Sample Chapters from

WIRELESS COMMUNICATIONS

by Andrea Goldsmith

Contents

Chapter 1

Overview of Wireless Communications

Wireless communications is, by any measure, the fastest growing segment of the communications industry. As such, it has captured the attention of the media and the imagination of the public. Cellular systems have experienced exponential growth over the last decade and there are currently around two billion users worldwide. Indeed, cellular phones have become a critical business tool and part of everyday life in most developed countries, and are rapidly supplanting antiquated wireline systems in many developing countries. In addition, wireless local area networks currently supplement or replace wired networks in many homes, businesses, and campuses. Many new applications, including wireless sensor networks, automated highways and factories, smart homes and appliances, and remote telemedicine, are emerging from research ideas to concrete systems. The explosive growth of wireless systems coupled with the proliferation of laptop and palmtop computers indicate a bright future for wireless networks, both as stand-alone systems and as part of the larger networking infrastructure. However, many technical challenges remain in designing robust wireless networks that deliver the performance necessary to support emerging applications. In this introductory chapter we will briefly review the history of wireless networks, from the smoke signals of the pre-industrial age to the cellular, satellite, and other wireless networks of today. We then discuss the wireless vision in more detail, including the technical challenges that must be overcome to make this vision a reality. We describe current wireless systems along with emerging systems and standards. The gap between current and emerging systems and the vision for future wireless applications indicates that much work remains to be done to make this vision a reality.

1.1 History of Wireless Communications

The first wireless networks were developed in the Pre-industrial age. These systems transmitted information over line-of-sight distances (later extended by telescopes) using smoke signals, torch signaling, flashing mirrors, signal flares, or semaphore flags. An elaborate set of signal combinations was developed to convey complex messages with these rudimentary signals. Observation stations were built on hilltops and along roads to relay these messages over large distances. These early communication networks were replaced first by the telegraph network (invented by Samuel Morse in 1838) and later by the telephone. In 1895, a few decades after the telephone was invented, Marconi demonstrated the first radio transmission from the Isle of Wight to a tugboat 18 miles away, and radio communications was born. Radio technology advanced rapidly to enable transmissions over larger distances with better quality, less power, and smaller, cheaper devices, thereby enabling public and private radio communications, television, and wireless networking.

Early radio systems transmitted analog signals. Today most radio systems transmit digital signals composed of binary bits, where the bits are obtained directly from a data signal or by digitizing an analog signal. A digital radio can transmit a continuous bit stream or it can group the bits into packets. The latter type of radio is called a **packet radio** and is characterized by bursty transmissions: the radio is idle except when it transmits a packet. The first network based on packet radio, ALOHANET, was developed at the University of Hawaii in 1971. This network enabled computer sites at seven campuses spread out over four islands to communicate with a central computer on Oahu via radio transmission. The network architecture used a star topology with the central computer at its hub. Any two computers could establish a bi-directional communications link between them by going through the central hub. ALOHANET incorporated the first set of protocols for channel access and routing in packet radio systems, and many of the underlying principles in these protocols are still in use today. The U.S. military was extremely interested in the combination of packet data and broadcast radio inherent to ALOHANET. Throughout the 1970's and early 1980's the Defense Advanced Research Projects Agency (DARPA) invested significant resources to develop networks using packet radios for tactical communications in the battlefield. The nodes in these ad hoc wireless networks had the ability to self-configure (or reconfigure) into a network without the aid of any established infrastructure. DARPA's investment in ad hoc networks peaked in the mid 1980's, but the resulting networks fell far short of expectations in terms of speed and performance. These networks continue to be developed for military use. Packet radio networks also found commercial application in supporting wide-area wireless data services. These services, first introduced in the early 1990's, enable wireless data access (including email, file transfer, and web browsing) at fairly low speeds, on the order of 20 Kbps. A strong market for these wide-area wireless data services never really materialized, due mainly to their low data rates, high cost, and lack of "killer applications". These services mostly disappeared in the 1990s, supplanted by the wireless data capabilities of cellular telephones and wireless local area networks (LANs).

The introduction of wired Ethernet technology in the 1970's steered many commercial companies away from radio-based networking. Ethernet's 10 Mbps data rate far exceeded anything available using radio, and companies did not mind running cables within and between their facilities to take advantage of these high rates. In 1985 the Federal Communications Commission (FCC) enabled the commercial development of wireless LANs by authorizing the public use of the Industrial, Scientific, and Medical (ISM) frequency bands for wireless LAN products. The ISM band was very attractive to wireless LAN vendors since they did not need to obtain an FCC license to operate in this band. However, the wireless LAN systems could not interfere with the primary ISM band users, which forced them to use a low power profile and an inefficient signaling scheme. Moreover, the interference from primary users within this frequency band was quite high. As a result these initial wireless LANs had very poor performance in terms of data rates and coverage. This poor performance, coupled with concerns about security, lack of standardization, and high cost (the first wireless LAN access points listed for $1,400 as compared to a few hundred dollars for a wired Ethernet card) resulted in weak sales. Few of these systems were actually used for data networking: they were relegated to low-tech applications like inventory control. The current generation of wireless LANs, based on the family of IEEE 802.11 standards, have better performance, although the data rates are still relatively low (maximum collective data rates of tens of Mbps) and the coverage area is still small (around 150 m.). Wired Ethernets today offer data rates of 100 Mbps, and the performance gap between wired and wireless LANs is likely to increase over time without additional spectrum allocation. Despite the big data rate differences, wireless LANs are becoming the prefered Internet access method in many homes, offices, and campus environments due to their convenience and freedom from wires. However, most wireless LANs support applications such as email and web browsing that are not bandwidth-intensive. The challenge for future wireless LANs will be to support many users simultaneously with bandwidth-intensive and delay-constrained applications such as video. Range extension is also a critical goal for future wireless LAN systems.

By far the most successful application of wireless networking has been the cellular telephone system. The roots of this system began in 1915, when wireless voice transmission between New York and San Francisco was first established. In 1946 public mobile telephone service was introduced in 25 cities across the United States. These initial systems used a central transmitter to cover an entire metropolitan area. This inefficient use of the radio spectrum coupled with the state of radio technology at that time severely limited the system capacity: thirty years after the introduction of mobile telephone service the New York system could only support 543 users.

A solution to this capacity problem emerged during the 50's and 60's when researchers at AT&T Bell Laboratories developed the cellular concept [4]. Cellular systems exploit the fact that the power of a transmitted signal falls off with distance. Thus, two users can operate on the same frequency at spatially-separate locations with minimal interference between them. This allows very efficient use of cellular spectrum so that a large number of users can be accommodated. The evolution of cellular systems from initial concept to implementation was glacial. In 1947 AT&T requested spectrum for cellular service from the FCC. The design was mostly completed by the end of the 1960's, the first field test was in 1978, and the FCC granted service authorization in 1982, by which time much of the original technology was out-of-date. The first analog cellular system deployed in Chicago in 1983 was already saturated by 1984, at which point the FCC increased the cellular spectral allocation from 40 MHz to 50 MHz. The explosive growth of the cellular industry took almost everyone by surprise. In fact a marketing study commissioned by AT&T before the first system rollout predicted that demand for cellular phones would be limited to doctors and the very rich. AT&T basically abandoned the cellular business in the 1980's focus on fiber optic networks, eventually returning to the business after its potential became apparent. Throughout the late 1980's, as more and more cities became saturated with demand for cellular service, the development of digital cellular technology for increased capacity and better performance became essential.

The second generation of cellular systems, first deployed in the early 1990's, were based on digital communications. The shift from analog to digital was driven by its higher capacity and the improved cost, speed, and power efficiency of digital hardware. While second generation cellular systems initially provided mainly voice services, these systems gradually evolved to support data services such as email, Internet access, and short messaging. Unfortunately, the great market potential for cellular phones led to a proliferation of second generation cellular standards: three different standards in the U.S. alone, and other standards in Europe and Japan, all incompatible. The fact that different cities have different incompatible standards makes roaming throughout the U.S. and the world using one cellular phone standard impossible. Moreover, some countries have initiated service for third generation systems, for which there are also multiple incompatible standards. As a result of the standards proliferation, many cellular phones today are multi-mode: they incorporate multiple digital standards to faciliate nationwide and worldwide roaming, and possibly the first generation analog standard as well, since only this standard provides universal coverage throughout the U.S.

Satellite systems are typically characterized by the height of the satellite orbit, low-earth orbit (LEOs at roughly 2000 Km. altitude), medium-earth orbit (MEOs at roughly 9000 Km. altitude), or geosynchronous orbit (GEOs at roughly 40,000 Km. altitude). The geosynchronous orbits are seen as stationary from the earth, whereas the satellites with other orbits have their coverage area change over time. The concept of using geosynchronous satellites for communications was first suggested by the science fiction writer Arthur C. Clarke in 1945. However, the first deployed satellites, the Soviet Union's Sputnik in 1957 and the NASA/Bell Laboratories' Echo-1 in 1960, were not geosynchronous due to the difficulty of lifting a satellite into such a high orbit. The first GEO satellite was launched by Hughes and NASA in 1963. GEOs then dominated both commercial and government satellite systems for several decades.

Geosynchronous satellites have large coverage areas, so fewer satellites (and dollars) are necessary to provide wide-area or global coverage. However, it takes a great deal of power to reach the satellite, and the propagation delay is typically too large for delay-constrained applications like voice. These disadvantages caused a shift in the 1990's towards lower orbit satellites [6, 7]. The goal was to provide voice and data service competetive with cellular systems. However, the satellite mobile terminals were much bigger, consumed much more power, and cost much more than contemporary cellular phones, which limited their appeal. The most compelling feature of these systems is their ubiquitous worldwide coverage, especially in remote areas or third-world countries with no landline or cellular system infrastructure. Unfortunately, such places do not typically have large demand or the resources the pay for satellite service either. As cellular systems became more widespread, they took away most revenue that LEO systems might have generated in populated areas. With no real market left, most LEO satellite systems went out of business.

A natural area for satellite systems is broadcast entertainment. Direct broadcast satellites operate in the 12 GHz frequency band. These systems offer hundreds of TV channels and are major competitors to cable. Satellite-delivered digital radio has also become popular. These systems, operating in both Europe and the US, offer digital audio broadcasts at near-CD quality.

1.2 Wireless Vision

The vision of wireless communications supporting information exchange between people or devices is the communications frontier of the next few decades, and much of it already exists in some form. This vision will allow multimedia communication from anywhere in the world using a small handheld device or laptop. Wireless networks will connect palmtop, laptop, and desktop computers anywhere within an office building or campus, as well as from the corner cafe. In the home these networks will enable a new class of intelligent electronic devices that can interact with each other and with the Internet in addition to providing connectivity between computers, phones, and security/monitoring systems. Such smart homes can also help the elderly and disabled with assisted living, patient monitoring, and emergency response. Wireless entertainment will permeate the home and any place that people congregate. Video teleconferencing will take place between buildings that are blocks or continents apart, and these conferences can include travelers as well, from the salesperson who missed his plane connection to the CEO off sailing in the Caribbean. Wireless video will enable remote classrooms, remote training facilities, and remote hospitals anywhere in the world. Wireless sensors have an enormous range of both commercial and military applications. Commercial applications include monitoring of fire hazards, hazardous waste sites, stress and strain in buildings and bridges, carbon dioxide movement and the spread of chemicals and gasses at a disaster site. These wireless sensors self-configure into a network to process and interpret sensor measurements and then convey this information to a centralized control location. Military applications include identification and tracking of enemy targets, detection of chemical and biological attacks, support of unmanned robotic vehicles, and counter-terrorism. Finally, wireless networks enable distributed control systems, with remote devices, sensors, and actuators linked together via wireless communication channels. Such networks enable automated highways, mobile robots, and easily-reconfigurable industrial automation.

The various applications described above are all components of the wireless vision. So what, exactly, is wireless communications? There are many different ways to segment this complex topic into different applications, systems, or coverage regions [8]. Wireless applications include voice, Internet access, web browsing, paging and short messaging, subscriber information services, file transfer, video teleconferencing, entertainment, sensing, and distributed control. Systems include cellular telephone systems, wireless LANs, wide-area wireless data systems, satellite systems, and ad hoc wireless networks. Coverage regions include in-building, campus, city, regional, and global. The question of how best to characterize wireless communications along these various segments has resulted in considerable fragmentation in the industry, as evidenced by the many different wireless products, standards, and services being offered or proposed. One reason for this fragmentation is that different wireless applications have different requirements. Voice systems have relatively low data rate requirements (around 20 Kbps) and can tolerate a fairly high probability of bit error (bit error rates, or BERs, of around 10^{-3}), but the total delay must be less than around 30 msec or it becomes noticeable to the end user. On the other hand, data systems typically require much higher data rates (1-100 Mbps) and very small BERs (the target BER is 10^{-8} and all bits received in error must be retransmitted) but do not have a fixed delay requirement. Real-time video systems have high data rate requirements coupled with the same delay constraints as voice systems, while paging and short messaging have very low data rate requirements and no delay constraints. These diverse requirements for different applications make it difficult to build one wireless system that can efficiently satisfy all these requirements simultaneously. Wired networks typically integrate the diverse requirements of different using a single protocol. This integration requires that the most stringent requirements for all applications be met simultaneously. While this may be possible on some wired networks, with data rates on the order of Gbps and BERs on the order of 10^{-12}, it is not possible on wireless networks, which have much lower data rates and higher BERs. For these reasons, at least in the near future, wireless systems will continue to be fragmented, with different protocols tailored to support the requirements of different applications.

