2007-12-03

# Optimizing Wireless Network Throughput: Methods and Applications

Pengchang Zhan
*Brigham Young University - Provo*

# OPTIMIZING WIRELESS NETWORK THROUGHPUT: METHODS AND APPLICATIONS

by

Pengcheng Zhan

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Electrical and Computer Engineering

Brigham Young University

December 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

Pengcheng Zhan

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

| | |
|---|---|
| Date | A. Lee Swindlehurst, Chair |
| Date | Brian D. Jeffs |
| Date | Randal W. Beard |
| Date | David G. Long |
| Date | Clark N. Taylor |

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Pengcheng Zhan in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

A. Lee Swindlehurst
Chair, Graduate Committee

Accepted for the Department

Michael J. Wirthlin
Graduate Coordinator

Accepted for the College

Alan R. Parkinson
Dean, Ira A. Fulton College of
Engineering and Technology

ABSTRACT

# OPTIMIZING WIRELESS NETWORK THROUGHPUT: METHODS AND APPLICATIONS

Pengcheng Zhan

Department of Electrical and Computer Engineering

Doctor of Philosophy

Ever since Marconi succeeded in his first demonstration on the possibility to communicate over the air overseas about a century ago, wireless communications have experienced dramatic improvements. Today's world sees the penetration of wireless communications into human life almost everywhere, from a simple remote control for TV to a cellular phone. With a better understanding of the adverse nature of the wireless propagation channels, engineers have been able to invent various clever techniques, i.e. Multiple Input Multiple Output (MIMO) technology, spread spectrum communications, Orthogonal Frequency Division Multiplexing (OFDM) to name a few, to achieve fast and reliable communications over each point-to-point link. Communications between multiple parties create networks. Limited Radio Frequency (RF) resources, e.g. transmit power, channel bandwidth, signaling time slots, etc., call for an optimal distribution of these resources among the users in the network. In this dissertation, two types of communication networks are of particular interest: cellular networks and mobile-relay-aided networks.

For a symmetric cellular network, where a fixed communication infrastructure is assumed and each user has similar average Signal-to-Noise Ratio (SNR), we study the performance of a Maximum SNR (Max-SNR) scheduler, which schedules the strongest user for service, with the effects of channel estimation error, the Modulation and Coding Scheme (MCS), channel feedback delay, and Doppler shift all taken into account. The degradation of the throughput of a Max-SNR scheduler due to outdated channel knowledge for a system with large Doppler shift and asymmetric users is analyzed and mathematical derivations of the capacity of the system based upon an Auto-Regressive (AR) channel model are presented in the dissertation as well. Unlike the schedulers proposed in the literature, which instantaneously keep track of the strongest user, an optimal scheduler that operates on the properties of Doppler and the average SNR of each user is proposed.

The high flexibility and easy deployment characteristics that Unmanned Aerial Vehicles (UAVs) possess endow them with the possibility to act as mobile relays to create secure and reliable communication links in severe environments. Unlike cellular communications, where the base stations are stationary, the mobility in a UAV-assisted network can be exploited to improve the quality of the communications. Herein, the deployment and optimal motion control problem for a mobile-relay-aided network is considered. A network protocol which achieves optimal throughput and maintains a certain Quality of Service (QoS) requirement is proposed from a cross-layer perspective. The handoff problem of the Access Point (AP) between various relays is studied and the effect of the mobility on the handoff algorithm is addressed.

# ACKNOWLEDGMENTS

Located in the place which has the most beautiful snow in the world, Brigham Young University has offered me five years of unforgettable journey of discovery in the research and study during my pursuit of the PhD degree, as well as the pretty mountain view and outdoor experience. Having been able to accomplish what I have reached today, I owe so much to so many people.

First of all, I want to thank Professor A. Lee Swindlehurst, my committee chair, without whose help, I wouldn't be anywhere near who I am today. In those days, which were one of the low times in my life, Professor Swindlehurst had given endless help to me. He has always been taking care of his students. It is his wisdom of research and kind personality that attract all my respect.

If I may say, Professor Swindlehurst opened the door to my research in wireless communication, I owe my thanks to those professors who equipped me with all the necessary knowledge to conduct the research in the related field. Professor Brian Jeffs endowed me with the ability to think in the statistical sense by teaching me the engineering statistics. Professor David Long has shown me the powerful tool of the statistical signal processing. Without the optimization theory I used to learn from Professor Travis Oliphant, I wouldn't be able to solve most of my research problems. Inspired by the achievements that Professor Randal Beard and his research team have accomplished, I was able to identify part of my research topics as Unmanned Aerial Vehicles (UAVs) related project. I also want to thank Professor Clark Taylor for joining my committee at short notice.

The whole idea of engineering is to put what you've learned into practice. With the help of Professor Swindlehurst, I was lucky enough to intern with Array-

Comm LLC, where I exercised my knowledge of specialty in the real world. My supervisor there, Todd Chauvin, has offered great patience as well as guidance in my research, which helped bridge the theory to practice. My colleague Ramesh Annavajjala has been extremely helpful for the discussions about the research projects that later became part of this dissertation. I also want to thank Dr. Kai Yu from Lucent Technology for his support in one of the projects we have conducted together.

A man's life always includes a lot of aspects; study and research is only one of them. Most of people would often forget to show the gratitude of their daily life. I want to thank all my friends for always being there for me, be it research related or life related. Christian Peel is one of those kind persons who not only helped me understand my research better but also taught me how to have fun. His introducing me to the hiking and rock climbing enriched my after school life in the pretty Utah county and had turned into part of my memory forever. Moving from Provo to San Jose has been a big change in my life, and fortunately enough, I have the company of my friends in ArrayComm LLC, including Yuan Liu, Zhongshan Wu and Weimin Sun, to name a few.

Last but not the least, I owe my thanks to my parents, without whose support and care, I wouldn't have accomplished so far.

# Table of Contents

## List of Tables

## List of Figures

# Chapter 1

# Introduction to Wireless Communication and Networks

About 110 years ago, Gulielmo Marconi's first demonstration of radio telegraphy marked the beginning of the epoch of wireless communication. However, it wasn't until 1979 that the first analog cellular system, i.e. the Nippon Telephone and Telegraph (NTT) system, became operational [1]. Ever since then, the ability to communicate without a tether has evolved dramatically. From TV remote controls to the Global Positioning System, from a Bluetooth mouse to cell phones and pagers, our daily life is inseparable from the ability to communicate wirelessly.

## 1.1 The Development of the Wireless Transmission Technology

Due to the inhospitable nature of the radio propagation environment, i.e. multiple propagation paths, time variation, large-scale path loss and so on [2], the wireless channel is unfriendly to reliable communication. Over decades, different clever methods have been proposed by engineers to increase the robustness of a point-to-point wireless link. One of the most famous of these technologies is error control coding. By introducing extra redundant bits for transmission, error control coding is able to: 1) detect the occurrence of the error in the received data (i.e. error detection codes), 2) correct the bit error corrupted in the noisy channel (i.e. error correction codes). Interleavers are designed to spread the codewords apart to avoid bursts of errors in order for the received data to be corrected. Error correction codes with an interleaver structure are shown to closely approach the Shannon capacity limits and have found popular applications in commercial wireless standards.

As the multimedia era approaches, the need for higher data rate communication has attracted significant interest in wideband communications. When the

channel bandwidth gets larger, the multipath nature of the wireless channel will cause frequency dependent fading. This effect will show up in the time domain as a self-interference phenomenon, i.e. previously transmitted data symbols will interfere with the current and/or future data symbols. Without compensating for this effect, the aforementioned Inter-Symbol Interference (ISI) causes an error floor in the wireless link performance. Techniques that attempt to mitigate ISI effects are referred to as equalizers. In the early days, equalization was mostly done in the time domain by modeling the ISI effects of the channel as a linear filter. With the linear ISI assumption, time domain equalization resembles deconvolution, which attempts to undo the effects induced by linear filtering. Depending on the cost function that an equalizer tries to optimize, different filters can be designed to accomplish the compensation task; examples include the Zero forcing (ZF) and Minimum Mean Square Error (MMSE) equalizers. When the equalizer is not constrained to be linear, decision feedback equalizers which exploit previously detected symbols to minimize the impact of ISI have also been developed. Adaptive equalizers that automatically track the time variation of the channels have been proposed to account for the time varying nature of the wireless channel. Modulating signals over multiple carriers opens the door to Multi-Carrier Communication, where a data stream is decomposed into a set of slower rate streams and modulated over several carriers for simultaneous transmission. This simple demultiplexing creates a group of separate data pipes, each of which is essentially a channel with a smaller effective bandwidth. The reduction of the bandwidth can eliminate ISI in the transmitted data. Data detection can be done separately for each slower stream and then multiplexed to estimate the original transmitted information. Multi-carrier communication techniques can be considered as another means of implementing an equalizer in the frequency domain.

Another modulation scheme that transmits signals by using a bandwidth that is several orders of magnitude larger than otherwise required is called *spread spectrum* modulation. From a single user point of view, this modulation scheme is highly bandwidth inefficient. However, when multiple users are present in the network, and multiple access interference is of concern, spread spectrum technology allows many

users to have access to a limited channel bandwidth without generating significant interference on other users, and is therefore very bandwidth efficient. In addition, spread spectrum modulation is widely used in military communication, where secure voice/data links are necessary. By spreading the signal energy over a large bandwidth, signals can be hidden below the background noise level to avoid interception. The de-spreading procedure at the receiver side will retrieve the desired transmitted signal energy and spread narrowband jammers to achieve a good operating SNR for signal detection. Rake receivers for spread spectrum modulation collect the energy from different propagation paths to gain a certain degree of diversity reception benefit.

The exponential decay of the transmit power with distance limits the possible coverage area of radio networks. Hence, 1) improvements in system capacity/throughput, 2) link reliability and 3) network coverage are usually major concerns in wireless network design. To enlarge the coverage area of the network, a simple solution would be to increase the transmit power; however, this will cause more interference to other co-channel users and limit the network capacity. To combat fading, coding across time or frequency will always cause a reduction in effective bandwidth. By introducing extra degrees of freedom in the spatial domain, i.e., through the use of multiple antennas, many of the problems associated with wireless communications can be greatly mitigated. Coherent combining at the transmitter, receiver or both increases the operating SNR, and improves coverage with no further power expenditure. With channel knowledge, receive antennas can properly combine independently faded signals to reduce signal fluctuation due to fading [3]. When Channel State Information at the Transmitter (CSIT) is available, similar techniques can be applied to improve link reliability; i.e., signals can be pre-combined at the transmitter to achieve transmit diversity. Even when CSIT is not possible, space-time coding [4, 5] creates a way to simultaneously obtain coding gain and diversity gain, as well as high spectral efficiency [6]. For a Multiple Input Multiple Output (MIMO) channel under favorable conditions, transmitting a few independent streams over the channel at the same time improves the link capacity/throughput linearly in the number of transmit antennas or receive antennas. Furthermore, interference cancellation can be performed to reduce

co-channel interference, and subsequently improve both link capacity (due to higher Signal-to-Interference Ratio (SIR)) and network capacity (due to possible aggressive frequency reuse factor). Multiple antenna technology to some extent resolves the conflicts among the three major design issues discussed above. How well these goals are balanced largely depends on the signaling scheme and receiver design [7].

## 1.2 Wireless Networks and OSI Models

The marriage between wireless and computer networks creates one of the most important fields of wireless network applications, i.e. the Wireless Local Area Network (WLAN). Wi-Fi is a typical WLAN implementation that provides connection in various hotspots, e.g. airports, universities or restaurants. To cover a much larger area and deal with more mobile users, Wireless Metropolitan Area Network (WMAN) technologies, e.g. IEEE 802.16, a.k.a Worldwide Interoperability for Microwave Access (WiMAX), are designed to provide fast data rate (63 Mbps per sector downlink, 28 Mbps per sector uplink for 10 MHz channel bandwidth), high vehicular speeds (60-120 km/h) and large coverage areas (up to 50 km) [8, 9, 10]. Other telecommunications standards, like Global System for Mobile Communications (GSM), Personal Communications Service (PCS), Digital Advanced Mobile Phone Service (D-AMPS) etc., have been widely used. Wireless personal area networks (WPAN) are used to maintain communication among computer devices (including telephones and personal digital assistants), and are becoming feasible with IrDA, Bluetooth and UWB technologies [11].

Despite all the advantages that wireless networks have, e.g. convenience, fast deployment, and low cost to name a few, they suffer from various problems such as co-channel interference, product compatibility, communication data rate and security issues. When different users are sharing the same Radio Frequency (RF) resources (including time slots, frequency, codes, etc.) to transmit signals simultaneously in the wireless network, they will to some extent block the reception of the rest of the users by increasing the noise level of the front end receiver as interference. This is handled by solving the problem of how to optimally allocate or distribute the

limited RF resources amongst the users, i.e. frequency reuse planning, orthogonal code design, smart time slot allocation, to name a few. In addition to the resource allocation problem, wireless networks also suffer from compatibility problems, where products from various vendors don't work together due to lower level designs which are inconsistent in the air interface. With regard to throughput, wireless networks usually cannot provide links that are as fast as their wired counterparts, e.g. ethernet connections, etc. In addition, since anyone can receive the broadcast signals of the wireless network, their vulnerability to eavesdropping is very straightforward.

By layering the network systems, the various issues pinpointed above can be decoupled and handled at different layer while remaining transparent to other layers. The primary example, the Open Systems Interconnection Basic Reference Model (OSI Model), is a layered, abstract description for communications and computer network protocol design [12, 13]. As is well known in the literature [12], the OSI models break a typical network down into seven layers, i.e. physical layer, data link layer, network layer, transport layer, session layer, presentation layer and application layer, from the bottom to the top. The correct functionality of each individual layer depends on the service provided by the layer that is exactly one layer below and is responsible for the requests from the layer that is exactly one layer above. Depending on different network setups, some of the layers can be combined and some of them can be omitted in each specific implementation. This dissertation is concerned with the design of a few lower layers, including the physical layer, the data link layer, and to some degree the network layer.

As the most fundamental layer of the network, the physical layer is responsible for providing the means to transmit raw data bits over the data link layer (one layer up) to connected network nodes [12]. Transmitters at the physical layer encode the data bits, map them into complex data symbols and modulate the symbols onto a certain set of carriers to be broadcast as an analog signal over some physical media. A number of services performed at this layer include: 1) inter-nodes bit stream transmission, 2) standard interface for the transmission media, 3) baseband signal modulation, 4) bit stream synchronization, 5) carrier sense and collision detection,

6) equalization and pulse shaping, 7) forward error correction, 8) channel coding, and 9) bit-interleaving, etc [12]. In a word, the details of the physical transmission schemes are all determined at this level. Besides, the formation of the physical network topology is also of concern to this specific layer, e.g. bus, ring, mesh or star network.

With the service grade guaranteed from the physical layer, the data link layer needs to provide the functional and procedural means to transfer data between network entities and might provide the means to detect and possibly correct errors that may occur in the Physical layer [13]. It typically consists of two sub-layers: the Media Access Control (MAC) layer and the Logical Link Control (LLC) layer. The MAC layer provides protocol and control mechanisms for multiple users to be able to share the resources in the wireless channels. Depending on the mode the networks are operating in, wireless networks can be categorized into packet switching networks and circuit switching networks. In the packet switching mode, data packets at each node are buffered and queued, resulting in various delay times. In the circuit switching mode, a dedicated circuit or channel is created for each pair of nodes, resulting in a constant delay. Network users apply different channel access techniques for different networking modes. In a packet switching system, typical channel access techniques include [14, 15]: 1) contention based random access methods, e.g. ALOHA, Carrier Sense Multiple Access (CSMA), Carrier Sense Multiple Access with Collision Detection (CSMA/CD), etc, 2) resource reservation (scheduled) packet-mode protocols: Dynamic Time Division Multiple Access (Dynamic TDMA), Packet Reservation Multiple Access (PRMA), Reservation ALOHA (R-ALOHA) to name a few. However, when networks operate in circuit switching mode, channel access methods include [14, 15]: 1) Frequency Division Multiple Access (FDMA), 2) Time Division Multiple Access (TDMA), 3) Code Division Multiple Access (CDMA), 4) Space Division Multiple Access (SDMA), etc.

## 1.3 Schedulers and Practicalities in Communication Systems

When multiple users are competing for network bandwidth, intelligent algorithms should be designed to distribute the resources among them for the purpose of maximizing the throughput of the network. A scheduler is one of the units that can fulfill this job. The designing of schedulers falls within the responsibilities of the MAC layer. Schedulers make user selection decisions based on various degrees of channel knowledge, e.g. channel-aware-only scheduler, channel-aware and queue-aware scheduler, etc. One particular type of the scheduler is called the *opportunistic multiuser scheduler*, which distributes resources based on the instantaneous channel condition of the active users and gives priority to the users with favorable channel conditions. Adaptive modulation and coding schemes are typically implemented at the physical layer for a system applying opportunistic schedulers.

For schedulers that only have channel knowledge of the active users available, common methods to distribute the time slot resources include: 1) Round Robin (RR) scheduling, which sequentially selects the user in the network for transmission (e.g. GSM system), 2) Proportional Fair (PF) scheduling, which schedules the user with the largest normalized instantaneous date rate obtained by dividing the instantaneous rate by its own average data rate (e.g. IS-856), 3) Maximum SNR (Max-SNR) scheduling, which selects the user with the strongest channel gain, etc. The different requirements and goals of various wireless system justify the existence of all these various schedulers. The RR scheduler has no bias toward any party in the network, and treats all users the same. However this scheduler wastes the time resource that could otherwise be dedicated to users with better channel conditions, thus causing a network throughput loss. The Max-SNR technique prioritizes the best total throughput for the network, but its greedy behavior would consistently render no channel access to those users with bad channel condition (due to possibly low transmit power in a system without power control). The fairness aspect of the network is therefore totally compromised. The PF scheduler is somewhere in the middle, since it takes into account both the instantaneous channel condition and the users' average channel statistics, and basically schedules the user whose channel is of the best quality

with respect to his own average channel conditions over a certain time window. The PF scheduler sacrifices throughput performance for fairness and represents a balance between the RR and Max-SNR scheduler.

In order for any type of optimality to be claimed, a certain criterion or cost function needs to be specified for the scheduler. When Shannon capacity is of concern, a scheduler picking the strongest user at every time instant for transmission has been shown to be optimal [16, 17]. The idea behind the Max-SNR scheduler is based on the simple observation that when the network users fade independently, at any one time instant, the probability that all users have poor reception is low. By allowing only the strongest user to transmit, the system is always riding the peak of the channel conditions across the network; the shared channel resource is used most efficiently and system throughput is maximized [18]. This gain in the sum network capacity due to the selection over multiple independently faded users is called *multiuser diversity gain*. The larger the number of users, the greater the multiuser diversity gain. Besides, another factor that contributes to the level of the gain is the tail of the user's channel fading distribution. The heavier it is, the greater the gain is, i.e. the Max-SNR scheduler will induce a higher gain in a Rayleigh faded environment than one with a Line-of-Sight (LOS) propagation path. Instead of combating and compensating for the fading in the point-to-point link, multiuser diversity is actually exploiting the fading in the network [18].

Unfortunately, this multiuser diversity gain does not come for free. The benefits of the gain require the Base Station (BTS) to have the ability to obtain measurements indicating the quality of the various channels. For cellular communication, if uplink communication is of concern, the BTS can measure the SNR for each user directly and schedule the strongest user. However for the downlink scenario, each user has to keep track of its own channel condition and feed it back through a separate frequency channel in a Frequency Division Duplexing (FDD) system, or channel reciprocity has to be used to distinguish various users' channel quality in a Time Division Duplexing (TDD) system. Therefore the accuracy of the channel estimation, the delay in feeding back the estimates, etc., are all factors that affect the network

throughput. In addition, when the channels of the users vary slowly over time, faster and larger fluctuations should be created on purpose to extract the largest possible multiuser diversity gain; e.g. [19] uses antenna arrays to induce bigger channel fluctuation.

Multiuser diversity has been studied extensively in the literature [20, 21, 22, 23], and most attempts at solving the problem usually consider only the system's Shannon capacity. Shannon capacity is a bound that limits the maximum error-free data transmission rate that can be achieved over an Additive White Gaussian Noise (AWGN) channel given a certain amount of channel bandwidth. With the development of channel coding techniques, Shannon capacity becomes more realistic, e.g. Turbo codes approach Shannon capacity within a few tenths of a dB [24].

However, the drawbacks of using Shannon capacity lie mainly in its idealized assumptions. In a real communication system, finite modulation schemes are used and the system's spectral efficiency will be capped by the highest modulation class, which means that the Shannon capacity will practically never be achievable in the high SNR regime. To accomplish or approach the Shannon capacity limit, usually code blocks with infinite length are assumed, however in a realistic communication system, the Forward Error Correction (FEC) module generates codewords with finite length, which violates the assumption of capacity-achieving schemes. Furthermore, current communication systems nowadays typically don't have the luxury of operating in an error-free mode. For example, systems employing Automatic Repeat reQuest (ARQ) or Hybrid ARQ, can operate at a relatively high block error rate.

Besides the impracticality of the criterion itself, there are still other issues that affect the analysis and design of a practical scheduler. As pointed out in the paragraphs above, channel quality measurements are the keys to scheduler implementation. First of all, when pilots are embedded in data for channel estimation, a portion of the channel bandwidth is wasted. The percentage of the pilots in each data frame and how often the pilots need to be transmitted both affect the system's capacity. Secondly, the transmit power of the pilots to a large extent determines the channel estimation accuracy, which then affects the decisions at the scheduler.

Thirdly, when users need to feedback the channel conditions (e.g. SNR values) to the BTS, this information needs to be quantized before being transmitted back using a reverse channel link, because the bandwidth of the control channel for feedback is usually very narrow. Consequently, the number of quantization bits will determine the accuracy of the knowledge obtained at the BTS side. The more bits the quantizer has, the easier it is for the BTS to distinguish two users' channel quality. If the quantization resolution is too rough, two users with different SNRs might not be distinguished and the scheduler may decide to schedule a non-optimal user, leading to capacity loss in the network. The degraded SNR report will also affect the modulation selection decision made at the BTS if adaptive modulation scheme is employed in the system. In addition to quantization issues, the delay that feedback induces is also another factor that needs to be accounted for when the scheduler is under inspection. The delay that information feedback takes will by itself lead to inaccurate channel knowledge at the BTS, because the wireless channel will change in the meantime. The smaller the mobility is, the greater the channel is correlated, and the better the fidelity of the channel knowledge. When a realistic performance analysis of a practical scheduler is of concern, all these aspects need to be considered jointly.

## 1.4 Doppler Effects on Communication System Performance

When the parties involved in the communications are mobile, the movement of the transceivers will result in a shift of the carrier frequency of the radio signal, referred to as the Doppler effect [25]. The presence of multiple propagation paths in the wireless channel means that there will in general be more than one frequency shift in the radio signal. When the number of paths increases, the transmitted unmodulated carrier (a single tone) will be spread out in the frequency domain, resulting in a phenomenon referred to as Doppler spread. The larger the motion of the user terminal, the wider the Doppler spread.

The most obvious and direct impact of the Doppler effect is its shifting the carrier frequency of the RF signals. Misalignment of the frequency of the demodulator reference signal and that of the received signal will leave the baseband signal with a

phase rotation, which in the frequency domain will lead to a shift of the received signal spectrum. If the shift is large enough and not well accounted for, it will prevent the receiver filter from capturing a proper amount of transmitted signal energy, causing an error floor in the bit error rate performance curve. For systems that employ Multi-Carrier Modulation (MCM) schemes like Orthogonal Frequency Division Multiplexing (OFDM), an uncompensated Doppler shift will cause the data streams on all the sub-stream to interfere with each other, i.e. referred to as Inter-carrier Interference (ICI), therefore reducing the effective SNR on individual sub-streams, deteriorating the reception quality.

In [26], the effects of mobility on communication system performance are studied. One of the basic influences of mobility is that it will impose a certain correlation pattern on the time-varying channel [3], which further determines if the successive symbols are experiencing similar fading conditions. For data service networks that are required to provide high Quality of Service (QoS) even for high mobility users, the system level design should take Doppler spread into account. A larger Doppler spread indicates a rapidly fading channel environment, and therefore requires more frequent channel estimation and tracking, which would cause more training pilots and lower the spectral efficiency. How long can a receiver performs its signal detection without estimating the channel again? How much capacity loss would be incurred when the channel is assumed to be constant when it is truly not? These questions can be better answered by more closely examining the Doppler effect in the communication problem setup. For example, [27] studies the coherent and non-coherent signal detection problem with partial channel knowledge in a channel with large Doppler spread.

As pointed out above, Doppler spread can cause a large performance degradation from the perspective of the physical layer. But what is the effect of Doppler spread on the scheduler from the MAC layer point of view? When the users in the network have various levels of mobility, the channel condition of each user can be very different. For schedulers that operate based on the instantaneous channel conditions for various users, the Doppler effect will outdate the channel knowledge so that when the user gets his own scheduled time slot, the optimality of the scheduler may not

hold any more. Therefore, when a fixed amount of time is assigned to the selected user without its receiver tracking the variation of the channel, picking the strongest user based on each user's individual instantaneous signal strength may not be the optimal approach. Instead, one should consider both the average SNR and Doppler spread in the MAC layer protocol.

## 1.5 Communications Over UAV-assisted Wireless Networks

Ad hoc networks are another type of wireless network, one that typically assumes no fixed topology and are defined as a collection of autonomous nodes or terminals that communicate with each other by forming a multihop radio network and maintaining connectivity in a decentralized manner [28]. Similar to cellular based networks, users in wireless ad hoc networks also need to combat the adverse nature of the communication channels and compete for limited resources. However, in ad hoc networks, each node acts both as a host and a router, and therefore routing algorithms are critical to reduce the overhead and multihop delay [29]. Two types of wireless ad hoc networks that are of greatest interest include: Mobile Ad-hoc Networks (MANETs) and Smart Sensor Networks (SSNs).

According to [30], a MANET is an autonomous system of mobile routers (and associated hosts) connected by wireless links, the union of which form an arbitrary graph. They are often used for establishing communications in emergency scenarios. Mobility is one of the most important characteristics of a MANET. However, the time-varying nature of the MANET topology requires the development of efficient distributed algorithms to determine the network organization, link scheduling, and routing algorithm, etc. Various constraints need to be taken into account when designing a MANET, e.g. power constraints for battery-limited applications, minimum probability of interception constraints for military applications, maximum hop constraints for delay sensitive applications like voice communications and so on.

Compared with MANETs, the nodes in a wireless SSN are not necessarily fully autonomous, but have basic communication capability and a certain level of intelligence, i.e. signal processing and data networking. Examples of wireless SSNs

can be found in both civilian and military applications, including traffic sensor networks, surveillance sensor networks, military sensor networks, etc. Wireless SSNs can be classified according to two different criteria [28]: 1) whether each individual node in the network is addressable, and 2) whether the data in the network is aggregated. The common tasks of the wireless SSN are to 1) determine the value of some parameter at a given location, 2) detect the occurrence of events of interest and estimate parameters of the detected event or events, 3) classify a detected object, and 4) track an object. These tasks lead to a few shared requirements on the wireless SSN, including: 1) a large number of sensors, 2) low energy consumption, 3) network self-organization, 4) collaborative signal processing, and 5) querying ability [28].

Unmanned Aerial Vehicles (UAVs), including battlefield UAVs, miniature UAVs, endurance UAVs, etc., have been designed to fulfill the tasks of providing military communication links, traffic surveillance, border and coast patrol, fire perimeter monitoring, search and rescue, and so on. UAVs are particularly well-suited for situations that are too dangerous for direct human monitoring. In environments that are hazardous to human operators, unmanned vehicles can be deployed to carry out tasks that would otherwise be impossible to accomplish. Recent improvements in battery, micro-controller, and sensor technologies have resulted in the development of autonomous vehicles that are inexpensive, dependable, and simple to operate. Research on exploiting the mobility and cooperation between a team of UAVs to obtain better tracking of a moving target can be found in [31, 32, 33]. Applications involving UAVs in telecommunications can also be found in the literature [34, 35].

The importance of being able to communicate in the battlefield is obvious, but the hostile environment and the need for constant mobility and reliability makes it hard to install fixed communication facilities. A MANET's independence from fixed communication infrastructure, its instant deployment and easy reconfigurability make it very suitable for battlefield communication network applications. UAVs are good candidates for relays in MANETs. Their altitude enables them to be seen by all members of the network, and their size makes them resistant to attacks.

Without using any existing network infrastructure and any centralized administration, the great benefits that MANETs promise are all based on robust and smart routing algorithms. In [36, 37, 38], various multihop wireless ad hoc routing protocols are discussed and compared. However, for applications involving a large number of nodes, a pure flat ad hoc structure, where no hierarchical topology is assumed, faces various problems such as long hop paths, heavy routing overhead, spatial concurrency constraints of neighbor nodes, etc [34]. In a word, a flat structure encounters scalability problems when the number of nodes in the ad hoc network increases, especially in the face of node mobility. A hierarchical architecture is essential for achieving a basic performance guarantee in a large scale MANET [35]. A diagram for a hierarchical UAV-aided network is shown in Fig. 1.1, where the colored nodes represent the users at the lower level, and UAVs in the air act as backbones, providing higher level connections. Any kind of topology can be assumed at the lower level. Usually due to the probable LOS propagation path and larger payloads, the links between UAVs can be assumed to be much stronger than the ones between lower level nodes.

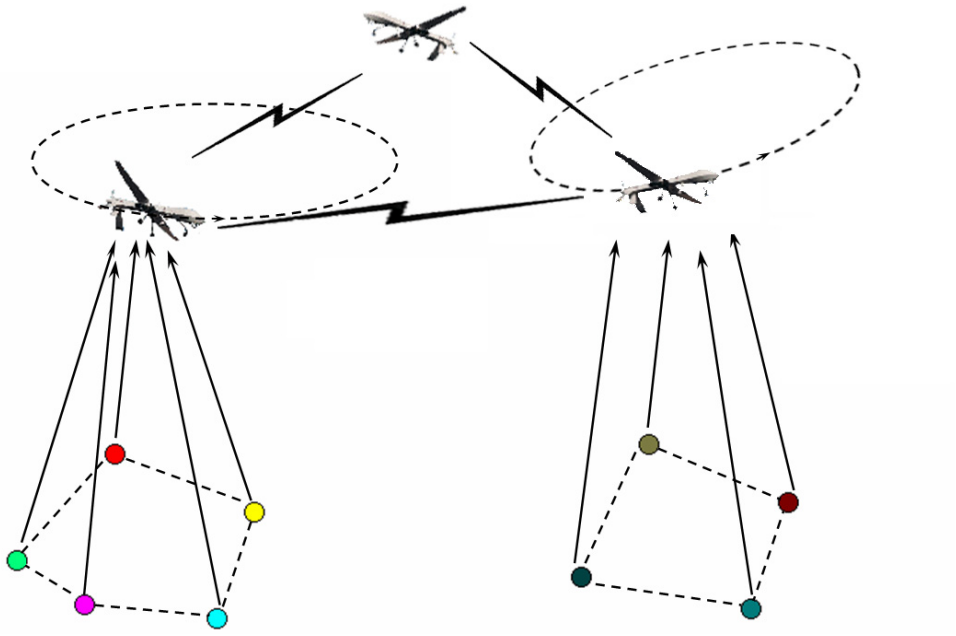


**Figure 1.1:** Hierarchical UAV-aided networks.

Applications using UAVs to help set up ad hoc network communications can be widely found in the literature. An architecture for implementing a rapidly deployable wireless network using UAVs is discussed in [39]. In [34, 40], a hierarchical UAV-assisted network is assumed to be deployed in the battlefield, and routing algorithms are studied to provide communications with higher throughput and to support the scalability of the network. The movement pattern of the UAVs can impact the networking performance of the system; therefore, in [41], algorithms for determining desirable mobility properties for UAVs was presented. In [42], artificial potential fields were used to guide the motion of UAVs that are used to bridge the connection between Autonomous Underwater Vehicles (AUVs) and Tactical Operation Centers. For delay-tolerant applications, e.g. bulk data transfer, [43] proposed a scheme to maximize network throughput by using UAVs to load data from the ground nodes, carry the data while flying to the destination, and deliver the data to the destination nodes. A Mobile Backbone Network Protocol (MBNP) is proposed in [44] that discusses network topological synthesis, proactive routing mechanisms, MAC layer power control, etc., in UAV-assisted hierarchical networks. Fuzzy control of the motion of the UAVs to improve the connectivity of mobile ground nodes is approached by mimicking the flocking rules of birds [45]. A feasibility study of using OFDM signaling technology in UAV wireless communications was conducted in [46]. Performance of the ad hoc ground network using swarm UAVs is studied from a physical layer perspective in [47], where an air-to-ground wireless channel is used to model the link condition, and the effects of the UAVs' positions, velocities, etc, on Bit Error Rate (BER) performances of different transmission strategies are analyzed.

In a word, networking using mobile backbones with a hierarchical structure has attracted significant interests. Many aspects of such networks need to be explored, including: routing algorithms, cross layer design, power control strategies, etc. An extra degree of freedom which can be manipulated to obtain better control and achieve higher performance is the motion of the UAVs. By optimally commanding the heading or designing the flying pattern of the UAVs, network throughput can be improved due to the reduction in the propagation distance of the RF signals or elimination

of shadowing, etc. Another issue that is key to UAV networks is the possibility of high Doppler shifts. The determination of time slot duration for scheduling needs to account for the Doppler effect and appropriate channel estimation is necessary for the network to provide acceptable data transmission rates.

## 1.6 Focus of the Research Topics in the Dissertation

This dissertation focuses on exploiting possible ways to optimize the performance of two kinds of networks, i.e. cellular networks for civilian applications and UAV-assisted hierarchical networks for military use. For cellular networks, schedulers which have instantaneous channel information for all users are studied. Practicalities like pilot insertion, channel estimation error, feedback delay, quantization error, etc, are considered when evaluating the spectral efficiency of a traditional Single Input Single Output (SISO) network with FEC blocks built in. Closed-form expressions are derived to quantify the throughput of the network in terms of spectral efficiency. As discussed above, when mobile relays are used in the network, the Doppler effect is not a trivial aspect that can be ignored for a fixed network alternative any more. The duration of allocated time slots affects the system throughput when channel estimation cannot be assumed to be perfect. The limited training capability of the system outdates the channel knowledge at the scheduler and leads to faulty decisions. The longer the slot allocation, the larger the loss it may induce. When the slot duration is relatively long and the training for the scheduled user is limited, opportunistic scheduling which exploits instantaneous channel knowledge of the users may not be optimal. The users' average link SNR combined with their respective Doppler spreads can be a good metric for scheduling algorithms, and therefore a Statistical Channel-aware (SCA) scheduler is proposed in the dissertation. In addition, the performance of a Max-SNR scheduler for an asymmetric network, where users have different average SNR, is studied and analytical results are obtained. For UAV-aided networks, Multiple Input Multiple Output (MIMO) antenna systems are assumed, and the optimal heading control problem for the mobile UAVs is of primary interest. A closed-form solution for commanding a team of UAVs to fly in a trajectory that

maximizes the data rate of a network with mobile nodes is found in the dissertation. Handoff of the mobile nodes in this specific mobile relay network setup is different from simply checking the receive signal strength and assigning the mobile user to the UAV offering best link quality. Constraints on the UAVs' mobility and minimum communication rate for each user create more complexity for this handoff problem. Extra UAVs need to be deployed in the network when the users cannot be accommodated by the current ones. The optimal UAV deployment problem is also part of the focus of the research. A complete network protocol is proposed to achieve the maximum transmission throughput.

The rest of the dissertation is organized as follows. Chapter 2 summarizes the various channel models for wireless communications to give a general view of the characteristics of the wireless propagation environment and the difficulties of communicating without wires. Chapter 3 reviews the state-of-the-art signaling strategies over the wireless link to exploit the multipath nature of the channels and maintain reliable and high-speed communications. In Chapter 4, a realistic performance analysis for a Max-SNR scheduler is conducted, including the effects of channel coding, channel estimation, Doppler shift, feedback delay etc. Chapter 5 is still concerned with a Max-SNR scheduler's performance in an asymmetric network consisting of users with high mobility. The large Doppler spread of the channel quickly outdates the channel knowledge. The effect of Doppler spread on the capacity of the system is studied, and a statistical scheduler which assigns users based solely on the average SNR and Doppler spread of the user is proposed. A hierarchical UAV-aided communication network is then examined in Chapter 6. Link level analysis including the average transmission rate and the probability of Symbol Error Rate (SER) is given in the chapter. Optimal heading controls for the UAV relays that maximize the throughput of the network are obtained. Cross layer issues, such as relay deployment and AP handoff problems are also proposed for a complete networking protocol for this mobile relay assisted network. Chapter 7 concludes the dissertation and proposes possible research topics for future work.

# Chapter 2

# Wireless Communication Channel

The most obvious difference between wire-line and wireless communication applications is their signal propagation environments. With more complicated channels that have a randomly time-varying impulse response, the transmitters and receivers in a wireless communication system need to combat more hostile transmission and propagation effects. The generally hostile nature of the wireless channel has roots in the mobility of both the propagation medium (time varying scatterers), multiple propagation paths (small scale fading), shadowing (medium scale fading), path loss (large scale fading), co-channel interference etc. The fading characteristics of the channel cause the channel realization to vary dramatically at different locations, on the order of a few wavelengths. Multiple copies of the transmitted signal arriving at the receiver at different symbol periods incur ISI; in other words, in the frequency domain, it deforms the transmitted signal spectrum and induces a frequency selective channel. All of the above issues make communication over the wireless channel less reliable. The other bottleneck in wireless communication applications is its limited channel capacity. To increase the robustness of the communication link, different approaches to combat fading (thus to reduce the variation of the received signal strength), to cancel self-symbol interference (thus to flatten channel responses) have been widely studied in the literature. On the other hand, to increase data rate per channel use, multiple antenna arrays have been proposed to create parallel data pipes for data multiplexing. In addition, spread spectrum technology has been developed to enable aggressive frequency reuse when cell planning is of concern so that higher network capacity can be achieved. This chapter will briefly discuss the characteristics of the

propagation environment for RF signals and introduce some wireless channel models for different system configurations.

## 2.1 Wireless Propagation Environment

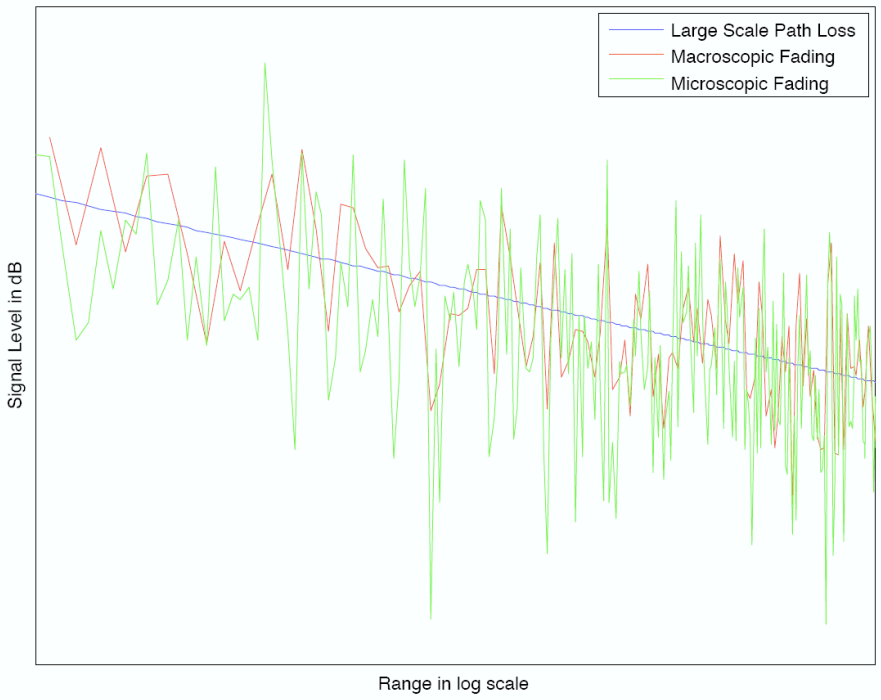



**Figure 2.1:** Signal strength as a function of T-R separation.

A Radio Frequency (RF) signal usually travels through a number of different paths before it reaches the receiver. This is often referred to as multipath propagation. The propagating mechanism can be generally classified as reflection, diffraction and scattering [2, 7]. The most important factor that determines the quality of the communication link is the received signal power. Concerning the level of the received signal power, three mutually independent, multiplicative propagation phenomena can usually be distinguished: multipath fading (microscopic/small-scale fading), shadowing (macroscopic fading) and large-scale path loss (mean propagation loss) [7, 48]. As can be seen in Fig. 2.1, as the Transmitter-Receiver (T-R) separation increases, the

signal power decreases, and the area-mean power decreases smoothly. Fading, especially small scale fading causes the received signal to fluctuate greatly over a small period of time or over short distances, and therefore leads to unreliable communications. The rest of the section will be dedicated to describing these three propagation characteristics.

### 2.1.1 Large Scale Path Loss

The large scale path loss is mainly the result of the inverse square law of the power loss in free space. Its effect on the received signal's attenuation is smooth with respect to the geometry of the path profile in the propagating environment. The large scale path loss is usually measured by averaging the power level over a relatively large area (tens or hundreds of meters) to quantify the so-called area-mean power level [48]. To estimate the radio coverage, mathematical models to predict the mean signal strength for arbitrary Transmitter-Receiver (T-R) separation distances are widely studied in the literature. Different large scale propagation models are developed for different propagation environments and mechanisms.

When there is a clear unobstructed LOS path between the transmitter and receiver, the received signal strength can be predicted using free space propagation models. This is typically the case for satellite communication systems, air-to-ground and microwave LOS links, to name a few. For a transmitter with transmit power $P_t$, the signal power at a receiver that is $d$ meters away, can be quantified using Friis free space equation:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L}\,, \tag{2.1}$$

where $G_t$ and $G_r$ are the antenna gain at the transmit and receive side respectively, $\lambda$ is the carrier's wavelength, and $L$ is miscellaneous loss due to other factors in the system, including transmission line attenuation, filter losses and so on [2]. Equation (2.1) is only valid when the receiver is in the far field (Fraunhofer region) of the transmitter. In other words, (2.1) holds only when the separation distance satisfies $d >> d_f$, where $d_f = \frac{2D^2}{\lambda}$ is the far field distance defined as a function of the

transmit antenna's linear dimension ($D$) and the RF carrier wavelength ($\lambda$). Hence, in a mobile radio system, a reference distance $d_0 > d_f$ and less than any practical distance is picked and the average power level is measured such that the received signal power at any point of practical interest in the system can be expressed as:

$$P_r(d) = P_r(d_0) \cdot \left(\frac{d_0}{d}\right)^2 , \tag{2.2}$$

where $d \geq d_0 \geq d_f$.

For those paths that involve more than one propagation medium, reflection, diffraction and scattering phenomena will impact system performance. When an electromagnetic (EM) wave impinges upon an object with a large dimension when compared to the wavelength of the EM wave, reflection occurs. If the wave is incident on a perfect dielectric, part of the energy is transmitted into the second medium and part of the energy is reflected back into the first medium. There is no absorption of the energy in the propagation. If the second medium is a perfect conductor, all the power of the EM wave is reflected back into the original medium. One of the many interesting models that take into account the reflection mechanism is the ground reflection (two-ray) model replicated below [49]. This model predicts the average signal strength when the environment contains two major propagation paths, i.e. a LOS path and a ground reflected path. The received power is modeled as:

$$P_r(d) = P_t G_t G_r \frac{h_t^2 h_r^2}{d^4} , \tag{2.3}$$

where $d >> \sqrt{h_t h_r}$, $P_t$ is the transmit power, $G_t$, $G_r$ is the antenna gain, $h_t$, $h_r$ are the transmit antenna height and receive antenna height respectively, and $d$ is the separation distance between transmitter and receiver. This model works fairly well for predicting signal strength over distances of several kilometers for mobile radio systems that use tall antennas ($h_t > 50$m), as well as for LOS microcell channels in urban environments [2, 49]. When EM waves hit an object with sharp irregularities (edges), diffraction will take place. Knife-edge models that are used to characterize diffraction process can be found in [50, 51, 52, 53]. If the EM wave travels through

medium containing dense obstacles that have smaller dimension when compared to the carrier's wavelength, scattering occurs. When a radio channel consists of large, distant objects that induce scattering, e.g. the urban mobile radio environment, the bistatic radar equation (2.4) can be used to model received signal strength:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2 \sigma_b}{(4\pi)^3 d_t^2 d_r^2}\,, \tag{2.4}$$

where $\sigma_b$ is the radar cross section (RCS), defined as the ratio of the power density of the signal scattered in the direction of the receiver to the power density of the radio wave incident upon the scattering object [2, 25, 54], and $d_t$, $d_r$ are the distances from the scatters to the transmitter and receiver respectively. Equation (2.4) is only valid when the scatterers are in the far field of both the transmitter and receiver. Several RCS values are determined in [55].

### 2.1.2 Shadowing Effects

Shadowing effects of buildings or natural features introduces additional fluctuations so that the received local-mean power varies around the area-mean. Considering the different nature and characteristics of the scatterers, the receiver will perceive vastly different power levels at different locations with the same T-R separation. This effect causes inaccuracies in the previous area-mean prediction models. The fluctuation of the signal strength in this case is called macroscopic fading as compared to the case discussed in the next section (microscopic fading). The received signal power is averaged over a few tens of wavelengths, typically $40\lambda$ so that the fluctuations of the instantaneous received power due to multipath effects are largely removed [48]. According to [3, 56, 57], the macroscopic fading approaches a Gaussian distribution when measured in log-scale, and therefore is referred to as log-normal shadowing:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,, \tag{2.5}$$

where $x$ (in dB) is the signal strength averaged over microscopic fading, $\mu$ is the area-mean power (in dB) predicted by the different models discussed above, and $\sigma$ is the standard deviation of $x$ (in dB).

### 2.1.3 Small-scale Multipath Propagation

Experiments conducted in the field reveal that when the receiver moves over a small distance (a few wavelengths), the instantaneous signal fluctuates drastically due to the superposition of a large number of independent scattered components. This is called small scale (microscopic) fading. Multiple copies of transmitted signals with different phases and amplitudes arriving at receivers at slightly different times combine together to yield a rapidly fluctuating signal that degrades the quality of the communication link. Multiple propagation paths are the cause of the small scale fading, and the three most important effects are [2]:

- Rapid fluctuation in the signal strength over a short distance or time interval
- Random frequency modulation due to different Doppler shifts on various propagation paths
- Time dispersion (echoes) caused by multipath propagation delays

Small-scale fading characterizes the case when the large-scale path loss can be ignored. Many physical factors can affect the small-scale fading, and the most important few include multiple propagation paths, relative motion between the transmitter and receiver, motion of the scatterers in the environment, transmitted signal bandwidth, etc. In the typical mobile communication setup, due to the relatively lower height of the mobile receiver, there is usually no strong LOS propagation path. In this scenario, when the number of independent EM waves is assumed to be large, the distribution of the received signal can be considered as a complex Gaussian process in both its in-phase and quadrature component with no cross correlation. The amplitude of the received signal is Rayleigh distributed as follows:

$$f(x) = \frac{2x}{\bar{\gamma}} e^{-\frac{x^2}{\bar{\gamma}}} , \tag{2.6}$$

where $x \geq 0$ and $\bar{\gamma}$ is the averaged received signal power. With a random variable transformation, the distribution of the signal power can be found to be exponentially distributed:

$$f(x) = \frac{1}{\bar{\gamma}} e^{-\frac{x}{\bar{\gamma}}}, \tag{2.7}$$

where $x \geq 0$. However, when the surrounding scatterers around the mobile receiver are smaller or fewer such that there is a dominant clear direct propagation path for the EM waves (LOS path), a mean component will be present in the received signal. In this case, the magnitude of the received signal is a random process with a Rician distribution:

$$f(x) = \frac{2x(K+1)}{\bar{\gamma}} e^{-K-\frac{(K+1)x^2}{\bar{\gamma}}} \mathcal{I}_0\left(2x\sqrt{\frac{K(K+1)}{\bar{\gamma}}}\right), \tag{2.8}$$

where $x \geq 0$ and $K$ is the Rician factor defined as the ratio of the dominant signal power ($s^2$) over the scattered signal power ($2\sigma^2$) in $K = \frac{s^2}{2\sigma^2}$, with the total power given by $\bar{\gamma} = s^2 + 2\sigma^2$. $\mathcal{I}_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{-xcos\theta} d\theta$ is the zero-th order modified Bessel function of the first kind. Using some transform techniques [58], the signal power can be shown to be distributed as non-central chi-square:

$$f(x) = \frac{(K+1)e^{-K}}{\bar{\gamma}} e^{-\frac{(1+K)x}{\bar{\gamma}}} \mathcal{I}_0\left(\sqrt{\frac{4K(K+1)}{\bar{\gamma}}x}\right), \tag{2.9}$$

where $x \geq 0$. Using $\mathcal{I}_0(0) = 1$, we see that (2.6) is a special case of (2.8) with Rician factor $K = 0$. Another popular distribution used to characterize the statistics of the fading signal is the Nakagami distribution [59]:

$$f(x) = \frac{2}{\Gamma(m)} \left(\frac{m}{\bar{\gamma}}\right)^m x^{2m-1} e^{-\frac{mx^2}{\bar{\gamma}}}, \tag{2.10}$$

where the parameter $m$ is defined as the ratio of moments, called the fading figure:

$$m = \frac{\bar{\gamma}^2}{\mathcal{E}\left[x^2 - \bar{\gamma}\right]^2}, \tag{2.11}$$

where $m \geq \frac{1}{2}$ and $\mathcal{E}(\cdot)$ is the expectation operator. As can be easily seen, Rayleigh fading is a special case of Nakagami fading when $m = 1$.

