SINR at the output of the joint temporal-spatial filter is given by:

$$\Gamma_i = \frac{P_i G_i |tr(\mathbf{X}_i^H \mathbf{s}_i \mathbf{a}_i^T)|^2}{\sum_{k \neq i} P_k G_k |tr(\mathbf{X}_i^H \mathbf{s}_k \mathbf{a}_k^T)|^2 + \sigma^2 tr(\mathbf{X}_i^H \mathbf{X}_i)}. \tag{3.51}$$

Adaptive modulation provides the system with the ability to adjust the effective bit rate (throughput), according to the interference and channel conditions. MQAM is a modulation method that has high spectrum efficiency. Without loss of generality, we assume that each user has the unit bandwidth and the throughput is continuous. Let T_i denote the i^{th} user's throughput, which is the number of bits sent within each transmitted symbol. The BER can be approximated as a function of the received SINR and throughput [54, 55, 40, 65] given by:

$$BER_i \approx c_1 e^{-c_2 \frac{\Gamma_i}{2^{T_i}-1}} \tag{3.52}$$

where $c_1 \approx 0.2$ and $c_2 \approx 1.5$ for MQAM when BER is small. From (3.52), for a specific BER, the i^{th} user's throughput is

$$T_i = \log_2 \left(1 + c_3^i \Gamma_i\right) \tag{3.53}$$

where $c_3^i = -\frac{c_2^i}{ln(BER_i/c_1^i)}$.

Resource Allocation Framework

Problem Formulations

In order to implement the proposed ideas and the framework for resource allocation, we propose to formulate and solve the problems heuristically at the micro user level and the macro system level as:

1. **Micro User Level:** According to the transmission history, the users calculate user satisfaction factor (USF) for their QoS. The users' tolerance for delay will affect the value of USF. At time n, according to the USF and his current channel condition, the i^{th} user feedbacks the system with an acceptable throughput range $[T_i^{min}(n),\ T_i^{max}(n)]$. The problems include how to define USF and how to update the acceptable throughput range.

2. **Macro System Level:** The system employs adaptive algorithms to optimally assign different users their shares of resources according to their throughput ranges and other constraints such as the system feasibility [53, 64] and the maximum power. We assume perfect estimations of channel conditions. The problem is given by:

$$\max_{\gamma_i, P_i} \sum_{i=1}^{K} T_i(n) \tag{3.54}$$

$$\text{s.t.} \begin{cases} \text{Feasibility: } (\mathbf{I} - \mathbf{DF})\mathbf{P} \geq \mathbf{u}, \\ \text{Throughput: } T_i^{min}(n) \leq T_i(n) \leq T_i^{max}(n), \\ \text{Maximum Power: } P_i \leq P_{max}, \end{cases}$$

where $\mathbf{P} = [P_1 \dots P_K]^T$, γ_i is the targeted SINR such that $\Gamma_i \geq \gamma_i$, $\mathbf{u} = [u_1 \dots u_K]^T$ with $u_i = \gamma_i \sigma^2 tr(\mathbf{X}_i^H \mathbf{X}_i)/(G_i |tr(\mathbf{X}_i^H \mathbf{s}_i \mathbf{a}_i^T)|^2)$, $\mathbf{D} = diag\{\gamma_1 \dots \gamma_K\}$, and

$$[\mathbf{F}_{ij}] = \begin{cases} 0 & \text{if } j = i, \\ \frac{G_j |tr(\mathbf{X}_i^H \mathbf{s}_j \mathbf{a}_j^T)|^2}{G_i |tr(\mathbf{X}_i^H \mathbf{s}_i \mathbf{a}_i^T)|^2} & \text{if } j \neq i. \end{cases}$$

User Satisfaction Factor

In this part, we will address how to quantify the USF which shall help adjust the resource allocation strategies, i.e., the system adapts its algorithms so that the resources are more likely to be allocated to the unsatisfied users in the future. Due to the concerns on bandwidth and real-time feature, only limited feedback is allowed, therefore, the USF should be represented efficiently, for example, by a simple real value.

Suppose that data stream is transmitted in frames. Each frame has the length of M. In our approach, the USF represents whether or not a user can transmit its frame within the desired time. We define N as the the transmit time with the strictest delay constraint. The time when the frame is completely transmitted is n' and $n' \geq N$. The current time is n. For each user, a parameter α is selected when he is admitted to the network, where α depicts the tolerance of delay for this user. We assume at each time $n \geq N$, the user has probability of $1 - \alpha$ to finish the current frame. Then we can depict the probability for the total frame transmit time n' as a geometric distribution:

$$P_r(n' = N + i) = (1 - \alpha)\alpha^i, \quad i = 0, 1, \dots. \tag{3.55}$$

Different types of payloads have different delay tolerances, which are categorized as:

1. **Strict Delay Constraint:** In this case, $\alpha = 0$, $P(n' = N) = 1$, which

means the frame must be transmitted before or at time N. It fits the voice payload.

2. **Soft Delay Constraint:** Here $0 < \alpha < 1$, the estimated time to transmit the frame is $\overline{N} = N - 1 + 1/(1-\alpha)$. It fits the video/image or data payload.

3. **No Delay Constraint:** $\alpha = 1$, so $P(n' = N + i) = 0, \forall\, i \geq 0$, which means the user can suffer arbitrary transmission delay. It fits some generic data payload that is not time sensitive.

In the traditional wireless network, when a user is admitted to the system, his parameters are predefined to the system. Then the system assigns the resources to the user, according to his parameters. There is no feedback from the user to the system during the transmission to reflect whether or not the user really gets the desired QoS, even if the wireless channels may fluctuate. So we need to define USF for user's real QoS satisfaction such that the system can adapt its resource allocation scheme under different conditions. Define $T_i^{his}(n-1) = \sum_{j=1}^{n-1} T_i(j)$. We define the i^{th} user's proposed USF at time n as:

$$\mathrm{USF}_i(n) = \alpha^{\left(\frac{M(n-1)}{T_i^{his}} - N\right)}. \tag{3.56}$$

If the i^{th} user maintains the average rate $T_i^{his}/(n-1)$, the estimated time to finish the frame is $n_i^{est} = M(n-1)/T_i^{his}$. So the physical meaning of USF is the probability that the user can transmit after n_i^{est} if $n_i^{est} \geq N$. If $n_i^{est} < N$, the user is over satisfied and USF > 1. The value of USF represents the user's QoS satisfaction level and has the following implications:

1. **USF $>$ 1:** user can finish transmission even before time N and is over satisfied. He can use a lower rate to transmit during the rest of times.

2. **USF** $= 1$**:** in this case, user's QoS is exactly satisfied. If he uses the average rate $M/\overline{N}$, he can finish the frame at time $\overline{N}$.

3. $0 \leq$ **USF** < 1**:** when USF becomes smaller, the user becomes more unsatisfied and has to transmit more in the rest of times.

Credit System, User Autonomy, Resource Awareness

Similar to the economy system, we introduce concepts of credit system, user autonomy, and resource awareness to resource allocation. At a specific time, since the channel varies, the user may transmit more or less than the desired throughput. A credit system is constructed to allow lending or borrowing resources and record user's transmission history. If the user experiences a bad channel, he will be more aggressive to transmit in the future when the channel becomes better. In the proposed approach, the user will provide a higher acceptable throughput range to demand more resources. On the other hand, if the channel is still bad, he will delay requesting resources until the channel becomes better. So the user has his own autonomy to decide when and how to use the resources.

In order to optimize the users' autonomy for resource usages, the users need to know their current channel conditions, i.e., they have resource awareness. If the channels are good, users prefer to spend more resources for transmission, else they will wait until the channels become better. Suppose $\widehat{T}_i^{max}$ and 0 be the maximum and minimum allowable throughput provided by the system for the i^{th} user. To quantify the resource awareness, we define:

$$\kappa_i(n) = \frac{T_i(n-1)}{M/\overline{N}}. \tag{3.57}$$

The physical meaning of $\kappa_i(n)$ is the ratio of the most current throughput at time $n-1$ over average desired throughput, which can represent the relative channel

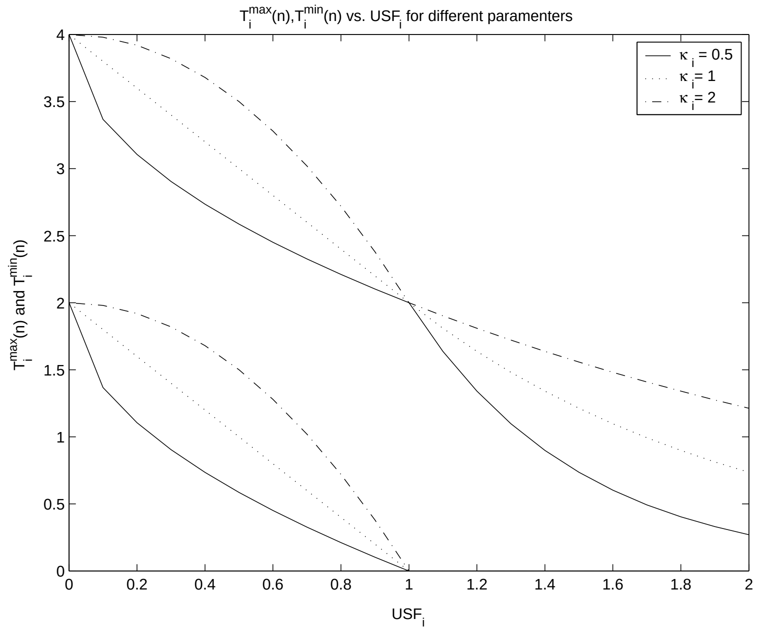


Figure 3.12: Throughput Range vs. USF for different κ_i

condition.

In the micro user level, the i^{th} user's goal is to report the system with the current acceptable throughput range $[T_i^{min}(n),\ T_i^{max}(n)]$, according to his USF and channel condition. If USF ≥ 1, there is no need for the user to transmit at the rate larger than $M/\overline{N}$. So we have $T_i^{min}(n) = 0$. We assume the USF is uniformly distributed from $[1, \infty]$ and we select exponential function for $T_i^{max}(n)$ as:

$$T_i^{max}(n) = (M/\overline{N})e^{-(\mathrm{USF}_i - 1)/\kappa_i(n)}. \tag{3.58}$$

So the average $T_i^{max}(n)$ for this USF_i is equal to the throughput $T_i(n-1)$. If $0 \leq \mathrm{USF} < 1$, we use power function to determine the throughput as:

$$T_i^{max}(n) = \widehat{T}_i^{max} - (\widehat{T}_i^{max} - M/\overline{N})(\mathrm{USF}_i)^{\kappa_i(n)} \tag{3.59}$$

$$T_i^{min}(n) = (M/\overline{N})(1 - (\mathrm{USF}_i)^{\kappa_i(n)}). \tag{3.60}$$

In Fig. 3.12, we give an example on how the throughput ranges change with different USF and channel conditions. Here $M/\overline{N} = 2$, $\widehat{T}_i^{max} = 4$, and $\kappa_i = 0.5, 1, 2$ respectively.

By jointly considering the USF and channel condition $\kappa_i(n)$, the adaptive algorithm for each user is given by:

Micro User Throughput Range Algorithm

1. Factors Calculation: calculate $\text{USF}_i(n)$, $\kappa_i(n)$.
2. Throughput Range Calculation: Update acceptable throughput range.
3. Feedback: report the range $[T_i^{min}(n)\ T_i^{max}(n)]$ to the macro system level for optimization.
4. Transmit Data: According to the rate $T_i(n)$ assigned by the system.

Adaptive Algorithm for Macro System Level

At the macro system level, the goal is to select the best throughput allocation method to different users to generate the maximum overall system throughput under the constraints. In [64], we developed a projected gradient method. In our approach, we will develop a much faster barrier method by using the idea from semi-definition programming [69].

The basic idea for the barrier method is to add barrier functions to the optimization goal such that the sum approaches negative infinity if the constraints are not satisfied. On the other hand, if the constraint is satisfied, the barrier function doesn't affect the optimization goal. The barrier function is commonly approximated by logarithmic barrier functions given by:

$$I_{constraint} \approx \Phi_1 + \Phi_2 + \Phi_3 + \Phi_4 \tag{3.61}$$

where Φ_1 is for $T_i(n) > T_i^{min}(n)$, Φ_2 is for $T_i(n) < T_i^{max}(n)$, Φ_3 is for feasibility, and Φ_4 is for P_{max}:

$$\Phi_1 = \begin{cases} \sum \ln(T_i(n) - T_{min}(n)),\ T_i(n) > T_{min}(n), \\ -\infty, \text{otherwise.} \end{cases} \tag{3.62}$$

$$\Phi_2 = \begin{cases} \sum \ln(T_{max}(n) - T_i(n)),\ T_i(n) < T_{max}(n), \\ -\infty, \text{otherwise.} \end{cases} \tag{3.63}$$

$$\Phi_3 = \begin{cases} \ln\ det\ (\mathbf{I} - \mathbf{DF}), \text{if } (\mathbf{I} - \mathbf{DF}) > 0, \\ -\infty, \text{otherwise.} \end{cases} \tag{3.64}$$

$$\Phi_4 = \begin{cases} \sum_{i=1}^{K} \ln\ (P_{max} - P_i), \text{if } P_i < P_{max}, \\ -\infty, \text{otherwise.} \end{cases} \tag{3.65}$$

The approach for barrier method is to solve the constrained optimization problem by a sequence of unconstrained problems. We rewrite (3.54) as:

$$\max_{\gamma_i, P_i}\ f = \widetilde{t} \sum_{i=1}^{K} T_i(n) + I_{constraint} \tag{3.66}$$

where $\widetilde{t}$ is a value that increases from iteration to iteration. The barrier functions become more and more like the ideal barrier function, when $\widetilde{t}$ is increasing. So the solution is more and more close to the optimal solution. Within each iteration, we use Newton method [69] to solve the unconstrained optimization problem. The algorithm is given by:

Barrier Method for Macro Throughput Maximization

1. Initial:

$\mathbf{\Gamma}$ = any feasible value, $\widetilde{t} = t_0 > 0$, $\beta > 1$, $\delta > 0$.

2. Repeat:

- Start at $\mathbf{\Gamma}$, compute $\mathbf{\Gamma}^*$ by maximizing f, using **Newton Method**:
  1. Compute Newton step $\mathbf{v}_{nt}$ and decrement λ^2.
     $\mathbf{v}_{nt} = -\bigtriangledown^2 f^{-1} \bigtriangledown f$
     $\lambda^2 = \bigtriangledown f^T \bigtriangledown^2 f^{-1} \bigtriangledown f$
  2. quit if λ^2 is stable.
  3. Line search: compute step size t' by backtracking line search.
  4. Update: $\mathbf{\Gamma}$=$\mathbf{\Gamma}-t'*\mathbf{v}_{nt}$.
- $\mathbf{\Gamma}=\Gamma^*$, calculate $\mathbf{P}$.
- if $m/\widetilde{t} < \delta$, return $\mathbf{\Gamma}$ and $\mathbf{P}$.
- $\widetilde{t} = \beta\widetilde{t}$.

where m is the iteration number for barrier method, δ determines the accuracy of the proposed algorithm, t' is the optimal step for the Newton method, t_0 is the initial value for barrier function, whose value determines the convergence rate of the first iteration, and β is the constant that $\widetilde{t}$ is multiplied in each iteration.

Simulation Results

We assume a linear array of omni directional antennas with $L = 4$ elements equispaced at half a wavelength. All $K = 80$ users are uniformly distributed within the range of $[r_0, r]$ with $r_0 = 50$m being the closest distance and $r = 1000$m being the cell radius. $H = 64$. The mobile users move in arbitrary directions with speeds uniformly distributed in the range $[0, 40]$ kph. We consider three phenomena in the propagation model: the path loss factor is 3.5 and a constant factor is chosen to yield a 30dB loss at 1m; the slow shadowing fading is modelled as a lognormal distribution with 3dB standard deviation; three paths with equal power Rayleigh fading with negligible delay spreads are considered. The fading is generated by the Jakes model with a $\pi/10$ angle spread. The update is taken every 10ms.

In Fig. 3.13 and Fig. 3.14, we show the throughput and USF for different types of users at different transmission times. Here the packet size $M = 30$, $\widehat{T}_i^{max} = 4$, $\widehat{T}_i^{min} = 0$, $\forall i$, and $N = 15$. We assume user 1 to 20 have $\alpha = 0$, user 21 to 40 have $\alpha = 0.9$, user 41 to 60 have $\alpha = 0.95$, and user 61 to 80 have $\alpha = 1$. The figure is brighter when the throughput is large and USF is large. The behaviors of different types of users are summarized as:

1. **$\alpha = 0$:** USF is equal to 0 or 1 and the transmission rate is always high because each user has to transmit his frame before the strict deadline.

2. **$0 < \alpha < 1$:** The transmission rate is determined by the channel condition. When the users with good channel conditions finish their frames early, their USF will be high so that they demand less throughput in the future.

3. **$\alpha = 1$:** USF is always equal to 1. The transmission is concentrated when the system is less busy. For example, the throughput is high around time 50 when most of users from No. 21 to 60 finish their transmissions.

From the simulation results, we can see that the proposed algorithms allocate system resources according to the service types, USF, and channel conditions.

Fig. 3.15 shows typical delay spreads for three schemes: our proposed scheme, Round Robin [41], and greedy scheduling (Traditional scheduling to maximize system throughput with $T_i(n) \in [0, \widehat{T}_i^{max}]$, $\forall i$). We order the users from the best

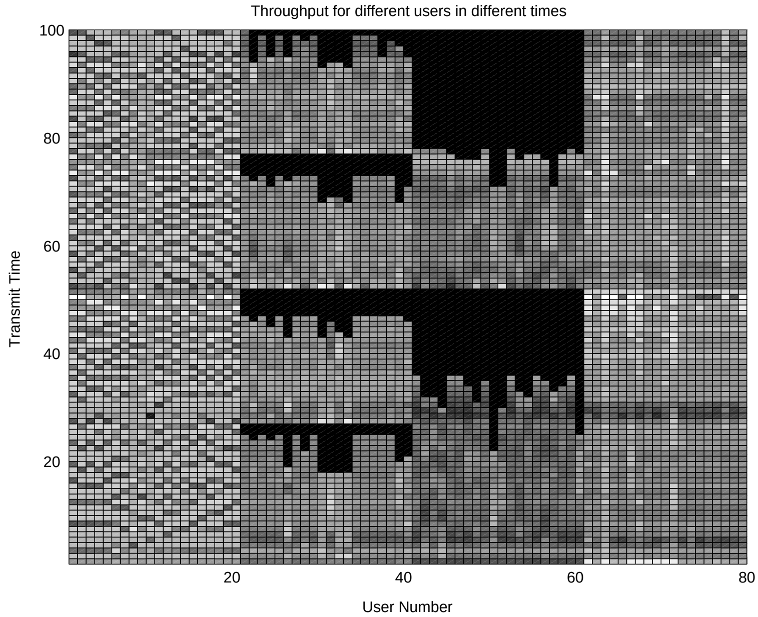


Figure 3.13: Throughput for Different Payloads vs. Transmit Times

channel to the worst. $M = 100$, $N = 55$, and $\alpha = 0.9$. Round Robin is a strict fair scheduling, but it has the poorest performance. Scheduling has the highest system throughput, but the users suffer arbitrary delays. While in the proposed scheme, the delays are more strict around the desired value.

In Fig. 3.16, we show the throughput loss for different types of services and different α. If all the users have arbitrary delay constraint $\alpha = 1$, system will have the largest average throughput and we use this value to compare with the other situations. When all the users have the same delay constraint (0%), if α is too small, the system will be infeasible. This is because the links can not survive in the long bad channel conditions. When $\alpha > 0.83$, the users can survive by borrowing from our credit system. On the other two curves, we have 25% and 50% of users with $\alpha = 1$. We can see that the strict delay will degrade the system performance most. The proposed algorithm can perform better if users' service types are more

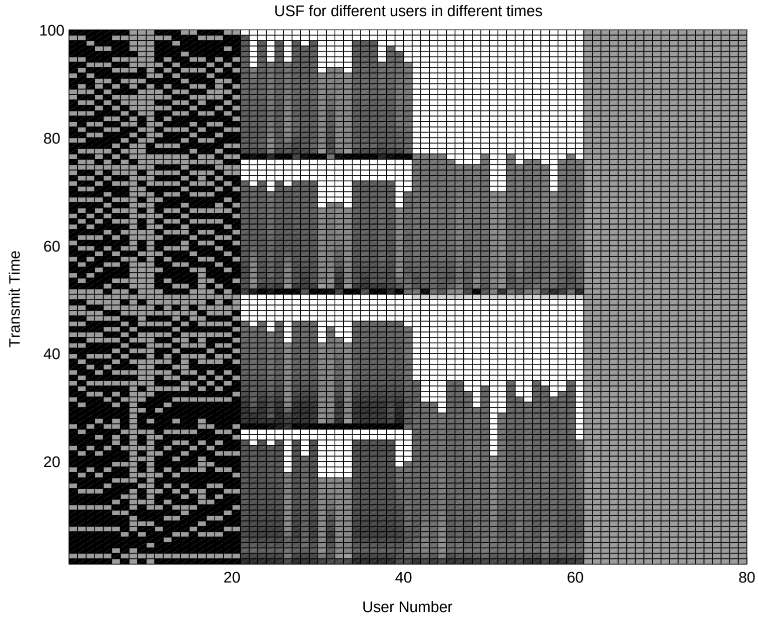


Figure 3.14: USF for Different Payloads vs. Transmit Times

diversified.

On the whole, the centralized resource allocation scheme has its advantages and disadvantages. After we present the distributed resource allocation schemes in the next chapter, we will compare the two approaches.

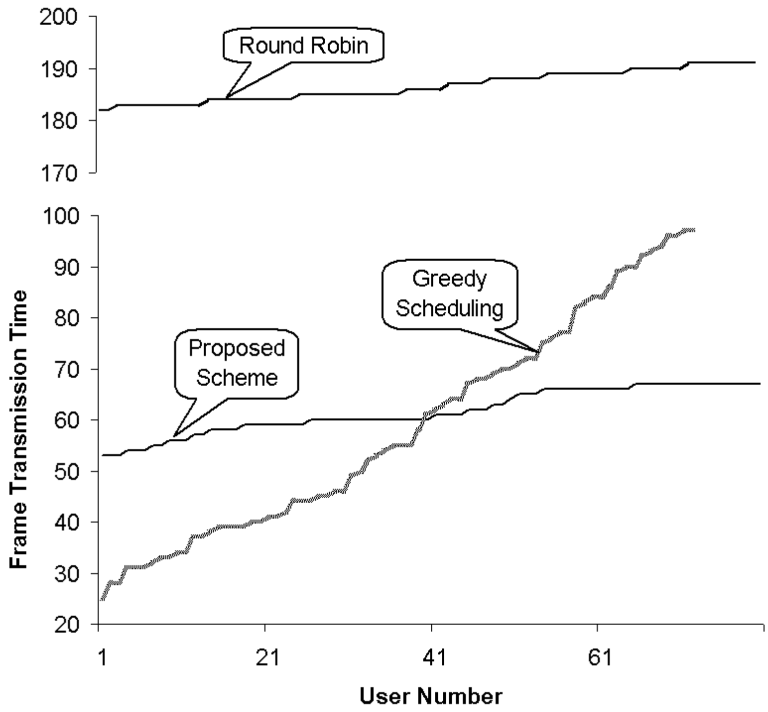


Figure 3.15: An Example of Delay Spread

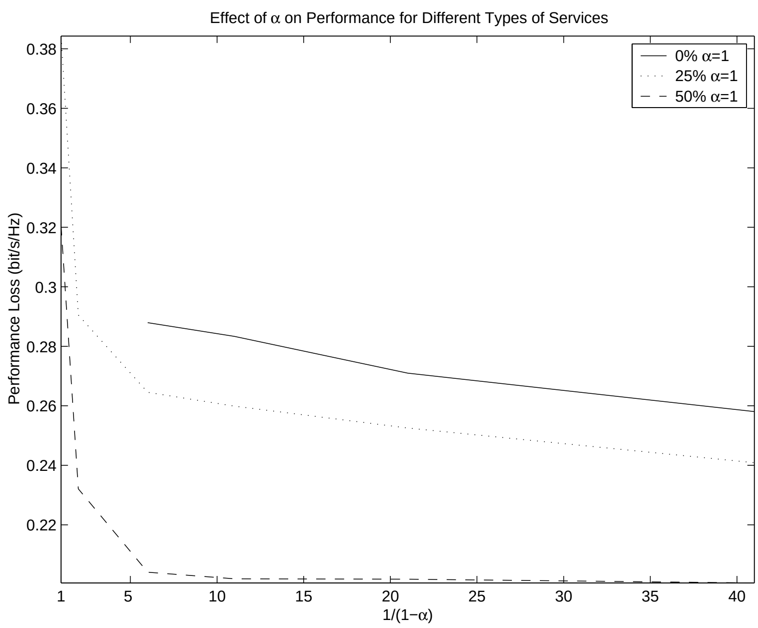


Figure 3.16: Throughput Loss For Different Types of Payloads

Chapter 4

Distributed Resource Allocation Using Game Theory

Distributed implementation of resource allocation is desired for large-scale system or multi-cell case. Game Theory is an effective mathematical tool for distributive system. In this chapter, we develop a game theory approach for wireless distributed resource allocation. In order to mediate the user's greediness and increase the system overall performance, we construct two interrelated games at the user level and system level, respectively. At the user level, each user tries to maximize his utility function in a non-cooperative power control game. The utility function has the physical meaning of throughput value minus power cost. At the system level, we develop a non-cooperative throughput game for each user to compete for the throughput. The game rule is designed to optimize the overall network throughput by controlling different users' greediness for throughput under the maximum power constraint. A simple distributed algorithm is constructed and a method is developed to initialize the proposed algorithm. An optimal centralized algorithm with high complexity is developed as a performance upper bound. From the analysis and simulation results, we show that the proposed games converge to a unique optimal Nash equilibrium at the user level and can be optimal or near optimal at the system level.

This chapter is organized as follows: In Section 4.1, we give the introduction and motivations. In Section 4.2, we present the basics for game theory. In Section 4.3, we give the system model and problem formulations. In Section 4.4, we construct two games for both the user level and the system level, respectively. Their characteristics are analyzed. The performance bound is developed to compare the performance. We have the numerical study. In Section 4.5, we compare the centralized system and distributed system.

4.1 Motivations

Over the past few decades, wireless communications and networking have witnessed an unprecedented growth, and have become pervasive much sooner than anyone could have imagined. One of the major challenges in wireless networks is to efficiently use the limited radio spectrum, which is restrained by the co-channel interference (CCI) and time varying nature of channels. CCI is caused by users' sharing of the same channel due to the multiple accesses in wireless networks. Because of the channel varying effects such as multipath fading, shadowing, path loss, propagation delay, and noise level, the signal-to-interference-noise-ratio (SINR) at a receiver output can fluctuate in the order of tens of dBs. Resource allocation such as power control and adaptive modulation is an important mean to combat these detrimental effects and increase the spectrum efficiency in the interference limited wireless networks. In power control [42, 43, 44, 45], the transmitted powers are constantly adjusted to ensure the link qualities and to combat CCI. Such a process improves the qualities of weak links. But at the same time, it increases CCI during the deep fading. Many works [46, 47, 48, 49, 22, 50, 51, 52, 53, 64] have been done to combine with other techniques such as beamforming, multi-user detection, and dynamic programming. The performances for such combined schemes are analyzed. In adaptive modulation [65] or adaptive coding [66], each link's throughput is adjusted, according to the channel conditions. The spectrum efficiency can be potentially increased. Joint consideration of power control and rate adaptation can further improve the system performance [53, 54, 55, 64]. So how to optimally perform resource allocation is an important issue we are facing today to control the interferences and enhance the performance of wireless networks.

Since individual mobile users do not have the knowledge of other users' conditions and cannot cooperate with each other, they act selfishly to maximize their own performances in a distributed fashion. Such a fact motivates us to adopt the game theory [56]. The resource allocation can be modelled as a non-cooperative game that deals largely with how rational and intelligent individuals interact with each other in an effort to achieve their own goals. In the resource allocation game, each mobile user is self-interested and trying to maximize his utility function, where the utility function represents the user's performance and controls the outcomes of the game. Many works [57, 58, 59, 60, 61, 62, 63] have been done in the power control literature. In most of the previous works, the utility function is defined as a function of power, throughput, and bit error rate (BER), which has the physical meaning of the number of information bits received successfully per Joule of energy expended.

For each user, he tries to maximize his individual interests, while the system wants to increase its efficiency, i.e. the overall system performance. Because of the users' greediness for the resources, the system may be balanced at the equilibrium with the poor overall system performance. This motives us to explore the interactions and mediations between the users' interests and the system efficiency. In the previous work, the game utility function will cause inefficiency from the system optimization point of view [57]. It is because joint power control and adaptive rate problem in multi-access networks has been shown to have non-linear and non-convex constraints [43, 54, 55] and the utility function itself is nonlinear as well. The system is probably balanced in the undesired local minima. Techniques such as pricing and repeated game have been explored to improve the efficiency. In [53, 64], it has been shown that joint power control and adaptive rate problem can be formulated to have a bilinear matrix inequality [68] constraint, if BER is fixed, i.e., the rates are linearly constrained if the powers are fixed, and the powers are linearly constrained if the rates are fixed. So this gives us motivation to design two games for the powers and the rates, respectively. In the power game, we assume the rates are fixed. Then in the rate game, the powers are assumed deterministic. The two games are interconnected, such that a higher system efficiency can be more likely to be achieved. In addition, compared with the previous games, our proposed games can guarantee BER performance and we define the system efficiency directly as the overall network throughput.

In order to achieve such a system efficiency, our primary concern is to design the utility functions and the rules of the games. One of the goals is to motive individual users to adopt a social behavior and enhance the system performance by sharing the resources. Consequently, we can make the distributed self-optimizing decisions compatible with the demand for a higher overall system performance. In this chapter, we implement the above idea, and link power control and adaptive modulation by designing games at both the user level and the system level. A non-cooperative power control game (NCPCG) is designed at the user level. We construct a user utility function that has the physical meaning of throughput value minus power cost. Each user tries to maximize his utility function, i.e., they want the desired throughput while paying less power. We will prove that the game converges to a unique optimal Nash equilibrium. At the system level, the optimization goal is to maximize the overall system throughput under the maximal transmitted power constraint. A non-cooperative throughput game (NCTG) is designed. We explain the system feasibility problem and show that there may be many Nash equilibriums in this game. A distributed algorithm is constructed by a proposed game rule and an initialization method. An optimal but complex centralized algorithm that achieves the system efficiency is developed as an upper bound to compare the performances. From the simulation results, we show that the proposed algorithms are optimal for the transmitted power at the user level, and can be optimal or near optimal for the network throughput at the system level.

4.2 Basics of Game Theory

Game theory is the study of mathematical models of conflict and cooperation between intelligent and rational decision makers. Rational means that each individual's decision-making behavior is consistent with the maximization of subjective expected utility. Intelligent means that each individual understands everything about the structure of the situation, including the fact that others are intelligent rational decision makers. In this section, we briefly review some basic concepts for game theory.

A **static game** is one in which all players make decisions (or select a strategy) simultaneously, without knowledge of the strategies that are being chosen by other players. Even though the decisions may be made at different points in time, the game is simultaneous because each player has no information about the decisions of others; thus, it is as if the decisions are made simultaneously. Simultaneous games are represented by the normal form and solved using the concept of a Nash equilibrium.

When players interact by playing a similar stage game (such as the prisoner's dilemma) numerous times, the game is called a **dynamic**, or repeated game. Unlike simultaneous games, players have at least some information about the strategies chosen on others and thus may contingent their play on past moves.

Repeated game is a special case of dynamic game. When players interact by playing a similar stage game (such as the prisoner's dilemma) numerous times, the game is called a repeated game. Unlike a game played once, a repeated game allows for a strategy to be contingent on past moves, thus allowing for reputation effects and retribution. In infinitely repeated games, trigger strategies such as tit for tat can encourage cooperation

A sequential game is **imperfect information** if a player does not know exactly what actions other players took up to that point. Technically, there exists at least one information set with more than one node. If every information set contains exactly one node, the game is one of perfect information. Intuitively, if it is my turn to move, I may not know what every other player has done up to now. Therefore, I have to infer from their likely actions and from Bayes rule which actions likely led to my current decision

A **cooperative game** is one in which players are able to make enforceable contracts. Hence, it is not defined as games in which players actually do cooperate, but as games in which any cooperation is enfoceable by an outside party (e.g., a judge, police, etc.). In termed non-cooperative games, contracts must be self-enforcing.

A market mechanism in which an object, service, or set of objects, is exchanged on the basis of bids submitted by participants. **Auctions** provide a specific set of rules that will govern the sale or purchase (procurement auction) of an object to the submitter of the most favorable bid. The specific mechanisms of the auction include first and second price auctions, and English and and Dutch auctions.

4.3 System Model and Problem Formulation

Consider K co-channel uplinks that may exist in distinct cells of wireless networks. Each link consists of a mobile and its assigned base station (BS). We assume coherent detection is possible so that it is sufficient to model this system by an equivalent baseband model. For uplink, the i^{th} user's sampled received signal at time n is given by:

$$x_i(n) = \sum_{k=1}^{K} \sqrt{P_k h_{ki}} s_k(n) + n_i(n) \tag{4.1}$$

where P_k is the k^{th} user's transmitted power, h_{ki} is the channel gain from the k^{th} user to the i^{th} BS, s_k is the transmitted symbol, and n_i is the sampled white Gaussian thermal noise. Here $h_{ki} = \alpha(L_{ki})^{-\eta}$, where L_{ki} is the distance from the k^{th} user to the i^{th} BS, α is a constant, and η is the path loss factor. We assume the average transmitted powers for different modulation constellations are normalized. Define $N_i = E(\|n_i\|^2)$. The i^{th} user's SINR is given by:

$$\Gamma_i = \frac{P_i h_{ii}}{\sum_{k \neq i} P_k h_{ki} + N_i}. \tag{4.2}$$

Adaptive modulation provides the links with the ability to match the effective bit rates (throughput), according to the interference and channel conditions. MQAM is a modulation method that has a high spectrum efficiency. Without loss of generality, we assume each user has a unit bandwidth. In [54, 65], for a desired throughput T_i of MQAM, the i^{th} user's BER can be approximated as a function of the received SINR Γ_i by:

$$\mathrm{BER}_i \approx c_1 e^{-c_2 \frac{\Gamma_i}{2^{T_i}-1}} \tag{4.3}$$

where $c_1 \approx 0.2$ and $c_2 \approx 1.5$ when BER_i is small. Rearrange (4.3), for a specific desired BER_i, the i^{th} link's required SINR for the desired throughput T_i can be expressed as:

$$\gamma_i = \frac{2^{T_i} - 1}{c_3^i} \tag{4.4}$$

where $c_3^i = -\frac{c_2^i}{\ln(\mathrm{BER}_i/c_1^i)}$. In this paper, we want to optimize the user performance and overall system throughput by jointly considering power control and adaptive modulation.

In wireless communication networks, because of the bandwidth limitation, it is impractical for the mobile users to communicate and cooperate with each other, so as to optimally utilize the wireless resources. Each individual mobile user tries to maximize his performance, based only on his perceived self-interest. All the users compete with each other for the wireless resources in a non-cooperative manner. However this will cause the system balanced in some undesired non-optimal equilibriums. Consequently, the whole system efficiency will be reduced. We need to design the game rules for the users' competitions such that the system will be balanced in the desired optimal and efficient resource allocation. This is the main goal of this paper.

Because each user controls his power to optimize his performance, the system wants to maximize the whole network throughput, and power and throughput are bi-linearly constrained, it is natural to divide the optimization efforts into the system level and the user level. We define the value function v_i as the connection between the two levels. The goals for both levels are given by:

1. **System Level:**

   The goal is to assign a user his value function v_i by a non-cooperative throughput game, such that the overall system throughput $\sum_{i=1}^{K} T_i$ is maximized, under the constraint $P_i \leq P_{max}$, $\forall i$, where P_{max} is the maximal transmitted power for each user, and v_i is related to T_i. When the system is balanced, T_i and P_i are functions of $\mathbf{v}$, where $\mathbf{v} = [v_1 \dots v_K]^T$. At the system level, the overall network throughput is optimized by the proposed NCTG, and the corresponding v_i, $\forall i$, are assigned to the users. The problem can be formulated as:

$$\max_{\mathbf{v}} \sum T_i(\mathbf{v}) \tag{4.5}$$
$$\text{subject to } P_i(\mathbf{v}) < P_{max}, \forall i.$$

2. **User Level:**

   The goal is to define a utility function u_i for each user that can describe his performance. Then each user can compete with other users in a non-cooperative power control game to maximize his utility function. There are some practical constraints such as the maximum transmitted power. The proposed NCPCG is formulated as:

$$\max_{P_i \leq P_{max}} u_i(P_i, \mathbf{P}_{-i}, v_i) \tag{4.6}$$

   where $\mathbf{P}_{-i} = [P_1 \dots P_{i-1} P_{i+1} \dots P_K]^T$, and v_i is the assigned value function that is related to throughput T_i. At the user level, the transmitted power P_i is optimized by the proposed NCPCG, while v_i is assigned from the BS.

4.4 Two Level Non-cooperative Approach

In this section, we will propose two games to accomplish the optimization goals in (4.5) and (4.6). First, we will define the utility function at the user level, construct NCPCG, and prove the convergence, uniqueness, and optimality of the equilibrium at the user level, respectively. Then, we will discuss the system feasibility problem for the system level optimization. Next we construct NCTG and a distributed algorithm at the system level. Finally, we develop a complex centralized algorithm as a performance upper bound.

User's Utility Function with Value and Cost

The idea to design the utility function at the user level is to define the concepts of value and cost. The users try to transmit specific throughput with the desired BER, which is tagged with some values v_i. These values represent what users need to pay. The higher the throughput and the lower the BER, the higher the values. The desired BER is determined by the users' service types, and the throughput is assigned from the BS by the system level optimization. The users obtain the values and achieve the desired throughput and BER by paying the costs of the transmitted powers. The costs may be high when CCI is large, i.e., the users have to have higher transmitted powers to compete with others and increase their SINR's. In addition, the transmitted powers are bounded by P_{max}. The difficulty lies in how to represent the values and costs such that the implementation can be very simple and distributed.

When the throughput T_i is equal to 0, no transmitted power is needed and the value should be zero. Otherwise, we define the value function as a function of the desired throughput as:

$$v_i = \begin{cases} \ln \frac{2^{T_i}-1}{c_3^i} + 1, & \text{if } T_i > 0; \\ 0, & \text{if } T_i = 0, \end{cases} \tag{4.7}$$

where v_i is a function of only throughput T_i and c_3^i defined in (4.4). c_3^i is related to the desired BER and is usually predefined and fixed. When the CCI is high, the cost for the desired value will increase. We represent the cost as $\ln \Gamma_i$, where Γ_i reflects the severeness of the CCI. Γ_i can be feeded back from the BS to the mobile, so the cost function can be easily implemented in a distributed manner. The users will try to get the values with less costs and the utility function is proportional to power, so we define the utility function as:

$$u_i = P_i(v_i - \ln \Gamma_i). \tag{4.8}$$

The desired utility functions should be maximized when the users pay exact costs of transmitted powers for the desired BER and throughput. When the powers are greater than necessary, the utility functions should be reduced, such that no extra transmitted powers will be wasted. If we differentiate the i^{th} utility function with its P_i and assume the interferences are fixed, we have

$$\frac{\partial u_i}{\partial P_i} = v_i - \ln \Gamma_i - 1 = 0. \tag{4.9}$$

Replace v_i by (4.7), the above equation is the same as (4.4). So the maximum of the utility function is achieved, when the minimal necessary power is applied for the desired BER and throughput T_i.

A simple two-user example for the utility function is shown in Fig. 4.1 to explain the idea and the physical meaning. Here we set the parameters as $h_{11} = 1$, $h_{21} = 0.01$, $P_{max} = P_2 = 50$, $N_1 = 0.01$, and the desired BER $= 10^{-3}$. We show u_1 as a function of P_1. For no transmission, obviously $P_1 = 0$ is the optimum. When the

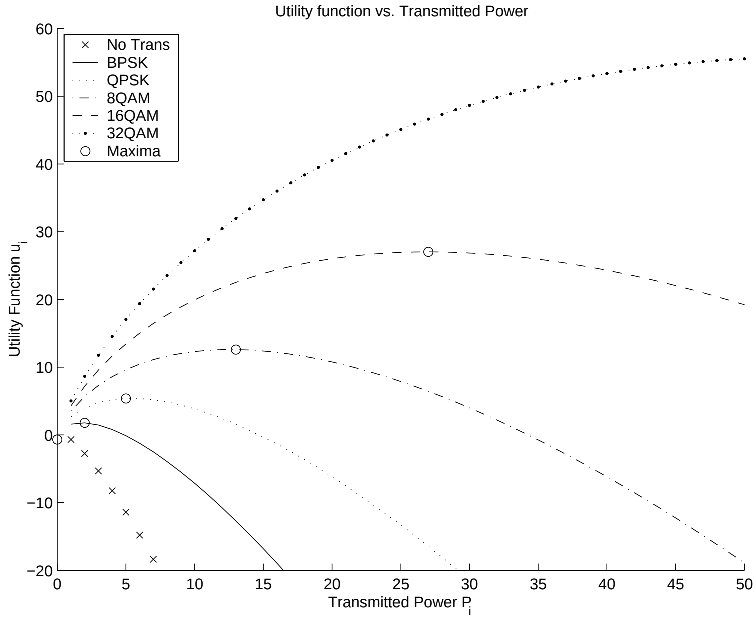


Figure 4.1: Utility vs. Transmitted Power with Fixed Interferences

adaptive modulation changes from BPSK to 16QAM, u_1 needs more transmitted power P_1 to achieve the maxima, because the value function increases. It is the user's goal to find the utility function's maxima, so that the desired BER and throughput can be satisfied. Since P_{max} is bounded, if the desired throughput is too large, for example 32QAM, even the maximum power cannot achieve the curve's maximum where (4.9) is satisfied. Under this condition, the desired BER for the throughput cannot be satisfied. It is the system's goal to assign the throughput to each user to prevent the above situation from happening.

Non-Cooperative Power Control Game at the User Level

In the wireless communication networks described in Section 4.3, a set of mobile users communicate simultaneously by sharing the same channel. Each user's performance depends on the manners in which the other users are utilizing the resources, which is a game to compete for the resources. In the proposed NCPCG,

Table 4.1: *User Level Power Control Adaptive Algorithm*

1. Obtain v_i for T_i from the BS.
2. Non-cooperative power control game: (NCPCG) $\max_{P_i \leq P_{max}} u_i(P_i, \mathtt{P}_{-i}, v_i), \ \forall i,$ given v_i.

each user aims to maximize his utility by adjusting his power without considering his interferences to other users. This will increase CCI and decrease SINR. Consequently, the power cost in the utility function in (4.8) will increase, which will prevent the users from acquiring more resources. Therefore, the system will be balanced in some equilibrium. We will analyze the characteristics of the equilibrium in the next subsection.

Several assumptions must be made for the proposed NCPCG. First, each user will strictly follow the game rule, i.e., he will not increase his power when the maximum of the utility function is achieved. Second, SINR can be accurately estimated in the BS and sent back to mobile via a reliable feedback channel without delay. Third, the utility function's maximal point is less than or equal to P_{max}, such that the desired BER and throughput can be achieved within the maximum transmitted power constraint. The first two assumptions can be easily implemented in practice, and we will construct two algorithms that can guarantee the third assumption in the later subsections.

At the user level, the value function v_i is iteratively calculated by communicating with the BS or is assigned by the BS. Then each user tries to adjust his transmitted power to maximize his utility function distributively. The adaptive algorithm for the user level is given in Table 4.4.

Characteristics of Nash Equilibrium in NCPCG

In this subsection, we will analyze the characteristics of the game equilibrium. We will prove the following theorems to show that the proposed NCPCG will converge to a unique optimal Nash equilibrium, if $\mathbf{v}$ is given.

Theorem 4.4.1 *In the NCPCG, given $\boldsymbol{v}$, there exists a Nash equilibrium: $u_i(P_i, \boldsymbol{P}_{-i}) \geq u_i(\widetilde{P}_i, \boldsymbol{P}_{-i})$, $\forall$ i, $\forall \widetilde{P}_i \leq P_{max}$, i.e. given the other users' powers, no user can improve his utility by changing his power alone.*

Proof: In [56], it has been shown that a Nash equilibrium exists, if $\forall$ i

1. Ω, the support domain of $u_i(P_i)$, $\forall i$, is a nonempty, convex, and compact subset of some Euclidean space $\Re^K$.
2. $u_i(\mathbf{P})$ is continuous in $\mathbf{P}$ and quasiconcave in P_i, where $\mathbf{P} = [P_1 \ldots P_K]^T$.

Each user can select power from any continuous real value in the close range $[0, P_{max}]$. So Ω is nonempty and compact subset of a Euclidean space $\Re^K$. For any power vectors $\mathbf{P}', \mathbf{P}'' \in \Omega$, we can easily show the convexity, i.e. $\forall \theta \in [0, 1]$, $\theta \mathbf{P}' + (1 - \theta)\mathbf{P}'' \in \Omega$. So the first condition is satisfied.