The exponential growth of cellular telephone use and wireless Internet access have led to great optimism about wireless technology in general. Obviously not all wireless applications will flourish. While many wireless systems and companies have enjoyed spectacular success, there have also been many failures along the way, including first generation wireless LANs, the Iridium satellite system, wide area data services such as Metricom, and fixed wireless access (wireless "cable") to the home. Indeed, it is impossible to predict what wireless failures and triumphs lie on the horizon. Moreover, there must be sufficient flexibility and creativity among both engineers and regulators to allow for accidental successes. It is clear, however, that the current and emerging wireless systems of today coupled with the vision of applications that wireless can enable insure a bright future for wireless technology.

1.3 Technical Issues

Many technical challenges must be addressed to enable the wireless applications of the future. These challenges extend across all aspects of the system design. As wireless terminals add more features, these small devices must incorporate multiple modes of operation to support the different applications and media. Computers process voice, image, text, and video data, but breakthroughs in circuit design are required to implement the same multimode operation in a cheap, lightweight, handheld device. Since consumers don't want large batteries that frequently need recharging, transmission and signal processing in the portable terminal must consume minimal power. The signal processing required to support multimedia applications and networking functions can be power-intensive. Thus, wireless infrastructure-based networks, such as wireless LANs and cellular systems, place as much of the processing burden as possible on fixed sites with large power resources. The associated bottlenecks and single points-of-failure are clearly undesirable for the overall system. Ad hoc wireless networks without infrastructure are highly appealing for many applications due to their flexibility and robustness. For these networks all processing and control must be performed by the network nodes in a distributed fashion, making energy-efficiency challenging to achieve. Energy is a particularly critical resource in networks where nodes cannot recharge their batteries, for example in sensing applications. Network design to meet the application requirements under such hard energy constraints remains a big technological hurdle. The finite bandwidth and random variations of wireless channels also requires robust applications that degrade gracefully as network performance degrades.

Design of wireless networks differs fundamentally from wired network design due to the nature of the wireless channel. This channel is an unpredictable and difficult communications medium. First of all, the radio spectrum is a scarce resource that must be allocated to many different applications and systems. For this reason spectrum is controlled by regulatory bodies both regionally and globally. A regional or global system operating in a given frequency band must obey the restrictions for that band set forth by the corresponding regulatory body. Spectrum can also be very expensive since in many countries spectral licenses are often auctioned to the highest bidder. In the U.S. companies spent over nine billion dollars for second generation cellular licenses, and the auctions in Europe for third generation cellular spectrum garnered around 100 billion dollars. The spectrum obtained through these auctions must be used extremely efficiently to get a reasonable return on its investment, and it must also be reused over and over in the same geographical area, thus requiring cellular system designs with high capacity and good performance. At frequencies around several Gigahertz wireless radio components with reasonable size, power consumption, and cost are available. However, the spectrum in this frequency range is extremely crowded.

Thus, technological breakthroughs to enable higher frequency systems with the same cost and performance would greatly reduce the spectrum shortage. However, path loss at these higher frequencies is larger, thereby limiting range, unless directional antennas are used.

As a signal propagates through a wireless channel, it experiences random fluctuations in time if the transmitter, receiver, or surrounding objects are moving, due to changing reflections and attenuation. Thus, the characteristics of the channel appear to change randomly with time, which makes it difficult to design reliable systems with guaranteed performance. Security is also more difficult to implement in wireless systems, since the airwaves are susceptible to snooping from anyone with an RF antenna. The analog cellular systems have no security, and one can easily listen in on conversations by scanning the analog cellular frequency band. All digital cellular systems implement some level of encryption. However, with enough knowledge, time and determination most of these encryption methods can be cracked and, indeed, several have been compromised. To support applications like electronic commerce and credit card transactions, the wireless network must be secure against such listeners.

Wireless networking is also a significant challenge. The network must be able to locate a given user wherever it is among billions of globally-distributed mobile terminals. It must then route a call to that user as it moves at speeds of up to 100 Km/hr. The finite resources of the network must be allocated in a fair and efficient manner relative to changing user demands and locations. Moreover, there currently exists a tremendous infrastructure of wired networks: the telephone system, the Internet, and fiber optic cable, which should be used to connect wireless systems together into a global network. However, wireless systems with mobile users will never be able to compete with wired systems in terms of data rates and reliability. Interfacing between wireless and wired networks with vastly different performance capabilities is a difficult problem.

Perhaps the most significant technical challenge in wireless network design is an overhaul of the design process itself. Wired networks are mostly designed according to a layered approach, whereby protocols associated with different layers of the system operation are designed in isolation, with baseline mechanisms to interface between layers. The layers in a wireless systems include the link or physical layer, which handles bit transmissions over the communications medium, the access layer, which handles shared access to the communications medium, the network and transport layers, which routes data across the network and insure end-to-end connectivity and data delivery, and the application layer, which dictates the end-to-end data rates and delay constraints associated with the application. While a layering methodology reduces complexity and facilitates modularity and standardization, it also leads to inefficiency and performance loss due to the lack of a global design optimization. The large capacity and good reliability of wired networks make these inefficiencies relatively benign for many wired network applications, although it does preclude good performance of delay-constrained applications such as voice and video. The situation is very different in a wireless network. Wireless links can exhibit very poor performance, and this performance along with user connectivity and network topology changes over time. In fact, the very notion of a wireless link is somewhat fuzzy due to the nature of radio propagation and broadcasting. The dynamic nature and poor performance of the underlying wireless communication channel indicates that high-performance networks must be optimized for this channel and must be robust and adaptive to its variations, as well as to network dynamics. Thus, these networks require integrated and adaptive protocols at all layers, from the link layer to the application layer. This cross-layer protocol design requires interdiciplinary expertise in communications, signal processing, and network theory and design.

In the next section we give an overview of the wireless systems in operation today. It will be clear from this overview that the wireless vision remains a distant goal, with many technical challenges to overcome. These challenges will be examined in detail throughout the book.

1.4 Current Wireless Systems

This section provides a brief overview of current wireless systems in operation today. The design details of these system are constantly evolving, with new systems emerging and old ones going by the wayside. Thus, we will focus mainly on the high-level design aspects of the most common systems. More details on wireless system standards can be found in [1, 2, 3] A summary of the main wireless system standards is given in Appendix D.

1.4.1 Cellular Telephone Systems

Cellular telephone systems are extremely popular and lucrative worldwide: these are the systems that ignited the wireless revolution. Cellular systems provide two-way voice and data communication with regional, national, or international coverage. Cellular systems were initially designed for mobile terminals inside vehicles with antennas mounted on the vehicle roof. Today these systems have evolved to support lightweight handheld mobile terminals operating inside and outside buildings at both pedestrian and vehicle speeds.

The basic premise behind cellular system design is frequency reuse, which exploits the fact that signal power falls off with distance to reuse the same frequency spectrum at spatially-separated locations. Specifically, the coverage area of a cellular system is divided into nonoverlapping cells where some set of channels is assigned to each cell. This same channel set is used in another cell some distance away, as shown in Figure 1.1, where C_i denotes the channel set used in a particular cell. Operation within a cell is controlled by a centralized base station, as described in more detail below. The interference caused by users in different cells operating on the same channel set is called intercell interference. The spatial separation of cells that reuse the same channel set, the reuse distance, should be as small as possible so that frequencies are reused as often as possible, thereby maximizing spectral efficiency. However, as the reuse distance decreases, intercell interference increases, due to the smaller propagation distance between interfering cells. Since intercell interference must remain below a given threshold for acceptable system performance, reuse distance cannot be reduced below some minimum value. In practice it is quite difficult to determine this minimum value since both the transmitting and interfering signals experience random power variations due to the characteristics of wireless signal propagation. In order to determine the best reuse distance and base station placement, an accurate characterization of signal propagation within the cells is needed.

Initial cellular system designs were mainly driven by the high cost of base stations, approximately one million dollars apiece. For this reason early cellular systems used a relatively small number of cells to cover an entire city or region. The cell base stations were placed on tall buildings or mountains and transmitted at very high power with cell coverage areas of several square miles. These large cells are called macrocells. Signal power was radiated uniformly in all directions, so a mobile moving in a circle around the base station would have approximately constant received power if the signal was not blocked by an attenuating object. This circular contour of constant power yields a hexagonal cell shape for the system, since a hexagon is the closest shape to a circle that can cover a given area with multiple nonoverlapping cells.

Cellular systems in urban areas now mostly use smaller cells with base stations close to street level transmitting at much lower power. These smaller cells are called microcells or picocells, depending on their size. This evolution to smaller cells occured for two reasons: the need for higher capacity in areas with high user density and the reduced size and cost of base station electronics. A cell of any size can support roughly the same number of users if the system is scaled accordingly. Thus, for a given coverage area a system with many microcells has a higher number of users per unit area than a system with just a few macrocells. In addition, less power is required at the mobile terminals in microcellular systems, since the terminals are closer to the base stations. However, the evolution to smaller cells has complicated network design. Mobiles traverse a small cell more quickly than a large cell, and therefore handoffs must be processed more quickly. In addition, location management becomes more complicated, since there are more cells within a given area where a mobile may be located. It is also harder to

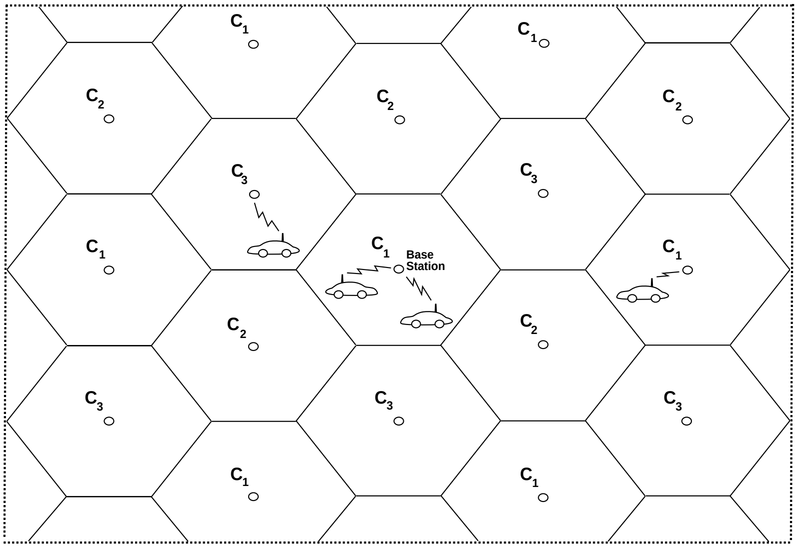


Figure 1.1: Cellular Systems.

develop general propagation models for small cells, since signal propagation in these cells is highly dependent on base station placement and the geometry of the surrounding reflectors. In particular, a hexagonal cell shape is generally not a good approximation to signal propagation in microcells. Microcellular systems are often designed using square or triangular cell shapes, but these shapes have a large margin of error in their approximation to microcell signal propagation [9].

All base stations in a given geographical area are connected via a high-speed communications link to a mobile telephone switching office (MTSO), as shown in Figure 1.2. The MTSO acts as a central controller for the network, allocating channels within each cell, coordinating handoffs between cells when a mobile traverses a cell boundary, and routing calls to and from mobile users. The MTSO can route voice calls through the public switched telephone network (PSTN) or provide Internet access. A new user located in a given cell requests a channel by sending a call request to the cell's base station over a separate control channel. The request is relayed to the MTSO, which accepts the call request if a channel is available in that cell. If no channels are available then the call request is rejected. A call handoff is initiated when the base station or the mobile in a given cell detects that the received signal power for that call is approaching a given minimum threshold. In this case the base station informs the MTSO that the mobile requires a handoff, and the MTSO then queries surrounding base stations to determine if one of these stations can detect that mobile's signal. If so then the MTSO coordinates a handoff between the original base station and the new base station. If no channels are available in the cell with the new base station then the handoff fails and the call is terminated. A call will also be dropped if the signal strength between a mobile and its base station drops below the minimum threshold needed for communication due to random signal variations.

The first generation of cellular systems used analog communications, since they were primarily designed in the 1960's, before digital communications became prevalent. Second generation systems moved from analog to digital due to its many advantages. The components are cheaper, faster, smaller, and require less power. Voice quality is improved due to error correction coding. Digital systems also have higher capacity than analog systems since they can use more spectrally-efficient digital modulation and more efficient techniques to share the cellular spectrum. They can also take advantage of advanced compression techniques and voice activity factors. In addition,

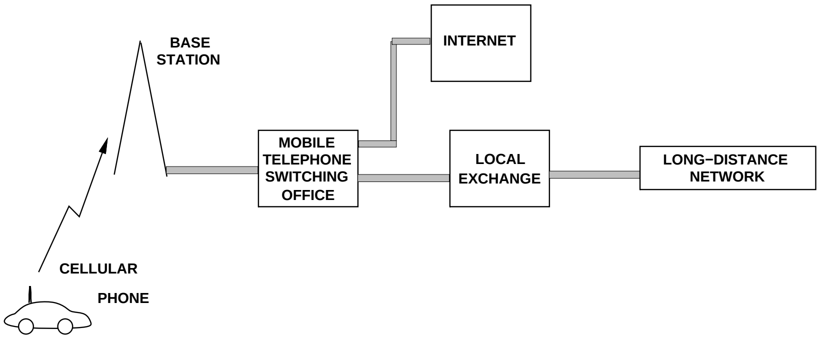


Figure 1.2: Current Cellular Network Architecture

encryption techniques can be used to secure digital signals against eavesdropping. Digital systems can also offer data services in addition to voice, including short messaging, email, Internet access, and imaging capabilities (camera phones). Due to their lower cost and higher efficiency, service providers used aggressive pricing tactics to encourage user migration from analog to digital systems, and today analog systems are primarily used in areas with no digital service. However, digital systems do not always work as well as the analog ones. Users can experience poor voice quality, frequent call dropping, and spotty coverage in certain areas. System performance has certainly improved as the technology and networks mature. In some areas cellular phones provide almost the same quality as landline service. Indeed, some people have replaced their wireline telephone service inside the home with cellular service.