## 2.2 Wireless Channel Modeling

### 2.2.1 Single Input Single Output (SISO) Channel

For cell planning, large scale propagation effects dominate the system design. However, when a specific link is under inspection, large scale path loss can typically be ignored, and only the small scale fading has to be taken into account for channel modeling. In this section, wireless channel models are discussed for systems that have only one antenna at both the transmitter and receiver. Channel here refers to the composite effect of the transmit pulse shaping filter, wireless propagation environment and the matched filter at the receiver. As explained above, due to multiple propagation paths, relative motion of scatterers in the environment, different angular distributions of the incoming waves, the wireless channel can be modeled as a function of three individual parameters, including time ($t$), delay ($\tau$) and spatial dimensions ($r$), and we denote it to be $h(t, \tau, r)$. Under the assumption of linearity, the received signal in the time domain can be expressed as:

$$y(t, r) = \int_{-\infty}^{\infty} h(t, \tau, r) x(t - \tau) d\tau \,. \tag{2.12}$$

Equivalently it can be expressed using the Fourier transform as follows [60, 61]:

$$y(t, r) = \frac{1}{2\pi} \int_{-\infty}^{\infty} h(t, f, r) X(f) e^{2\pi f t} df \,. \tag{2.13}$$

The variation of the channel over time ($t$), frequency ($f$), and space ($r$) leads to the selectivity concept of the channel. By introducing their corresponding Fourier transform pairs: Doppler domain ($\omega$), delay domain ($\tau$), and wave number ($k$) domain, the coherence concept of the channel can be defined as the "window" over which the channel's response does not vary significantly. For example, the coherence time can be defined as the time window over which a narrow band channel (no frequency or

spatial dependence) remains almost constant [61]. In other words, when $|t - t_0| \leq \frac{T_c}{2}$, the following equation is satisfied:

$$|h(t)| \approx V_0\,, \tag{2.14}$$

where $V_0$ is some constant voltage, $t_0$ is some arbitrary moment in time, and $T_c$ is the size of the time window of interest. The largest value of $T_c$ on average defines the coherence time of the channel. If the transmit symbol rate of the system is comparable to the coherence time $T_c$, it will be difficult for the receiver to recover the information. A much faster transmit rate ($T_s << T_c$) yields a slow fading channel, otherwise when a slower transmit rate is applied to average out the variation of the channel ($T_s >> T_c$), the channel is said to be fast fading. Similarly, when $|f - f_0| \leq \frac{B_c}{2}$, we have:

$$|h(f)| \approx V_0\,, \tag{2.15}$$

and the largest value of $B_c$ is defined to be the coherence bandwidth of the channel. When the bandwidth of the transmitted signal $B < B_c$, the signal is considered narrowband, and no severe distortion occurs in the time domain; otherwise the signal is said to be wide-band. When a channel has a coherence bandwidth that is larger than the signal bandwidth, it is said to be a frequency flat fading channel, otherwise it is referred to as a frequency selective fading channel. The coherence distance $D_c$ is defined as the largest one-dimensional displacement over which the channel is approximately constant:

$$|h(r)| \approx V_0\,. \tag{2.16}$$

when $|r - r_0| \leq \frac{D_c}{2}$ holds.

The above definitions assume that the channel is deterministic. A stochastic channel model serves to better describe the random nature of the wireless propagation environment. A detailed treatment of random time-variant linear channels can be found in [61, 62, 63]. A statistical perspective of the channel is to view it as a random process in terms of time, frequency and space. Correlation is usually used to describe

a random process when only one dependency of the channel is of interest, e.g. for a time dependent channel $h(t)$, its auto-correlation is defined as:

$$R_h(t_1, t_2) = \mathcal{E}\left[h(t_1) \cdot h^*(t_2)\right] , \tag{2.17}$$

where as mentioned above, $\mathcal{E}(\cdot)$ is the expectation operator. A random process is strictly stationary if the joint distribution $p(h_{t_1}, h_{t_2}, \cdots, h_{t_k})$ is identical to that of $p(h_{t_1+t}, h_{t_2+t}, \cdots, h_{t_k+t})$ for all $t$, and an arbitrary positive integer $k$. In other words, strict stationarity requires an arbitrary order of joint distribution invariance under time shifts [64]. This is a very difficult condition to verify for a practical process. A wide-sense stationary (WSS) process is a random process with weaker stationarity conditions. A process is WSS when its mean is time invariant (2.18):

$$\mathcal{E}(h(t)) = h\,, \tag{2.18}$$

and its covariance is time invariant (2.19):

$$R_h(t_1, t_2) = R_h(t_1 + t, t_2 + t)\,. \tag{2.19}$$

To fully characterize a WSS channel $h(t, f, r)$, its joint correlation should be defined:

$$R_h(t_1, f_1, r_1; t_2, f_2, r_2) = \mathcal{E}\left[h(t_1, f_1, r_1) \cdot h^*(t_2, f_2, r_2)\right] . \tag{2.20}$$

For the purpose of mathematical tractability, we assume the random channel is a jointly WSS process in $(t, f, r)$. According to the property of WSS processes, the WSS assumption leads to uncorrelated spectral components, i.e. the channel process is uncorrelated in the spectral domain $(\omega, \tau, k)$. Mathematically, this wide-sense stationary, uncorrelated scattering (WSS-US) assumption can be written as:

$$\begin{aligned} S_h(\omega_1, \tau_1, k_1; \omega_2, \tau_2, k_2) &= \mathcal{E}\left[h(\omega_1, \tau_1, k_1) \cdot h^*(\omega_2, \tau_2, k_2)\right] \\ &= S_h(\omega_1, \tau_1, k_1) \cdot \delta(\omega_1 - \omega_2) \cdot \delta(\tau_1 - \tau_2) \cdot \delta(k_1 - k_2)\,, \end{aligned} \tag{2.21}$$

where $S_h(\cdot,\cdot,\cdot)$ is the power spectral density of the channel process $h(t,f,r)$. According to the Wiener-Khinchine theorem, $R_h(\Delta t, \Delta f, \Delta r)$ and $S_h(\omega,\tau,k)$ are a Fourier transform pair. The relationships between the autocorrelation function and power spectral density (PSD) of the channel can be found in various textbooks [59, 61, 65]. The channel's dependency on multiple different parameters complicates the analysis; by integrating over the other two parameters in the spectral domain, the channel's Doppler spectrum, delay spectrum and wave-number spectrum can be defined respectively as follows:

$$S_h(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_h(\omega,\tau,k)d\tau dk\,, \tag{2.22}$$

$$S_h(\tau) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_h(\omega,\tau,k)d\omega dk\,, \tag{2.23}$$

and

$$S_h(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S_h(\omega,\tau,k)d\tau d\omega\,. \tag{2.24}$$

By setting the other two parameters to be zero in the original domain, a channel's time/frequency/space correlation can be defined as:

$$R_h(\Delta t) = R_h(\Delta t, 0, 0)\,, \tag{2.25}$$

$$R_h(\Delta f) = R_h(0, \Delta f, 0)\,, \tag{2.26}$$

and

$$R_h(\Delta r) = R_h(0, 0, \Delta r)\,. \tag{2.27}$$

$S_h(\omega)$ and $R_h(\Delta t)$, $S_h(\tau)$ and $R_h(\Delta f)$, $S_h(k)$ and $R_h(\Delta r)$ form Fourier transform pairs respectively.

With the stochastic channel model, the coherence definition introduced above can be quantitatively measured in terms of the spread in the spectrum domain. The delay/Doppler/wavenumber spread ($\sigma_\tau/\sigma_\omega/\sigma_k$) is defined as the root-mean-square (RMS) width of the corresponding spectrum. Mathematically, the definition men-

tioned above can be written as (2.28)-(2.30):

$$\sigma_\tau^2 = \overline{\tau^2} - (\bar{\tau})^2 \,, \tag{2.28}$$

where $\overline{\tau^n} = \frac{\int_{-\infty}^{\infty} \tau^n S_h(\tau) d\tau}{\int_{-\infty}^{\infty} S_h(\tau) d\tau}$,

$$\sigma_\omega^2 = \overline{\omega^2} - (\bar{\omega})^2 \,, \tag{2.29}$$

where $\overline{\omega^n} = \frac{\int_{-\infty}^{\infty} \omega^n S_h(\omega) d\omega}{\int_{-\infty}^{\infty} S_h(\omega) d\omega}$, and

$$\sigma_k^2 = \overline{k^2} - (\bar{k})^2 \,, \tag{2.30}$$

where $\overline{k^n} = \frac{\int_{-\infty}^{\infty} k^n S_h(k) dk}{\int_{-\infty}^{\infty} S_h(k) dk}$. As is well known in Fourier analysis, when a signal has a wide spread in one domain, its corresponding transform pair has a narrow spread in the other domain. Therefore, as a rule-of-thumb, the spreads defined above have a roughly reciprocal relationship with their corresponding coherence counterparts. If coherence bandwidth is defined as the frequency interval over which the channel's complex frequency transfer function has a correlation of at least 0.9, the coherence bandwidth is approximately $B_c \approx \frac{1}{50\sigma_\tau}$ [66, 67]. A more popular approximation of $B_c$ corresponding to a bandwidth interval having a correlation of at least 0.5 is $B_c \approx \frac{1}{5\sigma_\tau}$ [2, 59, 61, 67]. When coherence time $T_c$ is defined as the time duration over which the channel's response to a sinusoid has a correlation greater than 0.5, the relationship between $T_c$ and $\sigma_\omega$ is approximately $T_c \approx \frac{9}{16\pi\sigma_\omega}$ [68]. The relationship between the wavenumber spread and coherence distance can be found in [61].

Another important metric to characterize the variation of the random channel is fading rate variance defined as:

$$\sigma_t^2 = \mathcal{E}\left\{ \left| \frac{d\left[h(t) \cdot e^{-j\bar{\omega}t}\right]}{dt} \right|^2 \right\} ,$$

$$\sigma_f^2 = \mathcal{E}\left\{ \left| \frac{d\left[h(f) \cdot e^{-j2\pi\bar{\tau}f}\right]}{df} \right|^2 \right\} ,$$

and

$$\sigma_r^2 = \mathcal{E}\left\{\left|\frac{d\left[h(r)\cdot e^{-j\bar{k}r}\right]}{dr}\right|^2\right\}, \tag{2.31}$$

where $\bar{\omega}$, $\bar{\tau}$, $\bar{k}$ are the centroids of the corresponding spectra, and can be found by evaluating the function $f(x) = \frac{\int_{-\infty}^{\infty} x\cdot S_h(x)\cdot dx}{\int_{-\infty}^{\infty} S_h(x)\cdot dx}$. Exploiting the basic theorem in stochastic theory that relates the mean-square derivatives of complex processes to PSDs [61, 69],

$$\mathcal{E}\left\{\left|\frac{d^n h(t)}{dt^n}\right|^2\right\} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\omega^{2n}S_h(\omega)d\omega\,, \tag{2.32}$$

we can easily show (2.33)-(2.35):

$$\sigma_t^2 = \mathcal{E}\{P(t)\}\cdot\sigma_\omega^2\,, \tag{2.33}$$

$$\sigma_f^2 = (2\pi)^2\cdot\mathcal{E}\{P(f)\}\cdot\sigma_\tau^2\,, \tag{2.34}$$

and

$$\sigma_r^2 = \mathcal{E}\{P(r)\}\cdot\sigma_k^2\,, \tag{2.35}$$

where $P(\cdot)\equiv|h(\cdot)|^2$ is the power of the channel, and $\sigma_\omega$, $\sigma_\tau$, $\sigma_k$ are Doppler spread, delay spread and wavenumber spread respectively.

When the coherence bandwidth $B_c$ is larger than the signal bandwidth, the signal does not get severely distorted in the time domain, and the channel can be modeled using a linear random scalar response and the received signal can be written as (2.36):

$$y(t,r) = h(t,r)\cdot x(t) + z(t)\,, \tag{2.36}$$

where $\mathcal{E}(x(t)^2) = E_s$ is the power of the transmitted signal, and $z(t)\sim CN(0,\sigma^2)$ is zero mean circularly symmetric complex Gaussian noise. Depending on whether there is a dominant signal component or not, the received signal envelope is distributed as Rayleigh (2.6) or Rician (2.8).

However, when the signal bandwidth is larger than the coherence bandwidth $B_c$, which in other words means the delay spread of the channel ($\sigma_\tau$) spans multi-

ple transmit symbol durations $T_s$, the received signal will be composed of multiple delayed copies of the transmitted signal, and thus will possess ISI. This fading is called frequency selective fading, and the received signal in the time domain can be greatly distorted. Any low-pass signal $x(t)$ with two-sided bandwidth $W$ and time duration $t \in [0,\ T]$ may be represented by a $N = WT$ dimensional vector $\mathbf{x} = [x(0),\ x(1/W), \cdots,\ x(T - 1/W)]^T$ in the following equation [60, 70]:

$$x(t) = \sum_{n=0}^{WT-1} \mathbf{x}[n] \cdot \frac{\sin W \cdot (t - n/W)}{W \cdot (t - n/W)}, \tag{2.37}$$

where $\mathbf{x}[n]$ denotes the $n$-th component of the vector $\mathbf{x}$. The Fourier transform of $x(t)$ can be easily derived as:

$$X(f) = \begin{cases} \frac{1}{W} \sum_{n=0}^{WT-1} x[n] \cdot e^{-j\frac{2\pi f n}{W}} & |f| \leq \frac{W}{2} \\ 0 & |f| > \frac{W}{2} \end{cases}. \tag{2.38}$$

If we recall (2.13), and plug in (2.38), we have the noise corrupted received signal:

$$\begin{aligned} y(t,r) &= \int_{-\infty}^{\infty} H(t;f,r) \cdot X(f) \cdot e^{j2\pi f t} df + z(t) \\ &= \frac{1}{W} \sum_n x[n] \int_{-\infty}^{\infty} H(t;f,r) \cdot e^{j2\pi f[t - \frac{n}{W}]} df + z(t) \\ &= \frac{1}{W} \sum_n x[\frac{n}{W}] h(t; t - \frac{n}{W}, r) + z(t) \\ &= \sum_{n=0}^{L-1} x[t - \frac{n}{W}] h_n(t,r) + z(t), \end{aligned} \tag{2.39}$$

where $h_n(t,r) = \frac{1}{W} h(t; n/W, r)$ and $L = [W/B_c]$ is the number of resolvable multi-paths as seen by the transmitted signal. For each path, $h_n(t)$ is a complex Gaussian process as described in the narrowband case, and for different propagation paths $(m \neq n)$, $h_m(t)$ is uncorrelated with $h_n(t)$ due to the uncorrelated scattering assumption.

### 2.2.2 Multiple Input Multiple Output (MIMO) Channel

The extension of the SISO channel model to the Multiple Input Multiple Output (MIMO) channel to address the case when both the transmitter and receiver are equipped with multiple antennas is straightforward. The rest of the section will mostly follow the notation and the discussion from [7]. By defining the impulse response for each antenna pair (between the $j$-th transmit antenna and the $i$-th receive antenna) as $h_{i,j}(t,\tau)$, we can describe the whole system using the channel matrix $\mathbf{H}(t,\tau)$:

$$\mathbf{H}(t,\tau) = \begin{bmatrix} h_{1,1}(t,\tau) & h_{1,2}(t,\tau) & \cdots & h_{1,M_T}(t,\tau) \\ h_{2,1}(t,\tau) & h_{2,2}(t,\tau) & \cdots & h_{2,M_T}(t,\tau) \\ \vdots & \vdots & \ddots & \vdots \\ h_{M_R,1}(t,\tau) & h_{M_R,2}(t,\tau) & \cdots & h_{M_R,M_T}(t,\tau) \end{bmatrix}, \tag{2.40}$$

where $M_T$ and $M_R$ are the number of transmit and receive antennas respectively.

The relationship between the transmit and received signal without noise in (2.12) still holds in the MIMO case, with the only difference being that the transmit and receive signals are vectors instead of scalars. Therefore, modeling the MIMO channel to a large extent boils down to modeling the entries of the matrix $\mathbf{H}(t,\tau)$. Assuming a single scatterer present in the channel, one can model a MIMO channel as:

$$\mathbf{H}(\tau) = \int_{-\pi}^{\pi} \int_{0}^{\tau_{max}} S(\theta,\tau') \cdot \mathbf{a}(\theta) \cdot \mathbf{b}(\phi)^T \cdot g(\tau - \tau') \cdot d\tau' \cdot d\theta \,, \tag{2.41}$$

where $S(\theta,\tau)$ is the scattering amplitude drawn from some angle-delay scattering function, $\mathbf{a}(\theta)$ and $\mathbf{b}(\phi)$ are the array response at the receiver and transmitter respectively, $g(\tau)$ is the combined effect of the pulse-shaping filter at the transmitter and the matched filter at the receiver. Under the assumption that there is only one scatterer, $\theta$ and $\tau$ uniquely defines $\phi$ in the above equation. However, in practical applications, (2.41) has a number of limitations and cannot adequately model all the observed channel effects. The removal of the dependency of $\phi$ on $\theta$ and $\tau$ is necessary, and this leads to a requirement of a multiple scattering model, e.g. $S(\phi,\theta,\tau)$. One can show that in richly scattered environment the entries of $\mathbf{H}(t,\tau)$ are independent

zero mean circularly symmetric complex Gaussian random variables. Properly sampling the signals will induce a discrete time channel model, which will simplify the system design. Hereafter, we drop the dependency of the channel $\mathbf{H}(t,\tau)$ on $t$ and $\tau$, and simply denote the channel as $\mathbf{H}$.

Due to insufficient antenna spacing at the transmit and/or receive array, or the presence of a dominant channel component, etc., quite often one can find that the practical channel is not distributed as IID complex Gaussian. When the elements of the antenna arrays are closely spaced, or the environment has few scatterers, there will be correlation between the entries of the channel matrix $\mathbf{H}$, referred to as spatial fading correlation. If the entries of $\mathbf{H}$ are still jointly distributed as complex Gaussian with zero mean and covariance matrix $\mathcal{E}(\mathrm{vec}(\mathbf{H})\cdot\mathrm{vec}(\mathbf{H})^{\mathbf{H}}) = \mathbf{R}$, a correlated channel $\mathbf{H}$ can then be modeled as:

$$\mathrm{vec}(\mathbf{H}) = \mathbf{R}^{\frac{1}{2}} \cdot \mathrm{vec}(\mathbf{H}_w)\,, \tag{2.42}$$

where $\mathbf{H}_w$ denotes a matrix with each of its entries distributed as an iid complex Gaussian random variable, and the $\mathrm{vec}(\cdot)$ operator denotes taking the elements of a matrix and stacking them column by column [71]. To reduce the degrees of freedom required to describe the channel correlation matrix $\mathbf{R}$, a Kronecker product correlation model is often assumed:

$$\mathbf{R} = \mathbf{R}_t \otimes \mathbf{R}_r\,, \tag{2.43}$$

where $\mathbf{R}_t$ and $\mathbf{R}_r$ are the spatial correlation matrix seen from the transmitter side and receiver side respectively, and the $\otimes$ operator denotes the Kronecker product. With this assumption, a wireless channel $\mathbf{H}$ with given correlation matrices $\mathbf{R}_t$ and $\mathbf{R}_r$ can be modeled as:

$$\mathbf{H} = \mathbf{R}_r^{\frac{1}{2}} \cdot \mathbf{H}_w \cdot \mathbf{R}_t^{\frac{T}{2}}\,. \tag{2.44}$$

The above model assumes that there is no dominant propagation path in the environment, therefore each entry of the channel matrix has zero mean. However, when a line-of-sight propagation path exists, or when a certain signal component

dominates, each entry of $\mathbf{H}$ is no longer zero mean, and a Rician channel model is assumed [72]:

$$\mathbf{H} = \sqrt{\frac{K}{1+K}}\bar{\mathbf{H}} + \sqrt{\frac{1}{1+K}}\mathbf{H}_w\,, \tag{2.45}$$

where $\sqrt{\frac{K}{1+K}}\bar{\mathbf{H}} = \mathcal{E}(\mathbf{H})$ is the dominant component of the channel, the other term in the above equation denotes the fading component, and the Rician factor $K$ is defined as the ratio of the power of the dominant path and that of the fading paths.

# Chapter 3

# Signaling Technology over Wireless Channels

As explained in Chapter 2, two major adverse effects of wireless propagation environments are multiple propagation paths (multipath fading) and the time varying nature of the channels (Doppler spread). These effects render the communication links unreliable; e.g. for a fast moving user, multipath fading will lead to fast amplitude and phase fluctuations; for a wide-band signal user, it will create signal dispersion in the time domain and inter-symbol interference; for an analog television signal, it will create "ghost" images (slightly shifted to the right); for a multi-carrier signal, it will induce a varying frequency response at different subcarriers, to name a few [48]. The Doppler spread effect requires constant channel estimation to keep channel knowledge up to date, which therefore reduces the bandwidth efficiency and causes transmission overhead. Another major effect of Doppler spread is that it will cause ICI for a multi-carrier signal and frequency offset for a typical Single Carrier (SC) modulated signal.

Today, there are several popular communication systems in service, including Global System for Mobile Communications (GSM), Digital Enhanced Cordless Telecommunications (DECT), a cellular phone system based on Direct Sequence (DS) CDMA multiple access (IS95), wireless LANs and so on. These systems use different techniques to combat multipath fading and Doppler spread. For example, in a GSM system, the transmitter uses error correction coding to mitigate the effect of multipath fading. Encoders interleave the coded bits to avoid burst errors. And at higher levels, operators design extra fade margins into their cell planning. DECT systems apply diversity reception at the base station. For a IS95 system, wideband transmission technology averages channel behavior and avoids burst errors and deep

fades. For a wireless LAN setup, retransmission protocols such as ARQ and hybrid-ARQ are designed to request retransmission of corrupted data, and short data packets are constructed to combat the fading process. The Doppler spread effect is typically handled by transmitting signals at a much higher rate than the rate at which the channel decorrelates dictated by Doppler spread (GSM, DECT), constantly tracking the variation of the channel (GSM, ISM) and so on.

To create a robust communication link, different signaling techniques tailored for wireless propagation environments are exploited to combat the adverse nature of the channels. The rest of the chapter is dedicated to a summary of the most commonly used transmission and signal strategies for wireless communication, including channel coding, equalization, multi-carrier communication, MIMO technology, etc.

## 3.1 Error Correction Coding

The random nature of the propagation medium causes the link to be unpredictable, and therefore transmission over the link is most likely erroneous every now and then. Inspired by the idea that human error correction capability comes from the ability to make use of redundancy in the information, modern transmission schemes add extra bits to the information before sending them into the channels. However, this redundant information embedded in transmission would certainly reduce the capacity of the communication channel.

Shannon's ground-breaking work in [73], founded the fields of information [74, 75] and coding theory [76, 77, 78, 79], which focus on studying the properties of information and redundancy, as well as the methods to introduce properly controlled redundancy for the purpose of correcting the corrupted message before further processing. With redundancy introduced, the valid messages span only a subspace of all possible transmitted messages. This set of valid messages is referred to as the code, and the valid messages are called codewords. A rule of thumb in designing good codes is to generate codewords that are well-separated according to some metrics. During error detection, the receiver first checks the received message to see if it is a valid codeword. If so, it can be assumed that no error is present, because for a well de-

signed code, it is not likely that there will be enough errors to turn one codeword into another valid one. If not, after the detection of the presence of the error, the receiver has two ways to maintain robustness of the link, i.e. 1) request a retransmission 2) correct the errors. It is not uncommon for the receiver to recover the message by requesting retransmission, but in some situations, e.g. when the propagation delays due to the distance are too large for a delay-sensitive application, this technique is impractical. Under these circumstances, a codeword that is "closest" to the received signal is chosen to correct the errors in the message. This technique is also called Forward Error Correction (FEC).

An exhaustive search over all the codewords to come up with a "nearest" solution to the received signal is often not feasible for codewords with a large number of bits. More intelligent decoding methods that exploit the structure in generating the codewords need to be developed for realistic receiver design. A certain "reliability factor" delivered by a general channel about the qualifying information for each bit can be used to influence the concept and evaluation of "nearness" [76]. This sort of decoding technique is called "soft-decision" decoding. The benefit of a soft-decision technique over one based on hard-decisions can be seen from the analysis of the Binary Erasure Channel [75].

One of the most popular assumptions of coding theory is that the receiver has no prior information about the error pattern, and the errors are supposed to be random and uniform over all bits. However, bursty errors are more probable in a realistic wireless communication system due to short fades, etc. [76]. Various methods are designed to tackle this problem, including 1) introducing an interleaver to spread information over intervals longer than a typical error burst, 2) creating codes that exploit the bursty nature of the errors, and 3) grouping chunks of bits together to form code symbols (e.g. Reed-Solomon codes).

Generally speaking, there are two main types of popular codes: block codes and convolutional codes. Block codes are composed of a collection of fixed length vectors. The generation and analysis of block codes are based on rigorous algebra, which allows codes to be constructed with certain predetermined properties and also

leads to efficient decoding techniques. A subset of block codes that has the property that a linear combination of the codewords is still a codeword, is referred to as linear block codes. Some widely used block codes include: Hamming codes, Hadamard codes, Golay codes, cyclic codes, etc. On the other hand, a convolutional code is generated by passing the information sequence through a few shift register stages. When compared with block codes, the mathematical basis of the construction and analysis of convolutional codes is much weaker. The famous decoding algorithm for convolutional codes, i.e. the Viterbi algorithm, essentially involves searching the codespace with some restrictions to improve the efficiency.

Abundant discussions on the differences between block codes and convolutional codes can be found in the literature [80, 81, 82]. According to [76], block codes usually have higher code rates than convolutional codes. When channel capacity is of concern and no waste of spectral efficiency can be afforded, block codes are more favorable candidates than convolutional codes. When processing time is precious and computing resources are limited, block codes are also preferred due to the existence of efficient decoding algorithms. However, for applications where the communication channels have large bandwidth, a lower code rate can be afforded and higher computational complexity can be justified (e.g. deep space communications), convolutional codes provide a stronger error correcting solution.

To summarize, redundancy can be introduced into the transmission of raw information to allow the possibility of achieving error detection, correction, and even message security. With the fundamental pioneering work of Shannon, various codes have been developed to approach the limit of the channel capacity that Shannon predicted. In modern communications, where larger data rates and higher QoS are desired, benefits of channel coding are many, including: 1) lowering the necessary transmit power for a certain BER performance, 2) adaptively changing the code rate to optimize the system throughput based on instantaneous channel conditions, 3) combining with multiple antenna technology to achieve higher throughput and more reliable link quality.

## 3.2 Wide-band Communication

In wide-band communication systems, the bandwidth of the transmitted signal ($W$) is much larger than the coherence bandwidth ($B_c$) of the channel. For this type of communication, the channel has a fairly large time dispersion compared with the signaling time interval, so ISI is usually the factor that causes the error floor in the performance of the communication system. To combat the adverse effect of ISI, two signaling strategies are widely used, including: 1) time domain equalization that tries to undo the ISI, or 2) multicarrier modulation schemes that modulate the signal on a large number of subcarriers so the equivalent signal bandwidth for each subcarrier is smaller than the coherence bandwidth. The rest of this section is devoted to the review and brief discussion of these methods.

### 3.2.1 Time Domain Equalization

For certain channel realizations, the received signal can be modeled as a transmitted signal passing through a linear filter corrupted by additive white Gaussian noise at the receiver front end. A matched filter followed by appropriate sampling will convert the continuous signal model to a discrete model without loss of optimality. However, due to the self interference caused by multiple propagation paths, signal detection cannot be performed on a symbol-by-symbol basis without a preprocessing to cancel or reduce this interference, otherwise a large performance degradation will result. The preprocessing stage that undoes the interference is called an equalizer in the communication literature.

Equalizers can be classified into two categories: linear and nonlinear. Usually with higher complexity and implementation cost, nonlinear equalizers outperform linear equalizers. From a probability-of-error point of view, an equalizer based on Maximum Likelihood (ML) sequence detection is optimum, but the computational complexity is exponential in the length of the channel time dispersion [59]. This prohibitive cost calls for suboptimal equalizers which exploit criteria such as peak distortion or mean square error (MSE). Both of these criteria lead to linear equalizer implementations. To achieve better performance over a linear equalizer, but avoid

the exponential complexity of MLSE, equalizers that use previously detected symbols to cancel ISI are called Decision Feedback Equalizers (DFEs). To capture the time variation of the channel, equalizers that adaptively change their weights are designed to automatically track the channel and cancel the ISI.

A good treatment of equalization can be found in [83]. Here, the discussion follows the notation in [59]. The received signal can be expressed as:

$$r_l(t) = \sum_n I_n \cdot h(t - nT) + z(t)\,, \tag{3.1}$$

where $T$ denotes the symbol period of the system, $I_n$ is the complex symbol for the $n$-th symbol, $h(t) = \int_{-\infty}^{\infty} c(\tau) \cdot g(t-\tau) \cdot d\tau$ represents the response of the channel $c(t)$ to the signal pulse-shaping filter $g(t)$, and $z(t)$ is AWGN noise. It can be shown [59] that a sufficient statistic for estimating the information symbol $I_n$ is:

$$y_n \equiv y(nT) = \int_{-\infty}^{\infty} r_l(t) \cdot h^*(t - nT) \cdot dt\,, \tag{3.2}$$

which can be implemented at the receiver by matching the pulse $h(t)$ and sampling the output with the symbol period $T$. By defining $x_n \equiv x(nT) = \int_{-\infty}^{\infty} h^*(t) \cdot h(t+nT) \cdot dt$, we have the discrete signal model as follows:

$$y_k = \sum_n I_n \cdot x_{k-n} + \nu_k\,, \tag{3.3}$$

where $\nu_k$ is additive Gaussian noise, but with a power spectral density $\Phi_\nu(\omega) = N_0 \cdot X(e^{j\omega T})$ when $|\omega| \leq \frac{\pi}{T}$, and $X(e^{j\omega T})$ is the z-transform of $x_n$ evaluated at $e^{j\omega T}$. The received signal samples $y_k$ should be filtered to whiten the noise process. This can be easily achieved by factoring the z-transform $X(z)$ of the sequence $x_n$. Thus we can rewrite the whitened $y_k$ as:

$$v_l = \sum_{n=0}^{L} f_n \cdot I_{k-n} + \eta_k\,, \tag{3.4}$$

where $\eta_k$ is a white Gaussian noise process, and $f_k$ is the composite response of pulse shaping, channel, matched filter, sampler and whitening filter. The diagram for this signal model is shown in Fig. 3.1.

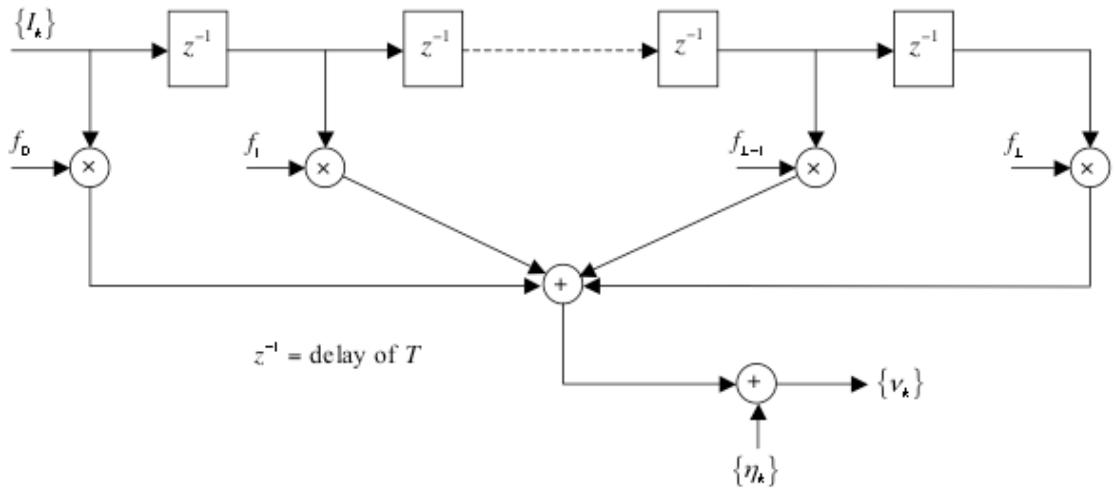



**Figure 3.1:** Discrete-time model for a system with ISI.

The Maximum Likelihood Sequence Estimator (MLSE), which is equivalent to estimating the state of a discrete-time finite-state machine, uses the Viterbi algorithm to determine the most probable path through the $M^L$-state channel trellis. It is the optimal equalizer in the probability-of-estimation-error sense. However, the high complexity of the MLSE limits its usage in practical systems. Suboptimal linear transversal filters that have linear computational complexity are much more popular in practical communication systems. Zero-forcing (ZF) equalizers that minimize the peak distortion defined by

$$D(\mathbf{w}) = \sum_{n=-\infty,\ n\neq 0}^{\infty} |q_n| = \sum_{n=-\infty,\ n\neq 0}^{\infty} \Big| \sum_{j=-\infty}^{\infty} w_j \cdot f_{n-j} \Big| , \tag{3.5}$$

where $\mathbf{w}$ is the weight vector of the linear transversal filter, and $q_n$ is the composite filter response of the equalizer and channel. When the dimension of the weight vector $\mathbf{w}$ is not limited, it can be shown that the equalizer satisfies $C(z) = \frac{1}{F(z)}$, and by

lumping the whitening filter and $C(z)$ together, an equivalent ZF equalizer is written as:

$$C'(z) = \frac{1}{F(z)F^*(z^{-1})} = \frac{1}{X(z)}\,. \tag{3.6}$$

As can be intuitively seen, the ZF equalizer directly inverts the channel. The SNR for the ZF equalizer has the form:

$$\gamma_\infty = \left[\frac{T^2 \cdot N_0}{2\pi}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \frac{d\omega}{\sum_{n=-\infty}^{\infty}|H(\omega + 2\pi n/T)|^2}\right]^{-1}, \tag{3.7}$$

where $H(\omega)$ is the Fourier transform of $h(t)$ defined above. A close look reveals that whenever the channel has a null or has really weak response in the band of interest, the SNR will approach zero and the equalizer performance is poor. This is mainly due to the noise amplification effect of the ZF equalizer. When the number of filter taps in the ZF equalizer is limited to a finite number, ISI cannot be fully cancelled. Furthermore, the design of the weights has to be done using numerical optimization techniques [84]. The MSE equalizer minimizes the mean square error of the estimates of the information symbols $I_k$:

$$J = \mathcal{E}|\epsilon_k|^2 = \mathcal{E}|I_k - \hat{I}_k|^2\,, \tag{3.8}$$

where $\hat{I}_k = \sum_{n=-\infty}^{\infty} w_n \cdot v_{k-n}$ is the estimated information symbol. When the number of taps of the MSE equalizer is unlimited, it can be shown that the combined MSE equalizer with the whitening filter built in has the following z-transform:

$$C'(z) = \frac{1}{X(z) + N_0}\,. \tag{3.9}$$

When the SNR of the system is large enough so that $N_0$ can be considered approximately 0, the MSE equalizer converges to the ZF equalizer. When the channel has a null in the band of interest, the MSE equalizer's response is limited by $1/N_0$, unlike the ZF equalizer case. In this case, the MSE equalizer outperforms the ZF equalizer by limiting noise amplification. The DFE utilizes the previously detected symbols,

and contains a feed-forward and feedback filter. A detailed treatment of the DFE can be found in [85, 86, 87, 88]. The marriage between adaptive filtering theory and equalization help minimize the risk of assuming an a priori known channel. In an adaptive equalizer, the filter's coefficients are adaptively chosen to track the variation of the channel. Details can be found in [84, 89, 90, 91].

### 3.2.2 Multi-carrier Modulation (MCM) Communications

As pointed out in Section 3.2.1, for a frequency selective fading channel, where a non-ideal linear filter can be used to model a specific realization of the channel, a SC modulated signal with a bandwidth larger than the coherence bandwidth of the channel will experience time domain distortion caused by ISI. An equalizer is necessary for the compensation of the ISI effect in order to guarantee reasonable system performance. However, an alternative approach would be to modulate a serial bit stream in parallel using multiple subcarriers, each of which occupies a relatively smaller bandwidth than the original SC system. This is called multi-carrier modulation. Since each carrier spans a smaller bandwidth over which the channel can be assumed to be constant, the frequency response for that subcarrier is therefore flat, inferring little or no ISI for each one of the parallel sub-streams. This MCM technique has gained more and more popularity in 4G mobile communication system proposals [92].

The first systems using MCM were military HF radio links in the late 1950s and early 1960s. However for a classical MCM system, the total available channel bandwidth is divided into a set of non-overlapped sub-channels for frequency multiplexing. Traditional methods used to separate different frequency bands are rooted in the Frequency Division Multiplexing (FDM) technique, where a set of filters are designed to completely separate the subbands. The drawback of this method is the reduction of bandwidth efficiency due to the roll-off factor of the shaping pulse, and the high complexity and cost to build a great number of matched filter banks when the number of subcarriers gets large [48].

Ever since Orthogonal Frequency Division Multiplexing (OFDM) was patented in 1970, MCM has seen more and more applications in practical communication sys-

tems. For example, applications of OFDM are found in: Digital Audio Broadcasting (DAB), Digital Terrestrial Television Broadcasting (DTTB), Universal Mobile Telephone System (UMTS), High Performance LAN (HIPERLAN) Phase II, and WiMAX (IEEE 802.16e), to name a few. Different from traditional technologies that force the subcarriers to be non-overlapping, therefore causing stricter filtering requirements and lower bandwidth efficiency, OFDM carefully selects the spacing between subcarriers such that orthogonality is maintained between them. OFDM uses the Discrete Fourier Transform (DFT) to modulate and demodulate parallel data. The individual spectra are now sinc functions and are not band limited [48]. FDM is achieved, not by bandpass filtering, but by baseband processing, therefore relieving the burden of designing filter banks for each subcarrier. The method of synthesizing OFDM signals can be found in [93].

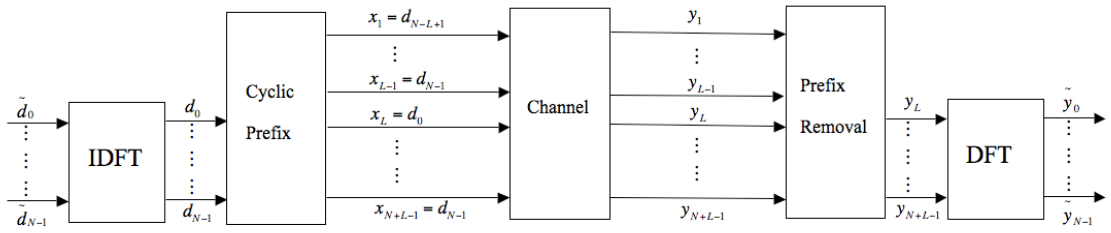



**Figure 3.2:** Block diagram for a typical OFDM system.

The system diagram for OFDM is shown in Fig. 3.2. The transmitter groups the information bits into data frames, each of which contains $N_f$ bits, and further divides each frame into $N$ groups. Each group has $N_i$ uncoded bits, and $\sum_{i=1}^{N} N_i = N_f$. A separate Modulation and Coding Scheme (MCS) can be applied to each group of $N_i$ bits to generate $\tilde{N}_i$ coded bits. A signal mapper maps these $N$ sets of coded bits into $N$ complex symbols, $\tilde{d}_0, \cdots, \tilde{d}_{N-1}$. After passing through the Inverse Discrete Fourier Transform (IDFT) module, $N$ transmit symbols $(d_0, \cdots, d_{N-1})$ have been created. By adding the last $L-1$ symbols $(d_{N-L+1}, \cdots d_{N-1})$ as a Cyclic Prefix (CP) to the transmit symbols, the ISI is to a large extent mitigated. $L$ is chosen to be at least

as long as the number of filter taps used to describe the channel. The receiver removes the CP, passes the received symbol through the Discrete Fourier Transform (DFT) module, and performs detection and decoding on the output of the DFT module to recover the transmitted information bits. The reason why the IDFT and DFT are used for signal modulation can be easily seen in the following derivation [18]. The introduction of the CP converts the linear convolution to a circular convolution, where the DFT can be used to perform frequency analysis. To be more rigorous about the claim, assume the wireless channel has an impulse response $\mathbf{h} = [h_0,\ h_1, \cdots,\ h_{L-1}]^T$. For the received signal samples with the CP removed, we have the following model:

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{bmatrix} = \begin{bmatrix} h_0 & 0 & \cdots & 0 & h_{L-1} & h_{L-2} & \cdots & h_1 \\ h_1 & h_0 & 0 & \cdots & 0 & h_{L-1} & \cdots & h_2 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & 0 & h_{L-1} & h_{L-2} & \cdots & h_1 & h_0 \end{bmatrix} \cdot \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{N-1} \end{bmatrix}. \quad (3.10)$$

It is obvious that the channel matrix in (3.10) is a circulant matrix. One of the most important properties of a circulant matrix is that its eigen-values are the DFT of the vector that creates the matrix, e.g. in this case, $\mathbf{h}$ [94]. Mathematically, if we denote the circulant channel matrix above as $\mathbf{H}$, we have:

$$\mathbf{H} = \mathbf{U}^{-1} \cdot \Lambda \cdot \mathbf{U}\,, \quad (3.11)$$

where $\mathbf{U}$ is a unitary matrix whose $(k, n)$-th element is given by $\frac{1}{\sqrt{N}} \cdot e^{-j\frac{2\pi \cdot n \cdot k}{N}},\ 0 \leq k,\ n \leq N-1$ and $\Lambda$ is a diagonal matrix whose diagonal elements are the DFT of the channel vector $\mathbf{h}$. Equation (3.10) can be rewritten as:

$$\mathbf{Y} = \mathbf{U}^H \cdot \Lambda \cdot \mathbf{U} \cdot \mathbf{d}. \quad (3.12)$$

If we recall $\mathbf{d} = \mathbf{U}^H \cdot \tilde{\mathbf{d}}$, the received signal after passing through the DFT module can be expressed as:

$$\tilde{\mathbf{Y}} = \mathbf{U} \cdot \mathbf{Y} = \mathbf{U} \cdot \mathbf{U}^H \cdot \Lambda \cdot \mathbf{U} \cdot \mathbf{U}^H \cdot \tilde{\mathbf{d}} = \Lambda \cdot \tilde{\mathbf{d}}\,. \tag{3.13}$$

As can be seen, the ISI is eliminated and the symbol detections in each frequency bin can be decoupled. A more detailed discussion on OFDM can be found in [95].

There are a few technical issues that require special attention when implementing an OFDM system, including carrier frequency offset estimation and Peak to Average Power Ratio (PAPR) reduction. The Doppler effect of the channel will introduce ICI, which ruins the orthogonality of the sub-carriers of the OFDM system, because with the shift, each sub-carrier is no longer equally placed with a separation $\frac{W}{N}$, and this will lead to interference between the sub-carriers. To eliminate this problem, careful frequency offset estimation needs to be performed for OFDM to work well. For a typical communication system that uses OFDM as a transmission strategy, the number of subcarriers $N$ is usually a large number. Under this assumption, the central-limit theorem may be used to show the combined signal on the $N$ subcarriers is a zero-mean Gaussian random process, which leads to the PAPR's proportionality to $\sqrt{N}$ [59]. The large signal peaks will result in the clipping of the signal voltage, will cause the amplifier to operate in a non-linear regime, and will distort the transmit signal. Abundant research work on reducing PAPR can be found in the literature [96, 97, 98].

## 3.3 Spread Spectrum Technology

Spread spectrum communication refers to the type of communication system where the bandwidth expansion factor, defined as the ratio of the signal bandwidth $W$ over the information rate $R_b$ (bits/s), is much larger than 1 [59, 99, 100]. Large bandwidth expansion factors, in other words, mean low spectral efficiency, and high redundancy in the transmission scheme. However, by using this specific modulation scheme, the receiver will benefit from a large processing gain by trading off the spectral

efficiency. Spread spectrum communication finds its popularity in both military and civilian communication applications. For military communications, where jamming is a possible detrimental effect, by spreading out the spectrum of the information bearing signal, it is possible to maintain reliable communication between the transceivers. For secure communication links that must not be intercepted by undesired listeners, spread spectrum communication is a good way to disguise the transmitted signal as thermal or background noise. As for civilian communications, one of the most famous multiple access schemes that allows users to occupy the time and frequency resources at the same time while transmitting using different codes, i.e. Code Division Multiple Access (CDMA), is a type of spread spectrum technology. Besides its applications in communication fields, spread spectrum signals can also be used in the radar and navigation fields for making the range and velocity measurements [59]. A significant amount of literature has been devoted to spread spectrum technology; for example, see the tutorial papers [101, 102].

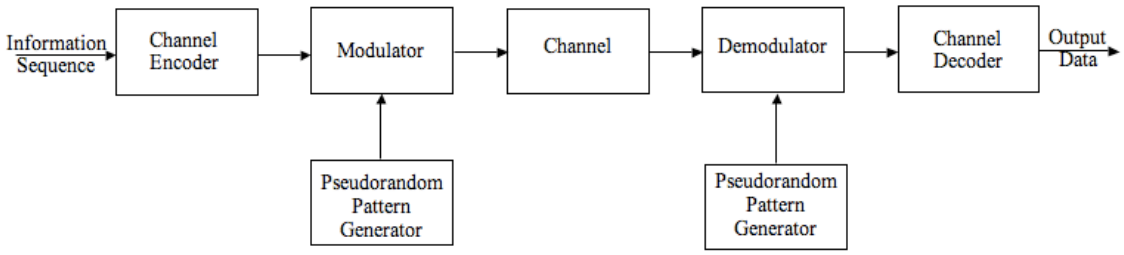



**Figure 3.3:** Block diagram for spread spectrum technology.

A block diagram for a spread spectrum digital communication system is shown in Fig. 3.3. As can be seen in the figure, the coded bits are modulated by a binary sequence of pseudo-random noise (PN) before transmission to the channel. Upon reception, the received signal has to be demodulated with exactly the same PN sequence before channel decoding can be performed. This set of PN sequences is called the "spreading code" in CDMA applications. For this communication system to work, synchronization of the PN sequence needs to be carefully maintained. To achieve this

goal, before each possible data bearing transmission, a fixed PN bit pattern that can be easily recognized by the receiver needs to be transmitted.