From (4.2) and (4.8), given $\mathbf{v}$, u_i is a continuous function of $\mathbf{P}$ and also a concave function of P_i that satisfies the quasiconcave condition: $\forall$ $P_i, \forall \widetilde{P}_i$, $0 \leq P_i \leq P_{max}$, $0 \leq \widetilde{P}_i \leq P_{max}$, and $\lambda \in [0, 1]$,

$$u_i(\lambda P_i + (1 - \lambda)\widetilde{P}_i, \mathbf{P}_{-i}) \geq \min(u_i(P_i, \mathbf{P}_{-i}), u_i(\widetilde{P}_i, \mathbf{P}_{-i})).$$

So the second condition is also satisfied, and there exists a Nash equilibrium for NCPCG. $\square$

Theorem 4.4.2 *Given $\boldsymbol{v}$, starting from any power allocation in Ω, NCPCG converges to a unique Nash equilibrium.*

Proof: From *Theorem 1*, there exists an equilibrium $\mathbf{P}^* = [P_1^* \dots P_K^*]^T$. From (4.2), $\mathbf{P}^*$ has to satisfy: $\mathbf{P}^* = \mathbf{\Gamma}(\mathbf{P}^*)$, where $\mathbf{\Gamma} = [\Gamma_1 \dots \Gamma_\mathrm{K}]^\mathrm{T}$. It has been shown in [42] that $\mathbf{\Gamma}$ is a *standard* function with the following properties:

1. Positivity: $\mathbf{\Gamma}(\mathbf{P}) > 0$.
2. Monotonicity: If $\mathbf{P} \geq \widehat{\mathbf{P}}$, then $\mathbf{\Gamma}(\mathbf{P}) \geq \mathbf{\Gamma}(\widehat{\mathbf{P}})$.
3. Scalability: $\forall\ \mu > 1$, $\mu\mathbf{\Gamma}(\mathbf{P}) \geq \mathbf{\Gamma}(\mu\mathbf{P})$.

Starting from any power allocation in Ω, the standard function converges to a unique fixed point $\mathbf{P} = \mathbf{\Gamma}(\mathbf{P})$, so the Nash equilibrium $\mathbf{P}^*$ is also unique. □

Theorem 4.4.3 *Given* $\boldsymbol{v}$, *the unique Nash equilibrium in NCPCG is Pareto efficient, i.e., no mobile user can ever be made happier without making at least one other mobile user less happy.*

Proof: In mathematics, for the Nash equilibrium with power vector $\mathbf{P}^*$, we want to prove that there exists no other power vector $\mathbf{P}$, such that $u_i(\mathbf{P}) \geq u_i(\mathbf{P}^*),\ \forall i$ and $\exists i,\ u_i(\mathbf{P}) > u_i(\mathbf{P}^*)$.

When the system is balanced, if the maximum of curve u_i as the function of P_i is smaller than or equal to P_{max}, the user will select P_i that maximize u_i, else because u_i is a concave function, u_i will be an increasing function on $[0, P_{max}]$. So P_{max} maximizes u_i and the user will select P_{max} as his transmitted power. The elements of Nash equilibrium $\mathbf{P}^*$ satisfy

$$\frac{\partial u_i}{\partial P_i^*} = 0 \text{ or } P_i^* = P_{max},\ \forall i. \tag{4.10}$$

In both cases, P_i^* maximizes u_i, $\forall i$, so there doesn't exist any $\mathbf{P} \in \Omega$, such that $u_i(\mathbf{P}) > u_i(\mathbf{P}^*), \exists i$, i.e., the Nash equilibrium $\mathbf{P}^*$ is Pareto efficient. □

System Feasibility

As we have mentioned in subsection 4.4, if the targeted throughput (i.e. v_i) is too large, the user cannot achieve the utility maximum, even by transmitting the maximum power. Under this condition, the user's desired BER cannot be satisfied, and we call the system not being feasible.

In order to prevent the system from not being feasible, we need to analyze the feasibility condition. First, we use the targeted SINR γ_i in (4.4) and require that the received SINR Γ_i be larger than or equal to this targeted SINR, i.e., $\Gamma_i \geq \gamma_i, \forall\ i$, in order to ensure the desired BER for the throughput T_i. Rewrite these inequalities in a matrix form, we have

$$(\mathbf{I} - \mathbf{DF})\mathbf{P} \geq \mathbf{Du} \tag{4.11}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{u} = [u_1, \ldots, u_K]^T$ with $u_i = N_i/(G_{ii}|h_{ii}|^2)$, $\mathbf{D} = diag\{\gamma_1, \ldots, \gamma_K\}$, and

$$[\mathbf{F}_{ij}] = \begin{cases} 0 & \text{if } j = i; \\ \frac{G_{ji}|h_{ji}|^2}{G_{ii}|h_{ii}|^2} & \text{if } j \neq i. \end{cases}$$

The inequality in (4.11) is a bilinear matrix inequality [68], i.e., the power vector is linearly constrained if the targeted SINR vector is fixed, and vice versa. By Perron-Frobenius theorem [67], there exists a positive power allocation if and only if the maximum eigenvalue of $\mathbf{DF}$, i.e., spectrum radius $\rho(\mathbf{DF})$, is inside unit circle. When $|\rho(\mathbf{DF})| < 1$, the optimal power solution is

$$\mathbf{P}' = (\mathbf{I} - \mathbf{DF})^{-1}\mathbf{Du}. \tag{4.12}$$

However, there is a probability that some elements in $\mathbf{P}'$ are larger than P_{max}. So the spectrum radius constraint is only a necessary condition for system feasibility. But it is usually used in literature [47] as an initial point. First, γ_i is selected to satisfy the spectrum radius constraint. The power is updated by SINR balancing [42, 43, 44]. If the targeted SINR γ_i cannot be achieved even by P_{max}, the corresponding γ_i must be reduced, until both the spectrum radius constraint and the maximum power constraint are satisfied, then the system is feasible.

To select targeted SINR and throughput under the above conditions is too complex in practice, because channel response matrix $\mathbf{F}$ needs to be estimated and the complexity to compute eigenvalues is high. Here for the adaptive MQAM, we provide a simple solution to guarantee the feasibility by the following theorem, which can be used as an initialization rule for the NCTG that we will propose in the next subsection.

Theorem 4.4.4 *Define the maximal achievable SINR as $\widehat{\Gamma}_i$ when $P_i = P_{max}$, $\forall\ i$. Then the value function*

$$v_i = \ln \frac{2^{\lfloor \log_2 \left(1 + c_3^i \widehat{\Gamma}_i\right) \rfloor} - 1}{c_3^i} + 1 \tag{4.13}$$

is always feasible, where $\lfloor\rfloor$ is the floor function to find the maximal integer smaller.

Proof: Define $\widehat{\gamma}_i = e^{v_i - 1} = (2^{\lfloor \log_2 \left(1 + c_3^i \widehat{\Gamma}_i\right) \rfloor} - 1)/c_3^i$. Since $\log_2$ is an increasing function, $\widehat{\Gamma}_i \geq \widehat{\gamma}_i, \forall\ i$. Because $\mathbf{D}, \mathbf{F} \in \Re^{K \times K}$ and $|\rho(\mathbf{DF})| < 1$, we can rewrite (4.12) as:

$$\mathbf{P}' = \sum_{j=0}^{\infty} (\mathbf{DF})^j \mathbf{Du}. \tag{4.14}$$

Since any component in $\mathbf{D}$, $\mathbf{F}$, and $\mathbf{u}$ is nonnegative. So all the components in $\mathbf{P}'$ are nondecreasing functions of γ_i. When we select the targeted SINR $\widehat{\gamma}_i \leq \widehat{\Gamma}_i$, any component of the power vector must be smaller than or equal to P_{max}. So we prove that the value functions in (4.13) satisfy the maximum power constraint. Consequently, the system must be feasible. □

Non-Cooperative Throughput Game at the System Level

At the system level, we will construct a non-cooperative throughput game for the users to compete with each other for the throughput distributively, while we want to make sure that the system is feasible. However we will show that the system may be balanced in multiple equilibriums. The goal is to optimize the overall system performance. In this subsection, we will discuss how to design the game rule for the users' NCTG to maximize the overall throughput, while the system feasibility is maintained. Then a simple distributed algorithm for throughput allocation is developed.

Since we want to prevent system from being not feasible, we define Λ as an indication function for system feasibility, which can be easily implemented in the BS. When the BS detects that all the required transmitted powers for the desired BER and throughput are less than or equal to P_{max}, Λ equals to 1, else it equals to 0, i.e.,

$$\Lambda = \begin{cases} 1, & \text{if } P_j \leq P_{max}, \ \text{BER}_j \text{ and } T_j \text{ are satisfied, } \forall \ j; \\ 0, & \text{otherwise.} \end{cases} \tag{4.15}$$

Since the users compete with each other for the throughput, we define each user's utility function $\overline{u}_i$ for NCTG as a product of his throughput T_i and Λ, i.e., the user's payoff will be zero, if his greediness for throughput will make the system infeasible. The game starts from any feasible initial values and is balanced when no user can increase his throughput. The proposed NCTG for each user can be expressed as:

$$\textbf{(NCTG)} \ \max_{T_i} \ \overline{u}_i = T_i \Lambda. \tag{4.16}$$

As we will show in the simulation results, there might be many Nash equilibriums. If the users with bad channels get high throughput, they will produce large CCI to other users. Consequently, the system overall throughput will be reduced.

So how to initialize the proposed game and how to design the game rule for each user to compete his throughput play a critical role on finding the global optimum. The idea comes from *Theorem 4*. We first let every user transmit the maximal power. The BS detects the received SINR, and decides what is the largest achievable throughput T_i by this received SINR to ensure the BER. Then the BS sends the corresponding value functions v_i, according to (4.13) back to the mobiles. The system is sure to be feasible but not necessarily optimal. By doing this, the users with good channels will get higher throughput.

After the users' powers are balanced in the desired values by NCPCG, the users decide if they can increase their throughput, while the system is still feasible. We need to find the criteria for the users when to send requests for throughput increasing. From (4.4), we define $\gamma_i(T_i)$ as the required SINR for the desired throughput T_i. When $T_i > 0$, if we assume the interferences, noise, and channel gains are fixed, from (4.2) and (4.4), the required power for throughput $T_i + 1$ will be $P_i(2^{(T_i+1)} - 1)/(2^{T_i} - 1)$, where P_i is the current power. We compare this desired power with βP_{max}, where β is a constant and $0 \leq \beta \leq 1$. If the result is larger, the user can send a request to the BS to increase his throughput by one. Here because the interferences from other users will increase when this user increases his power, β is a factor that takes into consideration of this effect. In the simulation results, we will show the effects of β on the performances. When $T_i = 0$, all the received powers are the interferences plus noise power defined as I_i. We can also estimate the channel gain h_{ii} during the initialization when this user transmits the maximal power. This user can calculate his estimated received SINR by transmitting power βP_{max}. If the value is larger than $\gamma_i(1)$, the user will send the throughput increase request. On the whole, define H_i as the throughput request factor, the criteria for the users to request their throughput to increase by one is $H_i > 1$, where

$$H_i = \begin{cases} \frac{\beta P_{max}(2^{T_i}-1)}{P_i(2^{(T_i+1)}-1)} & \text{if } T_i > 0; \\ \frac{\beta P_{max} h_{ii}}{I_i \gamma_i(1)} & \text{if } T_i = 0. \end{cases} \tag{4.17}$$

The algorithm is initialized by sending P_{max} from all users. The throughput and value functions are calculated by (4.13) and sent back to mobile users. Then after NCPCG converges at the user level, the users decide whether to request the BS to increase their throughput by the conditions in (4.17). The BS tries to increase the targeted throughput and value functions correspondingly. If the BS detects the system is feasible and there is no more request, the whole algorithm waits until the next update request. Otherwise, the BS goes back to the original value and refuses to increase the throughput for these users. If more than one users request at the same time, the BS prefers the user with the highest H_i, rejects all the other requests, and increases the throughput one by one. In the next time slot, the users will decides again if they will send the requests, until no more request is sent and the system is stable. The above game rule for NCTG and the distributed adaptive algorithm are summarized in Table 4.4.

Centralized System as a Performance Bound

The distributed algorithm in Table 2 may not be optimal. The first reason is that there is a probability that the users don't send requests, while the system might be feasible if they send. The second reason is the existence of Nash equilibriums that are not global optima. In order to understand the performance loss, we need to find the optimal solution as a performance upper bound, which may be too complex to be implement. The most straightforward idea is to let the system centrally decide how to allocate the throughput to the users with the assumption that all the channel responses are known. The problem becomes a constrained

Table 4.2: *Distributed System Algorithm*

`Initial:` $P_i = P_{max}$, $\forall\ i$, `calculate v by (4.13)`   `and send back to mobiles.`
`Iterations: When NCPCG converges`

`1. Power Increase Criteria at Users:`
 `If conditions in (4.17) are satisfied, send`
 `throughput increase request to the BS.`

`2. Feasibility Detection at the BS:`
 `Increase throughput for the requesting user`
 `with highest` H_i`, detect if still feasible.`

`3. Feedback to Users:`
 `If the system is not feasible:`
 `reduce throughput to original value.`
 `else if no more request: Wait.`

optimization problem: to maximize the overall throughput under the maximum power and maximum eigenvalue constraints, which can be written as:

$$\max_{T_i} \sum_{i=1}^{K} T_i(\mathbf{P}) \tag{4.18}$$
$$\text{s.t. } |\rho(\mathbf{DF})| < 1,\ P_i \le P_{max},\ \forall\ i.$$

Many adaptive algorithms are available in [64, 68, 69, 70]. Here we use the barrier method to solve the constrained problem in (4.18). The basic idea for the barrier method is to add barrier functions to the optimization goal such that the sum approaches negative infinity, if the constraints are not satisfied. On the other hand, if the constraint is satisfied, the barrier function doesn't affect the optimization goal. So we can solve the constrained optimization problem by a sequence of unconstrained optimization problems. The barrier function is commonly approximated by logarithmic barrier functions. We choose

$$I_{barrier} \approx \Phi_1 + \Phi_2, \tag{4.19}$$

$$\Phi_1 = \begin{cases} \ln\ det\ (\mathbf{I} - \mathbf{DF}), & \text{if } (\mathbf{I} - \mathbf{DF}) > 0; \\ -\infty, & \text{otherwise,} \end{cases} \tag{4.20}$$

$$\Phi_2 = \begin{cases} \sum_{i=1}^{K} \ln\ (P_{max} - P_i), & \text{if } P_i < P_{max},\ \forall i; \\ -\infty, & \text{otherwise,} \end{cases} \tag{4.21}$$

where Φ_1 is for the spectrum radius constraint, and Φ_2 is for the maximal power constraint.

The barrier method approach is to add the course barrier function first and solve the unconstrained problem, then in the next iteration the barrier function is refined and we solve the new unconstrained problem starting from the previous results. This iteration stops when the barrier function is close enough for the ideal

Table 4.3: *Centralized System Algorithm*

1.	Network Throughput Maximization: Iteration: • Maximize f and calculate P. • If $m/\widetilde{t} < \delta$, return T_i and P_i, $\forall i$. • $\widetilde{t} = \epsilon\widetilde{t}$ and $m = m + 1$.
2.	Calculate and Feedback v_i to Users.

constraints and the accuracy of the solutions is satisfied. We rewrite (4.18) as:

$$\max_{T_i} \ f = \widetilde{t}\sum_{i=1}^{K} T_i + I_{barrier} \tag{4.22}$$

where $\widetilde{t}$ is a value that is multiplied by ϵ in each iteration, where $\epsilon > 1$. The barrier function becomes more and more similar to the ideal barrier function, when $\widetilde{t}$ is increasing. So the solution is more and more closer to the optimal solution. Within each iteration, we can use any standard nonlinear optimization method [70] to solve the unconstrained optimization problem. Define m as the iteration number and δ as the accuracy of the proposed algorithm. The centralized algorithm is given in Table 3. Because the problem defined in (4.18) is non-linear and non-convex, there exist many local maxima. The multiple initializations are necessary to find the global optimum.

We call this kind of algorithms centralized algorithms, because the system fully decides how much throughput each user can have each time. The advantages of the centralized algorithms are obvious. First, the global optimum can be guaranteed. Second, the optimal resource allocations are assigned without the iterative feedbacks from the mobiles, so that it can fit the situation where the channels change fast. The disadvantages of the centralized algorithms are the complexity and the

Table 4.4: *Strategic Form for Two Users NCTG Example*

5	**0**	**0**	**0**	**0**	**0**
4	$\underline{\mathbf{5}}$	**0**	**0**	**0**	**0**
3	**4**	$\underline{\mathbf{5}}$	**0**	**0**	**0**
2	**3**	4	**5**	$\underline{6}$	**0**
1	**2**	**3**	**4**	**5**	**0**
(u_1, u_2)	1	2	3	4	5

difficulty to estimate the channel response matrix **F**. So this kind of algorithms only fit the situation like a CDMA system in a single micro-cell, where the user number is small and matrix **F** is directly available from the channel estimations. For the other situations such as large number of users or multi-cell situation, the centralized algorithms can be used as a performance upper bound employed later in the simulations, and the proposed distributed games can be implemented with low cost and comparative performance.

Simulation Results

We evaluate the performances of the proposed algorithms by two simulation setups. First, we consider a two-user case. Here we assume $h_{11} = h_{22} = 1$, $h_{21} = 0.01$, $h_{12} = 0.07$, $N_1 = N_2 = 1$, BER $= 10^{-3}$, and $P_{max} = 100$. In Fig. 4.2, we show the Nash equilibriums of NCPCG when the different throughput allocations are given. On any solid line, u_1 gets the maxima. On any dotted line, u_2 has the optima. Starting from any feasible power allocation, each user tries to maximize his utility function by controlling his power, such that the power allocation is closer to the corresponding lines. When the system is balanced, any intersection is a Nash equilibrium, where we denote the throughput as (user1's

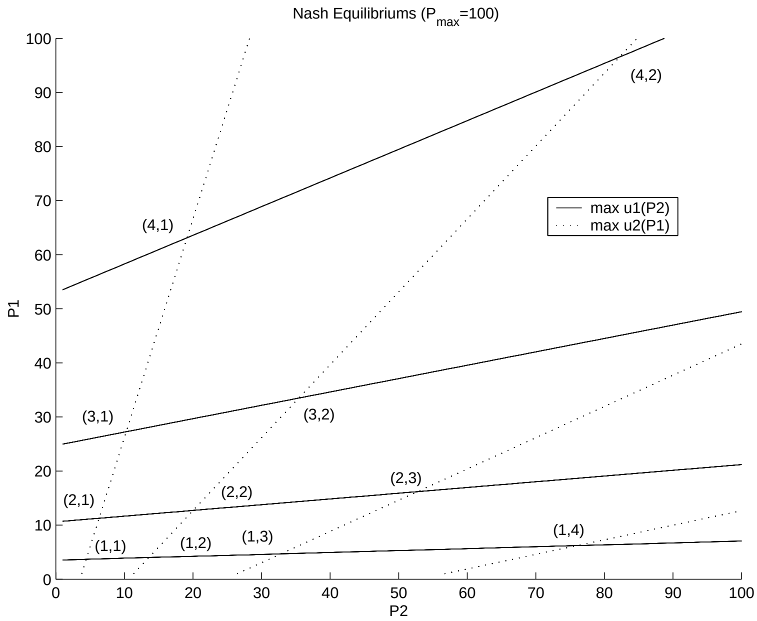


Figure 4.2: Nash Equilibriums of NCPCG

throughput, user2's throughput). We can see that the maxima for u_1 obtained from P_1 will increase with increasing of P_2. This is because the CCI increases. In Table 4, we list the strategic form of NCTG at the system level for all the nonzero throughput allocations. Each row lists user1's throughput and each column lists user2's throughput. The bold numbers are the overall throughput. If the system is not feasible, the overall throughput is 0. We can see that (4,2), (2,3), and (1,4) are Nash equilibriums, because no user can improve his throughput alone. However (2,3) and (1,4) are not desired Nash equilibriums for the optimal overall network throughput. The proposed distributed algorithm in Table 2 will be initialized at (3,2). If β is properly selected (in this case, $\beta > 0.32$), because $H_1 > H_2$ in (4.17), the algorithm will increase user1's throughput first and converge to the optimal Nash equilibrium (4,2). So we can see that we can achieve both power optimum and throughput optimum by playing the NCPCG at the user level and NCTG at

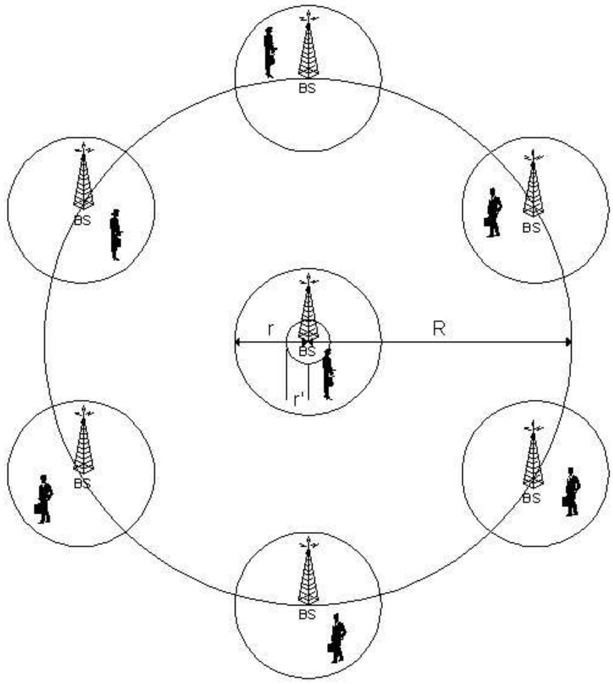


Figure 4.3: Simulation Setup II

the system level.

We setup another simulation to test the proposed algorithms. In Fig. 4.3, a network with 7 round cells are constructed. One cell is at the center and the other six are located at the degrees of $[0, 60, 120, 180, 240, 300]$. One BS is located at the center of each cell and one user is randomly located within each cell. The cell radius is $r = 1000$m, the minimal distance between the user and the BS is $r' = 50$m, and the distance between centers of two adjacent cells is $R = rR_u$m, where R_u is the reuse distance factor. We assume $P_{max} = 2$Watts, $\eta = 3.5$, $\alpha = 10^{-3}$, BER $= 10^{-3}$, and $N_i = 10^{-11}$Watts. We run the simulation 10^5 times.

In Fig. 4.4, we compare the system efficiencies (overall network throughput) vs. R_u for different β and optimal solution. When $\beta = 0$, the users don't send any requests for throughput increasing. On the other hand, when $\beta = 1$, the users are

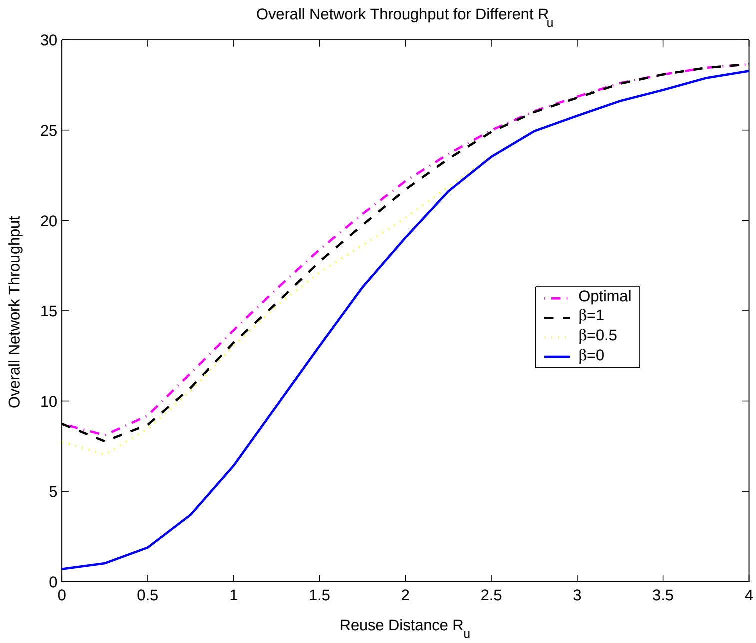


Figure 4.4: System Efficiency

most aggressive for throughput increasing requests.

- When R_u is small, CCI is severe. After initialized by sending the maximal transmitted power, most users get zero throughput. The overall network throughput is increased by the users' throughput increasing requests. When β is large enough, the proposed games can achieve the system efficiency when $R_u = 0$.

- When R_u is large, CCI is minor. After the initialization, most users get the desired throughput. The overall network throughput is refined when β is large. System efficiency can be achieved when R_u and β are large enough.

- When R_u is in the middle, the proposed games may fall into the local minima and produce the sub-optimal solutions, even when $\beta = 1$. The overall throughput will be improved by increasing β.

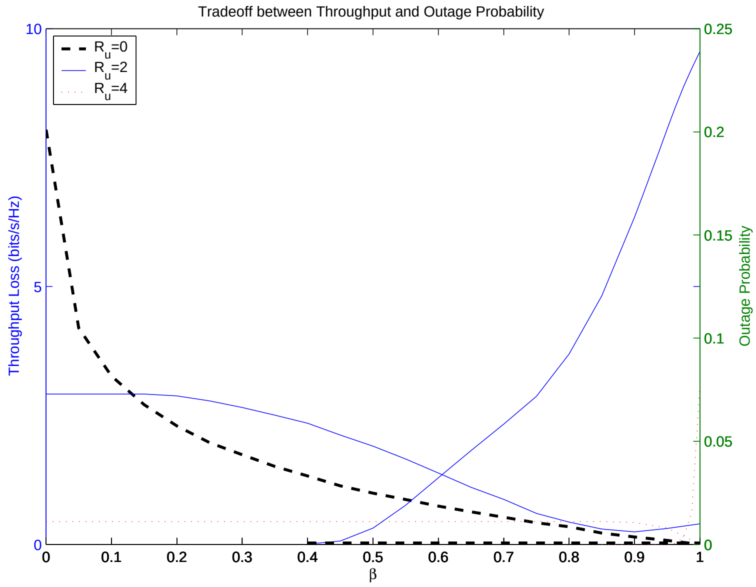


Figure 4.5: Tradeoff between Outage and Throughput Loss for β

The overall network throughput is minimal when $R_u \approx 0.25$, because different BS's and users are mixed together and CCI are most severe under this condition.

However, the overall throughput improvement by increasing β is under the cost of possible high outage probability, where the outage probability is defined as: the ratio of the number of turned down requests over the total number of requests. In Fig. 4.5, we show the throughput loss compared with the optimal solution and outage probability vs. β for different R_u.

- When $R_u = 0$, the outage probability is always zero and the throughput loss is monotonically decreasing with β. This is because the optimal solution is that only the user with the best channel condition transmits and there is no CCI from other users. So there is no penalty from other users if the transmitting user increases β and aggressively sends the request. It is optimal to select $\beta = 1$.

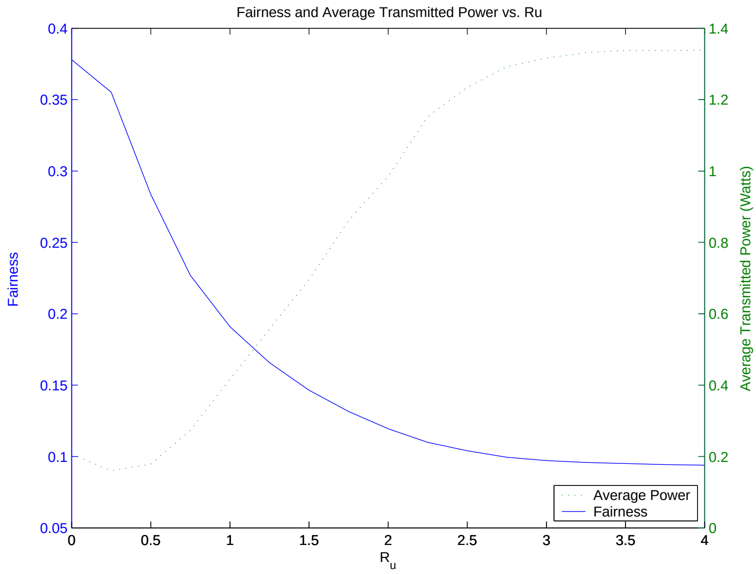


Figure 4.6: Fairness and Average Transmitted Power vs. Reuse Distance

- When $R_u = 2$, the outage probability monotonically increases with β. There is a tradeoff between the throughput loss and outage probability. The higher β, the lower throughput loss, but the higher outage probability. If the system wants a very low outage probability, we can select $\beta = 0.4$ with a performance loss of 2.35 bit/s/Hz.

- When $R_u = 4$, the outage probability is almost zero when $\beta < 0.97$, and the overall throughput loss is approximately 0.44 bit/s/Hz when $\beta < 0.9$. The tradeoff only occurs when β is large. The reason is that the users get almost optimal throughput after initialization. Consequently, the refinement only happens when the users are more aggressive for the requests.

In order to show how CCI influences the game results, we define the fairness factor as:

$$\varrho = \frac{1}{\bar{T}} \sqrt{\frac{1}{K-1} \sum_{i=1}^{K} (T_i/\widehat{T}_i - \bar{T})^2} \tag{4.23}$$

where $\widehat{T}_i$ is the maximal throughput, if only the i^{th} user is transmitting, and $\bar{T}_i = \text{mean}(T_i/\widehat{T}_i)$. The physical meaning of ϱ is the normalized variance of users' throughput compared with that of the single user case. The higher ϱ, the more unfair among users, i.e., the users throughput is more affected by CCI. ϱ is one of possible definitions to measure the fairness. In Fig. 4.6, we show the fairness and average transmitted power vs. R_u with $\beta = 1$. When R_u is small and CCI is severe, ϱ is large and the users with the better channel condition occupy most of the resources. The average transmitted power is also low, because most users cannot transmit. When R_u becomes large and CCI is reduced, the users with worse channel conditions can compete for their transmissions, while the users with better channel conditions are not so dominant. Consequently, ϱ is reduced and users transmit more fairly like the single user case. The average transmitted power is increased and saturated with increasing of R_u, because most users can transmit, according to its own channel conditions and regardless to the low CCI if R_u is large.

In Fig. 4.7, we show the average throughput per user vs. P_{max} for different R_u with $\beta = 1$. We can see that the average throughput increases slower when P_{max} is large. This is because the CCI is increasing especially when R_u is small. When R_u is decreasing, the point where average throughput per user saturates moves to the lower P_{max}. There is no need for higher P_{max}, if the performance curve is saturated already. So when R_u is decreasing, we can reduce P_{max} accordingly, such that no transmitted powers will be wasted.

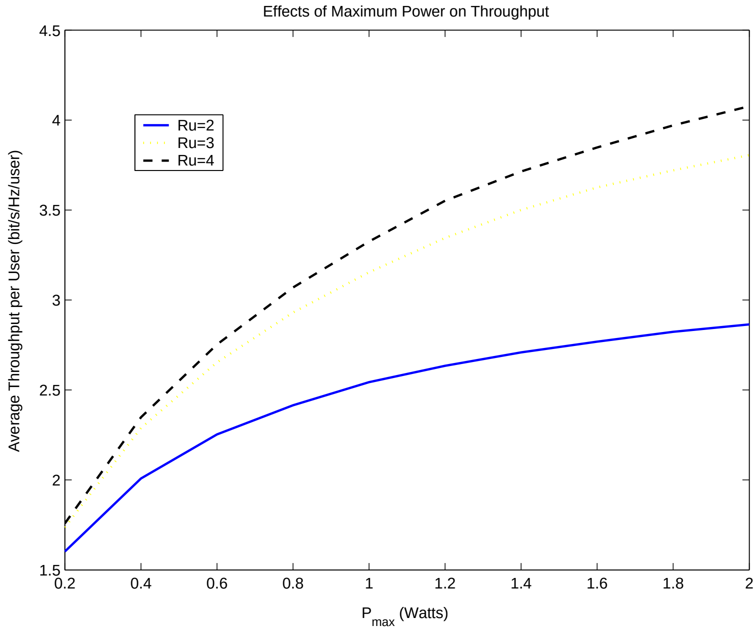


Figure 4.7: Average Throughput per User vs. P_{max}

4.5 Comparison of Centralized and Distributed Resource Allocation

Centralized and distributed resource allocation schemes can be classified by the criteria whether the resource allocation scheme needs the complete channel and user conditions or the scheme only needs the local information. Both centralized and distributed schemes have their advantages and disadvantages and fit different situations.

In the centralized resource allocation scheme, there exists a strong and powerful central node that can high computation capability. This node gathers information of all individual users and makes the decision on resource allocation together by solving the complicated constrained optimization problem. The advantages of this kind of scheme are obvious. First, it can deal with more sophisticated formulated problem such as the cross layer approach, because of the high computation power of the central node. Second, the resource allocation results can be generated fast without iterations. The central node calculates what is the best share of the resources and orders the users to apply the assigned share. So it is very fast in term of convergence and can be applied to the networks where the channels fluctuate more. Third, the result can be more efficient and accurate, because the central node can avoid local optima and assign global optimal resource allocation to users. However, the biggest disadvantage of the centralized scheme is the bottleneck when the number of users are large. Under this condition, the estimation and feedback overheads grow very fast and the computation burden grows even faster, such that the optimal solution is impossible to obtain even with the fastest computer. So this kind of centralized scheme only fits the networks with small number of users or with the topology suitable for centralized control, e.g. single micro-cell CDMA system.

For the distributed resource allocation scheme, only distributed nodes are available and each of them has low computation capability. Each node only has its local information and makes decision based only on its own benefit. The advantage of such kind of scheme is that it can handle large scale systems or the system with the distributed topology. The challenge of distributed scheme is how to design the optimization rule for each node such that the overall system performance can be optimized. The disadvantages of this type of scheme are the opposite of the centralized scheme. First, in order to achieve the global system optimization, each node is limited to do optimization convexly and linearly, which greatly limits the scope, accuracy, and efficiency of optimization for real wireless communication applications. Second, it takes time for distributed scheme to converge. For example, for power control [42], each node modifies its transmitted power according to the received SINR. When it is larger than the desired value, the power is reduced; otherwise, the power is increased. This process has been proved to converge within a few iteration for the stable channel. However, if the channel fluctuates, the convergence might be a problem. So this kind of distributed scheme fits the networks with large number of users or with the distributed topology, e.g. multi-cell system.

We have discussed the tradeoff between the centralized scheme and distributed scheme and shown the different scenarios to apply them. In order to explore the advantages of both systems, we can design a hybrid system. For example, within each cell, the centralized scheme allocates resources to the associated users for this cell. Among different cells, the distributed scheme let the cells compete with each other for the resources. Another example is for downlink and uplink. For downlink, the base station has the strong power and can make centralized optimal solutions by complicated computation. While for uplink, because of the distributed nature of users, the distributed solution is strongly preferred.

Chapter 5

Channel Assignment, Throughput Allocation, and Power Control for OFDMA

Orthogonal frequency division multiplexing (OFDM) is a communications technique that divides a communications channel into a number of equally spaced frequency bands. A subcarrier carrying a portion of the user information is transmitted in each band. Each subcarrier is orthogonal (independent of each other) with every other subcarrier, differentiating OFDM from the commonly used frequency division multiplexing (FDM).

In this chapter, we will review the OFDM technique and OFDMA networks. The basic problems for OFDMA are explained. Then we find three possible optimization solutions for different network scenarios. First for single cell OFDMA network, we apply cooperative game theory approach for channel assignment, throughput management, and power control. Second, for multicell OFDMA network where each cell has only co-channel user, we apply non-cooperative game theory approach for resource allocation. Finally, for multicell OFDMA network with multiple cochannel users per cell, we use subspace method to maximize the system capacity.

5.1 Introduction for OFDMA Networks

Frequency division multiplexing (FDM) is a technology that transmits multiple signals simultaneously over a single transmission path, such as a cable or wireless system. Each signal travels within its own unique frequency range (carrier), which is modulated by the data (text, voice, video, etc.).

Orthogonal FDM's (OFDM) spread spectrum technique distributes the data over a large number of carriers that are spaced apart at precise frequencies. This spacing provides the "orthogonality" in this technique which prevents the demodulators from seeing frequencies other than their own. The benefits of OFDM are high spectral efficiency, resiliency to RF interference, and lower multi-path distortion. This is useful because in a typical terrestrial broadcasting scenario there are multipath-channels (i.e. the transmitted signal arrives at the receiver using various paths of different length). Since multiple versions of the signal interfere with each other (inter symbol interference (ISI)) it becomes very hard to extract the original information.

OFDM is sometimes called multi-carrier or discrete multi-tone modulation. It is the modulation technique used for digital TV in Europe, Japan, and Australia. For example,

- DAB - OFDM forms the basis for the Digital Audio Broadcasting (DAB) standard in the European market.

- ADSL - OFDM forms the basis for the global ADSL (asymmetric digital subscriber line) standard.
- Wireless Local Area Networks - development is ongoing for wireless point-to-point and point-to-multipoint configurations using OFDM technology.

  In a supplement to the IEEE 802.11 standard, the IEEE 802.11 working group published IEEE 802.11a, which outlines the use of OFDM in the 5.8-GHz band.

For multiuser OFDM, OFDM based multi-access technique is called Orthogonal Frequency Division Multiple Access (OFDMA), which has been proposed as the wireless access and signaling scheme in several next generation wireless standards, as a means of achieving data rates of the order of 2-5 Mbits/see in macrocells. In OFDMA, the available spectrum is divided into multiple orthogonal narrowband subchannels (subcarriers) and information symbols are transmitted in parallel over these low-rate subchannels. This method results in reduced intersymbol interference and multipath delay spread, and thus improvement in capacity and attainable data rates. The rationale is that the fading on each individual subchannel is independent from user to user, so that adaptive resource allocation gives each their "best" subchannels and adapts optimally to these channels.

Optimization problem for OFDMA is still open for research. The degrees of freedom are subcarrier allocation, power, rate, coding, and BER. Variety of the optimization goals can be formulated

- Maximize the sum of average user rates
- Find all possible average rate vectors ("capacity" region)
- Find average rate vectors with minimum rate constraints

- Minimize power for some average rate vector
- Minimize outage probability for some constant rate vector

There exist many practical constraints such as

- Maximal power constraint for each mobile
- Minimal rate requirement for each mobile
- Total number of subcarrier
- How many users can share each subcarrier

So one of our research concentration is on this hot and open topic. We find some possible solutions for some formulated OFDMA optimization problems.

5.2 Cooperative Game Approach

We apply cooperative game theory to allocate subcarrier, throughput, and power for uplink single cell OFDMA systems. The goal is to maximize the system throughput, under the power and rate constraints while considering the fairness among users. Our approach is based on Nash bargaining solution. First, a two-user algorithm is developed to bargain subcarrier between users. Based on this algorithm, we develop a multiuser bargaining algorithm where coalitions among users are constructed by Hungarian method. The simulation results show that the proposed algorithms not only provide fair resource allocation among users, but also have similar system throughput as the greedy algorithm of maximizing the total throughput only. Moreover, the proposed algorithms have the complexity of only $O(N \log N)$, where N is the number of subcarrier.

This section is organized as follows: First, we give the motivation and sketch of the proposed scheme. Then, the system model is given. Basic facts for NBS of cooperative game theory are presented. The problem is formulated. A two-user algorithm and a multiuser algorithm are constructed. Simulations are conducted.

Motivation and Sketch

Orthogonal frequency division multiple access (OFDMA) is a promising multiplexing multi-access technique for high data rate transmissions over wireless radio channels. Efficient resource allocation involves bit loading, transmission power allocation, and subcarrier assignment, which can greatly improve system performances.

The resource allocation problem for a single user across parallel orthogonal channels with additive white Gaussian noise with the objective to maximize the total achievable rate subject to a total power constraint is optimally solved by the water-filling method. The throughput allocation in each subcarrier is then determined by the corresponding power allocation. The water-filling solution can also be applied in single-cell multiuser systems with a given set of allocated subcarrier to each user, since in that case power allocation for each user can be studied independently.

However, to optimally assign subcarrier to different users in a single-cell multiuser environment by considering the different users' link qualities is more difficult, because of the discrete nature of the subcarrier assignment problem. By adaptively assigning frequency subcarrier, we can take advantage of channel diversity among users in different locations, which is called multiuser diversity. This multiuser diversity stems from channel diversity including independent path loss and fading of users. Most of the existing works focus on improving the system efficiency by multiuser diversity, [71]-[88]. In [71], the authors studied the dual problem, namely, to find the optimal subcarrier allocation so as to minimize the total transmitted power and satisfy a minimum rate constraint for each user. The dual problem is further formulated as integer programming and a suboptimal solution is found by using the continuous relaxation. In [72], a low-complexity suboptimal algorithm is proposed, which decouples the problem into two sub-problems: (i) find the required power and number of subcarrier for each user and (ii) find the exact subcarrier and throughput allocation. In [73], the discrete subcarrier allocation problem is relaxed into a constrained optimization problem with continuous variables. The problem is shown to belong to the class of convex programming problems, thus allowing the optimal assignment to be found with numerical methods. In [74], the problem is formulated using a max-min criterion for downlink application. A suboptimal algorithm is developed assuming equal amount of power is allocated to each subcarrier. In [75, 76], real-time subcarrier allocation schemes are studied, which only use subcarrier allocation to enhance the performance while fixing modulation levels. The Hungarian method [82] can be used to solve such problems with a high computational complexity of $O(N^4)$, where N is the number of subcarrier. The sub-optimal algorithms are developed in [75, 76] to simplify the Hungarian algorithm and achieve similar performances. In [77], with an appropriate allocation strategy in both frequency and time domains, resources could be used more efficiently.

Most of the previous approaches maximize the total transmission rate or minimize the total transmitted power under some constraints. The formulated problem and their solutions are focused on the efficiency issue. However, these approaches benefit the users closer to the base station or with a higher power capability. The fairness issue is less studied. On the other hand, considering the fairness among users, max-min criterion has been studied for channel allocation in OFDMA systems [74]. However, it is not easy to take into account the notions that users might have different requirements within this framework. In addition, the max-min approach deals with the worst case of the system, penalizing users with better channels and reducing the system efficiency. Moreover, most of the existing solutions have high complexities, which prohibit them from practical implementation. Therefore, it is desired to develop an approach that considers the fairness of resource allocation, system efficiency, and complexity simultaneously.

In uplink single cell OFDMA systems, there are many distributed users that can cooperate in making the decisions on the subcarrier usage, such that each of them will operate at his optimum. Users can communicate via the base station and make joint agreements about their operating points. Such a fact motivates us to apply the cooperative game theory [78, 79, 81], which can achieve the crucial notion of fairness and maximize the system throughput. The Nash Bargaining Solution (NBS) is taken into our consideration for the resource allocation of OFDMA systems. It provides a fair operation point and a distributed implementation. Under certain conditions, the operation point can also be both unique and Pareto optimal.

Motivated by the above reasons, we apply the cooperative game theory for resource allocation in OFDMA systems. We want to maximize the system throughput, under the constraints of each user's minimal throughput requirement and maximal transmitted power. The approach is based on NBS which is not only conditional optimal from system optimization point of view, but shows fairness. First we develop a fast two-user bargaining algorithm to negotiate the usage of subcarrier. Based on this algorithm, we group the users into coalitions by using Hungarian method, such that minimal number of bargaining is required. The complexity is only $O(N \log N)$. From simulations, the proposed algorithms allocate resources fairly and efficiently compared to the greedy algorithm and max-min algorithm.

System Model

Consider an uplink scenario of a single cell OFDMA system. There are totally K users. The users want to share their transmissions among N different subcarrier. The i^{th} user's transmission rate is R_i and is allocated to different channel as $R_i = \sum_{j=1}^{N} r_{ij}$, where r_{ij} is the i^{th} user's transmission rate in the j^{th} subcarrier. Define the rate allocation matrix $\mathbf{r}$ with $[\mathbf{r}]_{ij} = r_{ij}$. Define the subcarrier assignment matrix $[\mathbf{A}]_{ij} = a_{ij}$, where $a_{ij} = 1$, if $r_{ij} > 0$; $a_{ij} = 0$, otherwise. Define power allocation matrix $[\mathbf{P}]_{ij} = P_{ij}$. For single cell OFDMA, no subcarrier can support the transmissions from more than one user, i.e., $\sum_{i=1}^{K} a_{ij} = 1, \forall j$.

Adaptive modulation provides each user with the ability to match each subcarrier's transmission rate r_{ij}, according to its channel conditions. MQAM is a modulation method with a high spectrum efficiency. In [65], BER as function of throughput and SINR is approximated by:

$$\mathrm{BER}_{ij} \approx c_1 e^{-c_2 \frac{\Gamma_{ij}}{2^{r_{ij}}-1}} \tag{5.1}$$

where $c_1 \approx 0.2$, $c_2 \approx 1.5$, and Γ_{ij} is the i^{th} user's SINR at the j^{th} subcarrier, given by:

$$\Gamma_{ij} = \frac{P_{ij} G_{ij}}{\sigma^2} \tag{5.2}$$

where G_{ij} is the subcarrier channel gain and P_{ij} is the transmitted power for the i^{th} user in the j^{th} subcarrier. Assume the thermal noise power for each subcarrier is the same and equal to σ^2. From (5.1), for a fixed BER, we have

$$r_{ij} = W \log_2 \left(1 + \frac{P_{ij} G_{ij} c_3}{\sigma^2}\right) \tag{5.3}$$

where W is the bandwidth and $c_3 = c_2 / \ln(c_1/\text{BER})$ with $\text{BER} = \text{BER}_{ij}, \forall i, j$.

Basics for Nash Bargaining Solution

In this section, we will briefly review the basic concepts and theorems for the cooperative game and NBS. Then we will give a overview on how to apply these ideas to OFDMA resource allocation.

The bargaining problem of cooperative game theory can be described as follows [78, 79, 81]: Let $\mathbf{K} = \{1, 2, \ldots, K\}$ be the set of players. Let $\mathbf{S}$ be a closed and convex subset of $\Re^K$ to represent the set of feasible payoff allocations that the players can get if they all work together. Let R^i_{min} be the minimal payoff that the i^{th} player would expect, otherwise, he will not cooperate. Suppose $\{R_i \in \mathbf{S} | R_i \geq R^i_{min}, \forall i \in \mathbf{K}\}$ is a nonempty bounded set. Then the pair $(\mathbf{S}, R^i_{min})$ is called a K-person bargaining problem.

Within feasible set $\mathbf{S}$, we define the notion of Pareto optimal as a selection criterion for the bargaining solutions.

Definition 5.2.1 *The point R_i, $\forall i$ is said to be **Pareto optimal**, if and only if there is no other allocation R'_i such that $R'_i > R_i, \forall i$. Pareto optimality means that it is impossible to find another resource allocation that leads to strictly superior performance for all users.*

There might be infinite number of Pareto optimal points. We need further criterion to select the bargaining results. One possible criterion is fairness. One commonly used fairness criterion is max-min [74], where the performance of the user with worst channel is maximized. This criterion penalizes the users with good channels and as a result generates inferior system performance. In our approach, we use the proportional fairness criterion [80] of NBS. The intuitive idea is that, after the minimal requirements are assigned to all users, the rest resources are allocated proportionally to users according to their channel conditions.