Spectral sharing in communication systems, also called multiple access, is done by dividing the signaling dimensions along the time, frequency, and/or code space axes. In frequency-division multiple access (FDMA) the total system bandwidth is divided into orthogonal frequency channels. In time-division multiple access (TDMA) time is divided orthogonally and each channel occupies the entire frequency band over its assigned timeslot. TDMA is more difficult to implement than FDMA since the users must be time-synchronized. However, it is easier to accommodate multiple data rates with TDMA since multiple timeslots can be assigned to a given user. Code-division multiple access (CDMA) is typically implemented using direct-sequence or frequency-hopping spread spectrum with either orthogonal or non-orthogonal codes. In direct-sequence each user modulates its data sequence by a different chip sequence which is much faster than the data sequence. In the frequency domain, the narrowband data signal is convolved with the wideband chip signal, resulting in a signal with a much wider bandwidth than the original data signal. In frequency-hopping the carrier frequency used to modulate the narrowband data signal is varied by a chip sequence which may be faster or slower than the data sequence. This results in a modulated signal that hops over different carrier frequencies. Typically spread spectrum signals are superimposed onto each other within the same signal bandwidth. A spread spectrum receiver separates out each of the distinct signals by separately decoding each spreading sequence. However, for non-orthogonal codes users within a cell interfere with each other (intracell interference) and codes that are reused in other cells cause intercell interference. Both the intracell and intercell interference power is reduced by the spreading gain of the code. Moreover, interference in spread spectrum systems can be further reduced through multiuser detection and interference cancellation. More details on these different techniques for spectrum sharing and their performance analysis will be given in Chapters 13-14. The design tradeoffs associated with spectrum sharing are very complex, and the decision of which technique is best for a given system and operating environment is never straightforward.

Efficient cellular system designs are **interference-limited**, i.e. the interference dominates the noise floor since otherwise more users could be added to the system. As a result, any technique to reduce interference in cellular systems leads directly to an increase in system capacity and performance. Some methods for interference reduction in use today or proposed for future systems include cell sectorization, directional and smart antennas, multiuser detection, and dynamic resource allocation. Details of these techniques will be given in Chapter 15.

The first generation (1G) cellular systems in the U.S., called the Advance Mobile Phone Service (AMPS), used FDMA with 30 KHz FM-modulated voice channels. The FCC initially allocated 40 MHz of spectrum to this system, which was increased to 50 MHz shortly after service introduction to support more users. This total bandwidth was divided into two 25 MHz bands, one for mobile-to-base station channels and the other for base station-to-mobile channels. The FCC divided these channels into two sets that were assigned to two different service providers in each city to encourage competition. A similar system, the European Total Access Communication System (ETACS), emerged in Europe. AMPS was deployed worldwide in the 1980's and remains the only cellular service in some of these areas, including some rural parts of the U.S.

Many of the first generation cellular systems in Europe were incompatible, and the Europeans quickly converged on a uniform standard for second generation (2G) digital systems called GSM [1]. The GSM standard uses a combination of TDMA and slow frequency hopping with frequency-shift keying for the voice modulation. In contrast, the standards activities in the U.S. surrounding the second generation of digital cellular provoked a raging debate on spectrum sharing techniques, resulting in several incompatible standards [10, 11, 12]. In particular, there are two standards in the 900 MHz cellular frequency band: IS-54, which uses a combination of TDMA and FDMA and phase-shift keyed modulation, and IS-95, which uses direct-sequence CDMA with binary modulation and coding [13, 14]. The spectrum for digital cellular in the 2 GHz PCS frequency band was auctioned off, so service providers could use an existing standard or develop proprietary systems for their purchased spectrum. The end result has been three different digital cellular standards for this frequency band: IS-136 (which is basically the same as IS-54 at a higher frequency), IS-95, and the European GSM standard. The digital cellular standard in Japan is similar to IS-54 and IS-136 but in a different frequency band, and the GSM system in Europe is at a different frequency than the GSM systems in the U.S. This proliferation of incompatible standards in the U.S. and internationally makes it impossible to roam between systems nationwide or globally without a multi-mode phone and/or multiple phones (and phone numbers).

All of the second generation digital cellular standards have been enhanced to support high rate packet data services [15]. GSM systems provide data rates of up to 100 Kbps by aggregating all timeslots together for a single user. This enhancement is called GPRS. A more fundamental enhancement, Enhanced Data Services for GSM Evolution (EDGE), further increases data rates using a high-level modulation format combined with FEC coding. This modulation is more sensitive to fading effects, and EDGE uses adaptive techniques to mitigate this problem. Specifically, EDGE defines six different modulation and coding combinations, each optimized to a different value of received SNR. The received SNR is measured at the receiver and fed back to the transmitter, and the best modulation and coding combination for this SNR value is used. The IS-54 and IS-136 systems currently provide data rates of 40-60 Kbps by aggregating time slots and using high-level modulation. This evolution of the IS-136 standard is called IS-136HS (high-speed). The IS-95 systems support higher data using a time-division technique called high data rate (HDR)[16].

The third generation (3G) cellular systems are based on a wideband CDMA standard developed within the auspices of the International Telecommunications Union (ITU) [15]. The standard, initially called International Mobile Telecommunications 2000 (IMT-2000), provides different data rates depending on mobility and location, from 384 Kbps for pedestrian use to 144 Kbps for vehicular use to 2 Mbps for indoor office use. The 3G standard is incompatible with 2G systems, so service providers must invest in a new infrastructure before they can provide 3G service. The first 3G systems were deployed in Japan. One reason that 3G services came out first in Japan is the process of 3G spectrum allocation, which in Japan was awarded without much up-front cost. The 3G spectrum in both Europe and the U.S. is allocated based on auctioning, thereby requiring a huge initial investment for any company wishing to provide 3G service. European companies collectively paid over 100 billion dollars

[1] The acronym GSM originally stood for Groupe Spéciale Mobile, the name of the European charter establishing the GSM standard. As GSM systems proliferated around the world, the underlying acronym meaning was changed to Global Systems for Mobile Communications.

in their 3G spectrum auctions. There has been much controversy over the 3G auction process in Europe, with companies charging that the nature of the auctions caused enormous overbidding and that it will be very difficult if not impossible to reap a profit on this spectrum. A few of the companies have already decided to write off their investment in 3G spectrum and not pursue system buildout. In fact 3G systems have not grown as anticipated in Europe, and it appears that data enhancements to 2G systems may suffice to satisfy user demands. However, the 2G spectrum in Europe is severely overcrowded, so users will either eventually migrate to 3G or regulations will change so that 3G bandwidth can be used for 2G services (which is not currently allowed in Europe). 3G development in the U.S. has lagged far behind that of Europe. The available 3G spectrum in the U.S. is only about half that available in Europe. Due to wrangling about which parts of the spectrum will be used, the 3G spectral auctions in the U.S. have not yet taken place. However, the U.S. does allow the 1G and 2G spectrum to be used for 3G, and this flexibility may allow a more gradual rollout and investment than the more restrictive 3G requirements in Europe. It appears that delaying 3G in the U.S. will allow U.S. service providers to learn from the mistakes and successes in Europe and Japan.

1.4.2 Cordless Phones

Cordless telephones first appeared in the late 1970's and have experienced spectacular growth ever since. Many U.S. homes today have only cordless phones, which can be a safety risk since these phones don't work in a power outage, in contrast to their wired counterparts. Cordless phones were originally designed to provide a low-cost low-mobility wireless connection to the PSTN, i.e. a short wireless link to replace the cord connecting a telephone base unit and its handset. Since cordless phones compete with wired handsets, their voice quality must be similar. Initial cordless phones had poor voice quality and were quickly discarded by users. The first cordless systems allowed only one phone handset to connect to each base unit, and coverage was limited to a few rooms of a house or office. This is still the main premise behind cordless telephones in the U.S. today, although some base units now support multiple handsets and coverage has improved. In Europe and Asia digital cordless phone systems have evolved to provide coverage over much wider areas, both in and away from home, and are similar in many ways to cellular telephone systems.

The base units of cordless phones connect to the PSTN in the exact same manner as a landline phone, and thus they impose no added complexity on the telephone network. The movement of these cordless handsets is extremely limited: a handset must remain within range of its base unit. There is no coordination with other cordless phone systems, so a high density of these systems in a small area, e.g. an apartment building, can result in significant interference between systems. For this reason cordless phones today have multiple voice channels and scan between these channels to find the one with minimal interference. Many cordless phones use spread spectrum techniques to reduce interference from other cordless phone systems and from other systems like baby monitors and wireless LANs.

In Europe and Asia the second generation of digital cordless phones (CT-2, for cordless telephone, second generation) have an extended range of use beyond a single residence or office. Within a home these systems operate as conventional cordless phones. To extend the range beyond the home base stations, also called phone-points or telepoints, are mounted in places where people congregate, like shopping malls, busy streets, train stations, and airports. Cordless phones registered with the telepoint provider can place calls whenever they are in range of a telepoint. Calls cannot be received from the telepoint since the network has no routing support for mobile users, although some CT-2 handsets have built-in pagers to compensate for this deficiency. These systems also do not handoff calls if a user moves between different telepoints, so a user must remain within range of the telepoint where his call was initiated for the duration of the call. Telepoint service was introduced twice in the United Kingdom and failed both times, but these systems grew rapidly in Hong Kong and Singapore through the mid 1990's. This rapid growth deteriorated quickly after the first few years, as cellular phone operators cut prices to compete with telepoint service. The main complaint about telepoint service was the incomplete radio coverage and lack of handoff. Since cellular systems avoid these problems, as long as prices were competitive there was little reason for people to use telepoint services. Most of these services have now disappeared.

Another evolution of the cordless telephone designed primarily for office buildings is the European DECT system. The main function of DECT is to provide local mobility support for users in an in-building private branch exchange (PBX). In DECT systems base units are mounted throughout a building, and each base station is attached through a controller to the PBX of the building. Handsets communicate to the nearest base station in the building, and calls are handed off as a user walks between base stations. DECT can also ring handsets from the closest base station. The DECT standard also supports telepoint services, although this application has not received much attention, probably due to the failure of CT-2 services. There are currently around 7 million DECT users in Europe, but the standard has not yet spread to other countries.

A more advanced cordless telephone system that emerged in Japan is the Personal Handyphone System (PHS). The PHS system is quite similar to a cellular system, with widespread base station deployment supporting handoff and call routing between base stations. With these capabilities PHS does not suffer from the main limitations of the CT-2 system. Initially PHS systems enjoyed one of the fastest growth rates ever for a new technology. In 1997, two years after its introduction, PHS subscribers peaked at about 7 million users, but its popularity then started to decline due to sharp price cutting by cellular providers. In 2005 there were about 4 million subscribers, attracted by the flat-rate service and relatively high speeds (128 Kbps) for data. PHS operators are trying to push data rates up to 1 Mbps, which cellular providers cannot compete with. The main difference between a PHS system and a cellular system is that PHS cannot support call handoff at vehicle speeds. This deficiency is mainly due to the dynamic channel allocation procedure used in PHS. Dynamic channel allocation greatly increases the number of handsets that can be serviced by a single base station and their corresponding data rates, thereby lowering the system cost, but it also complicates the handoff procedure. Given the sustained popularity of PHS, it is unlikely to go the same route as CT-2 any time soon, especially if much higher data rates become available. However, it is clear from the recent history of cordless phone systems that to extend the range of these systems beyond the home requires either similar or better functionality than cellular systems or a significantly reduced cost.

1.4.3 Wireless LANs

Wireless LANs provide high-speed data within a small region, e.g. a campus or small building, as users move from place to place. Wireless devices that access these LANs are typically stationary or moving at pedestrian speeds. All wireless LAN standards in the U.S. operate in unlicensed frequency bands. The primary unlicensed bands are the ISM bands at 900 MHz, 2.4 GHz, and 5.8 GHz, and the Unlicensed National Information Infrastructure (U-NII) band at 5 GHz. In the ISM bands unlicensed users are secondary users so must cope with interference from primary users when such users are active. There are no primary users in the U-NII band. An FCC license is not required to operate in either the ISM or U-NII bands. However, this advantage is a double-edged sword, since other unlicensed systems operate in these bands for the same reason, which can cause a great deal of interference between systems. The interference problem is mitigated by setting a limit on the power per unit bandwidth for unlicensed systems. Wireless LANs can have either a star architecture, with wireless access points or hubs placed throughout the coverage region, or a peer-to-peer architecture, where the wireless terminals self-configure into a network.