Two principal types of spread spectrum signals can be generated: Phase Shift Keying (PSK) or Frequency Shift Keying (FSK). For a system that can maintain phase coherence between the transmit and receive signal for a relatively long period of time when compared with the reciprocal of the channel bandwidth, PSK is usually appropriate for the application. PN sequences used in conjunction with PSK modulation to generate the signals are referred to as Direct Sequence Spread Spectrum (DSSS) signals or pseudo-noise (PN) spread spectrum signals. For those channels that vary quickly with time, e.g. Aircraft-to-Ground (ATG) communication or Aircraft-to-Aircraft (ATA) communication, phase coherence is hard to maintain. In these cases, FSK with a non-coherent envelop detector at the receiver is a good structure for the system design. The PN sequence is used to select the carrier frequency that modulates the transmit signal, and thus generates a Frequency Hopped Spread Spectrum (FHSS) signal. FHSS signals can be classified into two categories: *fast-hopped* signals, when the frequency hopping rate is faster than the symbol rate, and *slow-hopped* signals otherwise. A fast-hopped FHSS signal is usually used in anti-jamming applications to prevent "follower jammers" from having enough time to intercept the signal frequency. These are the two most widely used spread spectrum signals, but there are still other spreading technologies, e.g. the Time Hopping (TH) technique of [59].

The importance of the PN generator is straightforward, and for spread spectrum communication to work properly, the spreading codes in CDMA systems need to possess the following properties [103]:

- The number of runs of 0's and 1's is equal.
- The periodic autocorrelation function (ACF) has (nearly) only two values with a peak at 0 and near zero elsewhere.

The binary sequences that satisfy these constraints are called optimal binary sequences or pseudo-random sequences. The most popular codes used today are maxi-

mum length shift register sequences or m-squences for short. The signals are generated by an m-stage shift register with linear feedback. However the linearity in the feedback determines the vulnerability of the codes, because a jammer may be able to learn the feedback structure by listening to a certain number of transmissions. To reduce the vulnerability, nonlinear feedback structures can be used and feedback connections can be changed from time to time. Poor cross correlation properties between two m-sequence codes [104, 105] hampers the application of m-sequences in CDMA. PN sequences with better cross correlation properties can be created by combining two m-sequences to form Gold codes [106, 107, 108]. For a thorough discussion on code generation, refer to [109].

The spreading waveforms of the transmitter and receiver should be well synchronized. Small synchronization errors will produce a received signal with low SNR and lead to unreliable communication. The synchronization process for a spread spectrum system is composed of two phases: 1) code acquisition, i.e., creating an initial synchronization with the receiver's spreading code, and 2) code tracking, i.e., maintaining the code synchronization after it is established. For a DSSS system, the PN code at the receiver has to be synchronized to within a small fraction of the chip (one bit of a direct-sequence spread spectrum code) interval $T_c = \frac{1}{W}$ [59]. A large channel bandwidth $W$ usually makes this synchronization difficult. Uncertainty in the propagation delay estimate, oscillator instabilities, and Doppler frequency shifts are all factors that contribute to this difficulty. Furthermore, this synchronization has to be established even for low SNR applications or in the presence of jamming [99].

For most systems, this synchronization is obtained at the receiver by continuously searching for a known PN sequence sent by the transmitter. For the synchronization process to be more accurate when noise or interference is present, the receiver should process a certain number of chips $NT_c$ before a decision can be made. The shortest length of the transmitted synchronization sequence can be determined by:

$$T_{\text{syncseq}} = \frac{T_u}{\Delta_t} \cdot NT_c \,, \tag{3.14}$$

where $T_u$ is the timing uncertainty the system can handle, $\Delta_t$ is the resolution of the synchronization; e.g., for a coarse synchronization, $\Delta_t = \frac{1}{2}T_c$, and $N$ is the number of chips needed per synchronization decision. Theoretically, matched filters are the optimal solution for the initial synchronization, but this technique requires filters with extremely large time-bandwith products [99]. This implementation difficulty leads to an easier synchronization structure, i.e., a sliding correlator, where the correlator cycles through the time uncertainty in discrete time intervals of $\Delta_t$, and correlates the received signal with a known synchronization sequence [59]. When the correlator output exceeds a certain predetermined threshold, the receiver claims synchronization is achieved, otherwise the reference sequence is advanced in time by $\Delta_t$, and the same process is repeated. False alarms associated with a typical detection problem are also an issue for the synchronization algorithms. To lower the rate of false alarm, the receiver should keep tracking the output of the correlator to make sure it is always above the predetermined threshold. Once the initial (coarse) synchronization process is finished, finer synchronization to achieve better time resolution and code tracking to maintain this synchronization should be started. A tracking loop for a DSSS system is called a Delay-Locked Loop (DLL), which greatly resembles the idea of the phase-lock loop (PLL) [59]. A pair of shifted reference sequences is generated (shifted $\pm\delta$ in time), and cross correlated with the received signal. The difference between the two outputs is used to drive a Voltage-Controlled Clock (VCC). When the difference is zero, the synchronizer is in an equilibrium state, and the system is in synchronization. Other types of structures like the Tau-Dither Loop (TDL) can be found in [59, 99] and references therein. Since the information rate in the spread spectrum system is much lower than the channel bandwidth, the ISI is usually negligible and an equalizer is not required. A simple receiver called a Rake receiver can be exploited to achieve frequency diversity [18, 59].

## 3.4 MIMO Technology and Smart Antennas

With the development of Internet technology, multimedia services and digital cellular services, the potential market for large bandwidth communications calls for

higher capacity in wireless communications. Turbo and Low Density Parity Check (LDPC) codes approach the Shannon capacity of a SISO link. To further increase the system's capacity, extra degrees of freedom have to be exploited. The deployment of multiple antennas in wireless communications at the transmitter and receiver has attracted great interest from the research and industrial community, for the reason that it opens up spatial degrees of freedom and greatly increases the capacity limit of the wireless channel [110, 111]. Even in the case when multiple antennas are only available on one side of the communication link (either receiver or the transmitter), smart antenna technology [112] still significantly benefits system performance by providing the possibility of diversity reception/transmission, interference cancellation and so on. Tutorial papers about the role and architecture of MIMO in today's communication systems can be found in [113, 114]. The benefit of the multiple antennas does not come without cost. Besides the expenses of installing more antennas in the system, the required signal processing is much more complicated. In [115] and the reference papers cited therein, a good overview of space-time signal processing in multiple antenna systems is provided. A few of the more important aspects of multiple antenna communications will be recapped in the rest of the chapter including: diversity reception technology, space-time coding, and space-time multiplexing.

### 3.4.1 Diversity Reception Techniques

The idea behind the diversity reception technique is to acquire more than one independent or highly uncorrelated sample of a signal, and then use different criteria to either combine them or select a certain copy of the signal for further symbol detection. Clearly a set of independent samples will be much less likely to all fade simultaneously below a certain level when compared with a single sample:

$$\text{Prob}(x_1 < x \cap x_2 < x \cap \cdots \cap x_K < x) = \prod_{i=1}^{K} \text{Prob}(x_i < x) < \text{Prob}(x_j < x)\,, \quad (3.15)$$

for $1 \leq j \leq K$, where $K$ is the number of independent samples. This leads to the concept of diversity reception. The idea of diversity reception is straightforward,

but how to obtain uncorrelated signal samples requires some consideration of system resources. As discussed in Chapter 2, when the signal spans a bandwidth larger than the coherence bandwidth of the channel ($B_c$), frequency components more than $B_c$ Hz apart experience highly uncorrelated fading. This suggests the possibility of frequency diversity. However this is typically not feasible in cellular communication applications, due to the relatively large coherence bandwidth (a few hundred kilohertz up to a few megahertz). Other options like polarization diversity which depends on the depolarizing effect of the scatterers, and field diversity that uses the fact that the electric and magnetic component of the field at any receiving point are uncorrelated [116], have their own difficulties. Time diversity involves retransmitting the failed signal after a certain period of time, and finds its own applications in ARQ protocols. However, time diversity incurs a certain amount of delay overhead, and is not suitable for delay-sensitive applications such as voice communications. With the development of smart antenna technology and multiple antenna systems, space diversity that exploits uncorrelated signal samples from different antennas has become the most popular diversity reception technology in today's mobile communication systems.

As discussed above, uncorrelated samples can be obtained by collecting the signals from different antennas, but how to utilize them in an optimal sense varies differently for various applications and assumptions. A classical paper [117] and the references therein give a thorough treatment of the performance of different linear combining techniques. Depending on when the signal combining is performed, the diversity reception technology can be classified into two categories: pre-detection combining and post-detection combining. The general assumptions behind diversity reception technology include: additive noise is independent of the signal in each branch, the signal component is locally coherent, the noise component is locally incoherent with zero mean finite power, and the local mean square values of the signals are statistically independent [116]. Assuming co-phase processing has been perfectly

achieved, the combined signal can be written as:

$$s(t) = \sum_{i=1}^{K} a_i \cdot s_i(t)\,, \tag{3.16}$$

where $K$ is the number of independent signal branches and $s_i(t)$ is the signal impinging on each antenna. Different combining techniques differ in the choice of the coefficients $a_i$. In scanning and selection combining methods, only one of the $a_i$'s is set to one and the rest to zero. In the scanning diversity system, the combiner sets the $a_i$ of the first branch that has power exceeding a certain preset threshold to be unity. However, in the selection combiner, all the branches are compared, and the coefficient for the branch with the strongest signal is set to unity. A Equal Gain Combiner (EGC) sets all the coefficients to be identically 1's, while a Maximum Ratio Combiner (MRC) sets the coefficients according to each branch's SNR value [65]. When all the aforementioned assumptions hold, the MRC can be shown to yield the best output SNR. But when co-phase processing cannot be achieved, or when the noise in various branches is highly correlated, scanning/selection may outperform the EGC or even the MRC [116]. A detailed discussion and mathematical derivation can be found in [3, 65].

### 3.4.2 Space Time Coding

As pointed out above, diversity reception yields good performance by intelligently combining a number of highly uncorrelated copies of signal in a system with multiple antennas at the receiver. To be able to perform the co-phase processing, channel state information is required. When multiple antennas are available at the transmitter as well, a higher diversity order is possible. However, it is difficult for the transmit side to obtain the channel information correctly. One effective way to achieve this diversty without assuming channel state information at the transmit side is by employing space-time (ST) coding [6, 118]. Coding is carried out both in the space domain (between different antennas) and the time domain (between different time slots) to introduce redundancy between transmitted signals to better exploit the multipath fading characteristics of the channels and achieve more reliable transmis-

sions. There are different kinds of ST codes for different application requirements, including: Space-Time Block Codes (STBC), Space-Time Trellis Codes (STTC), and Layered Space-Time Codes (LST). The number of degrees of freedom in a propagation environment is fixed. One can choose to either exploit the transmit/receive diversity and power gain, or to create multiple data pipes for an even faster signaling rate per channel use. There is always a tradeoff that has to be made to balance the diversity gain and spectral efficiency gain. In other words, high data rate and low error probability are a pair of contradictory system goals. As a benchmark for different ST codes, the tradeoffs between diversity and multiplexing have been studied in [119]. The property that the orthogonal STBC (OSTBC) codes maintain makes it possible for the receiver to exploit the full diversity of the channel with a very simple Maximum Likelihood (ML) detection technique. However the lack of coding gain in STBC calls for more complicated joint error control coding, modulation, transmit and receive diversity design, which leads to STTC codes that are able to provide substantial coding gain, spectral efficiency and diversity improvement over flat fading channels [6, 120]. LST codes can attain a tight lower bound on the MIMO channel capacity and allow processing of multidimensional signals in the space domain by 1-D processing steps [6]. A general code design criteria for slow and fast Rayleigh fading channels based on the pairwise error probability bounds will be reviewed in the rest of this section and due to its simplicity and popularity, STBC will be discussed as well. A discussion of high complexity decoding design such as STTC will be omitted; related references can be found in [6]. LST codes will be briefly addressed in Section 3.4.3.

The basic ST code design criterion relies heavily on the pairwise error probability bound of the code matrix. A code matrix is obtained by arranging the vectors

of the transmitted signal serially in an array, as:

$$\mathbf{X} = [\mathbf{x}_1,\ \mathbf{x}_2, \cdots,\ \mathbf{x}_L] = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_L^1 \\ x_1^2 & x_2^2 & \cdots & x_L^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n_T} & x_2^{n_T} & \cdots & x_L^{n_T} \end{bmatrix}, \tag{3.17}$$

where $n_T$ is the number of antennas at the transmitter, $L$ is the number of ST symbols in the frame under consideration, and each column of the code matrix forms a space-time symbol at each signaling period. The probability that the receiver erroneously selects $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_L)$ as the estimate of the transmitted code matrix $\mathbf{X}$ given the channel condition for all $L$ symbol periods $\mathbf{H} = (\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_L)$ is:

$$P(\mathbf{X}, \hat{\mathbf{X}}|\mathbf{H}) = \mathcal{Q}\left(\sqrt{\frac{E_s}{2N_0} \cdot d_h^2(\mathbf{X}, \hat{\mathbf{X}})}\right), \tag{3.18}$$

where $E_s$ is the energy per symbol at each antenna, $N_0$ is the single-sided power spectral density of the noise, $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt$ is the complimentary error function, and $d_h^2(\mathbf{X}, \hat{\mathbf{X}}) = \sum_{t=1}^{L} \left\| \mathbf{H}_t \left( \mathbf{X}_t - \hat{\mathbf{X}}_t \right) \right\|^2$ is the modified Euclidean distance between the two code matrices $\mathbf{X}$ and $\hat{\mathbf{X}}$. Furthermore, (3.18) can be upper bounded as:

$$P(\mathbf{X}, \hat{\mathbf{X}}|\mathbf{H}) \leq \frac{1}{2} \cdot e^{-d_h^2(\mathbf{X}, \hat{\mathbf{X}}) \cdot \frac{E_s}{4N_0}}. \tag{3.19}$$

Depending on further assumptions about the channel condition, i.e. whether it is fast fading or slow fading and the number of transmit and receive antennas, the bound in (3.19) can be simplified to generate the following four design criterion for Rayleigh fading channels. Let $\mathbf{A}(\mathbf{X}, \hat{\mathbf{X}})$ denote the code distance matrix, and it is defined as the outer product of the difference between two code matrices: $(\mathbf{X} - \hat{\mathbf{X}}) \cdot (\mathbf{X} - \hat{\mathbf{X}})^H$. For a slow fading Rayleigh channel, when $n_R \cdot n_t < 4$, we have the following design criteria for ST codes: over all pairs of distinct code words, 1) maximize the minimum rank of the code distance matrix $\mathbf{A}(\mathbf{X}, \hat{\mathbf{X}})$, and 2) maximize the minimum determinant of $\mathbf{A}(\mathbf{X}, \hat{\mathbf{X}})$. The above design criteria is referred to as the rank and determinant criteria

or Tarokh/Seshadri/Calderbank (TSC) criteria. When $n_R \cdot n_t \geq 4$, the trace criteria described below should be used to generate the ST codes: over all pairs of distinct code words, 1) make sure the product of the rank of $\mathbf{A}(\mathbf{X}, \hat{\mathbf{X}})$ and $n_R$ is no less than 4, and 2) maximize the minimum trace of $\mathbf{A}(\mathbf{X}, \hat{\mathbf{X}})$. For a fast fading Rayleigh channel, (3.19) still holds, but the channel can't be assumed to be constant for $L$ consecutive symbol periods. Let $\rho(\mathbf{X}, \hat{\mathbf{X}})$ denote the set of symbols in which the codes $\mathbf{X}$ and $\hat{\mathbf{X}}$ differ. The cardinality of $\rho(\mathbf{X}, \hat{\mathbf{X}})$ is called the space-time symbol-wise Hamming distance between the two codewords, and is denoted as $\delta_H$. Similar to the above two design criterion, when $\delta_H \cdot n_R < 4$, we have the third design criteria for ST codes: over all pairs of distinct code words, 1) maximize the minimum space-time Hamming distance $\delta_H$, and 2) maximize the minimum product distance, $d_p^2 = \prod_{t \in \rho(\mathbf{X}, \hat{\mathbf{X}})} |\mathbf{x}_t - \hat{\mathbf{x}}_t|^2$ along the path $\delta_H$. However, when $\delta_H \cdot n_R$ is no less than 4, we have the following criteria: over all pairs of distinct code words, 1) make sure $\delta_H \cdot n_R$ is no less than 4, and 2) maximize the minimum Euclidean distance $d_E^2 = \sum_{t \in \rho(\mathbf{X}, \hat{\mathbf{X}})} |\mathbf{x}_t - \hat{\mathbf{x}}_t|^2$. For details, refer to [118]. However these code design criteria are developed based on a specific channel distribution, e.g. the Rayleigh fading channel. Designing codes that perform well for other channel distributions under the slow fading assumption can be found in [121].

The criteria mentioned above can be used to generate different kinds of codes that provide a certain diversity order and coding gain. Orthogonal space-time block codes are a type of code that can extract the full diversity of the system but achieve no coding gain. Signal processing decouples the decoding process at the receiver and makes the ML detection much simpler. With OSTBC, transmit diversity can be achieved without feeding back the channel condition from the receiver to the transmitter. The Alamouti scheme is the most popular STBC code designed for two transmit antennas system, and it can be applied to systems with an arbitrary number of receive antennas [6]. This scheme assumes the channel stays constant for

two consecutive symbols, and the code matrix has the following structure:

$$\mathbf{X} = \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix}, \tag{3.20}$$

where $x_1$ and $x_2$ denote the complex information symbols to be transmitted. It is obvious that redundancy has been introduced both in the space and time domains. For a system with $n_R$ receive antennas, it can be shown that the optimal decision statistics for estimating $x_1$ and $x_2$ are given as:

$$\tilde{x}_1 = \sum_{j=1}^{n_R} h_{j,1}^* r_1^j + h_{j,2}(r_2^j)^*$$

and

$$\tilde{x}_2 = \sum_{j=1}^{n_R} h_{j,2}^* r_1^j - h_{j,1}(r_2^j)^* .$$

The decoupling between the two decision statistics leads to the following ML decoding rules:

$$\hat{x}_1 = \arg\min_{\hat{x}_1 \in \mathcal{S}} \left[ \left( \sum_{j=1}^{n_R} (|h_{j,1}|^2 + |h_{j,2}|^2) - 1 \right) |\hat{x}_1|^2 + d^2(\tilde{x}_1, \hat{x}_1) \right]$$

and

$$\hat{x}_2 = \arg\min_{\hat{x}_2 \in \mathcal{S}} \left[ \left( \sum_{j=1}^{n_R} (|h_{j,1}|^2 + |h_{j,2}|^2) - 1 \right) |\hat{x}_2|^2 + d^2(\tilde{x}_2, \hat{x}_2) \right],$$

where $\mathcal{S}$ denotes the set of constellation points for the modulation, and $d(\cdot, \cdot)$ represents the Euclidean distance between two elements. The orthogonality between the transmitted sequences relieves the burden of joint symbol detection. The idea can be generalized for systems with an arbitrary number of transmit antennas. However for systems with more than 2 transmit antennas, the STBC code can only obtain full rate in very special circumstances. Therefore there is a loss in spectral efficiency as a tradeoff [6]. STBC has found applications in some commercial communication standards, e.g. 802.16e (WiMAX) [122].

### 3.4.3 Space Time Multiplexing

Layered Space Time (LST) codes are designed to approach the increased channel capacity that MIMO technology promises. Foschini proposed a structure that is able to achieve the lower bound on channel capacity [123]. It is straightforward to see that by multiplexing multiple data streams per channel use, the data rate of the system will be increased. The improvement in the data rate comes at the price of more complicated signal processing at the receiver side. However, LST codes convert the joint-stream multidimensional detection problem into simple 1-D spatial domain processing to reduce implementation complexity.

The most famous LST structure for an uncoded system is called the Vertical Bell Laboratories Layered Space-Time (VBLAST) algorithm. In this scheme, the input binary information bits are demultiplexed into $n_T$ parallel streams, and each stream is modulated independently for transmission. The signal processing chain related to each individual substream is named a layer. To further improve the performance of a VBLAST system, one dimensional codes can be implemented on each substream. Horizontal Layered Space Time (HLST) is one example of this type of architectures. It has two implementation types: 1) the input bit sequence is encoded together before being demultiplexed into $n_T$ substreams, and each substream is modulated and interleaved separately for each transmit antenna, and 2) the input bit sequence is demultiplexed first into $n_T$ substreams, and then each substream gets encoded, modulated, and interleaved separately for transmission. The purpose of interleaving the coded bits is to avoid error bursts so some of the errors can be corrected. A better structure is to distribute the modulated codeword across all the transmit antennas, in order to extract higher diversity order. This structure is called the Diagonal Layered Space Time (DLST) architecture. This structure resembles the HLST structure with the difference that before the interleaver module, a Space Interleaver (SI) module is inserted to interleave the codewords in the space domain. By changing the design of the Space Interleaver, a Threaded Layered Space Time (TLST) code can be obtained. For details and diagrams of these structures, please refer to [6, 7] and the references therein.

The benefit of the capacity increase that LST brings comes at the price of a more complicated receiver design. Due to the simultaneous transmission of multiple streams, cross stream interference occurs, and the optimal receiver design involves joint estimation of all the streams, and the complexity of this optimal ML receiver is exponential with respect to the number of transmit antennas. Simple linear receiver structures like Zero Forcing (ZF) and Minimum Mean Square Error (MMSE) receivers are easier to implement but have poorer performance when compared with nonlinear receivers that do successive interference cancellation.

For a narrowband MIMO channel, we have the following linear receive signal model:

$$\mathbf{r}_t = \mathbf{H} \cdot \mathbf{x}_t + \mathbf{n}_t \,, \tag{3.21}$$

where $\mathbf{r}_t$ denotes the $n_R$-by-1 received signal vector at time $t$, $\mathbf{x}_t$ is the $n_T$-by-1 transmitted complex symbol vector, $\mathbf{H}$ is the channel matrix, and $\mathbf{n}_t$ is the $n_R$-by-1 AWGN noise vector at the receive antennas, which has a single sided spectrum height of $N_0$ for each entry. For an uncoded system, a zero forcing receiver simply tries to invert the channel matrix, and then perform separate detection for each stream:

$$\hat{\mathbf{x}}_t = \mathcal{Q}_n \left\{ \mathbf{H}^\dagger \cdot \mathbf{r}_t \right\} \,, \tag{3.22}$$

where $\mathbf{H}^\dagger$ is the pseudo inverse of the channel matrix, equal to $(\mathbf{H}^H\mathbf{H})^{-1} \cdot \mathbf{H}^H$ when $\mathbf{H}$ has more rows than columns. $\mathcal{Q}_n \{\cdot\}$ is a "slicer" that performs hard decisions on the data. The drawback of ZF receivers is the noise amplification problem when the channel is nearly rank deficient. The MMSE receiver remedies the noise amplification problem, and is basically a regularization method that makes the matrix inversion better conditioned:

$$\hat{\mathbf{x}}_t = \mathcal{Q}_n \left\{ \left[ \mathbf{H}^H\mathbf{H} + N_0\mathbf{I}_{n_T} \right]^{-1} \cdot \mathbf{H}^H \cdot \mathbf{r}_t \right\} \,. \tag{3.23}$$

In low SNR regimes, $N_0\mathbf{I}_{n_T}$ dominates $\mathbf{H}^H\mathbf{H}$, and thus amplification of the noise is avoided. However, the receiver's performance can be further improved by use of Suc-

cessive Interference Cancellation (SIC), which estimates the streams sequentially in order of decreasing SNR. Every decoded stream will be remodulated and subtracted from the received signal for interference reduction. MMSE-SIC is optimal from information theoretic standpoint [18]. Decoding algorithms for VBLAST systems that use SIC technologies can be found in [124, 125, 126].

# Chapter 4

# Performance Analysis for Practical Channel-Aware Scheduling

Optimal utilization of the wireless spectrum requires efficient and intelligent radio resource allocation. Recently, the idea of opportunistically allocating transmission resources has gained wide-spread popularity in the context of a multi-user cellular system. The results of [16, 17] show that scheduling the best user, in terms of instantaneous received SNR, achieves the Shannon sum-capacity for both the uplink and downlink of a symmetric (equal average received SNR for all users) cellular network with single-antenna transceivers. In addition to employing channel codes with long block lengths, in order to achieve capacity, the channel state information (CSI) has to be perfect at both the User Terminals (UTs) and the BTS. However, for a practical system operating at non-negligible error probability, where perfect CSI is not available, the correctly delivered throughput (or spectral efficiency) is a more useful figure-of-merit. In this chapter, we consider the downlink performance (in terms of spectral efficiency) of a practical wireless system comprised of a single BTS and multiple UTs that schedules transmission resources over a given interval to the one UT that has the best channel conditions [19]. Our system model is different from the existing approaches [16, 17, 19] in the following aspects:

- The system employs a finite number of MCS.
- Short block length channel codewords are considered with non-negligible block error rate (BLER).
- A limited number of pilots are assumed for MMSE channel estimation.

- The effect of UT mobility and Doppler spread of the channel is explicitly modeled.

- A limited number of feedback bits are assumed for reporting channel quality.

While Shannon capacity is impractical to attain, by adjusting the number of bits per symbol and the code rate for various channel conditions, Shannon capacity can be closely approached in the vicinity of operating SNRs. To close the gap between Shannon capacity and the achievable system throughput, a number of different MCS schemes are usually implemented in the FEC block of a typical wireless communication system. For example, in the 802.16e standard [127], a number of MCS, ranging from QPSK rate 1/2 to 64QAM rate 5/6, are supported in the channel coding module. Depending on the data payload and channel allocation, the block length of the codewords can be relatively short, e.g. it can be as few as 96 QAM symbols (2 slots, 48 bytes) for 64-QAM rate $\frac{2}{3}$ in 802.16e [127]. Limited block lengths further hinder the achievement of Shannon capacity and its effect should be addressed when the system throughput is quantified. It is also reasonable to assume that practical communication systems operate at a relatively high BLER, e.g. 1% or 10% in a WiMAX system. The aforementioned aspects suggest that a better figure of merit when evaluating system performance would be the correctly delivered throughput, or spectral efficiency. For a number of purposes including message decoding, user selection, etc., CSI has to be obtained with a certain degree of accuracy at both the BTS and UTs. The most common way to obtain channel estimates is through embedded pilot symbols, but their use also creates overhead in the system. When a dedicated control channel is available to feedback the UT's channel or SNR estimate to the BTS, the finite precision of the quantizer and scarce feedback channel bandwidth limit the number of bits that are available for reporting the channel quality. Even without inaccuracies due to estimation error and feedback quantization, when the network consists of high mobility users, the channel can be largely de-correlated by the time the BTS receives the CSI report.

To study the effects of the aforementioned practical issues, including channel estimation, user mobility, finite feedback precision, etc., on reducing the system's spectral efficiency, in this chapter, a SISO block fading channel, where fading is constant over the duration of one codeword, is assumed. Using extensive simulations for a practical WiMAX system, we present a simple but accurate mathematical model for the system's BLER behavior. For each fixed MCS, the BLER model is given as a function of instantaneous receive SNR at the mobile users. With the proposed model, the spectral efficiency of a practical opportunistic scheduler can be analyzed in closed form. One of the main contributions in this chapter is that closed form expressions for system's spectral efficiency under various practical assumptions can be derived as a function of the number of users in the system, the number of MCS classes exploited at the FEC module, the system operating SNR, the level of quantization accuracy, the number of inserted pilot symbols, the SNR of the embedded pilot symbols, the feedback delay, and the Doppler spread induced by user mobility. Although the aforementioned analytical results are obtained for a symmetric network with equal average receive SNR for all users, the way to extend the analysis framework to an asymmetric network with dissimilar average SNR across the users is also briefly discussed. Our numerical results reveal that when compared with imperfect channel estimation (with a reasonable pilot SNR) and channel quantization, feedback delay and Doppler spread can significantly degrade the system performance due to the resulting channel decorrelation.

The rest of the chapter is organized as follows. Section 4.1 introduces the system models, where the system under inspection is described, and a piecewise log-linear model for the block error rate as a function of instantaneous received SNR is presented. In Section 4.2, the spectral efficiency of a symmetric network with perfect CSI at the scheduler side is studied. Section 4.3 models imperfect channel estimation and feedback delay and extends the results of Section 4.2. The effects of channel quantization on the throughput performance are discussed in Section 4.4. In Section 4.5, the possibility of extending the framework of the analysis for an asymmetric

network is briefly discussed. Results and discussion are given in Section 4.6. The chapter is summarized in Section 4.7.

## 4.1 System Model and Assumptions

### 4.1.1 Practical Opportunistic Scheduler and General System Assumptions

The practical opportunistic system under investigation is depicted in Fig. 4.1. In such a system, we study a downlink scheduling problem, where based on the

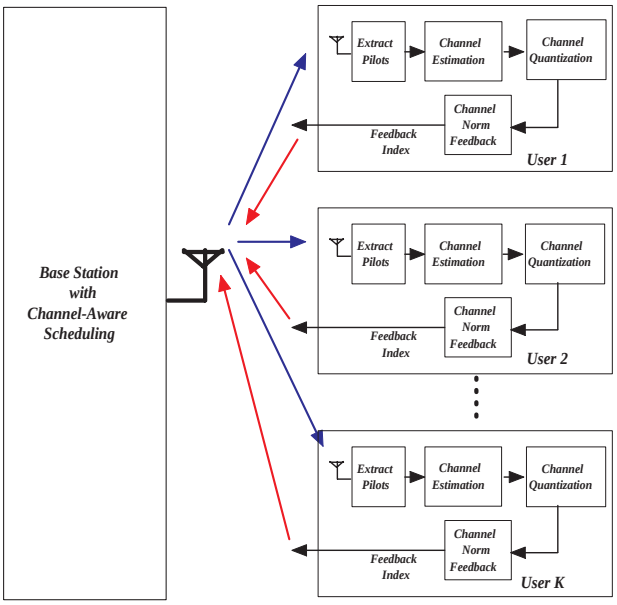



**Figure 4.1:** Block diagram for a practical opportunistic scheduler.

channel knowledge, BTS selects the optimal user and allocates the RF resource to it for a number of time slots. A single antenna is assumed for both the BTS and UTs. The BTS can acquire channel knowledge either by estimating the channel directly through the reverse link or getting feedback from the user through the control channel. Typically, as illustrated in Fig. 4.1, at the user side, channel estimation is performed using the embedded pilot symbols. A quantized instantaneous SNR value per user is then fed back to the BTS for scheduling purposes. As discussed above, in this

system setup, the finite pilot power, limited number of pilot symbols, finite feedback word length, Doppler spread, and feedback delay will all have the effect of limiting the accuracy of the channel SNR information available to the scheduler. Thus, the scheduling decision based on this information will likely cause a certain loss in system spectral efficiency. Hereafter, we use $K$ to denote the number of users in the downlink, and the random variable $\Gamma_j$ to represent the $j$-th user's instantaneous SNR. An iid Rayleigh fading scenario is assumed throughout the paper except Section 4.5. In other words, we assume a symmetric network, where each user has the same average SNR.

### 4.1.2 Piecewise Log-linear BLER Model

As is well known, the adverse nature of the wireless channel hampers reliable communications. To combat the unfavorable EM propagation environment, an FEC module, which introduces redundancy for the information sequences in the time domain, is almost a mandatory module in commercial communication systems. To further improve the link robustness, when the receiver does not successfully decode a code block, an ARQ technique will be implemented to transmit the block once again. The newly received block can be combined with the old one to yield a better estimate of the original information sequence for a higher probability of correct detection. The application of FEC encoders and ARQ techniques in communication systems make analytical BLER evaluation difficult. To account for the effect of the FEC module and the possibility of retransmission, a Monte-Carlo simulation approach is used to study BLER behavior of a coded system with or without ARQ depending on a specific system setup. Extensive simulations for various MCS classes in an AWGN channel reveal the feasibility of modeling the system's BLER behavior with a log-linear model as follows:

$$\mathrm{PER}_j = \begin{cases} 1 & \gamma < \gamma_{T,j} \\ e^{-b_j \cdot (\gamma - \gamma_{T,j})} & \gamma \geq \gamma_{T,j} \end{cases}, \tag{4.1}$$

where PER denotes Packet Error Rate (used interchangeably with BLER in the chapter), subscript $j$ indicates the $j$-th MCS, and $b_j$ and $\gamma_{T,j}$ are model parameters for the $j$-th MCS. The idea behind the model (4.1) is straightforward. Basically, it is

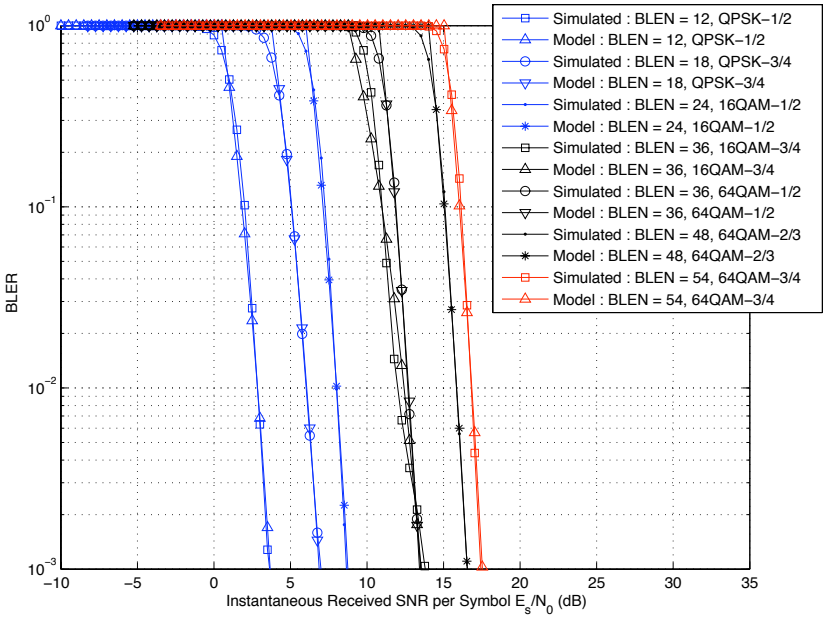



**Figure 4.2:** Comparison of the piecewise log-linear BLER model against Monte-Carlo simulations.

a piece-wise linear model in the log-domain. When instantaneous channel SNR is less than a certain threshold $\gamma_{T,j}$ for the $j$-th MCS, the system is assumed to be in outage, and hence the block error rate is always 1. When the channel is not in a deep fade, the BLER behavior of the system is assumed to have a linear shape in the log-domain for the sake of mathematical tractability. To determine the model parameters for each MCS, the parameters $\gamma_{T,j}$ and $b_j$ are found by fitting simulation results to the assumed model. In particular, the following optimization problem is solved to obtain the required parameters:

$$\{\gamma_{T,j}, b_j\} = \arg\max_{\gamma_{T,j}, b_j} \sum_{k=1}^{M} \left\{ (\gamma_{k,j} - \gamma_{T,j})^2 \cdot \mathbf{1}(\gamma_{\mathrm{k,j}} \geq \gamma_{\mathrm{T,j}}) + (1 - \ln P_{k,j})^2 \cdot \mathbf{1}(\gamma_{k,j} < \gamma_{T,j}) \right\},$$

where, for each modulation scheme, a set of paired data $(\gamma_{k,j}, P_{k,j})$ is collected by simulating a coded AWGN channel with or without ARQ implemented depending on the system configuration. In this chapter Convolutional Turbo Code (CTC) encoder is used for the simulations. The symbol $\mathbf{1}(\bullet)$ denotes an indicator function, which equals 1 when its parameter is true, or 0 otherwise. More specifically, $\gamma_{k,j}$ is the k-th SNR value simulated for the j-th modulation scheme, and $P_{k,j}$ is the corresponding block error rate. It can be shown that:

$$b_j = -\frac{\sum_{k=1}^{M} \ln P_{k,j}(\gamma_{k,j} - \gamma_{T,j}) \cdot \mathbf{1}(\gamma_{k,j} < \gamma_{T,j})}{\sum_{k=1}^{M} (\gamma_{k,j} - \gamma_{T,j})^2 \cdot \mathbf{1}(\gamma_{k,j} \geq \gamma_{T,j})}\,. \tag{4.2}$$

We search for $\gamma_{T,j}$ over all the simulated SNR values to find the parameters for our model. The $\gamma_{T,j}$ and $b_j$ found for the CTC encoder in a WiMAX system are listed in Table 4.1. Curve fit results in Fig. 4.2 show that the model in (4.1) very accurately characterizes the system BLER behavior. The term BLEN in the legend of Fig. 4.2 denotes the codeword length of the FEC block in bytes. However, more accurate model parameters can be achieved by jointly optimizing $b_j$ and $\gamma_{T,j}$ for each MCS.

**Table 4.1:** Parameters for the log-linear BLER model.

| No. Slots | QPSK $\frac{1}{2}$ | QPSK $\frac{3}{4}$ | 16QAM $\frac{1}{2}$ | 16QAM $\frac{3}{4}$ | 64QAM $\frac{1}{2}$ | 64QAM $\frac{2}{3}$ | 64QAM $\frac{3}{4}$ |
|---|---|---|---|---|---|---|---|
| 1 | (1.12, 4.02) | (2.12, 1.77) | (3.99, 1.53) | (9.49, 0.59) | (10.64, 0.42) | (22.49, 0.21) | (14.23, 0.06) |
| 2 | (1.12, 5.71) | (2.38, 2.76) | (3.99, 1.96) | (7.54, 0.46) | (11.94, 0.68) | (25.24, 0.35) | (31.86, 0.28) |

### 4.1.3 Adaptive Modulation and Coding Scheme

As discussed previously, adaptively switching between various MCS schemes based on instantaneous channel SNR will improve the system spectral efficiency. For a system to achieve reasonable performance, the thresholds that control the switching can be chosen to guarantee that the system's BLER operating point is strictly lower than a pre-determined value $P_T$. Using the model (4.1), a simple threshold for the $j$-th modulation scheme can be chosen as: $\gamma_j = -\frac{ln(P_T)}{b_j} + \gamma_{T,j}$. When the instanta-

neous maximum SNR of the network is determined and found to be within the range $[\gamma_j, \gamma_{j+1}]$, the corresponding user will be assigned the entire channel bandwidth and the $j$-th modulation scheme. For notational simplicity, we use $\gamma_0 \equiv 0$ and $\gamma_{N+1} \equiv \infty$.

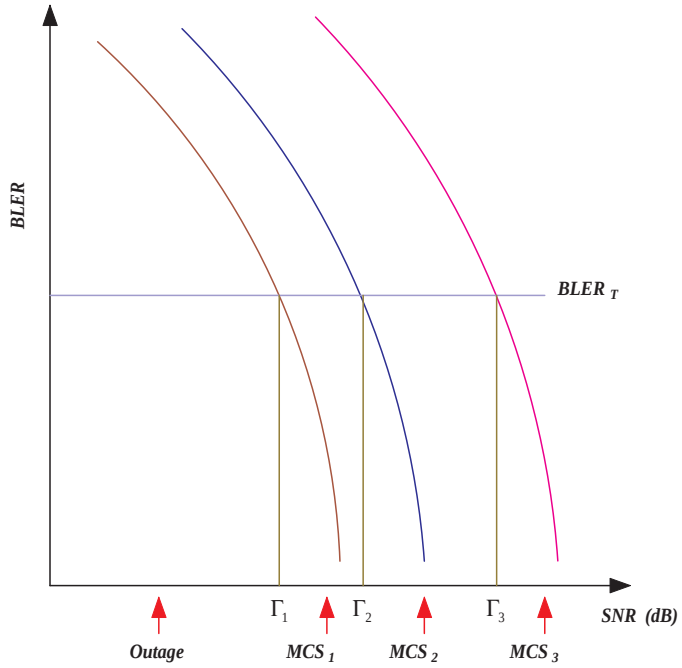



**Figure 4.3:** Adaptive modulation and coding scheme with a target BLER.

In general, the spectral efficiency of interest can be written as:

$$\bar{T} = \mathcal{E}\left(C_j(1 - \mathrm{PER}(j)) \cdot \mathbf{1}\left(\gamma_j \leq \gamma_{\max} < \gamma_{j+1}\right)\right), \tag{4.3}$$

where $C_j = \log_2 M_j$, $M_j$ is the number of constellation points for the $j$-th MCS scheme, $\mathrm{PER}(j)$ is the BLER model for the $j$-th MCS class, $\mathcal{E}(\bullet)$ is the expectation operator which calculates the statistical average of its parameter, $\mathbf{1}(\bullet)$ is the same indicator function defined above, and $\gamma_{\max}$ denotes the maximum SNR among the users. As can be observed from (4.3), $\bar{T}$ depends on the MCS switching thresholds (corresponding to the BLER operating point $P_T$), BLER model parameters ($b_j$ and $\gamma_{T,j}$), the accuracy of the CSI (how $\gamma_{\max}$ is determined), and so on. Depending on

the degree of CSI accuracy available at the scheduler, there are various expressions for $\gamma_{\max}$ above. Therefore, the rest of the chapter will be dealing with evaluating $\bar{T}$ for various assumptions about the CSI at scheduler.

## 4.2 Scheduler with Perfect Channel State Information

The analysis in this section is concerned with the spectral efficiency of a scheduler with Perfect Channel State Information (PCSI) of all users' SNR. In other words, in this case, perfect CSI can be assumed, and no feedback delay or quantization effects are taken into account. This will serve as the benchmark for the throughput of an opportunistic scheduling system. Based on the instantaneous feedback SNR for each user, the scheduler selects the user with the strongest SNR for transmission. Let $\gamma_{\max} = \max(\Gamma_1, \Gamma_2, \cdots, \Gamma_K)$. The Cumulative Distribution Function (CDF) of $\gamma_{\max}$ is given as [128]:

$$\begin{aligned} F_{\gamma_{\max}}(\mu) &= \text{Prob}(\Gamma_1 \leq \mu, \Gamma_2 \leq \mu, \cdots \Gamma_K \leq \mu) \\ &= (1 - e^{-\frac{\mu}{\bar{\gamma}}})^K \,, \end{aligned} \tag{4.4}$$

where $\bar{\gamma}$ is the average SNR for each user, i.e. the SNR operating point of the system. Carrying out the expansion of (4.4), the CDF can be written as:

$$F_{\gamma_{\max}}(\mu) = \sum_{l=0}^{K} (-1)^l \cdot \begin{pmatrix} K \\ l \end{pmatrix} \cdot e^{\frac{-l\mu}{\bar{\gamma}}} \,. \tag{4.5}$$

After taking the derivative of (4.5) with respect to $\mu$, the Probability Density Function (PDF) of the random variable $\gamma_{\max}$ is expressed as:

$$f_{\gamma_{\max}}(\mu) = \sum_{l=0}^{K} (-1)^{l+1} \cdot \begin{pmatrix} K \\ l \end{pmatrix} \cdot \frac{l}{\bar{\gamma}} \cdot e^{-\frac{l\mu}{\bar{\gamma}}} \,. \tag{4.6}$$

Using the PDF expression of $\gamma_{\max}$, the probability of each MCS class actually being chosen can be easily evaluated:

$$\text{Prob}\,(\text{MCS j}) = \begin{cases} \text{Prob}(\gamma_j \leq \gamma < \gamma_{j+1}) = F_{\gamma_{\max}}(\gamma_{j+1}) - F_{\gamma_{\max}}(\gamma_j) & j \neq N \\ \text{Prob}(\gamma_N \leq \gamma < \infty) = 1 - F_{\gamma_{\max}}(\gamma_N) & j = N \end{cases} . \quad (4.7)$$

This is different from a Shannon capacity analysis of this scheduling approach, which enjoys a simple expression as in (4.8):

$$C \equiv \int_0^\infty \log_2(1+\gamma) \cdot f_{\gamma_{\max}}(\gamma) d\gamma\,, \quad (4.8)$$

the closed form expression for $\bar{T}$ in the PCSI case is derived in Appendix A.1. The comparison between the Shannon capacity and system spectral efficiency is shown in Section 4.6.

## 4.3 Effects of Pilot Estimation and Feedback Delay

The spectral efficiency derived in the above section assumes that the scheduler has perfect knowledge of the users' SNR. However when there is error associated with the channel estimates at each user before they are fed back to the scheduler, the inaccurate estimates may cause the BTS to schedule a sub-optimal user, leading to a spectrum waste. The effect of Imperfect Channel State Information (ICSI) will be discussed in this section.

In a typical communication system, receivers usually will extract the pilots embedded in the data to perform the channel estimation. The insertion of the pilots has two effects: 1) finite pilot power induces inaccurate channel estimates, and 2) transmission of the pilots consumes bandwidth and deceases the spectral efficiency. Based on the estimated channel, the receivers calculate their SNR, and send this information to the scheduler. The time-varying nature of the wireless channel renders the feedback outdated due to the channel's Doppler spread. The first subsection will be dedicated to the effects of Pilot-aided channel Estimation and the Feedback Delay

(PEFD). The second subsection will address the system throughput in the PEFD setup.

### 4.3.1 Pilot-Aided Channel Estimation and Feedback Delay

Channel state information at the receiver (CSIR) is mostly achieved through pilot-aided estimation. Due to the limitation on the pilot power, channel estimates can be inaccurate. A very general model valid for Minimum Mean Square Error (MMSE), pilot-symbol-assisted, and additive-channel estimation error model is:

$$g(t) = \rho_e \sqrt{\frac{\Omega_g}{\Omega_h}} h(t) + \sqrt{(1 - |\rho_e|^2)\Omega_g}\omega(t)\,, \tag{4.9}$$

where $\rho_e = E[g(t)h^*(t)]/\sqrt{E[|g(t)|^2]E[|h(t)|^2]}$ is the normalized correlation between the true and the estimated channel gains, $\omega(t) \sim CN(0,1)$, $\Omega_g = E|g|^2$ and $\Omega_h = E|h|^2$. If following assumptions about the channel estimation can be made:

1. $N_p$ pilot symbols available for channel estimation,
2. channel remains constant over the duration of the pilots, and
3. receiver performs MMSE channel estimation,

then the pilots affect the estimation through the cross correlation $\rho_e$ in the form of:

$$\rho_e = \sqrt{\frac{\Omega_g}{\Omega_g + \frac{\sigma^2}{N_p E_p}}} = \sqrt{\frac{\overline{\gamma}_p}{1 + \overline{\gamma}_p}},$$

where $\sigma^2$ is the noise variance of the received signal, $E_p$ is the power per pilot, and $\overline{\gamma}_p = E_p N_p \Omega_g / \sigma^2$ is the average post-processing pilot SNR. Besides the fact that pilot power affects the channel estimates, another effect of the inserted pilots is the decrease in spectral efficiency. The system throughput should be scaled down by a factor of $\frac{N_d}{N_p + N_d}$ to account for the loss, where $N_d$ is the number of data symbols transmitted in between pilots.

The time-varying nature of the wireless channel outdates the channel estimates. To what extent the feedback SNR values or channel estimates describe the

current channel state information highly depends on the Doppler spread of the channel (or equivalently the coherence time of the channel). The feedback delay determines the correlation between the estimates and the current channel state as indicated in equation (4.10):

$$g(t-\tau) = \rho_d(\tau)g(t) + \sqrt{(1-|\rho_d(\tau)|^2)\Omega_g}\lambda(t)\,, \tag{4.10}$$

where $\lambda(t) \sim \mathcal{CN}(0,1)$ and is independent of $g(t)$. Assuming isotropic scattering, $\rho_d(\tau) = \mathcal{J}_0(2\pi f_d \tau)$, where $f_d$ is the Doppler spread of the channel, and $\mathcal{J}(\bullet)$ denotes the zero-th order Bessel function of the first kind. Equations (4.9) and (4.10) can be combined together to account for the channel estimation and feedback delay in one equation:

$$g(t-\tau) = \rho_d(\tau)\rho_e\sqrt{\frac{\Omega_g}{\Omega_h}}h(t) + \varsigma(t), \tag{4.11}$$

where $\varsigma(t) \sim \mathcal{CN}(0, (1-|\rho_d(\tau)|^2|\rho_e|^2)\Omega_g)$ and is independent of $h(t)$.