There exit many kinds of cooperative game solutions [81]. Among them, NBS provides a unique and fair Pareto optimal operation point under some conditions. NBS is briefly explained as follows:

Definition 5.2.2 *$\bar{\boldsymbol{r}}$ is said to be a **Nash Bargaining Solution** in $\boldsymbol{S}$ for $R^i_{min}, \forall i$, i.e., $\bar{\boldsymbol{r}} = \phi(\boldsymbol{S}, R^i_{min})$, if the following Axioms are satisfied:*

1. *Individual Rationality: $\bar{R}_i = \sum_{j=1}^{N} \bar{r}_{ij} \geq R^i_{min}, \forall i$.*
2. *Feasibility: $\bar{\boldsymbol{r}} \in \boldsymbol{S}$.*
3. *Pareto Optimality: For every $\widehat{\boldsymbol{r}} \in \boldsymbol{S}$, if $\sum_{j=1}^{N} \bar{r}_{ij} \geq \sum_{j=1}^{N} \widehat{r}_{ij}, \forall i$, then $\sum_{j=1}^{N} \bar{r}_{ij} = \sum_{j=1}^{N} \widehat{r}_{ij}, \forall i$.*
4. *Independence of Irrelevant Alternatives: If $\bar{\boldsymbol{r}} \in \boldsymbol{S'} \subset \boldsymbol{S}$, $\bar{\boldsymbol{r}} = \phi(\boldsymbol{S}, R^i_{min})$, then $\bar{\boldsymbol{r}} = \phi(\boldsymbol{S'}, R^i_{min})$.*
5. *Independence of Linear Transformations: For any linear scale transformation ψ, $\psi(\phi(\boldsymbol{S}, R^i_{min})) = \phi(\psi(\boldsymbol{S}), \psi(R^i_{min}))$.*
6. *Symmetry: If $\boldsymbol{S}$ is invariant under all exchanges of agents, then $\phi_j(\boldsymbol{S}, R^i_{min}) = \phi_{j'}(\boldsymbol{S}, R^i_{min}), \ \forall j, j'$.*

Axiom 4-6 are called axioms of fairness. The irrelevant alternative axiom asserts that eliminating feasible alternatives that would not have been chosen should not affect the solution. Axiom 5 asserts that the bargaining solution is scale invariant.

Symmetry axiom asserts that if the position of players are completely symmetric, then the solution should also treat them symmetrically.

From [81], there is exactly one NBS that satisfies the above axioms, which is shown in the following theorem.

Theorem 5.2.3 ***Existence and Uniqueness of NBS****: There is a unique solution function $\phi(\boldsymbol{S}, R^i_{min})$ that satisfies all six axioms in Definition 1. And this solution satisfies*

$$\phi(\boldsymbol{S}, R^i_{min}) \in \arg \max_{\bar{\boldsymbol{r}} \in \boldsymbol{S}, \bar{R}_i \geq R^i_{min}, \forall i} \prod_{i=1}^{K} \left(\bar{R}_i - R^i_{min} \right) \tag{5.4}$$

Until now, we have provided the mathematical background for the cooperative game theory. The cooperative game in the single cell OFDMA system can be defined as follows: Each of K users has R_i as his objective function, where R_i is bounded above and have a nonempty, closed, and convex support. The goal is to maximize all R_i simultaneously. R^i_{min} represents the minimal performance and is called the initial agreement point. Define $\mathbf{S}$ as the feasible set of rate allocation matrix $\mathbf{r}$ that satisfies $R_i \geq R^i_{min}, \forall i$. The problem, then, is to choose the operating point in $\mathbf{S}$ for users, such that this point is Perato optimal and fair.

Cooperative Game Approaches

Problem Formulation

Considering a channel for a specific subcarrier may be good for more than one users, there is competition among users to put their transmissions into the subcarrier with large G_{ij}. Moreover each mobile user's maximal transmitted power is bounded by some value P_{max} and each user has a minimal throughput requirement R^i_{min} if he is admitted to the system. In our approach, the optimization goal is to allocate different users' transmission to the different subcarrier such that NBS cost will be maximized, i.e.,

$$\max_{\mathbf{A},\mathbf{P}} U = \prod_{i=1}^{K} \left(R_i - R^i_{min}\right) \tag{5.5}$$

$$\text{subject to} \begin{cases} \sum_{i=1}^{K} a_{ij} = 1, \forall j; \\ R_i \geq R^i_{min}, \forall i; \\ \sum_{j=1}^{N} P_{ij} \leq P_{max}, \forall i. \end{cases}$$

We use the goal U as a product form for two reasons. First, it will be shown later that this form will ensure fairness of allocation. Second, cooperative game theories prove that there exists a unique and efficient solution under some conditions. The difficulty to solve (5.5) by traditional methods lies in two factors. First, the problem itself is a nonlinear constrained combinatorial problem. Second, distributed algorithms are desired for uplink OFDMA systems. We will use the bargaining concept and develop simple algorithms that can achieve the social optimal and fair resource allocation in next two subsections.

In Fig. 5.1, a two-user example is illustrated. R^i_{min} is assumed to be zero. **S** is the feasible range for R_1 and R_2. For the proposed cost function, the optimal point is (R_1, R_2). The physical meaning can be explained as follows: After assigned with the minimal throughput, "the remaining resources are divided between users in a ratio equal to the rate at which the utility can be transferred" [81]. The geometrical interpretation is by drawing a triangle such that its one side tangents the set **S** and one vertex is at (R^1_{min}, R^2_{min}). In our case, user 1 has better channel conditions than user 2. Compared with the greedy algorithm which maximizes the sum of throughput (i.e. $R_1 + R_2$) and has the optimal point at (R_1^*, R_2^*), our solution has slightly overall throughput loss, but keeps fairness. Compared with the max-min algorithm where the system satisfies the worst case situation and has the strictly fair at optimal point (R_1', R_2'), our solution has much higher overall throughput.

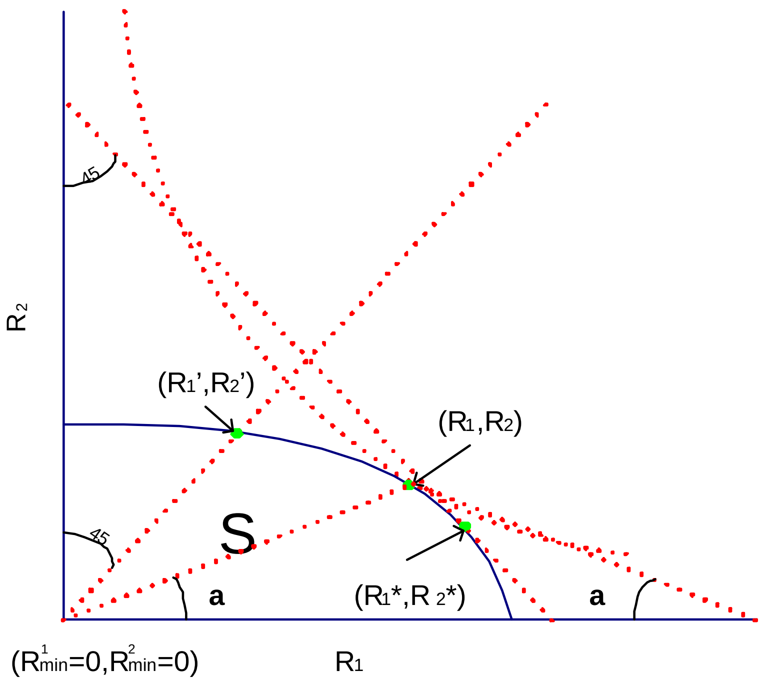


Figure 5.1: Two-User Illustrative Example

When $R^i_{min} = 0, \forall i$, the fairness is the same as the proportional fairness [80], i.e., any change in the distribution of rates will result in the sum of the proportional changes of the utilities to be non-positive as:

$$\sum_i \frac{R_i^* - R_i}{R_i} \leq 0, \ \forall R_i^* \in \mathbf{S}. \tag{5.6}$$

Bargaining Algorithm for Two-user Case

In this part, we will develop a two-user bargaining algorithm. The intuitive idea is to allow two users to negotiate and exchange their subcarrier such that benefits will be obtained. The idea is similar to bargaining in a real market. The difficulty is to determine how to optimally exchange subcarrier, which is a complex integer programming problem. In [73], the authors develop a low complexity algorithm. The idea is to use a simple two band partition for subcarrier assignment.

Table 5.1: Two-user Algorithm

1. Sort the subcarrier: Arrange the index from the largest to smallest $\frac{G_{1j}}{G_{2j}}$.
2. For j=1,. . . , N-1 User 1 occupies and water-fills subcarrier 1 to j; User 2 occupies and water-fills subcarrier j+1 to N. Waterfill both users to assigned subcarrier sets. Calculate U. **End**
3. Choose the j that generate the largest U **that satisfies the constraints.**

The authors prove that when SINR is high, the two band partition for two user subcarrier assignment is near optimal.

Based on the similar idea, we develop a fast algorithm for two users to exchange their subcarrier in Table 5.1. First we combine the two user's subcarrier and get the channel gains for all subcarrier. In this combined subcarrier set, the subcarrier is sorted by the order of user1's channel gain over user2's channel gain. The subcarrier allocation tries to find the optimal partition point to maximize the cost function $U = (R_1 - R^1_{min})(R_2 - R^2_{min})$, where user1 occupies and waterfill the first part of subcarrier set and user2 uses the rest. The algorithm has the complexity of $O(N^2)$ and can be further improved by using a binary search algorithm with a complexity of only $O(N \log N)$.

Cooperate Game Algorithm for Multiple Users

For multiuser cooperative game, one simple straight forward algorithm is to let the users bargain randomly, while the constraints are satisfied. The algorithm can be described by the following two steps:

1. Initialization: The goal is to assign all subcarrier and make sure that each user has at least throughput of R^i_{min} and the power constraint is satisfied. We develop a fast algorithm to allocate R^i_{min} to each user under the power constraint. If the user has throughput larger than R^i_{min}, he is removed from the assignment list. After every user has enough throughput, the rest of subcarrier is assigned to the users with the maximal channel gains.

2. Bargain: Negotiate between any two users to exchange the subcarrier by the two user algorithm in Table 5.1, such that the optimization product U is increased, until no improvement can be achieved.

We call this algorithm the random method. However the complexity explodes and the convergence is slow with the number of users increasing. This is because optimal cooperation among subsets of the users is not considered. Each user needs to carefully select who he should negotiate with. So we define a new concept for grouping the users:

Definition 5.2.4 *For a K-person game, any nonempty subset of the set of players is called a* ***coalition****.*

We call the users can negotiate effectively if there is a feasible change in the strategies of the members of the coalition that would benefit them all. There are many possible coalitions and most of them are not effective. In order to reduce the number of rounds for negotiate, effective coalitions should be carefully selected. In our approach, we concentrate on the coalition with the size of 2 and how to speed up the convergence of negotiations.

We quantify the convergence speed by the round of negotiations. For each round, two users are negotiated together to exchange their subcarrier. The problem is to decide how to form the coalition pairs, such that the overall system can be improved most. The negotiation iteration is continued until no user can be improved by negotiations.

Each user's channel gains are various over different subcarrier. A user may be preferred by many users to form coalitions, while only two user coalition is allowed. Thus, the problem to decide the coalition pairs can be stated as an assignment problem. Define a $K \times K$ assignment table $\mathbf{X}$. Each component represents whether or not there is a coalition between two users.

$$X_{ij} = \begin{cases} 1 & \text{if user i negotiates with user j;} \\ 0 & \text{otherwise.} \end{cases} \tag{5.7}$$

Define the benefit for the i^{th} user negotiates with the j^{th} user as b_{ij}. Obviously $b_{ii} = 0, \forall i$. For the other cases, from (5.5), each element of the cost table $\mathbf{b}$ can be expressed as:

$$\begin{aligned} b_{ij} = &\max(\log(\widetilde{R}_i - R^i_{min}) + \log(\widetilde{R}_j - R^j_{min}) \\ &- \log(\widehat{R}_i - R^i_{min}) - \log(\widehat{R}_j - R^j_{min}), 0). \end{aligned} \tag{5.8}$$

where $\widetilde{R}_i$ and $\widetilde{R}_j$ are the throughput if the negotiation happens, and $\bar{R}_i$ and $\bar{R}_j$ are the original throughput. So the problem is how to select the pairs of negotiations such that the overall benefit will be maximized. It is an assignment problem stated as:

$$\max_{\mathbf{X}} \sum_{i=1}^{K} \sum_{j=1}^{K} X_{ij} b_{ij} \tag{5.9}$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{K} X_{ij} = 1 & j = 1 \ldots K, \forall i; \\ \sum_{j=1}^{K} X_{ij} = 1 & i = 1 \ldots K, \forall j; \\ X_{ij} \in \{0, 1\} & \forall i, j. \end{cases}$$

If the number of user is odd, we can add a dummy user to make the total number of users even. To exchange with this user will always generate zero payoff. One of the popular solution for (5.9) is Hungarian method [82] which can always find the optimal coalition pairs. We change the maximization problem in (5.9) into a minimization problem by multiplying every b_{ij} by -1 and then adding the maximal value of $\mathbf{b}$. The algorithm is briefly explained as follows:

Step 1: Subtract the minimum element in each row from every entry in that row of a cost table.

Step 2: Subtract the minimum element in each column from every entry in that column of the resulting equivalent cost table. This step results in at least one zero in every row and column. If there is a complete set of assignments with zero elements is possible than the resultant equivalent cost table is the optimal solution otherwise go to next step.

Step 3: Draw a set of minimum number of lines through some of the rows and columns in such a way as to cover all the zeros. Subtract the minimum element from every element without a line through them and then add that minimum element that lies at the intersection of two lines. Now if there is a complete set of assignments with zero elements is possible than the resultant equivalent cost table is the optimal solution otherwise repeat this step (Step 3).

The complexity of Hungarian method is $O(K^4)$. Since the number of users is much less than the number of subcarrier, the complexity of the proposed algorithm is much lower than the schemes that apply Hungarian method directly to subcarrier domain [75, 76].

In each round, the optimal coalition pairs are determined by Hungarian method and then the users are set to bargain together using the two user algorithm. The

Table 5.2: Multiuser Algorithm

1. Initialize the channel assignment:
Assign at least R^i_{min} to each user.
2. Forming the coalition
- Method 1: Randomly form 2-user coalition, but
more stages to achieve stability for the whole system;
- Method 2: Hungarian algorithms.
3. Solve the 2-user NBS for each coalition
4. Need to bargain again?
If yes, go to step 2; else, algorithm ends.

whole algorithm stops when no bargaining can improve the performance, i.e., **b** is equal to a zero matrix.

Based on the above explanations, we develop cooperative game algorithm for multi-user resource allocation in single cell OFDMA systems as Table 5.2

Simulation Results

First, a two-user uplink OFDMA system is taken into consideration. We simulate the OFDM system with 32 subcarrier over 3.2 MHz band (or equivalently, an average of 100k bits/subcarrier). To evaluate the performances, we have simulated 10^5 sets of four-path frequency selective Rayleigh fading channels, which has an exponential power profile with 100ns root-mean-square (RMS) delay spread. The maximal power is $P_{max} = 0.5$Watts, and the desired BER is 10^{-2}. The thermal noise level is $N_0 = 10^{-10}$Watts. The propagation loss factor is 3. The distance between user 1 and base station is fixed at $D_1 = 50$m, while D_2 is varying from 10m to 200m. $R^i_{min} = 2\text{M bps}, \forall i$.

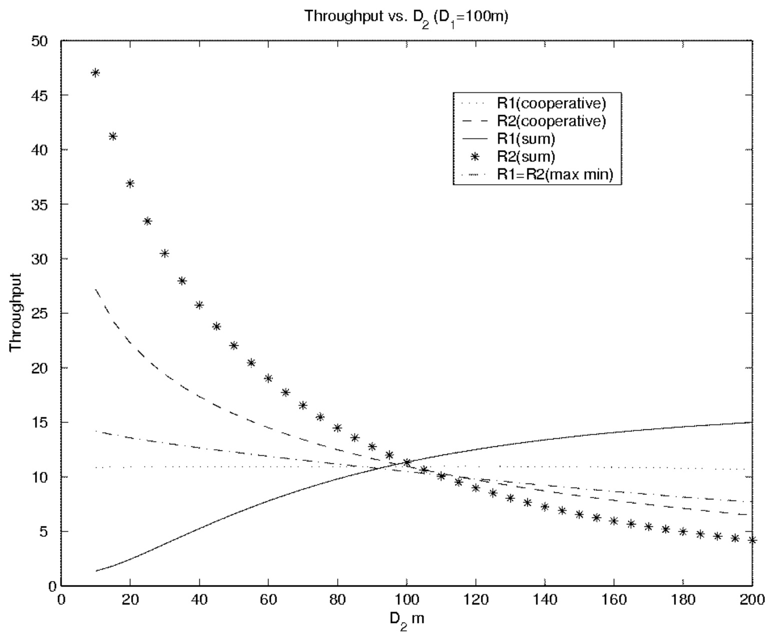


Figure 5.2: Each User's Throughput (Mbps) vs. D_2

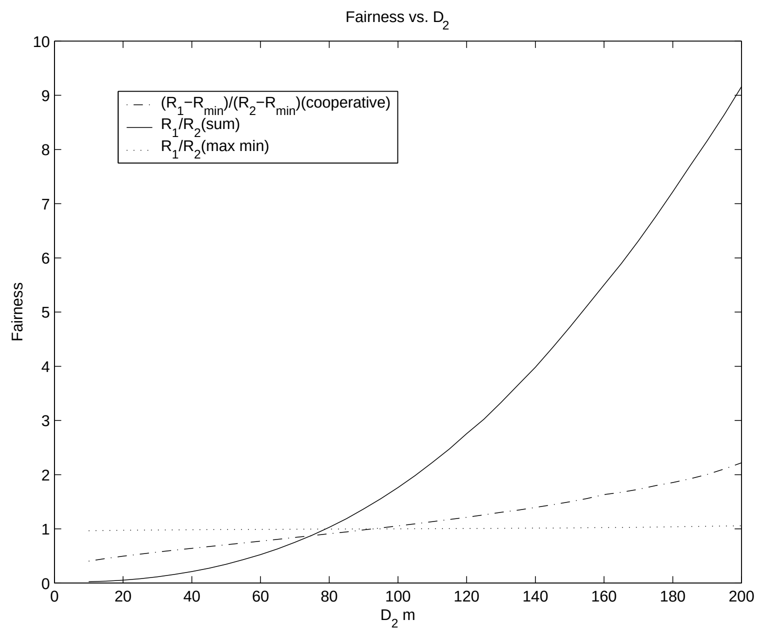


Figure 5.3: Fairness for Three Algorithms

In Fig. 5.2, the throughput of both users for the cooperative, greedy, and max-min algorithms is shown vs. D_2. For the greedy algorithm, the user close to the base station will have higher throughput and the throughput difference is very large when D_1 and D_2 are different. For the max-min algorithm, both users have the same throughput which is reduced when D_2 is increasing. While for the cooperative algorithm, user 1's throughput is almost the same regardless D_2 and user 2's throughput is reduced when D_2 is increasing. This shows that our algorithm is fair in the sense that the user's throughput is determined only by his channel condition and not by other users' conditions. In addition, the ratio of two users' throughput is shown in Fig. 5.3. For the max-min algorithm, the ratio is always equal to 1, which is strict fair. For the greedy algorithm, the ratio changes greatly for different D_2, which is very unfair. For the proposed algorithm, the ratio of $R_1 - R^1_{min}$ over $R_2 - R^2_{min}$ changes almost linearly with D_2, which shows the fairness of NBS.

In Fig. 5.4, we show the overall throughput $R_1 + R_2$ for three algorithms vs D_2. Because the max-min algorithm is for worst case situation, it has the worst performance. The cooperative algorithm has the performance between the greedy algorithm and max-min algorithm, while the greedy algorithm cannot guarantee the minimal throughput requirement R^i_{min}.

We setup the simulations with more users to test the proposed algorithms. All the users are randomly located within the cell of radius 200m. Each user is assigned with the minimal throughput $R^i_{min} = 500$k bps first, then we use the greedy, max-min, and cooperative algorithm to optimize the system performance. The other settings are the same as two-user case simulations.

In Fig. 5.5, we show the sum of all users' throughput vs. the number of

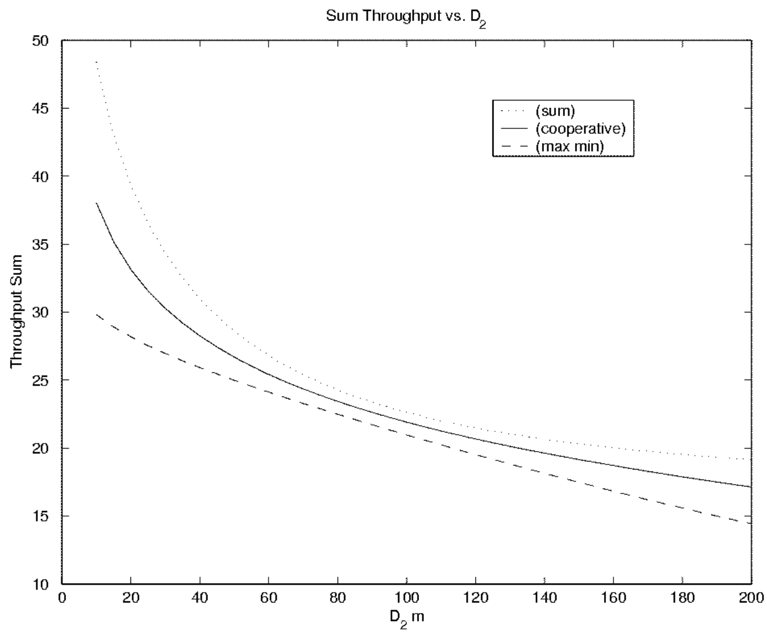


Figure 5.4: Overall Throughput (Mbps) $R_1 + R_2$

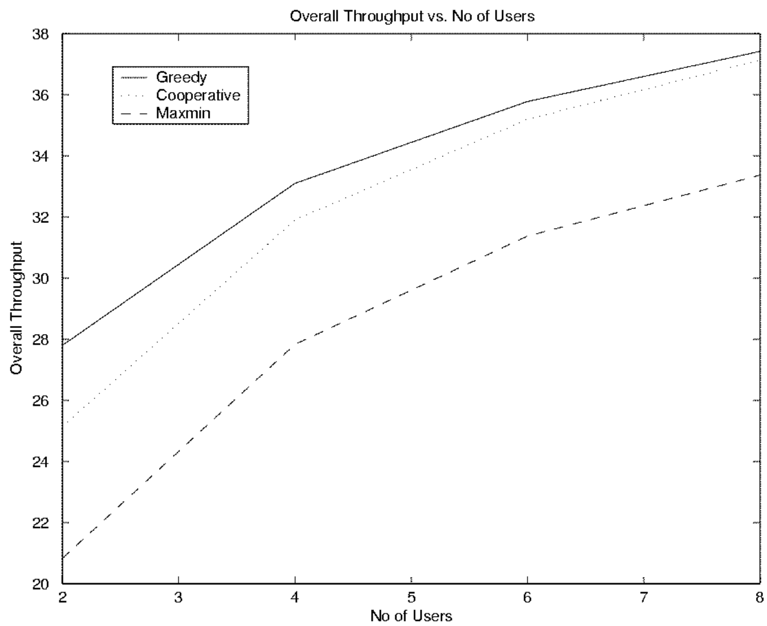


Figure 5.5: Overall Throughput (Mbps) vs. No. of Users

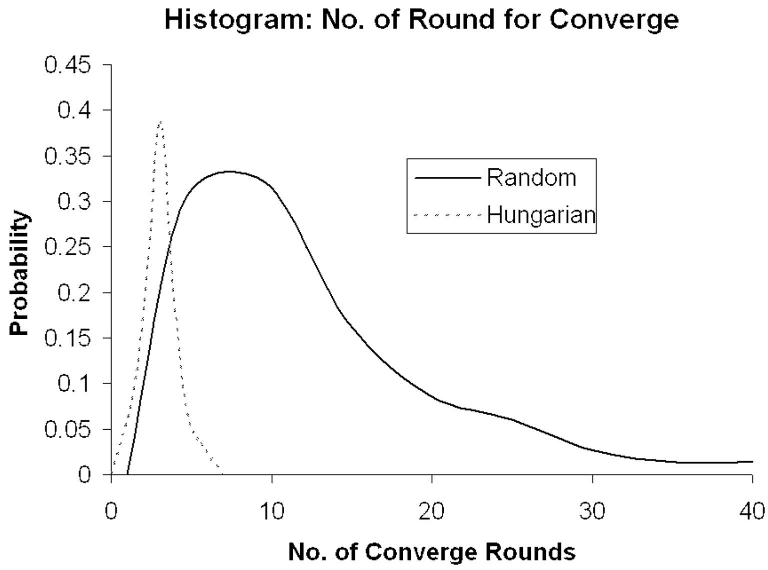


Figure 5.6: Histogram for Convergence

users in the system for the three algorithms. We can see that all three algorithms have better performances when the number of users increases. This is because of multiuser diversity. The performance improvement satiates gradually. The proposed cooperative algorithm has a similar performance to that of the greedy one and has a much better performance than that of the max-min algorithm. The performance gap between the greedy algorithm and the cooperative algorithm reduces when the number of users is large. This is because more bargain pair choices are available to increase the system performance.

In Fig. 5.6, we show the histogram of the number of rounds necessary for converge of the random method and Hungarian method. Hungarian method converges in about 1 to 6 rounds, while the random method may converge very slowly. The average converge rounds for the random method is 4.25 times to that of Hungarian method. By using Hungarian method, we can find the best pairs to negotiate. Consequently the convergence rate is much quicker.

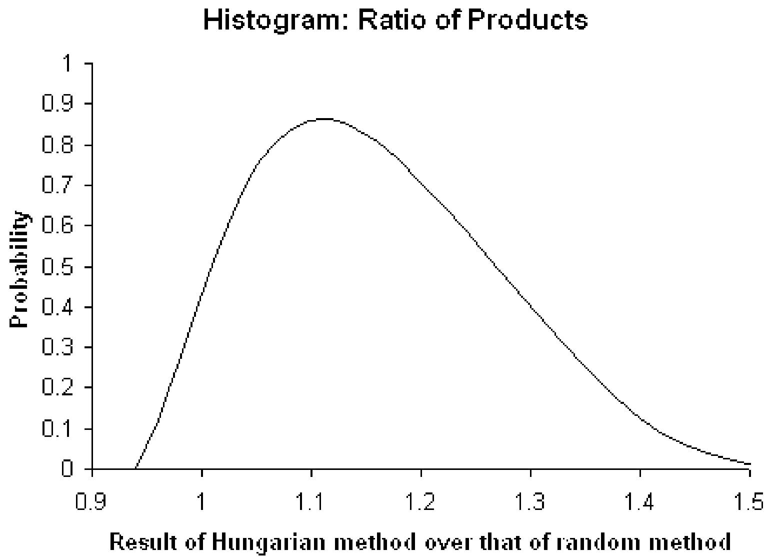


Figure 5.7: Histogram for Product Ratio

In Fig. 5.7, we show the ratio of $\Pi_{i=1}^{K}(R_i - R_{min}^i)$ of Hungarian method over that of the random method. Hungarian method converges to a better solution in most of times. The random algorithm may fall into some local optima. This is because the Nash six axioms may not be satisfied in the proposed system and the two user algorithm is suboptimal. However, most of times, the ratio is a small number, so the problem of local optima is not severe.

5.3 Non-cooperative Game Approach

In this section, we use noncooperative game approach to have sub-channel assignment, adaptive modulation, and power control for multi-cell OFDMA networks. The goal is to minimize the overall transmitted power under the constraints that each user has the desired throughput and each user's power is bounded. Our contribution is to model and solve this complicated problem by a distributed noncooperative game approach: Each user water-fills its power to different sub-channels regarding other users' powers as interferences. A noncooperative game is constructed for each user to compete with each other. A heuristic method is constructed as a mediator (judge) for the game. From the simulation results, the proposed scheme reduces the overall transmitted power greatly compared with the fixed channel assignment algorithm and pure water-filling algorithm.

The rest of this section is organized as follows: First, we give the motivation and sketch for the proposed scheme. Then, we give the system model and formulate the problem. The adaptive algorithm of the noncooperative approach is developed. We have simulation studies.

Motivation and Sketch

In a multi-cell OFDMA system, the resource allocation problem becomes more complicated, even if the assignment of sub-channels to users is predetermined. This is because users in different cells reuse the same sub-channels and cause interferences to each other. If the number of co-channel users is relatively large, the interference seen by a user in a sub-channel can be approximated by a Gaussian random variable by applying the central limit theorem. In this case, water-filling algorithm can provide a good solution. When the channel assignment is fixed, many iterative water-filling methods are proposed in [83, 84, 85, 86] to maximize the throughput with power constraints. However, if the sub-channel assignment to users is not predetermined, all possible combinations of co-channel users should be checked to determine the best resource allocation. In [87], the authors present heuristic distributed algorithms that are executed independently by each base station, which are based on iterative water-filling with removing sub-channels of low signal to interferences and noise ratio (SINR). In [88], the problem of channel allocation with modulation and power control in a multi-cell system is studied for generic multiple access schemes with orthogonal channels.

Since in the multi-cell case, individual mobile users do not have the knowledge of other users conditions and cannot cooperate with each other, they act selfishly to maximize their own performances in a distributed fashion. Such a fact motivates us to adopt the game theory [56]. The resource allocation can be modelled as a noncooperative game that deals largely with how rational and intelligent individuals interact with each other in an effort to achieve their own goals. In the resource allocation game, each mobile user is self-interested and trying to optimize his utility function, where the utility function represents the user's performance and controls the outcomes of the game.

In our approach, we want to minimize the overall transmitted power, under the constraint that each user has the desired throughput and each user's transmitted power is bounded. By noncooperative game theory approach, we find the following facts: If the co-channel interferences are small, users can share the sub-channels for transmission. In this case, by carefully designing the utility function, the noncooperative game for each user to compete the resources will be balanced in an optimal and unique Nash equilibrium point (NEP). If the co-channel interferences are severe for some sub-channels, NEP may not be optimal and there might be multiple NEPs. In order to deal with this condition, some users with bad channels or large interferences to others must be kicked out from using these sub-channels, so that the rest of the users can make use of the corresponding sub-channels. We design the utility function for each user, define the criterion as a game rule to kick out users, and develop the adaptive algorithms for resource allocation. From the simulation results, we can see that the proposed scheme can reduce the overall transmitted power greatly compared to the fixed channel assignment algorithm and pure water-filling algorithm.

System Model and Problem Formulation

The K co-channel links are taken into consideration that may exist in distinct cells of OFDMA networks. Each link consists of a mobile user and its assigned base station. Assume coherent detection is possible so that it is sufficient to model this multiuser system by an equivalent baseband model. The total number of OFDM sub-channels is L. For the uplink case, the sampled signal on the l^{th} sub-channel of the i^{th} user can be expressed as:

$$x_i^l(n) = \sum_{k=1}^{K} \sqrt{P_k^l G_{ki}^l} s_k^l(n) + n_i^l(n) \tag{5.10}$$

where P_k^l and G_{ki}^l is the transmitted power and propagation loss from the k^{th} user to the i^{th} base station in the l^{th} sub-channel, respectively, s_k^l is message symbol from the k^{th} user to the i^{th} base station at time n, and $n_i^l(n)$ is the sampled thermal noise. We assume that the channels change slowly. Without loss of generality, we assume $N_i^l = E(\|n_i^l\|^2) = N_0$. The i^{th} user's SINR at sub-channel l can be expressed as:

$$\Gamma_i^l = \frac{P_i^l G_{ii}^l}{\sum_{k \neq i} P_k^l G_{ki}^l + N_0}. \tag{5.11}$$

Rate adaptation such as adaptive modulation provides each sub-channel with the ability to match the effective bit rates, according to the interference and channel conditions. MQAM is a modulation method with high spectrum efficiency. Without loss of generality, we assume the output of different adaptive modulation constellation has unit power. In [54, 65], for a desired throughput r_i^l of MQAM, the BER of the l^{th} sub-channel of the i^{th} user can be approximated as a function of the received SINR Γ_i^l by:

$$\text{BER}_i^l \approx c_1 e^{-c_2 \frac{\Gamma_i^l}{2^{r_i^l}-1}} \tag{5.12}$$

where $c_1 \approx 0.2$ and $c_2 \approx 1.5$ with small BER_i^l. Rearrange (5.12), for a specific desired BER_i^l, the i^{th} user's transmission rate of the l^{th} sub-channel for the SINR Γ_i^l and the desired BER_i^l can be expressed as:

$$r_i^l = W \log_2(1 + c_3^i \Gamma_i^l) \tag{5.13}$$

where W is the bandwidth and $c_3^i = -\frac{c_2^i}{\ln(\text{BER}_i^l/c_1^i)}$. In our approach, for simplicity, we assume all the sub-channels and users have the same BER requirement, i.e., $\text{BER}_i^l = \text{BER}, \ \forall\, i, l$.

Each user requires the throughput R_i and distributed its throughput into L sub-channels, i.e., $\sum_{l=1}^{L} r_i^l = R_i$. Each user's transmitted power is bounded by P_{max}. Define the $K \times L$ channel assignment matrix $\mathbf{A}$ with $[\mathbf{A}]_{il} = 1$, if $r_i^l > 0$; $[\mathbf{A}]_{il} = 0$, otherwise. Therefore, our objective is to minimize the overall transmitted power under the throughput and power constraints, i.e.,

$$\min_{\mathbf{A},\mathbf{r}} f(\mathbf{r}) = \sum_{i=1}^{K} \sum_{l=1}^{L} P_i^l \tag{5.14}$$

$$\text{s.t.} \begin{cases} \sum_{l=1}^{L} r_i^l - R_i = 0, \ \forall i, \\ \sum_{l=1}^{L} P_i^l - P_{max} \leq 0, \ \forall i, \\ r_i^l, P_i^l \geq 0, \ \forall i, l. \end{cases}$$

The problem in (5.14) is very difficult to solve by centralized constrained nonlinear integer optimization, because the complexity and communication overhead grows fast as the number of users increases. This motivates us to develop a distributed algorithm with limited controls by using the game theory approach.

Noncooperative Game Approach

Our focus is to solve (5.14) by noncooperative game theory. First, we analyze the system feasible region. Then we will construct the game. A two-user two-sub-channel example is given to show insights. The properties of the NEP are analyzed. Finally, an iterative algorithm for multiple users with a game mediator is developed.

System Feasibility Region

In order to ensure the desired BER, for every sub-channel, every user should have SINR no less than the required SINR γ_i^l, i.e., $\Gamma_i^l \geq \gamma_i^l, \ \forall\, i, l$. Rewrite these inequalities in matrix form, we have

$$(\mathbf{I} - \mathbf{D}^l\mathbf{F}^l)\mathbf{P}^l \geq \mathbf{v}^l, \ \forall l, \tag{5.15}$$

where $\mathbf{I}$ is a $K \times K$ identity matrix, $\mathbf{v}^l = [v_1^l, \ldots, v_K^l]'$ with $v_i^l = N_0\gamma_i^l/G_{ii}$, $\mathbf{D}^l = diag\{\gamma_1^l, \ldots, \gamma_K^l\}$, and

$$[\mathbf{F}_{ij}^l] = \begin{cases} 0 & \text{if } j = i, \\ \frac{G_{ji}^l}{G_{ii}^l} & \text{if } j \neq i. \end{cases}$$

By Perron-Frobenius theorem, there exists a positive power allocation if and only if the maximum eigenvalue of $\mathbf{D}^l\mathbf{F}^l$, i.e. spectrum radius $\rho(\mathbf{D}^l\mathbf{F}^l)$, is inside unit circle. When $|\rho(\mathbf{D}^l\mathbf{F}^l)| < 1$, the optimal power solution is

$$\mathbf{P}^l = \begin{cases} (\mathbf{I} - \mathbf{D}^l\mathbf{F}^l)^{-1}\mathbf{v}^l, & |\rho(\mathbf{D}^l\mathbf{F}^l)| < 1; \\ +\infty, & \text{otherwise.} \end{cases} \tag{5.16}$$

The system feasibility region Ω is defined as the supporting domain where there exist solutions and power constraint in (5.14) is satisfied. The condition for (5.16) to have finite solutions is a necessary condition for existence of feasible Ω.

Noncooperative Game and Nash Equilibrium

Each user wants to minimize its transmitted power by allocating its throughput into the different sub-channels, regardless other users in a distributed way. Define

$\mathbf{r}_i = [r_i^1 \dots r_i^L]^T$, the noncooperative game can be written as:

$$\text{Game: } \arg\min_{\mathbf{r}_i \in \Omega} u_i = \sum_{l=1}^{L} P_i^l, \; s.t. \; \sum_{l=1}^{L} r_i^l = R_i. \tag{5.17}$$

where u_i is the utility function defined as the i^{th} user's transmit power. If the interferences from others are fixed, it is a water filling problem. Define

$$I_i^l = \frac{\sum_{k \neq i} P_k^l G_{ki}^l + N_0}{c_3^i G_{ii}^l}, \tag{5.18}$$

the solution is

$$P_i^l = (\mu_i - I_i^l)^+ \text{ and } r_i^l = \log_2(1 + \frac{P_i^l}{I_i^l}) \tag{5.19}$$

where $y^+ = max(y, 0)$. μ_i is solved by bisection search of

$$\sum_{l=1}^{L} \log_2 \left(1 + \frac{(\mu_i - I_i^l)^+}{I_i^l}\right) = R_i. \tag{5.20}$$

However the interferences from other users do change. Based on the game theory [56], the system will be balanced in a Nash equilibrium defined as:

Definition 5.3.1 *Define* $\boldsymbol{r}_i^{-1} = [\boldsymbol{r}_1 \dots \boldsymbol{r}_{i-1} \boldsymbol{r}_{i+1} \dots \boldsymbol{r}_L]$. *Nash Equilibrium Point* $\boldsymbol{r}_i$ *is defined as:*

$$u_i(\boldsymbol{r}_i, \boldsymbol{r}_i^{-1}) \leq u_i(\tilde{\boldsymbol{r}}_i, \boldsymbol{r}_i^{-1}), \; \forall i, \; \forall \tilde{\boldsymbol{r}}_i \in \Omega, \; \boldsymbol{r}_i^{-1} \in \Omega^{L-1}. \tag{5.21}$$

i.e., given the other users' throughput allocation, no user can reduce its transmitted power by changing its resource allocation alone.

Two-User Two-sub-channel Example

In order to explain the Nash equilibrium and show the idea of how we solve the problem. A simple two-user two-sub-channel example is illustrated as follows. The simulation setup is: BER $= 10^{-3}$, $N_0 = 10^{-3}$, $P_{max} = 10^4$, and

$$G_1 = \begin{bmatrix} 0.0631 & 0.0100 \\ 0.0026 & 0.2120 \end{bmatrix}, G_2 = \begin{bmatrix} 0.4984 & 0.0067 \\ 0.0029 & 0.9580 \end{bmatrix}.$$

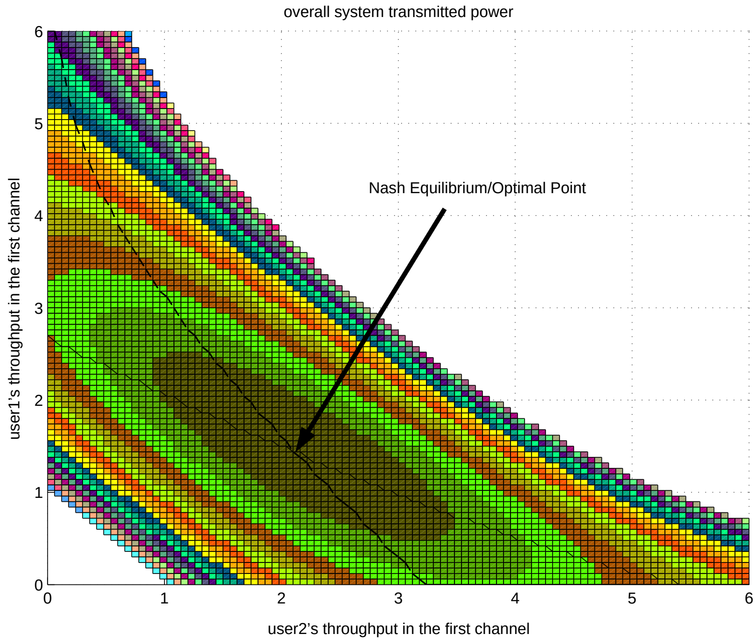


Figure 5.8: Two-user example: Unique NEP

Fig. 1 shows the overall power contour as a function of two users' throughput allocations, where $R_1 = R_2 = 6$. The axes are users' throughput in the first sub-channel. The two black curves show the minimal locations for the two users' own powers when the interference from the other user is fixed, respectively. Each user tries to minimize its power by adjusting its throughput allocation so that the operating point is more close to the curve. Consequently, the cross is a Nash equilibrium, where no user can reduce its power alone. We can see that the Nash equilibrium under this setup is unique and optimal for the overall power. It is worthy to mention that the feasible domain is not convex at all. Fig. 2 shows the situation when $R_1 = R_2 = 8$. Because the throughput is increased, the co-channel interferences are increased and the NEP is no longer the optimum. There exists more than one local optima and the global optimum occurs when user1 doesn't occupy the sub-channel 1. Fig. 3 shows the situation when $R_1 = R_2 = 8.5$. The

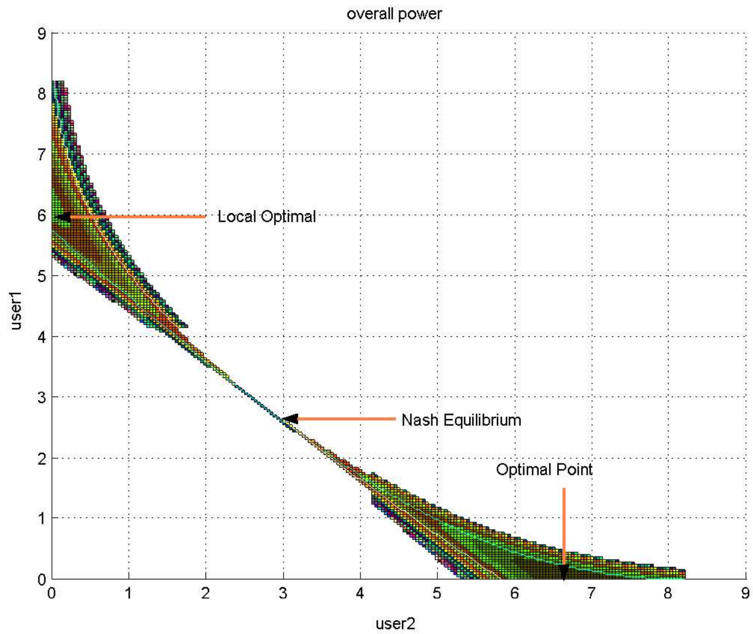


Figure 5.9: Two-user example: Multiple Local Optima

contour graph is not connected. There are two NEPs and two local optima. Under the above two conditions, we need to remove users from using the sub-channels. If we further increase $R_1 = R_2 = 10$, there exists no feasible area, i.e., both users cannot have a resource allocation that satisfies both power and throughput constraints. In this case, the throughput requirement should be reduced.

From the above observations, we can see that the behaviors of the optimal solution and NEP depend on how severe interferences are. In order to let NEP converge to the desired solution, we need to find a criterion as the game rule to decide whether the users can share the sub-channels. If not, who should be kicked out from using the sub-channels. Before we develop the proposed algorithm, following two theorems are proved for the properties of NEP.

Properties of Nash Equilibrium

Theorem 5.3.2 *There exists NEP in the proposed game defined in (5.17), if Ω is not empty.*

Proof In [56], it has been shown a NEP exists, if $\forall\ i$

1. Ω, the support domain of $u_i(\mathbf{r}_i)$, is a nonempty, convex, and compact subset of some Euclidean space $\Re^L$.

2. $u_i(\mathbf{r}_i)$ is continuous in $\mathbf{r}_i$ and quasiconvex in r_i^l.

We consider that each user allocates its transmitted power to different sub-channnels first. Since each sub-channel can be allocated by P_{max} and overall transmitted power for all sub-channels is linearly constraint by P_{max}, the supporting domain for power allocation is compact and convex. Because throughput is a linear function of transmitted power if the interferences are fixed, the supporting domain Ω for r_i^l, $\forall l$ is a convex and compact subset of some Euclidean space $(\Re^+)^L$. It is worthy mentioning that Ω^K is not convex and one example is shown in Fig. 1. But our proof only needs that Ω is convex and nonempty.