Dozens of wireless LAN companies and products appeared in the early 1990's to capitalize on the "pent-up demand" for high-speed wireless data. These first generation wireless LANs were based on proprietary and incompatible protocols. Most operated within the 26 MHz spectrum of the 900 MHz ISM band using direct sequence spread spectrum, with data rates on the order of 1-2 Mbps. Both star and peer-to-peer architectures were used. The lack of standardization for these products led to high development costs, low-volume production, and small markets for each individual product. Of these original products only a handful were even mildly successful. Only one of the first generation wireless LANs, Motorola's Altair, operated outside the 900 MHz band. This system, operating in the licensed 18 GHz band, had data rates on the order of 6 Mbps. However, performance of Altair was hampered by the high cost of components and the increased path loss at 18 GHz, and Altair was discontinued within a few years of its release.

The second generation of wireless LANs in the U.S. operate with 80 MHz of spectrum in the 2.4 GHz ISM band. A wireless LAN standard for this frequency band, the IEEE 802.11b standard, was developed to avoid some of the problems with the proprietary first generation systems. The standard specifies direct sequence spread spectrum with data rates of around 1.6 Mbps (raw data rates of 11 Mbps) and a range of approximately 150 m. The network architecture can be either star or peer-to-peer, although the peer-to-peer feature is rarely used. Many companies developed products based on the 802.11b standard, and after slow initial growth the popularity of 802.11b wireless LANs has expanded considerably. Many laptops come with integrated 802.11b wireless LAN cards. Companies and universities have installed 802.11b base stations throughout their locations, and many coffee houses, airports, and hotels offer wireless access, often for free, to increase their appeal.

Two additional standards in the 802.11 family were developed to provide higher data rates than 802.11b. The IEEE 802.11a wireless LAN standard operates with 300 MHz of spectrum in the 5 GHz U-NII band. The 802.11a standard is based on multicarrier modulation and provides 20-70 Mbps data rates. Since 802.11a has much more bandwidth and consequently many more channels than 802.11b, it can support more users at higher data rates. There was some initial concern that 802.11a systems would be significantly more expensive than 802.11b systems, but in fact they quickly became quite competitive in price. The other standard, 802.11g, also uses multicarrier modulation and can be used in either the 2.4 GHz and 5 GHz bands with speeds of up to 54 Mbps. Many wireless LAN cards and access points support all three standards to avoid incompatibilities.

In Europe wireless LAN development revolves around the HIPERLAN (high performance radio LAN) standards. The first HIPERLAN standard, HIPERLAN Type 1, is similar to the IEEE 802.11a wireless LAN standard, with data rates of 20 Mbps at a range of 50 m. This system operates in a 5 GHz band similar to the U-NII band. Its network architecture is peer-to-peer. The next generation of HIPERLAN, HIPERLAN Type 2, is still under development, but the goal is to provide data rates on the order of 54 Mbps with a similar range, and also to support access to cellular, ATM, and IP networks. HIPERLAN Type 2 is also supposed to include support for Quality-of-Service (QoS), however it is not yet clear how and to what extent this will be done.

1.4.4 Wide Area Wireless Data Services

Wide area wireless data services provide wireless data to high-mobility users over a very large coverage area. In these systems a given geographical region is serviced by base stations mounted on towers, rooftops, or mountains. The base stations can be connected to a backbone wired network or form a multihop ad hoc wireless network.

Initial wide area wireless data services had very low data rates, below 10 Kbps, which gradually increased to 20 Kbps. There were two main players providing this service: Motient and Bell South Mobile Data (formerly RAM Mobile Data). Metricom provided a similar service with a network architecture consisting of a large network of small inexpensive base stations with small coverage areas. The increased efficiency of the small coverage areas allowed for higher data rates in Metricom, 76 Kbps, than in the other wide-area wireless data systems. However, the high infrastructure cost for Metricom eventually forced it into bankruptcy, and the system was shut down. Some of the infrastructure was bought and is operating in a few areas as Ricochet.

The cellular digital packet data (CDPD) system is a wide area wireless data service overlayed on the analog cellular telephone network. CDPD shares the FDMA voice channels of the analog systems, since many of these channels are idle due to the growth of digital cellular. The CDPD service provides packet data transmission at rates of 19.2 Kbps, and is available throughout the U.S. However, since newer generations of cellular systems also provide data services, CDPD is mostly being replaced by these newer services. Thus, wide ara wireless data services have not been very successful, although emerging systems that offer broadband access may have more appeal.

1.4.5 Broadband Wireless Access

Broadband wireless access provides high-rate wireless communications between a fixed access point and multiple terminals. These systems were initially proposed to support interactive video service to the home, but the application emphasis then shifted to providing high speed data access (tens of Mbps) to the Internet, the WWW, and to high speed data networks for both homes and businesses. In the U.S. two frequency bands were set aside for these systems: part of the 28 GHz spectrum for local distribution systems (local multipoint distribution systems or LMDS) and a band in the 2 GHz spectrum for metropolitan distribution systems (multichannel multipoint distribution services or MMDS). LMDS represents a quick means for new service providers to enter the already stiff competition among wireless and wireline broadband service providers [1, Chapter 2.3]. MMDS is a television and telecommunication delivery system with transmission ranges of 30-50 Km [1, Chapter 11.11]. MMDS has the capability to deliver over one hundred digital video TV channels along with telephony and access to emerging interactive services such as the Internet. MMDS will mainly compete with existing cable and satellite systems. Europe is developing a standard similar to MMDS called Hiperaccess.

WiMAX is an emerging broadband wireless technology based on the IEEE 802.16 standard [20, 21]. The core 802.16 specification is a standard for broadband wireless access systems operating at radio frequencies between 10 GHz and 66 GHz. Data rates of around 40 Mbps will be available for fixed users and 15 Mbps for mobile users, with a range of several kilometers. Many laptop and PDA manufacturers are planning to incorporate WiMAX once it becomes available to satisfy demand for constant Internet access and email exchange from any location. WiMax will compete with wireless LANs, 3G cellular services, and possibly wireline services like cable and DSL. The ability of WiMax to challenge or supplant these systems will depend on its relative performance and cost, which remain to be seen.

1.4.6 Paging Systems

Paging systems broadcast a short paging message simultaneously from many tall base stations or satellites transmitting at very high power (hundreds of watts to kilowatts). Systems with terrestrial transmitters are typically localized to a particular geographic area, such as a city or metropolitan region, while geosynchronous satellite transmitters provide national or international coverage. In both types of systems no location management or routing functions are needed, since the paging message is broadcast over the entire coverage area. The high complexity and power of the paging transmitters allows low-complexity, low-power, pocket paging receivers with a long usage time from small and lightweight batteries. In addition, the high transmit power allows paging signals to easily penetrate building walls. Paging service also costs less than cellular service, both for the initial device and for the monthly usage charge, although this price advantage has declined considerably in recent years as cellular prices dropped. The low cost, small and lightweight handsets, long battery life, and ability of paging devices to work almost anywhere indoors or outdoors are the main reasons for their appeal.

Early radio paging systems were analog 1 bit messages signaling a user that someone was trying to reach him or her. These systems required callback over a landline telephone to obtain the phone number of the paging party. The system evolved to allow a short digital message, including a phone number and brief text, to be sent to the pagee as well. Radio paging systems were initially extremely successful, with a peak of 50 million subscribers in the U.S. alone. However, their popularity started to wane with the widespread penetration and competitive cost of cellular telephone systems. Eventually the competition from cellular phones forced paging systems to provide new capabilities. Some implemented "answer-back" capability, i.e. two-way communication. This required a major change in the pager design, since it needed to transmit signals in addition to receiving them, and the transmission distances to a satellite or distance base station is very large. Paging companies also teamed up with palmtop computer makers to incorporate paging functions into these devices [5]. Despite these developments, the market for paging devices has shrunk considerably, although there is still a niche market among doctors and other professionals that must be reachable anywhere.

1.4.7 Satellite Networks

Commercial satellite systems are another major component of the wireless communications infrastructure [6, 7]. Geosynchronous systems include Inmarsat and OmniTRACS. The former is geared mainly for analog voice transmission from remote locations. For example, it is commonly used by journalists to provide live reporting from war zones. The first generation Inmarsat-A system was designed for large (1m parabolic dish antenna) and rather expensive terminals. Newer generations of Inmarsats use digital techniques to enable smaller, less expensive terminals, around the size of a briefcase. Qualcomm's OmniTRACS provides two-way communications as well as location positioning. The system is used primarily for alphanumeric messaging and location tracking of trucking fleets. There are several major difficulties in providing voice and data services over geosynchronous satellites. It takes a great deal of power to reach these satellites, so handsets are typically large and bulky. In addition, there is a large round-trip propagation delay: this delay is quite noticeable in two-way voice communication. Geosynchronous satellites also have fairly low data rates, less than 10 Kbps. For these reasons lower orbit LEO satellites were thought to be a better match for voice and data communications.

LEO systems require approximately 30-80 satellites to provide global coverage, and plans for deploying such constellations were widespread in the late 1990's. One of the most ambitious of these systems, the Iridium constellation, was launched at that time. However, the cost of these satellites, to build, launch, and maintain, is much higher than that of terrestrial base stations. Although these LEO systems can certainly complement terrestrial systems in low-population areas, and are also appealing to travelers desiring just one handset and phone number for global roaming, the growth and diminished cost of cellular prevented many ambitious plans for widespread LEO voice and data systems to materialize. Iridium was eventually forced into bankruptcy and disbanded, and most of the other systems were never launched. An exception to these failures was the Globalstar LEO system, which currently provides voice and data services over a wide coverage area at data rates under 10 Kbps. Some of the Iridium satellites are still operational as well.

The most appealing use for satellite system is broadcasting of video and audio over large geographic regions. In the U.S. approximately 1 in 8 homes have direct broadcast satellite service, and satellite radio is emerging as a popular service as well. Similar audio and video satellite broadcasting services are widespread in Europe. Satellites are best tailored for broadcasting, since they cover a wide area and are not compromised by an initial propagation delay. Moreover, the cost of the system can be amortized over many years and many users, making the service quite competitive with terrestrial entertainment broadcasting systems.

1.4.8 Low-Cost Low-Power Radios: Bluetooth and Zigbee

As radios decrease their cost and power consumption, it becomes feasible to embed them in more types of electronic devices, which can be used to create smart homes, sensor networks, and other compelling applications. Two radios have emerged to support this trend: Bluetooth and Zigbee.

Bluetooth[2] radios provide short range connections between wireless devices along with rudimentary networking capabilities. The Bluetooth standard is based on a tiny microchip incorporating a radio transceiver that is built into digital devices. The transceiver takes the place of a connecting cable for devices such as cell phones, laptop and palmtop computers, portable printers and projectors, and network access points. Bluetooth is mainly for short range communications, e.g. from a laptop to a nearby printer or from a cell phone to a wireless headset. Its normal range of operation is 10 m (at 1 mW transmit power), and this range can be increased to 100 m by increasing the transmit power to 100 mW. The system operates in the unlicensed 2.4 GHz frequency band, hence it can be used

[2] The Bluetooth standard is named after Harald I Bluetooth, the king of Denmark between 940 and 985 AD who united Denmark and Norway. Bluetooth proposes to unite devices via radio connections, hence the inspiration for its name.

worldwide without any licensing issues. The Bluetooth standard provides 1 asynchronous data channel at 723.2 Kbps. In this mode, also known as Asynchronous Connection-Less, or ACL, there is a reverse channel with a data rate of 57.6 Kbps. The specification also allows up to three synchronous channels each at a rate of 64 Kbps. This mode, also known as Synchronous Connection Oriented or SCO, is mainly used for voice applications such as headsets, but can also be used for data. These different modes result in an aggregate bit rate of approximately 1 Mbps. Routing of the asynchronous data is done via a packet switching protocol based on frequency hopping at 1600 hops per second. There is also a circuit switching protocol for the synchronous data.

Bluetooth uses frequency-hopping for multiple access with a carrier spacing of 1 MHz. Typically, up to 80 different frequencies are used, for a total bandwidth of 80 MHz. At any given time, the bandwidth available is 1 MHz, with a maximum of eight devices sharing the bandwidth. Different logical channels (different hopping sequences) can simultaneously share the same 80 MHz bandwidth. Collisions will occur when devices in different piconets, on different logical channels, happen to use the same hop frequency at the same time. As the number of piconets in an area increases, the number of collisions increases, and performance degrades.

The Bluetooth standard was developed jointly by 3 Com, Ericsson, Intel, IBM, Lucent, Microsoft, Motorola, Nokia, and Toshiba. The standard has now been adopted by over 1300 manufacturers, and many consumer electronic products incorporate Bluetooth, including wireless headsets for cell phones, wireless USB or RS232 connectors, wireless PCMCIA cards, and wireless settop boxes.

The ZigBee[3] radio specification is designed for lower cost and power consumption than Bluetooth [5]. The specification is based on the IEEE 802.15.4 standard. The radio operates in the same ISM band as Bluetooth, and is capable of connecting 255 devices per network. The specification supports data rates of up to 250 Kbps at a range of up to 30 m. These data rates are slower than Bluetooth, but in exchange the radio consumes significantly less power with a larger transmission range. The goal of ZigBee is to provide radio operation for months or years without recharging, thereby targeting applications such as sensor networks and inventory tags.