### 4.3.2 System Throughput for PEFD

After each user estimates its own channel, the scheduler obtains each user's SNR estimate through the feedback channel. The scheduler's knowledge of the $i$-th user's SNR can be written as $\Gamma_{h,i} = \frac{|h_i|^2}{\Omega_h} \cdot \bar{\gamma}_h$, where $h_i$ is assumed to be the user channel estimate, $\bar{\gamma}_h$ is the mean value of the estimated user SNR, and $\Omega_h$ is the variance of the channel estimate. $\bar{\gamma}_h$ and $\Omega_h$ are assumed to be identical for all the users. The BTS selects $j^* = \arg\max_j \Gamma_{h,j}$, and assigns the whole channel to the $j^*$-th user. The transmit power at the BTS is assumed to be $P$. If we use $\hat{\gamma}_{h,\max}$ to denote the largest estimated SNR in the network, by taking into account the estimation error and overhead incurred by the pilots, we can write the system throughput as follows:

$$\begin{aligned}\bar{T} &= \frac{N_d}{N_p+N_d}\mathcal{E}\left(C_j(1-\mathrm{PER}(j))\cdot\mathbf{1}\left(\gamma_j \le \hat{\gamma}_{h,\max} < \gamma_{j+1}\right)\right) \qquad (4.12)\\ &= \frac{N_d}{N_p+N_d}\left\{\sum_{j=0}^{N-1} C_j \int_{\gamma_j}^{\gamma_{j+1}} f_{\hat{\gamma}_{\max}}(\gamma_h)\cdot\widetilde{\mathrm{PSR}}(j)\cdot d\gamma_h + C_N\int_{\gamma_N}^{\infty} f_{\hat{\gamma}_{\max}}(\gamma_h)\cdot\widetilde{\mathrm{PSR}}(j)\cdot d\gamma_h\right\},\end{aligned}$$

where $\widetilde{\mathrm{PSR}}(j) \equiv \int_0^\infty (1-\mathrm{PER}(j,\gamma_g)) f_{\Gamma_g|h}(\gamma_g) d\gamma_g$.

As we can see from the above, $\widetilde{PSR}(j)$ is actually the average packet success rate for each MCS class given the imperfect channel estimate. Since each user's SNR estimate is identically distributed, the subscript $i$ of $h_i$ is dropped for convenience. If we define $\Gamma_g$ as each user's true instantaneous SNR, according to (4.9), $\Gamma_g$ is distributed as non-central $\chi^2$ distribution with two degree of freedom. After some mathematical manipulations, $f_{\Gamma_g|h}(\gamma_g)$ can be written as:

$$f_{\Gamma_g|h}(\gamma_g) = \frac{1}{2\sigma^2} \cdot e^{-\frac{s^2+\gamma_g}{2\sigma^2}} \cdot I_0(\sqrt{\gamma_g}\frac{s}{\sigma^2}) \,, \tag{4.13}$$

where $s^2 = |\rho_{eq}|^2 \frac{|h|^2}{\Omega_h} \cdot \bar{\gamma}_g$, $2\sigma^2 = (1-|\rho_{eq}|^2) \cdot \bar{\gamma}_g$ are the non-centrality parameters and variance respectively, $\bar{\gamma}_g = \Omega_g \cdot \frac{E_s}{N_0}$ is the mean of the true channel SNR, $\frac{E_s}{N_0}$ is the SNR at the transmitter, and $\rho_{eq} = \rho_d(\tau) \cdot \rho_e$. Plugging (4.13) into the definition of $\widetilde{\mathrm{PSR}}(j)$ in (4.12), recalling (4.1), we can rewrite $\widetilde{\mathrm{PSR}}(j)$ as follows:

$$\widetilde{\mathrm{PSR}}(j) = \int_{\gamma_{T,j}}^{\infty} e^{-b_j \cdot (\gamma_g - \gamma_{T,j})} \cdot f_{\Gamma_g|h}(\gamma_g) \cdot d\gamma_g \,. \tag{4.14}$$

After some mathematical manipulations, we have:

$$\widetilde{\mathrm{PSR}}(j) = Q_1\left(\sqrt{p_{1,j}\beta}, q_{1,j}\right) - t_j \cdot Q_1\left(\sqrt{p_{2,j}\beta}, q_{2,j}\right), \ \beta = \frac{|h|^2}{\Omega_h} \tag{4.15}$$

$$p_{1,j} = \frac{2|\rho_{eq}|^2}{(1-|\rho_{eq}|^2)}, \ q_{1,j} = \sqrt{\frac{2\gamma_{T,j}}{(1-|\rho_{eq}|^2)\bar{\gamma}_g}}, \ t_j = e^{b_j\gamma_{T,j}} \cdot \frac{e^{-\frac{|\rho_{eq}|^2 \cdot |h|^2}{(1-|\rho e|^2)\Omega_h} \cdot \frac{b_j(1-|\rho_{eq}|^2)\bar{\gamma}_g}{1+b_j(1-|\rho_{eq}|^2)\bar{\gamma}_g}}}{1+b_j(1-|\rho_{eq}|^2)\bar{\gamma}_g}$$

$$p_{2,j} = \frac{2|\rho_{eq}|^2|h|^2}{(1-|\rho_{eq}|^2)\Omega_h(1+b_j(1-|\rho_{eq}|^2)\bar{\gamma}_g)}, \text{ and } q_{2,j} = \sqrt{\frac{2\gamma_{T,j}(1+b_j(1-|\rho_{eq}|^2)\bar{\gamma}_g)}{(1-|\rho_{eq}|^2)\bar{\gamma}_g}} \,.$$

With $\widetilde{\mathrm{PSR}}(j)$ calculated, the closed form expression for the spectral efficiency in the PEFD scenario is given in the Appendix A.2.

## 4.4 Effects of Quantized Feedback SNR

In order for the receiver to be able to feedback the estimated channel state information, the channel quality measure (either SNR or the channel estimate itself) has to be quantized due to the limited bandwidth of the feedback channel. The

SNR quantization will cause a loss of accuracy in representing each user's channel knowledge, which could induce scheduling errors and a loss of spectral efficiency. Besides the scheduling error, the loss in the resolution can also cause an MCS selection error. This section will discuss the Effect of Quantized Feedback SNR (EQFS) on multiuser scheduling.

### 4.4.1 Lloyd-Max Quantization

Assume that without feedback, the estimated SNR at the mobile can be expressed as $\gamma_h(t) = E_s|h(t)|^2/N_0$. After quantization, the SNR that the BTS receives is $\widehat{\gamma}_h(t) = \mathcal{Q}[\gamma_h(t)]$, where $\mathcal{Q}[\cdot]$ represents the specific quantizer that is employed. Since the decision making at the BTS relies on $\widehat{\gamma}_h(t)$, different choices for the quantization thresholds will affect the performance of the system. One of the optimal ways to determine the quantization thresholds is to use the Lloyd-Max quantizer [129], which minimizes the Mean Square Error (MSE) of the quantization error. Let us assume that the quantizer has a wordlength of $\log_2 N_B$ bits. In other words, the receiver has the ability to quantize a non-negative r.v $X$ into a r.v $\widehat{X}$ with $N_B$ levels. Let $y_1, \ldots, y_{N_B}$ be the reconstructed values, and $x_0 = 0, x_1, \ldots, x_{N_B} = \infty$ be the quantization thresholds. That is, $\widehat{X} = y_j$ if $X \in [x_{j-1}, x_j)$, $j = 1, \ldots, N_B$. Then, for $k = 1, \ldots, N_B$,

$$x_k = \frac{y_k + y_{k+1}}{2} \quad \text{and} \quad y_k = \frac{\int_{x_{k-1}}^{x_k} x f_X(x) dx}{\int_{x_{k-1}}^{x_k} f_X(x) dx}. \tag{4.16}$$

Note that the Lloyd-Max quantizer typically has non-uniform decision regions.

### 4.4.2 System Throughput for EQFS

Assume a $B$-bit quantizer is used at each user, therefore we have $N_B = 2^B$ quantization levels. Let $\hat{\gamma}_s$ denote the $s$-th quantized value, i.e. when the estimated channel SNR satisfies $a_s \leq \gamma_h < a_{s+1}$, $\hat{\gamma}_s$ will be assigned to represent its value, where $a_s$'s are the end points of the quantization intervals. Basically, the set of $\hat{\gamma}_s$, $1 \leq s \leq N_B$ is the reconstruction values, and the set of $a_s$, $1 \leq s \leq N + 1$ defines

the quantization thresholds, or decision regions. The spectral efficiency of the system can be expressed as:

$$\bar{T} = \sum_{j=1}^{N} C_j \cdot \frac{N_d}{N_p + N_d} \cdot \mathcal{E}\left\{(1 - \mathrm{PER}(j, \gamma_g, \gamma_h, \hat{\gamma}_h)) \cdot \mathbf{1}(\gamma_j \leq \hat{\gamma}_h < \gamma_{j+1})\right\},$$

where $\gamma_g$ is used to represent the true SNR for the strongest user, $\gamma_h$ is the estimated maximum SNR, and $\hat{\gamma}_h$ is the quantized maximum estimated SNR. A quantized SNR value $\hat{\gamma}_h$ is a discrete random variable, and given its continuous measurement $\gamma_h$, it has following PDF:

$$f(\hat{\gamma}_h|\gamma_h) = \sum_{l=1}^{N_q} \mathbf{1}(a_l \leq \gamma_h < a_{l+1}) \cdot \delta(\hat{\gamma}_h - \hat{r}_l)\,. \tag{4.17}$$

We can easily derive the joint PDF of $\gamma_h$ and $\hat{\gamma}_h$ as follows:

$$f_{\gamma_h,\hat{\gamma}_h}(\gamma_h, \hat{\gamma}_h) = \sum_{s=1}^{N_q} \sum_{l=0}^{K} (-1)^{l+1} \frac{l}{\bar{\gamma}_h} \begin{pmatrix} K \\ l \end{pmatrix} e^{-\frac{l\gamma_h}{\bar{\gamma}_h}} \cdot \mathbf{1}(a_s \leq \gamma_h < a_{s+1}) \cdot \delta(\hat{\gamma}_h - \hat{\gamma}_s)\,. \tag{4.18}$$

The closed form expression for the spectral efficiency of the system can be obtained using (A.11) and (A.12) derived in Appendix A.3.

## 4.5 Extension of the Analysis Framework to an Asymmetric Network

The above derivation assumes a symmetric network setup. For an asymmetric network, where users have dramatically different average SNR, the Max-SNR scheduler will schedule the user with the largest average SNR most often, causing unfairness in user scheduling. The Proportional Fair (PF) scheduler trades the system's throughput for fairness by scheduling the user with the largest ratio of instantaneous data rate to average transmission rate [130]. Another scheduler that resembles the behavior of the PF scheduler selects the user with the largest normalized instantaneous SNR for transmission. The analysis discussed in above sections can be extended for an asymmetric network with this PF-like scheduler (Quasi-PF scheduler), although more

complex mathematics are involved. This following paragraphs give some insights on how to extend the above analysis to this more general case.

By normalizing each user's instantaneous SNR with respect to its own average SNR, a new set of random variables $z_1, z_2, \cdots, z_K$ can be generated, where $z_i = \frac{\gamma_i}{\bar{\gamma}_i}$, $1 \leq i \leq K$, $\gamma_i$ is the instantaneous SNR and $\bar{\gamma}_i$ is the average SNR for the $i$-th user, respectively. The scheduler therefore selects the user that has the largest normalized SNR value $z_i$, and schedules that user for a number of time slots. Since each user's instantaneous SNR is independently distributed with an exponential distribution, the normalized $z_i$'s are therefore identically distributed with the same exponential distribution. Hence, in a system with a Quasi-PF scheduler, each user is allocated on average with an equal amount of time slots. The PDF of the $j$-th user's normalized SNR with a value $z$ being the maximum can be shown to be $p(j, z) = e^{-z} \cdot (1 - e^{-z})^{K-1}$. With the above PDF, one can easily calculate the system's spectral efficiency by considering each user separately and summing the results as follows:

$$C = \sum_{j=1}^{K} \int_0^{\infty} \log_2 (1 + \bar{\gamma}_j \cdot z) \cdot e^{-z} \cdot (1 - e^{-z})^{K-1} \cdot dz \,. \tag{4.19}$$

Similar manipulations can be used to modify all the derivations in the above sections by treating each user separately and averaging their performance to obtain the overall system throughput.

## 4.6 Results and Discussion

In this section, some results obtained by evaluating the expressions derived above will be shown to give some insights about the multi-user diversity gain in a system equipped with an opportunistic scheduler. The parameters used to generate these plots are given in the figures. Throughout the section, seven MCS schemes are assumed to be adaptively selected at the scheduler, and a WiMAX system with 2-slot allocation size for the FEC block is considered. In other words, the parameters $b_j$ and $\gamma_{T,j}$ assume the values in the second row of Table 4.1. Fig. 4.4 shows the performance gap between the ideal Shannon capacity and the supported spectral efficiency, under

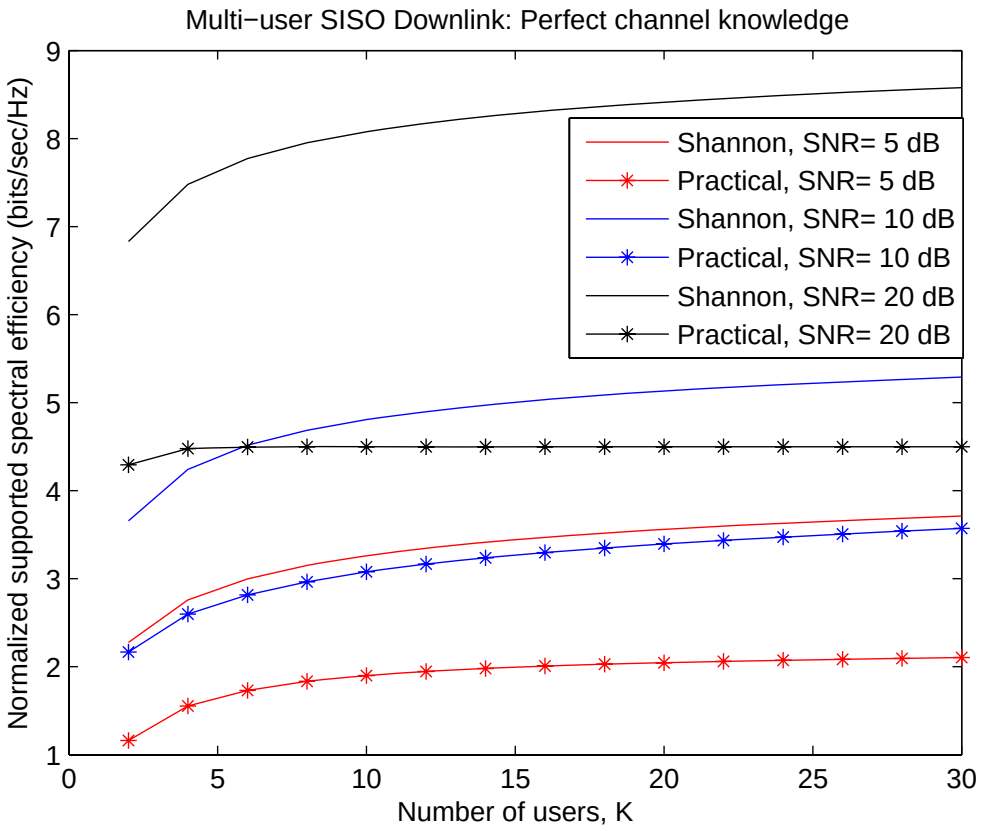



**Figure 4.4:** Shannon sum capacity versus the spectral efficiency achieved by a practical system constrained by a finite number of modulation and coding schemes (performance gap due to finite constellation size).

the assumption that the CSI at the BTS is perfect. It can be clearly observed that, as the number of users in the network increases, the Shannon capacity increases, but the practical supported spectral efficiency saturates at high SNR. This saturation is mainly due to the finite constellation size and can only be improved by packing more bits into each symbol.

As discussed in the above sections, the system's spectral efficiency is a function of the system's BLER operating point, because different BLER operating points will affect the switching thresholds of different MCS classes. A higher BLER operating point means larger error tolerance, which might cause the system's spectral efficiency to go down due to more frequent retransmission. However, on the other hand, it will push the FEC module to use MCS classes with larger constellation sizes and

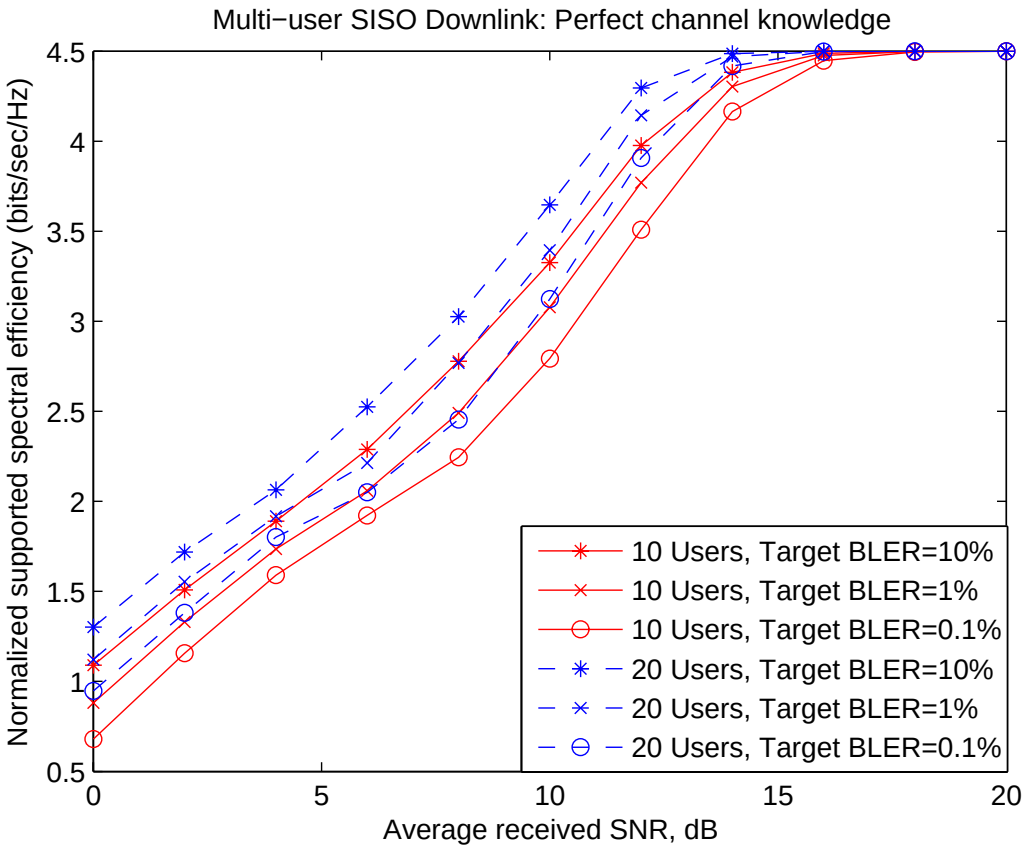



**Figure 4.5:** The spectral efficiency achieved by a practical system constrained by a finite number of modulation and coding schemes (effect of BLER operating point).

higher code rates more aggressively by lowering the MCS switching thresholds. This will lead to an improvement in the spectral efficiency. Therefore, the effect of the BLER operating point is two-fold and it is not obvious which BLER operating point is optimal. Fig. 4.5 reveals that for a few specific BLER operating points (i.e. 10%, 1% and 0.1%), the benefit of lowering switching thresholds beats the loss due to retransmission, and 10% target BLER operating point offers the highest spectral efficiency. Besides, it is obvious to see that the more users in the network, the better the system's spectral efficiency is.

Fig. 4.6 shows the spectral efficiency of a scheduler with imperfect SNR knowledge solely due to limited pilot power. And the loss due to feedback delay and quantization error is not shown. As expected, the higher the pilot power, the larger the spectral efficiency. Higher pilot power implies better channel estimates, which in

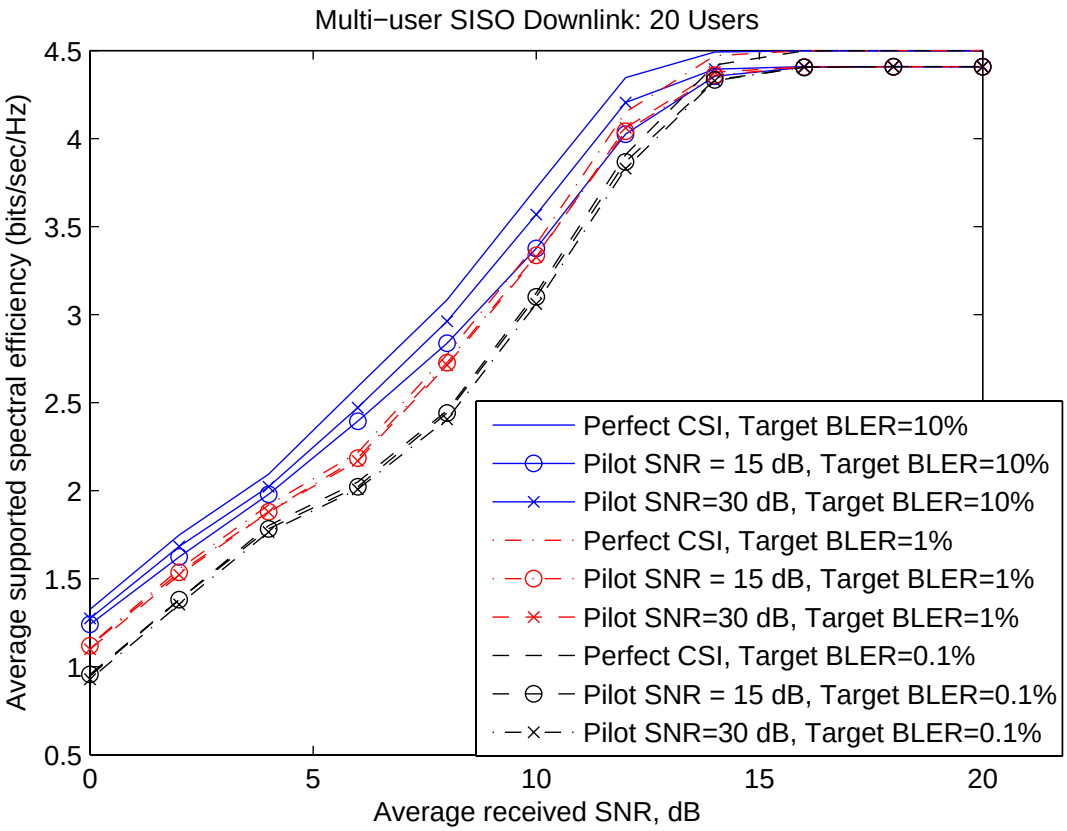



**Figure 4.6:** The spectral efficiency achieved by a practical system constrained by a finite number of modulation and coding schemes (effect of finite pilot power).

turn gives better SNR estimates, and thus a lower probability of mis-scheduling. A BLER operating point of 10% is still the best among the three candidates. Fig. 4.7 considers the imperfect SNR knowledge induced by the feedback delay and Doppler spread. Two BLER operating points are studied. For a fixed feedback delay of 5ms, the larger the Doppler spread, the bigger the spectral efficiency loss due to faster channel decorrelation. By the time the scheduler receives the SNR feedback, it is outdated. Decisions based on these SNR values will likely be suboptimal in the sense that either a non-optimal user will be selected for transmission or a "wrong" MCS class will be exploited at the FEC module. The gaps between the corresponding PCSI and ICSI curves when the Doppler equals 0 are due to the loss of spectral efficiency caused by the insertion of the pilots. Fig. 4.8 compares the spectral efficiency of a scheduler with perfect CSI versus one without. Within a wide range of average SNRs,

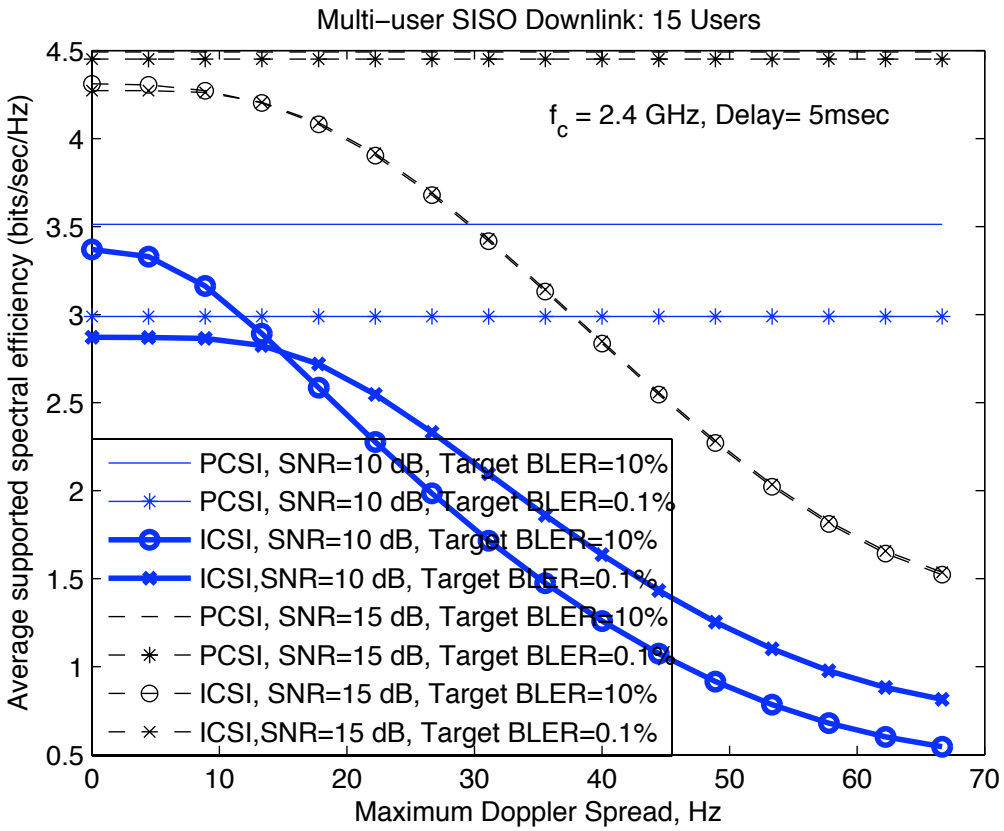



**Figure 4.7:** The spectral efficiency achieved by a practical system constrained by a finite number of modulation and coding schemes (effect of feedback delay and Doppler spread).

the multi-user diversity gain is observed as the gap between the set of curves denoting different numbers of users. As the average SNR increases, both curves saturate, and the multi-user diversity gain is no longer obvious.

## 4.7 Summary

In this chapter, we presented an analytical framework for the realistic performance analysis of a practical channel-aware scheduler. We have proposed a simple log-linear model for the BLER of a CTC operating at non-negligible BLER with a finite number of MCS classes. By taking into account the impact of channel estimation errors, Doppler spread due to user mobility, and the finite number of feedback bits available for channel quality reporting, we have derived closed-form expressions

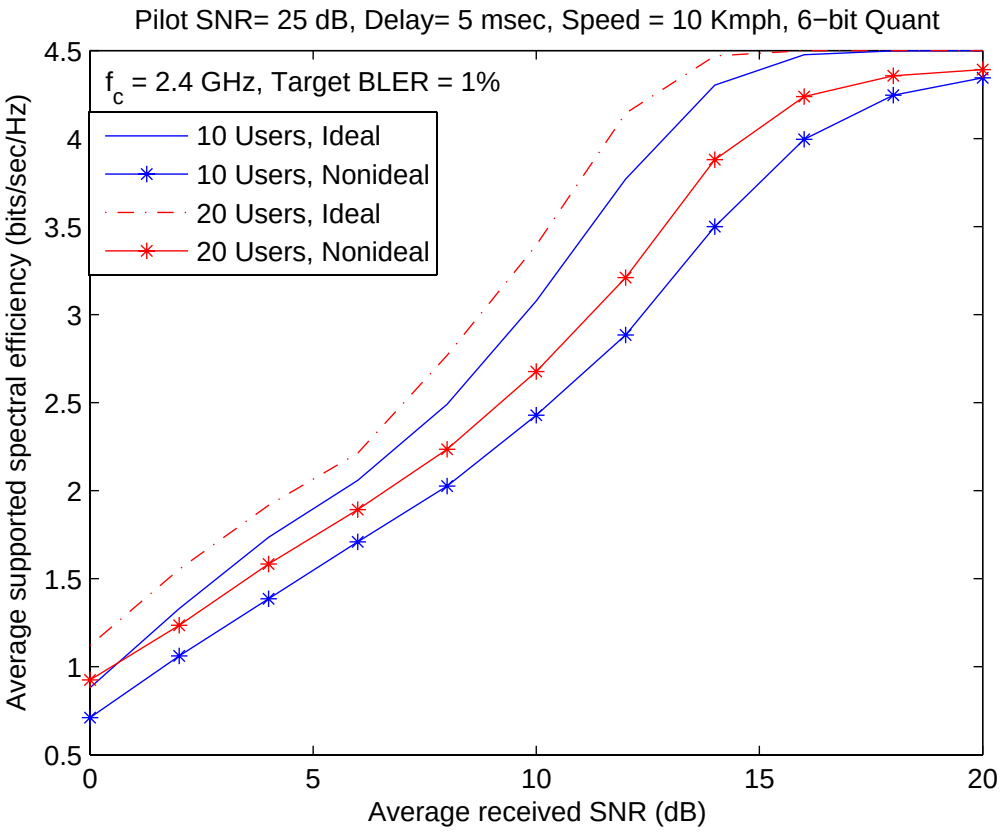



**Figure 4.8:** The spectral efficiency achieved by a practical system constrained by a finite number of modulation and coding schemes (all the practicalities accounted for).

for the average spectral efficiency of a multi-user downlink scheduler on symmetric Rayleigh fading channels. Our results showed that coarse granularity of the number of MCS levels limits the achievable spectral efficiency, whereas imperfect channel estimation, coupled with Doppler and limited feedback, might lead to scheduling a user who does not have the best instantaneous channel quality, thereby causing an additional loss in spectral efficiency. Extensions to an asymmetric network are also discussed for the completeness of the framework.

# Chapter 5

## Impact of User Mobility and Asymmetry on Multiuser Scheduler Performance

In this chapter, we are concerned with the performance of multiuser schedulers over an asymmetric network, where different users in the network have different average received SNRs and Doppler spreads. The throughput, in terms of the ergodic capacity of the channel, of a maximum-SNR scheduler (i.e., the scheduler that schedules the user with the largest instantaneous SNR) is investigated over Rayleigh fading channels with a single-antenna base station and user terminals. Closed-form expressions are presented to quantify the degradation in maximum-SNR scheduler performance due to lack of the knowledge of the CSI. For a comparative study, the performance of a Proportional Fair (PF)-like scheduler is also derived in closed-form. For an asymmetric network with a small number of users, a new simple scheduler is proposed which does not require instantaneous CSI, and its optimality is discussed.

In a single cell, with a single-antenna BTS and UTs, scheduling the strongest user at every transmission time, for either the uplink (UL) or the downlink (DL), is known to be optimal in the Shannon capacity sense [16, 17]. By allowing only the UT with the strongest instantaneous received SNR to transmit (i.e., the so-called maximum-SNR scheduler or, simply, Max-SNR scheduler), the shared channel resource is used most efficiently, and maximum system throughput results [19]. The resulting gain due to riding the peaks of the multiuser fading channel is termed multiuser diversity (MUD) gain [18], which, in a symmetric Rayleigh faded network (with equal average received SNRs for the UTs), grows doubly logarithmically with the number of UTs. However, when the network contains UTs with different average SNRs, the Max-SNR scheduler incurs unfairness, in the sense that the UT with the

largest average SNR gets scheduled more often. The simplest scheduler that guarantees fairness among the UTs is a Round-Robin (RR) scheduler, which schedules the UTs in equal proportion, regardless of the instantaneous UT channel quality. A Proportional-Fair (PF) scheduler trades off the system throughput for fairness across UTs by scheduling the UTs based on the instantaneous channel quality relative to the average channel quality over the time window of interest [18].

In this chapter, our first goal is to understand the impact of dissimilar average SNRs and Doppler spreads on the Max-SNR scheduler performance. Assuming DL transmission, accurate knowledge of channel state information (CSI) at the BTS is crucial to exploit the benefits of MUD. The UT also should have access to good quality CSI to obtain sufficient decoder performance. Lack or inaccuracy of the CSI leads to a significant performance degradation [22, 131], whereas the overhead in maintaining CSI penalizes the system throughput. Typically, a Max-SNR scheduler selects the UT with the strongest SNR based on the initial channel quality measurements (say, at time $t = 0$), and allocates radio resources to the scheduled UT[1] over a duration $T$ that is less than the coherence time $T_c$ of that user. Depending on the ability of the system to track the channel quality of the UTs, we study two scenarios in this chapter. These are: $(i)$ Perfect CSI, and $(ii)$ Outdated CSI. For the first case, we assume the system has perfect CSI only for the active UT during the whole allocated time duration $T$ (i.e., tracking the channel quality of the UT that is scheduled at $t = 0$). In this case, there is a performance degradation due to the fact that the SNR of another user may increase to a value that is better than the scheduled user during the allocation window $T$. For the second case, after making a scheduling decision based on the channel quality at $t = 0$, the BTS makes no attempt to monitor the channel quality over the duration $T$. In this case, there is an additional loss due to decorrelated CSI. The practical significance of both cases $i$) and $ii$) is that they quantify the performance loss due to the lack of periodic channel quality feedback.

Our results for an asymmetric network with a small number of UTs with different mobile speeds show that the Max-SNR scheduler may not necessarily perform

[1]The 'scheduled UT' is also termed the 'active UT'.

better than a scheduler that does not have access to instantaneous CSI. In particular, we propose a SCA scheduler that requires only knowledge of the average received SNRs and the Doppler spreads of the UTs. We formulate a simple constrained optimization problem that maximizes the average capacity of the proposed SCA scheduler in order to allocate the time interval assigned to the UTs. An algorithmic description of the SCA scheduler is also presented.

The rest of this chapter is organized as follows. In Section 5.1, we present the system model. A thorough study of the performance of the Max-SNR scheduler under different CSI assumptions is carried out in Section 5.2. Section 5.3 parallels the derivations in Section 5.2 for its PF-like counterpart, which achieves a sense of scheduling fairness. In Section 5.4, we expose the drawbacks of the Max-SNR scheduler for a small network with high mobility users, and propose our SCA scheduler. We conclude our work in Section 5.5.

## 5.1 System Model

We consider a single cell with a small group of $K$ independently faded asymmetric users. Assuming a single-antenna BTS and UTs, we focus on DL (i.e., BTS to UT) transmission. At time $t$, for UT $j$, $1 \leq j \leq K$, the instantaneous received SNR is denoted by $\gamma_j(t)$, whereas the corresponding statistical average is denoted by $\overline{\gamma_j(t)}$. We also assume that the users move at different speeds, and the resulting Doppler spread for the $j$th UT is given by $f_d^j = (v_j/c)f_c$, where $f_c$ is the carrier frequency, $c$ is the speed of light, and $v_j$ is the speed of UT $j$. The complex-valued time-varying channel gain from BTS to UT $j$ at time $t$ is denoted by $h_j(t)$, which is assumed to be a zero-mean complex-Gaussian (ZMCG) random variable (r.v.)[2] with second moment $\mathsf{E}[|h_j(t)|^2] = \Omega_{h_j}(t)$. With this definition, the received signal at UT $j$ is:

$$y_j(t) = h_j(t)x(t) + n_j(t)\,, \tag{5.1}$$

[2]For simplicity, we denote $\mathsf{X} \sim \mathcal{CN}(m, \sigma^2)$ to indicate that $\mathsf{X}$ is a CG r.v. with mean $m$ and variance $\sigma^2$.

where $x(t)$ is the transmitted signal with $\mathsf{E}(|x(t)|^2) = E_s(t)$ denoting the average transmit power, and $n_j(t) \sim \mathcal{CN}(0, \zeta_j^2)$. With this, we have $\gamma_j(t) = |h_j(t)|^2 E_s(t)/\zeta_j^2$ and $\overline{\gamma_j(t)} = \Omega_{h_j}(t) E_s(t)/\zeta_j^2$.

Due to the users' mobility, the channel for UT $j$ changes for values of $t$ greater than the coherence time $T_c^j$. Since $h_j(t)$ is ZMCG, $h_j(t)$ and $h_0(t)$ are related as [132]:

$$h_j(t) = \rho_j(t) \sqrt{\frac{\Omega_{h_j}(t)}{\Omega_{h_j}(0)}} h_j(0) + \sqrt{(1 - |\rho_j(t)|^2)\Omega_{h_j}(t)}\nu_j(t)\,, \tag{5.2}$$

where $\nu_j(t) \sim \mathcal{CN}(0, 1)$. Assuming the Jakes fading correlation model [3] for each user, we write $\rho_j(t) = \mathcal{J}_0(2\pi f_d^j t)$, where $\mathcal{J}_0(\cdot)$ is the zero-th order Bessel function of the first kind [133]. We denote by $T$ the time slot duration for this group of users (a BTS can serve different groups of users), within which the scheduler can make the decisions about which user to schedule, how much time to schedule it for and so on. We further assume that $T \leq \min_{1 \leq j \leq K}(T_c^j)$.

## 5.2 Max-SNR Scheduler

As mentioned before, the system performance is a function of the quality of the CSI at both the BTS and UTs. The UTs usually estimate the CSI by extracting the pilots inserted in the transmitted data. The BTS can obtain the CSI either by estimating the channel through the reverse link (assuming channel reciprocity), or via feedback from the UTs through a control channel. Keeping track of the CSI will incur a penalty in system throughput. This section addresses the Max-SNR scheduler performance under different scenarios, depending on the ability of both the BTS and the UTs to track the CSI.

### 5.2.1 Ideal Performance

Under this assumption that the BTS has perfect CSI for all the UTs during $0 \leq t \leq T$, the CDF of the maximum of the instantaneous SNR can be easily written

as:

$$P_{\gamma_{\max},t}(x) = \prod_{j=1}^{K} \left(1 - e^{-\frac{x}{\overline{\gamma_j(t)}}}\right) = \sum_{j=1}^{K} \sum_{l_j=0}^{1} (-1)^{\sum_{j=1}^{K} l_j} e^{-x \cdot \left(\sum_{j=1}^{K} \frac{l_j}{\overline{\gamma_j(t)}}\right)} . \tag{5.3}$$

By taking the derivative of (5.3) with respect to $x$, the pdf of the maximum of the instantaneous SNR at any $t \in [0, T]$ can be written as:

$$p_{\gamma_{\max},t}(x) = \sum_{j=1}^{K} \sum_{l_j=0}^{1} (-1)^{1+\sum_{j=1}^{K} l_j} \left(\sum_{j=1}^{K} \frac{l_j}{\overline{\gamma_j(t)}}\right) e^{-x \cdot \left(\sum_{j=1}^{K} \frac{l_j}{\overline{\gamma_j(t)}}\right)} . \tag{5.4}$$

The Shannon capacity of this scheduler is given by:

$$\begin{aligned} C(t) &= \int_0^{\infty} \log_2 (1 + x) p_{\gamma_{\max},t}(x) dx \\ &= \log_2(e) \sum_{j=1}^{K} \sum_{l_j=0}^{1} (-1)^{1+\sum_{j=1}^{K} l_j} e^{\left(\sum_{j=1}^{K} \frac{l_j}{\overline{\gamma_j(t)}}\right)} \mathcal{E}_1 \left(\sum_{j=1}^{K} \frac{l_j}{\overline{\gamma_j(t)}}\right) , \end{aligned} \tag{5.5}$$

where $\mathcal{E}_1(x) = \int_x^{\infty} e^{-t}/t \, dt$ is the exponential integral function [133]. When the transmission power and the average channel gains are time-invariant (i.e., no Doppler), $\overline{\gamma_j(t)}$ will remain time-invariant. In this scenario, the Shannon capacity in (5.5) remains constant over the period $[0, T]$, thanks to the scheduler's knowledge of all users' CSI for the whole time interval.

### 5.2.2 Channel Tracking of the Scheduled User

Due to high CSI requirements, the scheduler described in Section 5.2.1 is impractical. Here, we investigate the case when the scheduler has CSI for all the users *only at time* $t = 0$, based on which the initial scheduling decision is made for the duration $[0, T]$. In this period, only the active UT's channel is tracked at the UT side for the purpose of decoding the message, and is assumed to be perfectly known. Compared to the ideal scheduler, this system takes a performance hit due to not monitoring the channel quality of other UTs beyond $t = 0$. In this subsection, we quantify this loss.

From (5.2), conditioned on $h_j(0)$, it is easy to see that $\gamma_j(t) = |h_j(t)|^2 E_s(t)/\zeta_j^2$ is a non-central Chi-square r.v. with the parameters [59]:

$$s_j^2(t) = |\rho_j(t)|^2 |h_j(0)|^2 \frac{\Omega_{h_j}(t)}{\Omega_{h_j}(0)} \frac{E_s(t)}{\zeta_j^2} = \frac{|\rho_j(t)|^2 \overline{\gamma_j(t)}}{\Omega_{h_j}(0)} |h_j(0)|^2 \tag{5.6}$$

and

$$2\sigma_j^2(t) = (1 - |\rho_j(t)|^2)\Omega_{h_j}(t) \frac{E_s(t)}{\zeta_j^2} = (1 - |\rho_j(t)|^2)\overline{\gamma_j(t)}\,. \tag{5.7}$$

Therefore, the CDF of $\gamma_j(t)$ conditioned on $h_j(0)$ is [59, 134]:

$$P(\gamma_j(t) \leq x | h_j(0)) = 1 - \mathcal{Q}_1\left(\frac{s_j(t)}{\sigma_j(t)}, \frac{\sqrt{x}}{\sigma_j(t)}\right),$$

where $\mathcal{Q}_1(a, b)$ is the first-order Marcum-Q function [59]. It can be easily shown that the PDF of the maximum SNR, with UT $j$ being the strongest user at time $t = 0$ is:

$$\begin{aligned} p_{\gamma,j}(x, j, t = 0) &= \frac{e^{-\frac{x}{\overline{\gamma_j(0)}}}}{\overline{\gamma_j(0)}} \prod_{i=1, i\neq j}^{K} \left(1 - e^{-\frac{x}{\overline{\gamma_i(0)}}}\right) \\ &= \sum_{i=1, i\neq j}^{K} \sum_{l_i=0}^{1} (-1)^{(\sum_{i=1, i\neq j}^{K} l_i)} \frac{e^{-\left(\sum_{i=1, i\neq j}^{K} \frac{l_i}{\overline{\gamma_i(0)}} + \frac{1}{\overline{\gamma_j(0)}}\right)x}}{\overline{\gamma_j(0)}}\,. \end{aligned} \tag{5.8}$$

Upon defining

$$\begin{aligned} F(\alpha, k) &\triangleq \frac{e^{\alpha}}{k!} \sum_{j=0}^{K} (-1)^j \binom{k}{j} \alpha^j \int_{\alpha}^{\infty} e^{-t} t^{k-j-1} dt \\ &= e^{\alpha} \sum_{j=0}^{K} (-1)^j \binom{k}{j} \alpha^j \frac{G(\alpha, k-j)}{k(k-1)\cdots(k-j)}\,, \end{aligned} \tag{5.9}$$

where $G(\alpha, k) = \frac{1}{(k-1)!} \int_{\alpha}^{\infty} e^{-t} t^{k-1} dt$ is the incomplete gamma function [133], the Shannon capacity of a Max-SNR scheduler that tracks the channel of the active UT

for the whole time slot $T$ can be derived as:

$$C(t) = \sum_{j=1}^{K} \int_0^{\infty} C_{j|h_j(0)}(t) p_{\eta_j(0),j}(x, j, t=0) dx$$
$$= \sum_{j=1}^{K} \sum_{i=1, i\neq j}^{K} \sum_{l_i=0}^{1} \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{\log_2(e)}{\overline{\gamma_j}(0)} \times \frac{(b_j(t))^n (-1)^{\sum_{i=1,i\neq j}^{K} l_i} F(\alpha_j(t), k)}{(\sum_{i=1,i\neq j}^{K} \frac{l_i}{\overline{\gamma_i}(0)} + \frac{1}{\overline{\gamma_j}(0)} + b_j(t))^{n+1}}, \quad (5.10)$$

where $\alpha_j(t) = \frac{1}{(1-|\rho_j(t)|^2)\overline{\gamma_j}(t)}$, $b_j(t) = \frac{|\rho_j(t)|^2}{(1-|\rho_j(t)|^2)\overline{\gamma_j}(0)}$ and $C_{j|h_j(0)}(t)$ is the Shannon capacity of the $j$-th user being scheduled given its channel knowledge $h_j(0)$. Details of (5.10) can be found in Appendix B.1.

A Monte-Carlo simulation is carried out to simulate the Shannon capacity of the Max-SNR scheduler over a network with 3 asymmetric users. At time $t = 0$, each user's channel is generated randomly, and the strongest user is selected according to the users' instantaneous SNR at this time. After the user is selected, random samples are generated to evolve the channel according to (5.2). The Shannon capacity is measured by averaging over a large number of channel realizations. For simplicity, the second order statistics of each user's channel is assumed to be constant (i.e., $\Omega_{h_j}(t) = C,\ t \in [0, T]$), and each user's transmit power and noise variance are also set to be constant during the scheduled time slot. The carrier frequency is set to $f_c = 10$ GHz, the average SNRs of the users are set to [15, 25, 20] dB, the Doppler speed is set to [40, 100, 50] m/s, and the time slot duration $T$ is set to be 120 $\mu s$. The same set of parameters will be used throughput this chapter. As can be seen in Fig. 5.1, the simulated results agree very well with the analysis in (5.10). The performance loss can be quantified as the difference between (5.5) and (5.10).

### 5.2.3 Impact of Channel Tracking Inability

When the channel feedback frequency is severely constrained, the UTs provide only the initial CSI to the BTS scheduler. Once a scheduling decision is made, during the scheduling period the BTS will make no further attempts to track the channel of any of the UTs. More importantly, at the UT side, the receiver does not track its own channel. The initially obtained CSI is used throughout its scheduled time to decode

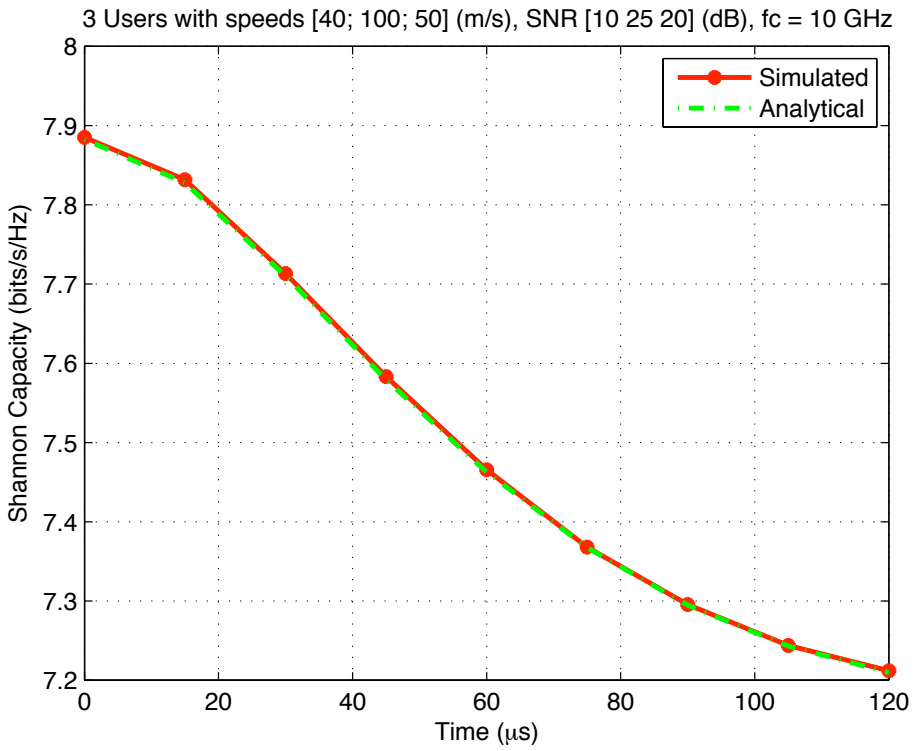



**Figure 5.1:** Shannon capacity of a Max-SNR scheduler. Only the scheduled user at $t = 0$ has access to perfect CSI over $t \in [0, T]$, where $T$ is the scheduling interval.

the message. Clearly, the required CSI load in this case is significantly less than for the approaches in Sections 5.2.1 and 5.2.2. However, not tracking the channel at all after $t = 0$ leads to further performance degradation over the one described in Section 5.2.2.