From (5.11) and (5.13), when the watterfilling is done for (5.17),

$$u_i = \sum_{l=1}^{L} \left(\frac{(2^{r_i^l} - 1)(\sum_{k \neq i} P_k^l G_{ki}^l + N_0)}{c_3^i G_{ii}^l} - \mu_i r_i^l \right). \tag{5.22}$$

Obviously, it is continuous and convex for $\mathbf{r}_i$. QED ∎

Theorem 5.3.3 *If the global minimum of (5.14) occurs when $r_i^l > 0, \forall A_{il} \neq 0$ and $\sum_{l=1}^{L} P_i^l < P_{max}$ and $\sum_{l=1}^{L} r_i^l = R_i$, $\forall i$, i.e., the assigned users can share all channels, the NEP satisfies the necessary Karush-Kuhn-Tucker (KKT) condition [89].*

Proof First, if $\sum_{l=1}^{L} P_i^l < P_{max}$ and $\sum_{l=1}^{L} r_i^l = R_i$, $\forall i$ at NEP, the iterative water-filling converges. For each user, the resource allocation is optimal if the

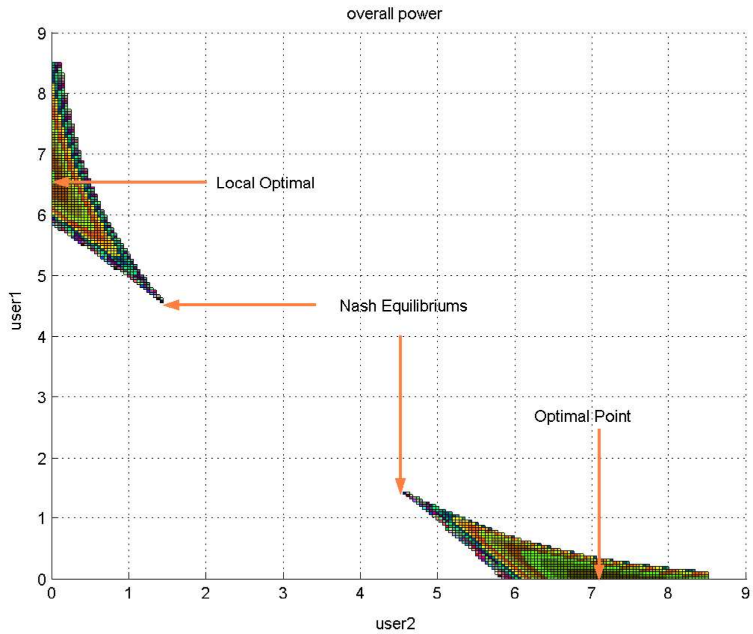


Figure 5.10: Two-user example: Multiple NEPs

interferences are considered as noises. By Lagrangian method, define $\nabla = \frac{\partial}{\partial \mathbf{r}_i}$, the following equation hold at the NEP when power is less than P_{max}.

$$\nabla(\sum_{l=1}^{L} P_i^l) - \mu_i \nabla(\sum_{l=1}^{L} r_i^l - R_i) = 0. \tag{5.23}$$

For the problem in (5.14), if $r_i^l > 0, \forall A_{il} \neq 0$ and $\sum_{l=1}^{L} P_i^l < P_{max}$, $\forall i$, the global optima will satisfy the KKT condition without considering the inequality constraints:

$$\sum_{i=1}^{K} \nabla(\sum_{l=1}^{L} P_i^l) - \sum_{i=1}^{K} \mu_i \nabla(\sum_{l=1}^{L} r_i^l - R_i) = 0. \tag{5.24}$$

Obviously, when the iterative water-filling converges, (5.24) will be satisfied from (5.23). So the KKT necessary condition is satisfied for NEP. ■

Resource Allocation Algorithm

Before developing the proposed algorithm, we analyze two extreme cases. In the first case, the groups of sub-channels are assigned to different cells without

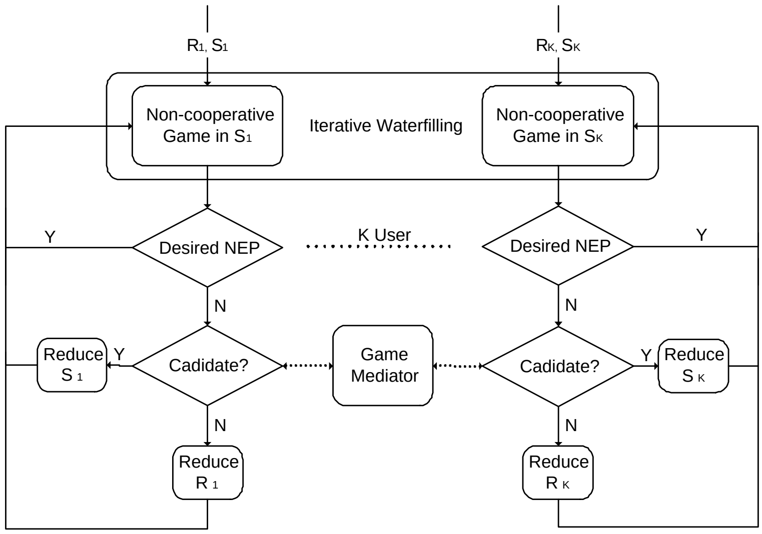


Figure 5.11: Noncooperative Game

overlapping such that there are no co-channel interferences among cells. We call it the fixed channel assignment scheme. However, this extreme method has the disadvantages of low spectrum efficiency because of the low frequency re-usage. The overall transmitted power in (5.14) solved by this method is far from minimum, because it doesn't take the advantage of the multiuser diversity and power control. In the second extreme case, all the users share all the sub-channels. We call it pure water-filling scheme. From Fig. 2 and Fig. 3, we can see that the system can be balanced at the undesired point, because of the severe inter-cell co-channel interferences. So the facts motive us to believe that the optimal resource allocation is between these two extreme cases, i.e., each sub-channel can be shared by only a group of users for transmission.

The basic idea of the proposed algorithm is to have a noncooperative game and if the game cannot converge to a good solution, a mediator is introduced on the sub-channel usage. Each user minimizes its own utility function, i.e. transmitted power, in a distributed game by applying water-filling. Then the system will be balanced in some NEP. If the co-channel interferences are not large, the NEP should be the desired solution. If the constraints of throughput and maximal transmitted power are not satisfied or NEP is not a local optimum, the co-channel interferences are too severe. From Theorem 2 and the previous observations, the system is probably balanced at a undesired solution. So a game mediator is needed to redefine the game, reducing the number of users that share the same sub-channels. We define the sub-channel set that the i^{th} user can allocate their throughput as transmission group S_i. In Fig. 4, we show the block diagram of the proposed algorithm from system point of view. We initially set S_i to have all the sub-channels. Then the noncooperative game is applied. When the system is iteratively balanced by the water-filling among users, we determine if the NEP is the desired one. If yes, we continue the water-filling. Otherwise, some user must remove some sub-channel from the transmission group. If the removal can make all users balanced in the desired NEP, the algorithm continues in the water-filling step. Otherwise, we continue the user removal step, until no user can be removed or the desired NEP is achieved. If no user can be removed and the desired NEP is still not achieved, we have to reduce the desired throughput requirement R_i.

The criterion for the user to remove a specific sub-channel is determined by the channel gain and the interferences plus noise level. If user i can not satisfy its constraints, the users who share the sub-channels in S_i will decide who will quit one channel. The channel with smaller channel gain and larger interferences plus noise will be selected, i.e., the j^{th} user will drop the l^{th} channel if

$$(l, j) = \arg\min_{l,j} \frac{P_j^l G_{jj}^l}{\sum_{k \neq j} P_k^l G_{kj}^l + N_0} \tag{5.25}$$

where $l \in S_i$ and user j shares a least one sub-channel with user i.

Table 5.3: *Distributed Resource Allocation Algorithm*

1. **Initialization**: R_i= predefined value, S_i includes all sub-channels
2. **Water Filling**:   each user have noncooperative game in (5.17).
3. **Desired NEP**:   if $\sum_{l=1}^{L} P_i^l < P_{max}$ and $\sum_{l=1}^{L} r_i^l = R_i$, not local   minimum on boundary, go to step 2;   otherwise, go to step 4.
4. **Sub-channel Removal/Throughput Reduction**:   remove sub-channel from transmission group by (5.25)   go to step 2. If no user can reduce his transmission   group, reduce R_i, go to step 2.

The criterion for whether or not the user can be removed from the transmission group is determined by three factors. 1) Each user must has at least one sub-channel to transmit. 2) No sub-channel is wasted, i.e., at least one user is assigned for each sub-channel. 3) User can not be kicked out from the sub-channel, if the user cannot transmit his throughput R_i using the rest of sub-channels, even though he occupies them alone.

The proposed distributed algorithm for each cell is shown in Table 5.3. In order to apply the proposed algorithm, we assume that base stations can accurately measure the channel gains and interferences plus noise power. Moreover the power and throughput allocation information can be reliably feeded back to mobiles without any delay. All these assumptions are reasonable for implementation in practice.

Simulation Results

To show the improvements of the proposed algorithm, we set up the simulations consisting of a two-cell case and a seven-cell case. In the two-cell case, one base station is situated at the center of each cell and one co-channel mobile per cell is generated as a uniform distribution within the corresponding cell for each simulation instance. The propagation model assumes the operation in a suburban environment and takes into consideration of path loss and shadowing. The received signal (in dB) at distance d from the base station is $L(d) = L(d_0) + 10\alpha \log_{10} \frac{d}{d_0}$, where $d_0 = 10$m is used as a reference point in measurements ($L(d_0) = 0$dB) and α is set to 3.5. Shadow fading for each user is modelled as an independent log-normal random variable with standard deviation $\sigma = 10$dB. The four-path Rayleigh model is taken into consideration to simulate the frequency selective fading channels, which has an exponential power profile with 100ns root-mean-square (RMS) delay spread. We consider a multi-cell OFDMA system with 32 sub-channels in total. The overall bandwidth is 6.4MHz. The total transmission power for every mobile is constrained by a maximal value of 10mW. The receiver thermal noise is -70dBm. The BER of the transmitted symbols is required to be 10^{-3} for every sub-channel and user, which corresponds to $c_3 = 0.2831$. We define the reuse factor R_u as the distance between two base stations D over the cell radius r. The smaller reuse distance, the more severe the co-channel interferences are.

In Fig. 5.12, we show the total transmitted power vs. rate constraint R_i for $R_u = 2$. Here we assume $R_i = R_j,\ \forall i, j$. When the rate requirement is increasing, the overall power is increasing. Compared with the fixed assignment algorithm, the proposed algorithm reduces about 80% of powers. This is because the fixed assignment algorithm wastes many resources by letting only one user

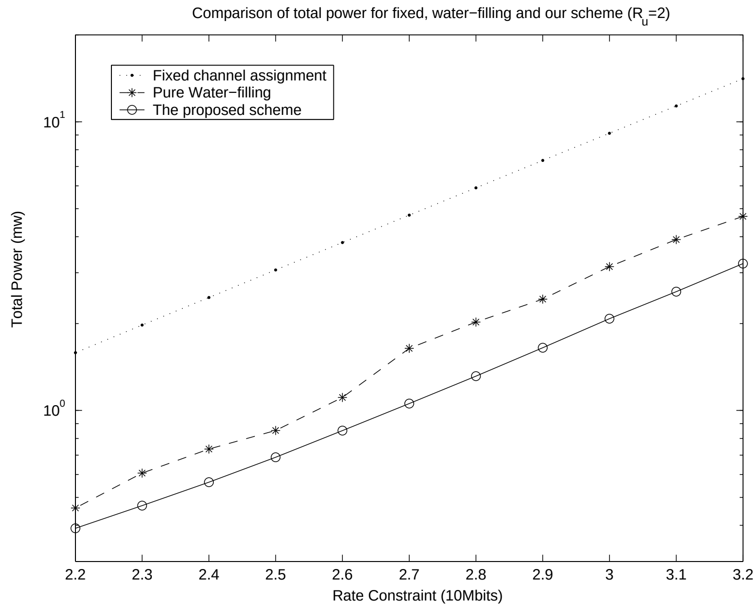


Figure 5.12: Total Power vs. Rate Constraint

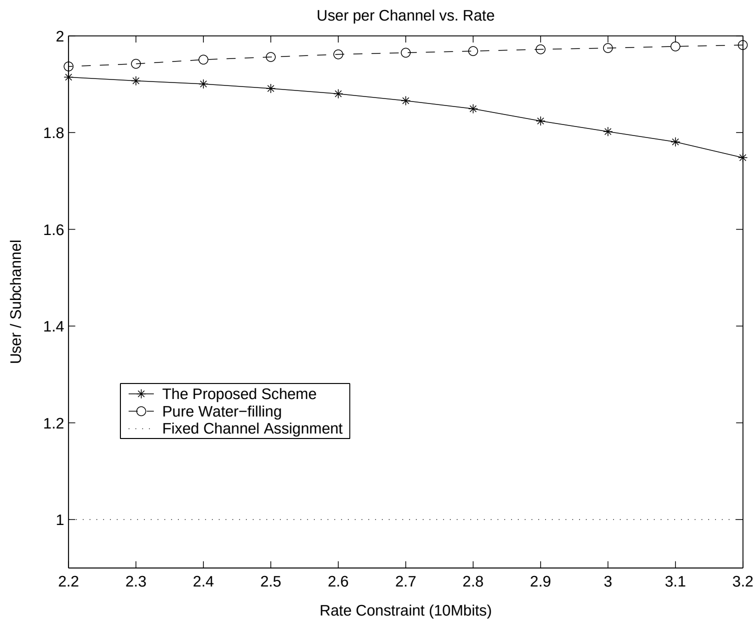


Figure 5.13: User per Channel vs. Rate Constraint

occupy any sub-channel. Compared with the pure water-filling algorithm, the proposed algorithm reduces about 25% of powers. The reduction is larger when the rate constraint is large. This is because some sub-channels cannot support more than one user especially when the rate constraint and co-channel interferes are large.

In Fig. 5.13, we show number of users per channel vs. rate constraint. The fixed channel assignment algorithm always has only one user per channel. The proposed algorithm has lower user per channel and the pure water-filling algorithm has higher user per channel when the rate requirement is larger. For pure water-filling algorithm, some sub-channels may not have allocated powers when the rate constraint is small, because of the low water-filling level. For the proposed algorithm, more users are kicked out from using the sub-channel when the rate constraint is large.

Seven-user simulation is setup as shown in Fig. 5.14. One cell is located in the middle and the other six cells are located at the angle of $[0, 30, 90, 150, 210, 270]$. The cell radius is $r = 100$m. The rate constraint is 10Mbits for each user. The other settings are the same as two-cell case.

In Fig. 5.15 and Fig. 5.16, we show the overall transmitted power and users per channel vs. reuse distance for the pure water-filling algorithm and the proposed algorithm, respectively. We can see that the proposed algorithm can reduce the overall power about 90% when the co-channel interferences are severe ($R_u = 2$), which will greatly improve the system performance. The proposed scheme kicks more users out and reduces number of users per sub-channel. When R_u is increasing, the co-channel interferences are reduced. Consequently, two schemes shows the similar overall transmitted power and user per channel.

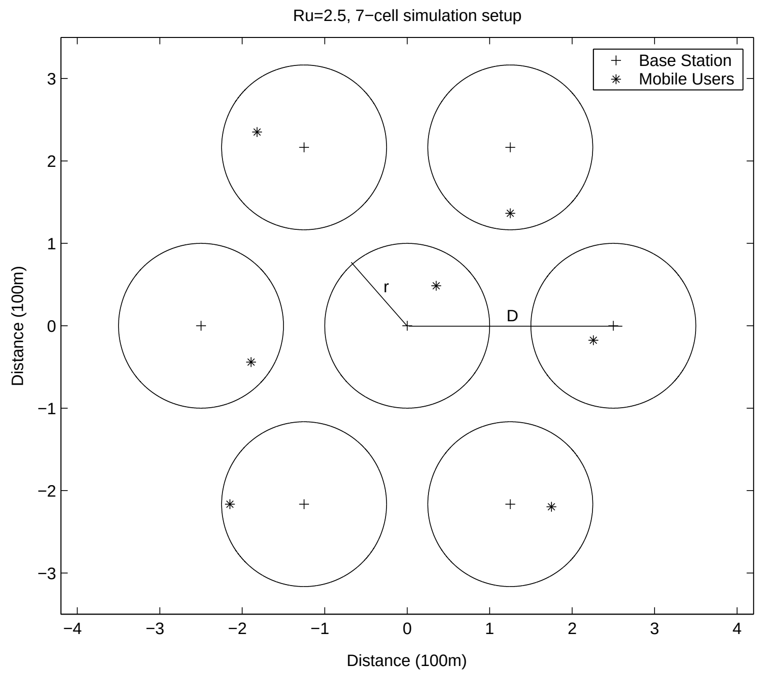


Figure 5.14: Simulation Setup

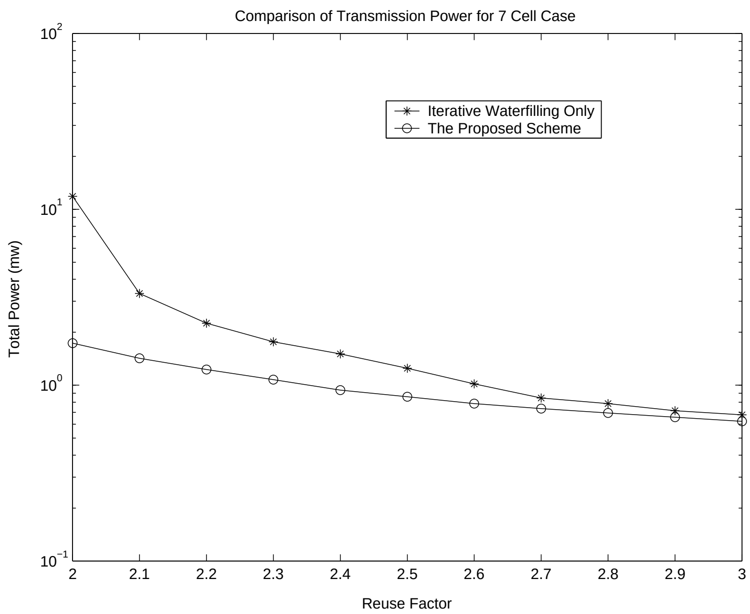


Figure 5.15: Overall Power vs. R_u for Multicell Case

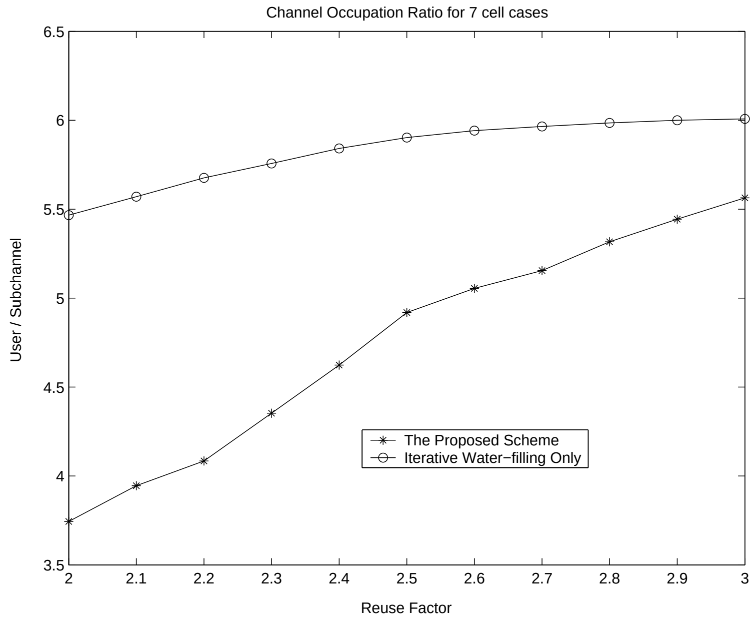


Figure 5.16: User per sub-channel vs. R_u for Multicell Case

5.4 Subspace Approach

Capacity optimization in a multi-cell OFDMA system where each cell has multiple users is investigated in this work. The objective is to find an assignment of users to the sets of subcarrier, their transmission rates for the subcarrier, and power allocation such that the total system capacity is increased, while users meet a minimum total rate requirement and a power constraint. Since the optimal solution involves an exhaustive search or complex nonlinear integer programming, we develop sub-optimal low complexity algorithms. We propose a two-step scheme: First an initial channel and data rate allocations are determined by two initialization algorithms. Then we refine the assigned rates by an iterative algorithm. From the preliminary simulation results, the proposed algorithms can efficiently allocate resources to increase the overall system capacity and reduce the allocation outages.

This section is organized as follows: First, we give the motivation and sketch for the proposed scheme. Then, we present the system model and problem definition. We propose two initialization algorithms. We develop an iterative capacity refinement algorithm by using subspace methods. Preliminary numerical studies are included.

Motivation and Sketch

In multiuser wireless systems, users to users channel variations, due to location differences and fading in time and frequency, can be utilized to improve system capacity. By assigning bandwidth according to users' channel responses, spectral efficiencies can be improved. This technique, which is known as multiuser diversity, in Orthogonal Frequency Domain Multi-Access (OFDMA) can be utilized over time and frequency. However, in order to maintain the basic link qualities, the allocation algorithm has to efficiently utilize the bandwidth to increase system capacity and at the same time meet the minimal data rate requirements of different users.

This problem has been of interest recently. In [71, 76], in single cell systems, suboptimal algorithms were proposed such that total transmit power was minimized and a minimum rate requirement for each user was to be satisfied. In [72], a suboptimal simple algorithm was proposed for single cell case. In [90], a similar problem in a single cell system was formulated as max-min user throughput optimization under a maximum transmit power policy. In [87], the objective was defined as maximizing total system throughput, in a multi-cell system, while the transmit power per user was limited. In that work, a suboptimal water-pouring based algorithm was proposed to solve that problem. A number of heuristic algorithms were proposed in [88] to find feasible channel assignments and transmit power allocation in multi-cell systems. Most of the previous works concentrate on either single cell channel assignment problem or multi-cell power control problem. Very few works address the multi-cell OFDMA resource allocation where each cell has multiple users, which is a very difficult high dimension assignment and non-linear problem. This motivates us to study this problem and try to find a possible solution.

In this work the problem of capacity optimization by dynamic allocation of subcarrier to users in a multi-cell OFDMA network is investigated. The objective is defined as to maximize the overall system capacity while a minimum rate requirement for each user can be satisfied and the transmitted power is constrained. Since this problem is NP hard, we propose a two-step suboptimal scheme. In the initialization step, we develop two algorithms: First, we start from an equal rate channel allocation across users; In another approach, we start from a maximum packing solution. In the refinement step, we improve system capacity by an iterative algorithm. Through the preliminary numerical studies, we will show that the proposed algorithms can efficiently allocate the resource to increase the system capacity and reduce the outage probability when the system cannot allocate the minimal rate requirements for all users.

System Model and Problem Definition

Assume there are N cells in the system and the i^{th} cell has M_i mobile users. There are totaly K subcarrier in the system. Within each cell, only one user is allocated to each subcarrier. Among different cells, multiple users share the same subcarrier. The allocations of users and powers to subcarrier are denoted by $K \times N$ matrices $\mathbf{A}$ and $\mathbf{P}$, respectively. $[\mathbf{A}]_{ki} = A_{ki}$ represents user number j that occupies the k^{th} subcarrier in the i^{th} cell. $A_{ki} \in [1, \ldots, M_i]$. $[\mathbf{P}]_{ki}$ is this user's power. For the uplink case, in the i^{th} cell, the j^{th} user occupies the k^{th} subcarrier, i.e., $A_{ki} = j$. The received SINR for uplink is given by:

$$\Gamma_i^k = \frac{P_i^k G_{ii}^k}{\sum_{l \neq i} P_l^k G_{li}^k + N_0} \tag{5.26}$$

where P_i^k is the transmit power from i^{th} cell for the k^{th} subcarrier, G_{li}^k is the interferer's propagation loss from the l^{th} cell to the i^{th} cell for the k^{th} subcarrier, and N_0 is the sampled thermal noise level. Without loss of generality, we assume the noise level is the same for all users. Suppose a target SINR γ_i ($\Gamma_i \geq \gamma_i$), in matrix form, we have

$$(I - \mathbf{D}^k \mathbf{F}^k)\mathbf{P}^k = \mathbf{u}^k \tag{5.27}$$

where $\mathbf{P}^k = [P_1^k, \ldots, P_N^k]^T$, $\mathbf{D}^k = \text{diag}(\gamma_1^k, \ldots, \gamma_N^k)$, $\mathbf{u}^k = [u_1^k, \ldots, u_N^k]^T$, $u_i^k = \gamma_i N_0 / G_{ii}$, and

$$[\mathbf{F}_{ij}^k] = \begin{cases} 0 & \text{if } j = i, \\ \frac{G_{ji}}{G_{ii}} & \text{if } j \neq i. \end{cases} \tag{5.28}$$

The above equation has a solution with possible power vector, if the spectral radius (the maximal eigenvalue) of $\rho(\mathbf{D}^k \mathbf{F}^k)$ is inside unit circle [67].

We assume that the channels change slowly and are stable over a frame with hundreds of symbols. Assume $A_{ki} = j$, the capacity is denoted by

$$c_{ij}^k = W \log(1 + \frac{\Gamma_i^k}{\Gamma}), \tag{5.29}$$

where Γ is a constant for capacity gap and W is the bandwidth. Without loss of generality, we assume $W = 1$.

The goal of our approach is to maximize the system overall capacity. Each user has a minimal rate requirement R_{ij} when he is admitted to the system. In practice, the transmitted power of each user is bounded by P_{max}. The users will water fill their powers to the carefully assigned channels according to the channel responses, interferences, and noises. This will involve complicated channel assignment and high dimension nonlinear optimization. In [64], it has been shown that the power is closely related to the spectral radius of $\mathbf{D}^k\mathbf{F}^k$. In our approach, to simplify the problem, we let the spectral radius to be bound by $1-\epsilon$, where ϵ is a small number. We can carefully select its value, such that the power constraint is satisfied. The constrained optimization problem can be expressed as:

$$\max_{\mathbf{A},\mathbf{P}} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \sum_{k=1}^{K} c_{ij}^k \tag{5.30}$$

$$s.t. \begin{cases} \text{Rate:} & c_{ij} = \sum_{k=1}^{K} c_{ij}^k \geq R_{ij},\ \forall i, j. \\ \text{Power:} & |\rho(\mathbf{D}^k\mathbf{F}^k)| \leq 1-\epsilon,\ \forall k. \end{cases}$$

Since finding the optimal solution to the problem in (5.30) directly is extremely complicated and may involve complicated nonlinear large dimension integer programming or even exhaustive search. For example, by using Monte Carlo method with multiple initializations or simulation annealing, we can achieve some local optima or even global optimum. However the complexity is too large even for performance analysis. So we try to solve it in two steps to reduce the complexity. In the first step, we initialize the resource allocation by fast suboptimal algorithms to allocate channels and powers. In the second step of refinement, for each subcarrier, we develop an iterative algorithm to increase the system capacity subject to the minimal rate and power constraints per user.

Initialization Algorithms

We present two algorithms for initializing resource allocations. In the first algorithm, we find a channel assignment that maximizes the equally achieved rate for users. In the second approach, we pack an initial set of users plus their channel and rate assignments such that total system capacity is optimized. Using any of the allocation schemes, we enhance the system capacity by an iterative algorithm in the following part.

Table 5.4: Initialization Algorithm A

1. For each subcarrier k from the pool of available subcarrier, search over the possible system to maximize the spectral radius of the matrix $\mathbf{F}^k$, i.e., $\gamma^k = \max_{\mathbf{A}} (1-\epsilon)/\rho(\mathbf{F}^k)$.
2. In the set of available subcarrier, start from the subcarrier with the best SINR and assign equal rates (a function of SINR) to assigned users in that carrier, and remove the subcarrier from the search.
3. Remove users that achieve the desired rate from the search.
4. Continue with a new subcarrier in Step 1, until all users have the minimal rate requirement.
5. Allocate the rest of subcarrier in a greedy way.

Equal SINR/Rate Allocation

In the first algorithm, we consider a system where each base station allocates one user to each subcarrier. The objective for user assignment is to select one user from each cell for each subcarrier and form the best set of users that maximizes capacity. For the case of equal SINR allocation to all users, this problem is equivalent to finding the best allocation of users that minimizes the spectral radius of the gain matrix, $\mathbf{D}^k\mathbf{F}^k$, for each carrier, i.e.,

$$\widehat{\mathbf{A}}^k = \arg\min_{\mathbf{A}^k} \rho(\mathbf{D}^k\mathbf{F}^k) \tag{5.31}$$

where $\mathbf{A}^k$ is the k^{th} column of $\mathbf{A}$ which consists of the indices of users allocated in different cells for the k^{th} subcarrier.

The optimal solution finds the best user for each subcarrier, the maximum equal SINR for each allocation, and the power allocation to achieve the maximum SINR. The power allocation can be calculated from (5.27). The difficulty is to find the best user assignment for each subcarrier, which involves an exhaustive search over all users. Here we present a suboptimal approach to find the user allocation. We find the user that minimizes the link gain for each subcarrier, i.e.,

$$\widehat{A}_{ki} = \arg\min_{A_{ki}} 1/G_{ii}^{k}. \tag{5.32}$$

After the above user allocation, we find the maximal achievable γ^k for this subcarrier. Then we try to find the best allocation for the next subcarrier. If a user's minimum rate requirement is satisfied, this user is excluded from further resource allocation, until all the users have their minimal rate requirements. Finally, the rest of the subcarrier is greedily allocated to the users with the best channel conditions. The algorithm is shown in Table 5.4, where ϵ is a small number. The maximal transmitted power P_{max} will determine its value and $\rho \leq 1-\epsilon$. This algorithm can be implemented in a distributive manner with limited communications between base stations.

Maximal Rate Packing

In the second algorithm, we find the best set of users for each subcarrier and each subcarrier is not necessarily occupied by all base stations. The basic idea is to pack each subcarrier with the best users in the networks as long as the capacity is increasing.

First, the algorithm finds the highest SINR user and subcarrier in the networks. This user maximizes the channel capacity for this subcarrier. Then we add users one by one to share the subcarrier. If adding users does not improve the total capacity for this subcarrier, the assignment is stopped and we continue for the remaining subcarrier. If any user is allocated more than its desired rate, he will be removed from the future optimization list. The algorithm continues until all users have the minimal rate requirements. When the algorithm is not able to find a solution, due to lack of resources, we report an outage. Otherwise, the rest of the subcarrier is assigned by a greedy method where we pack the subcarrier with the same method above, but users are not removed from the list, such that the total network capacity is increased.

In this approach, the maximal data rate is packed in the network for each subcarrier independent of their cells. That means some base stations may or may not assign a specific subcarrier to their users. In using a subcarrier, the base station sacrifices for the other cells with the hope that the other base stations will run out of users and reduce interferences in other subcarrier. The algorithm is shown in Table 5.5. To implement this algorithm, we need a centralized control and sufficient channel estimations. So the algorithm fits the situation where the number of cells is small and channel changes slowly.

Capacity Refinement Algorithm

We have presented initialization algorithms for the channel and power allocations for different subcarrier for different users in different cells. We will develop a two-step iterative algorithm to refine the allocation such that the system overall capacity can be improved under the rate and power constraints. In the first step, we improve the system feasibility. We find the gradient $\partial\rho(\mathbf{D}^k\mathbf{F}^k)/\partial\Gamma_i^k$ for the k^{th} subcarrier and then project this gradient onto the plane where the overall capacity for this subcarrier is fixed. Then we move along this modified gradient so that

Table 5.5: Initialization Algorithm B

1. Evaluate maximum data rate for each user at its maximum transmission power for each subcarrier. Start with the subcarrier and user with the highest rate
2. Add one user at a time and find the best user and assigned power that maximizes the total capacity.
3. Repeat Step 2 until total capacity does not increase by adding users.
4. Repeat from first step with a new subcarrier, until all users satisfy their rate requirement. If no solution, report an outage.
5. Allocate the rest of subcarrier in a greedy way.

$\rho(\mathbf{D}^k\mathbf{F}^k)$ is reduced, while the overall capacity of this subcarrier is maintained the same. In the second step, we increase each subcarrier's SINR for different users to increase the system performance until the system is almost infeasible. Here we consider the users whose rates are less than R_{ij} first, because their rates may be impaired by the first step. The two steps are executed iteratively to improve the system capacity. The iteration stops when reaching the boundary or some stable point.

In the first step, we find the gradient first. It has been shown that the existence of the derivative of the spectral radius $\rho(\mathbf{D}^k\mathbf{F}^k)$ by the following theorem [91].

Theorem 5.4.1 *let λ be a simple eigenvalue of $\boldsymbol{DF}$, with right and left eigenvectors $\mathbf{x}$ and $\mathbf{y}$, respectively. let $\widetilde{\boldsymbol{F}} = \boldsymbol{DF} + \boldsymbol{E}$, where $\boldsymbol{E}$ is a small perturbation. There exists a unique $\widetilde{\lambda}$, eigenvalue of $\widetilde{F}$ such that*

$$\widetilde{\lambda} = \lambda + \frac{\boldsymbol{y}^H \boldsymbol{E}\boldsymbol{x}}{\boldsymbol{y}^H \boldsymbol{x}} + \mathrm{O}(\|\boldsymbol{E}\|^2) \tag{5.33}$$

Proof [67] ∎

In our application, we only try to reduce the maximum absolute eigenvalue. Let $\mathbf{x}^k$ and $\mathbf{y}^k$ be the eigenvectors of the largest eigenvalue. Define $\mathbf{E}^k = \Delta\Gamma_i^k \mathbf{F}_i^k$, where

$$(\mathbf{F}_i^k)_{jl} = \begin{cases} 0, & j \neq i; \\ (\mathbf{F}^k)_{jl}, & j = i. \end{cases} \tag{5.34}$$

We can have the gradient to reduce spectral radius as:

$$g_i^k = \frac{\partial \rho(\mathbf{D}^k\mathbf{F}^k)}{\partial \Gamma_i^k} = \frac{(\mathbf{y}^k)^H \mathbf{F}_i^k \mathbf{x}^k}{(\mathbf{y}^k)^H \mathbf{x}^k}. \tag{5.35}$$

If we change each user's SINR, according to above gradient $\mathbf{g}^k = [g_1^k \dots g_N^k]^T$, the capacity of each subcarrier will be reduced. In [64], we project the gradient to a plane where the overall capacity is a constant. The plane, that is tangent to the curve where overall capacity is equal, can be expressed as:

$$\sum_{i=1}^{N} b_i x_i = C, \tag{5.36}$$

where $b_i^k = 1/(1+\Gamma_i^k)$ and C is a constant. The projected gradient is $\mathbf{h}^k$, i.e the gradients in (5.35) projected onto the plane in (5.36). However if we move along this gradient, some users' rates may be reduced below the minimal rate requirement. We will compensate back the rates in the second step. The first step is stopped when the resource allocation falls to a local optimum or hits the boundary, i.e., some user' SINR for some subcarrier is reduced to zero.

In the second step of the iterative algorithm, we will optimize the overall system capacity. The gradient of overall capacity with respect to targeted SINR is given by:

$$q_i^k = \frac{\partial c_i^k}{\partial \rho(\mathbf{D}^k \mathbf{F}^k)} = \frac{\partial c_i^k}{\partial \Gamma_i^k} / \frac{\partial \rho(\mathbf{D}^k \mathbf{F}^k)}{\partial \Gamma_i^k} = \frac{1}{g_i^k(\Gamma + \Gamma_i^k)}. \tag{5.37}$$

We will change the SINRs of the users whose rates are below the minimal requirements first, while keeping other users' SINR fixed, until all users' requirements are satisfied. Then we increase SINR of all users' according to this gradient to increase the overall system capacity, until we hit the boundary, i.e. $\rho(\mathbf{D}^k \mathbf{F}^k) = 1 - \epsilon$.

We repeat the above two steps until the results are stable. We observe that the second step may stop when the constraints are not satisfied. If the results satisfy the minimal rate constraint, we will return these results, otherwise we will return the results in the previous iteration. Then the channel and power allocation is selected for different users. The whole algorithm is operated within the feasible region and the solution is on the boundaries. Define μ and μ' as small constants, their values determines the converges speed and the accuracy of the final results. Since the algorithm is initialized with a feasible solution and users' targeted SINRs

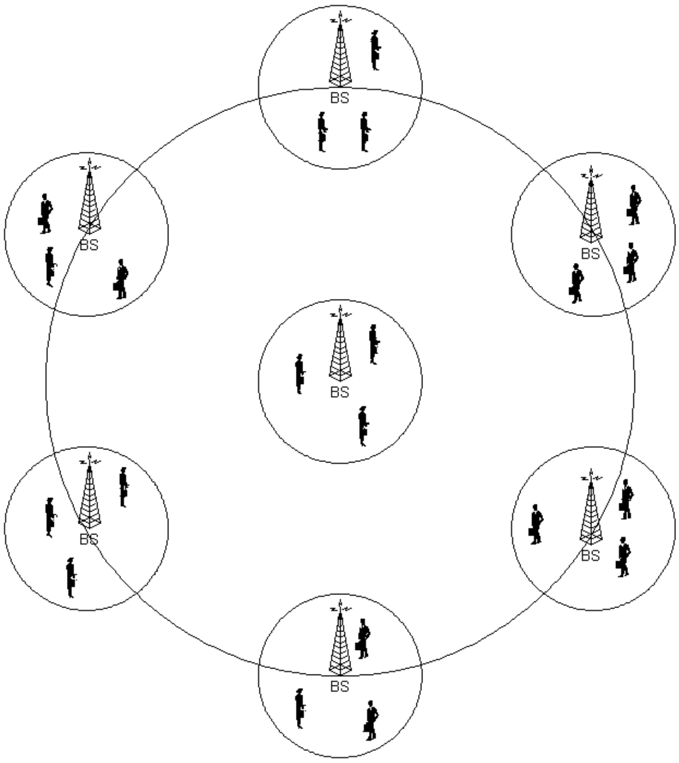


Figure 5.17: Simulation Setup

are modified within the feasible range, the proposed algorithm always converges. The iterative capacity Improvement algorithm is given in Table 5.6.

Similar to initialization algorithm B, the improvement algorithm needs a centralized control and many channel estimations. So it only fits small scale systems. Moreover, users' minimal rate may be reduced in the first step of the improvement algorithm and cannot be compensated back in the second step. Under this condition, there are outages when the minimal rate requirement is not satisfied, and we just simply switch back to the original settings determined by the initialization algorithms.

Simulations

In order to evaluate the performances of the proposed algorithms, a network

Table 5.6: Iterative Capacity Improvement Algorithm

```
Initialization:

  By either the initialization algorithms.
```

Iteration: Stop when Γ_i^k stable

1. $\rho(\mathtt{D}^k\mathtt{F}^k)$ Reduction:

do {

$\mathtt{g}^k = \bigtriangledown\rho(\mathtt{D}^k\mathtt{F}^k)$;

$\mathtt{h}^k = \mathtt{projection}(\mathtt{g}^k)$;

$\Gamma_i^k = \Gamma_i^k - \mu'.\mathtt{h}_i^k \ \ \forall \ \ i$;

while $(\Gamma_i^k$ not stable or not at boundary$)$

2. Capacity Improvement

do {

if $\exists j, \sum_{k=1}^{K} c_{ij}^k < R_{ij}$

$\Gamma_i^k = \Gamma_i^k + \mu.q_j^k, \forall k,$

$\Gamma_i^k = \Gamma_i^k,$ for other users;

otherwise

$\Gamma_i^k = \Gamma_i^k + \mu.q_i^k, \forall i.$

while $(\rho(\mathtt{D}^k\mathtt{F}^k) \leq 1 - \epsilon)$

Report the allocation results.

Channel assignment and Power update.

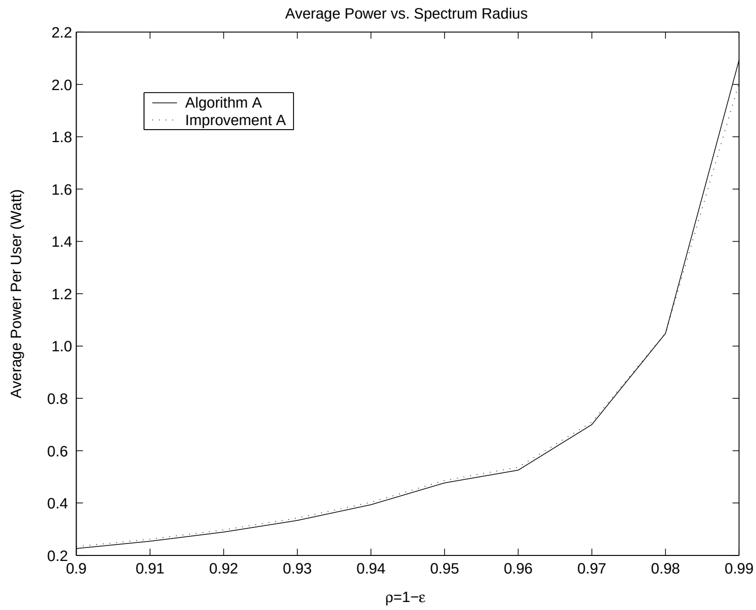


Figure 5.18: Average Power vs. Spectral Radius ρ

with $N = 7$ is simulated in Fig. 5.17. One base station is located as the center of each cell. Each cell has the same number of users $M_i = 3, \forall i$ and all the users are randomly located within each cell. The total number of subcarrier is $K = 32$ and each subcarrier is assumed to have unit bandwidth. Each cell's radius is 200m. The distance between base stations over the cell radius is 3. The maximal power is $P_{max} = 0.5$Watts. $\Gamma = 1$. Each user has the same thermal noise level -80dBm. The propagation loss factor is 3.5. The maximal doppler frequency shift is 100Hz and four-path frequency selective Rayleigh fading channels are simulated, which has an exponential power profile with 100ns root-mean-square (RMS) delay spread.

In Fig. 5.18, we show the average power per user vs. spectral radius. We show the result of algorithm A and improvement algorithm for algorithm A. We can find that even though the rate and power allocations are quite different, for the same spectral radius constraint, the average powers of two schemes are almost the same. This means that it is reasonable to replace the power constraint by the spectral

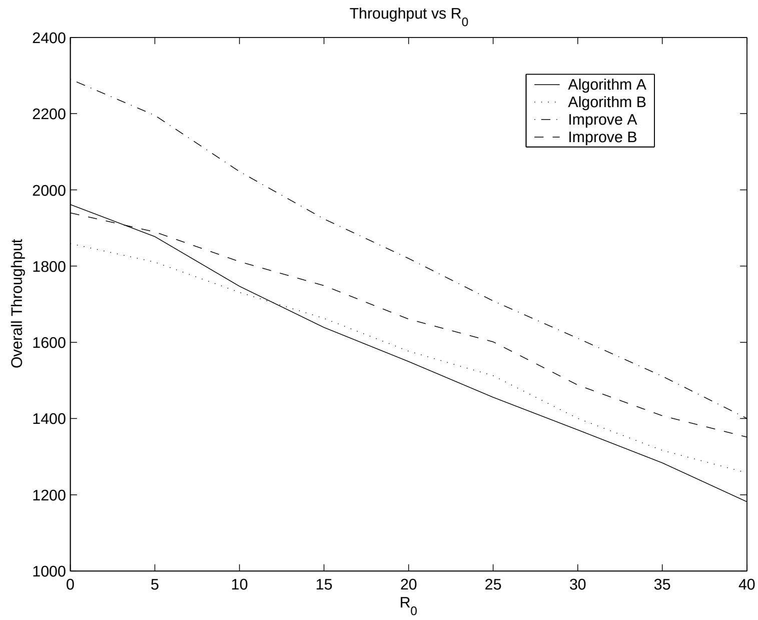


Figure 5.19: Overall Capacity vs. Minimal Rate R_0

radius constraint in the problem formulation in (5.30). The powers increase fast when the spectral radius approaches 1. We select $\rho = 0.95$ in our simulations such that the power constraint is limited to less than 0.5Watts.

In Fig. 5.19, we show the overall system capacity vs. the minimal rate requirement R_0 for each user for algorithm A, algorithm B, improved algorithm A, and improved algorithm B. We can see that the overall capacity is reduce when R_0 is increasing for all algorithms. This is because the system is more fair and has to give more resources to the users with bad channels. Algorithm B has slightly better performance than algorithm A when R_0 is large. This is because each subcarrier is optimally occupied by users. Algorithm A has a little bit better performance than algorithm B when R_0 is small. This is because when algorithm B satisfies most of users' minimal rate, the last few users will waste the resources because there are no other users that can share the subcarrier, while a large number of users get the minimal rate at the same time by the equal SINR algorithm A. The improvement algorithm can improve the performance of both algorithm A and algorithm B, while the improvement for algorithm A is much larger than that for algorithm B. This is because each carrier is occupied by much more users for algorithm A than for algorithm B. Consequently, the improvement algorithm can have much more room to reduce spectral radius and increase the overall capacity iteratively.

In Fig. 5.20, we show the outage percentage (the ratio of the number of users that cannot be satisfied with the minimal rate over the total number of users) vs. R_0. We can see that the outage percentage increases when R_0 is increasing. Algorithm B has much lower outage rate than algorithm A. This is because algorithm B can pack more rate for each subcarrier. Improvement algorithm can reduce the outage rate when R_0 is large. But when R_0 is not large enough, the improvement algorithm has higher outage rate. Under this condition, we will switch back to the original solution of the initialization algorithms.

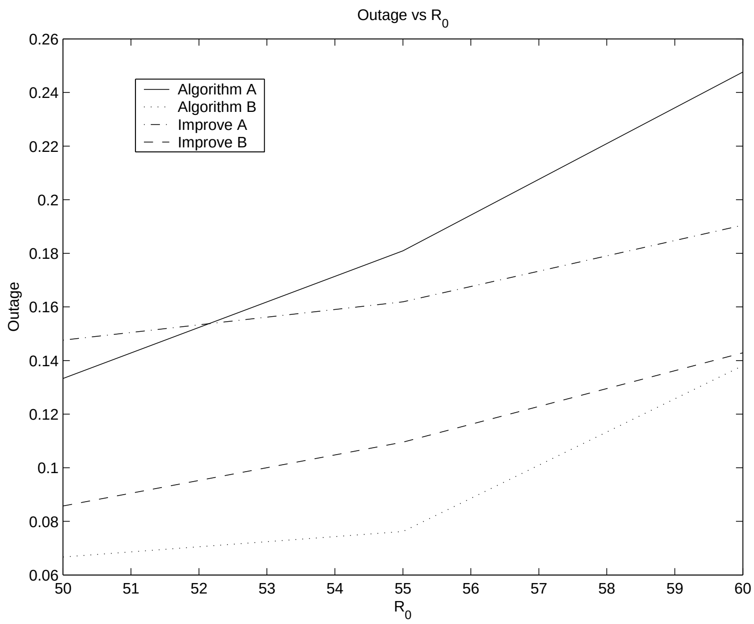


Figure 5.20: Outage vs. Minimal Rate R_0

Chapter 6

Cross Layer Approaches for Multiuser Communications

Current wireless networks are designed in layers according to OSI reference models. Each layer has its own design issue. For example, error control for physical layer, flow control for MAC layer, routing for network layer, and source coding for application layer. Each layer optimizes its own goal and the design can hardly be optimal from system point of view. Because of the increasing demand of wireless communication and limited bandwidth, it is more and more necessary for the system designer to implement more efficient protocols. Since there exists direct coupling between layers, it is natural to optimize the system performance by cross layer approach.

In this chapter, we briefly give the introduction on cross layer approach. We will give the motivations and possible implementation methods. Then we will give two examples for cross-layer approach. The first one is the joint source channel coding plus power control for multiuser communication. The second one is the join power control and blind beamforming.