1.4.9 Ultrawideband Radios

Ultrawideband (UWB) radios are extremely wideband radios with very high potential data rates [18, 6]. The concept of ultrawideband communications actually originated with Marconi's spark gap transmitter, which occupied a very wide bandwidth. However, since only a single low-rate user could occupy the spectrum, wideband communications was abandoned in favor of more efficient communication techniques. The renewed interest in wideband communications was spurred by the FCC's decision in 2002 to allow operation of UWB devices as system underlayed beneath existing users over a 7 GHz range of frequencies. These systems can operate either at baseband or at a carrier frequency in the 3.6-10.1 GHz range. The underlay in theory interferers with all systems in that frequency range, including critical safety and military systems, unlicensed systems such as 802.11 wireless and Bluetooth, and cellular systems where operators paid billions of dollars for dedicated spectrum use. The FCC's ruling was quite controversial given the vested interest in interference-free spectrum of these users. To minimize the impact of UWB on primary band users, the FCC put in place severe transmit power restrictions. This requires UWB devices to be within close proximity of their intended receiver.

UWB radios come with unique advantages that have long been appreciated by the radar and communications communities. Their wideband nature allows UWB signals to easily penetrate through obstacles and provides very precise ranging capabilities. Moreover, the available UWB bandwidth has the potential for very high data rates. Finally, the power restrictions dictate that the devices can be small with low power consumption.

Initial UWB systems used ultra-short pulses with simple amplitude or position modulation. Multipath can significantly degrade performance of such systems, and proposals to mitigate the effects of multipath include

[3] Zigbee takes its name from the dance that honey bees use to communicate information about new-found food sources to other members of the colony.

equalization and multicarrier modulation. Precise and rapid synchronization is also a big challenge for these systems. While many technical challenges remain, the appeal of UWB technology has sparked great interest both commercially and in the research community to address these issues.

1.5 The Wireless Spectrum

1.5.1 Methods for Spectrum Allocation

Most countries have government agencies responsible for allocating and controlling the use of the radio spectrum. In the U.S. spectrum is allocated by the Federal Communications Commission (FCC) for commercial use and by the Office of Spectral Management (OSM) for military use. Commercial spectral allocation is governed in Europe by the European Telecommunications Standards Institute (ETSI) and globally by the International Telecommunications Union (ITU). Governments decide how much spectrum to allocate between commercial and military use, and this decision is dynamic depending on need. Historically the FCC allocated spectral blocks for specific uses and assigned licenses to use these blocks to specific groups or companies. For example, in the 1980s the FCC allocated frequencies in the 800 MHz band for analog cellular phone service, and provided spectral licenses to two operators in each geographical area based on a number of criteria. While the FCC and regulatory bodies in other countries still allocate spectral blocks for specific purposes, these blocks are now commonly assigned through spectral auctions to the highest bidder. While some argue that this market-based method is the fairest way for governments to allocate the limited spectral resource, and it provides significant revenue to the government besides, there are others who believe that this mechanism stifles innovation, limits competition, and hurts technology adoption. Specifically, the high cost of spectrum dictates that only large companies or conglomerates can purchase it. Moreover, the large investment required to obtain spectrum can delay the ability to invest in infrastructure for system rollout and results in very high initial prices for the end user. The 3G spectral auctions in Europe, in which several companies ultimately defaulted, have provided fuel to the fire against spectral auctions.

In addition to spectral auctions, spectrum can be set aside in specific frequency bands that are free to use with a license according to a specific set of etiquette rules. The rules may correspond to a specific communications standard, power levels, etc. The purpose of these unlicensed bands is to encourage innovation and low-cost implementation. Many extremely successful wireless systems operate in unlicensed bands, including wireless LANs, Bluetooth, and cordless phones. A major difficulty of unlicensed bands is that they can be killed by their own success. If many unlicensed devices in the same band are used in close proximity, they generate much interference to each other, which can make the band unusable.

Underlay systems are another alternative to allocate spectrum. An underlay system operates as a secondary user in a frequency band with other primary users. Operation of secondary users is typically restricted so that primary users experience minimal interference. This is usually accomplished by restricting the power/Hz of the secondary users. UWB is an example of an underlay system, as are unlicensed systems in the ISM frequency bands. Such underlay systems can be extremely controversial given the complexity of characterizing how interference affects the primary users. Yet the trend towards spectrum allocation for underlays appears to be accelerating, mainly due to the scarcity of available spectrum for new systems and applications.

Satellite systems cover large areas spanning many countries and sometimes the globe. For wireless systems that span multiple countries, spectrum is allocated by the International Telecommunications Union Radio Communications group (ITU-R). The standards arm of this body, ITU-T, adopts telecommunication standards for global systems that must interoperate with each other across national boundaries.

There is some movement within regulatory bodies worldwide to change the way spectrum is allocated. Indeed, the basic mechanisms for spectral allocation have not changed much since the inception of the regulatory bodies in the early to mid 1900's, although spectral auctions and underlay systems are relatively new. The goal of changing spectrum allocation policy is to take advantage of the technological advances in radios to make spectrum allocation more efficient and flexible. One compelling idea is the notion of a smart or cognitive radio. This type of radio can sense its spectral environment to determine dimensions in time, space, and frequency where it would not cause interference to other users even at moderate to high transmit powers. If such radios could operate over a very wide frequency band, it would open up huge amounts of new bandwidth and tremendous opportunities for new wireless systems and applications. However, many technology and policy hurdles must be overcome to allow such a radical change in spectrum allocation.

1.5.2 Spectrum Allocations for Existing Systems

Most wireless applications reside in the radio spectrum between 30 MHz and 30 GHz. These frequencies are natural for wireless systems since they are not affected by the earth's curvature, require only moderately sized antennas, and can penetrate the ionosphere. Note that the required antenna size for good reception is inversely proportional to the square of signal frequency, so moving systems to a higher frequency allows for more compact antennas. However, received signal power with nondirectional antennas is proportional to the inverse of frequency squared, so it is harder to cover large distances with higher frequency signals.

As discussed in the previous section, spectrum is allocated either in licensed bands (which regulatory bodies assign to specific operators) or in unlicensed bands (which can be used by any system subject to certain operational requirements). The following table shows the licensed spectrum allocated to major commercial wireless systems in the U.S. today. There are similar allocations in Europe and Asia.

AM Radio	535-1605 KHz
FM Radio	88-108 MHz
Broadcast TV (Channels 2-6)	54-88 MHz
Broadcast TV (Channels 7-13)	174-216 MHz
Broadcast TV (UHF)	470-806 MHz
3G Broadband Wireless	746-764 MHz, 776-794 MHz
3G Broadband Wireless	1.7-1.85 MHz, 2.5-2.69 MHz
1G and 2G Digital Cellular Phones	806-902 MHz
Personal Communications Service (2G Cell Phones)	1.85-1.99 GHz
Wireless Communications Service	2.305-2.32 GHz, 2.345-2.36 GHz
Satellite Digital Radio	2.32-2.325 GHz
Multichannel Multipoint Distribution Service (MMDS)	2.15-2.68 GHz
Digital Broadcast Satellite (Satellite TV)	12.2-12.7 GHz
Local Multipoint Distribution Service (LMDS)	27.5-29.5 GHz, 31-31.3 GHz
Fixed Wireless Services	38.6-40 GHz

Note that digital TV is slated for the same bands as broadcast TV, so all broadcasters must eventually switch from analog to digital transmission. Also, the 3G broadband wireless spectrum is currently allocated to UHF TV stations 60-69, but is slated to be reallocated. Both 1G analog and 2G digital cellular services occupy the same cellular band at 800 MHz, and the cellular service providers decide how much of the band to allocate between digital and analog service.

Unlicensed spectrum is allocated by the governing body within a given country. Often countries try to match their frequency allocation for unlicensed use so that technology developed for that spectrum is compatible worldwide. The following table shows the unlicensed spectrum allocations in the U.S.

ISM Band I (Cordless phones, 1G WLANs)	902-928 MHz
ISM Band II (Bluetooth, 802.11b WLANs)	2.4-2.4835 GHz
ISM Band III (Wireless PBX)	5.725-5.85 GHz
NII Band I (Indoor systems, 802.11a WLANs)	5.15-5.25 GHz
NII Band II (short outdoor and campus applications)	5.25-5.35 GHz
NII Band III (long outdoor and point-to-point links)	5.725-5.825 GHz

ISM Band I has licensed users transmitting at high power that interfere with the unlicensed users. Therefore, the requirements for unlicensed use of this band is highly restrictive and performance is somewhat poor. The U-NII bands have a total of 300 MHz of spectrum in three separate 100 MHz bands, with slightly different restrictions on each band. Many unlicensed systems operate in these bands.

1.6 Standards

Communication systems that interact with each other require standardization. Standards are typically decided on by national or international committees: in the U.S. the TIA plays this role. These committees adopt standards that are developed by other organizations. The IEEE is the major player for standards development in the United States, while ETSI plays this role in Europe. Both groups follow a lengthy process for standards development which entails input from companies and other interested parties, and a long and detailed review process. The standards process is a large time investment, but companies participate since if they can incorporate their ideas into the standard, this gives them an advantage in developing the resulting system. In general standards do not include all the details on all aspects of the system design. This allows companies to innovate and differentiate their products from other standardized systems. The main goal of standardization is for systems to interoperate with other systems following the same standard.

In addition to insuring interoperability, standards also enable economies of scale and pressure prices lower. For example, wireless LANs typically operate in the unlicensed spectral bands, so they are not required to follow a specific standard. The first generation of wireless LANs were not standardized, so specialized components were needed for many systems, leading to excessively high cost which, coupled with poor performance, led to very limited adoption. This experience led to a strong push to standardize the next wireless LAN generation, which resulted in the highly successful IEEE 802.11 family of standards. Future generations of wireless LANs are expected to be standardized, including the now emerging IEEE 802.11a standard in the 5 GHz band.

There are, of course, disadvantages to standardization. The standards process is not perfect, as company participants often have their own agenda which does not always coincide with the best technology or best interests of the consumers. In addition, the standards process must be completed at some point, after which time it becomes more difficult to add new innovations and improvements to an existing standard. Finally, the standards process can become very politicized. This happened with the second generation of cellular phones in the U.S., which ultimately led to the adoption of two different standards, a bit of an oxymoron. The resulting delays and technology split put the U.S. well behind Europe in the development of 2nd generation cellular systems. Despite its flaws, standardization is clearly a necessary and often beneficial component of wireless system design and operation. However, it would benefit everyone in the wireless technology industry if some of the problems in the standardization process could be mitigated.

Bibliography

[1] T. S. Rappaport. *Wireless Communications: Principles and Practice*, 2nd ed. Prentice Hall, 2002.

[2] W. Stallings, *Wireless Communications and Networks*, 2nd Ed., Prentice Hall, 2005.

[3] K. Pahlavan and P. Krishnamurthy, *Principles of Wireless Networks A Unified Approach*, New Jersey: Prentice Hall, 2002.

[4] V.H. McDonald, "The Cellular Concept," *Bell System Tech. J*, pp. 15-49, Jan. 1979.

[5] S. Schiesel. Paging allies focus strategy on the Internet. *New York Times*, April 19, 1999.

[6] F. Abrishamkar and Z. Siveski, "PCS global mobile satellites," *IEEE Commun. Mag.*,, pp. 132-136, Sep. 1996.

[7] R. Ananasso and F. D. Priscoli, "The role of satellites in personal communication services," Issue on Mobile Satellite Communications for Seamless PCS, *IEEE J. Sel. Areas Commun.*, pp. 180-196, Feb. 1995.

[8] D. C. Cox, "Wireless personal communications: what is it?," *IEEE Pers. Commun. Mag.*, pp. 20-35, April 1995.

[9] A. J. Goldsmith and L.J. Greenstein. A measurement-based model for predicting coverage areas of urban microcells. *IEEE Journal on Selected Areas in Communication*, pages 1013–1023, September 1993.

[10] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr., and C. E. Wheatley III, "On the capacity of a cellular CDMA system," *IEEE Trans. Veh. Tech.*, pp. 303–312, May 1991.

[11] K. Rath and J. Uddenfeldt, "Capacity of digital cellular TDMA systems," *IEEE Trans. Veh. Tech.*, pp. 323-332, May 1991.

[12] Q. Hardy, "Are claims hope or hype?," *Wall Street Journal*, p. A1, Sep. 6, 1996.

[13] A. Mehrotra, *Cellular Radio: Analog and Digital Systems*, Artech House, 1994.

[14] J. E. Padgett, C. G. Gunther, and T. Hattori, "Overview of wireless personal communications," Special Issue on Wireless Personal Communications, *IEEE Commun. Mag.*, pp. 28–41, Jan. 1995.

[15] J. D. Vriendt, P. Laine, C. Lerouge, X. Xu, "Mobile network evolution: a revolution on the move," *IEEE Commun. Mag.*, pp. 104-111, April 2002.

[16] P. Bender, P. Black, M. Grob, R. Padovani, N. Sundhushayana, A. Viterbi, "CDMA/HDR: A bandwidth efficient high speed wireless data service for nomadic users," *IEEE Commun. Mag.*, July 2000.

[17] I. Poole, "What exactly is . . . ZigBee?," *IEEE Commun. Eng.*, pp. 44-45, Aug.-Sept. 2004

[18] L. Yang and G.B. Giannakis, “Ultra-wideband communications: an idea whose time has come,” *IEEE Signl. Proc. Mag.*, Vol. 21, pp. 26 - 54, Nov. 2004.

[19] D. Porcino and W. Hirt, “Ultra-wideband radio technology: potential and challenges ahead,” *IEEE Commun. Mag.*, Vol. 41, pp. 66 - 74, July 2003

[20] S.J. Vaughan-Nichols, “Achieving wireless broadband with WiMax,” *IEEE Computer*, Vol. 37, pp. 10-13, June 2004.