Using (5.2) in (5.1), the instantaneous SNR of UT $j$ given $h_j(0)$ is:

$$\begin{aligned}\eta_j(t) &\triangleq \frac{|\rho_j(t)|^2 |h_j(0)|^2 \frac{\Omega_{h_j}(t)}{\Omega_{h_j}(0)} E_s(t)}{(1-|\rho_j(t)|^2)\Omega_{h_j}(t)E_s(t) + \zeta_j^2} \\ &= \frac{|\rho_j(t)|^2 |h_j(0)|^2 E_s(t)}{\Omega_{h_j}(0)\left[(1-|\rho_j(t)|^2)E_s(t) + \frac{\zeta_j^2}{\Omega_{h_j}(t)}\right]} \\ &= \phi_j(t)\eta_j(0)\,, \end{aligned} \tag{5.11}$$

where

$$\phi_j(t) \triangleq \frac{|\rho_j(t)|^2 E_s(t)}{E_s(0)/\zeta_j^2 \Omega_{h_j}(0) \left[(1-|\rho_j(t)|^2)E_s(t) + \frac{\zeta_j^2}{\Omega_{h_j}(t)}\right]}$$
$$= \frac{|\rho_j(t)|^2 \overline{\gamma_j(t)}}{\overline{\gamma_j(0)} \left[(1-|\rho_j(t)|^2)\overline{\gamma_j(t)} + 1\right]} . \tag{5.12}$$

Averaging over $\eta_j(t)$ and the users, we obtain the Shannon capacity in the absence of channel tracking as (see Appendix B.2):

$$C(t) = \mathsf{E}_j \left( \mathsf{E}_{\eta_j(t)} \left( \log_2(1+\eta_j(t)) \right) \right) \tag{5.13}$$
$$= \sum_{j=1}^{K} \frac{\log_2 e}{\overline{\gamma_j(0)}} \sum_{i=1, i\neq j}^{K} \sum_{l_i=0}^{1} \frac{(-1)^{\sum_{i=1,i\neq j}^{K} l_i}}{\sum_{i=1,i\neq j}^{K} \frac{l_i}{\overline{\gamma_i(0)}} + \frac{1}{\overline{\gamma_j(0)}}}$$
$$\cdot \; e^{\frac{\sum_{i=1,i\neq j}^{K} \frac{l_i}{\overline{\gamma_i(0)}} + \frac{1}{\overline{\gamma_j(0)}}}{\phi_j(t)}} \mathcal{E}_1\left(\frac{\sum_{i=1,i\neq j}^{K} \frac{l_i}{\overline{\gamma_i(0)}} + \frac{1}{\overline{\gamma_j(0)}}}{\phi_j(t)}\right) .$$

We perform a Monte-Carlo simulation for this scheduler, and the result shown in Fig. 5.2 verifies the accuracy of the analysis. We also observe from Fig. 5.2 that the loss in capacity is significant when the scheduler does not keep track of the user's channel (6 bits/s/Hz loss in Fig. 5.2 as compared to 0.95 bits/s/Hz loss in Fig. 5.1 at the end of the scheduling time $T$). The difference between (5.10) and (5.13) characterizes additional loss due to the UT not tracking its own channel.

## 5.3 Quasi-Proportional Fair (PF) Scheduler

A Max-SNR scheduler maximizes the network throughput by greedily scheduling the strongest user in the network. However, the benefit of riding the peak sacrifices fairness among the users in the sense that a user with a stronger average SNR gets scheduled more often. To mitigate this issue, various algorithms have been proposed, including: Max Fairness algorithm, Proportional Rate Constraint algorithm and so on [130]. One simple and effective approach to balance the throughput of the network and the scheduling fairness is Proportional Fair (PF) scheduling. The PF scheduler selects the user with the largest instantaneous data rate relative to its

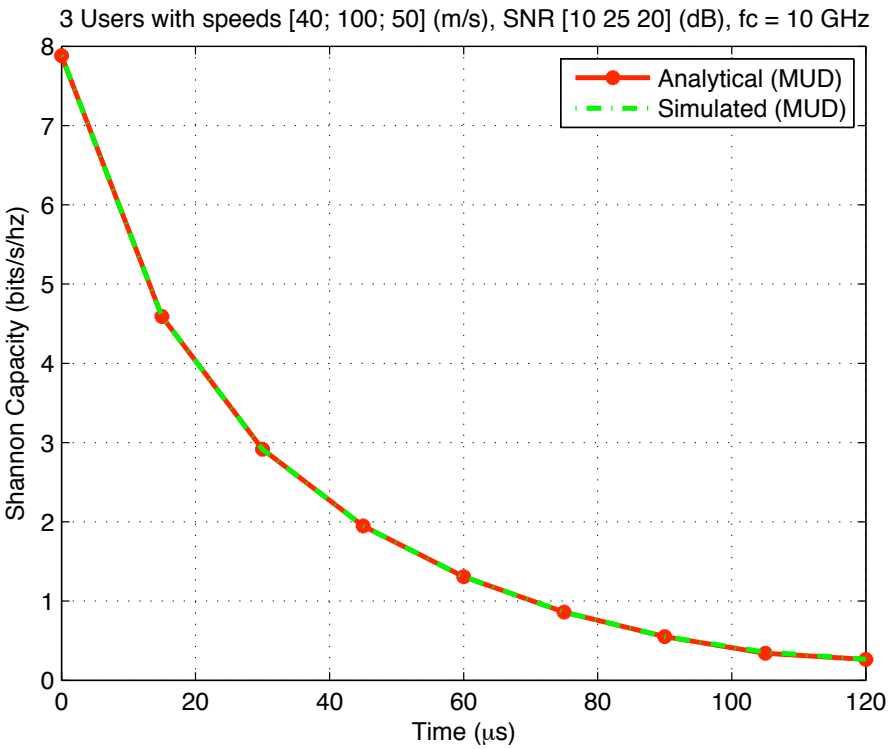



**Figure 5.2:** Shannon capacity of the Max-SNR scheduler. Channel knowledge is assumed for the selected user only at the scheduling time.

average throughput to avoid selecting users with "bad" channel. Details about the PF scheduler can be found in [130] and the references therein. However, due to the logarithm relationship between rate and average SNR, a theoretical analysis for the performance of a PF scheduler is complicated. Herein, we investigate a Quasi-PF scheduler, which selects the strongest user with respect to his own average SNR to achieve fair scheduling. Simply put, a Quasi-PF scheduler normalizes each user's instantaneous SNR $\gamma_j(t)$ with regard to his average channel SNR $\overline{\gamma_j(t)}$, and schedules the user with the maximum normalized value. After normalization, all the users' relative SNR values are i.i.d distributed, and therefore each user gets scheduled with equal probability, which achieves the fairness of the scheduler. Similar to Section 5.2, based on various assumptions at both the scheduler and receiver sides, different system performance can be achieved. The following subsections will be dedicated to deriving closed-form expressions to quantify system throughput under various assumptions.

### 5.3.1 Ideal Performance

When the base station has the luxury of possessing the instantaneous channel information for all users and the entire allocated time duration $[0,\ T]$, it is capable of switching users at every time instant. Since the users' SNR is i.i.d, the Shannon capacity of the network remains as a constant. The key factor in deriving the throughput is to realize that where the $j$-th user has a maximum SNR of $x \cdot \overline{\gamma_j}$, its PDF can be expressed as:

$$\begin{aligned} p(z_j \text{ is max}, z_j = x) &= \int_0^x \cdots \int_0^x p_{z_1,z_2,\cdots z_K}(x_1, x_2, \cdots, x_K) dx_1 \cdot dx_{j-1} \cdot dx_{j+1} \cdots dx_K \\ &= e^{-x} \cdot \left(1 - e^{-x}\right)^{K-1}, \end{aligned} \tag{5.14}$$

where $z_j = \gamma_j(t)/\overline{\gamma_j(t)}$ is the normalized instantaneous SNR for user j. The Shannon capacity of the network can thus be expressed as:

$$\begin{aligned} C(t) &= \sum_{j=1}^{K} \mathcal{E}\left\{1 + \gamma_j(t)\right\} \\ &= \sum_{j=1}^{K} \int_0^{\infty} \log_2\left(1 + x \cdot \overline{\gamma_j}\right) e^{-x} \left(1 - e^{-x}\right)^{K-1} dx \\ &= \frac{\log_2 e}{K} \sum_{j=1}^{K} \int_0^{\infty} \frac{\overline{\gamma_j}}{1 + x \cdot \overline{\gamma_j}} \left[1 - \left(1 - e^{-x}\right)^K\right] dx\,. \end{aligned} \tag{5.15}$$

Using the Binomial Theorem to expand $(1 - e^{-x})^K$, after some mathematical manipulations we have:

$$C(t) = \frac{\log_2 e}{K} \sum_{j=1}^{K} \sum_{i=1}^{K} (-1)^{i+1} \begin{pmatrix} K \\ i \end{pmatrix} e^{\frac{i}{\overline{\gamma_j(t)}}} \mathcal{E}_1\left(\frac{i}{\overline{\gamma_j(t)}}\right). \tag{5.16}$$

### 5.3.2 Channel Tracking of the Scheduled User

Obviously, in order for the BTS to have channel state information for all users and for all time, a large amount of feedback information has to be passed through the feedback link, which results in extra bandwidth consumption and causes extra power

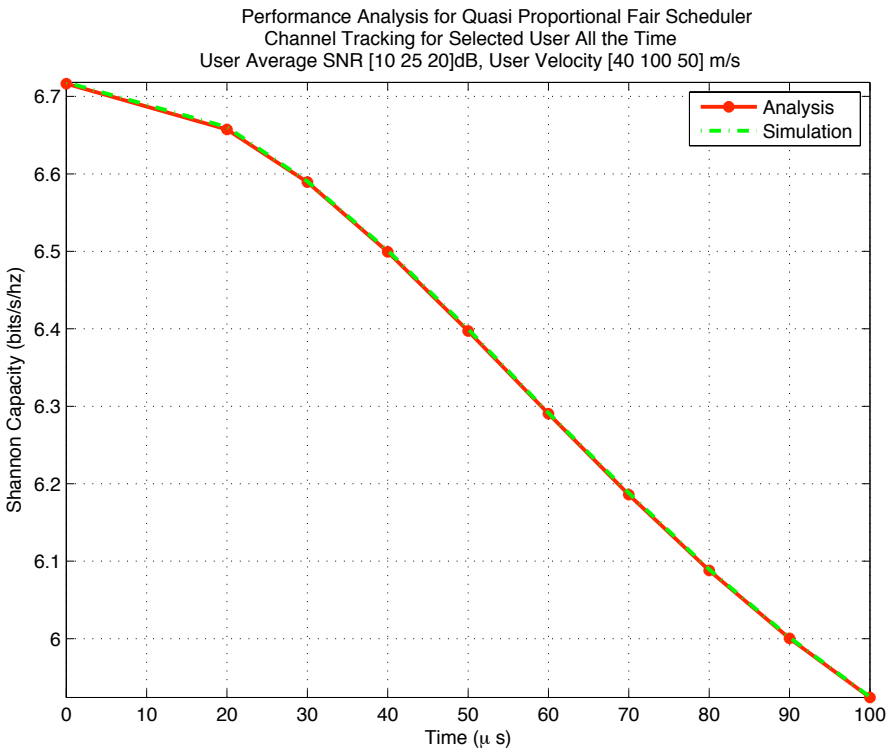



**Figure 5.3:** Shannon capacity of the Quasi-PF scheduler. Only the scheduled user at $t = 0$ has access to perfect CSI over $t \in [0, T]$, where $T$ is the scheduling period.

expenditure at the mobile UT side. A more practical scenario as discussed before is to have the BTS make scheduling decisions according to a one-shot collection of user SNRs, and give the scheduled user access for a certain fixed amount of time. In this case, we assume each UT keeps track of its own channel variation for the purpose of decoding the information message. As pointed out in Section 5.2.2, the post-processing SNR at each scheduled UT is distributed as a non-central chi-square with the parameters given in (5.6) and (5.7). For simplicity of notation, we assume the second order statistics of the channel do not vary with time, and the transmit power remains constant. Therefore (5.6) can be simplified as:

$$s_j^2(t) = |\rho_j(t)|^2 \cdot \eta_j(0)\,, \tag{5.17}$$

where $\eta_j(0)$ is the instantaneous SNR value for the scheduled user at the initial time. Using the same method presented in Section 5.2.2, and the Marcum Q function

expansion in [134], after some mathematical manipulations (see Appendix B.5), the Shannon capacity of this system setup can be found as:

$$C(t) = \log_2 e \sum_{j=1}^{K} \sum_{l=0}^{K-1} (-1)^l \begin{pmatrix} K-1 \\ l \end{pmatrix} \times \sum_{n=0}^{\infty} \frac{\left(\frac{|\rho_j|^2(t)}{1-|\rho_j|^2(t)}\right)^n}{\left(\frac{|\rho_j|^2(t)}{1-|\rho_j|^2(t)} + l + 1\right)^{n+1}} \sum_{m=0}^{n} F\left(\frac{1}{(1-|\rho_j|^2(t)) \cdot \overline{\gamma_j(0)}}, m\right), \quad (5.18)$$

where the function $F(\cdot,\cdot)$ is defined in Section 5.2.2. A simulation with exactly the same setup as Section 5.2.2 is shown in Fig. 5.3. The perfect agreement between the two different curves supports the validity of the derivation. However, the implementation of the analysis involves the evaluation of the incomplete gamma function. When the channel correlation is very high, e.g. $\rho_j(t) = 0.999$, numerical implementation of (5.18) will require a large number of terms in the summation before convergence is reached, and numerical issues with incomplete gamma function will lead to inaccurate results. More research can be done to pursue more robust numerical implementation methods to evaluate system capacity accurately even for high channel correlation.

### 5.3.3 Performance without Channel Tracking

When the UT does not track its time varying channel and just assumes the initially estimated channel for purpose of decoding, an analysis similar to that in Section 5.2.3 shows this lack of channel knowledge at the receiver side will lead to an SNR loss as described in (5.11). By plugging equation (5.14) in, the capacity of the scheduler can be easily derived as (see Appendix B.6):

$$C(t) = \frac{\log_2 e}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} (-1)^{j+1} \cdot \begin{pmatrix} K \\ j \end{pmatrix} \cdot e^{\frac{j}{\alpha_i \cdot \overline{\gamma_i(t)}}} \cdot \mathcal{E}_1\left(\frac{j}{\alpha_i \cdot \overline{\gamma_i(t)}}\right), \quad (5.19)$$

where $\alpha_i \equiv \frac{|\rho_i(t)|^2}{(1-|\rho_i(t)|^2)\cdot\overline{\gamma_i(t)}+1}$. A simulation with the same setup is shown in Fig. 5.4. As expected, the system performance degrades largely due to the lack of accurate channel information for decoding.

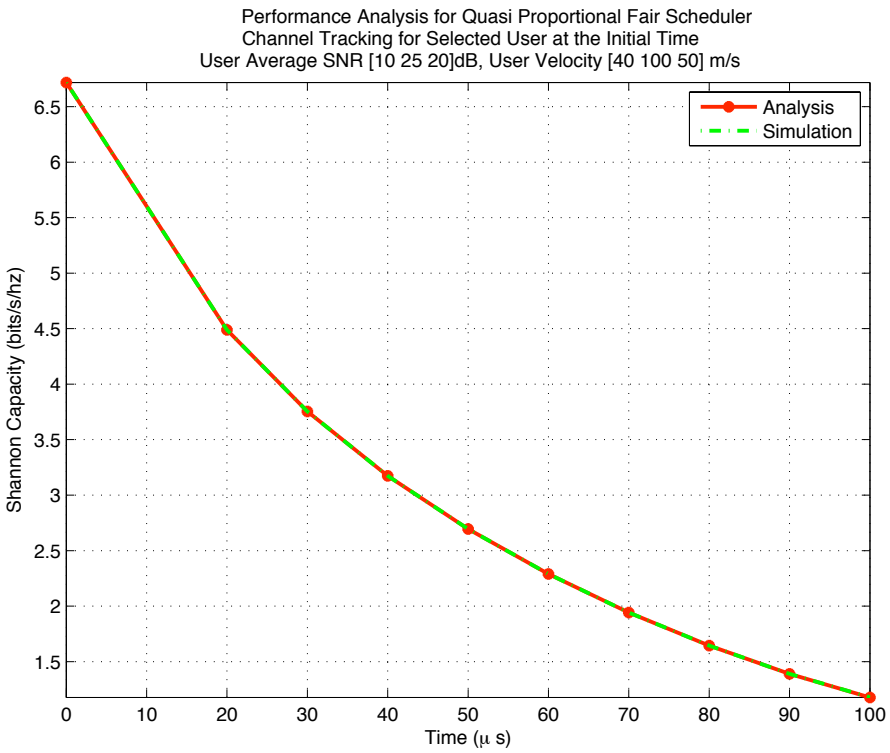



**Figure 5.4:** Shannon capacity of the Quasi-PF scheduler. Channel knowledge is assumed for the selected user only at the scheduling time.

### 5.3.4 Performance Comparison Between Quasi-PF Scheduler and Max-SNR Scheduler

A comparison between the Max-SNR and the Quasi-PF schedulers in terms of capacity and fairness is shown in Fig. 5.5 and Fig. 5.6. Perfect channel tracking is assumed at the UT side and the BTS only has CSI for all users at the initial time. As can be clearly seen in Fig. 5.5, the Max-SNR scheduler outperforms the Quasi-PF scheduler in the sense of maximizing capacity. However more work is needed to determine which scheduler's capacity performance has a faster degradation effect with respect to time, as the Quasi-PF tends to schedule all users equally. The distribution of the Doppler frequency throughout the whole network needs to be taken into account to analyze the aforementioned performance degradation problem. Fig. 5.6 shows that the Quasi-PF scheduler schedules users with equal probability, while the Max-SNR

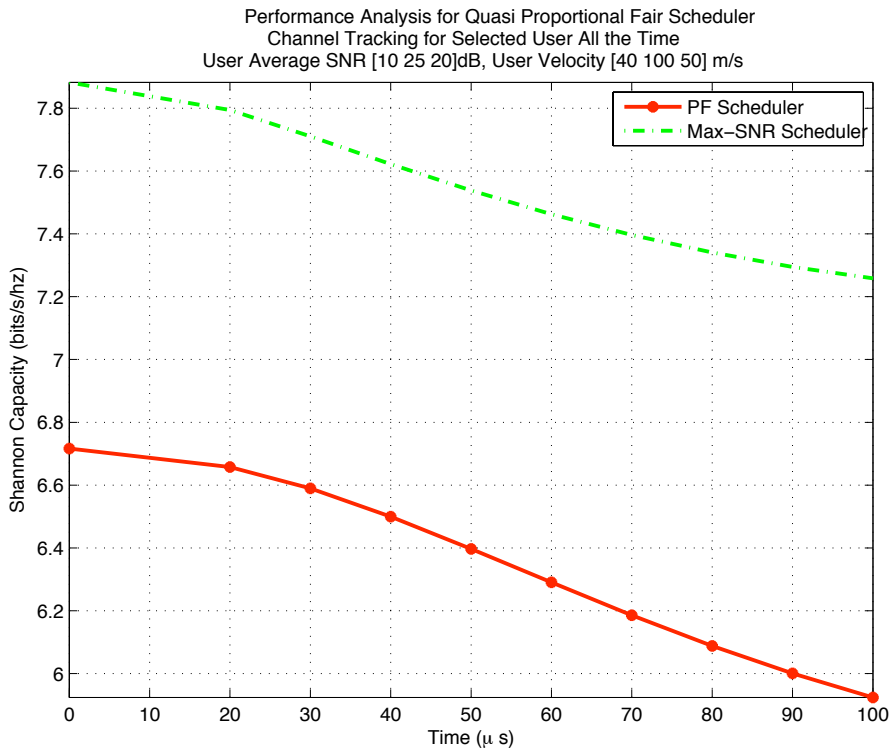



**Figure 5.5:** Shannon capacity of Quasi-PF and Max-SNR scheduler. Only the scheduled user at $t = 0$ has access to perfect CSI over $t \in [0, T]$.

scheduler is highly biased toward the user with highest average SNR. The probability of each user being scheduled in a Max-SNR scheduler setup is given in (7.2).

## 5.4 SCA Scheduler

As seen from Fig. 5.2, when the strongest user (in terms of average SNR) in the network has a large Doppler spread, the Shannon capacity of the Max-SNR scheduler decays quickly with time. We conjecture that for a small network, where the number of asymmetric users is not large, scheduling the strongest user without tracking its channel for a period of time may not provide better throughput than simply selecting a user with a similar average SNR, but with a smaller Doppler spread. This is attributed to the observation that high mobility will outdate the channel knowledge very quickly. Under these conditions, Doppler spread as well as average SNR should be key factors to be considered for making scheduling decisions. Here, we consider a

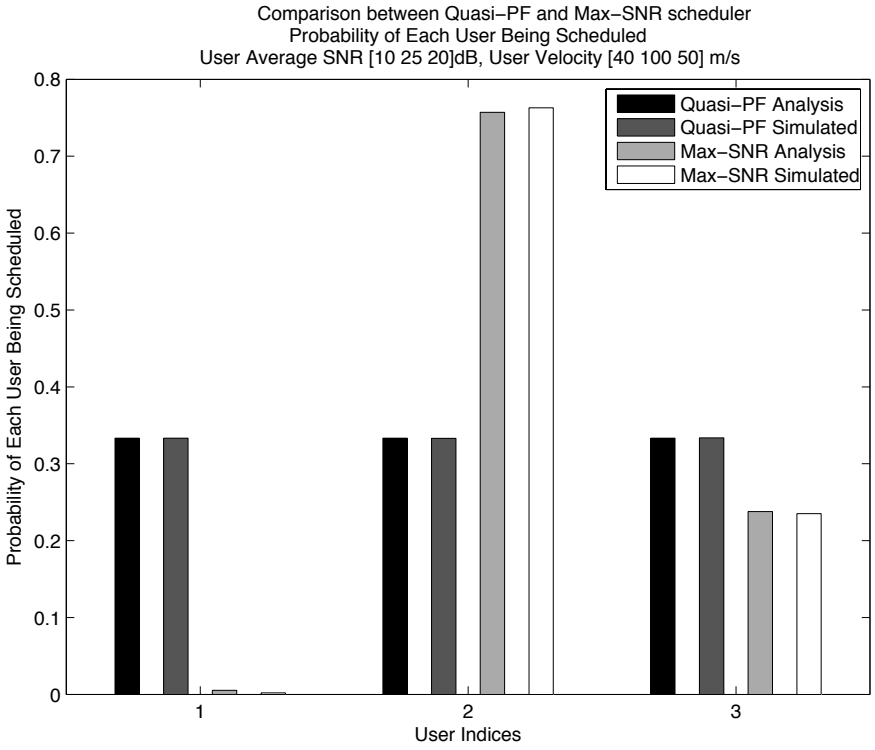



**Figure 5.6:** Probability of each user being scheduled for Quasi-PF and Max-SNR scheduler. Only the scheduled user at $t = 0$ has access to perfect CSI over $t \in [0, T]$.

simple SCA scheduler that requires only the average SNRs and the Doppler spreads as inputs. By taking into account both users instantaneous CSI and long term rate of channel de-correlation, the SCA scheduler can select user according to a more practical criterion which is based on each user's Shannon capacity averaged over the scheduled time interval. As motivation, consider a scheduler that selects a user, say $j$, and allocates time slot $T$ to it regardless of how good its channel is when compared to the other users. The selected user's Shannon capacity is given by (see B.4):

$$\begin{aligned} C_j(t) &= \mathsf{E}\left(\log_2\left(1+\phi_j(t)\eta_j(0)\right)\right) \\ &= \log_2 e \times e^{\frac{(1-|\rho_j(t)|^2)\overline{\gamma_j(t)}+1}{|\rho_j(t)|^2\overline{\gamma_j(t)}}} \cdot \mathcal{E}_1\left(\frac{(1-|\rho_j(t)|^2)\overline{\gamma_j(t)}+1}{|\rho_j(t)|^2\overline{\gamma_j(t)}}\right). \end{aligned} \tag{5.20}$$

In Fig. 5.7, we plot the Shannon capacity of the Max-SNR scheduler and this simple scheduler according to (5.13) and (5.20), respectively. From Fig. 5.7, we

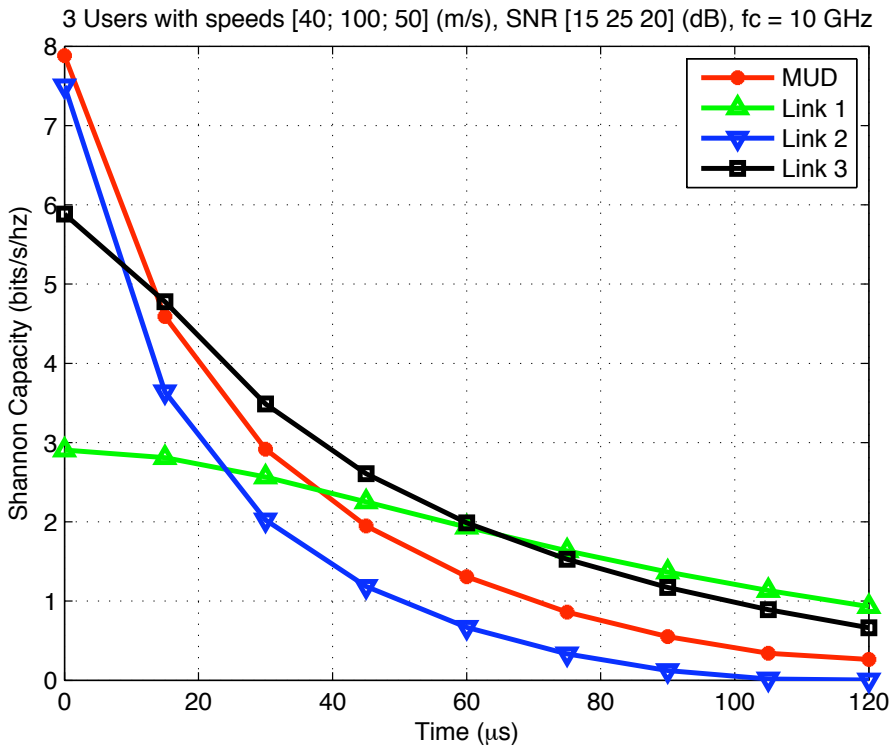



**Figure 5.7:** Shannon capacity of multiuser schedulers. Channel knowledge is known for the selected user only at the scheduling time.

conclude that: a) selecting the strongest at user $t = 0$, and scheduling it for $T = 120\mu s$ without tracking its channel is not as good as simply selecting user '3' for the whole time duration, and b) the gain of MUD over selecting the user with the largest average SNR (user '2') is marginal. We attribute these observations to two reasons: 1) the number of users in the network is not large, hence the possible MUD gain is not large to begin with, and 2) due to the asymmetric network setup, with high probability, the Max-SNR scheduler selects user '2' who has the strongest average SNR. However, due to high Doppler spread, the channel for user '2' de-correlates much quicker than the rest of the users. Therefore, in a small network with some high mobility users, MUD without CSI for the whole scheduling time may lead to suboptimal performance.

To illustrate that a Max-SNR scheduler with insufficient CSI is not always optimal, Shannon capacity for a Max-SNR scheduler in (5.13) and that for each individual user without a scheduler in (5.20) are both averaged over $[0,\ T]$, and

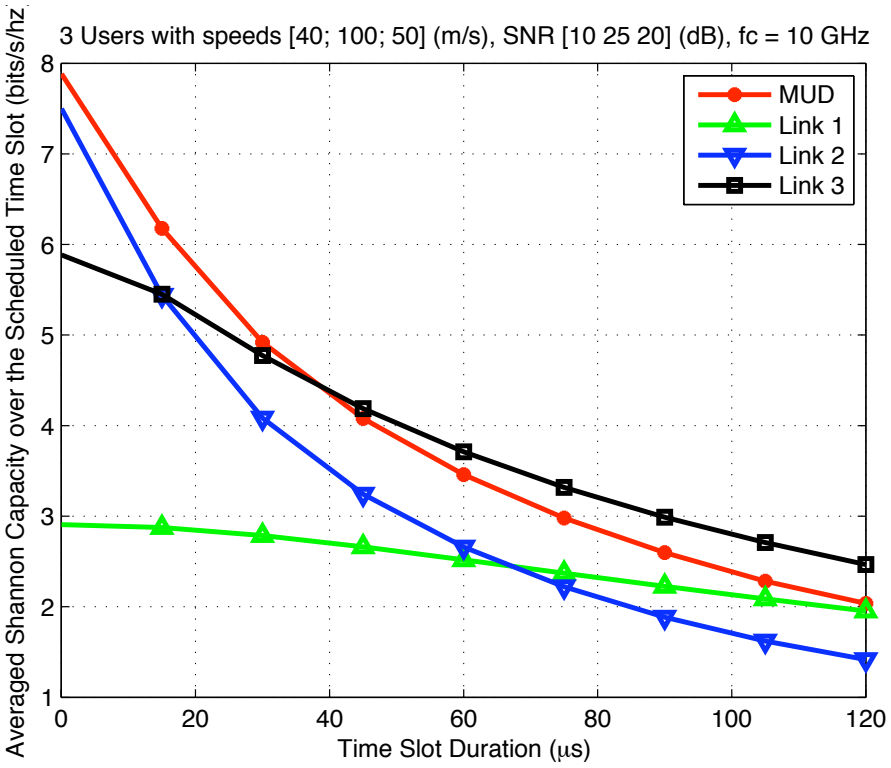



**Figure 5.8:** Average Shannon capacity of Max-SNR scheduler, and the capacity of a given user link.

used as the metrics for user selection. Fig. 5.8 plots such metrics mentioned above as a function of scheduled time slot duration. It is clear from Fig. 5.8 that there are crossover points for different curves in the figure. Therefore, besides the users' average SNRs and Doppler spreads, the slot duration is also an important parameter to determine the optimal scheduler type (Max-SNR versus SCA), and user selection. When $T = 120\mu s$, user 3 is the optimal user to be scheduled. However, it neither corresponds to the user with the largest average SNR nor corresponds to the user with the smallest Doppler spread. When $T \lesssim 40\mu s$, the Max-SNR scheduler that extracts the MUD gain would be the right choice. For the rest of this paper, we dedicate our attention to proposing a SCA scheduler where the MUD gain is only marginal to the system. However, which user to schedule, and for how much time, will be the focus of the rest of this section.

### 5.4.1 Scheduling a Single User

In this section, we study the type of SCA which schedules only one user for the whole period $T$. We assume the selected user only estimates its own channel at time $t = 0$ at the receiver side for the purpose of decoding the message. At the scheduler side, no CSI is required to determine the prospective user. Only average SNR and Doppler spread information is needed to make the scheduling decision. As explained in the last paragraph, the optimal scheduling strategy is a function of $T$. To determine which user is the optimal user to be scheduled, integration over $[0,\ T]$ is involved. When the scheduled time duration $T$ is smaller than the minimum of all the intersecting points at which the capacity curve (averaged over $[0,\ T]$, i.e. Fig.5.8) of the user with the largest average SNR crosses the remaining curves, the user with the largest average SNR is the optimal user to schedule. In this case, an SCA scheduler bases its decision only on the users' average SNR. For example, in Fig. 5.8, if $T$ is less than the intersecting point of the blue curve and the black curve, scheduling user '2' for the whole period $T$ is optimal.

Determining the intersecting points in Fig. 5.8 is difficult since integration is involved. As can be proved, the intersecting point in Fig. 5.7 always lower bounds the desired one. Therefore, we propose to determine the intersecting point in Fig. 5.7. When $T < T_c^j$, we can write $2\pi f_d^j T < 4$, and, from [135], we use the polynomial approximation of Bessel functions: $\mathcal{J}_0(x) \approx \sum_{m=0}^{n} C_{nm} x^{2m}$, with $n = 2$ and $C_{nm} = \frac{(-1)^m n^{1-2m}(n+m-1)!}{2^{2m}(n-m)!(m!)^2}$. Since each term in (5.20) has the form $e^x \mathcal{E}_1(x)$, we can prove that the latter is monotonically decreasing in $x$ (see B.3). It can be shown that the following holds true when the capacity curves of users $i$ and $j$ intersect:

$$\frac{|\rho_j(t)|^2}{|\rho_i(t)|^2} = \frac{\overline{\gamma_i(t)}(\overline{\gamma_j(t)} + 1)}{\overline{\gamma_j(t)}(\overline{\gamma_i(t)} + 1)} \triangleq \epsilon_{ji}. \tag{5.21}$$

Using polynomial approximation for $\rho_j(t) = \mathcal{J}_0(2\pi f_d^j T)$, we have

$$\left( \frac{\sum_{m=0}^{n} C_{nm} x_j^{2m}}{\sum_{m=0}^{n} C_{nm} x_i^{2m}} \right)^2 = \epsilon_{ji}, \tag{5.22}$$

where $x_i = 2\pi f_d^i T$. Upon setting $x_j/x_i = f_d^j/f_d^i \triangleq k_{ji}$, an approximation to (5.22) with $n = 2$ is given by

$$\sum_{m=1}^{2} C_{2m}(k_{ji}^{2m} - \sqrt{\epsilon_{ji}})x_i^{2m} + C_{20}(1 - \sqrt{\epsilon_{ji}}) = 0. \tag{5.23}$$

In (5.23), we determine the smaller positive root $x_i$, which can be divided by $2\pi f_d^i$ to obtain $T_{i,j}$, $i \neq j$. If user $v$ is the one with the largest average SNR, we set $T_B = \min(T_{v,1}, \cdots, T_{v,v-1}, T_{v,v+1}, \cdots T_{v,K})$. If $T \leq T_B$, it is guaranteed that scheduling user $v$ for the whole period $T$ will be optimal.

### 5.4.2 Scheduling More Than One User

When the SCA scheduler has the luxury to schedule more than one user during $T$, switching between the users can provide better throughput when compared to the case where only one user is allowed to be scheduled. For simplicity, we assume that the overhead in switching the users is neglected. With all pairs of intersecting points $T_{i,j}$, $i \neq j$ available, one sub-optimal SCA scheduler is described as follows:

1. Sort the users according to their average SNR in descending order. The indices in the sorted set $\mathcal{I} = \{\mathsf{I}_1, \ldots, \mathsf{I}_K\}$ are such that $\overline{\gamma_{\mathsf{I}_1}} \geq \cdots \geq \overline{\gamma_{\mathsf{I}_K}}$

2. Schedule user $\mathsf{I}_1$ for $T_1 = \arg\min_j T_{\mathsf{I}_1,j}$. If $T_1 \geq T$, schedule user $\mathsf{I}_1$ for the entire period $T$

3. For $m$-th scheduled user, when $m < K$, if $\sum_{j=1}^{m-1} T_j \geq T$, the scheduler has finished its job of selecting users. Otherwise, we schedule user $\mathsf{I}_m$ for $T_m = \min\left(\underset{j, j \neq \cup_{l=1}^{m-1} \mathsf{I}_l}{\arg\min}\ T_{\mathsf{I}_m,j},\ T - \sum_{j=1}^{m-1} T_j\right)$

4. For the $K$-th scheduled user, if $\sum_{j=1}^{K-1} T_j < T$, schedule it for $T_K = T - \sum_{j=1}^{K-1} T_j$. Otherwise, the scheduling is done

The capacity achieved by the SCA scheduler is shown in Fig.5.9. This plot consists of piecewise continuous curves that are from different users' capacity curves (shown in Fig. 5.7). Since each user's signal fades independently, when the $j$-th user

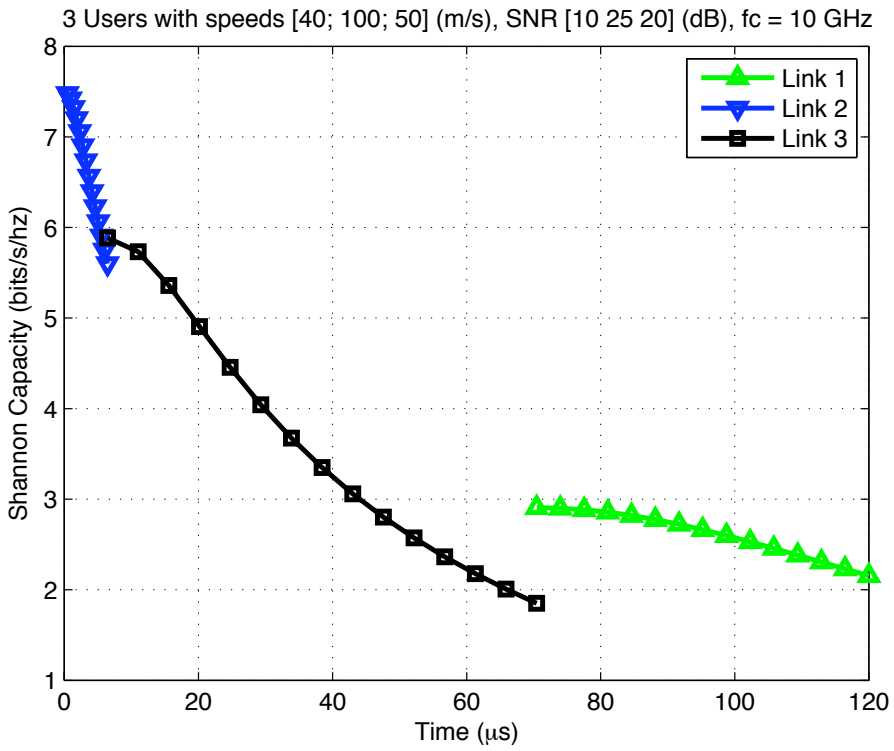



**Figure 5.9:** Shannon capacity of the SCA scheduler. Channel knowledge is known for the selected user only at the scheduling time.

is scheduled, the system's behavior can be captured by the portion of $j$-th user's curve in Fig. 5.7 from $[0,\ T_j]$. The discontinuity is due to switching the users. It is obvious that the SCA scheduler offers better throughput than scheduling only one of the users during $T$. The portion of time allocated to each user may not be optimal in maximizing the total throughput of the network over $[0,\ T]$, because the crossover point by itself is not a sufficient condition for achieving maximum capacity, as will be clarified in the following paragraphs.

When multiple users need to be scheduled (i.e., $T \geq T_B$), to optimize the total network throughput for the whole period of time, the following optimization problem

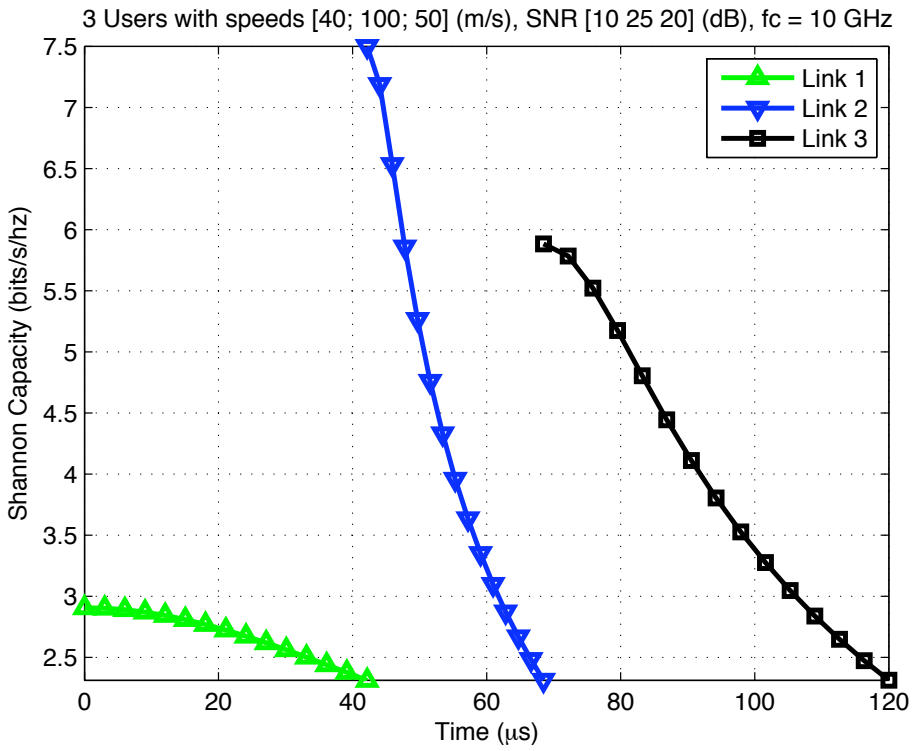



**Figure 5.10:** Shannon capacity of the optimal scheduler. Channel knowledge is known for the selected user only at the scheduling time.

is formulated:

$$[T_1, \cdots, T_K] = \arg\max_{[x_1,\cdots,x_K]} f([x_1, \cdots, x_K])$$
$$\text{s.t.} \sum_{j=1}^{K} x_j = T, \text{ and } x_j \geq 0, \text{ for } 1 \leq j \leq K\,, \tag{5.24}$$

where

$$f([x_1, \cdots, x_K]) = \sum_{j=1}^{K} \int_0^{x_j} \log_2 e \; e^{\frac{(1-|\rho_j(t)|^2)\overline{\gamma_j(t)}+1}{|\rho_j(t)|^2\overline{\gamma_j(t)}}} \mathcal{E}_1\left(\frac{(1-|\rho_j(t)|^2)\overline{\gamma_j(t)}+1}{|\rho_j(t)|^2\overline{\gamma_j(t)}}\right) dt\,. \tag{5.25}$$

The gradient of the above cost function can be evaluated by plugging $x_j$ into (5.20):

$$\frac{\partial f}{\partial x_j} = \log_2 e \; e^{\frac{(1-|\rho_j(x_j)|^2)\overline{\gamma_j(x_j)}+1}{|\rho_j(x_j)|^2\overline{\gamma_j(x_j)}}} \mathcal{E}_1\left(\frac{(1-|\rho_j(x_j)|^2)\overline{\gamma_j(x_j)}+1}{|\rho_j(x_j)|^2\overline{\gamma_j(x_j)}}\right). \tag{5.26}$$

Any gradient based numerical optimization technique can be used to find the optimal time allocation for each user (i.e., $[T_1, \cdots, T_K]$). The Shannon capacity of this optimal scheduler is plotted in Fig. 5.10. A comparison between Fig. 5.10 and Fig. 5.9 shows the improved performance of this optimal scheduler. The averaged Shannon capacity over $[0,\ 120]$ $\mu$s is 3.3592 bits/s/Hz for the scheduler in Fig. 5.9, whereas it is 3.6023 bits/s/Hz for the one in Fig. 5.10. The 0.2431 bits/s/Hz spectral efficiency improvement (7.24%) in this case can be a great performance improvement when the bandwidth of system under consideration is large. Under that assumption, the complexity induced by the SCA scheduler can be justified.

## 5.5 Summary

In this chapter, we investigated the impact of user mobility and asymmetry on multiuser scheduler performance. Closed-form expressions were derived for the Max-SNR scheduler performance under various assumptions on the level of CSI. A scheduler that selects the strongest normalized instantaneous SNR among the users to fulfill the fairness of the scheduling is also analyzed. Our simulations showed that channel-aware multiuser scheduling is not always optimal for a small network with a large Doppler spread. Over such networks, we proposed a simple SCA scheduler that achieves significant improvements to the system throughput.

# Chapter 6

# Relay Communications with Unmanned Aerial Vehicles: Performance and Optimization

## 6.1 Introduction

Recently, UAVs have attracted considerable attention in many military as well as civilian applications [31, 34, 39, 40, 44]. An attractive feature of using UAVs for networked communications is that they can be quickly deployed as relays to extend coverage and improve network connectivity [43, 45, 47, 136, 137]. Employing UAVs in this manner is especially helpful in situations where nodes are widely scattered or obstacles such as hills or large buildings deteriorate the quality of the link between a BTS and an Access Point (AP). The advantages of using relays in more generic wireless network scenarios have been the subject of considerable interest recently (*e.g.,* see [138, 139]).

Ayyagari in [39] presented a network architecture that deployed airborne unmanned relay platforms to form equivalent "cellular towers" in the sky for implementing rapidly deployable, broadband wireless networks. In [34, 40], the authors are concerned with the routing issues of a hierarchical network with UAV nodes relaying messages at higher levels. Hierarchical State Routing (HSR) algorithms are modified to reduce routing overhead and improve the throughput. Rubin proposed a Mobile Backbone Network Protocol (MBNP) that synthesized the topology of the backbone network, which made use of unmanned vehicles including UAVs [44] and dealt with the routing and resource allocation problems for a mobile backbone network. [45] studied the UAV placement and navigation problem with the end goal of improving network connectivity by mimicking the flocking rules that aerial living beings follow. Using graph theory, [136] approached a similar problem by optimizing various connectivity

criteria. [46] investigated the feasibility of using OFDM transmission techniques for UAV wireless communication. Like [46], Palat in [47] focused on the physical layer aspects of UAV relay communications, and studied the performance of distributed transmit beamforming and distributed Orthogonal Space Time Block Coding (OSTBC) schemes under the ideal and non-ideal UAV flight conditions. [43] considered a special scenario of relay communications for delay-tolerant applications, where the UAV relays carry data and deliver them upon approaching the user terminals.

Inspired by the previous work on UAV communications, this chapter investigates a network with multiple UAVs relaying messages from the ground AP to a remote BTS. Unlike [34, 40, 44], we are not concerned with routing algorithms. Instead, various aspects of the network are studied, including: the physical layer communication link properties, i.e. Ergodic Normalized Transmission Rate (ENTR, or spectral efficiency of the transmission scheme) and Symbol Error Rate (SER) analysis for each AP-to-UAV link, the MAC layer handoff algorithm that the APs use to switch between different UAV relays for better performance as the network evolves over time, and the network layer UAV relay deployment problems (placement and optimal motion control). This chapter differs from the previously cited literature in the assumptions made about the network and the criterion used for optimization. In this chapter, we consider a tactical communication scenario, where a set of obstructed APs in a remote area are trying to communicate with a BTS, and teams of UAVs are deployed to help setup the communication links. Relatively abundant bandwidth resources on the UAVs are assumed. Due to the LOS propagation environment between UAVs and between the UAV and the BTS, an error-free link is assumed between them. We focus our investigation on the sum uplink data rate from all the APs to the UAVs, which is more likely to be sensitive to UAV positioning than the downlink from the UAV to the BTS. In particular, we investigate how to find the headings for the UAV relays so that the overall transmission rate is maximized under the constraint that the data rate for each AP is above a certain threshold. The mobility of the APs and relays changes the topology of the network. The varying link strength suggests that the APs may need to be switched to other relays for better transmission. As

will be explained later, this requires a handoff algorithm for APs due to the special motion constraints of the relay. When the current UAV configuration is insufficient to accommodate all the APs on the ground, one must determine where to deploy a new UAV relay and how to command its motion pattern to achieve some desired goal. This problem will also be addressed in this chapter.

The chapter is organized as follows. Section 6.2 describes the mathematical models assumed in this work, including the channel model, and the modulation and coding schemes employed. In Section 6.3, we derive a closed-form expression for the average uplink data rate, and analyze the symbol error rate for each AP-UAV link. We also formulate the optimization problem to find the optimal heading of the UAV for the network. Section 6.4 improves the network throughput by allowing APs to switch relays when necessary. A handoff algorithm is studied for this specific network. In Section 6.5, we study the case where new UAVs need to be added to the network, and develop an approach for determining their positions, headings, and AP assignments. Section 6.6 presents some simulation results for the network protocol we briefly propose. Section 6.7 concludes the chapter and gives some insights into possible future work. Some of the critical derivations can be found in Appendix C.

## 6.2 System Description

### 6.2.1 System Model

We assume a multi-user uplink scenario, with a UAV deployed to relay messages from a collection of APs to a given BTS. We further assume that the APs employ orthogonal transmissions, so that inter-user interference can be ignored. We consider a general multi-antenna setting, where each AP has $M$ transmit antennas, and the UAV relay has $N$ receive antennas. With these assumptions, the signal from the $k$th AP to the UAV can be written as

$$\mathbf{y}_k = \sqrt{\frac{E_k}{M}}\mathbf{H}_k\mathbf{s}_k + \mathbf{n}_k, \tag{6.1}$$

where $\mathbf{y}_k$ is the received signal at the UAV from AP $k$, $\mathbf{H}_k$ is the $N \times M$ channel matrix between AP $k$ and the UAV, $\mathbf{s}_k$ is the transmitted signal from AP $k$, $E_k$ represents the transmit power for AP $k$, and $\mathbf{n}_k$ is additive temporally and spatially white noise.