6.1 Motivations

Future wireless networks will provide ubiquitous communications among people and devices. Services such as wireless Internet access, N^{th} generation cellular network, wireless Ad Hoc networks, sensor networks, wireless entertainment, smart homes/spaces, and automated highways are emerging, which demands great efforts for new wireless network design.

The challenges exists for design of new networks. Wireless channels are a difficult and capacity limited broadcast communication medium. Traffic patterns, user locations, and network conditions are constantly changing. Applications are heterogeneous with hard constraints that must be met by the networks. Energy and delay constrains change design principles across all layers of the protocol stack. Cross layer approach is a good way to handle all these challenges.

With advance of technology, the system performances have been enhanced in different layers. For hardware, better batteries and better circuits/processors are available. To maintain link quality, antenna array processing, adaptive modulation and coding, and advanced DSP techniques are implemented in recent years. Dynamic resource allocation and mobility support enhance the network layer design. In application, soft and adaptive QoS are taken into considerations. All these techniques improve the system performance. However there are some fundamental tradeoffs such as rate vs. coverage, delay, cost, and energy. So fundamental design breakthroughs are needed in the next generation wireless network designs.

The design objective for cross layer approach is to provide end-to-end QoS. The challenge for this QoS provision is the system dynamics. Many techniques can be applied to combat these dynamics. For example, scheduling can help shape these dynamics; adaptivity can compensate for or exploit these dynamics; diversity provides robustness to unknown dynamics. Also energy must be allocated and many constraints are implemented across all layers during the cross layer design. Some techniques are listed as

1. Adaptive techniques

    - Link, MAC, network, and application adaptation
    - Resource management and allocation (power control)
    - Synergies with diversity and scheduling

2. Diversity techniques

    - Link diversity (antennas, channels, etc.)
    - Access diversity
    - Route diversity
    - Application diversity
    - Content location/server diversity

3. Scheduling

    - Application scheduling/data prioritization
    - Resource reservation
    - Access scheduling

Some key questions are remained for cross layer approach, which gives us motivations for research: What is the right framework for cross layer design? What are the key cross layer design synergies? How to manage its complexity? What information should be exchanged across layers, and how should this information be used? How do the different timescales affect adaptivity? What are the diversity versus throughput tradeoffs? What criterion should be used for scheduling? How to balance the needs of all users/applications? How to implement distributed control over wireless networks?

On the whole, cross layer design needs to meet requirements and constraints of future wireless networks. Key synergies in cross layer design must be identified. The design must be tailored to the application. Cross layer design should include adaptivity, scheduling, and diversity across protocol layers. Energy can be a precious resource that must be shared by different protocol layers.

6.2 Multimedia Transmission over Wireless Networks

In this section, we will first briefly review joint source channel coding. Then we discuss the proposed downlink resource allocation for multimedia CDMA networks.

6.2.1 Joint Source Channel Coding

Shannon's separation theorem states that source coding (compression) and channel coding (error protection) can be performed separately and sequentially, while maintaining optimality. However, this is true only in the case of asymptotically long block lengths of data. In many practical applications, the conditions of the Shannon's separation theorem neither hold, nor can be used as a good approximation. For example, in real-time communication like videoconferencing, any delay greater than 100ms is not tolerable. Joint source-channel coding takes advantage of this fact and jointly optimize the source-channel coders when the assumptions

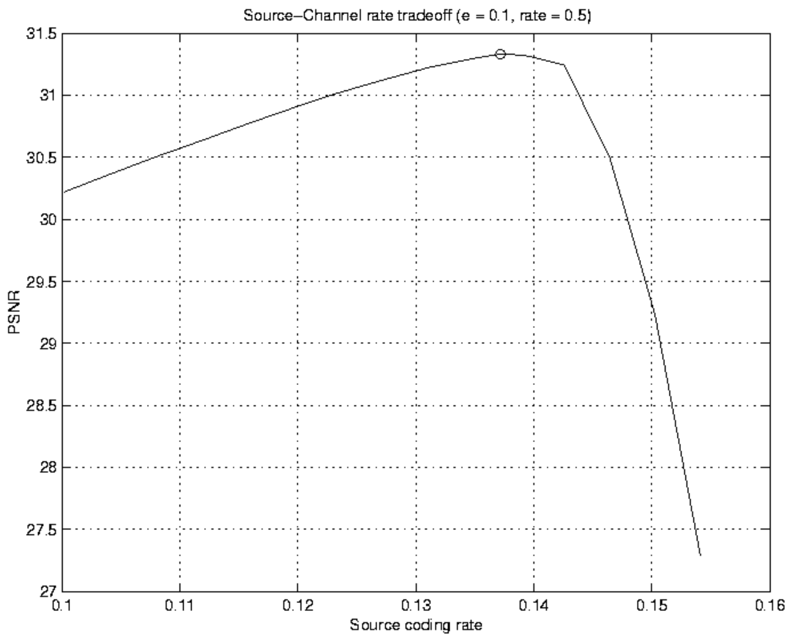


Figure 6.1: An Example of Tradeoff for Source Rate and Channel Rate

are invalidated, and thus achieving performance gains. So considerable interest has developed in various schemes of joint source-channel coding.

The easiest example is a channel capacity-limited video communication system: both the source and channel coders need bits, spending more bits on the source means not enough channel protection, which leads to channel errors, received video quality is bad; spending more bits on the channel means enough protection and no transmission errors, but then you have overcompressed the source material and received video quality is again bad. There is a trade-off and balance point where the channel capacity is optimally allocated between source and channel to achieve the best received video quality. In Fig. 6.1, we show an example of tradeoff for source rate and channel rate, where the total transmission bandwidth is fixed. We can see that the received PSNR will increase with the increasing of source rate. This is because the source introduced errors are reduced. But when the source

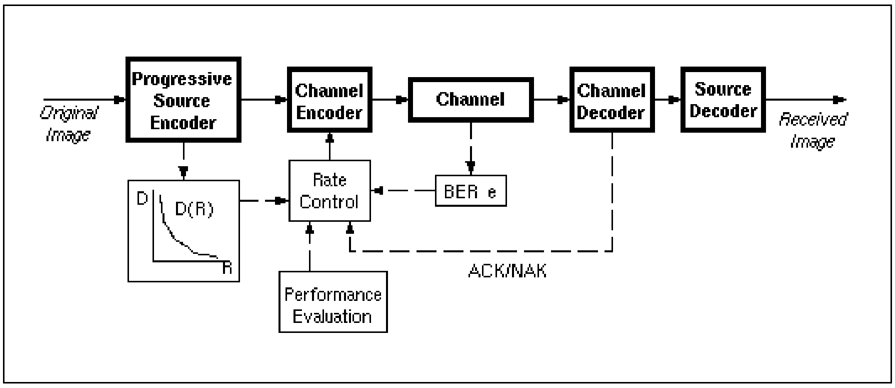


Figure 6.2: Block Diagram for a Typical Joint Source Channel Coding

rate is larger than some threshold, the reconstructed images quality drop quickly. This is because the channel introduced errors are dominant.

The joint source channel coding research concentrates a low-complexity mechanism for the determination of rate allocation for source-channel coding of progressive sources. The system diagram is shown in the Fig. 6.2, where the source and channel coder can be any family of channel codes and any progressive source coder. In addition, a rate-compatibility condition on the error control code can be applied. The rate control unit solves the optimization problem to devise an efficient and fast solution.

The challenge for the join source channel coding is how to model the problem over wireless network and how to develop a fast and efficient algorithm to allocate the source rate and channel rate. In the following section, we propose a multiuser joint source channel coding system with power control over CDMA networks.

6.2.2 Multiuser Cross Layer Approach

In interference-limited CDMA networks, the maximal number of users of real time applications can be increased by smoothly increasing the end-to-end distortions. In our approach, each user can accept a range of carefully controlled distortions. We develop a system protocol to control each user's distortion by adapting the resources like source coding rates, channel coding rates, transmit rates, and transmitted powers. The formulated problem is to reduce the overall system distortion in downlink single-cell systems, under the constraints of users' maximal distortions and the maximal transmitted power from the base station. In order to solve such a difficult problem, inspired by an event of daily life, we develop a heuristic and fast algorithm to allocate these resources for different users. The idea is to initialize the resource allocation with the maximal distortion for all users, and then allocate the remaining transmitted power quota first to the user who can most easily be satisfied for reducing its distortion. This user must have a high distortion, have a good channel condition, or generate small interferences to others. The allocation process is continued, until the transmitted power is used up. A performance bound is developed. We also analyze the dynamic system model with different arrival rates, holding times, and speech activities. From the simulation results, the proposed algorithm fundamentally reduces the distortions and the necessary maximal transmitted power when the number of users is large, compared with a traditional voice over CDMA scheme (with no distortion control).

The organization of this subsection is as follows: First, we have the motivation and sketch for the proposed scheme. Then the system model is given. The cross-layer protocol for voice transmission is described. The problem is formulated and the proposed algorithm is developed. A performance bound is also developed. We evaluate the performance for dynamic system. Simulations studies are presented.

Motivation and Sketch

In Code Division Multiple Access (CDMA) systems, all users transmit simultaneously over the same frequency band by using different spread spectrum codes. Because perfect separation between codes is not achievable under real wireless channels, the capacity and the maximal number of users are limited by interferences. Resource allocation such as rate adaptation and power control is an important means to combat the interferences, increase the number of users, and maintain the received signals' qualities. In joint source channel coding, rate adaptation by modifying the source rates, channel coding rates, and transmit rates can adjust the source encoders' output qualities and the protections for channel errors. Consequently, the reconstructed signals' qualities can be carefully controlled, according to the channel conditions. Power control is a technique to maintain the received SINR. So the problem is how to increase the system performance, by cleverly allocating these resources under some practical constraints.

Distortion based CDMA resource allocation is a hot topic in literature. In [111, 112], packet reservation multiple access (PRMA) was introduced to integrate voice and date wireless transmission. In [113], a source encoding assisted multiple access protocol was developed to selectively drop source packets and increase the system capacity during congestions. In [114], the resource allocation problems were formulated for different Quality of Service (QoS) requirements. In [115], a video transmission scheme was presented over multi-access networks. In [116], the overall powers were minimized for uplink multi-cell CDMA systems. In [60, 63], the system utility was maximized by dynamic pricing and cooperation between mobiles and base stations. In [117], the problem was formulated as a constrained optimization problem by approximations to have a simple solution. There are few existing works for modelling joint source channel coding and power control in multi-access networks. Moreover, the solutions are either nonlinear optimization programming, or Lagrangian or convex optimization methods by using convex or linear approximations. However it is hard to find an algorithm with a good performance and a low complexity. Therefore, the goal is to construct a multiuser across-layer resource allocation protocol and develop a fast algorithm with a relatively good performance.

In our approach, we propose a distortion management protocol and develop a heuristic and fast resource allocation algorithm in a power-limited downlink single cell CDMA system. The goal is to reduce the overall system distortion, under the constraint of maximal transmitted power from the base station and the maximal distortion for each user. If the network is lightly loaded, we will assign the minimal distortion to everybody. Otherwise, even with the maximal transmitted power, the minimal distortion cannot be achieved by everybody. Under this condition, we assign the maximal distortion to each user first. If there is transmitted power left, we will assign some extra power to the user who can be satisfied and reduce his distortion most easily. To deserve an assignment that reduces his distortion, a user must have a small rate (high distortion), have a good channel condition, or generate small interferences to others. The previous step is continued until the power is used up. The idea is similar to a daily pizza party with limited pizzas. We will let everybody eat the minimal quantity of pizzas. Then we will assign the pizzas left to kids first, then to old people and ladies, finally to young gentlemen. Here the power is similar to pizzas and the distortion resembles the index for hunger. Because of the similarity, we call the proposed algorithm “pizza party” in [118]. We develop a performance bound to compare the proposed algorithm. We also explore the dynamic system case where the call arrival rate is modelled as Poisson distribution, the holding time is modelled as exponential distribution, and speech activities is modelled as a Markov process. From simulations, the proposed algorithm fundamentally reduces the distortions and the necessary maximal transmitted power when the number of users is large, compared with a traditional voice over CDMA scheme (with no distortion control).

System Model

Consider N users for downlink of a single cell CDMA system. W is the total bandwidth which is fixed. R_i is the i^{th} user's transmit rate. So W/R_i is the processing gain. The system is assumed to be synchronous and each user is assigned a unique pseudo-random code within each cell. Because of the multipath environment, the orthogonality may not be guaranteed [120, 121]. Each mobile user is subject to intra-cell interferences from other users. Over one bit period, the received signal at the i^{th} mobile is given by:

$$y_i(t) = \sum_{j=1}^{N} \sqrt{P_j G_i} \sum_{l=1}^{L} \alpha_i^l (t - \tau_i^l) b_j s_j (t - \tau_i^l) + n_i(t) \tag{6.1}$$

where P_j is the transmitted power from the base station for the j^{th} mobile, G_i is the path loss to the i^{th} user, α_i^l is the l^{th} multipath fading to the i^{th} user, τ_i^l is the corresponding delay, b_j is the transmit bit, s_j is the signature of the j^{th} user, and n_i is the noise plus inter-cell interferences.

The chip rate matched filter is applied with sampling at the chip rate. The Rake receiver is used with finger weight equal to the complex conjugates of the multipath fading. The sum of multipath fading powers is assumed to be unit. The mobiles' thermal noise plus inter-cell interference are assumed to be the same for all users and have variance σ^2. The SINR of mobile i at the output of Rake receiver is give by:

$$\Gamma_i = \frac{W}{R_i} \frac{P_i G_i}{G_i \sum_{\substack{j=1 \\ j \neq i}}^{N} \theta_{ji} P_j + \sigma^2} \tag{6.2}$$

where θ_{ji} is the orthogonality factor which represents the fraction of the received downlink power that is converted by multipath into the intra-cell interference. The higher the value, the more the orthogonality is loss. We assume the fading profiles are the same and $\theta = \theta_{ji},\ \forall i, j$. In [121], for the independent Rayleigh fading, the average orthogonality factor is approximated by:

$$\theta = 0.81 - \frac{0.81 \sum_{l=1}^{L} (E(|\alpha_i^l|^2))^2}{(\sum_{l=1}^{L} E(|\alpha_i^l|^2))^2}. \tag{6.3}$$

Protocol Description

Fig. 6.3 shows the block diagram of the proposed cross-layer protocol to manage the interferences by controlling different users' source rates, channel coding rates, transmit rates, and transmitted powers. The protocol is located at the base station and allocates rates and powers to all the users based on the speech activities and the channel conditions. The protocol is operated in such a way that the distortion due to channel-induced errors should be within a range of acceptable small values, so that the system will behave according to the rate distortion curve of the speech encoder. In doing so, the protocol considers the effects on the reconstructed signal qualities and takes into consideration the subjectively more annoying random nature of channel-induced errors. For example, when the channel is bad, there are more transmitted bits assigned to channel protection and less bits for source coding. This reduces the channel errors but introduces source coding distortions. For the reconstructed received voice packet, this kind of distortions are subjectively better, behave according to the rate distortion curve, and can

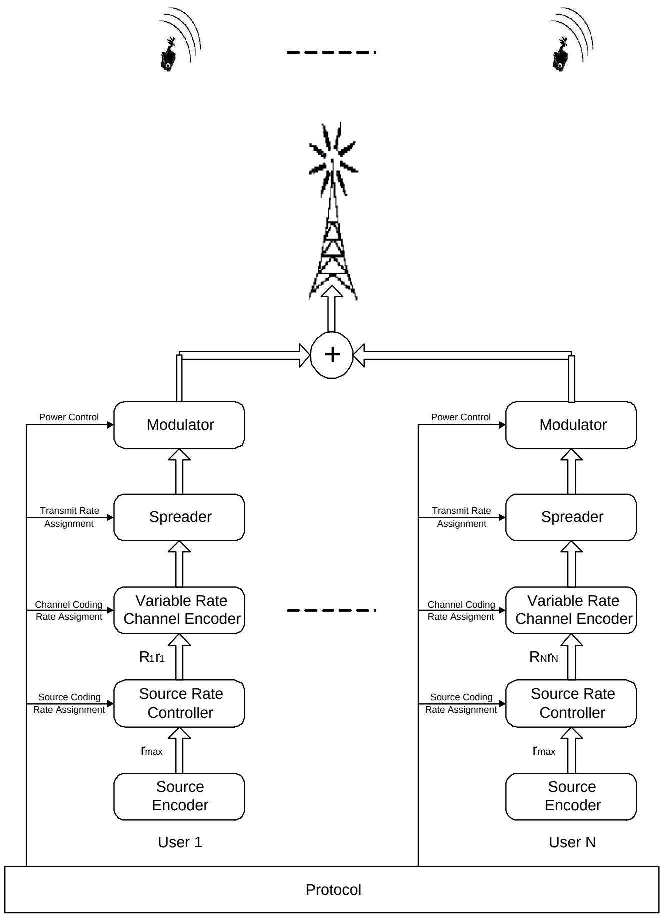


Figure 6.3: Block Diagram for the Proposed Protocol

be predictively controlled by the proposed protocol. In the rest of this part, the modules in the system will be described.

In the proposed system, the real time source encoder has the key property that the output rate can be externally controlled. This can be implemented by using either variable rate or embedded encoders. In the first case, the coder generates one bit stream for each of the possible encoding rates. Only one of these will be selected and transmitted based on the rate assignment. Using embedded encoders presents the advantage that only one bit stream is generated, making the adaptation to the rate assignment simply by dropping as many bits as necessary from the end of the bit stream. Although the "bit dropping mechanism" is exclusive to the embedded stream, his term is used loosely to represent a reduction in the source rate, regardless of the particular source encoder implementation. The source coder is assumed to have the maximal output rate r_{max} bits/s and the source rate controller has the output rate $r_i R_i$ bits/s ($r_i R_i \leq r_{max}$), where r_i is the variable channel coding rate and R_i is the CDMA transmit rate. Then the data streams are encoded by channel coding with rate r_i. The processing gain for the CDMA spreader is W/R_i. BPSK modulation is applied with power control in the modulator.

Define $f_i(r_i R_i)$ as the distortion-rate function of the i^{th} user's source encoder transmitting at rate $r_i R_i$. In most well designed encoders, f_i is a convex and decreasing function. The minimum distortion occurs at maximum source rate r_{max}. Furthermore, the source encoder distortion-rate function [127], [128] is approximated by:

$$f_i = \delta 2^{2k(r_{max} - r_i R_i)} \tag{6.4}$$

where δ is the minimal distortion and k is a parameter depending on the encoders. This is a very general form that applies to the case of Gaussian source with squared-error distortion or when the high-rate approximation holds. In the case of realistic encoders, we find that (6.4) constitutes a good and tight upper bound on the real distortion-rate curve. Furthermore, the parameter k can be determined through simulations for any encoder, so that (6.4) can be a tight bound on the real distortion-rate operating curve. Define $D = 2^{2kr_{max}}$, the normalized distortion is given by:

$$D_i(r_i, R_i) = \frac{f_i}{\delta} = D2^{-2kr_i R_i}. \tag{6.5}$$

For simplicity, all the transmitted bits are assumed equally important for error protection purposes. Because channel induced errors are more perceptibly annoying than the source encoding distortion, the design goal is that channel induced errors would account for a small percentage of the overall end-to-end distortion, i.e., a desired Frame Error Rate (FER) must be satisfied. The design will be constrained by the condition of meeting a target SINR so as to achieve the FER. A reduction in source encoding rate allows for an decrease in channel code rate or a decrease in transmit rate, as a result increases the channel protection. To maintain the design goal, the target SINR is a function of the source encoding rate, or equivalently, a function of both channel coding and transmit rate. Therefore, it is possible to increase the overall end-to-end network distortion slightly and reduce the interferences greatly while meeting the FER requirement by clearly managing the users' source, channel coding, and transmit rates.

In our approach, Rate Compatible Punctured Convolutional (RCPC) code [119] is applied for channel coding. A family of RCPC codes is described by the mother code of rate $\frac{1}{M}$. The output of the coder is punctured periodically by puncture tables. The puncturing period Q determines the range of channel coding rates

$$r_i = \frac{Q}{Q+l}, \quad l = 1, \ldots, (M-1)Q, \tag{6.6}$$

between $\frac{1}{M}$ and $\frac{Q}{Q+1}$ with different channel protection abilities.

The rate compatibility restriction on the RCPC puncturing tables ensures that all code bits of high rate codes are used by the lower rate codes. This allows incremental redundancy and continuous rate variation of error protection within a data frame. Moreover only one kind of Viterbi receiver is needed for the RCPC codes with different rates, which reduces the system complexity.

Furthermore, through simulations using different configurations of RCPC codes, the targeted SINR as a function of channel coding rate, when transmit rate is fixed, can be approximated accurately by

$$\gamma_i = 2^{Ar_i + B} \tag{6.7}$$

where γ_i is the required targeted SINR for the desired FER, A and B are the fixed parameters of the error control coding, and r_i is the channel coding rate. For desired FER, $\Gamma_i \geq \gamma_i, \ \forall i$.

From (6.5) and (6.7), both distortion and targeted SINR (which is related to transmitted power in (6.2)) are the functions of rates. So the protocol can control the source coding rate, channel coding rate, and transmit rate to control the distortion, reduce the transmitted power, and increase the system performance.

Real Time Distortion Management

Problem Formulation

In practice, the transmitted power from the base station is bounded, because there exists an implementation limitation and co-channel interferences should not be introduced too much to other cells. When the system is lightly loaded, each user could have the minimal distortion and the necessary total transmitted power could be still less than the maximal transmitted power available from the base station antenna. When the system becomes more loaded, even with the maximal transmitted power, the system cannot let every user have the minimal distortion. Under this condition, it is necessary to have a graceful distortion control: Some users, with high satisfactions of distortions, with bad channel conditions, or who introduce too many interferences to others, will sacrifice their performances slightly and increase their distortions in a controlled way. By doing these, the system will use the limited transmitted power to reduce the interferences, optimize the overall system performance, and increase the total number of users. The problem is to decide who will sacrifice and how the users increase their distortions.

First, the transmit rate R_i is assumed fixed and only the channel coding rate r_i is modified. Later, it will be shown how to modify both R_i and r_i. The goal to minimize the overall system distortion, under the constraints that each user's distortion is smaller than a maximal acceptable distortion and the overall transmitted power $P_{sum} = \sum_{i=1}^{N} P_i$ from the base station is bounded. The problem is formulated as:

$$\min_{r_i} \sum_{i=1}^{N} D_i \tag{6.8}$$

$$\text{subject to} \begin{cases} \text{Distortion Range:} & 1 \leq D_i \leq D_{max}, \forall\ i, \\ \text{Transmitted Power:} & P_{sum} \leq P_{max}, \end{cases}$$

where P_{max} is the maximal transmitted power and D_{max} is the maximal acceptable distortion. Without loss of generality, all users are assumed to have the same D_{max} for simplicity.

The problem in (6.8) is a nonlinear nonconvex problem and there might be many local minima. It is very difficult to solve it by Lagrangian method or nonlinear integer programming. Moreover, the computation complexity will grow quickly with the number of users increasing. In order to implement the protocol in the real CDMA system with large number of users, it is necessary to develop a fast algorithm with a relatively good performance.

Resource Allocation Algorithm

The intuitive idea to develop a fast algorithm comes from a daily event. For example, in a pizza party with limited available pizzas, if the number of people is small, everybody will have enough food and there might be some pizzas left. However if the number of people is large and there is no way that everybody will be well satisfied, it is necessary to decide how to allocate the pizzas. One possible solution is to let everybody eat the minimal pizzas. (we assume there are enough pizzas for this requirement.) Then we will let kids eat one more slice of pizza, because they eat less and are easy to be happy. If there are any pizzas left, we will give one slice per time to the people who can be satisfied easily then. (Probably older people will get pizzas next, then ladies, and finally young gentlemen.) By allocating pizzas in such a way, we can use the limited pizzas to let the overall people's satisfaction high.

By using the same idea above, the overall transmitted power is viewed as pizzas and the user's distortion as the index for hunger. In order to decide who is easy to be satisfied, we need to find the differential of the overall transmitted power with respect to each user's distortion. Define

$$T_i = \frac{2^{Ar_i+B}R_i}{W} = \frac{P_i G_i}{G_i \sum_{j\neq i} \theta_{ji} P_j + \sigma^2}. \tag{6.9}$$

If the processing gain is large, i.e., W/R_i is large, T_i is a small number. Since $\theta_{ji} < 1$, $\theta_{ji}T_i$ is also a small number. A simple approximation for P_{sum} is give by

$$P_{sum} = \mathbf{1}^T[\mathbf{I}-\mathbf{F}]^{-1}\mathbf{u} \approx \mathbf{1}^T[\mathbf{I}+\mathbf{F}]\mathbf{u} = \sum_{i=1}^{N} \frac{\sigma^2 T_i}{G_i} + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{\sigma^2 \theta_{ji} T_i T_j}{G_j} \tag{6.10}$$

where $\mathbf{1} = [1 \dots 1]^T$, $\mathbf{u} = [u_1, \dots u_N]^T$ with $u_i = \sigma^2 T_i / G_i$, and

$$[\mathbf{F}]_{ij} = \begin{cases} 0 & \text{if } j = i, \\ \theta_{ji} T_i & \text{if } j \neq i. \end{cases}$$

The gradient of the overall transmitted power with respect to each user's distortion can be written as a function of the following three differentials:

$$g_i = \frac{\partial P_{sum}}{\partial D_i} = \frac{\partial P_{sum}}{\partial T_i}\frac{\partial T_i}{\partial r_i} / \frac{\partial D_i}{\partial r_i} \tag{6.11}$$

where

$$\frac{\partial P_{sum}}{\partial T_i} = \frac{\sigma^2}{G_i} + \sum_{j \neq i}^{N} \frac{\sigma^2 \theta_{ji} T_j}{G_j}, \tag{6.12}$$

$$\frac{\partial T_i}{\partial r_i} = \frac{AR_i 2^{Ar_i+B} \ln 2}{W}, \tag{6.13}$$

$$\frac{\partial D_i}{\partial r_i} = -2kDR_i 2^{-2kr_iR_i} \ln 2. \tag{6.14}$$

So the final gradient can be written as:

$$g_i = C2^{(A+2kR_i)r_i} \left(\frac{1}{G_i} + \sum_{j \neq i}^{N} \frac{\theta_{ji} T_j}{G_j} \right) \tag{6.15}$$

where C is a negative constant. The absolute value of g_i is determined by the three factors: the current rates (the term before the parentheses), the channel gain (the first term inside the parentheses), and the interferences to others (the second term inside the parentheses).

If P_{max} is large enough for every user in the cell to have the minimal distortion, $D_i = 1$ is assigned to everybody and there is might be some overall transmitted power left.

If P_{max} is not large enough for everybody to have the minimal distortion, $D_i = D_{max}$ $\forall i$ will be initially assigned. If the power is still not enough, it means that there are not enough power to satisfy the group's minimal needs and an outage is reported. If there is some power left, we will see who will be most easily to be satisfied by determining the gradient $\partial P_{sum}/\partial D_i$. If the absolute value of the gradient is small, that means this user is a "kid" who can eat little and become happy. For this user, from (6.15), the current rates is low (i.e. the distortion is high), the channel gain is good, or the interferences to others are small, consequently this user deserves a smaller distortion. In other words, this

Table 6.1: Pizza Party Algorithm

1. Initialization: If everybody can get $D_i = 1$, then allocate the powers and stop; else allocate D_{max} to everybody. If $P_{sum} > P_{max}$, report an outage.
2. Repeat: • Calculate $\lvert g_i \rvert$. • Increase the rate of the user with smallest $\lvert g_i \rvert$. • If $P_{sum} > P_{max}$, return the previous rate allocation and break.
3. Rate and Power Assignment.

user can reduce his end-to-end distortion while creating the smallest strain on the available resources. So a higher r_i (consequently higher source rate) is assigned to this user to let the distortion become small. Then the gradient is estimated and the rate is assigned again. This process is continued, according to the order of the gradients, until the power is used up. By doing this, the distortions are reduced by consuming the minimal resources step by step.

On the whole, the proposed algorithm is given in Table 6.1. As mentioned before, (6.8) is extremely difficult to solve by traditional methods in which the complexity grows fast with the number of user N increasing. In the proposed algorithm, the complexity lies in calculating the overall transmitted power in (6.10) and computing the gradients in (6.15) . The complexity is $O(N^2)$ and so the proposed algorithm can be easily implemented in practice.

Joint Consideration with Transmit Rate

This part considers the case where the transmit rate R_i can also be adapted. It will be shown that there is no need to adapt both the transmit rate and the channel coding rate, as long as $A \ln 2 \geq 1$. The new gradient is developed by adapting the transmit rate only.

If both R_i and r_i are adapted for resource allocation, it is necessary to find out how to select R_i and r_i. The goal is to minimize the distortion, under the constraint that the demand for the transmitted power is fixed, i.e., $T_i = C'$, where C' is a constant. The problem is:

$$\min_{r_i, R_i} D2^{-2kr_iR_i} \tag{6.16}$$

$$\text{subject to } T_i = R_i 2^{Ar_i+B}/W = C'.$$

Write the Lagrangian function as:

$$J = D2^{-2kr_iR_i} + \lambda(R_i 2^{Ar_i+B}/W - C'). \tag{6.17}$$

where λ is the Lagrangian multiplier. The solution is

$$r_i = \frac{1}{A \ln 2}. \tag{6.18}$$

Astonishingly, r_i is a fixed value and there is no need to adapt r_i, if $A \ln 2 \geq 1$. In the simulation setup, this condition is always held. Therefore R_i can be adapted only. The gradient of the overall transmitted power with respect to the distortion is now given by:

$$g_i = \frac{\partial P_{sum}}{\partial D_i} = \frac{\partial P_{sum}}{\partial T_i}\frac{\partial T_i}{\partial R_i} / \frac{\partial D_i}{\partial R_i} = C'' 2^{\frac{2kR_i}{A \ln 2}} \left(\frac{1}{G_i} + \sum_{j \neq i}^{N} \frac{\theta_{ji} T_j}{G_j} \right). \tag{6.19}$$

where C'' is a negative constant. The same algorithm in Table 6.1 can be applied by varying the transmit rate instead. However if the transmit rate (i.e. processing gain) is changed, the available codes for users will be reduced, which will limit the total number of users admitted to the system. Consequently, the adaption of transmit rate can only apply to the system with abundant codes but limitation on adapting the channel coding.

Performance Bound

Since the optimal solution for the constrained integer programming problem in (6.8) is hard to calculate, in order to evaluate the performance of the proposed algorithm, we provide a performance bound that is able to be calculated, has better results than the optimal solution, and is not be implementable in practice. If the proposed algorithm has the similar performance as the bound, we can conclude that the proposed algorithm is at least near optimal. If the channel coding is assumed as a continuous variable, the problem in (6.8) becomes a nonlinear constrained problem. Then some nonlinear optimization methods can be used to solve it. In this part, an algorithm is developed to calculate the performance bound by applying the barrier method with Newton method [122].

First the transmit rate is assumed fixed and the channel coding rate is assumed to be a continuous real variable. The modified problem definition from (6.8) can be expressed as:

$$\min_{r_i} \sum_{i=1}^{N} 2^{-2kR_i r_i} \tag{6.20}$$

$$\text{subject to} \begin{cases} r_i^{min} \leq r_i \leq r_i^{max}, \forall\ i, \\ P_{sum} \leq P_{max}, \end{cases}$$

where $r_i^{min} = \frac{\log 2(\frac{D}{D_{max}})}{2kR_i}$, $r_i^{max} = \frac{r_{max}}{R_i}$, and r_i is a real number between r_i^{min} and r_i^{max}. From (6.9) and (6.10), the power constraint is a nonlinear function of r_i.

In order to solve (6.20), a barrier method with Newton method [122] is applied. The basic idea for the barrier method is to add barrier functions to the optimization goal such that the constrained optimization problem becomes the unconstrained optimization problem. The sum of optimization goal and barrier functions approaches infinity if the constraints are not satisfied. On the other hand, if the constraint is satisfied, the barrier function does not affect the optimization goal. The barrier function is commonly approximated by logarithmic barrier functions. In the proposed problem, the barrier function is given by:

$$I_{constaint} \approx \Phi_1 + \Phi_2 + \Phi_3, \tag{6.21}$$

$$\Phi_1 = \begin{cases} \sum_{i=1}^{N} \ln(r_i - r_{min}), & r_i > r_{min}, \\ \infty, & \text{otherwise.} \end{cases} \tag{6.22}$$

$$\Phi_2 = \begin{cases} \sum_{i=1}^{N} \ln(r_{max} - r_i), & r_{max} > r_i, \\ \infty, & \text{otherwise.} \end{cases} \tag{6.23}$$

$$\Phi_3 = \begin{cases} \ln(P_{max} - P_{sum}), & P_{max} > P_{sum}, \\ \infty, & \text{otherwise.} \end{cases} \tag{6.24}$$

Φ_1 and Φ_2 are for the channel coding rate range. Φ_3 is for the overall power. The barrier method approach is to solve the constrained optimization problem by a sequence of unconstrained problems, where the new problem is initialized by the results in the previous iteration. Rewrite (6.20) as:

$$\min_{r_i} \ f = \tilde{t} \sum_{i=1}^{N} 2^{-2kR_i r_i} + I_{constraint} \tag{6.25}$$

where $\tilde{t}$ is a value that increases from iteration to iteration. The barrier functions become more and more like the ideal barrier function as $\tilde{t}$ increasing. So the solution is more and more optimal. Within each iteration, Newton method [122] is used to solve the unconstrained optimization problem. Define $\mathbf{r} = [r_1 \dots r_N]^T$, the algorithm is given in Table 6.2, where m is the iteration number for barrier method, ϵ determines the accuracy of the proposed algorithm, t' is the optimal step for the Newton method, t_0 is the initial value for barrier function, whose value determines

Table 6.2: Barrier Method for Performance Bound

1. Initial:
$\mathbf{r}$ = any feasible value, $\widetilde{t} = t_0 > 0$, $\beta > 1$, $\epsilon > 0$.
2. Repeat:
• Start at $\mathbf{r}$, compute new $\mathbf{r}$ by minimizing f, using **Newton Method**:
1. Compute Newton step $\mathbf{v}_{nt}$ and decrement λ^2.
$\mathbf{v}_{nt} = - \bigtriangledown^2 f^{-1} \bigtriangledown f$
$\lambda^2 = \bigtriangledown f^T \bigtriangledown^2 f^{-1} \bigtriangledown f$
2. quit if λ^2 is stable.
3. Line search: compute step size t' by backtracking line search.
4. Update: $\mathbf{r}$=$\mathbf{r}$+t'*$\mathbf{v}_{nt}$.
• if $m/\widetilde{t} < \epsilon$, return $\mathbf{r}$.
• $\widetilde{t} = \beta\widetilde{t}$.

the convergence rate of the first iteration, and β is the constant that $\widetilde{t}$ is multiplied in each iteration.

The performance bound algorithm in Table 6.2 is not implementable in practice. First, the rate is assumed to be continuous, which is not true in real channel coding coder. Because of the continuous assumption, this algorithm will find a performance bound with a better performance than the optimal solution in (6.8). Second, the complexity of this algorithm is much higher than the proposed algorithm in Table 6.1. The complexity lies in the factors that, in order to find the solution, one iteration is needed for Newton method and another iteration is needed for barrier method. Third, because the problem in (6.20) is non-linear and non-convex, there might be many optima. Multiple initialization or even annealing is necessary to find the global optimum. So the algorithm in Table 6.2 is hard to be implemented in practice but can be used to compare the performance of the proposed fast implementable algorithm in Table 6.1. In the simulation results, it will be shown that the proposed algorithm has the similar performance as the performance bound. Consequently, the proposed algorithm is at least near optimal.

Dynamic System Model

In previous sections, the resource allocation algorithm is developed to reduce the overall distortions with a fixed number of users in the system. In this part, the dynamic traffic case is considered where the number of the admitted users is changing. Furthermore, the different speech activities are considered. A Monte Carlo method is constructed to analyze the system performance.

The probability of the arrival rate for each call is assumed to be a Poisson distribution with mean value λ. The holding time for each call is modelled as an exponential distribution with parameter μ. The number of admissible users is bounded by the processing gain. Define N_{max} as the maximal number of admitted users. The average distortion per call $E(D)$ is assumed as a performance measure for admission policy. Specifically, N_{max} is selected such that $E(D) \leq d$, where d is a given threshold. Moreover N_{max} is less than or equal to the processing gain.

Suppose N is the number of users in the system. From [113], N is a truncated Poisson (ρ, N_{max}) random variable, where $\rho = \lambda/\mu$. For $n = 0, 1, \ldots, N_{max}$, the stationary probability that the system has n ongoing users is given by:

$$P[N = n] = \frac{\frac{\rho^n}{n!}}{\sum_{i=0}^{N_{max}} \frac{\rho^i}{i!}}. \tag{6.26}$$

By PASTA property [123], the blocking probability is given by $P_b = P[N = N_{max}]$.

The probability of a user in the system with n users is called the batch probability. The batch probability that an arbitrary call is from a batch of n calls is given by [124]:

$$\widehat{p}_n = n \frac{\frac{\rho^n}{n!}}{\sum_{i=0}^{N_{max}} i \frac{\rho^i}{i!}}. \tag{6.27}$$

From the previous, the overall distortion with the fixed number of users is known, i.e $E(D|N = n)$. By using the distribution in (6.27), the average distortion per call is expressed as:

$$E[D] = E[E(D|N = n)] = \sum_{n=0}^{N_{max}} E(D|N = n)\widehat{p}_n. \tag{6.28}$$

In addition, the speech activity is also considered. In two-way conversation calls, the silent period is roughly 65% of all the time [125]. Moreover the operating ranges on the rate distortion curves vary greatly between the silent and talk speech periods [125]. So the number of transmitted bits that is necessary for the perceptual quality changes widely with time. It is natural to classify the speech activities and apply variable rate codings. In our approach, the speech activities are modelled as a two-state on-off Markov chain [126] and the transition probabilities are shown in Fig. 6.4. The transition probability $\varepsilon = 1 - e^{-T/t_1}$ and $\kappa = 1 - e^{-T/t_2}$, where T is the frame duration, t_1 is the average talkspurt duration, and t_2 is the average silence duration.

When the speech is in silence state, there is not need to have high transmission rates. The minimal distortion is increased to δ^s and the maximal source rate is decreased to r^s_{max}. Define $D^s = 2^{2kr^s_{max}}$, the normalized distortion in the silent mode is expressed as:

$$D_i^s = \frac{f_i}{\delta^s} = D^s 2^{-2kr_i R_i}, \; r_i R_i \leq r^s_{max}, \; \forall i. \tag{6.29}$$

The normalized D_i^s is used by the same way as D_i in the proposed algorithms. The

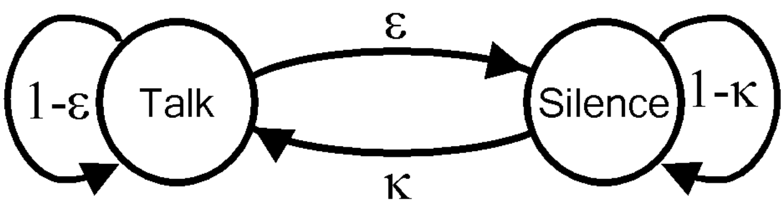


Figure 6.4: Two-State Markov Model for Speech Activity

difference is that the distortions are normalized with different minimal distortions for talk and silence modes. Consequently, many transmission bandwidths can be saved during the silent mode without reducing the speech perceptual quality. The system capacity can be greatly increased as well.

As mentioned in the previous, the analytical result for $E(D|N = n)$ is impossible to obtain. In order to evaluate the system performance, a Monte Carlo method is shown in Table 6.3. The simulation is run for a sufficient large number of runs, so that the stable performance results are obtained with a sufficient accuracy.

Simulation Results

We focus our study on real time voice communications. Eighteen sequences, both male and female speakers, from the NIST speech corpus [129] are used. These sequences are encoded using the GSM AMR (Advance Multi-Rate) Narrow-band Speech Encoder [130]. This encoder operates with 20 ms frames, 5 ms look-ahead and includes an error concealment mode. Of the eight possible encoding rates: 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15 and 4.75 kbps, the six highest ones are used.

To determine the end-to-end distortion, we choose a perceptually weighted log-spectral distortion measure [132] calculated by numerical approximation of the function

$$SD(\widehat{A}(f), A(f)) = \sqrt{\int |W_B(f)|^2 \left| 10 \log \frac{|\widehat{A}(f)|^2}{|A(f)|^2} \right|^2 df} \tag{6.30}$$

Table 6.3: Monte Carlo Method for Dynamic System

1. Random Generation: Generate the calls and channel conditions.
2. Admission Control: Given λ and μ, calculate N_{max} and blocking probability P_b by threshold d.
3. Speech Activity: Determine the state of speech and assign different normalized distortion.
4. Pizza Party Algorithm: Find the rate and power allocation by the proposed algorithm.
5. Accumulate Results Accumulate results and repeat from step 1, until the sufficient accuracy.

where $A(f)$ and $\widehat{A}(f)$ are the FFT-approximated spectra of the original and the reconstructed speech frames, and $W_B(f)$ is the subjective sensitivity weighting function [131]:

$$W_B(f) = \frac{1}{25 + 75(1 + 1.4(f/1000)^2)^{0.69}}. \tag{6.31}$$

This distortion is measured on a frame-by-frame basis and then averaged over all frames, including outliers to further capture the effects of channel errors. This measurement is chosen not only because of its good mathematical properties, but also because of its good correspondences to subjective measure. A normalized distortion measure is reported, which is computed as the ratio of the spectral distortions to that of the speech sequence encoded at the highest rate (12.2kbps) without channel noise.

Also, for the proposed system, BPSK modulation is assumed. For RCPC channel coder, a memory 4, puncturing period 8, mother code rate 1/4 (variable rate in the proposed system) RCPC code in [119] is decoded with a soft Viterbi decoder. The total bandwidth W is 1.5616MHz. The channel is assumed to be affected by normalized Rayleigh fading (average power loss equal to 1), and normalized path loss (with propagation constants assumed equal to 1) with a path loss exponent equal to 3. The cell radius is 500m. θ_{ji} is assume to be the same for all the users and is set to 0.9. Background noise level was assumed equal to 10^{-6}. $k = 3.3 \cdot 10^{-5}$. $r_{max} = 12.2$kbps.

One important point worth of noticing is that the constraint on the channel induced errors not only is necessary for (6.8) but also is advantageous, because it assures that the increase in distortion is smooth, controllable, and predictable. This is because the dominant process is the reduction in source encoding rate, thus the system behavior follows the rate-distortion curve. Channel induced distortion is kept at a sufficiently small value by appropriately setting the rates and powers. In contrast, this is not the case for the traditional voice over CDMA approach where the increase in distortion is a consequence of the uncontrolled increase in channel-induced errors. In this case, the system behavior is much less predictable, because the random process of errors in the channel will dominate, and distortion is more subjectively annoying.

Fig. 6.5 shows the distortion as a function of SINR for six possible operating modes, where each mode is characterized by the pair (source encoding rate, channel code rate). Without adaptive source coding, each user's distortion has to follow a specific curve. With adaptive source coding, each user can follow the minimum of different curves, so that the distortion can be greatly reduced. Fig. 6.6 shows an example of the simulations to find an approximation for the target SINR-channel coding rate function. The figure shows the target SINR, in $\log_2$ scale, as a function

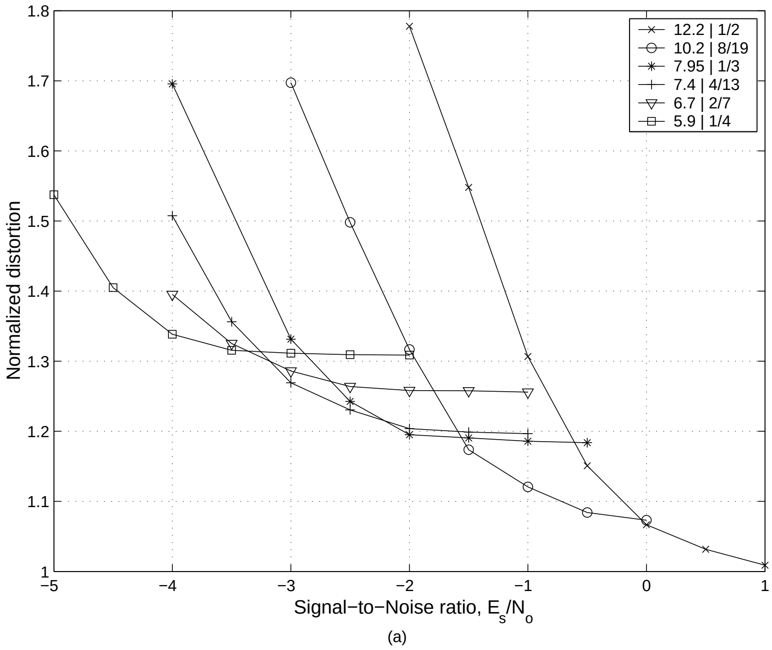


Figure 6.5: Rate Distortion Curves

of the source encoding rate, where the channel induced errors are less than 3% of that of the corresponding source encoding distortion (channel induced distortion contributes 3 % to the end-to-end distortion). The figure confirms that (6.7) is a good approximation.