[21] S.M. Cherry, “WiMax and Wi-Fi: Separate and Unequal,” *IEEE Spectrum*, Vol. 41, pg. 16, March 2004.

Chapter 1 Problems

1. As storage capability increases, we can store larger and larger amounts of data on smaller and smaller storage devices. Indeed, we can envision microscopic computer chips storing terraflops of data. Suppose this data is to be transfered over some distance. Discuss the pros and cons of putting a large number of these storage devices in a truck and driving them to their destination rather than sending the data electronically.

2. Describe two technical advantages and disadvantages of wireless systems that use bursty data transmission rather than continuous data transmission.

3. Fiber optic cable typically exhibits a probability of bit error of $P_b = 10^{-12}$. A form of wireless modulation, DPSK, has $P_b = \frac{1}{2\overline{\gamma}}$ in some wireless channels, where $\overline{\gamma}$ is the average SNR. Find the average SNR required to achieve the same P_b in the wireless channel as in the fiber optic cable. Due to this extremeley high required SNR, wireless channels typically have P_b much larger than 10^{-12}.

4. Find the round-trip delay of data sent between a satellite and the earth for LEO, MEO, and GEO satellites assuming the speed of light is 3×10^8 m/s. If the maximum acceptable delay for a voice system is 30 milliseconds, which of these satellite systems would be acceptable for two-way voice communication?

5. Figure 1.1 indicates a relatively flat growth for wireless data between 1995 and 2000. What applications might significantly increase the growth rate of wireless data users.

6. This problem illustrates some of the economic issues facing service providers as they migrate away from voice-only systems to mixed-media systems. Suppose you are a service provider with 120KHz of bandwidth which you must allocate between voice and data users. The voice users require 20Khz of bandwidth, and the data users require 60KHz of bandwidth. So, for example, you could allocate all of your bandwidth to voice users, resulting in 6 voice channels, or you could divide the bandwidth to have one data channel and three voice channels, etc. Suppose further that this is a time-division system, with timeslots of duration T. All voice and data call requests come in at the beginning of a timeslot and both types of calls last T seconds. There are six independent voice users in the system: each of these users requests a voice channel with probability .8 and pays $.20 if his call is processed. There are two independent data users in the system: each of these users requests a data channel with probability .5 and pays $1 if his call is processed. How should you allocate your bandwidth to maximize your expected revenue?

7. Describe three disadvantages of using a wireless LAN instead of a wired LAN. For what applications will these disadvantages be outweighed by the benefits of wireless mobility. For what applications will the disadvantages override the advantages.

8. Cellular systems are migrating to smaller cells to increase system capacity. Name at least three design issues which are complicated by this trend.

9. Why does minimizing reuse distance maximize spectral efficiency of a cellular system?

10. This problem demonstrates the capacity increase as cell size decreases. Consider a square city that is 100 square kilometers. Suppose you design a cellular system for this city with square cells, where every cell (regardless of cell size) has 100 channels so can support 100 active users (in practice the number of users that can be supported per cell is mostly independent of cell size as long as the propagation model and power scale appropriately).

    (a) What is the total number of active users that your system can support for a cell size of 1 square kilometer?

(b) What cell size would you use if you require that your system support 250,000 active users?

Now we consider some financial implications based on the fact that users do not talk continuously. Assume that Friday from 5-6 pm is the busiest hour for cell phone users. During this time, the average user places a single call, and this call lasts two minutes. Your system should be designed such that the subscribers will tolerate no greater than a two percent blocking probability during this peak hour (Blocking probability is computed using the Erlang B model: $P_b = (A^C/C!)/(\sum_{k=0}^C A^k/k!)$, where C is the number of channels and $A = U\mu H$ for U the number of users, μ the average number of call requests per unit time, and H the average duration of a call. See Section 3.6 of Rappaport, EE276 notes, or any basic networks book for more details).

(c) How many total subscribers can be supported in the macrocell system (1 square Km cells) and in the microcell system (with cell size from part (b))?

(d) If a base station costs $500,000, what are the base station costs for each system?

(e) If users pay 50 dollars a month in both systems, what will be the montly revenue in each case. How long will it take to recoup the infrastructure (base station) cost for each system?

11. How many CDPD data lines are needed to achieve the same data rate as the average rate of Wi-Max?

Chapter 3

Statistical Multipath Channel Models

In this chapter we examine fading models for the constructive and destructive addition of different multipath components introduced by the channel. While these multipath effects are captured in the ray-tracing models from Chapter 2 for deterministic channels, in practice deterministic channel models are rarely available, and thus we must characterize multipath channels statistically. In this chapter we model the multipath channel by a random time-varying impulse response. We will develop a statistical characterization of this channel model and describe its important properties.

If a single pulse is transmitted over a multipath channel the received signal will appear as a pulse train, with each pulse in the train corresponding to the LOS component or a distinct multipath component associated with a distinct scatterer or cluster of scatterers. An important characteristic of a multipath channel is the time delay spread it causes to the received signal. This delay spread equals the time delay between the arrival of the first received signal component (LOS or multipath) and the last received signal component associated with a single transmitted pulse. If the delay spread is small compared to the inverse of the signal bandwidth, then there is little time spreading in the received signal. However, when the delay spread is relatively large, there is significant time spreading of the received signal which can lead to substantial signal distortion.

Another characteristic of the multipath channel is its time-varying nature. This time variation arises because either the transmitter or the receiver is moving, and therefore the location of reflectors in the transmission path, which give rise to multipath, will change over time. Thus, if we repeatedly transmit pulses from a moving transmitter, we will observe changes in the amplitudes, delays, and the number of multipath components corresponding to each pulse. However, these changes occur over a much larger time scale than the fading due to constructive and destructive addition of multipath components associated with a fixed set of scatterers. We will first use a generic time-varying channel impulse response to capture both fast and slow channel variations. We will then restrict this model to narrowband fading, where the channel bandwidth is small compared to the inverse delay spread. For this narrowband model we will assume a quasi-static environment with a fixed number of multipath components each with fixed path loss and shadowing. For this quasi-static environment we then characterize the variations over short distances (small-scale variations) due to the constructive and destructive addition of multipath components. We also characterize the statistics of wideband multipath channels using two-dimensional transforms based on the underlying time-varying impulse response. Discrete-time and space-time channel models are also discussed.

3.1 Time-Varying Channel Impulse Response

Let the transmitted signal be as in Chapter 2:

$$s(t) = \Re\left\{u(t)e^{j2\pi f_c t}\right\} = \Re\left\{u(t)\right\}\cos(2\pi f_c t) - \Im\left\{u(t)\right\}\sin(2\pi f_c t), \tag{3.1}$$

where $u(t)$ is the complex envelope of $s(t)$ with bandwidth B_u and f_c is its carrier frequency. The corresponding received signal is the sum of the line-of-sight (LOS) path and all resolvable multipath components:

$$r(t) = \Re\left\{\sum_{n=0}^{N(t)} \alpha_n(t)u(t-\tau_n(t))e^{j(2\pi f_c(t-\tau_n(t))+\phi_{D_n})}\right\}, \tag{3.2}$$

where $n = 0$ corresponds to the LOS path. The unknowns in this expression are the number of resolvable multipath components $N(t)$, discussed in more detail below, and for the LOS path and each multipath component, its path length $r_n(t)$ and corresponding delay $\tau_n(t) = r_n(t)/c$, Doppler phase shift $\phi_{D_n}(t)$ and amplitude $\alpha_n(t)$.

The nth resolvable multipath component may correspond to the multipath associated with a single reflector or with multiple reflectors clustered together that generate multipath components with similar delays, as shown in Figure 3.1. If each multipath component corresponds to just a single reflector then its corresponding amplitude $\alpha_n(t)$ is based on the path loss and shadowing associated with that multipath component, its phase change associated with delay $\tau_n(t)$ is $e^{-j2\pi f_c\tau_n(t)}$, and its Doppler shift $f_{D_n}(t) = v\cos\theta_n(t)/lambda$ for $\theta_n(t)$ its angle of arrival. This Doppler frequency shift leads to a Doppler phase shift of $\phi_{D_n} = \int_t 2\pi f_{D_n}(t)dt$. Suppose, however, that the nth multipath component results from a reflector cluster[1]. We say that two multipath components with delay τ_1 and τ_2 are **resolvable** if their delay difference significantly exceeds the inverse signal bandwidth: $|\tau_1 - \tau_2| >> B_u^{-1}$. Multipath components that do not satisfy this resolvability criteria cannot be separated out at the receiver, since $u(t-\tau_1) \approx u(t-\tau_2)$, and thus these components are **nonresolvable**. These nonresolvable components are combined into a single multipath component with delay $\tau \approx \tau_1 \approx \tau_2$ and an amplitude and phase corresponding to the sum of the different components. The amplitude of this summed signal will typically undergo fast variations due to the constructive and destructive combining of the nonresolvable multipath components. In general wideband channels have resolvable multipath components so that each term in the summation of (3.2) corresponds to a single reflection or multiple nonresolvable components combined together, whereas narrowband channels tend to have nonresolvable multipath components contributing to each term in (3.2).

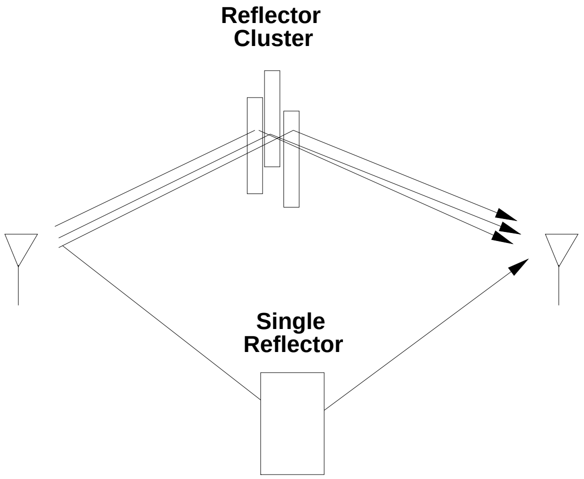


Figure 3.1: A Single Reflector and A Reflector Cluster.

Since the parameters $\alpha_n(t)$, $\tau_n(t)$, and $\phi_{D_n}(t)$ associated with each resolvable multipath component change over time, they are characterized as random processes which we assume to be both stationary and ergodic. Thus, the received signal is also a stationary and ergodic random process. For wideband channels, where each term in

[1] Equivalently, a single "rough" reflector can create different multipath components with slightly different delays.

(3.2) corresponds to a single reflector, these parameters change slowly as the propagation environment changes. For narrowband channels, where each term in (3.2) results from the sum of nonresolvable multipath components, the parameters can change quickly, on the order of a signal wavelength, due to constructive and destructive addition of the different components.

We can simplify $r(t)$ by letting

$$\phi_n(t) = 2\pi f_c \tau_n(t) - \phi_{D_n}. \tag{3.3}$$

Then the received signal can be rewritten as

$$r(t) = \Re\left\{ \left[\sum_{n=0}^{N(t)} \alpha_n(t) e^{-j\phi_n(t)} u(t - \tau_n(t)) \right] e^{j2\pi f_c t} \right\}. \tag{3.4}$$

Since $\alpha_n(t)$ is a function of path loss and shadowing while $\phi_n(t)$ depends on delay and Doppler, we typically assume that these two random processes are independent.

The received signal $r(t)$ is obtained by convolving the baseband input signal $u(t)$ with the equivalent lowpass time-varying channel impulse response $c(\tau, t)$ of the channel and then upconverting to the carrier frequency[2]:

$$r(t) = \Re\left\{ \left(\int_{-\infty}^{\infty} c(\tau, t) u(t - \tau) d\tau \right) e^{j2\pi f_c t} \right\}. \tag{3.5}$$

Note that $c(\tau, t)$ has two time parameters: the time t when the impulse response is observed at the receiver, and the time $t - \tau$ when the impulse is launched into the channel relative to the observation time t. If at time t there is no physical reflector in the channel with multipath delay $\tau_n(t) = \tau$ then $c(\tau, t) = 0$. While the definition of the time-varying channel impulse response might seem counterintuitive at first, $c(\tau, t)$ must be defined in this way to be consistent with the special case of time-invariant channels. Specifically, for time-invariant channels we have $c(\tau, t) = c(\tau, t + T)$, i.e. the response at time t to an impulse at time $t - \tau$ equals the response at time $t + T$ to an impulse at time $t + T - \tau$. Setting $T = -t$, we get that $c(\tau, t) = c(\tau, t - t) = c(\tau)$, where $c(\tau)$ is the standard time-invariant channel impulse response: the response at time τ to an impulse at zero or, equivalently, the response at time zero to an impulse at time $-\tau$.