### 6.2.2 Channel Model

We assume Rayleigh fading channels $\mathbf{H}_k$ with large scale path loss related to the distance between each AP and the UAV, *i.e.,*

$$\mathbf{H}_k = \frac{\mathbf{H}^k_{\text{norm}}}{d_k^{\alpha_k}}, \tag{6.2}$$

where $d_k$ is the distance between the $k$th AP and the UAV, and $\mathbf{H}^k_{\text{norm}}$ is a normalized complex Gaussian matrix which when stacked in an $NM \times 1$ vector has the distribution $\mathcal{CN}(\mathbf{0}, \mathbf{R}^k_H)$. For free space transmission, the path-loss exponent $\alpha_k$ is unity. [2]. Values of $\alpha_k > 1$ occur in obstructed environments, while $\alpha_k < 1$ is common in wave-guided environments. Note that log-normal shadow fading can easily be incorporated into the channel model and the analysis below. Assume the coordinates of the $k$th AP and UAV are given as $[x_k\ y_k\ h_k]^T$ and $[x_u\ y_u\ h_u]^T$ respectively, so that $d_k$ can be calculated as

$$d_k = \sqrt{(x_u - x_k)^2 + (y_u - y_k)^2 + (h_u - h_k)^2}. \tag{6.3}$$

We use the well-known Kronecker model [140, 141] to describe the correlation matrix $\mathbf{R}^k_H$ of the MIMO wireless channel, *i.e.,* $\mathbf{R}^k_H = \mathbf{R}^k_{Tx} \otimes \mathbf{R}^k_{Rx}$, where $\mathbf{R}^k_{Tx}$ are $\mathbf{R}^k_{Rx}$ are the normalized transmit and receive channel correlation matrices for the link between the $k$th AP and the UAV, respectively. When the APs are located in multipath scattering environments, we would expect low spatial correlation at the AP side. At the UAV side, however, high spatial correlation is expected since there are few if any scatterers close to the airborne UAV. The normalized channel matrix can be expressed as

$$\mathbf{H}^k_{\text{norm}} = (\mathbf{R}^k_{Rx})^{1/2}\mathbf{G}[(\mathbf{R}^k_{Tx})^{1/2}]^T, \tag{6.4}$$

where the $N \times M$ matrix $\mathbf{G}$ contains independent and identically distributed (IID) $\mathcal{CN}(0,1)$ elements, $(\cdot)^T$ denotes transpose, $(\cdot)^{1/2}$ is defined such that $\mathbf{R}^{1/2}(\mathbf{R}^{1/2})^H = \mathbf{R}$, and $(\cdot)^H$ is the Hermitian transpose.

### 6.2.3 Adaptive Modulation

We assume that the system employs adaptive modulation based on the current channel SNR for each link, denoted by $\gamma_k$. For a given desired SER, the required SNR thresholds are predetermined using the SER expression given in [59, 7]:

$$P_e \approx \overline{N}_e Q\left(\sqrt{\frac{\gamma_k d_{\min}^2}{2}}\right), \tag{6.5}$$

where $P_e$ is the symbol error probability, $\overline{N}_e$ is the number of nearest neighbor constellation points, and $d_{\min}$ is the minimum separation distance between points in the underlying constellation. Assume that $\gamma^i$ and $\gamma^{i+1}$ are the predetermined SNR thresholds for the $i$th and $(i+1)$th modulation schemes respectively. If $\gamma^{i+1} > \gamma_k \geq \gamma^i$, the $i$th modulation scheme will be used to transmit the message. If $\gamma_k < \gamma^1$, no transmit scheme will be chosen, which indicates there will be no transmission between the transmitter and the receiver.

### 6.2.4 Orthogonal Space-time Block Coding

We assume that only the receiver knows the channel matrix. Hence, orthogonal space-time block codes (OSTBC) [118] are used to transmit the data. For example, in the $2 \times 2$ case, the well-known Alamouti code [5] is employed. Since adaptive modulation is used, the receiver needs to determine/predict a suitable modulation scheme and feed this information back to the transmitter. In this chapter, we assume that this feedback is perfect, *i.e.,* the transmitter knows which modulation scheme to use.

## 6.3 System Analysis

Below, we first investigate the single link SNR and data rate. We then extend the results to the multi-link scenario. Note that we drop the subscript $k$ for separate APs in the single link analysis. The subscript will be reintroduced when multiple links are taken into account.

### 6.3.1 Single Link SNR

It is well known that OSTBC exploits the diversity of the MIMO channels, and the instantaneous uplink SNR at the UAV can be expressed as

$$\begin{aligned}\gamma &= \|\mathbf{H}\|_F^2 \frac{E}{M\sigma_N^2} \\ &= \|\mathbf{H}\|_F^2 \rho \ ,\end{aligned} \tag{6.6}$$

where $\rho$ is defined as $\rho = \frac{E}{M\sigma_N^2}$, $\sigma_N^2$ is the noise power, and $\|\cdot\|_F$ denotes the Frobenius norm. Plugging (6.2) into (6.6), we obtain

$$\gamma = \frac{\|\mathbf{H}_{\text{norm}}\|_F^2}{d^{2\alpha}}\rho. \tag{6.7}$$

In [142], using the inverse Laplace transform, the PDF of $\|\mathbf{H}_{\text{norm}}\|_F^2$ is derived as

$$f(x) = \sum_{j=1}^{P}\sum_{k=1}^{m_j} A_{jk}\frac{x^{k-1}}{(k-1)!\sigma_j^k}e^{-\frac{x}{\sigma_j}}u(x), \tag{6.8}$$

where $\sigma_j$ $(j = 1, 2, \cdots, P)$ are the distinct non-zero eigenvalues of $\mathbf{R}_H$, $m_j$ denotes the multiplicity of $\sigma_j$, and $A_{jk}$ can be determined by solving a system of linear equations [142]. Defining

$$\begin{aligned}g(n, \alpha, x) &= \int \frac{x^n}{n!}e^{\alpha x}dx \\ &= \frac{1}{n!}\frac{e^{\alpha x}}{\alpha^{n+1}}\sum_{i=0}^{n}(-1)^{n-i}\frac{n!}{i!}(\alpha x)^i \ ,\end{aligned} \tag{6.9}$$

the CDF of $\|\mathbf{H}_{\text{norm}}\|_F^2$ can be expressed as:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \sum_{j=1}^{L}\sum_{k=1}^{m_j} \frac{A_{jk}}{\sigma_j^k}[g(k-1, -\frac{1}{\sigma_j}, x) - g(k-1, -\frac{1}{\sigma_j}, 0)]\,. \tag{6.10}$$

### 6.3.2 Single Link ENTR and SER

Due to the random nature of the channel matrices, the instantaneous transmission rate is different for different channel realizations. Therefore, we define the ENTR and use it as the criteria to quantify the performance of the link. The ENTR $R(t)$ is defined as

$$R(t) = \beta \cdot E(\log_2 K(t))\,, \tag{6.11}$$

where

$$K(t) = K^1 u(\gamma(t) - \gamma^1) + \sum_{i=1}^{L-1}(K^{i+1} - K_i)u(\gamma(t) - \gamma^{i+1})\,, \tag{6.12}$$

and where $\beta$ is a scalar that takes into account the rate loss when OSTBC is used, and $u(\cdot)$ is the unit step funtion. Note that for $2 \times 2$ Alamouti coding, $\beta = 1$. In (6.12), $K^i$ is the number of constellation points for the $i$th modulation scheme, and $L$ is the total number of modulation schemes used in the system. Defining $C^i(t) = \frac{\gamma^i}{\rho} d^{2\alpha}(t)$, it is straightforward to show that the expression for the ETNR of the AP-UAV link (i.e. the uplink between an AP and UAV relay) can be written as

$$\begin{aligned} R(t) &= \beta \cdot \{\sum_{i=1}^{L-1} \log_2 K^i \int_{C^i(t)}^{C^{i+1}(t)} f(x)dx + \log_2 K^L \int_{C^L(t)}^{\infty} f(x)dx\} \\ &= \beta \cdot \{\sum_{i=1}^{L-1} \log_2 K^i [F(C^{i+1}(t)) - F(C^i(t))] + \log_2 K^L [1 - F(C^L(t))]\}. \end{aligned} \tag{6.13}$$

To verify the above analysis, we simulate a case where the AP and UAV are separated by a distance of 3640m, and both have two antennas. The AP transmit power is 2 W, and the noise power spectral density at the UAV relay is $10^{-16}$ W/Hz. The path-loss exponent $\alpha$ is assumed to be 1.5, the carrier frequency is 1 GHz, and the system bandwidth is 20 kHz. Seven different MPSK modulation schemes are used in the simulations, i.e. from BPSK to 128-PSK. We assume a rich scattering

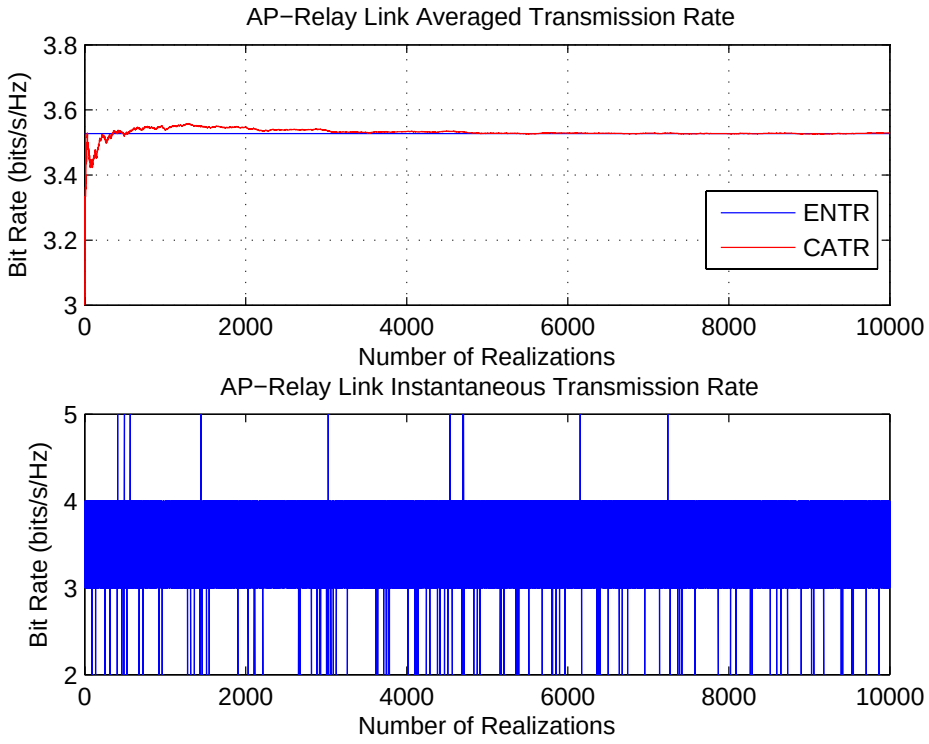



**Figure 6.1:** Link rate simulation results.

environment at the AP side, so that the correlation matrix at the AP side is given by

$$\mathbf{R}_{Tx} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

At the UAV side, high spatial correlation is assumed:

$$\mathbf{R}_{Rx} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

$10^5$ channel realizations were run to generate the plot in Fig. 6.1. The upper plot shows the ENTR and the Calculated Averaged Transmission Rate (CATR) defined as $\frac{\sum_{i=1}^{N} S^i}{N}$, where $N$ is the number of the channel realizations, and $S^i$ is the instantaneous spectral efficiency of the $i$-th channel realization. Clearly, the CATR quickly converges

to the ENTR expression, which verifies our derivation. The lower plot shows the instantaneous transmission rate of the link.

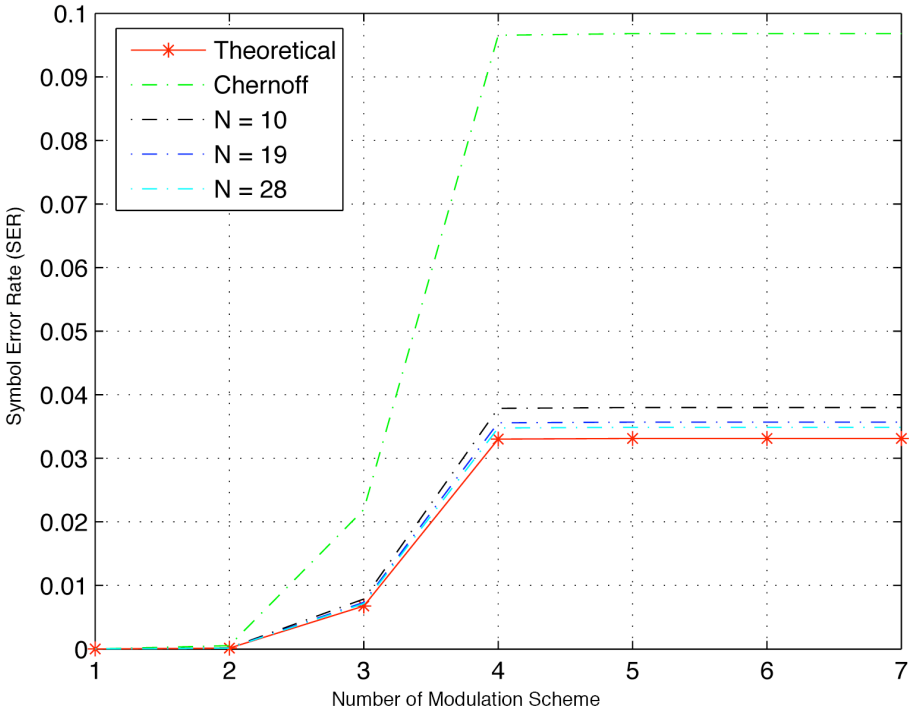



**Figure 6.2:** Upper bound on the symbol error rate for each AP-UAV communication link.

A closed-form expression for the single link SER has also been derived in [137]:

$$P_s = \frac{1}{\pi}\{\sum_{i=1}^{N-1}\int_0^{\frac{\pi}{2}}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\frac{\overline{N}_e(i)A_{jk}}{\sigma_j^k}[g(k-1,-(\frac{\rho d_{\min}^2(i)}{4\sin^2\theta}+\frac{1}{\sigma_j}),C_{i+1}(t))$$
$$-g(k-1,-(\frac{\rho d_{\min}^2(i)}{4\sin^2\theta}+\frac{1}{\sigma_j}),C_i(t))]d\theta$$
$$-\int_0^{\frac{\pi}{2}}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\frac{\overline{N}_e(N)A_{jk}}{\sigma_j^k}g(k-1,-(\frac{\rho d_{\min}^2(N)}{4\sin^2\theta}+\frac{1}{\sigma_j}),C_N(t))d\theta\}\,. \quad (6.14)$$

The complexity in integrating the SER expression (6.14) can be reduced by resorting to the evaluation of SER bounds given in (C.7). Fig. 6.2 shows that the analytical

expression derived in (6.14) agrees very well with our simulation results and illustrates the upper bound obtained above tightly bounds the SER expression. The derivation of these equations can be found in Appendix C.1.

### 6.3.3 Heading Optimization in the Multi-link Scenario

Using a constant speed model, the UAV dynamics are governed by

$$\begin{aligned} \mathbf{x}_u^t &= \mathbf{x}_u^{t-1} + Vcos(\delta_{t-1})\Delta \\ \mathbf{y}_u^t &= \mathbf{y}_u^{t-1} + Vsin(\delta_{t-1})\Delta\,, \end{aligned} \tag{6.15}$$

where $V$ is the UAV speed, $\delta_{t-1}$ is the UAV heading at time step $t-1$, $\Delta$ is the length of the time step, and where we have added superscripts to $x_u$ and $y_u$ to indicate that the UAV position varies with time. Note that in general the APs are also mobile. We will assume that the UAV can track the AP positions and predict their locations from time step $t-1$ to $t$. The change in distance between the APs and the UAV over one time step can be expressed as a function of the UAV heading $\delta_{t-1}$ by plugging (6.15) into the equation for $d_k$ in (6.3).

The average data rates $R_k(t)$ for each UAV $k$ are a function of $d_k$, and hence a function of the UAV heading as well, and it makes sense to choose the UAV heading that maximizes the overall system date rate, *i.e.,*

$$\begin{aligned} \arg\max_{\delta_t} \;\; & R_T(t) = \sum_{k=1}^{K} R_k(t) \quad s.t. \\ & R_k \geq R_{\min} \\ & |\delta_t - \delta_{t-1}| \leq \Delta\delta\,, \end{aligned} \tag{6.16}$$

where $R_{\min}$ is the minimum data rate requirement for each UAV-AP link, $\Delta\delta$ and defines the maximum turning radius of the UAV in one time step. The first constraint guarantees a minimum level of performance for each AP, assuming that each AP-UAV link uses the same bandwidth. If the bandwidth can be allocated dynamically for different APs, then the total rate would be the weighted sum of each AP-UAV link

data rate. For some scenarios, there is no solution to the above problem. This means one UAV is not enough to provide coverage for the whole system, and additional UAVs are needed in order to achieve the minimum requirements.

The above optimization problem is very complicated, and does not admit a simple solution. A key result of this chapter is derived in the Appendix, where it is shown that under some mild conditions, $R_k(t)$ can be approximated as a sinusoid plus a constant offset:

$$R_k(t) = \beta_k(A_k cos(\delta_t - \theta_k^0) + C_k)\,. \tag{6.17}$$

Expressions for the constants $A_k$ and $C_k$ can be found in the Appendix. Using this approximation, the complexity of the optimization problem is significantly reduced. The total network throughput $R_T(t)$ is thus also approximated as a sinusoid plus a constant offset, and if no constraints were imposed on the UAV turning radius, the optimal UAV heading would be given by

$$\delta_t = \arctan\ \frac{\sum_{k=1}^{K} \beta_k A_k \sin\theta_k^0}{\sum_{k=1}^{K} \beta_k A_k \cos\theta_k^0}\,,$$

as derived in (C.18). To solve the optimization with the heading constraint, we simply compute $\delta_t$ as above, and determine if it falls within the turning radius. If yes, this solution is used as the UAV's heading for the next time interval. If not, the two boundary points are checked, and the one that results in the largest rate is chosen.

To validate our derivation, we simulated a scenario with two APs randomly positioned on the ground within a 2000m-by-2000m square and one UAV located at $[0\ 0\ 3600]^T$ in the air. Most of the simulation parameters are the same as in the previous example, except that the bandwidth of each AP is assumed to be 200 kHz and the update time interval is set to 15s for the purpose of clearly illustrating the idea that smart heading control does affect the system throughput in a significant sense. In order to make the simulation more realistic, we use Lee's channel model described in [143] to generate $\mathbf{R}_{Tx}$ and $\mathbf{R}_{Rx}$. Besides the parameters mentioned above, we set

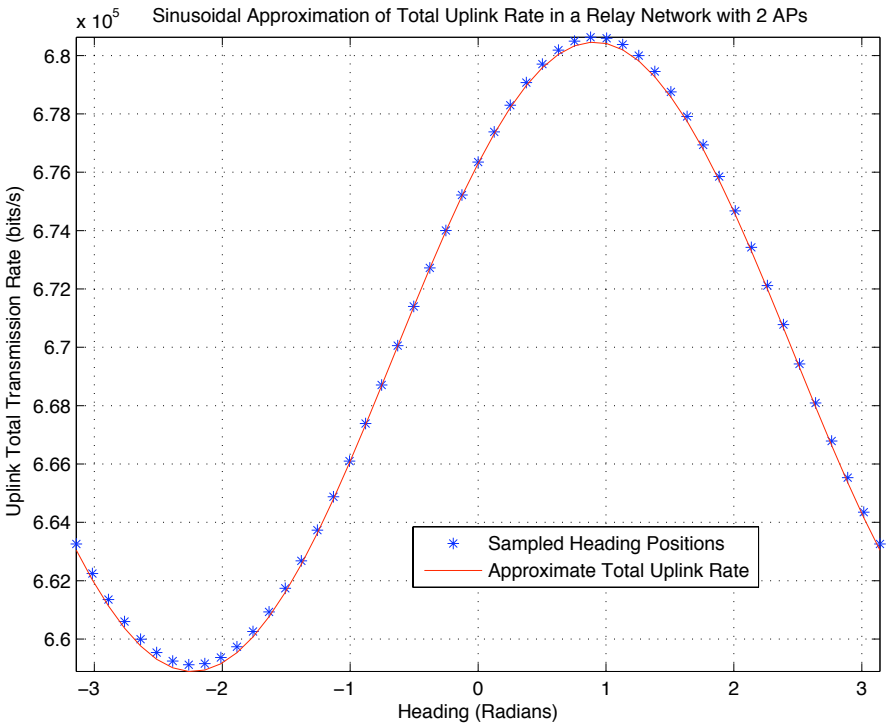



**Figure 6.3:** Sinusoid approximation for the uplink total communication rate of a subnetwork.

the antenna separation at the UAV to be $2\lambda$ and the antenna separation at the APs as $\frac{1}{2}\lambda$, where $\lambda$ is the wavelength of the transmitted EM wave. We also assume that 40 scatterers are uniformly placed on a circle with radius $100\lambda$ around each AP. The simulation results are plotted in Fig. 6.3. It is clear that the total uplink transmission rate is well approximated by the sinusoidal expression derived in the Appendix. The importance of optimizing the UAV's motion can be seen from the 20k bps date rate difference yielded by simply assuming a better heading. This difference by itself is capable of supporting an additional user for voice communication in most commercial standards. The accuracy of the approximation can be further improved when the update time interval is smaller and the conditions stated in the Appendix are better satisfied.

## 6.4 Handoff Algorithm and Optimal Motion Control in the Mobile Relay Assisted Network

Mobility of both the APs and the relays causes the average link SNR strength for each AP-UAV link to vary for every update time interval. As time goes by, the original association of the APs and UAVs may not be optimal any more, and the switching of the APs to a better hosting relay would improve the network throughput. In this section, we will study the AP handoff problem in the context of a mobile-relay-assisted network. $N$ airborne relays are assumed to be in service, each hosting a set of APs with an index set $\mathcal{I}_i,\ 1 \leq i \leq N$, whose elements are the indices to the APs that the $i$-th UAV is offering service to. We suppose there are $L$ total APs requesting service, so that $\cup_{i=1}^{N} \mathcal{I}_i = Q \equiv \{z | z = 1, 2, \cdots, L\}$ and $\cap_{i=1}^{N} \mathcal{I}_i = \emptyset$. In other words, the current $N$ UAVs can host all the $L$ APs for the time interval of interest. If not, new UAVs need to be deployed, and this will be addressed in the following sections. This section will be dedicated to the AP handoff problem in this mobile relay network setup. In addition, more discussion about optimal motion control for the mobile relay will be treated as well.

Various handoff algorithms based on Received Signal Strength (RSS) are discussed in [1]. The basic idea behind the handoff algorithms in a cellular network is that the mobile terminal, the AP in this case, measures the received signal strength from various BTS over a time window, and associates itself with the BTS that provides the strongest link. A similar idea can be used in the handoff algorithm for this mobile UAV assisted network with some modifications. The motion constraint for the UAV relay and minimum rate constraint for each AP in (6.16) complicates the handoff procedure as will be discussed in the rest of the section.

Let us define a subnet as the part of the network, with a UAV as the center node, serving a set of APs in a star topology. In other words, the network under consideration consists of $N$ subnets, and each subnet hosts a set of APs indicated by index set $\mathcal{I}_i,\ 1 \leq i \leq N$. In the above two sections, we have approximated the link communication rate as a sinusoid with some specific offset. A close look at (C.10) reveals that in a single UAV, single AP scenario, the optimal heading for the UAV

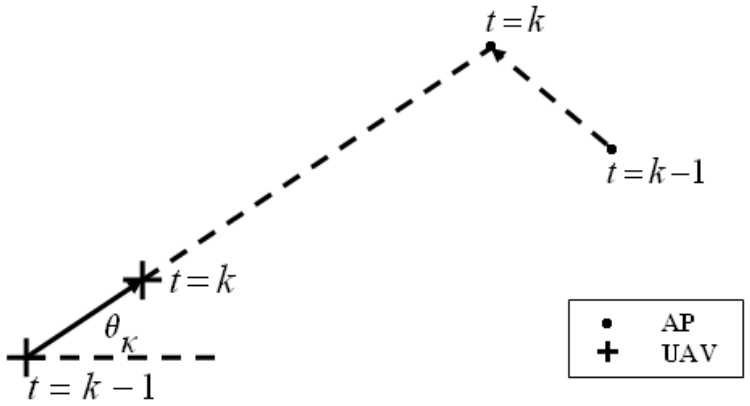



**Figure 6.4:** Optimal UAV heading for single AP-UAV communication link.

is the direction pointing to the position of the AP at the next time step. This is no surprise since it is obvious that in this case, the optimal position of the UAV is directly above the AP. Section 6.3.3 gives an approximate closed-form solution for optimal heading command for each subnet configuration when a certain constraint requirement is met, i.e. when it falls in the area reachable by the UAV and it satisfies the minimum rate constraint. If not, boundary points need to be checked to yield the optimal heading solution. To have a better understanding of the requirements and the so called "boundary", we will introduce a few new concepts as follows. The scope of the following definitions is each subnet. The *link allowable region* for the $i$-th AP in the $j$-th subnet is defined as the heading range $\Omega_j^i$, s.t. $\forall\ \delta \in \Omega_j^i$, $R_j^i \geq R_{\min}$ is satisfied, where $R_j^i$ is the data rate that the $j$-th subnet can provide for the $i$-th AP. According to the sinusoidal approximation in (6.17), we find the link allowable region for the $i$-th AP to be:

$$\Omega_j^i \equiv \begin{cases} [0\ 2\pi] & \beta_i C_i - \beta_i |A_i| \geq R_{\min} \\ \emptyset & \beta_i C_i + \beta_i |A_i| \leq R_{\min} \\ [t_1\ \ t_2] & (t_1 - \theta_j^{0,i}) \cdot (t_2 - \theta_j^{0,i}) < 0 \\ [0\ \ t_1] \cup [t_2\ \ 2\pi] & (t_1 - \theta_j^{0,i}) \cdot (t_2 - \theta_j^{0,i}) > 0 \end{cases}, \tag{6.18}$$

where

$$\theta_j^{0,i} = \begin{cases} \lfloor \arctan\left(\frac{y_u^{k-1}-y_i^k}{x_u^{k-1}-x_i^k}\right) \rfloor_{2\pi} & A_i > 0 \\ \lfloor \pi + \arctan\left(\frac{y_u^{k-1}-y_i^k}{x_u^{k-1}-x_i^k}\right) \rfloor_{2\pi} & A_i < 0 \end{cases}, \tag{6.19}$$

and $t_1$ and $t_2$ are defined as:

$$\begin{aligned} t_1 &= \min\{\psi_j^i,\ 2\pi - \psi_j^i\}, \\ t_2 &= \max\{\psi_j^i,\ 2\pi - \psi_j^i\}, \\ \text{and } \psi_j^i &\equiv \lfloor \arccos \frac{R_{\min} - \beta_i \cdot C_i}{|A_i|} + \theta_j^{0,i} \rfloor_{2\pi}, \end{aligned} \tag{6.20}$$

where we use $\lfloor \cdot \rfloor_{2\pi}$ to denote the mod-$2\pi$ operation, i.e. $\lfloor x \rfloor_{2\pi} \equiv mod(x, 2\pi)$. Therefore solution to the optimization problem is a subset of the intersection of all $\Omega_i$'s for each subnet. The *reachable region* is defined as the set of heading angles that are within the turning radius of the UAV, i.e. the command set that doesn't violate the heading constraint in (6.16). Mathematically, the reachable region is written as the heading constraint set:

$$\mathcal{C}_j \equiv \left\{\delta \middle| \left|\delta - \delta_{k-1}^j\right| \leq \Delta\delta_j\right\}, \tag{6.21}$$

where $\delta_{k-1}^j$ is the previous heading for the $j$-th relay and $\Delta\delta_j$ is determined by the turning radius of the $j$-th relay. The intersection between $\Omega_j^i$ and $\mathcal{C}_j$ defines the *admissible region* for the $i$-th AP with respect to the $j$-th relay:

$$\Xi_j^i \equiv \Omega_j^i \cap \mathcal{C}_j. \tag{6.22}$$

A non-empty admissible region is a necessary but insufficient condition for the $j$-th relay to host the $i$-th AP. For the $j$-th relay to simultaneously support the set $\mathcal{I}_j$, to satisfy both the minimum rate constraint and turning radius constraint, its *feasible region*, defined as the intersections of all the hosted APs' admissible regions:

$$\mathcal{S}_j \equiv \cap_{i=1}^{|\mathcal{I}_j|} \Xi_j^{\mathcal{I}_j(i)}, \tag{6.23}$$

has to be non-empty, where $|\mathcal{I}_j|$ is the cardinality of the set $\mathcal{I}_j$. For a potential entry of AP $p$ into the $q$-th subnet, not only the admissible region $\Xi_q^p$ has to be nonempty, but also it should be compatible with the set of APs that the $q$-th relay is currently hosting, i.e. $\Xi_q^p \cap \mathcal{S}_q \neq \emptyset$. This is a sufficient condition for an AP to register with a potential relay. If we recall the approximate optimal solution given in (C.18), we can conclude that for each subnet $j$ hosting $\mathcal{I}_j$ APs, when the relay's feasible region $\mathcal{S}_j$ is nonempty, and optimal solution falls within the range defined by $\mathcal{S}_j$, the optimal solution is achieved by commanding the $j$-th relay to fly at the angle determined by (C.18). If $\mathcal{S}_j$ is nonempty but the angle given by the aforementioned equation does not fall in the range, the boundary point of $\mathcal{S}_j$ is checked to yield the optimal solution. If $\mathcal{S}_j$ itself is empty, it means the APs in $\mathcal{I}_j$ are not compatible with each other in the $j$-th subnet, either some of the APs have to handoff to other relays currently in service, or new relays have to be deployed to accommodate their communication requirements.

Here we consider the problem introduced at the beginning of the section, the AP handoff problem, under the assumption that the current number of UAVs is sufficient to cover the service requirement. As discussed before, an AP cannot just handoff to any UAV that provides a stronger link, its admissible region with respect to the potential host relay has to be nonempty for it to carry out the handoff procedure. The registration of the AP to a new subnet will change the feasible region of that relay node, thereby affecting the other possible APs' switching over to this specific subnet in the future. Hence, the order in which APs switch themselves in the handoff process will affect network performance. An optimal algorithm that solves this handoff problem, involving a joint optimization over all the subnets, can be

formulated as follows:

$$\arg\max_{\mathcal{I}_1,\cdots,\mathcal{I}_N} \sum_{j=1}^{N}\sum_{i=1}^{|\mathcal{I}_j|} R_j^{\mathcal{I}_j(i)} \qquad s.t.$$
$$\cup_{j=1}^{N}\mathcal{I}_j = \mathcal{Q},$$
$$\cap_{j=1}^{N}\mathcal{I}_j = \emptyset,$$
$$\mathcal{S}_j \neq \emptyset, \qquad \forall j \in [1,\cdots,N]\,. \tag{6.24}$$

Once the AP-relay associations $\mathcal{I}_1,\cdots,\mathcal{I}_N$ are determined, the optimal headings for the relays can be obtained using the method presented above. Obviously, this optimization problem is hard to solve, and an exhaustive search through all the combinations would involve a prohibitive amount of computation, in the worst case $N^L$ possibilities, for each AP has $N$ possible relays as its potential BTS. Some knowledge about the positions of the relays and APs can be used to significantly narrow the search space, but the complexity is still exponential.

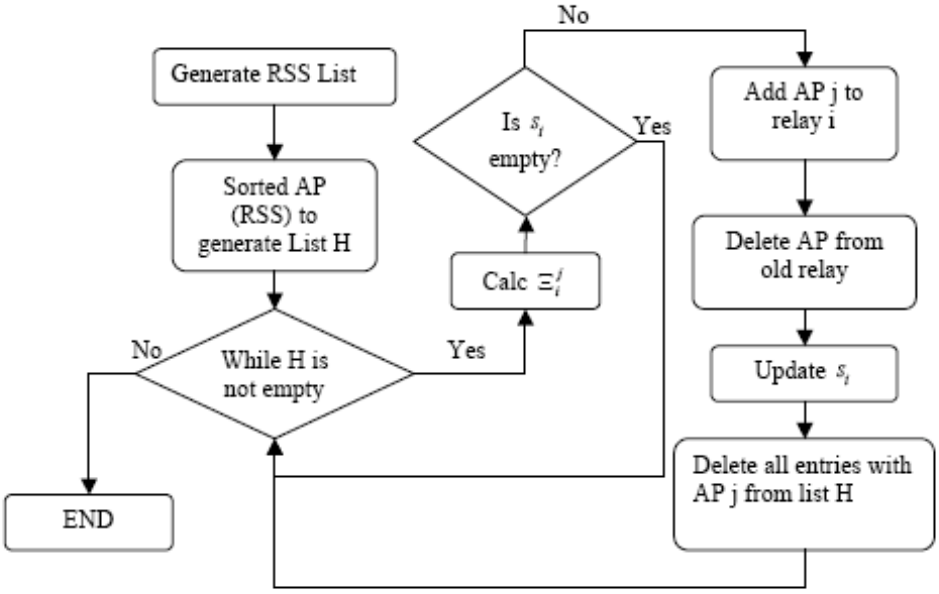



**Figure 6.5:** Flowchart of handoff algorithm for AP "$j$" to hand over to relay "$i$" in the UAV assisted network.

An ad-hoc handoff algorithm with less complexity is presented below and an example will be given to clarify the procedure. Each AP in the network is assumed to continuously monitor the RSS from all the relays. When better link than the current one is detected, a handoff can be initialized either by the AP or the relay that is currently hosting the AP. A list of candidate APs that have the potential to gain better link quality by switching to another relay is created. All pairs of candidate APs and their potential new hosting relays are sorted in order of decreasing RSS, and this list of pairs is denoted as $H$. Entries of $H$ are examined one by one to see if the admissible region of the candidate AP intersects its corresponding potential relay's feasible region. If so, the AP can be handed off to the new relay, the new relay's feasible region gets replaced by the aforementioned intersection, the feasible region of the old relay that hosted the AP is updated, and the rest of the entries in $H$ that are associated with this specific AP are deleted. Otherwise, we delete the entry under inspection, and proceed to the next entry in $H$. This process repeats itself until the list $H$ is empty. The flowchart for the handoff algorithm is shown in Fig. 6.5. An example is given here to better explain the proposed handoff procedure. Here we assume there are 5 airborne relays, and 10 APs requesting service on the ground. At a certain time instant, AP3, AP5 and AP7 are found to benefit from a possible handoff, with candidate relays (R1, R3, R5) for AP3, (R2, R4) for AP5 and (R1, R2, R5) for AP7 respectively. The list $H$ is generated by sorting the RSS of all the possible pairs, and we have, for example:

$$\begin{aligned} H = \{&(R2, AP5), (R1, AP3), (R5, AP3), (R3, AP3), \\ &(R2, AP7), (R4, AP5), (R1, AP7), (R5, AP7)\}. \end{aligned} \tag{6.25}$$

When $H(1)$ is under inspection, the admissible region $\Xi_2^5$ is tested to see if it intersects with the feasible region of Relay 2, i.e. $\mathcal{S}_2$. A positive answer to the above test leads to a smaller $H$:

$$H = \{(R1, AP3), (R5, AP3), (R3, AP3), (R2, AP7), (R1, AP7), (R5, AP7)\}.$$

Now the pair $(R1, AP3)$ is under consideration, if AP3 is not allowed to hand off to Relay 1 due to an empty intersection between $\Xi_1^3$ and $\mathcal{S}_1$, this entry is deleted and $(R5, AP3)$ is the next pair of interest. This process keeps repeating itself until $H$ becomes empty. This ad-hoc algorithm is not optimal in the sense that the order in which the handoff is carried out would possibly preclude other possible handoffs in the waiting list, and therefore affect the total throughput optimization of the network. While other orders of operation are also possible, the above presented order which always values better link quality first is a reasonable solution.

## 6.5 The Deployment of New UAV Relays

The discussion in the above section is based on the assumption that each UAV in the air has a list of its served APs. The problem of how to partition the APs into various subnets (clusters), remains unaddressed. Due to both the mobility of the APs and relays, the signal strength of each link is always changing. At some point of time, it is possible that there are APs that cannot be served by any of the relays currently in the air for the desired link quality. Under this circumstance, additional UAVs need to be put in the air to maintain the communications. However, it remains an open problem regarding where to place the UAV and how many UAVs need to be deployed. This problem will be discussed below.

There are various approaches to solve this problem. A straightforward method can be obtained by posing the UAV deployment and AP assignment problems as an adaptive clustering problem. As [144] pointed out, a pure flat architecture for an ad hoc network suffers from scalability problem, and to guarantee a basic level of performance in terms of a reasonable throughput and delay, a hierarchical architecture is more attractive. The relays in this chapter act as backbones, and the scenario we study assumes a natural hierarchical architecture. Assigning APs to each relay forms a set of subnets (equivalently, clusters) and therefore assigning the APs has a good analogy to the adaptive clustering problem. As is obvious, the hierarchical cluster structure can benefit the system design in the sense that 1) system resources can be spatially reused, e.g. frequency reuse and code reuse for non-overlapping clusters, 2)

routing information can be restricted to a relatively small set of nodes, and 3) it makes the ad hoc network seem smaller and more stable from the AP's point of view [35]. Of course, there are difficulties as well, e.g. explicit control message exchange between APs, ripple effects of re-clustering, possibly unbounded computational rounds, etc. Abundant studies have been conducted on this clustering problem. Most of them take a routing algorithm's perspective, and focus on distributed algorithms. [35] gives a good overview of different clustering methods. It categorizes the algorithms according to the objectives of the clustering algorithms into six groups: 1) Dominant Set (DS) based, 2) low-maintainence, 3) mobility-aware, 4) energy efficient, 5) load-balancing, and 6) combined-metric based clustering methods. More references about these types of clustering algorithms can be found therein.

In our application, we assume centralized control; i.e. a central point where all information is collected and handoff/deployment decision are made is assumed. For most of the clustering algorithms mentioned in [35], cluster heads are elected among all the nodes, but however our problem setup doesn't require UAV relays to be right on top of one of the APs. In addition, since there is a limit to the communication range, there is an extra constraint on the maximum radius of each cluster that needs to be accounted for. Therefore, adaptive clustering methodologies from a routing perspective are not directly applicable to our UAV relay deployment problem. Since global information for the whole network is available, one possibility would be to use "K-means" algorithms to adaptively group APs and place the mobile relays at the center of each cluster for every update instant [145]. Even though this approach relieves the constraint on relays' positions, i.e. the positions of the relays are no longer constrained to be on top of the nodes, they still suffer from two major drawbacks: 1) no global optimality is guaranteed 2) the number of clusters have to be decided before the application of the algorithms [146, 147]. The maximum radius of the cluster constraint is once again another issue that hampers the direct applications of "K-means" algorithms to this specific deployment problem. Other methods involving the ideas of cell planning can easily take into account the size constraint of the

cluster, but running an optimization involving all the nodes for every update instant is computationally intensive.

Here we pose the UAV deployment and clustering problem differently for the sake of avoiding the above predicaments and maintaining simplicity. For bonding various APs into subnets, we first suppose that only one relay is present to start with. At the first update interval, the base generates a list of APs that aren't served by this UAV. For the "non-service" list, the deployment method presented below will give the number of relays necessary and the "optimal" positions for them. As for other update intervals, handoff techniques will allow APs to switch between relays for better link quality. The introduction of the handoff technique makes it possible to relieve the controller from optimizing over all APs for every update interval, and is therefore cheaper to implement. Even though an ideal global optimization would yield the best possible clustering topology and keep a minimum number of required relays, besides the high computational cost required to accomplish this solution, the dynamic constraint on UAV movement may prohibit the realization of this topology. Later, if there are unaccommodated APs, the deployment process will be conducted only for them to narrow down the potential search space. When relays are no longer serving any AP, they will be removed from the network. The handoff algorithm discussed in the previous section automatically changes the topology of the network as necessary, and this, by its own nature, is also a special type of adaptive clustering method.

The reasoning in the above paragraph leaves only the optimal UAV relay deployment problem to be solved. To simplify this problem in our scenario, recall that we assume no interference between different APs. Furthermore, we assume perfect communication between the UAVs and the BTS. The QoS requirement is that a minimum rate $R_{\min}$ is guaranteed for every AP, but if $R_{\min} > B \cdot \log_2 K^L$, where $B$ is the bandwidth allocated to each user and $K^L$ is the largest possible number of constellation points used for transmission, then by using the adaptive modulation scheme described before, it is impossible to fulfill the link quality requirement. In other words, this specific communication system can not support a data rate that is greater

than $B \cdot \log_2 K^L$. Otherwise, the minimum rate constraint of each user can always be satisfied by deploying more relays or increasing transmit power (however, this is not the circumstance under consideration). Here we assume $R_{\min}$ is predetermined to be a value that allows a meaningful solution for the relay assignment and rate optimization problems, and we will discuss the relay deployment problem including the number of new relays and the position for each of them.

As assumed above, only one relay (call it $U_1$) is present to begin with during the formation of the network. Section 6.4 gives a sufficient condition for it to simultaneously host a set of APs, i.e. $\cap_{j=1}^{|\mathcal{I}_1|} \Xi_1^{\mathcal{I}_1(j)} \neq \emptyset$. Determining which APs to serve with this initial relay, or equivalently determining $\mathcal{I}_1$, is an interesting problem, because some of the AP admissible regions are conflicting, making them incompatible in the same subnet. There are numerous methods to solve this problem. However, it is desirable to have the UAV support as many APs as possible. How many APs and which APs should the current UAV support? To solve this question, an optimization problem is posed, i.e.

$$\arg\max_{\mathcal{I}_1} |\mathcal{I}_1|, \qquad s.t. \;\; \mathcal{I}_1 \subseteq \mathcal{Q} \;\&\; \mathcal{S}_1 \neq \emptyset. \tag{6.26}$$

Exhaustive search can be used to fulfill this task for the small network considered in our simulation. Note that when the number of APs increases, a more efficient optimization technique is required to save the computational cost. We define the non-service list $\mathcal{M} \equiv \mathcal{Q} - \cup_{j=1}^{N} \mathcal{I}_j$ as the set of APs that cannot be supported by the current number of relays (here $N = 1$), and therefore a non-empty $\mathcal{M}$ requires more UAV(s) to be added to the network.

To accommodate the APs in list $\mathcal{M}$, the number of new relays that will be deployed and their respective positions need to be decided at the base. To answer these two questions, first of all, the coverage area of each relay should be examined. For simplicity, the channel model we assume implies a circular coverage shape for the relay. At the fringe of the $j$-th relay's coverage, according to (6.7), the average

received SNR can be expressed as

$$\bar{\gamma}_j = \frac{E\left(\|\mathbf{H}_{\text{norm}}\|_F^2\right)}{\left(d_{j,u}^0\right)^{2\alpha}}\rho, \tag{6.27}$$

where

$$E\left(\|\mathbf{H}_{\text{norm}}\|_F^2\right) = \int_0^\infty x f(x) dx\,, \tag{6.28}$$

and $d_{j,u}^0$ is the distance from the AP at the fringe of the coverage area to the $j$-th relay. For all the APs in the coverage area to be able to communicate at a minimum rate $R_{\min}$, $d_{j,u}^0$ has to be chosen carefully such that $\bar{\gamma}_j$ will ensure that $R(t)$ in (6.13) for all APs hosted by the $j$-th relay is greater than $R_{\min}$, and meanwhile maximum SER is kept under a predetermined threshold (to satisfy an extra initial SER QoS constraint). The complicated expression in (6.13) does not provide any insight for determining $d_{j,u}^0$ analytically, however numerical results can easily be obtained. To design the system with some margin and also for the sake of a closed-form solution, we simply require that at the fringe of each UAV's coverage, the APs can communicate at the highest constellation level that it is capable of transmitting at within a predetermined SER level. That is

$$\bar{\gamma}_j = \frac{2\left[Q^{-1}\left(\frac{Pe}{Ne(K^L)}\right)\right]^2}{d_{\min}^2(K^L)}\,, \tag{6.29}$$

according to (6.5), where $\overline{N}_e(\cdot)$ and $d_{\min}(\cdot)$ are a function of the constellation numbers respectively, and $P_e$ is the maximum tolerable SER. The radius of coverage for $j$-th relay is further determined as $d_{j,u}^0 = \left(\frac{E\left(\|\mathbf{H}_{\text{norm}}\|_F^2\right)\cdot\rho}{\bar{\gamma}_j}\right)^{\frac{1}{2\alpha}}$. We can futher evaluate $E\left(\|\mathbf{H}_{\text{norm}}\|_F^2\right)$ by using the PDF derived in (6.8). Note that in a fixed communication environment, the expectation in (6.28) is a constant for each update interval.

Since the shape of the relay coverage area is assumed to be circular, for AP to be served by a given relay, a necessary and sufficient condition is that the relay is in the circular area with center at the position of the AP, and radius equal to $r_{j,u}^0 = \sqrt{\left(d_{j,u}^0\right)^2 - h^2}$, where $h$ is the altitude difference between the relay and the AP under consideration. If a set of such circles is drawn for all the APs in the set $\mathcal{M}$, the ones that overlap can share one common UAV relay as shown in Fig. 6.6.

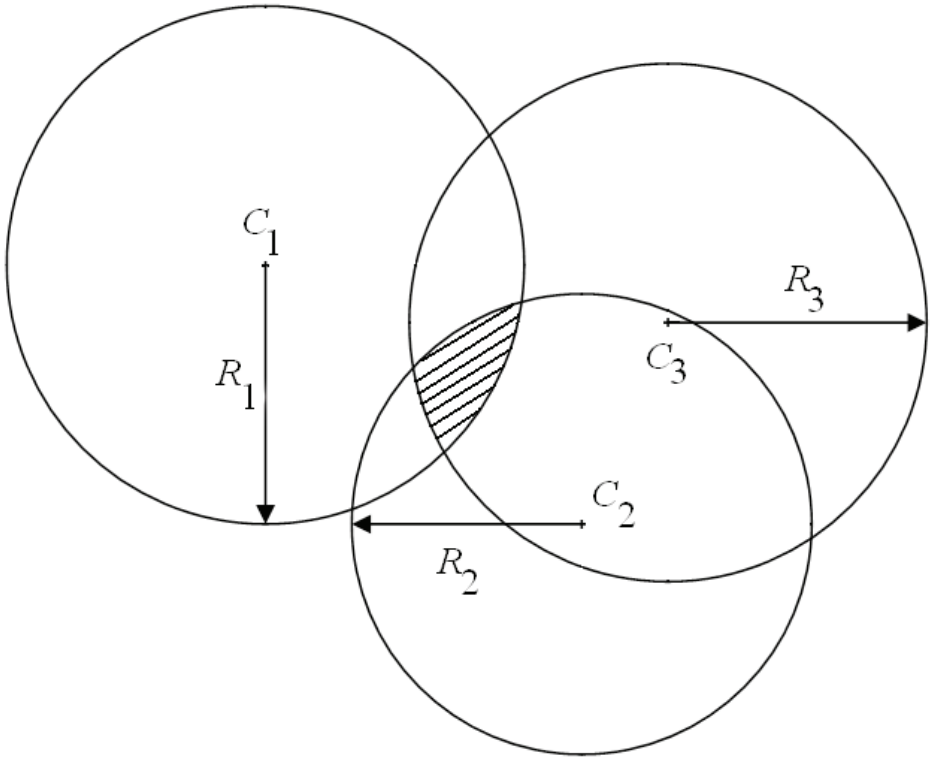



**Figure 6.6:** Optimal UAV deployment for multiple unaccommodated APs.

The problem of determining the required number of UAVs to setup a network can be converted to finding a grouping method such that the circles in each group jointly intersect. It is simple to decide if two circles intersect, i.e. if the distance between the two centers is smaller than the sum of the two radii, they have to intersect. However a crude algorithm that determines the maximum number of circles that jointly intersect involves an exhaustive search and is prohibitive when the number of circles is moderately large. But some preprocessing can be done in this deployment to narrow down the search space; e.g. APs that are too far from each other can be predetermined to not be able to share a common relay. The possibly prohibitive search usually only happen once at the beginning of the formation of the network, when only one UAV is assumed in the air and there are possibly many APs that have not been accommodated. After that, the dimension of a typical non-service list $\mathcal{M}$ is relatively small. A smarter algorithm implemented in matlab code can be found in [148]. With the

number of new relays and the lists of APs that can share relays decided, the positions for each potential UAV relay need to be decided. Different optimization problems can be used for this problem, such as the placement of relays can be targeting at maximizing the sum capacity of the APs they will host, or maximizing the sum rate of the APs. Noting $\sum_{j=1}^{n} \log\left(1+\bar{\gamma}_j\right) = \log\left(\prod_{j=1}^{n}\left(1+\bar{\gamma}_j\right)\right) \leq n \cdot \left(\log\left(1+\sum_{j=1}^{n}\gamma_j/n\right)\right)$, an ad hoc approach to maximize the sum capacity is to maximize the sum of average SNRs for each AP. Again, as illustrated in Fig. 6.6, the UAV should be put in the shadowed area, and the exact position is calculated by solving the following optimization problem:

$$\begin{aligned} \arg\max_{x_u, y_u} \quad & \sum_{j=1}^{l} \frac{E\left(\|\mathbf{H}^j_{\text{norm}}\|_F^2\right)}{d_{j,u}^{2\alpha_j}(x_u, y_u, h_j))} \\ s.t. \quad & d_{j,u}(x_u, y_u, h_j) \leq d^0_{j,u} \ , \end{aligned} \tag{6.30}$$

where $h_j$ is the altitude difference between the relay and the $j$-th AP, and $l$ is the number of APs that can share one common relay. This optimization problem can be efficiently solved by any gradient based technique. After the positions of newly deployed relays are obtained, the handoff algorithm is run again to see if better links can be created by switching the APs from their previous service provider to the new relays.