First, the transmit rate is at 24.4kbps and processing gain is at 64. Fig. 6.7 shows the normalized distortion vs. the number of calls with different transmitted powers for the proposed scheme. The figure also includes, for comparison purposes, results for an equivalent traditional CDMA system that shares the same configuration as the proposed scheme but operates without changing mode. For the case of this traditional system, all calls operates in the (12.2 Kbps, 1/2) mode. From these results, we can draw several conclusions. When the number of users is small, all the schemes with different powers works the same. This is because there is enough

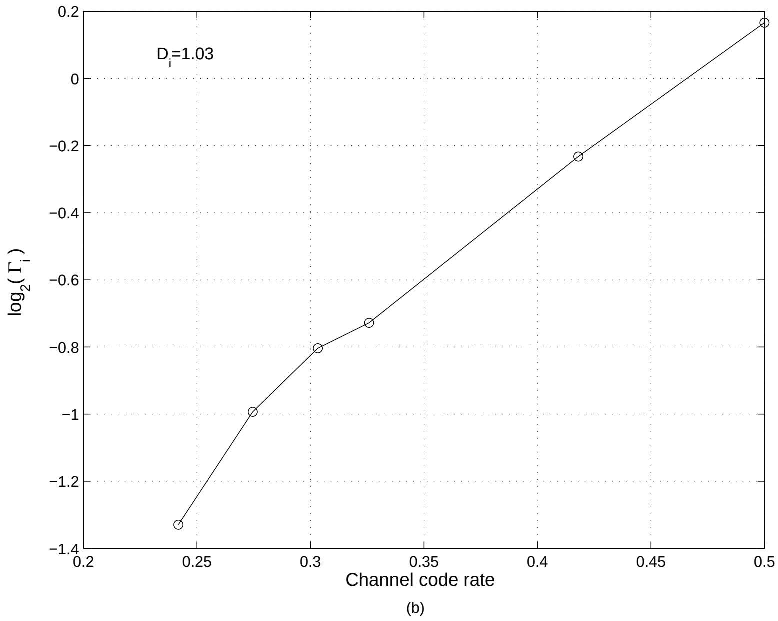


Figure 6.6: Required SINR vs. Rate

power for everybody to have the minimal distortion. When the user number is increased, the proposed scheme can reduce the normalized distortion fundamentally, when compared to the traditional system. This is because the proposed scheme controls the distortion smoothly by adapting the source and channel coding rates. In particular, if, for example $P_{max} = 350$, the proposed system can support 30 users with 6 % less distortion, 40 with 12 % and 50 users with 37 % less distortion. When the transmitted power is increased, the distortion will be reduced. In Fig. 6.8, we compared the normalized distortion as a function of the maximal available power for a fixed number of users in the system ($N = 30$, $N = 40$, and $N = 50$) that represents different network loading conditions. It shows the proposed system can deliver the same level of average end-to-end distortion by a much lower maximum transmitted power. We also show the case to modify the transmit rate only and fix the channel coding rate as $\frac{2}{7}$. In this case, the proposed algorithm has

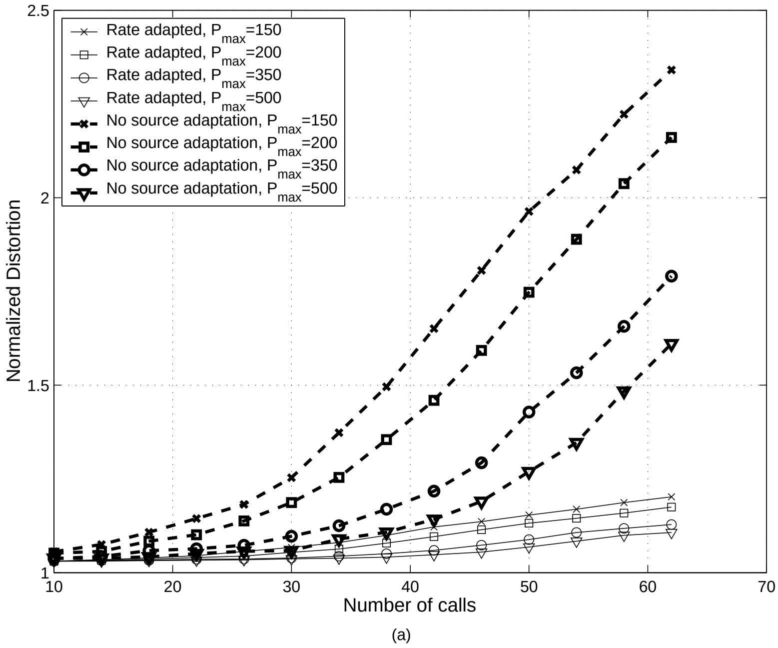


Figure 6.7: Norm. Distortion vs. No. of Calls

a slightly performance loss with small P_{max}.

In order to evaluate the performance of the proposed fast algorithm, we compare the results with the performance of the bound algorithm in (6.2) developed in previous. We define the relative difference as the average distortion of the proposed algorithm minus the average distortion obtained by the bound algorithm, then the result is divided by the average distortion of the proposed algorithm. In Fig. 6.9, we show the relative difference vs. number of users with fixed transmit rate and by adapting only channel coding rate. We apply multiple initial to get the global optimization by the bound algorithm. Since the channel rate is assumed continuous for the bound algorithm, the global optimum is always better than the global optimum defined in problem formulation (6.8). Our proposed algorithm is suboptimal. But when we compare the performance of the proposed algorithm and

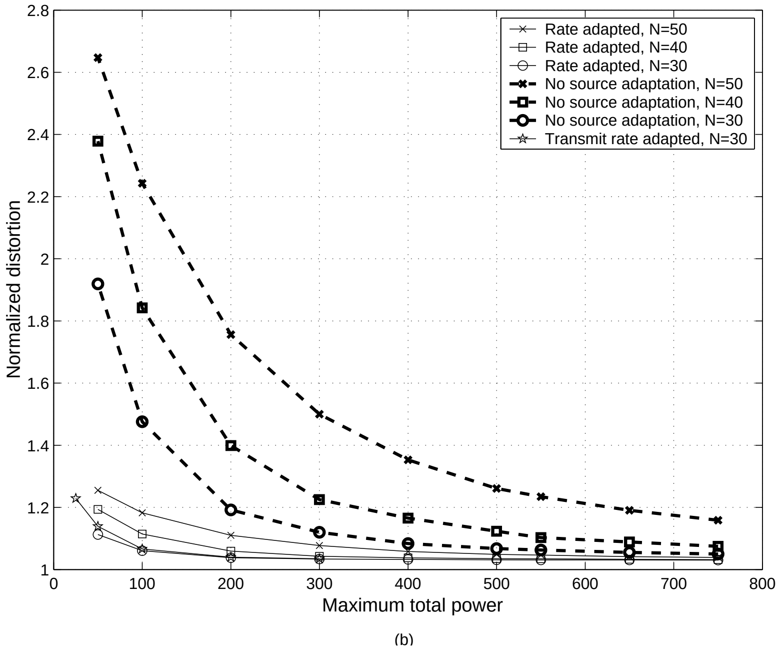


Figure 6.8: Norm. Distortion vs. P_{max}

the bound algorithm, the difference is very small. So this proves that the proposed algorithm is at least near optimal. The performance gets worse when the number of users increases, because there exist more and more local optimal and the bound algorithm can performance better with more users.

For the dynamic system, the arrival rate λ over the holding time μ is offered load ρ. The average talkspurt duration is 1s and the average silence duration is 1.35s. $D^s =$ and $r^s_{max} =$. In Fig. 6.10 and Fig. 6.11, we show the distortions and outage probabilities of normalize distortion vs. offered load for $d = 1.2$, where outage is when the system cannot allocate the resource for the maximal distortion requirement. We compare the proposed algorithm vs. the fixed algorithm for $P_{max} = 150, 200, 350$, respectively. We can see that both the normalized distortions and the outage probabilities increase with the offered load increasing, while the

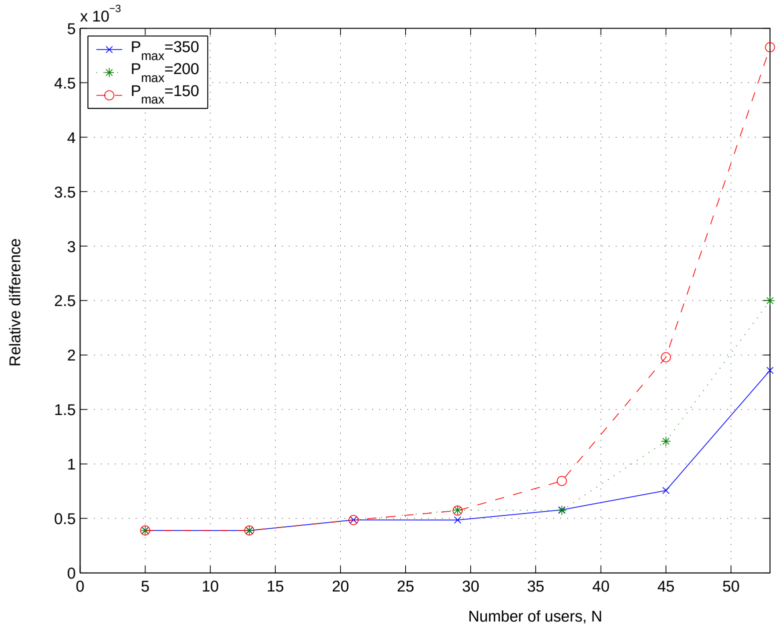


Figure 6.9: Relative Difference vs. Number of Users

proposed algorithm provides much lower distortions and outage probabilities.

6.3 Joint Power Control and Blind Beamforming

In this section, first we will discuss the basics for blind methods for estimation. Then we propose a joint power control and blind beamforming algorithm for wireless networks.

6.3.1 Basics for Blind Methods

The important role that channel estimation and equalization play in digital communication systems is well known. As a majority of communication systems often struggle with limited bandwidth constraint, it is desirable for the receiver to obtain optimum channel equalizers without consuming much channel bandwidth. By

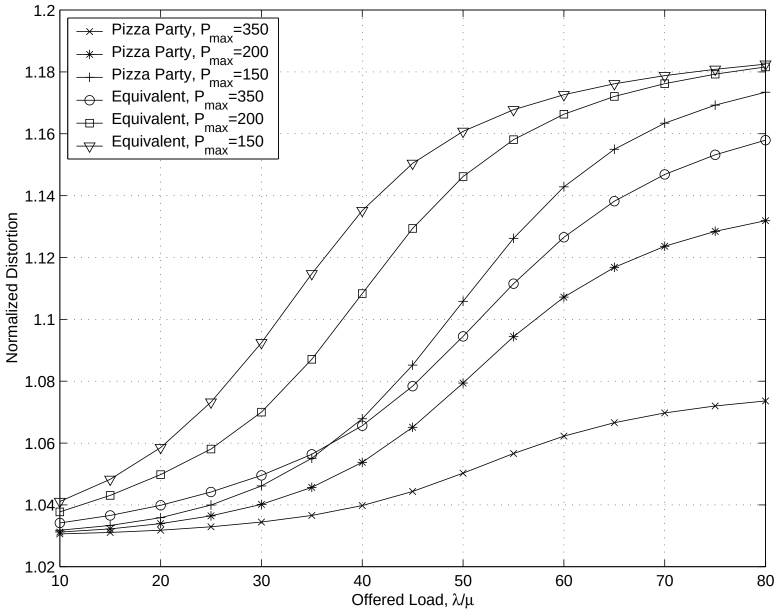


Figure 6.10: Dynamic System: Normalized Distortions with $d = 1.2$

eliminating training data and maximizing channel capacity for true information transmission, blind channel equalization presents a bandwidth efficient solution to distortion compensation. Its importance also lies in the practical need for some communication receivers to equalize unknown channels without the assistance and the expense of training sequences. Compared with the more traditional approach of training based equalization, blind equalization is a theoretically challenging problem that is gaining appeal.

The blind estimation problem is to recover a set of n independent signals and the channels response $\mathbf{A}$ from $m \geq n$ observed instantaneous mixture of these signals without knowledge of channel and original transmitted signals $\mathbf{s}(t)$. Let $\mathbf{x}(t)$ denote the $m \times 1$ vector of observations at time t. The signal is corrupted by an additive noise vector $\mathbf{n}(t)$. We have

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t). \tag{6.32}$$

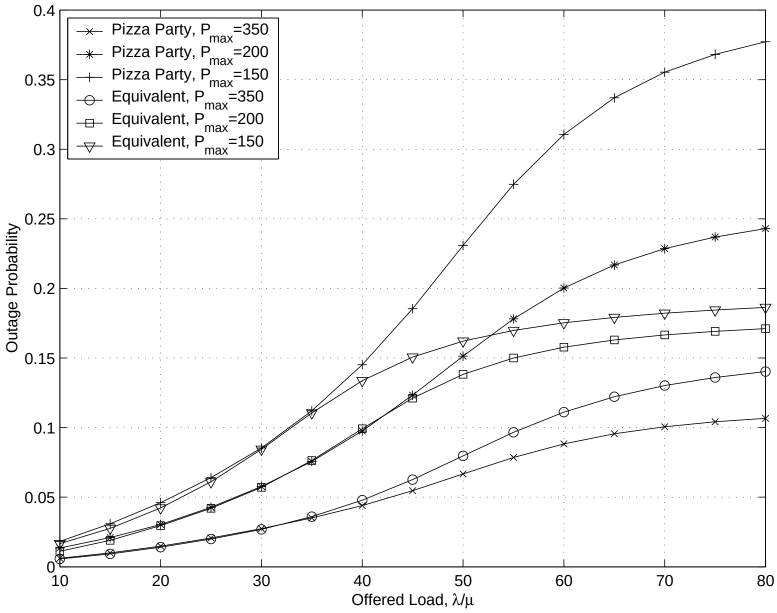


Figure 6.11: Dynamic System: Outage Probabilities with $d = 1.2$

The problem of blind estimation is to estimate $\mathbf{A}$ and $\mathbf{s}(t)$ from $\mathbf{x}(t)$ directly without need of training sequences. The major existing methods to solve the above problem are listed as:

- Constant modulus algorithm (CMA)

  CMA is one of the most popular and effective blind equalization algorithm for linear equalizers, which restores the constant envelope property of the transmitted signal and increases the SNR. This algorithm thus employs a priori knowledge about the envelope of the transmitted signal and has the nice characteristic that no training sequence is required.

- Iterative approach

  This method attempts to estimate the unknown channel and the channel input symbols in alternation. maximum likelihood algorithm can be implementable via expectation-maximization (EM) algorithm.

- High order statistics (HOS)

  HOS algorithms select the channel response either by lease square cumulant matching or by solving equations that the channel response must satisfy.

- Subspace approach

  This approach using the sinusoid nature of the channel responses and estimate the phase of the channel response by using subspace approach.

- Using known structure of transmitted signal or channel

  Many of the nature of the transmitted signal and channel structure can be explored to fasten the convergence rate, such as the discrete alphabet of transmitted signal and sinusoid nature of channel responses.

- SIMO, MIMO

  If the number of output signals exceeds the number of input signals, only second order statistics are necessary to identify linear discrete channels.

In our approach, we use an iterative blind estimation approach by using discrete finite constellation property of the transmitted signal. We extend this approach for multiuser case so that all the users can communication simultaneously by careful power control.

6.3.2 Distributed Joint Scheme

Traditional joint power control and beamforming achieve the targeted Signal-to-Interference-Noise-Ratio (SINR) at the receivers by assuming the knowledge of the measurements of channel parameters and SINR. Blind beamforming is an effective technique for beamforming and channel estimation without the need of training sequences, thus not consuming extra bandwidth. In our approach, we propose a novel joint power control and blind beamforming algorithm that reformulates the power control problem in such a way that it does not need any prior knowledge and additional measurements in the physical layer. In contrast to the traditional schemes that optimize SINR and, as a result, minimize bit error rate (BER), our proposed algorithm achieves the desired BER by adjusting a quantity available from blind beamforming. By sending this quantity to the transmitter through a feedback channel, the transmit power is iteratively updated in a distributed manner in the wireless networks with co-channel interferences. Our proposed algorithm is more robust to estimation errors. We have shown in both analysis and simulation that our algorithm converges to the desired solution. In addition, a Cramer-Rao lower bound is derived to compare with the performance of our proposed joint power control and blind beamforming system.

The organization of this subsection is as follows: First, we give the motivation and sketch for the proposed scheme. Then, we present the system model and the traditional joint power control and beamforming problem. We choose a blind beamforming algorithm. Then we give the reformulated joint power control and blind beamforming problem. An adaptive algorithm is developed and a system is constructed. The convergence and uniqueness of the solution are analyzed. The CRB is derived to compare the performance. We evaluate our algorithm via numerical studies.

Motivation and Sketch

One of the major challenges for the system design is the limited available radio frequency spectrum. Channel reuse is a common method to increase the wireless system capacity by reusing the same channel beyond some distance. However this introduces CCI that degrades the link quality. Two promising approaches to combat CCI are power control and antenna array processing. Power control is one direct approach toward minimizing CCI. The transmit powers are constantly adjusted. They are increased if the SINRs at the receivers are low and are decreased if the SINRs are high. Such a process improves the quality of weak links and reduces the unnecessary transmit powers. Antenna array processing techniques such as beamforming can be applied to receive and transmit multiple signals that are separated in space. Hence, multiple co-channel users can be supported in each cell to increase the capacity by exploring the space diversity.

Many works have been reported in the literature for employing power control and beamforming to reduce CCI. Traditional beamformers such as minimum mean square error (MMSE) and minimum variance distortion response (MVDR) methods have been commonly employed [92]. In [93, 94], general frameworks for power control are constructed. Beamforming is a physical layer technique that can greatly increase receivers' SINR by using the signal processing algorithms, while power control is a media access control layer technique that can effectively control users' transmit powers to share the channels. Many joint power control and beamforming algorithms are proposed in [47, 95, 96, 97, 98]. Most of the existing works assume the availability of prior channel information and measurement of SINR.

As a majority of communication systems often struggle with the limited bandwidth constraint, it is desirable for the receiver with multiple antennas to steer to the desired direction and to estimate the transmit signals without consuming much channel bandwidth. By eliminating the training sequence overhead, used for estimation, and maximizing the channel capacity for information transmission, blind estimation and beamforming [99, 100, 101, 102, 103, 104, 105, 107] offer a bandwidth efficient solution to signal separation and estimation. Its importance also lies in the practical need for some communication receivers to equalize unknown channels without the assistance and the expense of training sequences.

Current methods of joint power control and beamforming [47, 95, 96, 97, 98] assume perfect measurement of channel parameters and SINR at the receivers, which is very difficult to obtain in practice. Blind beamforming can estimate and separate, without the use of training sequences, the transmitted signals that suffer from the channel distortion and additive noise. The difficulties for joint power control and blind beamforming are to formulate such a cross layer problem into a joint optimization problem, and develop an algorithm that can be self-trained and adaptively adjust the system parameters. In our approach, we present a novel joint power control and blind beamforming algorithm for a multi-cell multi-antenna system. Based on a reformulated joint problem, our proposed algorithm optimizes the Bit Error Rate (BER) using a quantity directly available from the blind beamforming and estimation, which avoids additional measurements mentioned above. Mobiles' transmit powers are updated in a distributed manner such that the CCI is effectively reduced. Convergence properties of the proposed algorithm are discussed. A Cramer-Rao lower bound (CRB) is derived to show the effect of power control on the symbol estimation performance in the networks. Simulation results illustrate that our algorithm converges to the desired solution and is more robust to channel estimation error compared with traditional joint power control and training based beamforming algorithm.

System Model, Beamforming and Power Control

Consider K distinct cells in wireless networks where co-channel links exist. Each cell consists of one base station and its assigned D mobiles. Antenna arrays with M elements are used only at the base station and $M \geq D$. We assume coherent detection is possible so that it is sufficient to model this multiuser system by an equivalent baseband model. Each link is affected by the slow Rayleigh fading. The propagation delay is far less than one symbol period. For uplink case, the i^{th} base station antenna array's output vector is given by:

$$\mathbf{x}_i(t) = \sum_{k=1}^{K} \sum_{d=1}^{D} \sqrt{G_{ki}^d P_k^d} \alpha_{ki}^d \mathbf{a}_{ki}^d(\theta_{ki}^d) \cdot g_k^d(t - \tau_{ki}) s_k^d(t - \tau_{ki}) + \mathbf{n}_i(t) \tag{6.33}$$

where G_{ki}^d is path loss, α_{ki}^d is fading coefficient, P_k^d is transmit power, $\mathbf{a}_{ki}^d(\theta_{ki}^d)$ is the i^{th} base station array response vector to the signal from the d^{th} mobile in the k^{th} cell at direction θ_{ki}^d, $g_k^d(t)$ is shaping function, $s_k^d(t)$ is message symbol, τ_{ki} is the delay, and $\mathbf{n}_i(t)$ is thermal noise vector. We assume the synchronous transmission for all the users within the same cell, i.e. $\tau_{ii} = 0, \forall i$. The synchronous assumption is reasonable because the symbol timing can be effectively controlled within each cell. We assume the CCI from other cells is asynchronous for the desired signals within the cell and $\tau_{ki}, k \neq i$ is uniformly distributed within the symbol duration. We assume the channels are flat fading and stable within a frame of hundreds of symbols. Define the impulse response from the d^{th} mobile in the k^{th} cell to the p^{th} element of the i^{th} base station as: $h_{ki}^{dp} = \alpha_{ki}^d a_{ki}^{dp}(\theta_{ki}^d) r_{ki}^{dp}$, where r_{ki}^{dp} includes the effect of the transmitter, receiver filter, and shaping function $g_k^d(t - \tau_{ki})$. In the vector form, it is given by $\mathbf{h}_{ki}^d = [h_{ki}^{1d}, \ldots, h_{ki}^{Md}]^T$. The sampled received vector for this DK users and MK antenna outputs multi-cell system at time n is given by:

$$\mathbf{X}(n) = \mathbf{AS}(n) + \mathbf{n}(n) \tag{6.34}$$

where $\mathbf{X}(n) = [\mathbf{x}_1^T(n), \mathbf{x}_2^T(n) \ldots \mathbf{x}_K^T(n)]^T$, $\mathbf{S}(n) = [\mathbf{S}_1^T(n), \mathbf{S}_2^T(n) \ldots \mathbf{S}_K^T(n)]^T$, $\mathbf{S}_i(n) =$

$[s_i^1(n), \dots s_i^D(n)]^T$, $\mathbf{n}(n)$ is the sampled thermal noise vector, and

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{21} & \dots & \mathbf{A}_{K1} \\ \mathbf{A}_{12} & \mathbf{A}_{22} & \dots & \mathbf{A}_{K2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{A}_{1K} & \mathbf{A}_{2K} & \dots & \mathbf{A}_{KK} \end{bmatrix}_{MK \times DK} \tag{6.35}$$

where $\mathbf{A}_{ij} = [\sqrt{P_i^1 G_{ij}^1}\mathbf{h}_{ij}^1 \dots \sqrt{P_i^D G_{ij}^D}\mathbf{h}_{ij}^D]$.

Let $\mathbf{w}_i^d$ be the beamforming weight vector for the d^{th} mobile in the i^{th} cell. Without loss of generality, we normalize the beamformer weight vector $\|(\mathbf{w}_i^d)^H \mathbf{h}_{ii}^d\|^2 = 1$, which will not change the receivers' SINRs. We assume the transmitted signals from different sources are uncorrelated and zero mean, and the additive noise is spatially and temporally white with variance $\mathbf{N}_i = \sigma^2 \mathbf{I}_{M \times M}$, where σ^2 is the thermal noise variance. The d^{th} user's SINR at the its associated i^{th} base station's beamformer output is:

$$\Gamma_i^d = \frac{P_i^d G_{ii}^d}{\sum \sum_{(k,j) \neq (i,d)} P_k^j G_{ki}^j \|(\mathbf{w}_i^d)^H \mathbf{h}_{ki}^j\|^2 + (\mathbf{w}_i^d)^H \mathbf{N}_i \mathbf{w}_i^d}. \tag{6.36}$$

The issue in question here is how to find the users' beamforming vectors and transmit powers such that each user has the desired link quality and does not introduce unnecessary CCI to other users. In the rest of this part, we will briefly illustrate the traditional joint power control and beamforming.

An adaptive antenna array is designed to receive the signals from the desired directions and attenuate signals' radiations from other directions of no interest. The outputs of the array elements are weighted by a beamformer. In order to suppress the interferences, the beamformer places its nulls in the directions of interference sources and steers to the direction of the target signal. Some most popular beamformers are MMSE and MVDR beamformers[92]. In our approach, we will compare joint power control and MVDR beamforming method with our proposed blind scheme, because MVDR beamformer is commonly used in the literature [47].

If the channel responses $\mathbf{h}_{ii}^d$ can be estimated, the beamforming vector can be calculated by the MVDR method, which minimizes the total interferences at the output of a beamformer, while the gain for the desired d^{th} user in the i^{th} cell is kept as a constant. The MVDR problem can be defined as:

$$\min_{\mathbf{w}_i^d} \|(\mathbf{w}_i^d)^H \mathbf{x}_i\|^2 \ , \tag{6.37}$$

$$\text{subject to } \|(\mathbf{w}_i^d)^H \mathbf{h}_{ii}^d\|^2 = 1, \ i = 1, ..., M.$$

Define correlation matrix as $\mathbf{\Phi}_i = E[\mathbf{x}_i \mathbf{x}_i^H]$. The optimal weight vector is given by:

$$\widehat{\mathbf{w}}_i^d = \frac{\mathbf{\Phi}_i^{-1} \mathbf{h}_{ii}^d}{(\mathbf{h}_{ii}^d)^H \mathbf{\Phi}_i^{-1} \mathbf{h}_{ii}^d}. \tag{6.38}$$

In traditional power control schemes, the overall transmit powers of all links are minimized, while each link's transmit power is selected so that its SINR is equal to or larger than a fixed and predefined targeted SINR threshold γ_i^d required to maintain the link quality. The power control problem can be defined as:

$$\min_{P_i^d} \sum_{i=1}^{K} \sum_{d=1}^{D} P_i^d \ , \tag{6.39}$$

$$\text{subject to } (\mathbf{I} - \mathbf{BF})\mathbf{P} \geq \mathbf{u}$$

where $\mathbf{u} = [u_1^1, \ldots, u_1^D, \ldots, u_K^1, \ldots, u_K^D]^T$, $\mathbf{P} = [P_1^1, \ldots P_1^D, \ldots P_K^1, \ldots, P_K^D]^T$, $\mathbf{I}$ is the identical matrix, $\mathbf{B} = diag\{\gamma_1^1, \ldots \gamma_1^D, \ldots \gamma_K^D, \ldots, \gamma_K^D\}$, and

$$[\mathbf{F}]_{kj} = \begin{cases} 0 & \text{if } j = k, \\ \frac{G_{i'i}^{d'} \|(\mathbf{w}_i^d)^H \mathbf{h}_{i'i}^{d'}\|^2}{G_{ii}^d} & \text{if } j \neq k \end{cases} \tag{6.40}$$

where $i = \lfloor k/D \rfloor$, $d = mod(k, D)$, $i' = \lfloor j/D \rfloor$, $d' = mod(j, D)$, and $k, j = 1 \ldots KD$.

If the spectral radius $\rho(\mathbf{BF})$ [67], i.e. the maximum eigenvalue of BF, is inside the unit circle, the system has feasible solutions and there exists a positive power allocation vector to achieve the desired targeted SINRs. By Perron-Frobenius theorem [67, 91], the optimum power vector for this problem is $\widehat{\mathbf{P}} = (\mathbf{I} - \mathbf{BF})^{-1}\mathbf{u}$. Many adaptive algorithms [94, 47, 106] have been developed to reduce the system complexity by the following distributed iteration:

$$P_i^d(n+1) = \frac{\gamma_i}{G_{ii}^d} I_i^d \tag{6.41}$$

where $I_i^d = (\mathbf{w}_i^d)^H \mathbf{N}_i \mathbf{w}_i^d + \sum_{(k,j)\neq(i,d)}^{K,D} \|(\mathbf{w}_i^d)^H \mathbf{h}_{ki}^j\|^2 P_k^j G_{ki}^j$ and I_i^d can be easily estimated at the receivers. The power allocation is balanced at the equilibrium when the power update in (6.41) has converged.

The level of CCI depends on both channel gain and transmit power. The optimal beamforming vector may vary for different powers. Hence the beamforming and power control should be considered jointly. In [47], a joint power control and beamforming scheme has been proposed. An iterative algorithm is developed to jointly update the transmit powers and beamformer weight vectors. The algorithm converges to the jointly optimal transmit power and beamforming solution. The joint iterative algorithm can be summarized by the following two steps:

Beamforming in Physical Layer: MVDR Algorithm,

Power Update in MAC Layer: $\mathbf{P}^{n+1} = \mathbf{BFP}^n + \mathbf{u}$,

where power update step can be implemented by using only local interference measurement. But the algorithm assumes the knowledge of SINR, and directions of the desired signals or the perfect measurements of channel responses, which are very difficult to get in practice.

Joint Power Control and Blind Beamforming

In this part, first we consider how to choose a blind beamforming algorithm that can be used for joint optimization with power control. Then we reformulate the joint power control and blind beamforming problem as a cross layer approach. Finally an adaptive iterative algorithm is developed.

Choosing a Blind Beamforming Algorithm

The traditional beamforming needs the measurement of spatial responses of the array. A common practice is the use of training sequences [92]. However it costs bandwidth which is very precious and limited in wireless networks. Moreover the measurement errors can greatly reduce the performance of beamforming. This gives us the motivation to use blind beamforming method to separate and estimate the multiple signals arriving at the antenna array. Since beamforming and power control are two different layer techniques, we need to find the blind beamforming algorithms that allow us to have joint optimization across the layers. In [103, 104], a maximum likelihood approach named iterative least squares projection (ILSP) algorithm is proposed. The algorithm explores the finite alphabet property of digital signals. The channel estimation and symbol detection can be implemented at the same time. In addition, a quantity is available for BER performance and can be used for power control optimization. In this part, we will briefly review the ILSP algorithm.

Consider the same channel module in (6.34). The d^{th} mobile inside the i^{th} cell generates binary data $s_i^d(n)$ with power P_i^d transmitted over a low delay spread Rayleigh fading channel. The channel and antenna array response is $\mathbf{h}_{ii}^d$. The sampled antenna output at the i^{th} base station is given by:

$$\mathbf{x}_i(n) = \sum_{d=1}^{D} \mathbf{h}_{ii}^d \sqrt{P_i^d G_{ii}^d} s_i^d(n) + \mathbf{v}_i(n). \tag{6.42}$$

Where $\mathbf{v}_i(n)$ includes the i^{th} base station antenna thermal noise and all the CCI from the other cells, i.e.,

$$\mathbf{v}_i(n) = \mathbf{n}_i(n) + \sum_{k=1,k\neq i}^{K} \sum_{d=1}^{D} \mathbf{h}_{ki}^d \sqrt{P_k^d G_{ki}^d} s_k^d(n) \tag{6.43}$$

where $\mathbf{n}_i(n)$ is the $M \times 1$ sampled thermal noise vector.

The ILSP algorithm works with a shifting window on data blocks of size N. Assume that the channel is constant over the N symbol periods. In the i^{th} cell, we obtain the following formulation of the l^{th} data block

$$\mathbf{X}_i(l) = \mathbf{A}_i \mathbf{S}_i(l) + \mathbf{V}_i(l) \tag{6.44}$$

where l is block number, $\mathbf{X}_i(l) = [\mathbf{x}_i(lN+1)\ \mathbf{x}_i(lN+2) \dots \mathbf{x}_i((l+1)N)]$, $\mathbf{V}_i(l) = [\mathbf{v}_i(lN+1)\ \mathbf{v}_i(lN+2) \dots \mathbf{v}_i((l+1)N)]$, $\mathbf{S}_i(l) = [\mathbf{s}_i(lN+1)\ \mathbf{s}_i(lN+2) \dots \mathbf{s}_i((l+1)N)]$, $\mathbf{s}_i(n) = [s_i^1(n) \dots s_i^D(n)]^T$, and $\mathbf{A}_i = [\sqrt{P_i^1 G_{ii}^1}\mathbf{h}_{ii}^1\ \dots \sqrt{P_i^D G_{ii}^D}\mathbf{h}_{ii}^D]$. We assume that the number of users is known or has been estimated.

The ILSP algorithm uses the finite alphabet property of the input to implement a least squares algorithm that has good convergence properties for the channel with low delay spread. The algorithm is carried out in two steps to alternatively estimate $\mathbf{A}_i$ and $\mathbf{S}_i$ as:

$$\min_{\mathbf{A}_i,\mathbf{S}_i} f(\mathbf{A}_i, \mathbf{S}_i; \mathbf{X}_i) = \|\mathbf{X}_i(l) - \mathbf{A}_i \mathbf{S}_i(l)\|^2. \tag{6.45}$$

The first step is a least square minimization problem where $\mathbf{S}_i$ is unstructured and its amplitude is continuous without considering the discrete nature of modulations, while $\mathbf{A}_i$ is fixed and equal to estimated $\widehat{\mathbf{A}}_i$. In the second step, each element of the solution $\mathbf{S}_i$ is projected to its closest discrete values $\widehat{S}_i$. Then a better estimate of $\widehat{\mathbf{A}}_i$ is obtained by minimizing $f(\mathbf{A}_i, \widehat{\mathbf{S}}_i; \mathbf{X}_i)$ with respect to $\mathbf{A}_i$, keeping $\widehat{\mathbf{S}}_i$ fixed. We continue this process until estimates of $\widehat{\mathbf{A}}_i$ and $\widehat{\mathbf{S}}_i$ are converge. The ILSP algorithm is given in Table 6.4:

Table 6.4: *ILSP Algorithm*

1.	Initial $\widehat{\mathtt{A}}_{i,0}$, Step $m = 0$;
2.	$m = m + 1$ a. $\bar{\mathtt{S}}_{i,m} = \mathtt{A}^{+}_{i,m-1}\mathtt{X}_i$, where $\mathtt{A}^{+}_{i,m-1} = (\widehat{\mathtt{A}}^{H}_{i,m-1}\widehat{\mathtt{A}}_{i,m-1})^{-1}\widehat{\mathtt{A}}^{H}_{i,m-1}$ b. projection onto finite alphabet $\widehat{\mathtt{S}}_{i,m} = proj[\bar{\mathtt{S}}_{i,m}]$ c. $\widehat{\mathtt{A}}_{i,m} = \mathtt{X}_i\widehat{\mathtt{S}}^{+}_{i,m}$, where $\widehat{\mathtt{S}}^{+}_{i,m} = \widehat{\mathtt{S}}^{H}_{i,m}(\widehat{\mathtt{S}}_{i,m}\widehat{\mathtt{S}}^{H}_{i,m})^{-1}$
3.	Repeat until $(\widehat{\mathtt{A}}_{i,m}, \widehat{\mathtt{S}}_{i,m}) \approx (\widehat{\mathtt{A}}_{i,m-1}, \widehat{\mathtt{S}}_{i,m-1})$.

Reformulation of Joint Power Control and Beamforming

In traditional joint power control and beamforming, the user's received SINR is larger than or equal to a targeted value to maintain the link quality such as the desired BER. In our approach, we proposed another quantity available from the ILSP algorithm to directly ensure each user's BER. For simplicity, we use BPSK modulation for the analysis and simulation. The other PAM or MQAM modulation methods can be easily extended in a similar way. It has been shown in [104], the error probability of ILSP algorithm is approximated by:

$$P_r(s_i^d) = Q\left(\sqrt{\frac{2}{Var[\widehat{s}_i^d(n)]}}\right) \tag{6.46}$$

where each estimated signal $\widehat{s}_i^d(n)$ has $E[\widehat{s}_i^d(n)] = s_i^d(n)$, i.e, ILSP is an unbiased estimator with variance

$$Var[\widehat{s}_i^d(n)] = 2\sigma_i^2(\mathbf{A}_i^H\mathbf{A}_i)^{-1}_{dd} \tag{6.47}$$

where, in our case, $\sigma_i^2 = E[\mathbf{v}_i(n)^H \mathbf{v}_i(n)]$ and can be estimated by:

$$\sigma_i^2 \approx \frac{1}{N}\|\mathbf{X}_i - \widehat{\mathbf{A}}_i\widehat{\mathbf{S}}_i\|^2 = \frac{1}{N}\|\mathbf{V}_i\|^2. \tag{6.48}$$

In [104], (6.47) is developed for single cell environment with additive white Gaussian noise. In our case, we need to perform optimization in multicell scenario with CCI. Because there are a large number of co-channel interference sources with similar received powers, by the central limit theorem, we can assume $\mathbf{v}_i(n)$ approaches a zero mean Gaussian vector. So (6.47) is still hold in our case. From the simulation results later, we can show that this assumption is valid.

In our proposed joint power control and blind beamforming scheme, the key issue is the quantity $Var[\widehat{s}_i^d(n)]$ which is directly related to error performance. $Var[\widehat{s}_i^d(n)]$ is a function of σ_i^2 and $\mathbf{A}_i$, so it is also a function of all $P_i^d, \forall\ i, d$. We want the maximum variance for each user's $Var[\widehat{s}_i^d(n)]$ to be less than or equal to a predefined value var_0, so that each user's BER is less than the desired value. However if var_0 is too small, each user's transmit power will be too large and cause too much CCI. Under this condition, the system may not be feasible, i.e., no matter how large the transmit powers are, the receivers cannot achieve desired BER. So we need a feasibility constraint for var_0. The reformulated joint power control and blind beamforming problem is given by:

$$\min_{P_i^d} \sum_{i=1}^{K}\sum_{d=1}^{D} P_i^d \tag{6.49}$$

$$\text{subject to } \begin{cases} Var(\widehat{s}_i^d(n)) \le var_0,\ \forall i, d, \\ var_0 \text{ is feasible.} \end{cases}$$

In order to solve this problem, we need to develop a distributed algorithm such that each user can adapt its transmit power by using only local information. We need to evaluate the feasible range of var_0 such that the system is feasible, i.e., there exists a possible power allocation vector. The convergence and optimality of the adaptive algorithm will be considered.

Adaptive Iterative Algorithm

In this part, we assume var_0 is feasible for the system. We will discuss the feasibility issue in the next. In ILSP algorithm, the iteration stops when the estimated channel response matrix and symbol matrix have converged. In the algorithm, we use the final channel response matrix $\widehat{\mathbf{A}}_i$ to substitute $\mathbf{A}_i$ in (6.47). Then the estimation of $Var(\widehat{s}_i^d(n))$ is calculated by:

$$var_i^d = 2\sigma_i^2(\widehat{\mathbf{A}}_i^H \widehat{\mathbf{A}}_i)^{-1}_{dd}. \tag{6.50}$$

In the uplink, the value of var_i^d is obtained in the base station and compared with the desired var_0. If var_i^d is too large, it means that the BER for the d^{th} user is too large and consequently the d^{th} user's power needs to be increased. If var_i^d is too small, it is unnecessary to have such a high power for the d^{th} user. Consequently, the power needs to be reduced. The power update stops when transmit powers have converged in the consecutive iterations, i.e., $var_i^d \approx var_0$. Each user's power is updated by the simple feedback of $\lambda = var_i^d/var_0$ from the base station. The power update scheme can be easily implemented in a distributed manner. In each iteration, the power is updated by:

$$P_i^d(m+1) = \lambda P_i^d(m) \tag{6.51}$$

where m is the iteration number.

With the above power update equation, we develop the following joint adaptive power control and blind beamforming algorithm. The algorithm is initialized by some feasible power allocation vector $\mathbf{P}(0)$ and some approximate channel estimation $\widehat{\mathbf{A}}_{i,0}$ [103]. The user's BER may be larger than the desired value during the initialization. In each iteration, first, ILSP blind estimate algorithm is applied to estimate the antenna array responses and the transmitted signals. Then var_i^d is calculated. The new transmit power is updated by (6.51). The iteration is stopped by comparing the power vector of the two consecutive iterations. When the algorithm stops, each user's desired BER will be satisfied. The adaptive algorithm is summarized in Table 6.5:

Table 6.5: *Joint Power Control and Blind Beamforming Algorithm*

1. Given $\mathtt{P}(0)$, var_0, $m=0$ and $\widehat{\mathtt{A}}_i = \widehat{\mathtt{A}}_{i,0}$.
2. Received data block at base station i, i. ILSP Blind Estimation to get $\widehat{\mathtt{A}}_i$ ii. For each mobile d inside i^{th} cell, $var_i^d = 2\widehat{\sigma}_i^2(\widehat{\mathtt{A}}_i^H\widehat{\mathtt{A}}_i)^{-1}_{dd}$ $\lambda = \frac{var_i^d}{var_0}$ $P_i^d(m+1) = \lambda P_i^d(m)$ iii. $\widehat{\mathtt{A}}_{i,0} = \widehat{\mathtt{A}}_i$
3. $m = m+1$. Go to step 2; Repeat until $\mathtt{P}_i(m) \approx \mathtt{P}_i(m-1)$, $\forall i$.

With the adaptive algorithm, we can construct a joint power control and blind beamforming system as shown in Fig. 6.12. The variance calculator module calculates the estimation var_i^d from the ILSP module. The updating information of transmit powers is computed by the power update module. Then the simple power update information is sent back to mobiles via the feedback channels. When the algorithm converges, the output data from the ILSP module will have the desired BER.

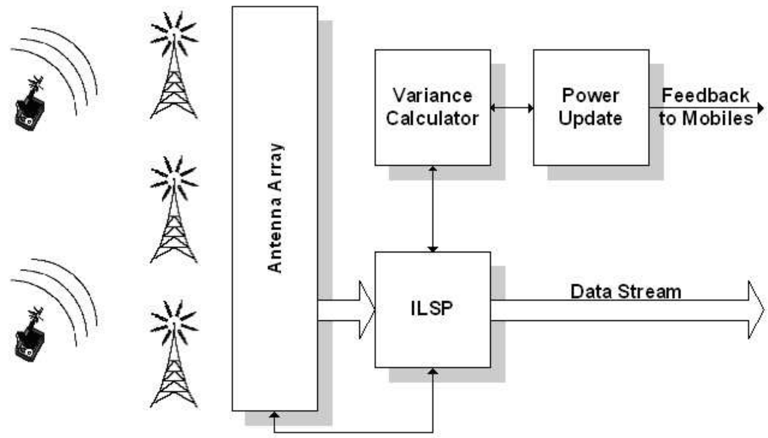


Figure 6.12: Joint Power Control and Blind Beamforming System

Analysis and Convergence of the Algorithm

Convergence Analysis

In this part, we analyze the condition for our proposed algorithm to converge, i.e., we find the feasible range for var_0. Then we prove that the power update converges to a unique solution when system is feasible, while the blind beamforming may not converge to a unique solution. So our proposed joint power control and blind beamforming algorithm may have local minima because of the inherited characteristics of the blind estimation. We will propose a method to avoid the local minima. From the simulation results, we can show that even with the possible local minima, the proposed algorithm performs comparably well with the traditional joint power control and beamforming algorithm.

Consider the transmission from the d^{th} mobile to its associated i^{th} base station with $\mathbf{h}_{ii}^d$ and G_{ii}^d being the channel response and link gain, respectively, and $\mathbf{A}_i$ being the channel response matrix. We want to find the expression $Var[\widehat{s}_i^d(n)]$ in (6.47). Then we will analyze the conditions for the convergence of our algorithm.

We have

$$[\mathbf{A}_i^H \mathbf{A}_i]_{jk} = \sqrt{P_i^j P_i^k G_{ii}^j G_{ii}^k} (\mathbf{h}_{ii}^j)^H \mathbf{h}_{ii}^k. \tag{6.52}$$

The $det(\mathbf{A}_i^H \mathbf{A}_i)$ can be expanded by the following alternating sum form:

$$det(\mathbf{A}_i^H \mathbf{A}_i) = P_i^1 G_{ii}^1 \ldots P_i^D G_{ii}^D f_1(\mathbf{h}_{ii}) \tag{6.53}$$

where $\mathbf{h}_{ii} = [\mathbf{h}_{ii}^1, \ldots, \mathbf{h}_{ii}^D]$, and $f_1(\mathbf{h}_{ii})$ is a real function of channel responses $\mathbf{h}_{ii}^d, \forall d$. Then it follows from cofactor method of matrix inverse [67] that

$$(\mathbf{A}_i^H \mathbf{A}_i)_{dd}^{-1} = \frac{f_2^d(\mathbf{h}_{ii}) \prod_{j=1, j\neq d}^{j=D} P_i^j G_{ii}^j}{f_1(\mathbf{h}_{ii}) \prod_{j=1}^{j=D} P_i^j G_{ii}^j} = \frac{f_3(\mathbf{h}_{ii})}{P_i^d G_{ii}^d} \tag{6.54}$$

where $f_2^d(\mathbf{h}_{ii})$ is a real functions of channel responses $\mathbf{h}_{ii}^j, j \neq d$, and $f_3(\mathbf{h}_{ii}) = f_2^d(\mathbf{h}_{ii})/f_1(\mathbf{h}_{ii})$.

Because the channels are not reused in the adjacent cells in most of the communication system, we assume the CCI plus thermal noise in (6.43) is Gaussian noise with the variance:

$$\sigma_i^2 = \sum_{j\neq i}^{K} \sum_{d=1}^{D} \|\mathbf{h}_{ji}^d\|^2 G_{ji}^d P_j^d + M\sigma^2. \tag{6.55}$$

Now we can calculate $Var[\widehat{s}_i^d(n)]$ as:

$$Var[\widehat{s}_i^d(n)] = \frac{2\sigma_i^2}{(\mathbf{A}_i^H \mathbf{A}_i)_{dd}} = \frac{2\sigma_i^2}{P_i^d G_{ii}^d} f_3(\mathbf{h}_{ii}). \tag{6.56}$$

An interesting result is that $Var(\widehat{s}_k^d(n))$ is independent of the transmit powers of the other mobiles in the same cell. So the main concern for power control is inter-cell CCI. Substitute into (6.51), the power update equation can be expressed as:

$$P_i^d(n+1) = \frac{\sum_{j\neq i}^{K} \sum_{d=1}^{D} \|\mathbf{h}_{ji}^d\|^2 G_{ji}^d P_j^d + M\sigma^2}{G_{ii}^d var_0} f_3(\mathbf{h}_{ii}). \tag{6.57}$$

In matrix form, we define a matrix $\mathbf{Q}$ as

$$[\mathbf{Q}]_{kj} = \begin{cases} G_{i'i}^{d'} f_4^{kj} / G_{ii}^d & \text{if } i' \neq i, \\ 0 & \text{otherwise,} \end{cases} \tag{6.58}$$

where $i = \lfloor k/D \rfloor$, $d = mod(k, D)$, $i' = \lfloor j/D \rfloor$, $d' = mod(j, D)$, and $f_4^{kj} = \|\mathbf{h}_{ji}^d\|^2 f_3(\mathbf{h}_{ii})$. The matrix expression of (6.57) for the whole network can be written as:

$$\mathbf{P}(n+1) = \frac{1}{var_0}\mathbf{Q}\mathbf{P}(n) + \mathbf{u}, \tag{6.59}$$

where $\mathbf{P} = [P_1^1 \dots P_1^D, \dots, P_K^1 \dots P_K^D]^T$, $\mathbf{u} = [u_1, \dots, u_{DK}]^T$, and

$$u_j = \frac{f_3(\mathbf{h}_{ii})M\sigma^2}{G_{ii}^d var_0}. \tag{6.60}$$

By Perron-Frobenius theorem [67], the power update in (6.59) has the equilibrium

$$\mathbf{P} = (\mathbf{I} - \frac{1}{var_0}\mathbf{Q})^{-1}\mathbf{u}. \tag{6.61}$$

If $(\mathbf{I} - \frac{1}{var_0}\mathbf{Q})$ is positive definite, i.e., the spectrum radius $|\rho(\mathbf{Q})| < var_0$, the positive power vector exists and the power update converges. Under this condition, the system is converge when $Var[\widehat{s}_i^d(n)] = var_0$. From the simulation results, we will see that our algorithm converges rapidly to the desired var_0, if $|\rho(\mathbf{Q})| < var_0$.