We see from (3.4) and (3.5) that $c(\tau, t)$ must be given by

$$c(\tau, t) = \sum_{n=0}^{N(t)} \alpha_n(t) e^{-j\phi_n(t)} \delta(\tau - \tau_n(t)), \tag{3.6}$$

where $c(\tau, t)$ represents the equivalent lowpass response of the channel at time t to an impulse at time $t - \tau$. Substituting (3.6) back into (3.5) yields (3.4), thereby confirming that (3.6) is the channel's equivalent lowpass

[2] See Appendix A for discussion of the lowpass equivalent representation for bandpass signals and systems.

time-varying impulse response:

$$\begin{aligned}
r(t) &= \Re\left\{\left[\int_{-\infty}^{\infty} c(\tau,t)u(t-\tau)d\tau\right] e^{j2\pi f_c t}\right\} \\
&= \Re\left\{\left[\int_{-\infty}^{\infty}\sum_{n=0}^{N(t)} \alpha_n(t)e^{-j\phi_n(t)}\delta(\tau-\tau_n(t))u(t-\tau)d\tau\right] e^{j2\pi f_c t}\right\} \\
&= \Re\left\{\left[\sum_{n=0}^{N(t)} \alpha_n(t)e^{-j\phi_n(t)}\left(\int_{-\infty}^{\infty}\delta(\tau-\tau_n(t))u(t-\tau)d\tau\right)\right] e^{j2\pi f_c t}\right\} \\
&= \Re\left\{\left[\sum_{n=0}^{N(t)} \alpha_n(t)e^{-j\phi_n(t)}u(t-\tau_n(t))\right] e^{j2\pi f_c t}\right\},
\end{aligned}$$

where the last equality follows from the sifting property of delta functions: $\int \delta(\tau-\tau_n(t))u(t-\tau)d\tau = \delta(t-\tau_n(t)) * u(t) = u(t-\tau_n(t))$. Some channel models assume a continuum of multipath delays, in which case the sum in (3.6) becomes an integral which simplifies to a time-varying complex amplitude associated with each multipath delay τ:

$$c(\tau,t) = \int \alpha(\xi,t)e^{-j\phi(\xi,t)}\delta(\tau-\xi)d\xi = \alpha(\tau,t)e^{-j\phi(\tau,t)}. \tag{3.7}$$

To give a concrete example of a time-varying impulse response, consider the system shown in Figure 3.2, where each multipath component corresponds to a single reflector. At time t_1 there are three multipath components associated with the received signal with amplitude, phase, and delay triple $(\alpha_i, \phi_i, \tau_i)$, $i = 1,2,3$. Thus, impulses that were launched into the channel at time $t_1 - \tau_i, i = 1,2,3$ will all be received at time t_1, and impulses launched into the channel at any other time will not be received at t_1 (because there is no multipath component with the corresponding delay). The time-varying impulse response corresponding to t_1 equals

$$c(\tau,t_1) = \sum_{n=0}^{2} \alpha_n e^{-j\phi_n}\delta(\tau-\tau_n) \tag{3.8}$$

and the channel impulse response for $t = t_1$ is shown in Figure 3.3. Figure 3.2 also shows the system at time t_2, where there are two multipath components associated with the received signal with amplitude, phase, and delay triple $(\alpha_i', \phi_i', \tau_i')$, $i = 1,2$. Thus, impulses that were launched into the channel at time $t_2 - \tau_i', i = 1,2$ will all be received at time t_2, and impulses launched into the channel at any other time will not be received at t_2. The time-varying impulse response at t_2 equals

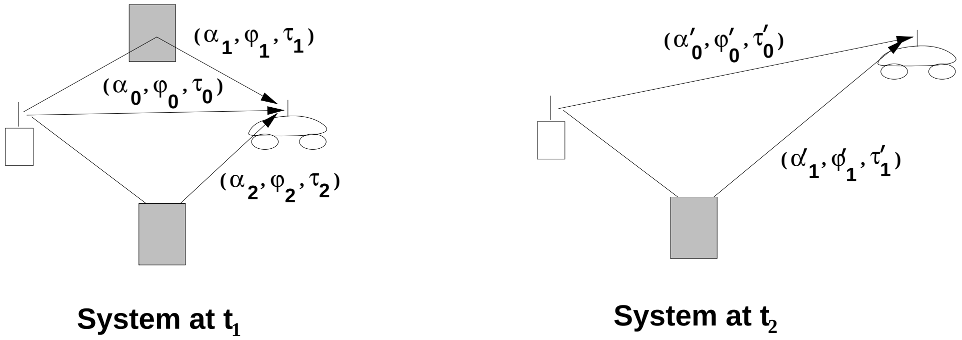
$$c(\tau,t_2) = \sum_{n=0}^{1} \alpha_n' e^{-j\phi_n'}\delta(\tau-\tau_n') \tag{3.9}$$

and is also shown in Figure 3.3.

If the channel is time-invariant then the time-varying parameters in $c(\tau,t)$ become constant, and $c(\tau,t) = c(\tau)$ is just a function of τ:

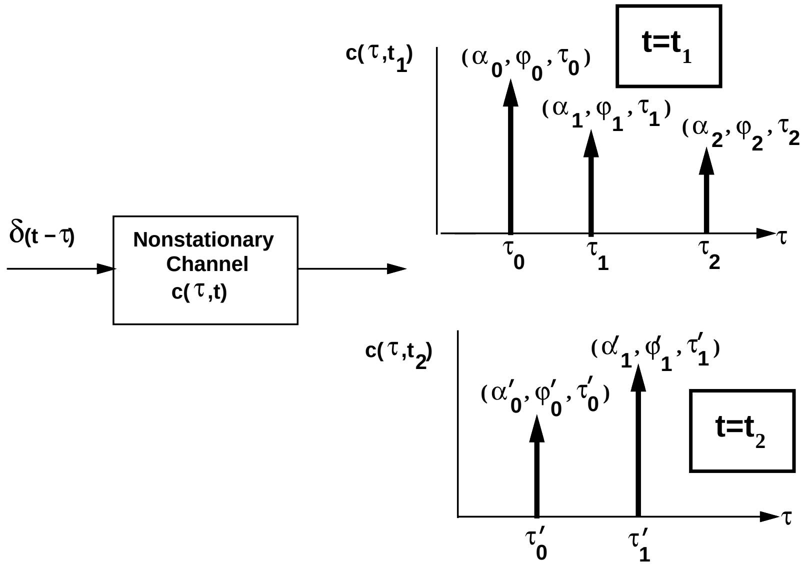
$$c(\tau) = \sum_{n=0}^{N} \alpha_n e^{-j\phi_n}\delta(\tau-\tau_n), \tag{3.10}$$

for channels with discrete multipath components, and $c(\tau) = \alpha(\tau)e^{-j\phi(\tau)}$ for channels with a continuum of multipath components. For stationary channels the response to an impulse at time t_1 is just a shifted version of its response to an impulse at time $t_2, t_1 \neq t_2$.



$(\alpha_1, \varphi_1, \tau_1)$
$(\alpha_0, \varphi_0, \tau_0)$
$(\alpha_2, \varphi_2, \tau_2)$
$(\alpha'_0, \varphi'_0, \tau'_0)$
$(\alpha'_1, \phi'_1, \tau'_1)$
System at t_1
System at t_2

Figure 3.2: System Multipath at Two Different Measurement Times.



$\delta(t-\tau)$
Nonstationary Channel $c(\tau,t)$
$c(\tau,t_1)$
$t=t_1$
$(\alpha_0, \varphi_0, \tau_0)$
$(\alpha_1, \varphi_1, \tau_1)$
$(\alpha_2, \varphi_2, \tau_2)$
τ_0
τ_1
τ_2
τ
$c(\tau,t_2)$
$(\alpha'_1, \phi'_1, \tau'_1)$
$(\alpha'_0, \varphi'_0, \tau'_0)$
$t=t_2$
τ'_0
τ'_1
τ

Figure 3.3: Response of Nonstationary Channel.

Example 3.1: Consider a wireless LAN operating in a factory near a conveyor belt. The transmitter and receiver have a LOS path between them with gain α_0, phase ϕ_0 and delay τ_0. Every T_0 seconds a metal item comes down the conveyor belt, creating an additional reflected signal path in addition to the LOS path with gain α_1, phase ϕ_1 and delay τ_1. Find the time-varying impulse response $c(\tau, t)$ of this channel.

Solution: For $t \neq nT_0$, $n = 1, 2, \ldots$ the channel impulse response corresponds to just the LOS path. For $t = nT_0$ the channel impulse response has both the LOS and reflected paths. Thus, $c(\tau, t)$ is given by

$$c(\tau, t) = \begin{cases} \alpha_0 e^{j\phi_0}\delta(\tau - \tau_0) & t \neq nT_0 \\ \alpha_0 e^{j\phi_0}\delta(\tau - \tau_0) + \alpha_1 e^{j\phi_1}\delta(\tau - \tau_1) & t = nT_0 \end{cases}$$

Note that for typical carrier frequencies, the nth multipath component will have $f_c\tau_n(t) >> 1$. For example, with $f_c = 1$ GHz and $\tau_n = 50$ ns (a typical value for an indoor system), $f_c\tau_n = 50 >> 1$. Outdoor wireless systems have multipath delays much greater than 50 ns, so this property also holds for these systems. If $f_c\tau_n(t) >> 1$ then a small change in the path delay $\tau_n(t)$ can lead to a very large phase change in the nth multipath component with phase $\phi_n(t) = 2\pi f_c\tau_n(t) - \phi_{D_n} - \phi_0$. Rapid phase changes in each multipath component gives rise to constructive and destructive addition of the multipath components comprising the received signal, which in turn causes rapid variation in the received signal strength. This phenomenon, called *fading*, will be discussed in more detail in subsequent sections.

The impact of multipath on the received signal depends on whether the spread of time delays associated with the LOS and different multipath components is large or small relative to the inverse signal bandwidth. If this channel delay spread is small then the LOS and all multipath components are typically nonresolvable, leading to the narrowband fading model described in the next section. If the delay spread is large then the LOS and all multipath components are typically resolvable into some number of discrete components, leading to the wideband fading model of Section 3.3. Note that some of the discrete components in the wideband model are comprised of nonresolvable components. The delay spread is typically measured relative to the received signal component to which the demodulator is synchronized. Thus, for the time-invariant channel model of (3.10), if the demodulator synchronizes to the LOS signal component, which has the smallest delay τ_0, then the delay spread is a constant given by $T_m = \max_n \tau_n - \tau_0$. However, if the demodulator synchronizes to a multipath component with delay equal to the mean delay $\overline{\tau}$ then the delay spread is given by $T_m = \max_n |\tau_n - \overline{\tau}|$. In time-varying channels the multipath delays vary with time, so the delay spread T_m becomes a random variable. Moreover, some received multipath components have significantly lower power than others, so it's not clear how the delay associated with such components should be used in the characterization of delay spread. In particular, if the power of a multipath component is below the noise floor then it should not significantly contribute to the delay spread. These issues are typically dealt with by characterizing the delay spread relative to the channel power delay profile, defined in Section 3.3.1. Specifically, two common characterizations of channel delay spread, average delay spread and rms delay spread, are determined from the power delay profile. Other characterizations of delay spread, such as excees delay spread, the delay window, and the delay interval, are sometimes used as well [6, Chapter 5.4.1],[28, Chapter 6.7.1]. The exact characterization of delay spread is not that important for understanding the general impact of delay spread on multipath channels, as long as the characterization roughly measures the delay associated with significant multipath components. In our development below any reasonable characterization of delay spread T_m can be used, although we will typically use the rms delay spread. This is the most common characterization since, assuming the demodulator synchronizes to a signal component at the average delay spread, the rms delay spread is a good measure of the variation about this average. Channel delay spread is highly dependent on the propagation environment. In indoor channels delay spread typically ranges from 10 to 1000 nanoseconds, in suburbs it ranges from 200-2000 nanoseconds, and in urban areas it ranges from 1-30 microseconds [6].

3.2 Narrowband Fading Models

Suppose the delay spread T_m of a channel is small relative to the inverse signal bandwidth B of the transmitted signal, i.e. $T_m << B^{-1}$. As discussed above, the delay spread T_m for time-varying channels is usually characterized by the rms delay spread, but can also be characterized in other ways. Under most delay spread characterizations $T_m << B^{-1}$ implies that the delay associated with the ith multipath component $\tau_i \leq T_m \forall i$, so $u(t-\tau_i) \approx u(t) \forall i$ and we can rewrite (3.4) as

$$r(t) = \Re\left\{u(t)e^{j2\pi f_c t}\left(\sum_n \alpha_n(t)e^{-j\phi_n(t)}\right)\right\}. \tag{3.11}$$

Equation (3.11) differs from the original transmitted signal by the complex scale factor in parentheses. This scale factor is independent of the transmitted signal $s(t)$ or, equivalently, the baseband signal $u(t)$, as long as the narrowband assumption $T_m << 1/B$ is satisfied. In order to characterize the random scale factor caused by the multipath we choose $s(t)$ to be an unmodulated carrier with random phase offset ϕ_0:

$$s(t) = \Re\{e^{j(2\pi f_c t+\phi_0)}\} = \cos(2\pi f_c t - \phi_0), \tag{3.12}$$

which is narrowband for any T_m.

With this assumption the received signal becomes

$$r(t) = \Re\left\{\left[\sum_{n=0}^{N(t)} \alpha_n(t)e^{-j\phi_n(t)}\right] e^{j2\pi f_c t}\right\} = r_I(t)\cos 2\pi f_c t + r_Q(t)\sin 2\pi f_c t, \tag{3.13}$$

where the in-phase and quadrature components are given by

$$r_I(t) = \sum_{n=1}^{N(t)} \alpha_n(t)\cos\phi_n(t), \tag{3.14}$$

and

$$r_Q(t) = \sum_{n=1}^{N(t)} \alpha_n(t)\sin\phi_n(t), \tag{3.15}$$

where the phase term

$$\phi_n(t) = 2\pi f_c \tau_n(t) - \phi_{D_n} - \phi_0 \tag{3.16}$$

now incorporates the phase offset ϕ_0 as well as the effects of delay and Doppler.