One thing worth pointing out is that the QoS that is under consideration so far is the ENTR. In other words, every AP's minimum communication rate is guaranteed. However the quality of this communication in terms of SER is not accounted for except when a new UAV is deployed, only its coverage area considers the SER QoS. The extension of all the algorithms presented above to handle SER QoS is straightforward. By using the SER expression derived in (6.14), a one-to-one mapping between the link SER and the average link SNR $\bar{\gamma}$ can be created, at least numerically. A maximum SER will give a minimum average SNR value (call it $\bar{\gamma}_{\min}$) that allows the QoS to be satisfied. With $\bar{\gamma}_{\min}$ generated, at every time update interval, the average received SNR needs to be compared to this value for every AP. If an AP's received SNR is less than $\bar{\gamma}_{\min}$, it will be put in the list $\mathcal{M}$. In other words, the trigger for expanding

the "non-service" list $\mathcal{M}$ are both the empty admissible region and a small average received SNR. Other than this, every algorithm remains as described above.

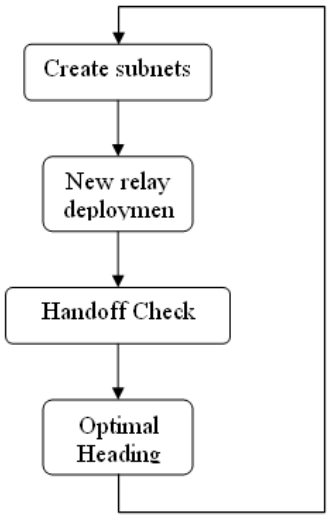



**Figure 6.7:** Flowchart of the network protocol.

The above two sections discuss methods for acquiring optimal heading commands for rate maximization and pose a few optimization problems for deploying relays in the network. As shown in Fig. 6.7, the proposed protocol includes the assignment of heading commands for each UAV in the current network, new UAV deployment, and the handoff of the APs between different UAVs. In summary, first, for each update time interval, a subnet creation is required (either through a clustering algorithm at the initial time or using the handoff algorithm for the subsequent update interval). As mentioned before, a subnet is defined as a star topology with one relay (UAV) as the center node. The whole network is thus divided into several different subnets. Each subnet then checks to see whether the APs within the subnet can be hosted by their associated relay. If not, new UAV(s) needs to be deployed. With the new UAV(s) added in, a handoff check is performed for every AP. The last step of

each iteration is to re-generate a list of new subnets, and give the optimal heading for each UAV. The base can follow this protocol to deploy and control mobile relays to achieve a certain level of "optimal" network communication.

## 6.6 Simulation Results

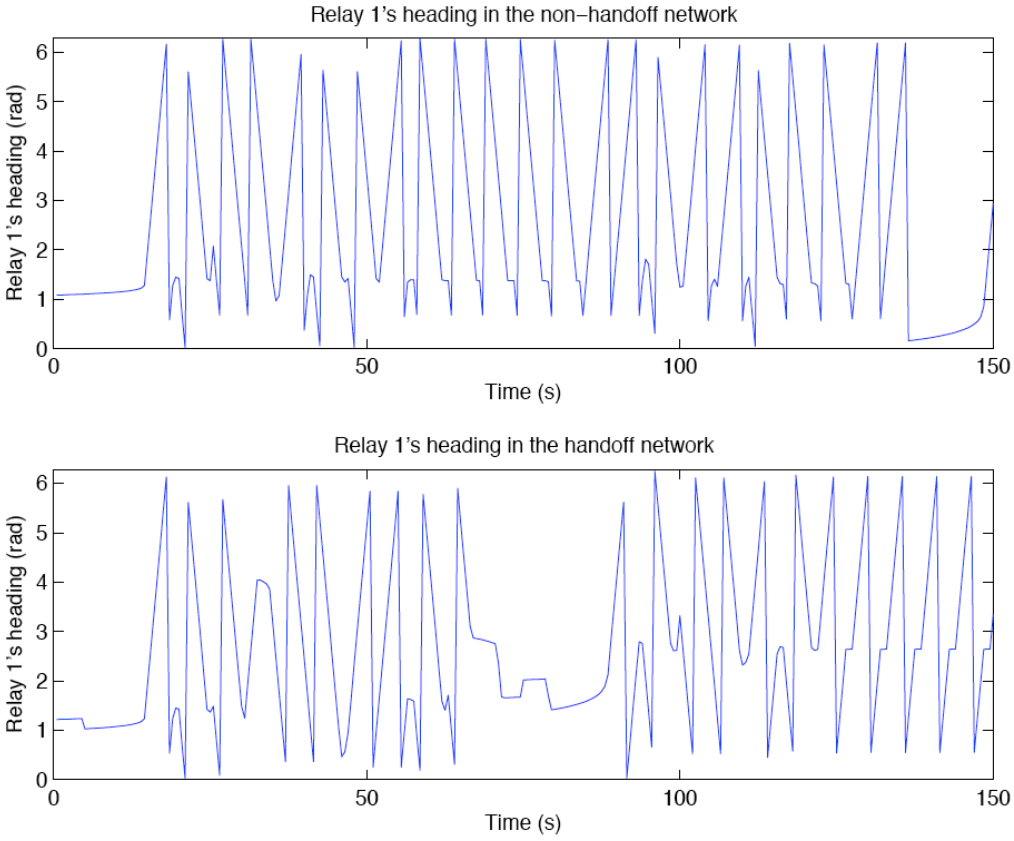



**Figure 6.8:** Heading of UAV-1 in the network.

Simulation results for the proposed network protocol are shown in Fig. 6.8 ∼ Fig. 6.13. In the simulations, 10 APs are assumed to be moving on the ground with random initial directions, and all APs move in a straight line. Each of the APs has 3 antennas and 2W transmit power, and they are all assumed to be moving at 10m/s. To simplify the simulation, all the APs are assumed to have the same propagation environment, i.e. $\alpha = 1.5$, the number of scatterers in the environment is 40, and the

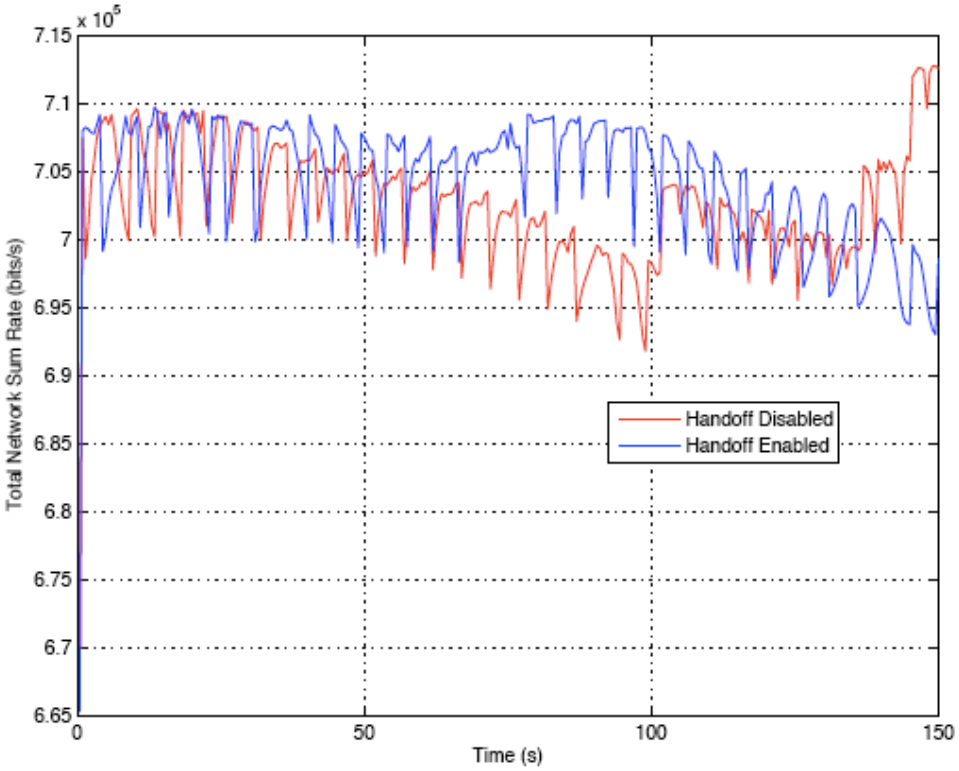



**Figure 6.9:** Sum uplink transmission rate.

equivalent scattering radius is 100 $\lambda$. We consider a narrowband scenario where each AP has a bandwidth of 20kHz. All the UAVs are assumed to have 2 antennas with $2\lambda$ separation and fly at a height of 3600m. UAVs fly at a speed of 50m/s with the heading constraint $\Delta\delta \leq \frac{\pi}{9}$. The minimum transmission rate constraint is set to be $R_{i,u} \geq 6.61 \times 10^4$ bits/s. The update time interval is 0.5s, and the simulation is run for 150s. The APs are randomly initialized to be positioned on the ground within a 2500m $\times$ 2500m square, and only one UAV is put into the system at $[0\ 0\ 3600]^T$ to begin with. From the derivations above, it is clear that controlling the heading of the UAVs will maximize the possible throughput. Therefore, the simulation results here compare scenarios where no AP handoff is allowed to the scenarios where AP handoff is implemented. Fig. 6.8 plots the instantaneous heading for UAV-1. As we can see there is no sharp changes in the UAV's heading between consecutive time slots, thanks to the second constraint in (6.16) we apply in our optimization. As shown

in Fig. 6.9, the total network throughput has similar behavior. The small wiggles are due to the changes in the heading of the UAVs. During most of the simulation time, the network with handoff capability gives higher throughput, however after $t \approx 135$s, the non-handoff network gives better results. This is due to the fact that more UAVs are put into the network to boost the system performance. This effect can be clearly seen in Fig. 6.10. Since no handoff is assumed, the red curve

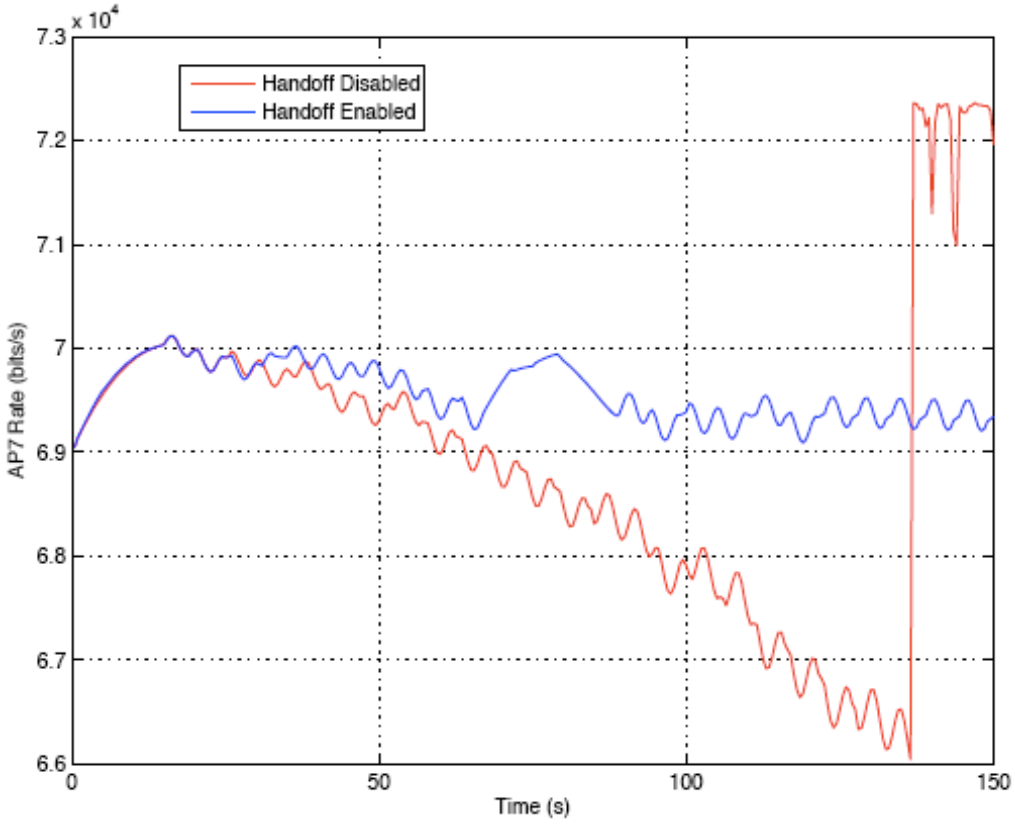



**Figure 6.10:** Link 7 transmission rate.

representing the 7th AP's transmission rate keeps going down, until it reaches the lowest supportable transmission rate. To guarantee the QoS (the first constraint in (6.16)), one more UAV needs to be added into the network, which happens to be right on top of the 7th AP, which causes the sharp rate increase. However generally speaking, the network with handoff will provide a better communication link for

each AP as indicated in Fig. 6.11. The topologies of the non-handoff network and the handoff-enabled network at a few stages of the network evolution are shown in Fig. 6.12 and Fig. 6.13 respectively. The dots in the figures represent the positions of the APs, and the stars stand for the positions of the UAVs.

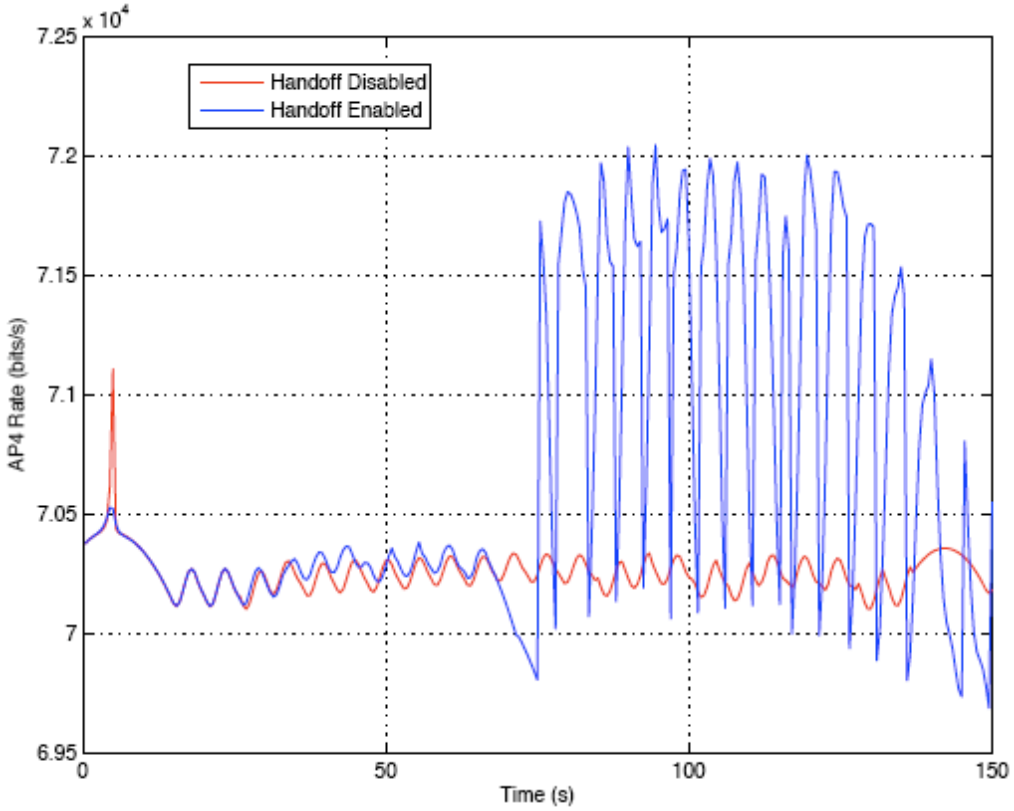
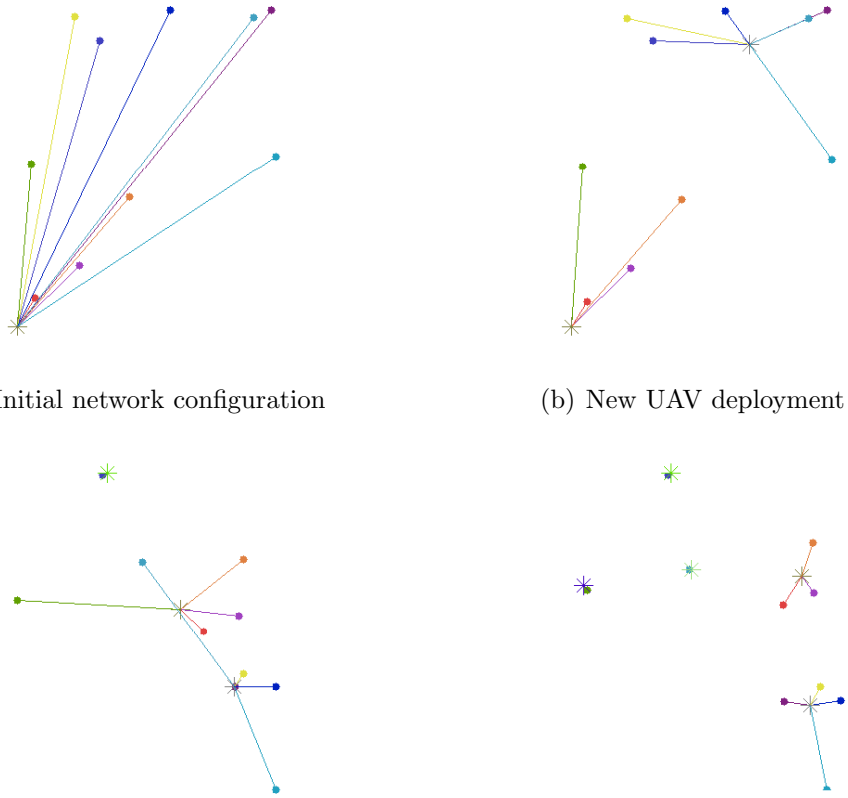



**Figure 6.11:** Link 4 transmission rate.

As shown in Fig. 6.12(a) and Fig. 6.13(a), all the APs attempt to communicate with the single UAV relay at the beginning. However, as we can see in Fig. 6.12(b) and Fig. 6.13(b), due to the rate constraints, not all the APs can be hosted in this case. Therefore one UAV is added to change the network topology. As one of the APs moves farther and farther, the rate constraints can not be achieved, and this causes one UAV to be added in right above it as shown in Fig. 6.12(b) and Fig. 6.12(c). Another handoff happens in the network as the APs and UAVs move (Fig. 6.13(c)).



(a) Initial network configuration

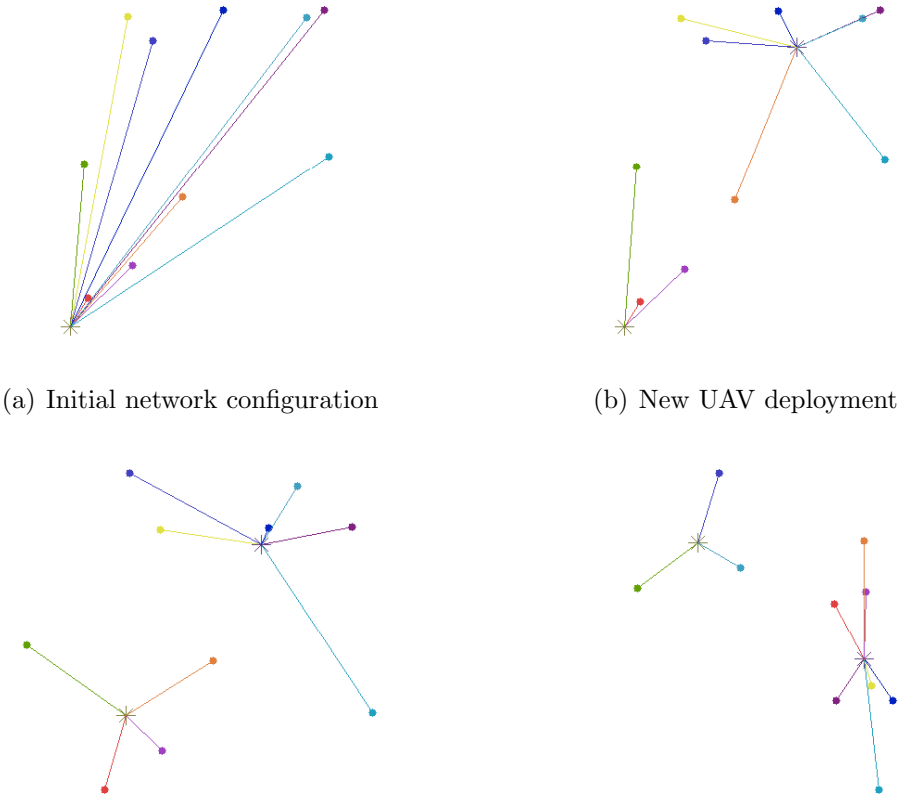
(b) New UAV deployment

(c) Handoff during evolution

(d) Final network topology

**Figure 6.12:** UAV network simulation without handoff algorithm.

The final network topology is plotted in Fig. 6.12(d) and Fig. 6.13(d). As we can see, if handoffs are not allowed in the network, more UAVs need to be added in the network in order to maintain a minimum communication rate for each link. Even though the total network throughput can be increased due to the deployment of more UAVs, precious UAV resources are wasted. The advantages of the proposed protocol are two-fold. It considers the best position of the UAV deployment, so that new UAVs are only added when necessary, as compared to the non-handoff network scenario where UAVs are usually added just to support one AP. Moreover, by using the sinusoidal approximation shown in the Appendix, the heading command given to the UAV for each subnet is close to optimal.



(a) Initial network configuration

(b) New UAV deployment

(c) Handoff during evolution

(d) Final network topology

**Figure 6.13:** UAV network simulation with handoff algorithm implemented.

## 6.7 Summary

In this chapter, we have derived expressions for the SNR and the ENTR for each AP-UAV link with adaptive modulation. A sinusoidal approximation has been derived to approximate the single link data rate. We have examined a network with a star topology, and proposed a method that finds the optimal heading of a UAV to achieve the highest overall data rate in an uplink multi-user system. Furthermore, a more complicated scenario has been considered where new UAVs need to be added into the network to achieve the system requirement. A handoff algorithm has been proposed to deal with subnet switching with the heading constraint imposed. A few optimization problems have been posed to solve UAV deployment problems.

Further work can be done to make the presented protocol more practical, including modifying the assumptions, solving the optimization problems more efficiently, and imposing more constraints, etc. The chapter assumes unlimited bandwidth resources are available at the relays. This is justifiable when the number of APs is not too large and each of them does not individually require too much bandwidth. If limited bandwidth has to be considered, it is possible that when assigning APs, some kind of load balancing measure needs to be part of the cost metric in the optimization formulation. When polarized antennas are installed on the relays, larger capacity can be achieved and a different channel model might be more accurate in characterizing the correlation between the UAV's antennas. For some of the posed optimization problems, more elegant solutions need to be formed to reduce the computational cost. An additional issue worth considering is how the UAVs reach their deployment locations. We propose a method to obtain a set of coordinates for the relays, but how they reach those positions remains unexplored. A possible constraint that requires the initial deployment location to be within a certain distance from air base is feasible in controlling the network setup time. In addition, there are other aspects to improve the quality of the communication that are worth considering, e.g. trajectory design for the relays, battery consumptions of the relays, AP movement prediction, etc., to name a few.

# Chapter 7

# Conclusions and Future work

## 7.1 Summary of the Dissertation

The past century has seen tremendous advances in wireless technologies. The penetration of these technologies into human society has brought significant convenience into our life, including: cellular telephony, broadband access, bluetooth, etc. They have changed and continue changing the way we perceive and process information from the world. These smart point-to-point link-level signaling strategies create communication networks, which bring together people that are far apart. As the demand for higher data rate applications (e.g. multimedia services, wireless internet access, etc.) increase and the number of data service subscribers burgeons, the throughput of wireless networks needs to be expanded. Radio resources such as transmit power, bandwidth, etc., are scarce. But fortunately, by carefully manipulating these limited resources, network throughput can be further improved to support higher quality of service and larger number of users. This dissertation focuses on different ways to optimize the performance of the network, be it cellular networks or relay networks.

Inspired by the ability of an opportunistic scheduler to provide higher data rates, the first part of the dissertation is concerned with the performance of the Max-SNR scheduler. Different from most of the work in the literature where ideal assumptions on the communication systems are made to evaluate throughput, some practical issues that may be encountered when implementing a real system are modeled and addressed to derive a reasonably realistic metric for network performance. A simple exponential BER model is developed for the system with a turbo encoding block. With the help of this model, multiuser diversity gain is studied in a more

thorough way. Simulations are shown to quantify the loss due to the impossibility of achieving the ideal system assumptions.

Still concerned with the performance of a Max-SNR scheduler, the second part of the dissertation is dedicated to the analysis of an asymmetric network, where users have large disparities in their velocities and average SNR. Distinct from most papers that address the performance of a proportional fair scheduler, analytical results that characterize the multiuser diversity gain of the Max-SNR scheduler in this setup are presented in the dissertation. When a mobile base station is deployed, or the users have great mobility, Doppler effects are predominant. If the luxury of keeping track of the channel cannot be afforded, channel knowledge will be outdated and decoding with inaccurate channel information will lead to an error floor in system performance. In these scenarios, the slot duration should be designed with the Doppler effect taken into account. In other words, if channel tracking is not frequent, the slot duration allocated to each user should not be too long. Given a certain fixed slot duration, and when channel knowledge is only assumed at the scheduling time, the Max-SNR scheduler may not be the optimal solution. A statistical channel-aware scheduler is proposed to select users based on relatively long term channel statistics, i.e. average SNR and Doppler spreads.

The last part of the dissertation studies a tactical communication scenario involving UAV-aided communication networks. For difficult environments where fixed infrastructure is not available or direct links are obstructed, relays can be deployed to help support cooperative communications. Due to their relatively low cost, simple operability, etc., UAVs are usually good candidates to act as mobile relays. Trajectory design has to be considered for these UAVs, specifically how to determine heading commands for each UAV to accomplish their relay mission. The UAV dynamics impose certain constraints on their mobility, which further defines the range within which UAVs can move within a given amount of time. For reliable communication to be possible, a certain grade of QoS needs to be guaranteed, e.g. minimum communication rate, probability of symbol error rate, etc. The optimal UAV motion control problem is solved in the dissertation by maximizing the network transmission

rate under the dynamics constraint and the minimum QoS constraints. A hierarchical star topology with each UAV as the center node serving a group of access points is assumed. Different sub-networks are defined and as the ground nodes and UAV relays move, the topology of the sub-network will change. As the network evolves, ground nodes may need to switch to other sub-network for better reception and transmission. Hence, the handoff problem is discussed under the same set of constraints as above. When the current set of UAV relays is not enough to support all the users on the ground, more relays need to be deployed. The optimal deployment problem is addressed in the dissertation by posing another optimization problem.

## 7.2 Possible Future Work

Wireless communication network is an area with abundant possibilities for research ideas. Extensions can be easily made to the results achieved in this dissertation. The most straightforward extension of the realistic multiuser scheduler performance study is to drop the SISO assumption. The use of multiple transmit antennas in wireless networks has gained considerable attention, so consideration of the effect of multiple antennas on multiuser diversity gain under the realistic assumptions is certainly of interest. When a MIMO system is under consideration, various transmission schemes are possible, including beamforming, spatial multiplexing, etc. The specific transmission technique that is chosen will affect the performance of the network and the multiuser diversity gain. It will be interesting to find an unifying approach independent of the transmission strategy to study the multiuser diversity gain for multiple MIMO users with practical impairments accounted for. Besides, as mentioned in Chapter 4, the extension of the analysis to handle the proportional fair scheduler in an asymmetric network is also a promising research topic for future work.

In addition to the aforementioned extensions, there are other aspects of the system that need more inspection, such as nonlinearity of the power amplifier, IQ imbalance and Local Oscillator (LO) phase noise, to name a few. One of the RF system specifications to measure the overall quality of transmission is called Error

Vector Magnitude (EVM) and its relationship with IQ imbalance and LO phase noise have been derived in [149]. Further investigations can be found in [150, 151, 152, 153, 154, 155]. According to the literature above, the gain/phase imbalance and phase noise effects can be modeled as the transmitted signals being multiplied by a matrix and then offset by a dc vector before being broadcast into wireless channels. The estimates of this matrix and vector will further increase the uncertainty of the composite transfer function of the system, which would first of all impact the link performance for a single user system and then affect the throughput of a network with a scheduler making decisions based on the estimated channel knowledge. The extension of the EVM calculation methods to MIMO system itself is an interesting research topic. In addition to that, any multiuser system analysis addressing these EVM causing effects would push the results closer to reality.

The Doppler effect is always of considerable importance when users with high mobility are present in the network. The large velocities will complicate the transceiver design in the sense that 1) carrier offset has to be compensated for, 2) inter-carrier Interference will be present in multicarrier communication, and 3) training and channel tracking need to be performed more frequently. A faster channel varying rate outdates the channel knowledge more quickly, and reduces the realizable channel capacity. Besides its degradation in point-to-point link throughput, it will also cause the scheduler to make decisions based on information that does not truly reflect the user's channel behavior for the scheduled time slot. Thus there is additional network capacity loss due to mis-scheduling. This dissertation only considers a network with SISO users. A similar analysis can be carried out for the MIMO case. In my research work, the evolution of the channel is assumed to follow a Markov process, and the channel decorrelation with regard to time is captured by Jakes model. More realistic channel measurements or other models can be considered to obtain more practical results. The performance of SCA scheduler should be studied more, due to its convenience of not requiring the users' instantaneous channel knowledge. Better yet, when the users' channel knowledge is available, long term channel statistics should also be taken into account when deciding the user to serve. Typically,

when the number of the users is large enough, the study of asymptotic behavior of the scheduler can be justified. Asymptotic analysis on the multiuser diversity can be found in the literature, but extending that analysis to account for the design of the scheduler in systems with large Doppler effects can still be explored.

The work presented on UAV-assisted networks gives a way to optimally determine the trajectory of the mobile relays. However, the heading command is based on knowledge of the instantaneous positions of both UAVs and ground nodes, as well as the ability to constantly change UAVs' heading. If these assumptions cannot be justified, and the UAVs can only be constrained to fly in simple patterns like circles or ellipses, the design of optimal trajectories for mobile relay deployment must be modified. To be more specific, it would be necessary to find optimal design parameters for the motion pattern of the mobile relays, such as the center, radius, and orientation of the trajectory. The expressions for the Shannon capacity and probability of a user being selected when relays are at at each point on the trajectory would need to be derived and verified. Possible objective functions and constraint functions based on these expressions should be created. By optimizing the objective functions subject to the constraints, optimal deployment can be accomplished. Some preliminary studies have been started on this idea and results obtained are shown in the following few paragraphs.

We are interested in the applications where a Max-SNR scheduler is implemented on the UAVs, and optimal trajectory design of the UAVs to maximize network throughput is the focus. To simplify the problem, we first start with a Rayleigh fading assumption, although each user can have a different average receive SNR. In other words, an asymmetric network is considered in this problem. When the relay is moving in its constrained circular flying pattern, each user's receive SNR changes accordingly. When the max-SNR scheduler is used, the system's Shannon capacity can be shown to be:

$$C_K = \log_2 e \cdot \sum_{j=1}^{K} \sum_{l_j=0}^{1} e^{\sum_{j=1}^{K} \frac{l_j}{\overline{\gamma}_j}} \cdot E_1 \left\{ \sum_{j=1}^{K} \frac{l_j}{\overline{\gamma}_j} \right\}, \tag{7.1}$$

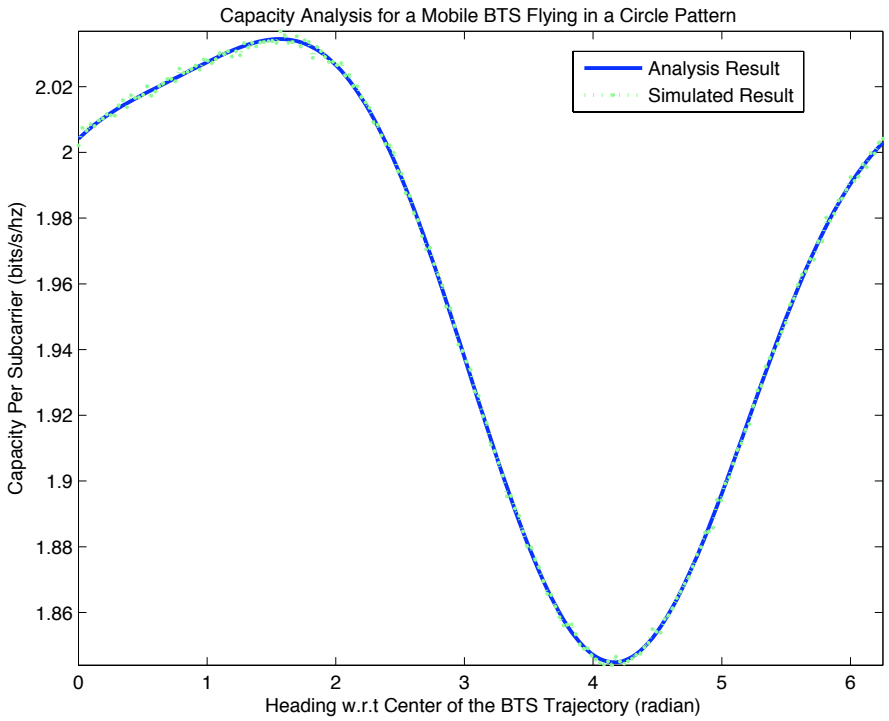



**Figure 7.1:** System Shannon capacity when BTS is flying in a circle pattern.

where $K$ is the number of the users in the network, and $\bar{\gamma}_j$ is the average receive SNR per user. The probability of the $j$-th user being scheduled is:

$$\text{Prob}(user\ j) = \sum_{i \neq j} \sum_{l_i=0}^{1} \frac{(-1)^{\sum_{i \neq j} l_i}}{1 + \sum_{i \neq j} l_j \cdot \frac{\gamma_j}{\gamma_i}} \,. \tag{7.2}$$

Simulation plots in Fig. 7.1 and Fig. 7.2 validate the above claims. In this simulation, 5 users locations are randomly generated and assumed fixed after creation. The mobile relay is flying in a circle centered at position $[0\ 0\ 10000]^T$m, with a turning radius of 2000m. At each point of the trajectory, a Monte-Carlo simulation is performed to measure average Shannon capacity and user selection probability. As can be clearly seen in 7.1, there is a point on the trajectory that achieves maximum Shannon capacity, and due to the time-varying average receive SNR, the probability of a specific user being scheduled varies considerably. Hence, if the BTS is not constrained to fly in a certain pattern, it is obvious that an optimal heading command

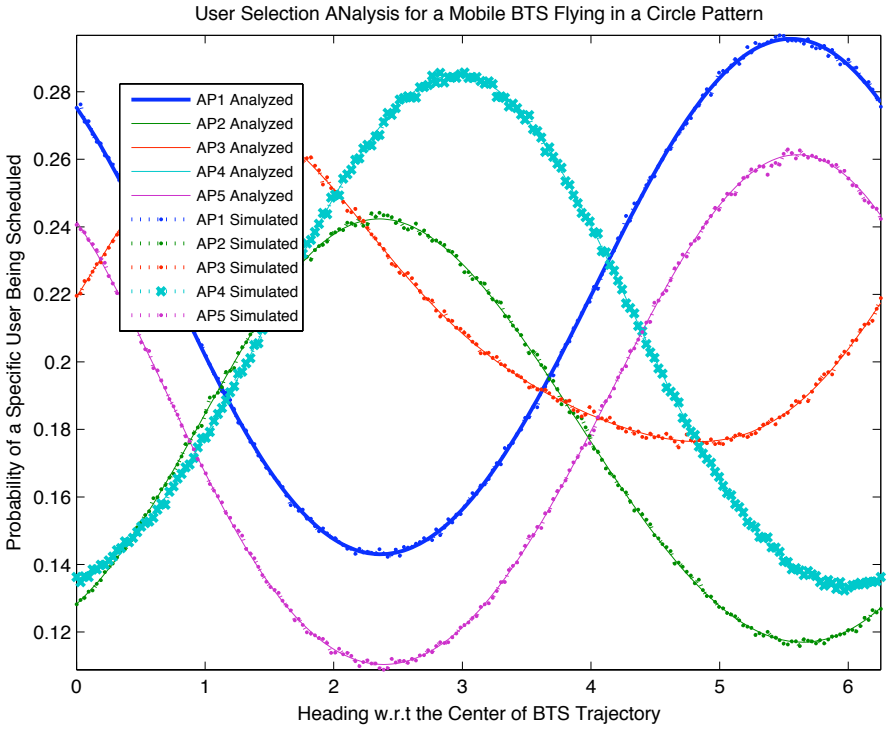



**Figure 7.2:** Probability of a specific user being scheduled when base station is flying in a circle pattern.

can be obtained for each time instant. However, if the BTS has to fly in a pattern, the average Shannon capacity over the trajectory can be evaluated by integrating (7.1) over the trajectory. By the same reasoning, the probability of a specific user being scheduled can also be obtained by integrating (7.2). The difference between different patterns lies only in the integration path. Since the control parameters (radius, center location for the circle case; axis lengths, center location and orientation angle for the elliptical case) affect the trajectory integration path, the optimization problem can be solved to obtain them. Different cost functions can be used depending on the goal of the deployment. For example if the idea is simply to maximize capacity, average Shannon capacity can be used as the objective function; however, to guarantee a certain fairness in the system, we can maximize the same average Shannon capacity objective function subject to the constraint that the variance of the probability of

each user being scheduled is less than a threshold. Of course, if a Rician channel is assumed, the whole problem presented above has a different set of solutions.

In addition, antenna designs for multiple element arrays on UAVs is an area worth of pursuing, since due to the small angular spread observed at the UAV side, high spatial correlation is usually present. Polarized antennas can be deployed for the MIMO setup to achieve more capacity and reliability. Clustering algorithms that automatically group nodes together to form smaller network entities and routing algorithms that propagate messages without generating too much delay and overhead can also be investigated in UAV-assisted mobile networks.

## Bibliography


[1] G. Stuber, *Principles of Mobile Communication*. Springer, 2001. 1, 121

[2] T. S. Rappaport, *Wireless Communications, Principles and Practice*. Prentice Hall PTR, 1996. 1, 20, 21, 22, 23, 24, 30, 112

[3] W. Jakes, *Microwave Mobile Communications*. Wiley, 1974. 3, 11, 23, 55, 88

[4] J. Guey, M. Fitz, M. Bell, and W. Kuo, "Signal design for transmitter diversity wireless communication systems over rayleigh fading channels," in *Proc. IEEE VTC*, vol. 1, 1996, pp. 136–140. 3

[5] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, 1998. 3, 113

[6] B. Vucetic and J. Yuan, *Space-Time Coding*. John Wiley & Sons, 2003. 3, 55, 56, 58, 59, 60

[7] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge University Press, 2003. 4, 20, 33, 60, 113

[8] WiMAX Forum, "Mobile WiMAX Part I: A technical overview and performance evaluation," *http://www.wimaxforum.org/technology/downloads/*, 2006. 4

[9] ——, "Mobile WiMAX Part II: A comparative analysis," *http://www.wimaxforum.org/technology/downloads/*, 2006. 4

[10] M. Thelander, "WiMAX opportunities and challenges in a wireless world," *White Paper for CDMA Development Group in Signals Research Group LLC*, 2005. 4

[11] J. A. Gutierrez, E. H. Callaway, and R. Barrett, *IEEE 802.15.4 Low-Rate Wireless Personal Area Networks: Enabling Wireless Sensor Networks*. IEEE Press, 2003. 4

[12] H. Zimmermann, "OSI reference model the ISO model of architecture for open systems interconnection," in *IEEE Transactions on Communications*, vol. 28, no. 4, 1980, pp. 425–432. 5, 6

[13] A. S. Tanenbaum, *Computer Networks*. Prentice Hall, 2002. 5, 6

[14] P. E. Green, *Network Interconnection and Protocol Conversion*. IEEE Press, 1988. 6

[15] A. S. Acampora, *An Introduction to Broadband Networks*. Plenum Press, 1994. 6

[16] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *IEEE International Conference on Communications*, vol. 53, June 1995, pp. 331–335. 8, 63, 85

[17] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proceedings of IEEE International Symposium on Information Theory*, 1997, p. 27. 8, 63, 85

[18] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005. 8, 47, 52, 62, 85, 86

[19] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," in *IEEE Transactions on Information Theory*, vol. 48, June 2002, pp. 1277–1294. 9, 63, 85

[20] H. Zhou, D. Yang, W. Qi, and M. Ma, "On performance of mutiuser diversity in SISO and MIMO wireless communication," in *Proceedings on Personal, Indoor and Mobile Radio Communications*, vol. 3, 2003, pp. 2872 – 2875. 9

[21] J. Jiang, M. Buehrer, and W. Tranter, "Antenna diversity in multiuser data networks," in *IEEE Transactions on Communications*, vol. 52, 2004, pp. 490–497. 9

[22] Q. Ma and C. Tepedelenlioglu, "Practical multiuser diversity with outdated channel feedback," in *IEEE Transactions on Vehicular Technology*, vol. 54, 2005, pp. 1334–1345. 9, 86

[23] G. Song and Y. Li, "Asymptotic throughput analysis for channel-aware scheduling," in *IEEE Transactions on Communications*, vol. 54, 2006, pp. 1827–1834. 9

[24] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1," in *IEEE International Conference on Communications*, vol. 2, 1993, pp. 1064–1070. 9

[25] P. Peebles, *Radar Principles*. Wiley, 1998. 10, 23

[26] M. Chu and W. Stark, "Effect of mobile velocity on communications in fading channels," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 1, pp. 202–210, 2000. 11

[27] K. Gomadam and S. Jafar, "Modulation and detection for simple receivers in rapidly time varying channels," in *IEEE Transactions on Communications*, vol. 55, 2007, pp. 529–539. 11

[28] ITL Advanced Network Technologies Division, "Wireless ad hoc networks," *http://w3.antd.nist.gov/wahn_bkgnd.shtml*. 12, 13

[29] C. Elliott and B. Heile, "Self-organizing, self-healing wireless networks," in *IEEE International Conference on Personal Wireless Communications*, 2000, pp. 355–362. 12

[30] F. Baker, "An outsider's view of MANET," in *Internet Engineering Task Force document*, 2002. 12

[31] P. Zhan, D. Casbeer, and A. L. Swindlehurst, "A centralized control algorithm for target tracking with UAVs," in *39th IEEE Asilomar Conference*, October 2005. 13, 109

[32] D. Casbeer, P. Zhan, and A. Swindlehurst, "A non-search optimal control solution for a team of muavs in a reconnaissance mission," in *IEEE ICASSP*, 2006, pp. 1169–1172. 13

[33] P. Zhan, D. Casbeer, and A. Swindlehurst, "Adaptive mobile sensor positioning for multi-static target tracking," in *IEEE Transactions on Aerospace and Electronic Systems*, 2007, submitted. 13

[34] K. Xu, X. Hong, M. Geerla, H. Ly, and D. L. Gu, "Landmark routing in large wireless batlefield networks using UAVs," in *IEEE MILCOM 2001*, vol. 1, September 2005, pp. 561–573. 13, 14, 15, 109, 110

[35] J. Yu and P. Chong, "A survey of clustering schemes for mobile ad hoc networks," *IEEE Communications Survey & Tutorials*, vol. 7, no. 1, pp. 32–48, 2005. 13, 14, 128

[36] T. Larsson and N. Hedman, "Routing protocols in wireless ad-hoc networks–a simulation study," Master's thesis, Lulea University of Technology, Stockholm, 1998. 14

[37] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Mobile Computing and Networking*, 1998, pp. 85–97. [Online]. Available: citeseer.ist.psu.edu/broch98performance.html 14

[38] E. M. Royer and C.-K. Toh, "A review of current routing protocols for ad hoc mobile wireless networks," *IEEE Personal Communications*, vol. 6, no. 2, pp. 46–55, 1999. 14

[39] A. Ayyagari, J. Harrang, and S. Ray, "Airborne information and reconnaissance network," in *IEEE Proceedings of Military Communications Conference*, Oct. 1996, pp. 230 – 234. 15, 109

[40] D. Gu, G. Pei, H. Ly, M. Gerla, B. Zhang, and X. Hong, "UAV aided intelligent routing for ad-hoc wireless network in single-area theater," *IEEE WCNC*, vol. 3, pp. 1220–1225, 2000. 15, 109, 110

[41] E. Kuiper and S. Nadjm-Tehrani, "Mobility model for UAV group reconnaissance applications," in *IEEE International Conference on Wireless and Mobile Communications*, 2006, pp. 33–39. 15

[42] D. Horner and A. Healey, "Use of artificial potential fields for UAV guidance and optimization of WLAN communications," *Autonomous Under Vehicles, IEEE/OES*, pp. 88–95, 2004. 15

[43] C. Cheng, P. Hsiao, H. Kung, and D. Vlah, "Maximizing throughput of UAV-relaying networks with the load-carry-and-deliver paradigm," *IEEE WCNC*, pp. 4417–4424, 2007. 15, 109, 110

[44] I. Rubin, A. Behzad, H. Ju, R. Zhang, X. Huang, Y. Liu, and R. Khalaf, "Ad hoc wireless networks with mobile backbones," *IEEE PIMRC*, vol. 1, pp. 566–573, 2004. 15, 109, 110

[45] P. Basu, J. Redi, and V. Shurbanov, "Coordinated flocking of UAVs for improved connectivity of mobile ground nodes," *IEEE MILCOM*, vol. 3, pp. 1628–1634, 2004. 15, 109

[46] Z. Wu, H. Kumar, and A. Davari, "Performance evaluation of OFDM transmission in UAV wireless communication," *Proceedings of the thirty-seventh southeastern symposium on system theory*, pp. 6–10, 2005. 15, 110

[47] R. Palat, A. Annamalau, and J. Reed, "Cooperative relaying for ad-hoc ground networks using swarm UAVs," *IEEE MILCOM*, vol. 3, pp. 1588–1594, 2005. 15, 109, 110

[48] JPL, "JPL's wireless communication reference website," *http://www.wireless-communication.nl/reference/contents.htm*. 20, 21, 23, 37, 45, 46

[49] M. Feuerstein, K. Blackard, T. S. Rappaport, S. Seidel, and H. Xia, "Path loss, delay spread, and outage models as functions of antenna height for microcellular system design," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 3, pp. 487–498, August 1994. 22

[50] K. Bullington, "Radio propagation at frequencies above 30 megacycles," *Proceedings of the IEEE*, pp. 1122–1136, 1947. 22

[51] J. Epstein and D. Peterson, "An experimental study of wave propagation at 840 m/c," *Proceedings of the IRE*, vol. 41, no. 5, pp. 595–611, 1953. 22

[52] J. Deygout, "Multiple knife-edge diffraction of microwaves," *IEEE Transactions on Antennas and Propagation*, vol. AP-14, no. 4, pp. 480–489, 1966. 22

[53] W. Lee, *Mobile Communications Engineering.* McGraw Hill Publications, 1985. 22