When var_0 is too small and less than $\rho(\mathbf{Q})$, the system is not feasible and the adaptive algorithm diverges. In order to prevent the algorithm from diverging, the system will detect the severity of CCI. If the system detect $\rho(\mathbf{Q})$ approaches var_0 or the transmit powers increase very fast, var_0 will be increased so that users will reduce their transmit powers and CCI will be alleviated.

Following the same proof in [106], we can prove that the power update in (6.57) converges to a unique solution. Suppose $\widehat{\mathbf{P}}$ and $\mathbf{P}^*$ are two different converge power allocation vectors. Without loss of generality, we assume $\beta = max_l(\widehat{P}_l^d / P_l^{d*}) > 1$, such that $\beta\mathbf{P}^* \geq \widehat{\mathbf{P}}$. We can find an index i such that $\beta P_i^{d*} = \widehat{P}_i^d$. We have

$$\widehat{P}_i^d = \frac{\sum_{j \neq i}^K \sum_{d=1}^D \|\mathbf{h}_{ji}^d\|^2 G_{ji}^d \widehat{P}_j^d + M\sigma^2}{G_{ii}^d var_0} f_3(\mathbf{h}_{ii})$$

$$
\begin{aligned}
&\leq \frac{\sum_{j\neq i}^{K}\sum_{d=1}^{D}\|\mathbf{h}_{ji}^{d}\|^{2}G_{ji}^{d}\beta P_{j}^{d*}+M\sigma^{2}}{G_{ii}^{d}var_{0}}f_{3}(\mathbf{h}_{ii}) \\
&< \beta\frac{\sum_{j\neq i}^{K}\sum_{d=1}^{D}\|\mathbf{h}_{ji}^{d}\|^{2}G_{ji}^{d}P_{j}^{d*}+M\sigma^{2}}{G_{ii}^{d}var_{0}}f_{3}(\mathbf{h}_{ii}) \\
&= \beta P_{i}^{d*}. \qquad (6.62)
\end{aligned}
$$

The above contradiction implies that the power update equation (6.51) will converge to a unique solution. However because the solution of blind beamforming may not be unique [104], our proposed joint scheme may fall into local minima. In order to prevent such local minima, we propose the following scheme to avoid the local minima.

When the two users are not well separated in the angle, i.e., the array response $\mathbf{A}_i$ is ill-conditioned. The ILSP algorithm can converge to some fixed points that are not the global minima. In this case, instead of projecting unstructured continuous estimated symbols to the closest discrete values in ILSP algorithm, we enumerate over all Ω^D possible vectors $\mathbf{S}_i^j \in \Omega^D$ and choose the one that minimizes

$$
\widehat{\mathbf{S}}_i(n) = arg \min_{\mathbf{S}_i^j \in \Omega^D} \|\mathbf{X}_i(n) - \mathbf{A}_i\mathbf{S}_i^j\|^2, \forall j \qquad (6.63)
$$

where Ω is the modulation constellation alphabet. This enumerating method has a better performance but a higher complexity. If the global minimum is still not achieved, it has been shown in [103], usually one or two re-initializations with random guess are sufficient to yield the global minimum. So we can have two or three parallel structures with different initial values to calculate ILSP algorithm. Then we select the minimal one. The probability of staying in a local minimum will be greatly reduced.

Cramer-Rao Lower Bound

In our proposed joint power control and blind beamforming system, the performance of each user's BER is determined by the noise variance, channel conditions, and power allocation. When the additive noise is a zero mean Gaussian random process, the estimation performance of the unbiased estimator is bounded by the CRB. In this part, we derive the covariance matrix for the parameters of the thermal noise variance, the input symbols, and the power allocation vector for the CRB. The results will help us analyze the effects of power control on the users' symbol estimation performances in this multi-cell system.

For simplicity, we assume the data are modulated as BPSK, i.e., $\mathbf{S}(n) \in \Omega^{KD}$, where $\Omega = \{\pm 1\}$. Similar to the performance analysis of ILSP in [104], we assume the channel responses are known (the algorithm itself doesn't need such information). The parameters for Fisher information matrix is $\vartheta = [\sigma^2, \mathbf{S}(1), \ldots, \mathbf{S}(N), \mathbf{P}]$. The likelihood function L of the received data $\mathbf{X}(n)$ is given by:

$$L[\mathbf{X}(1)\ldots\mathbf{X}(N)] = \frac{1}{(\pi\sigma^2)^{MKN}} exp\{-\frac{1}{\sigma^2}\sum_{n=1}^{N}[\mathbf{X}(n) - \mathbf{A}\mathbf{S}(n)]^H[\mathbf{X}(n) - \mathbf{A}\mathbf{S}(n)]\}. \tag{6.64}$$

The Fisher information matrix is calculated by:

$$\begin{aligned} \mathbf{I}(\theta)_{ij} &= -E\left[\frac{\partial^2 ln(L)}{\partial\theta_i\partial\theta_j}\right] \\ &= \begin{bmatrix} \frac{MKN}{\sigma^4} & 0 & \ldots & 0 & 0 \\ 0 & \mathbf{Q} & \ldots & 0 & \mathbf{R}(1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{Q} & \mathbf{R}(N) \\ 0 & \mathbf{R}(1) & \ldots & \mathbf{R}(N) & \mathbf{R}_P \end{bmatrix} \end{aligned} \tag{6.65}$$

where $\mathbf{Q}$, $\mathbf{R}(n)$, and $\mathbf{R}_P$ are derived in the Appendix.

In order to see the effect of the proposed power control on the symbol estimation

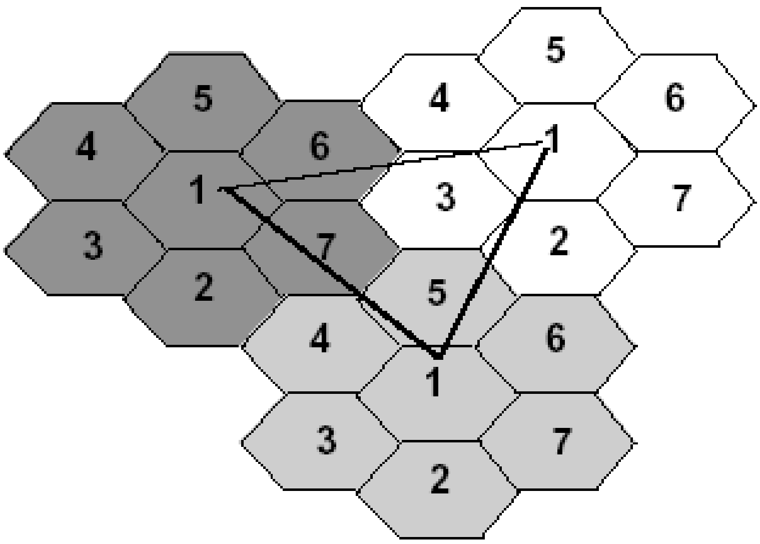


Figure 6.13: Simulation Setup

errors, we define the average mean square error (AMSE) as a performance measure of the symbol estimation:

$$AMSE = \frac{1}{N}\sum_{n=1}^{N}\frac{\|\widehat{\mathbf{S}}(n)-\mathbf{S}(n)\|^2}{\|\mathbf{S}(n)\|^2}. \tag{6.66}$$

Because we use BPSK modulation, $\|\mathbf{S}(n)\|^2 = DK, \forall\ n$ and AMSE is the variance bounded by CRB. The CRB for the symbol estimation can be obtained directly from the inverse of Fisher information matrix, i.e.,

$$AMSE \geq \frac{1}{NDK}\sum_{n=1}^{N}\sum_{j=1}^{DK}(\mathbf{I}^{-1}(\theta))_{\mathbf{S}^j(n)\mathbf{S}^j(n)} \tag{6.67}$$

where $\mathbf{S}^j(n)$ is the j^{th} element of $\mathbf{S}(n)$. How close AMSE is to the CRB will show the relative efficiency of our proposed algorithm.

Simulation Results

A network with 50 cells is simulated as shown in Fig. 6.13. Each hexagonal cell's radius is 1000m. Two adjacent cells do not share the same channel. In each cell, one base station is placed at the center. Two mobiles are placed randomly with uniform distribution. Each mobile transmits BPSK data over Rayleigh fading channels. Each base station employs four elements antenna array. The noise level is $\sigma = 1$. The transmit frame has $N = 1000$ data symbols. Our shaping function is raised cosine function.

Path loss is due to the decay of the intensity of a propagating radio wave. In our simulations, we use the two slope path loss model [110] to obtain the average received power as a function of distance. According to this model, the average path loss is given by:

$$G = \frac{C}{r^a(1 + r\lambda_c/(4h_b h_m))^b} \tag{6.68}$$

where C is a constant, r is the distance between the mobile and the base station, a is the basic path loss exponent (approximately two), b is the additional path loss component (ranging from two to six), h_b is the base station antenna height, h_m is the mobile antenna height, and λ_c is the wavelength of the carrier frequency. We assume the mobile antenna height is 2m and the base station antenna height is 50m. The carrier frequency is 900-MHz.

In Fig. 6.14, we show the analytical and numerical performance of ILSP, compared with MVDR with perfect channel estimation. The numerical results with CCI match the analytical results well especially at high SINR range, which proves our assumption that $\mathbf{V}_i(n)$ can be treated as Gaussian noise when the number of CCI is large. Our proposed joint power control and blind beamforming has only about 1-2dB performance loss over traditional power control and MVDR beamforming with perfect channel estimation. However MVDR beamforming needs additional training sequence to estimate the channel and SINR with prior information that may not be available in practice.

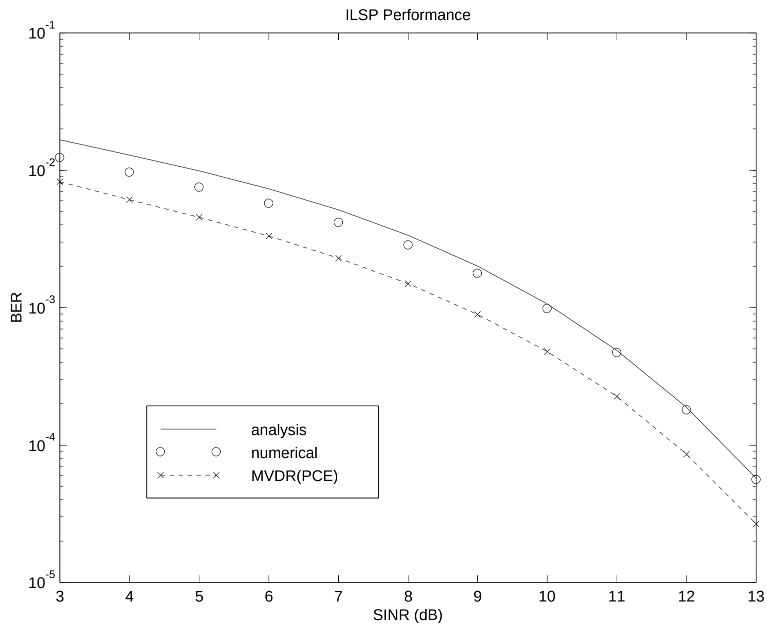
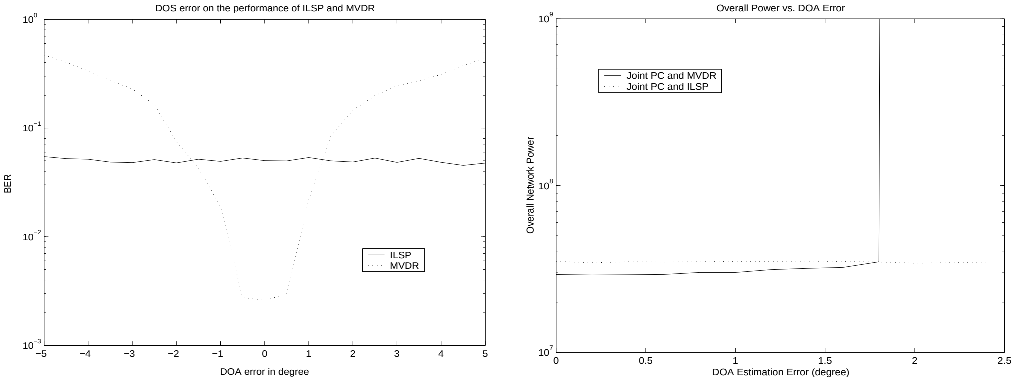


Figure 6.14: ILSP Performance

In reality, perfect channel estimation is hard to obtain. In Fig. 6.15, we show the effect of directions of arrivals (DOA) estimation error on the traditional joint power control and MVDR beamforming and our algorithm. In Fig. 6.15 (a), we compare the BER performance, while the transmit power allocation is the same for both algorithms. We can see from the curves that when the channel estimation error for DOA is greater than about 2 degree, the blind beamforming algorithm outperforms the traditional MVDR. In Fig. 6.15 (b), we compare the overall transmit power, while BER performance is the same for both algorithms. We can see that the blind beamforming algorithm needs a little bit more transmit powers when the DOA estimation error is small. However the traditional power control with MVDR method will diverge when the DOA estimation error is about 2 degree. Our proposed joint power control and beamforming algorithm will always converge regardless the DOA variations. When the mobiles are moving, DOA



(a) BER (b) Overall Power

Figure 6.15: Effects of DOA Estimation Error

are changing and this will cause the channel estimation errors. The traditional MVDR beamformer may not be aware of the changing and still use the obsolete $\mathbf{h}_{ii}^d$ in (6.38). This will greatly increase BER and transmit powers of the joint power control and MVDR method. The proposed blind scheme will automatically track and adapt to the changes and so it is more robust to channel estimation errors. Consequently, our algorithm is more robust in applications where usually only the inaccurate channel and SINR estimations are available. It is worthy to mention that the proposed scheme is more sensitive to fast channel varying and the complexity is much higher compared to the traditional training sequence based algorithm. However our scheme saves the transmission bandwidth by eliminating the training sequences and is more robust to channel estimation errors.

In Fig. 6.16, we show the numerical results of BER and overall transmit power vs. var_0 for the proposed joint blind beamforming and power control algorithm. When var_0 is decreasing from a large number, BER decreases and overall power increases slightly. Within a reasonable BER range such as BER $= 10^{-3}$ to BER $= 10^{-5}$, we can calculate the threshold of var_0 for the desired BER. Af-

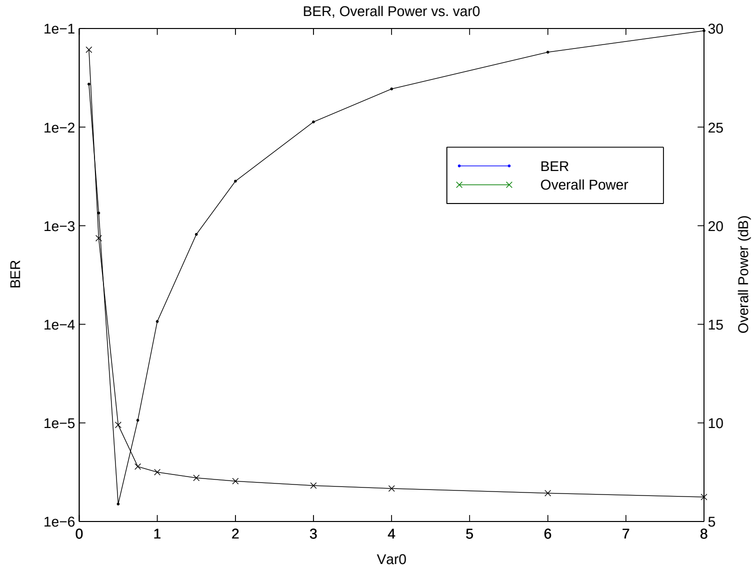


Figure 6.16: BER, Overall Power vs. var_0

ter var_0 decreases to a specific value, overall transmit power increases and BER decreases quickly. This is because the CCI is too large and $var_0 \rightarrow \rho(\mathbf{Q})$. After var_0 is smaller than some value, the algorithm diverges. Consequently, there is no feasible power control solution, i.e., no matter how large the transmit powers are, the receivers cannot ensure the desired BER. This proves that our algorithm behaves exactly the same as the traditional power control algorithm, except that our algorithm directly ensures BER instead of each user's SINR. There is a tradeoff between the overall transmit power and BER, while var_0 is the bridge between the two quantities.

In Fig. 6.17, we show the distribution of the number of iterations required for the convergence of our proposed algorithm with different values of var_0. The convergence criteria is that the maximum difference of users' transmit powers between two consecutive iterations is less than 3%. When var_0 is within the range that the system is feasible, we can see that our algorithm converges within a small number

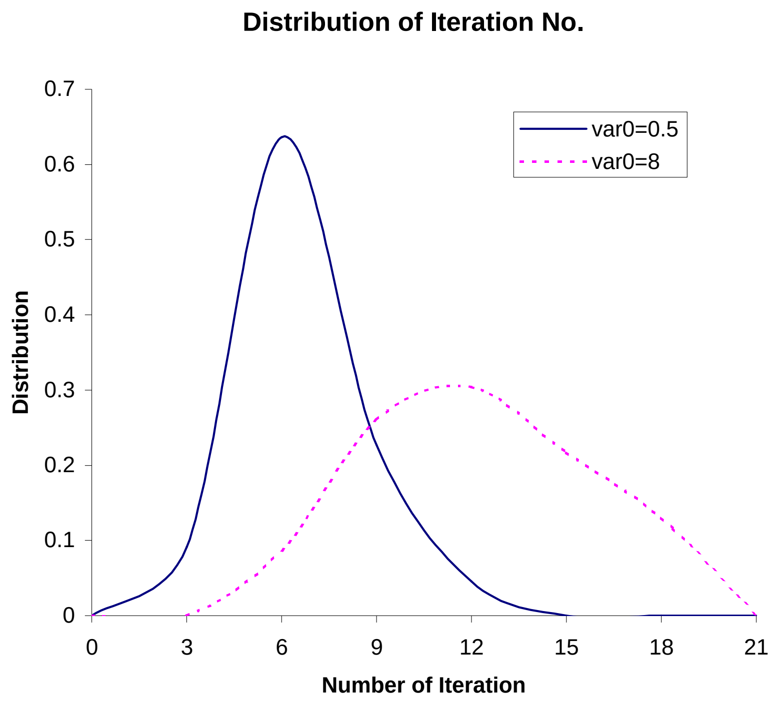


Figure 6.17: Convergence of the Algorithm

of iterations, which demonstrates that our algorithm is robust in the wireless communication systems if the channel gains and topologies have been changed. When var_0 is large, i.e., the desired BER is large, the algorithm converges slower. This is because the transmit powers are small, when var_0 is large. Consequently, the var_i^d estimation is poor and more iterations are needed for the convergence.

In Fig. 6.18, we compare the AMSE and CRB vs. var_0. When var_0 is large and the transmit powers of users are small, the CCI is small. The performance of ILSP is close to CRB. The difference is because discrete alphabets are used for transmitted symbols, while there is no such assumption for CRB. When var_0 is decreasing, the CCI and our algorithm's AMSE are decreasing because of the increasing transmit powers. In this situation, the CRB is much lower than our algorithm performance. This is because we assume all the channel conditions

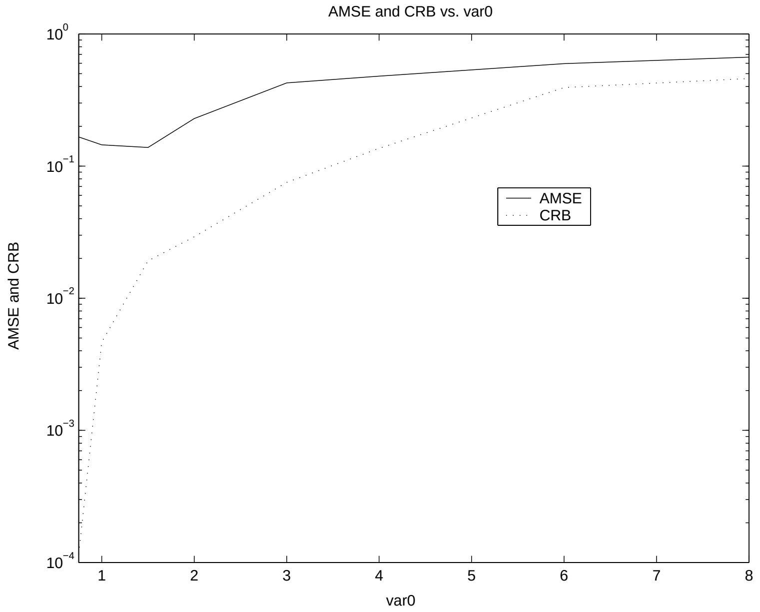


Figure 6.18: AMSE and CRB vs. var0

including $A_{ij}, i \neq j$ are known for CRB, while our algorithm only estimates A_{ii} and treats transmitted signals from other cells as noise. If an algorithm can take consideration of all $A_{ij}, \forall\ i, j$, its performance will be much better and closer to CRB, however the complexity will be unacceptably high. When var_0 is smaller than some value, our algorithm diverges. The transmit powers also diverge to arbitrary large values. But the CRB goes extremely low because SINR can be very high, if we know all the channel responses.

APPENDIX

From (6.64), the log-likelihood function is:

$$ln(L) = -MKNln(\pi) - MKNln(\sigma^2) - \frac{1}{\sigma^2}\sum_{n=1}^{N}[\mathbf{X}^H(n) - \mathbf{S}^T(n)\mathbf{A}^H][\mathbf{X}(n) - \mathbf{A}\mathbf{S}(n)]. \quad (6.69)$$

We take partial derivatives of (6.69) with respect to σ^2, $\mathbf{S}(n)$, and $\mathbf{P}$:

$$\frac{\partial ln(L)}{\partial \sigma^2} = -\frac{MKN}{\sigma^2} + \frac{1}{\sigma^4}\sum_{n=1}^{N}\mathbf{e}(n)^H\mathbf{e}(n), \tag{6.70}$$

$$\frac{\partial ln(L)}{\partial \mathbf{S}(n)} = \frac{2}{\sigma^2}Re\{\mathbf{A}^H\mathbf{e}(n)\}, \tag{6.71}$$

$$\frac{\partial ln(L)}{\partial P_i^d} = \frac{2}{\sigma^2}\sum_{n=1}^{N}Re\{\mathbf{S}^T(n)\frac{d\mathbf{A}^H}{dP_i^d}\mathbf{e}(n)\}, \tag{6.72}$$

and

$$\frac{\partial ln(L)}{\partial \mathbf{P}} == \frac{1}{\sigma^2}\sum_{n=1}^{N}Re\{diag(\mathbf{S}^T(n))diag(\frac{1}{\mathbf{P}})\mathbf{A}^H\mathbf{e}(n)\}, \tag{6.73}$$

where $\mathbf{e}(t) = \mathbf{X}(t) - \mathbf{AS}(t)$, and $diag(\frac{1}{\mathbf{P}}) = diag(1/P_1^1 \dots 1/P_1^D, \dots, 1/P_K^D)$. Using the several results that are proven in [107, 109], we have

$$E\left[\left(\frac{\partial ln(L)}{\partial \sigma^2}\right)^2\right] = \frac{MKN}{\sigma^4}, \tag{6.74}$$

$$E\left[\left(\frac{\partial ln(L)}{\partial \sigma^2}\right)\left(\frac{\partial ln(L)}{\partial \mathbf{S}(n)}\right)^T\right] = E\left[\left(\frac{\partial ln(L)}{\partial \sigma^2}\right)\left(\frac{\partial ln(L)}{\partial \mathbf{P}}\right)^T\right] = 0, \tag{6.75}$$

$$\mathbf{Q} = E\left[\left(\frac{\partial ln(L)}{\partial \mathbf{S}(n)}\right)\left(\frac{\partial ln(L)}{\partial \mathbf{S}(r)}\right)^T\right] = \frac{2}{\sigma^2}Re\{\mathbf{A}^H\mathbf{A}\}\delta_{n,r}, \tag{6.76}$$

$$\begin{aligned}\mathbf{R}_P &= E\left[\left(\frac{\partial ln(L)}{\partial \mathbf{P}}\right)^2\right] \qquad (6.77)\\ &= \frac{1}{2\sigma^2}\sum_{n=1}^{N}Re\{diag(\mathbf{S}^T(n))diag(\frac{1}{\mathbf{P}})\mathbf{A}^H\mathbf{A}diag(\frac{1}{\mathbf{P}})diag(\mathbf{S}(n))\},\end{aligned}$$

and

$$\mathbf{R}(n) = E\left[\left(\frac{\partial ln(L)}{\partial \mathbf{S}(n)}\right)\left(\frac{\partial ln(L)}{\partial \mathbf{P}}\right)^T\right] = \frac{1}{\sigma^2}Re\{\mathbf{A}^H\mathbf{A}diag(\frac{1}{\mathbf{P}})diag(\mathbf{S}(n))\}. \tag{6.78}$$

Chapter 7

Conclusions and Future Work

In this chapter, we will have summery for our research works first. We will show that why this research topic deserves a detailed study, what we have done for these topics, and what are our contributions. We want to construct a unified framework with universal a view of wireless resource allocation, show the ways to model and formulate the problems, and show the techniques that are possible applicable for the proposed problems.

Then we will list some possible future work such as: effective bandwidth and capacity, video transmission, dynamic programming over Hidden Markov Model (HMM), dynamic reinforcement learning for cooperation in multiuser system, repeated game approach, utility and pricing for multimedia transmission, and swarm intelligence for Ad Hoc networks with limited resources. We discuss briefly on what these problems are, how we could model them, what the challenges are, and how we could solve them.

7.1 Summery and Conclusions

In this dissertation, we describe the overview of wireless resource allocation. We explain what are the challenges and what are the constraints. For different network situations and different users' payload types, the optimization problem can be formulated in all kinds of different ways. The generalized constrained optimization problem is formulated and the possible solutions are discussed by using different mathematic tools.

In order to improve the system performance while maintaining the QoS for users, we explore the multi-dimension diversity. First, since users experience different channel conditions, multiuser diversity is applied to efficiently allocate resources to users. Second, since each user's channel condition fluctuates over time, we explore the time diversity such that each user can "water fill" its resources during different periods of fadings. Third, for high speed data transmission, OFDM takes advantages of frequency diversity to achieve the high spectrum efficiency. Fourth, we apply antenna array processing to have space diversity to increase system capacity by separating users with different directions of arrivals. All these diversity can be combined together to combat the detrimental effects such as time varying channel, cochannel interference, heterogeneous QoS requirement, etc.

In addition, we also consider the fairness issue in the resource allocation problem. We consider the fairness definitions such as max-min, proportional, and time average fairness. The goal is to keep fairness of resource allocation among users while keeping the system performance high.

We have discuss how to allocate resources among users. Moreover we also consider how to allocate resources within each user across layers. The advantages of this cross layer approach are that it deals with end-to-end QoS directly, reduces layer to layer overhead, and optimizes the performance globally for each user instead of optimizing within each layer. The challenges of this approach are how to model the problem and how to find an efficient way to solve the problem.

With the advanced signal processing technique, we can further improve the system performance. For example, antenna array processing, multiuser detection, space-time processing, etc. All these techniques can be applied in the existing framework.

The solutions for the proposed problems can be classified into four different categories: analysis, optimal control, game theory, and dynamic programming. Each solution has its advantages and disadvantages under different conditions. We combine them to solve the specific problems according to the wireless network scenarios.

In Chapter 1, we give the introduction for overview of the wireless network. We explain some basics related to the dissertation. We present the motivations and contributions of this dissertation.

In Chapter 2, we formulate the wireless resource allocation problem as a generalized constrained optimization problem. We give the four possible mathematics solutions and compare them.

In Chapter 3, we present the centralized resource allocation with time average fairness. We explore the multiuser and time diversity. In addition, we apply the space diversity by antenna array processing. Finally, we introduce some concepts of economy to resource allocation problems and find a solution to implement these ideas. We have three works in this topic and their conclusions are

- A joint power and throughput optimization framework is proposed to study the performance of adaptive resource allocation in wireless networks. The adaptive power minimization algorithms are constructed under the fairness constraint, by using adaptive modulation with antenna diversity to fully utilize the spectrum, to combat time-varying wireless channels and to reduce CCI. The proposed scheme can be interpreted as "water filling" each user's throughput in time domain and allocating the network throughput to different users each time. A joint power and throughput management system is built to adaptively allocate the resources. From the simulation results, the algorithms reduce the total transmitted power of mobile users by up to 7dB, which is critical in terms of battery life. The spectral efficiency is increased by up to 1.2 bit/s/Hz, which, in turn, increases the network performance.

- By adaptively managing the link quality and transmitted power, we minimize the overall transmitted power while each user's time average link quality is maintained as a constant to ensure fairness. We develop the schemes to ensure fairness and encourage some users to sacrifice their resource demands in a short period of time, with the incentive that the system performance can be improved and their sacrifices can be compensated in the future. It can be conceived that the wireless network resources are "water filling" in time domain and for different users to reduce the overall network transmitted power.

  In uplink cases, the proposed adaptive algorithm for uplink reduces 60% of the overall transmitted power of mobile users compared with that of the fixed SINR threshold scheme [47], which is very critical in terms of battery lives in mobile sets. In downlink cases, the proposed adaptive algorithm significantly saves the overall transmitted power of base stations by 60% compared with that of the algorithms in [95], which in turn increases the capacity of wireless networks. The maximal achievable SINR is extended by 4dB to 6dB toward higher SINR areas with better link qualities. When combining with beamforming, our scheme can combat CCI's in different DOA's and different channel conditions over time, which leads to a better utilization of the space-time characteristics of wireless communication.

- We propose a resource allocation framework for heterogeneous types of services. We define the QoS measure for delay sensitive applications. We introduce the concepts of credit system, user autonomy, and resource awareness. The users can borrow or lend resources from the credit system and decide when and how to use their resources within their transmission time. An adaptive algorithm is developed for the user level to feedback users' demands for throughput according to their USF and current channel conditions. An adaptive algorithm is developed for the system level to adapt resource allocation strategy according to the users' feedbacks. From the simulation results, the proposed algorithms allocate the resources to different types of users to maximize the system performance and guarantee QoS. The links can survive in the long bad channel conditions.

In Chapter 4, we optimize resource allocation by jointly considering power control and adaptive modulation using game theory. In order to achieve the system efficiency and maximize the overall network throughput, we construct NCPCG and NCTG at the user level and the system level, respectively. At the user level, the users compete for the transmitted powers and are balanced in the unique optimal Nash equilibrium. At the system level, the users compete for throughput and a game rule is designed to increase the system efficiency. A centralized adaptive algorithm with a high complexity is constructed as a performance bound. From the simulations, we can see the proposed NCPCG converges to the optimal power allocation, and with properly selected parameters, NCTG can converge to the optimal or near optimal throughput allocation for the system efficiency. Finally we compare the centralized and distributed approaches and propose the concept of hybrid systems.

In Chapter 5, we further explore frequency diversity by using OFDMA. The channel assignment is a big challenge for OFDMA resource allocation. We provide three different solutions for three different network situations by using cooperative game, non-cooperative game, and subspace methods, respectively, which are listed as:

- In cooperative game approach, we use NBS and cooperative game theory to develop a fast and fair algorithm for adaptive subcarrier, throughput, and power allocation for single cell uplink OFDMA systems. The proposed algorithm is consisted of a fast two-user bargaining algorithm and a Hungarian method for determining bargaining pairs among users. From the simulation results, the proposed algorithm shows similar performances to that of the greedy algorithm and much better performance than that of the max-min algorithm, while keeping the fairness. The most highlights of the proposed algorithm are the bargaining idea and the amazing $O(N \log N)$ complexity.
- In non-cooperative game approach, the goal is power minimization under the constraints of throughput and maximal transmitted power in multi-cell OFDMA systems. We develop a distributed game theory approach to adaptively assign the sub-channels, throughput, and powers. From the simulation results, the proposed distributed algorithm reduces the overall transmitted power up to 80% compared with the fixed assignment scheme for two-cell

case, and up to 90% compared with the pure water-filling scheme for seven-cell case when the co-channel interferences are large. As a result, the systems performances can be greatly improved.

- In subspace approach, we study how to increase the capacity for multi-cell OFDMA systems where each cell has multiple users. The difficulties are the channel assignment within each cell and power control among cells. The goal is to develop a less complex scheme to optimize the system capacity under the constraint of minimal rate and power constraints for each user by adaptive channel assignment and power allocations. We develop two algorithms for initial resource allocation and one iterative algorithm to improve the performance. From the preliminary simulation results, the proposed algorithms can provide good solutions for this complicated resource allocation problem.

In Chapter 6, we want to further optimization within each user across different layers. We give the motivations why we should do cross layer approach. We give two solutions: multimedia transmission over wireless networks and joint power control and blind beamforming:

- In our application, MAC, and physical cross layer approach, we develop a protocol to smoothly control each user's distortion by varying the source coding rate, channel coding rate, transmit rate, and transmitted power in a downlink single cell CDMA system. We develop a fast algorithm to reduce the system overall distortion under the maximal transmitted power and maximal user's distortion constraints, according to different users' current rates, channel conditions, and interferences to others. Compared with the traditional voice over CDMA scheme, the proposed scheme can greatly reduce the distortion and the required transmitted power, which, in turn, will increase the maximal number of admissible users.

- We have proposed a novel joint power control and blind beamforming algorithm that reformulates the power control problem in terms of a quantity directly related to the error performance of the estimation. First, this approach optimizes BER instead of a theoretically indirect SINR. Secondly, the algorithm does not require additional measurements of channel or SINR, which saves valuable limited bandwidth. Third, our scheme can be easily implemented in a distributed manner. Fourth, our scheme is more robust to channel estimation error. The proof of convergence of the algorithm is derived and supported by simulation results. Performance results show that our algorithm performs well in the situations where the radio spectrum is limited or the good estimations are hard to obtain.

On the whole, we give an overview of the wireless resource allocation, construct an optimization framework, and provide some possible solutions for different networks. Hope our works can help the designers of the future wireless networks implement more efficient systems.

In the following sections, we provides some possible future work in wireless resource allocation.

7.2 Effective Bandwidth and Capacity

The next-generation wireless networks such as the third generation (3G) and the fourth generation (4G) wireless systems are targeted at supporting diverse quality of service (QoS) requirements and traffic characteristics. The success in the deployment of such networks will critically depend upon how efficiently the wireless networks can support different traffic flows with QoS guarantees. To achieve this goal, mechanisms for guaranteeing QoS (e.g., admission control and resource reservation) need to be efficient and practical.

Efficient and practical mechanisms for QoS support require accurate and simple channel models. Towards this end, it is essential to model a wireless channel in terms of QoS metrics such as data rate, delay and delay-violation probability. However, the existing channel models (e.g., Rayleigh fading model with a specified Doppler spectrum) do not explicitly characterize a wireless channel in terms of these QoS metrics. To use the existing channel models for QoS support, we first need to estimate the parameters for the channel model, and then extract QoS metrics from the model. This two-step approach is obviously complex, and may lead to inaccuracies due to possible approximations in extracting QoS metrics from the models.

To address this issue, we plan to use a link-layer channel model termed the effective capacity (EC) model [29]. In this approach, the authors first model a wireless link by two EC functions, namely, the probability of non-empty buffer, and the QoS exponent of the connection. Then, the authors propose a simple and efficient algorithm to estimate these EC functions. The physical-layer analogs of these two link-layer EC functions are the marginal distribution (e.g., Rayleigh/Ricean distribution) and the Doppler spectrum, respectively. The key advantages of EC link-layer modelling and estimation are

1. ease of translation into QoS guarantees, such as delay bounds.
2. simplicity of implementation.
3. accuracy, and hence, efficiency in admission control and resource reservation.

Simulation results show that the actual QoS metric is closely approximated by the estimated QoS metric obtained from the proposed channel estimation algorithm, under a wide range of conditions. This demonstrates the effectiveness of the EC link-layer model, in guaranteeing QoS.

Conventional channel models directly characterize the fluctuations in the amplitude of a radio signal. These models are called physical layer channel models, to distinguish them from the proposed link layer channel model. The authors also consider small-scale fading model for the physical-layer channel. Small-scale fading models describe the characteristics of generic radio paths in a statistical fashion. Small-scale fading refers to the dramatic changes in signal amplitude and phase that can be experienced as a result of small changes (as small as a half-wavelength) in the spatial separation between a receiver and transmitter. Small-scale fading can be slow or fast, depending on the Doppler spread. The statistical time-varying nature of the envelope of a flat-fading signal is characterized by distributions such as Rayleigh, Ricean, Nakagami, etc.

Physical-layer channel models provide a quick estimate of the physical-layer performance of wireless communications systems (e.g., symbol error rate vs. signal-to-noise ratio (SNR)). However, physical-layer channel models cannot be easily translated into complex link-layer QoS guarantees for a connection, such as bounds on delay. The reason is that, these complex QoS requirements need an analysis of the queueing behavior of the connection, which is hard to extract from physical-layer models. Thus it is hard to use physical-layer models in QoS support mechanisms, such as admission control and resource reservation.

Recognizing that the limitation of physical-layer channel models in QoS support, is the difficulty in analyzing queues using them, the authors propose moving the channel model up the protocol stack, from the physical-layer to the link-layer. The resulting model is called an Effective capacity link model, because it captures a generalized link-level capacity notion of the fading channel.

To summarize, the Effective capacity link model that we plan to apply to resource allocation, aims to characterize wireless channels in terms of functions that can be easily mapped to link-level QoS metrics, such as delay bound. Furthermore, a novel channel estimation algorithm is needed that allows practical and accurate measurements of the Effective capacity model functions.

In our proposed future work, we plan to model the multiuser communication over this effective capacity link model to guarantee the link level QoS. The challenges are how to formulate this cross layer problem and how to find an efficient and distributed algorithm to adapt each user's effective bandwidth and effective capacity under some practical constraints.

7.3 Video Transmission

Transmitting real-time compressed videos over CDMA networks has become an emerging service. Compressed video exhibits a highly bursty rate variation due to the various complexities of different video contents and intra/inter coding mode. Many recent research works have concentrated on different aspects. A variable bandwidth retransmission scheme in an MC-CDMA system was proposed in [133]. Deep and Feng in [134] proposed a channel allocation policy by dynamically assigning more codes to an I frame in a multi-user MC-CDMA system. A joint rate and power allocation scheme for 3D-ESCOT scalable video codec was studied in [135]. An overview of current and future video over wireless was presented in [136]. The performance of CBR H.263 video over Nakagami fading channels in IS-95 CDMA systems for single-cell and multi-cell environment was studied in [137]. Chan et. al. [138] analyzed the capacity of a CDMA system supporting homogeneous H.263 video traffic. An multirate DS-CDMA system supporting heterogeneous services for QoS balance was studied in [139]. A scheme minimizing the overall power consumption of source/channel coding and transmission power was proposed in [140]. Thus, the problem of how to perform the rate adaptation, code allocation, and power control for distortion management becomes an important research topic.

In the future work, we will study the resource allocation problem to provide the subscribers with satisfactory received qualities and achieve minimal system overall distortions, while the number of multicodes and transmit power are bounded. We will design a protocol to transmit realtime FGS video sequences over downlink multi-code CDMA systems. We will develop a fast distortion management algorithm to allocate resources to each user. From the simulation results, we will show that our scheme can increases the average PSNR, and reduce the distortion compared to the modified greedy algorithm [113].

Figure 7.1 shows a block diagram of our proposed distortion management protocol to transmit FGS video over multicode CDMA. The protocol is implemented at the base station. The system resources, such as the number of codes and power, are managed to reduce the overall distortion. All users have their own FGS encoders to encode different real-time video programs. Those FGS encoders send the rate-distortion (R-D) information to the proposed protocol. The protocol assigns a variable number of codes to each user according to his/her resource needs and channel conditions. For example, an I frame requires more codes than a P frame. Also, according to the feedback of downlink channel estimations, the protocol assigns the channel coding rates and power allocations to each code. In allocating

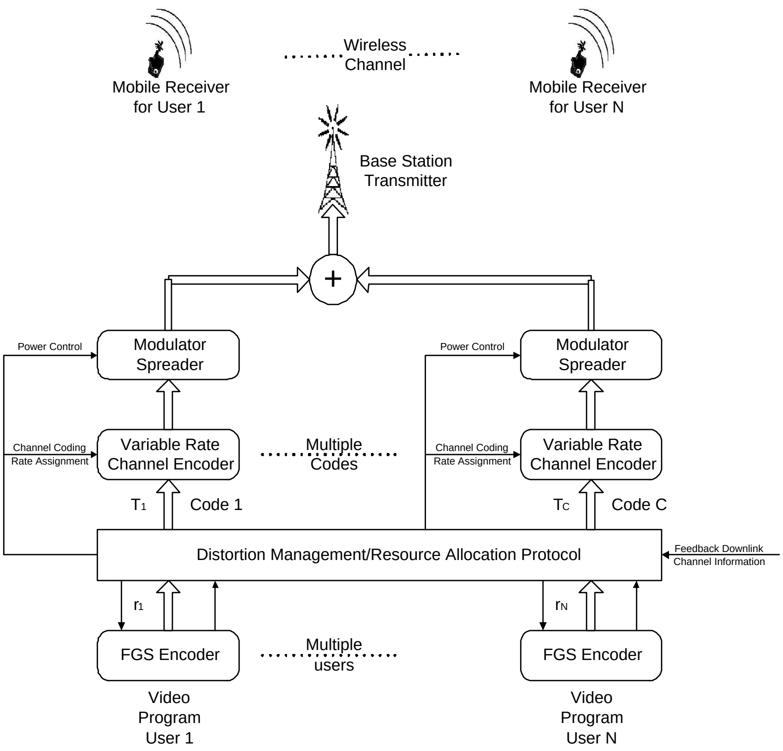


Figure 7.1: Block Diagram for the Proposed Protocol

resources, our goal is to maintain good video qualities, even when transmitting through a noisy channel with interference. Channel-induced errors affect qualities in an unpredictable way: for the same channel conditions, random errors may affect the received qualities in very different ways for different users. To avoid the uncertainty and maintain controllable video qualities, we use adaptive channel coding and power control to achieve a sufficiently small Bit Error Rate (BER). Because the number of codes and the overall transmitted power are limited, the challenge for the proposed protocol is how to efficiently allocate these resources such that the overall system distortion can be minimized.

7.4 Dynamic Programming over HMM

One of the possible future work is to model the resource allocation problem for single user over Hidden Markov Model (HMM) and solve it by dynamic programming. In this section, we give the brief review of HMM model and propose how to formulate the problem.

The Hidden Markov Model is a finite set of states, each of which is associated with a (generally multidimensional) probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are "hidden" to the outside; hence the name Hidden Markov Model.

In order to define an HMM completely, following elements are needed.

- The number of states of the model, N.
- The number of observation symbols in the alphabet, M. If the observations are continuous then M is infinite.
- A set of state transition probabilities $\Lambda = \{a_{ij}\}$.

  $$a_{ij} = p(q_{t+1} = j | q_t = i), 1 \leq i, j \leq N$$

  where q_t denotes the current state.

  Transition probabilities should satisfy the normal stochastic constraints,

  $$a_{ij} \geq 0, 1 \leq i, j \leq N$$

  and

  $$\sum_{j=1}^{N} a_{ij} = 1, 1 \leq i \leq N.$$

- A probability distribution in each of the states, $B = \{b_j(k)\}$

$$b_j(k) = p(o_t = v_k | q_t = j), 1 \leq j \leq N, 1 \leq k \leq M$$

where v_k denotes the k^{th} observation symbol in the alphabet, and o_t the current parameter vector. Following stochastic constraints must be satisfied

$$b_j(k) \geq 0, 1 \leq j \leq N, 1 \leq k \leq M$$

and

$$\sum_{k=1}^{M} b_j(k) = 1, 1 \leq j \leq N.$$

If the observations are continuous then we will have to use a continuous probability density function, instead of a set of discrete probabilities. In this case we specify the parameters of the probability density function. Usually the probability density is approximated by a weighted sum of M Gaussian distributions N,

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N(\mu_{jm}, \sum_{jm}, o_t)$$

where

$$c_{jm} = \text{weighting coefficients},$$

$$\mu_{jm} = \text{mean vectors},$$

and

$$\sum_{jm} = \text{covariance matrices}.$$

c_{jm} should satisfy the stochastic constrains

$$c_{jm} \geq 0, 1 \leq j \leq N, 1 \leq m \leq M$$

and

$$\sum_{m=1}^{M} c_{jm} = 1, 1 \leq j \leq N.$$

- The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = p(q_1 = i), 1 \leq i \leq N.$$

  Therefore we can use the compact notation

$$\lambda = (\Lambda, B, \pi)$$

  to denote an HMM with discrete probability distributions, while

$$\lambda = (\Lambda, c_{jm}, \mu_{jm}, \sum_{jm}, \pi)$$

  to denote one with continuous densities.

Once we have an HMM, there are three problems of interest.

1. The Evaluation Problem

   Given an HMM λ and a sequence of observations $O = o_1, o_2, \ldots, o_T$, what is the probability that the observations are generated by the model, $p(O|\lambda)$?

2. The Decoding Problem

   Given a model λ and a sequence of observations $O = o_1, o_2, \ldots, o_T$, what is the most likely state sequence in the model that produced the observations?

3. The Learning Problem

   Given a model λ and a sequence of observations $O = o_1, o_2, \ldots, o_T$, how should we adjust the model parameters $\{\Lambda, B, \pi\}$ in order to maximize $p(O|\lambda)$

Evaluation problem can be used for isolated (word) recognition. Decoding problem is related to the continuous recognition as well as to the segmentation. Learning problem must be solved, if we want to train an HMM for the subsequent use of recognition tasks.