[54] H. Griffiths, "From a different perspective: principles, practice and potential of bistatic radar," in *Proceedings of International Radar Conference*, 2003, pp. 1–7. 23

[55] S. Seidel, T. Rappaport, S. Jain, M. Lord, and R. Singh, "Path loss, scattering and multipath delay statistics in four European cities for digital cellular and microcellular radiotelephone," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 4, pp. 721–730, 1991. 23

[56] D. Cox, R. Murray, and A. Norris, "800 MHz attenuation measured in and around suburban houses," *AT&T Bell Laboratory Technical Journal*, vol. 673, no. 6, 1984. 23

[57] R. Bernhardt, "Macroscopic diversity in frequency reuse systems," in *IEEE Journal on selected areas in communications*, vol. SAC 5, 1987, pp. 862–878. 23

[58] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd ed. Prentice Hall, 1993. 25

[59] J. G. Proakis, *Digital Communications.* McGraw Hill, 2001. 25, 29, 30, 41, 42, 48, 49, 50, 51, 52, 90, 113

[60] V. Lau and Y. Kwok, *Channel Adaptive Technologies and Cross Layer Designs for Wireless Systems with Multiple Antennas.* Wiley, 2006. 26, 32

[61] G. D. Durgin, *Space-Time Wireless Channels.* Prentice Hall PTR, 2003. 26, 27, 29, 30, 31

[62] P. Bello, "Characterization of randomly time-variant linear channels," in *IEEE Transactions on Communications Systems*, vol. 11, 1963, pp. 360–393. 27

[63] T. Kailath, "Sampling models for linear time-variant filters," in *M.I.T Research Lab. of Electronics, Cambridge, Mass. Report*, May 1959. 27

[64] R. Tsay, *Analysis of Financial Time Series.* John Wiley & Sons, 2002. 28

[65] A. Goldsmith, *Wireless Communications.* Cambridge University Press, 2005. 29, 55

[66] W. Lee, *Mobile Cellular Communications.* McGraw Hill Publications, 1989. 30

[67] B. Sklar, "Rayleigh fading channels in mobile digital communication systems: Part I: Characterization," in *IEEE Communications Magazine*, 1997, pp. 90–100. 30

[68] D. Greenwood and L. Hanzo, "Characterisation of mobile radio channels," in *Mobile Radio Communications, by R. Steele*, 1994, p. Chapter 2. 30

[69] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw Hill, 1991. 31

[70] J. Wozencraft and I. Jacobs, *Principles of Communication Engineering*. Wiley, 1965. 32

[71] T. Moon and W. Sterling, *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, 1999. 34

[72] F. Farrokhi, A. L. G. Foschini, and R. Valenzuela, "Spectral efficiency of wireless systems with multiple transmit and receive antennas," in *The 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, 2000, pp. 373–377. 35

[73] C. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, 1948. 38

[74] N. Abramson, *Information Theory and Coding*. New York: McGraw Hill, 1963. 38

[75] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 1991. 38, 39

[76] M. Purser, *Introduction to Error-correcting Codes*. Artech House, 1995. 38, 39, 40

[77] S. G. Wilson, *Digital Modulation and Coding*. Prentice Hall, 1996. 38

[78] I. Glover and P. Grant, *Digital Communications*, 2nd ed. Prentice Hall, 2003. 38

[79] T. Moon, *Error Correction Coding*. Wiley, 2005. 38

[80] W. Peterson and E. Weldon, *Error-Correcting Codes*. Cambridge, Mass: MIT Press, 1972. 40

[81] A. Viterbi and J. Omura, *Principles of Digital Communication and Coding*. New York: McGraw Hill, 1979. 40

[82] A. Michelson and A. Levesque, *Error-Control Techniques for Digital Communication*. New York: Wiley-Interscience, 1985. 40

[83] J. Cioffi, *http://www.stanford.edu/class/ee379a*. Class Reader for EE379a-Digital Communication:Signal Processing, Stanford University, Stanford, CA, 2002. 42

[84] R. Lucky, "Automatic equalization for digital communications," in *Bell Syst. Tech. J.*, vol. 44, 1965, pp. 547–588. 44, 45

[85] D. George, R. Bowen, and J. Storey, "An adaptive decision-feedback equalizer," in *IEEE Transactions on Communication Technology*, vol. COM-19, 1971, pp. 281–293. 45

[86] R. Price, "Nonlinearly feedback-equalized PAM vs. capacity," *Proceedings of IEEE international conference on communications*, pp. 22–12–22–17, 1972. 45

[87] J. Saltz, "Optimum mean-square decision feedback equalization," *Bell Syst. Tech. J.*, vol. 52, pp. 1341–1373, 1973. 45

[88] J. Proakis, "Advances in equalization for intersymbol interference," *Advances in communication systems*, vol. 4, 1975, a. J. Viterbi (ed.). 45

[89] R. Lucky, "Techniques for adaptive equalization of digital communication," in *Bell Syst. Tech. J.*, vol. 45, 1966, pp. 255–286. 45

[90] B. Widrow, "Adaptive filters, I: Fundamentals," in *Tech Report, No. 6764-6*, Stanford University, Stanford, CA, Dec. 1966. 45

[91] ——, "Adaptive filters," in *Aspects of Network and System Theory*, vol. R. E. Kalman and N. Declairs (eds.), Holt, Rinehart and Winston, New York, 1970. 45

[92] S. Hara and R. Prasad, *Multicarrier Techniques for 4G mobile communications*. Artech House, 2003. 45

[93] R. W. Chang, "Synthesis of band-limited orthogonal signals for multichannel data transmission," *Bell Syst. Tech. J.*, vol. 45, Dec 1966. 46

[94] R. Gray, *Toeplitz and Circulant Matrices: A Review*. Now Publishers, 2006. 47

[95] A. Bahai and B. Saltzberg, *Multi-carrier Digital Communications- Theory and Applications of OFDM*. Kluwer Academic Publishers, 2002. 48

[96] J. Tellado and J. Cioffi, "Efficient algorithms for reducing par in multicarrier systems," in *Proceedings of IEEE International Symposium on Information Theory*, 1998, p. 191. 48

[97] S. Boyd, "Multitone signals with low crest factor," in *IEEE Transactions on Circuits and Systems*, vol. CAS-33, 1986, pp. 1018–1022. 48

[98] B. Popvic, "Synthesis of power efficient multitone signals with flat amplitude spectrum," *IEEE Transactions on Communications*, vol. 39, pp. 1031–1033, 1991. 48

[99] R. Ziemer and R. Peterson, *Digital Communications and Spread Spectrum Systems*. Macmillan Publishing Company, 1985. 48, 51, 52

[100] A. Viterbi, *CDMA: Principles of Spread Spectrum Communication*. Wiley, 1995. 48

[101] R. Pickholtz, D. Schilling, and L. Milstein, "Theory of spread-spectrum communications - a tutorial," *IEEE Transactions on Communications*, vol. 30, no. 5, pp. 855–884, 1982, part 2. 49

[102] ——, "Revisions to 'theory of spread-spectrum communications-a tutorial'," *IEEE Transactions on Communications*, vol. 32, no. 2, pp. 211–212, 1984. 49

[103] J. S. Lee and L. E. Miller, *CDMA Systems Engineering Handbook*. Artech House, 1998. 50

[104] D. Sarwate and M. Pursley, "Crosscorrelation properties of pseudorandom and related sequences," *Proceedings of the IEEE*, vol. 68, no. 5, pp. 593–619, 1980. 51

[105] ——, "Correction to 'crosscorrelation properties of pseudorandom and related sequences'," *Proceedings of the IEEE*, vol. 68, no. 12, pp. 1554–1554, 1980. 51

[106] R. Gold, "Optimal binary sequences for spread spectrum multiplexing," in *IEEE Transactions on Information Theory*, vol. IT-13, 1967, pp. 619–621. 51

[107] ——, "Maximal recursive sequences with 3-valued recursive cross correlation functions," in *IEEE Transactions on Information Theory*, vol. IT-14, 1968, pp. 154–156. 51

[108] T. Kasami, "Weight distribution formula for some class of cyclic codes," in *UIUC Coordinated Science Lab Tech Report*, vol. No. R-285, 1966. 51

[109] S. Golomb, *Shift Register Sequences*. Holden-Day, 1967. 51

[110] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, pp. 311–335, 1998. 53

[111] E. Telatar, "Capacity of multi-antenna gaussian channels," in *European Transactions on Telecommunications*, vol. 10, Nov./Dec. 1999, pp. 585–595. 53

[112] A. J. Paulraj, D. Gesbert, and C. Papadias, "Smart antennas for mobile communications," *The Encyclopedia of Electrical and Electronics Engineering*, vol. J. Webster Ed, 2000, wiley. 53

[113] D. Gesbert and J. Akhtar, "Breaking the barriers of Shannon's capacity: An overview of MIMO wireless system," in *Telektronikk*, vol. 98, 2002, pp. 53–64. 53

[114] D. Gesbert, M. Shafi, D. Shiu, P. Smith, and A. Naguib, "From theory to practice: An overview of MIMO space-time coded wireless systems," in *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 3, 2003, pp. 281–302. 53

[115] A. J. Paulraj and C. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Magazine*, vol. 14, no. 6, pp. 49–83, 1997, submitted. 53

[116] J. Parsons, *The Mobile Radio Propagation Channel*. Wiley, 2000. 54, 55

[117] D. G. Brennan, "Linear diversity combining techniques," *Proceedings of the IRE*, vol. 47, no. 6, pp. 1075–1102, 1959. 54

[118] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999. 55, 58, 113

[119] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," in *IEEE Transactions on Information Theory*, vol. 48, no. 2, 2002, pp. 359–383. 56

[120] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998. 56

[121] S. Tavilar and P. Viswanath, "Approximately universal codes over slow-fading channels," *IEEE Transactions on Information Theory*, vol. 52, no. 7, pp. 3233–3258, July 2006. 58

[122] L. Nuaymi, *WiMAX : Technology for Broadband Wireless Access*. Wiley, 2007. 59

[123] G. Foschini, "Layered space-time architecture for wireless communications in a fading environment when using multiple antennas," *Bell Labs. Tech. J.*, vol. 6, no. 2, pp. 41–59, 1996. 60

[124] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *International Symposium on Signals, Systems, and Electronics*, September 1998, pp. 295–300. 62

[125] G. Golden, G. Foschini, R. Valenzuela, and P. Wolniansky, "Detection algorithm and initial laboratory results using the V-BLAST spacetime communication architecture," in *Electron. Lett.*, vol. 35, no. 1, 1999, pp. 14–15. 62

[126] G. Foschini, G. Golden, P. Wolniansky, and R. Valenzuela, "Simplified processing for wireless communication at high spectral efficiency," *IEEE Journal on Selected Areas in Communications-Wireless Communication Series*, vol. 17, pp. 1841–1852, 1999. 62

[127] IEEE., "Standard 802.16e-2005. part16: Air interface for fixed and mobile broadband wireless access systemsamendment for physical and medium access control layers for combined fixed and mobile operation in licensed band," Dec 2005. 64

[128] H. A. David, *Order Statistics*. Wiley, 1970. 71

[129] S. Lloyd, "Least squares quantization in PCM," in *IEEE Transactions on Information Theory*, vol. 28, 1982, pp. 129–137. 76

[130] J. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX-Understanding Broadband Wireless Networking*. Prentice Hall, 2007. 77, 93, 94

[131] D. Piazza and L. Milstein, "Multiuser diversity-mobility tradeoff: Modeling and performance analysis of a proportional fair scheduling," in *IEEE Global Telecommunications Conference*, vol. 1, 2002, pp. 906–910. 86

[132] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics, 2003. 88

[133] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. Applied Mathematics Series 55. New York: National Bureau of Standard, 1964. 88, 89, 90

[134] M. Simon and M. Alouini, *Digital Communication over Fading Channels*. Wiley, 2004. 90, 97, 164, 167

[135] R. Millane and J. Eads, "Polynomial approximations to Bessel functions," in *IEEE Transactions on Antennas and Propagation*, vol. 51, Jun. 2003, pp. 1398 – 1400. 103

[136] Z. Han, A. L. Swindlehurst, and K. J. R. Liu, "Smart deployment/movement of unmanned air vehicle to improve connectivity in MANET," in *IEEE Wireless Communications and Networking Conference*, 2006, submitted. 109

[137] P. Zhan, K. Yu, and A. Swindlehurst, "Wireless relay communications using an unmanned aerial vehicle," in *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2006, pp. 1–5. 109, 117

[138] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, 2005. 109

[139] H. Bölcskei, R. U. Nabar, O. Oyman, and A. J. Paulraj, "Capacity scaling laws in MIMO relay networks," *IEEE Transactions on Wireless Communications*, 2006, to appear. 109

[140] D.-S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, March 2000. 112, 175

[141] K. Yu, M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson, and M. Beach, "Modeling of wideband MIMO radio channels based on NLOS indoor measurements," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 655–665, May 2004. 112

[142] R. U. Nabar, H. Bölcskei, and A. J. Paulraj, "Outage properties of space-time block codes in correlated Rayleigh or Ricean fading environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2002, pp. 2381–2384. 114

[143] R. B. Ertel, P. Cardieri, K. W. Sowerby, T. S. Rappaport, and J. H. Reed, "Overview of spatial channel models for antenna array communication systems," *IEEE Personal Communications*, vol. 5, no. 1, pp. 10–22, February 1998. 119

[144] C. Perkins, *Ad Hoc Networking*. Addison-Wesley, 2001. 127

[145] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Berkeley, University of California Press*, vol. 1, pp. 281–297, 1967. 128

[146] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002, pp. 881–892. 128

[147] C. Ding and X. He, "K-means clustering via principal component analysis," *Proc. of Int'l Conf. Machine Learning (ICML)*, pp. 225–232, July 2004. 128

[148] A. Vakulenko, "Overlapping region of the collection of circles," *http://www.mathworks.com/matlabcentral/fileexchange/*, 2005. 132

[149] M. S. Heutmaker, "The error vector and power amplifier distortion," in *IEEE Wireless Communications Conference*, 1997, pp. 100–104. 146

[150] R. Hassun, M. Flaherty, R. Matreci, and M. Taylor, "Effective evaluation of link quality using error vector magnitude techniques," in *IEEE Wireless Communications Conference*, Oct. 1997, pp. 89–94. 146

[151] A. Mashhour and A. Borjak, "A method for computing error vector magnitude in gsm edge systems-simulation results," in *IEEE Communications Letters*, vol. 5, 2001, pp. 88–91. 146

[152] A. Georgiadis, "Gain, phase imbalance, and phase noise effects on error vector magnitude," in *IEEE Transactions on Vehicular Technology*, vol. 53, no. 2, 2004, pp. 443–449. 146

[153] M. O'Droma and N. Mgebrishvili, "On quantifying the benefits of SSPA linearization in UWC-136 systems," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2470–2476, 2005. 146

[154] A. Haider and A. Chatterjee, "Low-cost alternate EVM test for wireless receiver systems," in *Proceedings of the IEEE VLSI Test Symposium*, 2005, pp. 352–498. 146

[155] R. Liu, Y. Li, H. Chen, and Z. Wang, "EVM estimation by analyzing transmitter imperfections mathematically and graphically," in *Analog Integrated Circuit Signal Processing*, vol. 48, 2006, pp. 257–262. 146

[156] M. Simon and M.-S. Alouini, "A unified approach to the performance analysis of digital communication over generalized fading channels," *Proceedings of The IEEE*, vol. 86, no. 9, pp. 1860–1877, 1998. 173

[157] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 840 – 845, July 2003. 174

# Appendix A

# Derivation for Correctly Supported Spectral Efficiency

## A.1 Derivation of PCSI Spectral Efficiency

The spectral efficiency for the PCSI case can be derived as follows:

$$\begin{aligned}
\bar{T} &= \mathcal{E}\left(C_j(1-\mathrm{PER}(j))\cdot\mathbf{1}\ (\gamma_j \le \gamma < \gamma_{j+1})\right) \\
&= \sum_{j=0}^{N} C_j \int_{\gamma_j}^{\gamma_{j+1}} (1-\mathrm{PER}(j))\cdot f_{\gamma_{\max}}(\gamma)d\gamma \\
&= \sum_{j=0}^{N-1} C_j \left(F_{\gamma_{\max}}(U_j) - F_{\gamma_{\max}}(L_j)\right) + C_N\cdot\left(1-F_{\gamma_{\max}}(L_N)\right) \\
&- \sum_{j=0}^{N-1} C_j \sum_{l=0}^{K} J_l \cdot e^{b_j\cdot\gamma_{T,j}}\cdot W_{j,l}\left(e^{-\frac{U_j}{W_{j,l}}} - e^{-\frac{L_j}{W_{j,l}}}\right) \\
&+ C_N \sum_{l=0}^{K} J_l \cdot e^{b_j\cdot\gamma_{T,j}}\cdot W_{N,l} e^{-\frac{L_N}{W_{N,l}}}\,, \qquad \text{(A.1)}
\end{aligned}$$

where

$$J_l \equiv (-1)^{l+1}\frac{l}{\bar{\gamma}}\begin{pmatrix} K \\ l \end{pmatrix}$$

and

$$W_{j,l} \equiv \frac{1}{b_j + \frac{l}{\bar{\gamma}}}\,.$$

## A.2 Derivation of ICSI-PEFD Spectral Efficiency

If we recall (4.12), two integrals need to be calculated to evaluate the spectral efficiency, i.e. $\int_{\gamma_j}^{\gamma_{j+1}} f_{\hat{\gamma}_{\max}}(\gamma_h)\cdot\widetilde{\mathrm{PSR}}(j)\cdot d\gamma_h$ when $0 \le j \le N-1$, and $\int_{\gamma_N}^{\infty} f_{\hat{\gamma}_{\max}}(\gamma_h)\widetilde{\mathrm{PSR}}(N)\cdot$

$d\gamma_h$ otherwise. Let's define:

$$F(a,b,c) \equiv \int_0^a e^{-x} Q_1\left(\sqrt{bx}, c\right) dx\,. \tag{A.2}$$

According to [134], the Marcum $Q$ function can be expanded as:

$$Q_1\left(\sqrt{bx}, c\right) = \sum_{n=0}^{\infty} e^{-\frac{bx}{2}} \frac{1}{n!} \left(\frac{bx}{2}\right)^n \sum_{k=0}^{n-1} e^{-\frac{c^2}{2}} \frac{1}{k!} \left(\frac{c^2}{2}\right)^k . \tag{A.3}$$

We have:

$$F(a,b,c) = \sum_{n=0}^{\infty} \frac{m_n}{\left(1+\frac{b}{2}\right)^{n+1}} \Gamma_{Inc}\left(a\left(1+\frac{b}{2}\right), n+1\right) , \tag{A.4}$$

where

$$m_n = \left(\frac{b}{2}\right)^n \left\{ \sum_{k=0}^{n} e^{-\frac{c^2}{2}} \cdot \frac{1}{k!} \cdot \left(\frac{c^2}{2}\right)^k \right\} ,$$

$$\Gamma_{Inc}(a,n) = \int_0^a \frac{e^{-t} \cdot t^{n-1}}{\Gamma(n)} dt$$

is the incomplete Gamma function, and

$$\Gamma(n) = \int_0^{\infty} e^{-t} \cdot t^{n-1} dt\,.$$

Using (A.4), the first integral is expressed in closed form as follows:

$$\int_{\gamma_j}^{\gamma_{j+1}} f_{\hat{\gamma}_{\max}}(\gamma_h) \cdot \widetilde{\mathrm{PSR}}(j) \cdot d\gamma_h = \sum_{l=0}^{K} (-1)^{l+1} \binom{K}{l} \left\{ \left[ F\left(\frac{l\cdot\gamma_{j+1}}{\bar{\gamma}_h}, \frac{p_{1,j}}{l}, q_{1,j}\right) \right.\right.$$
$$\left.\left. - F\left(\frac{l\cdot\gamma_j}{\bar{\gamma}_h}, \frac{p_{1,j}}{l}, q_{1,j}\right)\right] - t_j \left[ F\left(\frac{l\cdot\gamma_{j+1}}{\bar{\gamma}_h}, \frac{p_{2,j}}{l}, q_{2,j}\right) - F\left(\frac{l\cdot\gamma_j}{\bar{\gamma}_h}, \frac{p_{2,j}}{l}, q_{2,j}\right)\right] \right\} . \tag{A.5}$$

Now we define:

$$\mathbf{G}(a,b,c) \equiv \int_a^{\infty} e^{-x} \cdot Q\left(\sqrt{bx}, c\right) \cdot dx\,. \tag{A.6}$$

Using similar reasoning, and the same definition for $m_n$, we have:

$$\mathbf{G}(a,b,c) = \sum_{n=0}^{\infty} \frac{m_n}{\left(1+\frac{b}{2}\right)^{n+1}} \Gamma_{Inc}^{C}\left(a\left(1+\frac{b}{2}\right), n+1\right), \tag{A.7}$$

where

$$\Gamma_{Inc}^{C}(a,n) = \int_a^{\infty} \frac{e^{-t}\cdot t^{n-1}}{\Gamma(n)} dt$$

is the tail of the incomplete Gamma function. With the above definitions, after some mathematical simplification, the second integral can be expressed as:

$$\begin{aligned}\int_{\gamma_N}^{\infty} f_{\hat{\gamma}_{\max}}(\gamma_h)\widetilde{\mathrm{PSR}}(N)d\gamma_h = \sum_{l=0}^{K}(-1)^{l+1}\binom{K}{l}\cdot\Bigg[G\left(\frac{l\gamma_N}{\bar{\gamma}_h}, \frac{p_{1,N}}{l}, q_{1,N}\right) \\ - t_N\cdot G\left(\frac{l\gamma_N}{\bar{\gamma}_h}, \frac{p_{2,N}}{l}, q_{2,N}\right)\Bigg].\end{aligned} \tag{A.8}$$

### A.3 Derivation of ICSI-EQFS Spectral Efficiency

Recalling (4.17), we can write:

$$\begin{aligned}&\mathcal{E}\left\{(1-\mathrm{PER}(j,\gamma_g,\gamma_h,\hat{\gamma}_h))\cdot\mathbf{1}(\gamma_j\le\hat{\gamma}_h<\gamma_{j+1})\right\}\\ &=\int_0^{\infty}\int_0^{\infty}\widetilde{\mathrm{PSR}}(j,\gamma_h)\cdot\mathbf{1}(\gamma_j\le\hat{\gamma}_h<\gamma_{j+1})f_{\gamma_h,\hat{\gamma}_h}(\gamma_h,\hat{\gamma}_h)d\gamma_h d\hat{\gamma}_h\,.\end{aligned} \tag{A.9}$$

After some tedious mathematical simplifications, we have:

$$\begin{aligned}&\mathcal{E}\left\{(1-\mathrm{PER}(j,\gamma_g,\gamma_h,\hat{\gamma}_h))\cdot\mathbf{1}(\gamma_j\le\hat{\gamma}_h<\gamma_{j+1})\right\}\\ &=\sum_{l=0}^{K}(-1)^{l+1}\frac{l}{\bar{\gamma}_h}\binom{K}{l}\int_{\gamma_j'}^{\gamma_{j+1}'}\widetilde{PSR}(j,\gamma_h)\cdot e^{-\frac{l\gamma_h}{\bar{\gamma}_h}}d\gamma_h\\ &=\sum_{l=0}^{K}(-1)^{l+1}l\binom{K}{l}\int_{\frac{\gamma_j'}{\bar{\gamma}_h}}^{\frac{\gamma_{j+1}'}{\bar{\gamma}_h}}\widetilde{PSR}(j,\beta)\cdot e^{-l\beta}d\beta\,,\end{aligned} \tag{A.10}$$

where $\beta = \frac{|h|^2}{\Omega_h}$, $\gamma'_j$ and $\gamma'_{j+1}$ are defined as follows. Let $\hat{\gamma}_j$ denote the smallest quantized value that is greater than or equal to $\gamma_j$, and $\hat{\gamma}_{j+1}$ denote the largest quantized value that is smaller than or equal to $\gamma_{j+1}$. Then $\gamma'_j$ is defined as the lower bound of the quantization range for $\hat{\gamma}_j$, and $\gamma'_{j+1}$ is defined as the upper bound of the quantization range for $\hat{\gamma}_{j+1}$. If we recall (A.5), when $j \neq N$, we have:

$$\begin{aligned}
&\mathcal{E}\left\{(1-\mathrm{PER}(j,\gamma_g,\gamma_h,\hat{\gamma}_h))\cdot\mathbf{1}(\gamma_j\leq\hat{\gamma}_h<\gamma_{j+1})\right\} \qquad \text{(A.11)}\\
&=\sum_{l=0}^{K}(-1)^{l+1}\begin{pmatrix}K\\ l\end{pmatrix}\cdot\left\{\left[F\left(\frac{l\cdot\gamma'_{j+1}}{\bar{\gamma}_h},\frac{p_{1,j}}{l},q_{1,j}\right)-F\left(\frac{l\cdot\gamma'_j}{\bar{\gamma}_h},\frac{p_{1,j}}{l},q_{1,j}\right)\right]\right.\\
&\left.-t_j\left[F\left(\frac{l\cdot\gamma'_{j+1}}{\bar{\gamma}_h},\frac{p_{2,j}}{l},q_{2,j}\right)-F\left(\frac{l\cdot\gamma'_j}{\bar{\gamma}_h},\frac{p_{2,j}}{l},q_{2,j}\right)\right]\right\}.
\end{aligned}$$

Recall (A.8), when $j = N$, we have:

$$\begin{aligned}
&\mathcal{E}\left\{(1-\mathrm{PER}(N,\gamma_g,\gamma_h,\hat{\gamma}_h))\cdot\mathbf{1}(\gamma_N\leq\hat{\gamma}_h<\infty)\right\} \qquad \text{(A.12)}\\
&=\sum_{l=0}^{K}(-1)^{l+1}\begin{pmatrix}K\\ l\end{pmatrix}\cdot\left[G\left(\frac{l\gamma'_N}{\bar{\gamma}_h},\frac{p_{1,N}}{l},q_{1,N}\right)-t_N\cdot G\left(\frac{l\gamma'_N}{\bar{\gamma}_h},\frac{p_{2,N}}{l},q_{2,N}\right)\right].
\end{aligned}$$

# Appendix B

# Derivations for the Shannon Capacity of a MUD System

## B.1 Shannon Capacity for the Max-SNR Scheduler with Channel Tracking for the Scheduled User

In order to calculate (5.10), $C_{j|h_j(0)}(t)$ needs to be evaluated:

$$\begin{aligned} C_{j|h_j(0)}(t) &= \int_0^\infty \log_2(1+x) \cdot f_{\gamma_j(t)|h_j(0)}(x) \cdot dx \\ &= \log_2(e) \cdot \int_0^\infty \frac{1}{1+x} \cdot \mathcal{Q}_1(\frac{s_j(t)}{\sigma_j(t)}, \frac{\sqrt{x}}{\sigma_j(t)}) \cdot dx \,. \end{aligned} \tag{B.1}$$

According to [134], we have:

$$\mathcal{Q}_1(\frac{s}{\sigma}, \frac{\sqrt{x}}{\sigma}) = \sum_{n=0}^{\infty} e^{-\frac{s^2}{2\sigma^2}} \cdot \frac{(\frac{s^2}{2\sigma^2})^n}{n!} \cdot \sum_{k=0}^{n} e^{-\frac{x}{2\sigma^2}} \cdot \frac{(\frac{x}{2\sigma^2})^k}{k!} \,. \tag{B.2}$$

Plugging (B.2) in (B.1), $C_{j|h_j(0)}(t)$ can be written out explicitly as:

$$C_{j|h_j(0)}(t) = \log_2 e \cdot \sum_{n=0}^{\infty} e^{-\frac{s_j^2(t)}{2\sigma_j^2(t)}} \cdot \frac{\left(\frac{s_j^2(t)}{2\sigma_j^2(t)}\right)^n}{n!} \cdot \sum_{k=0}^{n} F(\alpha_j(t), k) \,. \tag{B.3}$$

Let us define:

$$C_j(t) = \int_0^\infty C_{j|h_j(0)}(t) p_{\eta_j(0),j}(x, j, t=0) dx \,. \tag{B.4}$$

By substituting (B.3) in the integral of (B.4), we have:

$$C_j(t) = \frac{\log_2(e)}{\overline{\gamma_j}(0)} \sum_{i=1, i\neq j}^{K} \sum_{l_i=0}^{1} (-1)^{\sum_{i=1, i\neq j}^{K} l_i} \times \sum_{n=0}^{\infty} \frac{(b_j^t(0))^n}{(\sum_{i=1, i\neq j}^{K} \frac{l_i}{\overline{\gamma_i}(0)} + \frac{1}{\overline{\gamma_j}(0)} + b_j^t(0))^{n+1}} \sum_{k=0}^{n} F(\alpha_j(t), k)\,. \tag{B.5}$$

With (B.5) obtained above, some algebraic simplifications will lead to (5.10).

## B.2 Shannon Capacity for the Max-SNR Scheduler without Channel Tracking for the Scheduled User

If we recall (5.8), and the Shannon capacity of the system can be expressed as:

$$\begin{aligned} C(t) &= \sum_{j=1}^{K} \int_0^{\infty} \log_2(1+\phi_j(t)x) p_{\gamma,j}(x, j, t=0) dx \\ &= \sum_{j=1}^{K} \int_0^{\infty} \frac{\partial \log_2(1+\phi_j(t)x)}{\partial x} \cdot \mathrm{CCDF}(x, j) dx \\ &= \log_2 e \sum_{j=1}^{K} \int_0^{\infty} \frac{\phi_j(t)}{1+\phi_j(t)x} \cdot \mathrm{CCDF}(x, j) dx\,, \end{aligned} \tag{B.6}$$

where $\mathrm{CCDF}(x, j)$ is the Complementary Cumulative Distribution Function, and is defined as:

$$\mathrm{CCDF}(x, j) = \int_x^{\infty} p_{\gamma,j}(y, j, t=0) dy\,. \tag{B.7}$$

After some mathematical manipulations, we have:

$$\mathrm{CCDF}(x, j) = \frac{1}{\overline{\gamma_j}(0)} \sum_{i\neq j, i=1}^{K} \sum_{l_i=0}^{1} (-1)^{\sum_{i=0}^{K} l_i} \frac{1}{\frac{1}{\overline{\gamma_j}(0)} + \sum_{i\neq j, i=1}^{K} \frac{l_i}{\overline{\gamma_i}(0)}} e^{-\left[\frac{1}{\overline{\gamma_j}(0)} + \sum_{i\neq j, i=1}^{K} \frac{l_i}{\overline{\gamma_i}(0)}\right] x}\,. \tag{B.8}$$

Plugging (B.8) in (B.6), by changing the integration variable, (5.13) is obtained.

## B.3 Monotonicity of $e^x\mathcal{E}_1(x)$

The monotonicity of $e^x\mathcal{E}_1(x)$ can be shown by taking the derivative of it with respect to $x$ as follows:

$$\frac{\partial[e^x\mathcal{E}_1(x)]}{\partial x} = e^x\mathcal{E}_1(x) + e^x\frac{\partial\mathcal{E}_1(x)}{\partial x}\,. \tag{B.9}$$

If we recall the definition of $\mathcal{E}_1(x)$ in Section 5.2.1, by changing the integration variable, $\mathcal{E}_1(x)$ can be rewritten as:

$$\mathcal{E}_1(x) = \int_1^\infty \frac{e^{-tx}}{t}dt\,. \tag{B.10}$$

Plugging (B.10) in (B.9), we have:

$$\begin{aligned}\frac{\partial[e^x\mathcal{E}_1(x)]}{\partial x} &= e^x\mathcal{E}_1(x) + e^x\int_1^\infty \frac{1}{t}(-t)e^{-xt}dt \\ &= e^x\int_1^\infty \frac{1}{t}(1-t)e^{-xt}dt < 0\,.\end{aligned} \tag{B.11}$$

The inequality in (B.9) is due to the fact that the limit of the integration variable suggests $t \geq 1$. Hence in the range over which integration is performed, the integrand is always less than or equal to 0.

## B.4 Shannon Capacity for SCA Schedulers

SCA schedulers select users according to their long term channel statistics regardless how good their instantaneous channel conditions are. Hence, if user $j$ is selected, the Shannon capacity of the system can be written as:

$$\begin{aligned}C_j(t) &= \mathsf{E}\left(\log_2\left(1+\phi_j(t)\eta_j(0)\right)\right) \\ &= \int_0^\infty \log_2\left(1+\phi_j(t)x\right)\frac{1}{\overline{\overline{\gamma_j(0)}}}e^{-\frac{x}{\overline{\overline{\gamma_j(0)}}}}dx\,.\end{aligned} \tag{B.12}$$

The Rayleigh assumption on each user's channel magnitude infers that each user's SNR has an exponential distribution, therefore leads to the above equality. The CCDF of $j$-th user's SNR can be easily shown as $\text{CCDF}(x,j) = e^{-\frac{x}{\overline{\gamma_j(0)}}}$. (B.12) can be reorganized as follows:

$$\begin{aligned} C_j(t) &= \mathsf{E}\left(\log_2\left(1+\phi_j(t)\eta_j(0)\right)\right) \\ &= \int_0^\infty \frac{\partial \log_2\left(1+\phi_j(t)x\right)}{\partial x} e^{-\frac{x}{\overline{\gamma_j(0)}}} dx \\ &= \log_2 e \int_0^\infty \frac{\phi_j(t)}{1+\phi_j(t)x} e^{-\frac{x}{\overline{\gamma_j(0)}}} dx\,. \end{aligned} \tag{B.13}$$

By letting $y = 1+\phi_j(t)x$, the above integral is equivalent to:

$$C_j(t) = \log_2 e \int_1^\infty \frac{\phi_j(t)}{y} e^{-\frac{y-1}{\phi_j(t)\overline{\gamma_j(0)}}} \frac{dy}{\phi_j(t)}\,. \tag{B.14}$$

Let $z = \frac{y}{\phi_j(t)\overline{\gamma_j(0)}}$, $C_j(t)$ can be further simplified as:

$$\begin{aligned} C_j(t) &= \log_2 e \times e^{\frac{1}{\phi_j(t)\overline{\gamma_j(0)}}} \int_{\frac{1}{\phi_j(t)\overline{\gamma_j(0)}}}^\infty \frac{1}{z} e^{-z} dz \\ &= \log_2 e \times e^{\frac{1}{\phi_j(t)\overline{\gamma_j(0)}}} \mathcal{E}_1\left(\frac{1}{\phi_j(t)\overline{\gamma_j(0)}}\right)\,. \end{aligned} \tag{B.15}$$

Plugging (5.12) in (B.15), (5.20) can be derived.

## B.5 Shannon Capacity for the Quasi PF Scheduler with Channel Tracking for the Scheduled User

The Shannon capacity under this assumption can be obtained in a similar manner as Section B.1

$$C(t) = \sum_{j=1}^{K} \int_0^\infty C_{j|h_j(0)}(t) p(z_j \text{ is max},\ z_j = x) dx\,, \tag{B.16}$$

where $z_j = \gamma_j(0)/\overline{\gamma_j(0)}$ is the normalized instantaneous SNR of the $j$-th user, and therefore is a function of $h_j(0)$. $C_{j|h_j(0)}(t)$ under this assumption has already been evaluated in (B.3). Plugging (B.3) and (5.14) in (B.16), we have:

$$\begin{aligned} & C(t) = \log_2 e \times \\ & \sum_{j=1}^{K} \left\{ \sum_{n=0}^{\infty} \frac{1}{n!} \left( \int_0^{\infty} \left( b_j \overline{\gamma_j(0)} \right)^n x^n e^{-b_j \overline{\gamma_j(0)} x} e^{-x} \left(1 - e^{-x}\right)^{K-1} dx \right) \cdot \sum_{m=0}^{n} F(\alpha_j, m) \right\} \\ & = \log_2 e \sum_{j=1}^{K} \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{l=0}^{K-1} \left( \begin{array}{c} K-1 \\ l \end{array} \right) (-1)^l \frac{n! \beta_j^n}{(\beta_j + l + 1)^{n+1}} \sum_{m=0}^{n} F(\alpha_j, m) \,, \end{aligned} \tag{B.17}$$

where $\beta_j = b_j \overline{\gamma_j(0)}$ and the integration in the bracket is evaluated using binomial expansion. Reorganizing the terms in the above equation yields (5.18).

## B.6 Shannon Capacity for the Quasi PF Scheduler without Channel Tracking for the Scheduled User

If we recall (5.11), we know when the scheduled user is not tracking its own channel for decoding purpose, its SNR at any time $t$ within coherence time is in proportion to its SNR at time $t = 0$. A revisit of (5.15) gives us:

$$\begin{aligned} C(t) &= \sum_{j=1}^{K} \mathcal{E} \left\{ 1 + \gamma_j(t) \right\} \\ &= \sum_{j=1}^{K} \int_0^{\infty} \log_2 \left( 1 + \phi_j(t) \overline{\gamma_j(0)} x \right) e^{-x} \left(1 - e^{-x}\right)^{K-1} dx \,. \end{aligned} \tag{B.18}$$

A close observation reveals that (B.18) resembles (5.15) with the only difference that $\overline{\gamma_j}$ is replaced with $\phi_j(t)\overline{\gamma_j(0)}$. Therefore (5.19) can be readily obtained from (5.16) by replacing $\overline{\gamma_j}$ with $\phi_j(t)\overline{\gamma_j(0)}$.

# Appendix C

# Derivation for Link Performance and Channel Norm CDF

## C.1 Link Level SER Analysis

This section is dedicated to the derivation of the SER analysis at the link level. Once the error analysis for each link has been performed, the SER of the whole system can be calculated.

### C.1.1 Closed Form SER Expression

The SER can be expressed as in (C.1):

$$P_s = \sum_{i=1}^{N-1} \int_{C_i(t)}^{C_{i+1}(t)} \overline{N}_e(i) \cdot Q(\sqrt{\frac{x\rho d^2_{\min}(i)}{2}}) f(x)dx + \int_{C_N(t)}^{\infty} \overline{N}_e(N) \cdot Q(\sqrt{\frac{x\rho d^2_{\min}(i)}{2}}) f(x)dx\,. \tag{C.1}$$

In [156], an alternative definite integral form for the Gaussian Q-function is given as

$$Q(x) = \frac{1}{\pi} \int_0^{\frac{\pi}{2}} \exp\left(-\frac{x^2}{2\sin^2\theta}\right) d\theta, \quad x \geq 0. \tag{C.2}$$

Using this alternative form and interchanging the order of the integrations, the SER can be rewritten as in (C.3). Recalling the definition in (6.9), it is straightforward to

derive the SER expression given in (6.14):

$$P_s = \frac{1}{\pi}(\sum_{i=1}^{N-1}\int_0^{\frac{\pi}{2}}\int_{C_i(t)}^{C_{i+1}(t)} \overline{N}_e(i)\cdot\exp(-\frac{x\rho d^2_{\min}(i)}{4\sin^2\theta})f(x)dx\cdot d\theta$$
$$+\int_0^{\frac{\pi}{2}}\int_{C_N(t)}^{\infty} \overline{N}_e(N)\cdot\exp(-\frac{x\rho d^2_{\min}(i)}{4\sin^2\theta})f(x)dx\cdot d\theta)\,. \tag{C.3}$$

### C.1.2 SER Upper Bound

In order to relieve the computational burden when evaluating (6.14), an upper bound for the SER is derived by resorting to the results of [157]. In Chiani's work, an improved exponential bound for the Q function is given as:

$$Q(x) \le \frac{1}{2}\sum_{i=1}^{N} a_i \exp(-\frac{b_i x^2}{2}), \tag{C.4}$$

where

$$a_i = \frac{2(\theta_i - \theta_{i-1})}{\pi} \tag{C.5}$$

and

$$b_i = \frac{1}{\sin^2\theta_i}. \tag{C.6}$$

Note that this bound is much better than the popular Chernoff bound. After some manipulation, the upper bound for the SER is found to be given by:

$$P_s \le \sum_{i=1}^{N-1}\frac{\overline{N}_e(i)}{2}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\sum_{n=1}^{Q}\frac{a_n A_{jk}}{\sigma_j^k}[g(k-1,-(\frac{b_n\rho d^2_{\min}(N)}{4}+\frac{1}{\sigma_j}),C_{i+1}(t))$$
$$-g(k-1,-(\frac{b_n\rho d^2_{\min}(N)}{4}+\frac{1}{\sigma_j}),C_i(t))]$$
$$-\frac{\overline{N}_e(N)}{2}\sum_{j=1}^{p}\sum_{k=1}^{m_j}\sum_{n=1}^{Q}\frac{a_n A_{jk}}{\sigma_j^k}g(k-1,-(\frac{b_n\rho d^2_{\min}(N)}{4}+\frac{1}{\sigma_j}),C_N(t))\,. \tag{C.7}$$

As we can see in Fig. 6.2, when $N$ increases, the SER bound closely approaches the theoretical value.

## C.2 Approximation of $F(y)$ and the Rate $R(t)$

In this section, we show that the CDF of the Frobenius norm of the channel, $F(y)$, can be approximated by a sinusoid under certain assumptions. Let us assume a one ring model scenario [140] (i.e., the APs are surrounded by the effective scatterers on a ring, and the UAV has no scatterers around it), and a Kronecker structure for the channel correlation matrix (6.4). Under such assumptions, the channel is ill-conditioned with only one dominant eigen-mode. Assuming there is no spatial correlation at the APs, the channel correlation matrix $R_H^k$ between the UAV and the $k$th AP has only one distinct non-zero eigenvalue $\sigma$ with multiplicity $m$, where $m$ is the number of antennas at the AP side. Therefore, the Laplace transform of the pdf of $\|\mathbf{H}_{\text{norm}}\|_F^2$ can be expressed as

$$\psi(s) = \frac{1}{(1+\sigma)^m}, \tag{C.8}$$

and the CDF can be written as

$$F(y) = \left(1 - \sum_{i=0}^{m-1} \frac{(\frac{y}{\sigma})^i}{i!} e^{-\frac{y}{\sigma}}\right) u(y). \tag{C.9}$$

Now assume that at time $t-1$ the UAV is at position $(x_u^{t-1}, y_u^{t-1}, h_u)$, and at the next time $t$, the $k$th AP is at $(x_k^t, y_k^t, 0)$. Recall $C^i = \frac{\gamma^i}{\rho_k} d_k^{2\alpha_k}$, and $d_k^2 = (x_u^t - x_k^t)^2 + (y_u^t - y_k^t)^2 + h_u^2$ as described in Section 6.2 and Section 6.3 respectively. By plugging the constant speed model (6.15) into these expressions, we obtain

$$C^i = K^i \left(1 + \frac{2r_k}{L_k} cos(\delta - \theta_k^0)\right)^{\alpha},$$

where

$$
\begin{aligned}
K^i &= \frac{\gamma^i}{\rho_k} L_k^{\alpha_k}, \\
L_k &= (x_u^{t-1} - x_k^t)^2 + (y_u^{t-1} - y_k^t)^2 + h_u^2 + V^2\Delta^2\,, \\
r_k &= \sqrt{(x_u^{t-1} - x_k^t)^2 + (y_u^{t-1} - y_k^t)^2}\; V\Delta\,,
\end{aligned}
$$

and

$$
\theta_k^0 = \arctan\, \frac{y_u^{t-1} - y_k^t}{x_u^{t-1} - x_k^t}. \tag{C.10}
$$

Consider a function $f(x) = e^{k(1+x)^\alpha}$, where $k$ and $\alpha$ are both constants. When $x$ is small, linearizing $f(x)$ around $x = 0$ using the Taylor expansion, we have $f(x) \approx e^k + \alpha k e^k x$. Therefore

$$
e^{-\frac{y}{\sigma}} = e^{-\frac{\mathbf{C}^i}{\sigma}} \approx e^{-\frac{K^i}{\sigma}} - \alpha_k \frac{K^i}{\sigma}\; e^{-\frac{K^i}{\sigma}} \frac{2r_k}{L_k} cos(\delta - \theta_k^0), \tag{C.11}
$$

and

$$
\frac{y}{\sigma} = \frac{C^i}{\sigma} \approx \frac{K^i}{\sigma} + \frac{\alpha K^i}{\sigma} \frac{2r_k}{L_k} cos(\delta - \theta_k^0). \tag{C.12}
$$

Let us define $a = \frac{K^i}{\sigma}$, $b = \alpha_k \frac{K^i}{\sigma} cos(\delta - \theta_k^0)$, $q = \frac{2r_k}{L_k}$, $c = e^{-\frac{K^i}{\sigma}}$, $d = \alpha \frac{K^i}{\sigma} e^{-\frac{K^i}{\sigma}} cos(\delta - \theta_k^0)$. If we recall the binomial expansion theorem, we have:

$$
(a+b)^n = \sum_{j=0}^{n} \left( \begin{array}{c} n \\ \\ j \end{array} \right) a^j b^{n-j}\,. \tag{C.13}
$$

Note that in most of the scenarios we consider, $L >> 2r_k$ and therefore $q$ is close to 0. In such scenarios, each term in (C.9) can be written as

$$
\begin{aligned}
\frac{1}{i!} \left(\frac{y}{\sigma}\right)^i e^{-\frac{y}{\sigma}} &\approx \frac{1}{i!}(a+bq)^i(c-dq) \\
&= \frac{1}{i!} \left( \sum_{j=0}^{i} C_i^j a^k b^{i-j} q^{i-j} \right) (c - dq)\,.
\end{aligned} \tag{C.14}
$$

Since $q$ is a number close to zero, any terms involving $q$ with higher than second order can be neglected. Hence

$$\begin{aligned}\frac{1}{i!}\left(\frac{y}{\sigma}\right)^i e^{-\frac{y}{\sigma}} &\approx \frac{1}{i!}\left(a^i + ia^{i-1}bq\right)(c - dq) \\ &\approx \frac{1}{i!}\left[a^i c + a^{i-1}(ibc - ad)q\right] \\ &\approx p_1 + p_2\cos(\delta - \theta_i^0), \end{aligned} \tag{C.15}$$

where $p_1 = \frac{1}{i!}(a^i c)$ and $p_2 = \frac{1}{i!}\alpha a^i cq(i-a)$.

The above derivation shows that each term in (C.9) is a sinusoid of the same frequency with some DC offset. Therefore, the sum of these terms is also a sinusoid with the same frequency but a different DC offset and phase. Note that the above analysis can be easily extended to the case where the channel has more than one dominant eigen-mode.

For the $k$th AP, the CDF of the channel's Frobenius norm using the $i$th modulation scheme can be written as

$$F_k(i) = Q_1^{k,i} + Q_2^{k,i} cos(\delta - \theta_k^0)\,, \tag{C.16}$$

where $Q_1^{k,i} = 1 - \sum_{j=1}^{L}\sum_{k=1}^{m_j} A_{jk}\sum_{i=0}^{k-1} P_1^i$ and $Q_2^{k,i} = -\sum_{j=1}^{L}\sum_{k=1}^{m_j} A_{jk}\sum_{i=0}^{k-1} P_2^i$. The overall rate of AP $k$ can be obtained using (6.13):

$$R_k = \beta_k(A_k cos(\delta - \theta_k^0) + C_k)\,, \tag{C.17}$$

where $C_k = \sum_{i=1}^{L-1}\log_2 K^i \cdot (Q_1^{k,i+1} - Q_1^{k,i}) + \log_2 K^L - \log_2 K^L \cdot Q_1^{k,L}$ and $A_k = \sum_{i=1}^{L-1}\log_2 K^i \cdot (Q_2^{k,i+1} - Q_2^{k,i}) - \log_2 K^L \cdot Q_2^{k,L}$. If no extra constraint is superimposed

on UAV's heading change, the sum rate of the system can be written as

$$
\begin{aligned}
R_T &= \sum_{k=1}^{K} R_k = r\cos\left(\delta-\theta\right)+C\,,\\
r &= \sqrt{(\sum_{k=1}^{K}\beta_k A_k\cos\theta_k^0)^2+(\sum_{k=1}^{K}\beta_k A_k\sin\theta_k^0)^2}\,,\\
\theta &= \arctan\ \frac{\sum_{k=1}^{K}\beta_k A_k\sin\theta_k^0}{\sum_{k=1}^{K}\beta_k A_k\cos\theta_k^0}\,,
\end{aligned}
$$

and

$$
C=\sum_{k=1}^{K}\beta_k\mathbf{C}_k\,. \tag{C.18}
$$