Hidden Markov models (HMMs), or probabilistic functions of Markov chains, have been used extensively in various systems, including speech and image recognition, telecommunications, and queuing systems. The major reasons for the model's popularity is its ability to approximate a large variety of stochastic processes and its relative simplicity.

Digital signal transformations in the presence of noise and fading when combined with other channel impairment leads to bursty errors on the channel. HMMs are widely used to describe the bursty nature of communication channel errors. For digital wireless channels, a Markov chain can model the channel states. By allowing error probabilities to be state-dependent, we can model channel states of different error probabilities. In an HMM, a set of channel states (including the state descriptions) and the matrix of transition probabilities among states are defined.

We plan to propose a cross layer resource allocation for signal user over HMM channel. The optimal decision for resource allocation is based on the average payoff over times. We plan to model voice payload or video payload over this HMM channel. The solution might be forward algorithm, Viterbi algorithm, Baum-Welch algorithm, gradient based method, or maximum mutual information criterion.

7.5 Dynamic Reinforcement Learning for Multiuser Cooperation

For multiuser communication system, one user's action will change its interferences to others and cause others to adapt their strategies. This problem is very hard to be analyzed by dynamic programming, because the distribution of the interferences are unknown. Dynamic reinforcement learning can provide a robust and natural means for agents to learn how to coordinate their action choices in multi-agent systems. So it is nature to introduce this technique to dynamic resource allocation over wireless networks.

Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward, but also the next situation and, through that, all subsequent rewards. These two characteristics—trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning.

Reinforcement learning is defined not by characterizing learning algorithms, but by characterizing a learning problem. Any algorithm that is well suited to solving that problem we consider to be a reinforcement learning algorithm. A full specification of the reinforcement learning problem in terms of optimal control of Markov decision processes must wait until Chapter 3, but the basic idea is simply to capture the most important aspects of the real problem facing a learning agent interacting with its environment to achieve a goal. Clearly such an agent must be able to sense the state of the environment to some extent and must be able ato take actions that affect that state. The agent must also have a goal or goals relating to the state of the environment. Our formulation is intended to include just these three aspects—sensation, action, and goal—in the simplest possible form without trivializing any of them.

Reinforcement learning is different from supervised learning, the kind of learning studied in most current research in machine learning, statistical pattern recognition, and artificial neural networks. Supervised learning is learning from examples provided by some knowledgable external supervisor. This is an important kind of learning, but alone it is not adequate for learning from interaction. In interactive problems it is often impractical to obtain examples of desired behavior that are both correct and representative of all the situations in which the agent has to act. In uncharted territory—where one would expect learning to be most beneficial—an agent must be able to learn from its own experience.

One of the challenges that arises in reinforcement learning and not in other kinds of learning is the tradeoff between exploration and exploitation. To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions it has to try actions that it has not selected before. The agent has to exploit what it already knows in order to obtain reward, but it also has to explore in order to make better action selections in the future. The dilemma is that neither exploitation nor exploration can be pursued exclusively without failing at the task. The agent must try a variety of actions and progressively favor those that appear to be best. On a stochastic task, each action must be tried many times to reliably estimate its expected reward. The exploration–exploitation dilemma has been intensively studied by mathematicians for many decades (see Chapter 2). For now we simply note that the entire issue of balancing exploitation and exploration does not even arise in supervised learning as it is usually defined.

Another key feature of reinforcement learning is that it explicitly considers the whole problem of a goal-directed agent interacting with an uncertain environment. This is in contrast with many approaches that address subproblems without addressing how they might fit into a larger picture. For example, we have mentioned that much of machine learning research is concerned with supervised learning without explicitly specifying how such an ability would finally be useful. Other researchers have developed theories of planning with general goals, but without considering planning's role in real-time decision-making, or the question of where the predictive models necessary for planning would come from. Although these approaches have yielded many useful results, their focus on isolated subproblems is a significant limitation.

Reinforcement learning takes the opposite tack, by starting with a complete, interactive, goal-seeking agent. All reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments. Moreover, it is usually assumed from the beginning that the agent has to operate despite significant uncertainty about the environment it faces. When reinforcement learning involves planning, it has to address the interplay between planning and real-time action selection, as well as the question of how environmental models are acquired and improved. When reinforcement learning involves supervised learning, it does so for very specific reasons that determine which capabilities are critical, and which are not. For learning research to make progress, important subproblems have to be isolated and studied, but they should be subproblems that are motivated by clear roles in complete, interactive, goal-seeking agents, even if all the details of the complete agent cannot yet be filled in.

One of the larger trends of which reinforcement learning is a part is that towards greater contact between artificial intelligence and other engineering disciplines. Not all that long ago, artificial intelligence was viewed as almost entirely separate from control theory and statistics. It had to do with logic and symbols, not numbers. Artificial intelligence was large LISP programs, not linear algebra, differential equations, or statistics. Over the last decades this view has gradually eroded. Modern artificial intelligence researchers accept statistical and control-theory algorithms, for example, as relevant competing methods or simply as tools of their trade. The previously ignored areas lying between artificial intelligence and conventional engineering are now among the most active of all, including new fields such as neural networks, intelligent control, and our topic, reinforcement learning. In reinforcement learning we extend ideas from optimal control theory and stochastic approximation to address the broader and more ambitious goals of artificial intelligence.

7.6 Repeated Game Approach

When players interact by playing a similar stage game (such as the prisoner's dilemma) numerous times, the game is called a repeated game. Unlike a game played once, a repeated game allows for a strategy to be contingent on past moves, thus allowing for reputation effects and retribution. In infinitely repeated games, trigger strategies such as tit for tat can encourage cooperation. The basic philosophy is that : even though each user could do better in the short run by defecting instead of cooperating, for a patient user this short-run gain is outweighted by the prospect unrelenting future "punishment" from other users.

A repeated game can be defined as follows: For a stage game: $G = \{A_1, ..A_n; u_1, ..., u_n\}$ where A_i is the outcome space and u_i is the utility, the outcome of G is $a = (a_1, ..., a_n) \in A = A_1 \times ... \times A_n$. For repeated game $G(T)$, G is repeated T times. a^t is the outcome of the t^{th} repetition of G history prior to t^{th} repetition: $h^{t-1} = (a^1, ..., a^{t-1}) \in A^{t-1}$. The strategy for player i in $G(T)$ is $\sigma_i = (\sigma_i^1, ..., \sigma_i^t, ..., \sigma_i^T)$, where $\sigma_i^t : A^{t-1} \to A_i$ maps the history into an action. Stage-game payoffs: $\{u_i(a^1), ..., u_i(a^T)\}$ The payoffs for $G(T)$ can be given by the following two case:

1. average

$$U_i(h^T) = \frac{1}{T} \sum_{t=1}^{T} u_i(a^t) \tag{7.1}$$

2. discounted sum

$$U_i(h^T) = \sum_{t=1}^{T} \delta^{t-1} u_i(a^t),\ \forall \delta \in [0, 1] \tag{7.2}$$

For the game that repeated finite times, we call it finite repeated game. To analyze this kind of game, subgame perfection is the most important concept.

Consider a game G of perfect information consisting of a tree T linking the information sets $i \in I$ (each of which consists of a single node) and payoffs at each terminal node of T. A subtree T_i is the tree beginning at information set i, and a subgame G_i is the subtree T_i and the payoffs at each terminal node of T_i.

Definition 7.6.1 *A Nash equilibrium of G is subgame perfect if it specifies Nash equilibrium strategies in every subgame of G. In other words, the players act optimally at every point during the game.*

By using backward recursion, we can determines credible behavior in finite-horizon extensive-form games.

If the game continues infinitely, we call this game infinitely repeated game. The most import theory for this game is Fork Theorem:

Theorem 7.6.2 *Folk Theorem. For any $\{v_1, ..., v_n\} in V*$ if players discount the future sufficiently little $(\exists \Delta \in (0, 1) s.t. \forall \delta \in (\Delta, 1))$, there exists a Nash equilibrium of $G(\infty)$ where for all i, i's average payoff is v_i.*

By using this theorem, we can construct games that force the users to cooperate to produce better system performance, provided that each user is patient enough for long term payoff.

One of the example for repeated game is Organization of the Petroleum Exporting Countries (OPEC). Each country produces its own amount of oil per day. The organization controls the countries' amounts so that the overall profits are maximized. If some countries deviate from the assigned amount and make the oil prices drops, the other country observes the price until a threshold. If the price below this threshold, the countries believe that some other countries produce too many, so these countries will produce more as well, so this will drive the price even lower until the market comes back. Because of this mechanism, the countries have to decide if it is profitable to deviate from the assigned price because they have to pay the penalty of low price in the future. So this is an example why repeated game can force users to cooperate with each other.

In the wireless resource allocation application, we can apply the repeated game to let users efficiently share the bandwidth. For the static noncooperative game, the disadvantage is that there exist many non-optimal Nash equilibriums. For example, for the signal cell TDMA system, the user who occupy the channel first will always hold the channel regardless his channel condition. For the repeated game approach, we can define the game utility function and punishment methods such that each user will act according to the predefined optimal way to share the resource in order to avoid future punishments. The challenge is how to define the punishment stage and how to let the users be balanced in the desired states.

7.7 Utility and Pricing for Multimedia Transmission

The concept of utility is commonly used in microeconomics and refers to the level of satisfaction the decision maker receives as a result of its actions. Formally, a utility function is defined as:

Definition 7.7.1 *A function that assigns a numerical value to the elements of the action set* $A(\mu : A \to \Re^1)$ *is a utility function, if for all* $x, y \in A$*,* x *is at least as preferred compared to* y *if and only if* $u(x) \geq u(y)$*.*

The utility function that describes a particular set of preference rules is not unique. Any function that puts the elements of A in the desired order is a candidate for a utility function. The challenges are how to use the utility function to represent the users' satisfaction of QoS and how to let the system be balanced in the desired Nash equilibrium to generate the optimal system performances.

The first challenge for this approach is how to define a meaningful utility function such that it can represent the true satisfaction of the users. The problem itself is related on how to model the cross layer approach. We list some of the possible directions:

- In [31], the authors define the utility function as:

  $$u = \frac{LRP_c}{Mp} \tag{7.3}$$

  where L is the information bits in frames of size M, R is the bits/section, P_c denotes the probability of correct reception of a frame at the receiver, p is the transmitted power. The physical meaning of the utility is the number of information bits received successfully per Joule of energy expended.

- Distortion based utility function:

  The utility function represent the end-to-end quality of service directly, which can the most direct physical meaning but might be very difficult to be represented. By using some approximation, some clear and nice utility functions might be obtained.

- Utility function for delay sensitive application

  This kind of utility function will involve the dynamics and transmission histories. The utility function also depends on how the users can tolerant the delay.

- Utility function for routing purpose

  For ad Hoc network, the routing itself can be modelled as utility base optimization. We can definite the utility as functions of power, throughput, routing cost, etc, so that we can solve the problem in a distributed way.

The second challenge for utility based approach is that the utility function might not be linear or convex, consequently, there might be many local optima or Nash equilibriums. Most of them are not optimal from system optimization point of view. We need to find a way to force the users to have nice solutions of Nash equilibrium.

In the noncooperative game, each user aims to maximize its own utility by adjusting its own resource usage, but ignoring the interferences it imposes on other users. The self optimizing behavior of an individual user is said to create an externality when it degrades the quality for every other user in the system. The system performance can be greatly reduced because of individual user's greediness. So we need an efficient way to improve the system efficiency. Among the many ways to deal with externality, pricing (taxing) has been used as an effective tool both by economists and researchers in the field of computer networks. Typically, pricing is motivated by two different objectives:

1. it generates revenue for the system

2. it encourages users to use system resources more efficiently.

We plan to use pricing as a control signal to motivate users to adopt a social behavior. An efficient pricing mechanism makes decentralized decisions compatible with overall system efficiency by encouraging efficient sharing of resources rather than the aggressive competition of the purely noncooperative game. A pricing policy is called incentive compatible if pricing enforces a Nash equilibrium that improves social welfare, where social welfare can be roughly defined as the sum of the utilities.

It is possible to use various pricing polices, such as flat rate, access based, usage based, priority based, etc. This situation raises the question of which pricing policy is appropriate. The service provider determines both the pricing policy and the specific prices for the user of resources based on the system, the kind of resources it offers and the type of the demand for these services. An efficient price will reflect accurately the costs of usage of a resource and must take into account the nature of the demand for the offered service. Usage based pricing is an approach commonly encountered in literature. In usage based pricing, the price a user pays for using the resources is proportional to the amount of resources consumed by the user.

The utility function with pricing can be defined as the user's measure of QoS satisfaction minus the price. The resource allocation proceeds with an exchange of price and demand information. The base station announces a price per unit transmitted power and a price per code. Each user responds by requesting the amount of each resource that maximizes her individual surplus, defined as utility minus cost. The goal is to set prices to maximize total utility or revenue.

Unlike other approaches to resource allocation, pricing can allocate resources according to perceived user utility, thereby increasing the overall utility of the network. Other attractive properties include the accommodation of a wide range of traffic flows, and potential simplification or elimination of explicit admission control policies

However to set the appropriate price to let the distributed decision compatible with the centralized optimal decision is a hard problem. For example, to solve the following throughput maximization problem

$$\sum_{i=1}^{N} T_i \tag{7.4}$$
$$s.t. \ \ P_i \le P_{max}, \ \forall i$$

where N is the total number of users, T_i and P_i are user's throughput and power respectively, and P_{max} is the maximal transmitted power from users. The Lagrangian function can be written as

$$J = \sum_{i=1}^{N} (T_i - \lambda_i(P_i - P_{max})) \tag{7.5}$$

where λ_i is Lagrangian multiplier. The solution can be solved by differentiating the above equation, setting the results to zeros, and solving the equations by combining the constraint function.

If we define the utility function for each user as

$$u_i = T_i - \lambda_i(P_i - P_{max}), \tag{7.6}$$

as long as we can announce the optimal Lagrangian multiplier λ_i for each user, the distributed optimization can achieve the global optimal for system optimization point of view. However, to obtain optimal λ_i, we need all the channel conditions and centralized computation is necessary. So the challenge for the pricing based utility method is how to find an efficient way to calculate the price to optimize the system performances. Some heuristic and fast algorithms are desired to calculate the price.

7.8 Ad Hoc Networks with Limited Resources

Swarm Intelligence is a new computational and behavioral metaphor for solving distributed problems; it is based on the principles underlying the behavior of natural systems consisting of many agents, such as ant colonies and bird flocks . The approach emphasizes distributedness, direct or indirect interactions among relatively simple agents, flexibility, and robustness. Applications include optimization algorithms, communications networks, and robotics.

There has been growing general interest in infrastructureless or "ad hoc" wireless networks recently as evidenced by such activities as the MANET (Mobile Ad hoc NETworking) working group within the Internet Engineering Task Force (IETF). Other examples are plans unveiled for NASAs Earth orbit satellite constellation networks, and the Mars network, consisting of a "web" of satellites, rovers, and sensors within a ubiquitous information network1. The main issues inherent in such an ad hoc network are the following:

- Dynamic network topologies, presenting challenges in routing and link bandwidth allocation
- Providing consistent quality of service levels subject to a changing environment

- Conservation of power, which is essential to users of mobile wireless networks
- Global vs. local longevity, i.e., how routing may be desirable on more "long-lived" routes

Intelligent network routing, bandwidth allocation, and power control techniques are thus critical for such networks that have heterogeneous nodes with different data rate requirements and limited power and bandwidth. Such techniques coordinate the nodes to communicate with one another while exercising power control, using efficient protocols, and managing spectral occupancy to achieve the desired Quality of Service (QoS). They also let the network adapt to the removal and addition of different high and low rate communication sources, changing activity patterns, and incorporation of new services.

A large body of work exists on the general problem of network routing. Wireless networks present particular difficulties arising from the dynamic nature of their topology, due to node movement, radio interference, node failures, and new additions. A variety of routing protocols have been offered and the best-performing schemes generally depend on the specific characteristics of the operating environment (such as distribution of connectivity and topology change rates).

Although in its infancy, swarm intelligence [147, 148, 149, 150] is being intensely studied for applications in communication network routing. France Telecom and British Telecommunications (BT) have applied swarm intelligence to their phone networks. MCI Worldcom is also seriously investigating swarm intelligence for telephone network management in the United States.

The potential advantages of swarm intelligence over conventional centralized telecommunications approaches are enormously compelling. The New Scientist [148] recently gave 3 troubling details about problems with BTs network, and the companys investigation of swarm intelligence as a potential solution. BT's 24 million users are coordinated through a conventional web controller that, in 1995, was comprised of 30 programs with average memory requirements of 350 gigabytes. "Much of [the controller's]... time is spent just checking that all the elements of the network are working. It must also be constantly updated as new subscribers, new services, and new problems emerge. As it gets older it becomes harder to adapt, and a failure at the center could have potentially disastrous effects across the whole network. The distributed nature of swarm intelligence avoids the troubling bottlenecks that result from continuous use of such a centralized controller.

Presently, routing algorithms developed for sensor networks usually assume equal data volume and priority from every sensor in the network. This is often not the case however. For example, seismic and acoustic sensor networks typically have relatively low data rates while imaging and spectrometric ones need to collect high-resolution images, requiring high data rates. In a sensor network that has heterogeneous sensors with different data rate requirements and limited power and bandwidth, an intelligent sensor network routing algorithm is required not only to coordinate the existing sensors to communicate with one another by methods of power control, efficient protocols, and spectral management to achieve desired sensing goals, but also to sense and adapt to the removal and addition of different high and low rate sensors and changing activity.

Swarm based routing algorithms such as [151] derive from recent understandings of basic principles underlying the operation of biological swarms, such as ants or honeybees. These swarms, often containing thousands or tens of thousands of elements, routinely perform extraordinarily complex tasks of global optimization and resource allocation using only local information. The swarm can perform such complex tasks due to intelligence emergent from the collective of all its elements. This is while each such individual element has relatively little intelligence, incapable of understanding or modifying the swarm behavior on a global or often even a broad regional scale. As an example, while the direction-finding (routing) efficiency of an individual ant appears to be poor due to its random behavior, in fact the routing efficiency of the ant colony super-organism is extremely high as judged by the survivability of the species through finding their way to various food sources. The underlying principles governing these swarms are such that operating within a highly dynamic, random, and often-hostile environment is the routine and the norm, rather than the exception. As such, they offer tremendous insight and guidance into the development of algorithms designed to intelligently control systems with similar underlying characteristics, such as those of a wireless communication network.

Swarm-intelligent routing methods will enhance the reliability and timeliness of data transfer within a heterogeneous multi-node wireless communication network. They will significantly contribute to achieving the goal of robust pervasive communication coverage of a network. They will furthermore reduce the overhead in network growth due to their inherently scalable features.

BIBLIOGRAPHY

[1] T.S. Rappaport, "Wireles communications: principles and practice, 2nd ed." {em Prentice Hall", 2002.

[2] J.G. Proakis, "Digital Communications, 3rd ed.", *McGRAW-HILL*, 1995.

[3] William C. Jakes, "Microwave Mobile Communications," *John Wiley,* New York, 1974.

[4] S. Grandhi, R. Yates, and D. Goodman, "Resource allocation for cellular radio systems", *IEEE Trans. on Vehic. Tech.*, vol.46, no.3, pp.581-588, Aug. 1997.

[5] R. Yates, "A framework for uplink power control in cellular radio systems", *IEEE J. Sel. Areas Commun.*, vol.13, no.7, pp.1341-1348, Sep. 1995.

[6] A. Yener, R. D. Yates and S. Ulukus, "Interference management for cdma systems through power control, multiuser detection and beamforming", *IEEE Trans. Commun.*, vol.49, no.7, pp.1227-1239, Jul. 2001.

[7] W. T. Webb and R. Steele, "Variable rate QAM for mobile radio", *IEEE Trans. Commun.* vol.43, pp.2223-2230, Jul. 1995.

[8] X. Tang, M. S. Alouini, and A. J. Goldsmith, "Effect of channel estimation error on M-QAM BER performance in Rayleigh fading", *IEEE Trans. on Commun.*, vol.47, no.12, pp.1856-1864, Dec. 1999.

[9] M. S. Alouini and A. J. Goldsmith, "A unified approach for calculating error rates of linearly-modulated signals over generalized fading channels", *IEEE Trans. Commun.* , vol.47, no.9, pp.1324-1334, Sep. 1999.

[10] S. Falahati, A. Svensson, T. Ekman, and M. Sternad, "Adaptive modulation systems for predicted wireless channels", *Proc. 7th Wireless World Research Forum Workshop*, Eindhoven, Netherlands, Dec. 2002.

[11] S. Falahati, A. Svensson, T. Ekman, and M. Sternad, "Effect of prediction errors on adaptive modulation systems for wireless channels", *Proc. Radiovetenskap och Kommunikation*, pp.234-238, Stockholm, Sweden, Jun. 2002.

[12] L. Hanzo, C. H. Wong, and M. S. Yee, "Adaptive wireless transceivers: Turbo-coded, Turbo-equalized and Space-Time coded TDMA, CDMA, and OFDM systems", *John Wiley and Sons*, 2002.

[13] K. J. Hole, H. Holm, and G. E. Oien, "Adaptive multidimensional coded modulation over flat fading channels", *IEEE Journal on Selected Areas in Communications*, vol.18, no.7, pp.1153 -1158, Jul. 2000.

[14] T. M. Cover and J. A. Thomas, "Elements of information theory", *Wiley-Interscience*, 1991.

[15] A. Ephremides and B. Hajek, "Information theory and communication networks: an unconsummated union", *IEEE Trans. on Information Theory*, vol.44, no.6, pp.2416-2434, 1998.

[16] Z. Han and K. J. R. Liu, "Joint adaptive power and modulation management in wireless networs with antenna diversity", *Sam2002*, Rosslyn, Virgina, Aug. 2002.

[17] J. S. Blogh and L. Hanzo, "Adaptive modulation and adaptive antenna array assisted network performance of multi-user detection aided UTRA-like FDD/CDMA systems", *Proc. of VTC'2002*, pp. 1806-1810, Vancouver, 2002.

[18] S. L. Kim, Z. Rosberg, and J. Zander, "Combined power control and transmission rate selection in cellular networks", *Proc. of VTC*, pp.1653-1657, Amsterdam, Fall, 1999.

[19] Z. Han and K. J. R. Liu, "Adaptive SINR threshold allocation for joint power control and beamforming over wireless network", *IEEE VTC.*, pp.1548-1552, Fall 2001.

[20] P. Harley, "Short distance attenuation measurements at 900MHz and 1.8 GHz using low antenna heights for microcells", *IEEE Trans. Sel. Areas Commun.*, vol.37, pp.220-222, Nov. 1988.

[21] S. Ulukus and R. Yates, "Stochastic power control for cellular radio systems", *IEEE Trans. Commun.*, vol.46, no.6, pp.784-798, 1998.

[22] S. Kandukuri and S. Boyd, "Optimal power control in interference-limited fading wireless channels with outage-probability specifications", *IEEE Trans. on Wireless Commun.*, vol.1, no.1, pp.44-55, Jan. 2001.

[23] M. Frodigh, "Performance bounds for power control supported DCA-algorithms in highway micro cellular radio systems", *IEEE Trans. Veh. Tech.*, vol. 44, no. 2, May 1995.

[24] N. Prasad and M. Varanasi, "Analysis of decision feedback detection for MIMO Rayleigh fading channels and optimum allocation of transmitter powers and QAM constellations", *Pro. of 39th Annual Allerton Conf. on Comm. Control, and Comput.*, Monticello, IL, Oct., 2001.

[25] A. Goldsmith and S. Chua, "Variable-rate variable-power MQAM for fading channels", *IEEE Trans. on Comm.*, vol.45, no.10, pp.1218-1230, Oct. 1997.

[26] S. Papavassiliou and L. Tassiulas, "Improving the capacity of wireless networks through integrated channel base station and power assignment", *IEEE Trans. on Vehicular Tech.*, vol. 47, no. 2, pp. 417-427, 1998.

[27] A. Yener, D. Yates, and S. Ulukus, "Interference management for CDMA systems through power control, multiuser detection, and beamforming", *IEEE Trans. Commun.*, vol.49, no.7, pp.1227-1239, Jul. 2001.

[28] R. Negi and J. Cioffi, "Transmission over fading channels with channel side information and delay constraint", *Proc. of IEEE Globecom*, Rio de Janiero, Brazil, Dec. 1999.

[29] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630-643, Jul. 2003.

[30] S. Kandukuri and N. Bambos, "Multi-modal dynamic multiple access (MDMA) in wireless packet networks", *Proc. of INFOCOM 2001*, vol. 1, pp. 199-208, Anchorage, Apr. 2001.

[31] C. Saraydar, N. Mandayam, and D. Goodman, "Efficient power control via pricing in wireless data networks", *IEEE Trans. on Comm.*, vol. 50, no. 2, pp. 291-303, Feb. 2002.

[32] M. Xiao, N. B. Shroff, and E. K. P. Chong "Utility-based power control in cellularwireless systems", *Infocom* 2001.

[33] M. V. Clark, V. Erceg, and L. J. Greenstein, "Reuse efficiency in urban microcellular networks", *Proc. IEEE VTC*, pp. 421-425, Atlanta, Apr. 1996.

[34] R. A. Monzingo and T. W. Miller, *Introduction to adaptive arrays*, New York: Wiley, 1980.

[35] D. Gerlach and A. Paulraj, "Adaptive transmitting antenna arrays with feedback", *IEEE Signal Processing Lett.*, vol. 1, pp. 150-152, Oct. 1994.

[36] P. Zetterberg and B. Ottersten, "The spectrum efficiency of a base station antenna array system for spatially selective transmissions", *IEEE Transaction on Vehicular Technology,* vol. 44, pp. 651-660, Aug. 1995.

[37] G. G. Raleigh, S. D. Diggavi, V. K. Jones, and A. Paulraj, "A blind adaptive transmit antenna algorithm for wireless communication", *Proc. IEEE ICC*, pp.1949-1953, Seattle, 1995.

[38] D. P. Bertsekas, "Nonlinear programming, 2nd edition", *Athena Scienific*, 1999.

[39] G. L. Stuber, *Principles of mobile communications*, Norwell, MA: Kluwer, 1996.

[40] S. Ulukus, E. Biglieri and M. Z. Win, "Optimum modulation and multicode formats in CDMA systems with multiuser receivers", *IEEE Infocom*, Anchorage, AK, Apr. 2001.

[41] S. Ramakrishna and J.M. Holtzman, "A scheme for throughput maximization in a dual- class CDMA system", *IEEE Journal on Selected Areas in Commun.*, vol.16, no.6, pp.830- 844, Aug. 1998.

[42] R. Yates, "A framework for uplink power control in cellular radio systems", *IEEE JSAC.*, vol.13, no.7, pp.1341-1348, Sep. 1995.

[43] J. Zander and S.L. Kim, "Radio resource managment for wireless networks", *Artech House Publishers*, 2001.

[44] G. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence", *IEEE Trans. Veh. Tech.*, vol. 42, pp. 641-646, Nov. 1993.

[45] D. Mitra, "An asynchronous distributed algorithm for power control in cellular radio systems", *Fourth WINLAB Workshop on Third Generation Wireless Information Networks*, New Brunswick, NJ (1993).

[46] N. Bambos and S. Kandukuri, "Power controlled multiple access (PCMA) in wireless communication networks", *Proc. of IEEE INFORCOM*, vol.2, 2000.

[47] F.Rashid-Farrokhi, L. Tassiulas, and K.J. R. Liu, "Joint optimal power control and beamforming in wireless network using antenna arrays", *IEEE Trans. Commun.*, vol.46, no.10, pp.1313-1323, Oct. 1998.

[48] D. Tse and S. Hanly, "Multi-access fading channels: part I: Polymatroid structure, optimal resource allocation and throughput capacities", *IEEE Trans. on Information Theory*, vol.44, no.7, pp.2796-2815, Nov. 1998.

[49] K.K. Leung and L.C. Wang, "Controlling QoS by integrated power control and link adaptation in broadband wireless networks", *European Trans. on Telecomm.*, no.4, pp.383-394, Jul. 2000.

[50] A. Yener, R. D. Yates, and S. Ulukus, "Interference management for cdma systems through power control, multiuser detection and beamforming", *IEEE Trans. Commun.*, vol.49, no.7, pp.1227-1239, Jul. 2001.

[51] W. Yu, W. Rhee, and J.M. Cioffi, "Optimal power control in multiple access fading channels with multiple antennas", *Proc. of IEEE ICC*, vol.2, pp. 575 -579, 2001.

[52] K.K. Leung, C.W. Sung, W.S. Wong, and T. M. Lok, "Convergence theorem for a general class of power control algorithms", *Proc. of IEEE ICC*, vol.3, pp. 811 -815, 2001.

[53] Z. Han and K.J.R. Liu, "Power minimization under constant throughput constraint in wireless networks with beamforming", *Proc. of IEEE VTC.*, Vancouver, Sep. 2002.

[54] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems", *IEEE Trans. Commun.,* vol.47, no.6, pp.884–895, Jun. 1999.

[55] E. Armanious, D.D. Falconer, and H. Yanikomeroglu, "Adaptive modulation, adaptive coding, and power control for fixed cellular broadband wireless systems: some new insights", *WCNC*, Mar. 2003.

[56] D. Fudenberg and J. Tirole, "Game theory", *MIT Press*, Cambridge, MA, 1991.

[57] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Efficient power control via pricing in wireless data networks", *IEEE Trans. Commun.* , vol.50, no.2, pp.291-303, Feb. 2002.

[58] A. B. MacKenzie and S. B. Wicker, "Game theory in communications: motivation, explanation, and application to power control", *GLOBECOM*, Nov., 2001.

[59] M. Xiao, N. B. Shroff, and E. K. P. Chong, "Utility-based power control (UBPC) in cellular wireless systems", *INFOCOM*, Anchorage, AK, Apr. 2001.

[60] J.W. Lee, R. Mazumdar, and N.B. Shroff, "Downlink power allocation for multi-class CDMA wireless networks", *Proceedings of the INFOCOM*, N.Y., Jun. 2002.

[61] N. Feng, S. Mau, and N. Mandayam, "Pricing and power control for joint network-centric and user-centric radio resource management", *39th Allerton conference on Communication, Control, and Computing*, Allerton, IL, Oct, 2001.

[62] H. Ji and C. Huang, "Non-cooperative power control in cellular radio systems", *Wireless Networks*, vol. 4, pp.233-240, 1998.

[63] P. Zhang, S. Jordan, P. Liu, and M.L. Honig, "Power control of voice users using pricing in wireless networks", *SPIE ITcom*, Aug. 2001.

[64] Z. Han and K.J.Ray Liu, "Throughput maximization using adaptive modulation in wireless networks with fairness constraint", *Proc. of WCNC*, New Orleans, Mar. 2003.

[65] S.T. Chung and A.J. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view", *IEEE Trans. Commun.* , vol.49, pp.1561-1571, Sep. 2001.

[66] P Ormeci, X. Liu, D. L. Goeckel, and R. D. Wesel, "Adaptive bit-interleaved coded modulation," *IEEE Trans. on Commun.*, vol.49, no.9, pp.1572-1581, Sep. 2001.

[67] G. Golub, "Matrix computations, 3rd edition", *Johns Hopkins University Press*, 1996.

[68] M. Mesbahi, M.G. Safonov, and G.P. Papavassilopoulos, "Bilinearity and complementarity in robust control", *Advances in Linear Matrix Inequality Methods in Control*, SIAM, 2000.

[69] L. Vandenberghe and S. Boyd, "Semidefinite programming", *SIAM*, pp. 49-95, Mar. 1996.

[70] D.P. Bertsekas, "Nonlinear programming: 2nd edition", *Athena Scientific*, MA, 1995.

[71] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Jour. on Sel. Areas in Comm.*, vol.17, no.10, pp.1747-1758, Oct. 1999.

[72] H. Yin and H. Liu, "An efficient multiuser loading algorithm for OFDM-based broadband wireless systems," *IEEE Global Telecomm. Conference*, vol.1, pp. 103-107, 2000.

[73] W. Yu and J. M. Cioffi, "FDMA capacity of Gaussian multi-access channels with ISI," *IEEE Trans. on Comm.*, vol. 50, pp. 102-111, Jan. 2002.

[74] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," *Proc. of IEEE VTC'00*, pp. 1085-1089, 2000.

[75] C. Y. Wong, C. Y. Tsui, R. S. Cheng, and K. B. Letaief, "A real-time subcarrier allocation scheme for multiple access downlink OFDM transmission," *Proc. of IEEE VTC'99*, pp. 1124-1128, 1999.

[76] S. Pietrzyk and G. J. M. Janssen, "Multiuser subcarrier allocation for QoS provision in the OFDMA systems," *IEEE VTC fall*, Sep. 2002.

[77] I. Koutsopoulos and L. Tassiulas, "Carrier assignment algorithms in wireless broadband networks with channel adaptation," *Proc. of IEEE ICC'01*, pp. 1401-1405, 2001.

[78] H. Yaiche, R.R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband newworks", *IEEE/ACM Trans. on Networking*, vol. 8, no. 5, pp. 667-678, Oct. 2000.

[79] D. Grosu, A. T. Chronopoulos, M.Y. Leung, "Load balancing in distributed systems: an approach using cooperative games", *Proc. of IPDPS 2002*, pp. 52-61, 2002.

[80] F. Kelly, "Chargin and rate control for elastic traffic", *European Transactions on Telecommunications*, vol.28, no.1, pp.33-37, 1997.

[81] G. Owen, "Game theory, 3rd ed." *Academic Press*, 2001.

[82] H.W. Kuhn, "The Hungarian method for the assignment problem", *Naval Research Logistics*, Quarterly 2, 1955.

[83] W. Yu, G. Ginis, and J.M. Cioffi, "Distributed multiuser power control for digital subscriber lines", *IEEE Journal on Sel. Areas in Commun.*, vol.20, no.5, pp.1105-1114, Jun. 2002.

[84] W. Yu, W. Rhee, S. Boyd, and J.M. Cioffi, "Iterative water-filling for gaussian vector multiple access channels", to appear in *IEEE Transactions on Information Theory*.

[85] P. Viswanath, D. Tse, and V. Anantharam, "Asymptotically optimal water-filling in vector multiple access channels", *IEEE Transactions on Information Theory*, vol.12, no.4, pp.241 267, Jan. 2001.

[86] O. Kaya and S. Ulukus, "Optimum power control for fading CDMA with deterministic sequences", *40th Annual Allerton Conference on Communications, Control and Computing*, Oct. 2002.

[87] H. J. Su and E. Geraniotis, "A distributed power allocation algorithm with adaptive modulation for multi-cell OFDM sytems," *5th IEEE International Symposium on Spread Spectrum Techniques and Applications*, vol. 2, pp. 474-478, 1998.

[88] I. Koutsopoulos and L. Tassiulas, "Channel state-adaptive techniques for throughput enhancement in wireless broadband networks," *Proceedings of IEEE INFOCOM*, pp. 757-766, 2001.

[89] M.S. Barzaraa, "Nonlinear programming: theory and algorithms, 2nd ed.", *John Wiley & Sons*, 1993.

[90] W. Yu, W. Rhee, S. Boyd, and J.M. Cioffi, "Iterative water-filling for gaussian vector multiple access channels", to appear in *IEEE Transactions on Information Theory*.

[91] A. Mercado and K.J. R. Liu, "NP-hardness of the stable matrix in unit interval family problem in discrete time", *Systems and Control Letters*, pp. 261-265, April 2001.

[92] S. Haykin, "Adaptive filter theory", 3rd edition, *Prentice Hall*, 1996.

[93] S. Grandhi, R.Vijayan, D.J. Goodman, and J. Zander, "Centralized power control for cellular radio systems", *IEEE Trans. Veh. Technol.*, vol.42, pp. 466-468, Nov. 1993.

[94] R. Yates, "A framework for uplink power control in cellular radio systems", *IEEE J. Sel. Areas Commun.*, vol.13, no.7, pp.1341-1348, Sep. 1995.

[95] F. Rashid-Farrokhi, K.J. R. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems", *IEEE J. Sel. Areas Commun., Special Issue on Signal Processing for Wireless Communications*, vol. 16, no. 8, pp. 1437-1450, Oct. 1998.

[96] A. Mercado and K.J. R. Liu, "Adaptive QoS for wireless multimedia networks using power control and smart antennas", *IEEE Trans. Vehicular Technology*, vol. 51, no. 5, pp.1223-1233, Sep. 2002.

[97] Y.C. Liang, F.P.S. Chin, and K.J. R. Liu, "Downlink beamforming for DS-CDMA mobile radio with multimedia services", *IEEE Trans. Commun.*, vol.49, no.7, pp.1288-1298, Jul. 2001.

[98] A. Yener, R. D. Yates, and S. Ulukus, "Interference management for CDMA systems through power control, multiuser detection and beamforming", *IEEE Trans. on Comm.*, vol.49, no.7, pp.1227-1239, Jul. 2001.

[99] L. Tong, G. Xu, and T. Kailath, "Blind channel identification based on second order statistics: a frequency domain approach", *IEEE Trans. Information Theory*, vol.41, pp.329-334, Mar. 1995.

[100] B. Sampath, K.J.R. Liu, and Y.G. Li, "Deterministic blind subspace MIMO equalization", *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 5, pp.538-551, May 2002.

[101] T. Nguyen and Z. Ding, "Blind CMA beamforming for narrowband signals with multipath arrivals", *Int. Journal of Adaptive Control and Signal Processing*, vol.12, pp.157-172, Mar. 1998.

[102] Y. Li and K. J. R. Liu, "Adaptive blind multi-channel equalization for multiple signal separation", *IEEE Trans. on Information Theory*, vol. 44, pp. 2864-2876, Nov. 1998.

[103] S. Talwar, M. Viberg, and A. Paulraj, "Blind separation of synchronous co-channel digital signals using an antenna array. part I algorithms", *IEEE Trans. Sig. Proc.*, vol.44, no.5, pp. 1184-1197, May 1996.

[104] S. Talwar and A. Paulraj, "Blind separation of synchronous co-channel digital signals using an antenna array. part II Performance analysis", *IEEE Trans. Sig. Proc.*, vol.45, no.3, pp.706-718, Mar. 1997.

[105] Z. Ding and Y. Li, "Blind equalization and identification", *Signal Processing and Communications Series*, Marcel Dekker, 2001

[106] G.J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence", *IEEE Trans. Vehicular Technol.*, vol.42, Nov. 1993.

[107] H.Liu and G. Xu, "Closed-form blind symbol estimation in digital communications", *IEEE Trans. Signal Processing*, vol.43, no.11, pp.2714-2723, Nov. 1995.

[108] Z. Han and K.J. R. Liu, "Joint power control and blind beamforming in wireless networks", *IEEE ICC03*, 2003.

[109] W.M. Steedly and R.L. Moses, "The Cramer-Rao bound for pole and amplitude coefficient estimates of damped exponential signals in noise", *IEEE Trans. Sig. Proc.*, vol.41, no.3, pp.706-718, Mar. 1993.

[110] P. Harley, "Short distance attenuation measurements at 900MHz and 1.8 GHz using low antenna heights for microcells", *IEEE Trans. Sel. Areas Commun.*, vol.37, pp.220-222, Nov. 1988.

[111] D.J. Goodman, R.A. Valenzuela, K.T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communication", *IEEE Trans. Commun.*, vol.37, pp.885-890, Aug. 1989.

[112] P. Narasimhan and R. Yates, "A new protocol for the integration of voice and data over PRMA", *IEEE J. Sel. Areas Commun.*, vol.14, no.4, pp.623631, May 1996.

[113] M. Alasti and N. Farvardin, "SEAMA: a source encoding assisted multiple access protocol for wireless communications", *IEEE J. Sel. Areas in Comm.*, Sep. 2000.

[114] A.Sampath, P.S. Kumar, and J.M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system", *PIMRC'95*, 1995.

[115] S. Zhao, Z. Xiong, and X. Wang, "Joint error control and power allocation for video transmission over cdma networks with multiuser detection", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no.6, pp.425-437, Jun. 2002.

[116] A. Kwasinski, Z. Han, K. J. R. Liu, and N. Farvardin, "Power minimization under real time source distortion constraints in wireless networks", *IEEE WCNC*, 2003.

[117] C.W. Sun and W.S. Wong, "Power control and rate management for wireless multimedia cdma systems", *IEEE Trans. on Comm.*, pp.1215-1226, Jul. 2001.

[118] Zhu Han, Andres Kwasinski, K.J. Ray Liu, and Nariman Farvardin, "Pizza Party Algorithm for Real Time Distortion Management in Downlink Single-Cell CDMA Systems", *Allerton Conference*, 2003.

[119] J. Hagenauer, "Rate compatible punctured convolutional (RCPC) codes and their applications", *IEEE Transactions on Communications*, vol. 36, pp.389-399, Apr. 1988.

[120] F. Adachi, "Effects of orthogonal spreading and rake combining on DS-CDMA forward link in mobile radio", *IEICE Trans. on Communications*, vol.E80-B, no.11, pp.1703-1712, Nov. 1997.

[121] O. Awoniyi, N.B. Mehta, and L.J. Greenstein, "Characterizing the ortogonality factor in WCDMA downlinks", *IEEE Trans. on Wireless Communications*, vol.2, no.4, pp.621-625, Jul. 2003.

[122] S. Boyd and L. Vandenberghe, "Convex Optimization", *Cambridge University Press*, 2003. (http://www.stanford.edu/ boyd/cvxbook.html)

[123] R. W. Wolff, "Stochastic modeling and the theory of queues", *Englewood Cliffs, NJ: Prentice-Hall*, 1988.

[124] C.M. Barnhart, J.E. Wieselthier, and A. Ephremides, "An approach to voice admission control in multihop wireless networks", *INFOCOM*, pp.246-255, vol.1, 1993.

[125] B.S. Atal, V. Cuperman, and A. Gersho, "Speech and audio coding for wireless and network application", *Boston: Kluwer Academic*, 1993.

[126] P.T. Brady, "A model for generating on-off speech patterns in two-way conversations", *Bell Sys. Tech. J.*, vol. 48, pp. 2445-2472, Sep. 1969.

[127] A Gersho and R. M. Gray, "Vector quantization and signal compression", *Kluwer Academic Publishers*, 1992.

[128] N. S. Jayant and P. Noll, "Digital coding of waveforms", *Prentice-Hall Inc.*, 1984.

[129] DARPA TIMIT, "Acoustic-phonetic continuous speech corpus CD-ROM", Document NISTIR 4930, NIST Speech Disk 1-1.1.

[130] ETSI/GSM, "Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90 version 7.2.1 Release 1998)", Document ETSI EN 301 704 V7.2.1 (2000-04).

[131] J.S. Collura, A. McCree, and T.E. Tremain, "Perceptually based distortion measurements for spectrum quantization", *Proc. 1995 IEEE Workshop on Speech Coding for Telecommunications*, pp. 49 -50, Sep. 1995.

[132] L. Rabiner and B.H. Juang, "Fundamentals of speech recognition", *Prentice Hall PTR.*, 1993.

[133] M.R. Hueda, et. al, "Enhanced-performance video transmission in multicode CDMA wireless systems using a feedback error control scheme", *IEEE Globecom 2001* , pp.619 -626, Nov. 2001.

[134] B. Deep and W.-C. Feng "Adaptive code allocation in multicode-CDMA for transmitting H.263 video", *IEEE WCNC*, 1999.

[135] S. Zhao, Z. Xiong, and X. Wang, "Optimal Resource Allocation for Wireless Video Over CDMA Networks", *ICME* 2003.

[136] L. Hanzo, et. al, "Wireless Video Communications, Second to Third Generation Systems and Beyond", *IEEE Press*, 2001.

[137] N.H.L. Chan and P.T. Mathiopoulos, "Efficient video transmission over correlated Nakagami fading channels for IS-95 CDMA systems", *IEEE JSAC* , vol.18, no.6, , pp.996-1011, Jun. 2000.

[138] Y. S. Chan and J.W. Modestino, "A Joint Source Coding-Power Control Approach for Video Transmission over CDMA Networks", *to appear in IEEE JSAC.*

[139] L. Wang , et. al., "Optimal power allocation for a multi-rate packet CDMA system with multi-media traffic", *IEEE WCNC99.*

[140] Q. Zhang, Z. Ji, W. Zhu, and Y.-Q. Zhang, "Power-minimized bit allocation for video communication over wireless channels", *IEEE Trans. on Circ. and Sys. for Video Techn.* , Jun. 2002.

[141] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard", *IEEE Trans. on Circuits and Systems for Video Technology.* , vol.11, no.3, pp. 301-317, March 2001.

[142] H.M. Radha, M. van der Schaar, and Y. Chen, "The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP", *IEEE Transactions on Multimedia.* , Mar 2001.

[143] M. van der Schaar and H.M. Radha, "A hybrid temporal-SNR fine-granular scalability for internet video", *IEEE Trans. on Circuits and Systems for Video Technology.* , March 2001.

[144] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications", *IEEE Trans. on Circuits and Systems for Video Technology* , pp. 172-185, Feb. 1999.

[145] L. Zhao, J. Kim, C.-C. J. Kuo, "MPEG-4 FGS video streaming with constant-quality rate control and differentiated forwarding", *Proc. of SPIE, Conf. on Visual Comm. and Image Proc., 2002.*

[146] X.M. Zhang, et. al., "Constant Quality constrained rate allocation for FGS video coded videos", *IEEE Trans. on Circuits and Systems for Video Technology.* , pp. 121-130, Feb. 2003.

[147] E. Bonabeau and G. Thraulaz, "Swarm smarts", *Scientific American*, pp. 72-79, March 2000.

[148] M. Ward, "There's an ant in my phone", *New Scientist*, 24 January 1998.

[149] E. Bonabeau, M. Dorigo, and G. Thraulaz, "Swarm intelligence: from natural to artificial systems", *Oxford University Press*, 1999.

[150] M. Resnick, "Turtles, termites, and traffic jams", *MIT Press*, 1997.

[151] G. Di Caro and M. Dorigo, “AntNet: distributed stigmergetic control for communications networks”, *Journal of Artificial Intelligence Research*, vol. 9, pp. 317-365, 1998.