If $N(t)$ is large we can invoke the Central Limit Theorem and the fact that $\alpha_n(t)$ and $\phi_n(t)$ are stationary and ergodic to approximate $r_I(t)$ and $r_Q(t)$ as jointly Gaussian random processes. The Gaussian property is also true for small N if the $\alpha_n(t)$ are Rayleigh distributed and the $\phi_n(t)$ are uniformly distributed on $[-\pi, \pi]$. This happens when the nth multipath component results from a reflection cluster with a large number of nonresolvable multipath components [1].

3.2.1 Autocorrelation, Cross Correlation, and Power Spectral Density

We now derive the autocorrelation and cross correlation of the in-phase and quadrature received signal components $r_I(t)$ and $r_Q(t)$. Our derivations are based on some key assumptions which generally apply to propagation models without a dominant LOS component. Thus, these formulas are not typically valid when a dominant LOS component exists. We assume throughout this section that the amplitude $\alpha_n(t)$, multipath delay $\tau_n(t)$ and Doppler frequency $f_{D_n}(t)$ are changing slowly enough such that they are constant over the time intervals of interest: $\alpha_n(t) \approx \alpha_n$, $\tau_n(t) \approx \tau_n$, and $f_{D_n}(t) \approx f_{D_n}$. This will be true when each of the resolvable multipath components is associated with a single reflector. With this assumption the Doppler phase shift is[3] $\phi_{D_n}(t) = \int_t 2\pi f_{D_n} dt = 2\pi f_{D_n} t$, and the phase of the nth multipath component becomes $\phi_n(t) = 2\pi f_c \tau_n - 2\pi f_{D_n} t - \phi_0$.

We now make a *key* assumption: we assume that for the nth multipath component the term $2\pi f_c \tau_n$ in $\phi_n(t)$ changes rapidly relative to all other phase terms in this expression. This is a reasonable assumption since f_c is large and hence the term $2\pi f_c \tau_n$ can go through a 360 degree rotation for a small change in multipath delay τ_n. Under this assumption $\phi_n(t)$ is uniformly distributed on $[-\pi, \pi]$. Thus

$$\mathrm{E}[r_I(t)] = \mathrm{E}[\sum_n \alpha_n \cos\phi_n(t)] = \sum_n \mathrm{E}[\alpha_n]\mathrm{E}[\cos\phi_n(t)] = 0, \tag{3.17}$$

[3] We assume a Doppler phase shift at $t = 0$ of zero for simplicity, since this phase offset will not affect the analysis.

where the second equality follows from the independence of α_n and ϕ_n and the last equality follows from the uniform distribution on ϕ_n. Similarly we can show that $\mathrm{E}[r_Q(t)] = 0$. Thus, the received signal also has $\mathrm{E}[r(t)] = 0$, i.e. it is a zero-mean Gaussian process. When there is a dominant LOS component in the channel the phase of the received signal is dominated by the phase of the LOS component, which can be determined at the receiver, so the assumption of a random uniform phase no longer holds.

Consider now the autocorrelation of the in-phase and quadrature components. Using the independence of α_n and ϕ_n, the independence of ϕ_n and ϕ_m, $n \neq m$, and the uniform distribution of ϕ_n we get that

$$\begin{aligned}
\mathrm{E}[r_I(t)r_Q(t)] &= \mathrm{E}\left[\sum_n \alpha_n \cos\phi_n(t) \sum_m \alpha_m \sin\phi_m(t)\right] \\
&= \sum_n \sum_m \mathrm{E}[\alpha_n\alpha_m]\mathrm{E}[\cos\phi_n(t)\sin\phi_m(t)] \\
&= \sum_n \mathrm{E}[\alpha_n^2]\mathrm{E}[\cos\phi_n(t)\sin\phi_n(t)] \\
&= 0. \qquad (3.18)
\end{aligned}$$

Thus, $r_I(t)$ and $r_Q(t)$ are uncorrelated and, since they are jointly Gaussian processes, this means they are independent.

Following a similar derivation as in (3.18) we obtain the autocorrelation of $r_I(t)$ as

$$A_{r_I}(t,\tau) = \mathrm{E}[r_I(t)r_I(t+\tau)] = \sum_n \mathrm{E}[\alpha_n^2]\mathrm{E}[\cos\phi_n(t)\cos\phi_n(t+\tau)]. \qquad (3.19)$$

Now making the substitution $\phi_n(t) = 2\pi f_c\tau_n - 2\pi f_{D_n}t - \phi_0$ and $\phi_n(t+\tau) = 2\pi f_c\tau_n - 2\pi f_{D_n}(t+\tau) - \phi_0$ we get

$$\mathrm{E}[\cos\phi_n(t)\cos\phi_n(t+\tau)] = .5\mathrm{E}[\cos 2\pi f_{D_n}\tau] + .5\mathrm{E}[\cos(4\pi f_c\tau_n + -4\pi f_{D_n}t - 2\pi f_{D_n}\tau - 2\phi_0)]. \qquad (3.20)$$

Since $2\pi f_c\tau_n$ changes rapidly relative to all other phase terms and is uniformly distributed, the second expectation term in (3.20) goes to zero, and thus

$$A_{r_I}(t,\tau) = .5\sum_n \mathrm{E}[\alpha_n^2]\mathrm{E}[\cos(2\pi f_{D_n}\tau)] = .5\sum_n \mathrm{E}[\alpha_n^2]\cos(2\pi v\tau\cos\theta_n/\lambda), \qquad (3.21)$$

since $f_{D_n} = v\cos\theta_n/\lambda$ is assumed fixed. Note that $A_{r_I}(t,\tau)$ depends only on τ, $A_{r_I}(t,\tau) = A_{r_I}(\tau)$, and thus $r_I(t)$ is a wide-sense stationary (WSS) random process.

Using a similar derivation we can show that the quadrature component is also WSS with autocorrelation $A_{r_Q}(\tau) = A_{r_I}(\tau)$. In addition, the cross correlation between the in-phase and quadrature components depends only on the time difference τ and is given by

$$A_{r_I,r_Q}(t,\tau) = A_{r_I,r_Q}(\tau) = \mathrm{E}[r_I(t)r_Q(t+\tau)] = -.5\sum_n \mathrm{E}[\alpha_n^2]\sin(2\pi v\tau\cos\theta_n/\lambda) = -\mathrm{E}[r_Q(t)r_I(t+\tau)]. \qquad (3.22)$$

Using these results we can show that the received signal $r(t) = r_I(t)\cos(2\pi f_c t) + r_Q(t)\sin(2\pi f_c t)$ is also WSS with autocorrelation

$$A_r(\tau) = \mathrm{E}[r(t)r(t+\tau)] = A_{r_I}(\tau)\cos(2\pi f_c\tau) + A_{r_I,r_Q}(\tau)\sin(2\pi f_c\tau). \qquad (3.23)$$

In order to further simplify (3.21) and (3.22), we must make additional assumptions about the propagation environment. We will focus on the **uniform scattering environment** introduced by Clarke [4] and further developed by Jakes [Chapter 1][5]. In this model, the channel consists of many scatterers densely packed with respect to angle, as shown in Fig. 3.4. Thus, we assume N multipath components with angle of arrival $\theta_n = n\Delta\theta$, where $\Delta\theta = 2\pi/N$. We also assume that each multipath component has the same received power, so $\mathrm{E}[\alpha_n^2] = 2P_r/N$, where P_r is the total received power. Then (3.21) becomes

$$A_{r_I}(\tau) = \frac{P_r}{N}\sum_{n=1}^{N}\cos(2\pi v\tau\cos n\Delta\theta/\lambda). \tag{3.24}$$

Now making the substitution $N = 2\pi/\Delta\theta$ yields

$$A_{r_I}(\tau) = \frac{P_r}{2\pi}\sum_{n=1}^{N}\cos(2\pi v\tau\cos n\Delta\theta/\lambda)\Delta\theta. \tag{3.25}$$

We now take the limit as the number of scatterers grows to infinity, which corresponds to uniform scattering from all directions. Then $N \to \infty$, $\Delta\theta \to 0$, and the summation in (3.25) becomes an integral:

$$A_{r_I}(\tau) = \frac{P_r}{2\pi}\int\cos(2\pi v\tau\cos\theta/\lambda)d\theta = P_r J_0(2\pi f_D\tau), \tag{3.26}$$

where

$$J_0(x) = \frac{1}{\pi}\int_0^\pi e^{-jx\cos\theta}d\theta$$

is a Bessel function of the 0th order[4]. Similarly, for this uniform scattering environment,

$$A_{r_I,r_Q}(\tau) = \frac{P_r}{2\pi}\int\sin(2\pi v\tau\cos\theta/\lambda)d\theta = 0. \tag{3.27}$$

A plot of $J_0(2\pi f_D\tau)$ is shown in Figure 3.5. There are several interesting observations from this plot. First we see that the autocorrelation is zero for $f_D\tau \approx .4$ or, equivalently, for $v\tau \approx .4\lambda$. Thus, the signal decorrelates over a distance of approximately one half wavelength, under the uniform θ_n assumption. This approximation is commonly used as a rule of thumb to determine many system parameters of interest. For example, we will see in Chapter 7 that obtaining independent fading paths can be exploited by antenna diversity to remove some of the negative effects of fading. The antenna spacing must be such that each antenna receives an independent fading path and therefore, based on our analysis here, an antenna spacing of $.4\lambda$ should be used. Another interesting characteristic of this plot is that the signal recorrelates after it becomes uncorrelated. Thus, we cannot assume that the signal remains independent from its initial value at $d = 0$ for separation distances greater than $.4\lambda$. As a result, a Markov model is not completely accurate for Rayleigh fading, because of this recorrelation property. However, in many system analyses a correlation below .5 does not significantly degrade performance relative to uncorrelated fading [8, Chapter 9.6.5]. For such studies the fading process can be modeled as Markov by assuming that once the correlation is close to zero, i.e. the separation distance is greater than a half wavelength, the signal remains decorrelated at all larger distances.

[4] Note that (3.26) can also be derived by assuming $2\pi v\tau\cos\theta_n/\lambda$ in (3.21) and (3.22) is random with θ_n uniformly distributed, and then taking expectation with respect to θ_n. However, based on the underlying physical model, θ_n can only be uniformly distributed in a dense scattering environment. So the derivations are equivalent.

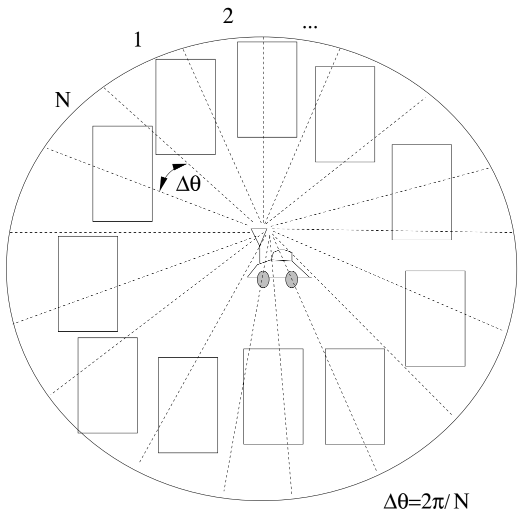


Figure 3.4: Dense Scattering Environment

The power spectral densities (PSDs) of $r_I(t)$ and $r_Q(t)$, denoted by $S_{r_I}(f)$ and $S_{r_Q}(f)$, respectively, are obtained by taking the Fourier transform of their respective autocorrelation functions relative to the delay parameter τ. Since these autocorrelation functions are equal, so are the PSDs. Thus

$$S_{r_I}(f) = S_{r_Q}(f) = \mathcal{F}[A_{r_I}(\tau)] = \begin{cases} \frac{P_r}{2\pi f_D} \frac{1}{\sqrt{1-(f/f_D)^2}} & |f| \leq f_D \\ 0 & \text{else} \end{cases} \tag{3.28}$$

This PSD is shown in Figure 3.6.

To obtain the PSD of the received signal $r(t)$ under uniform scattering we use (3.23) with $A_{r_I,r_Q}(\tau) = 0$, (3.28), and simple properties of the Fourier transform to obtain

$$S_r(f) = \mathcal{F}[A_r(\tau)] = .25[S_{r_I}(f - f_c) + S_{r_I}(f + f_c)] = \begin{cases} \frac{P_r}{4\pi f_D} \frac{1}{\sqrt{1-\left(\frac{|f-f_c|}{f_D}\right)^2}} & |f - f_c| \leq f_D \\ 0 & \text{else} \end{cases}, \tag{3.29}$$

Note that this PSD integrates to P_r, the total received power.

Since the PSD models the power density associated with multipath components as a function of their Doppler frequency, it can be viewed as the distribution (pdf) of the random frequency due to Doppler associated with multipath. We see from Figure 3.6 that the PSD $S_{r_i}(f)$ goes to infinity at $f = \pm f_D$ and, consequently, the PSD $S_r(f)$ goes to infinity at $f = \pm f_c \pm f_D$. This will not be true in practice, since the uniform scattering model is just an approximation, but for environments with dense scatterers the PSD will generally be maximized at frequencies close to the maximum Doppler frequency. The intuition for this behavior comes from the nature of the cosine function and the fact that under our assumptions the PSD corresponds to the pdf of the random Doppler frequency $f_D(\theta)$. To see this, note that the uniform scattering assumption is based on many scattered paths arriving uniformly from all angles with the same average power. Thus, θ for a randomly selected path can be regarded as a uniform random variable on $[0, 2\pi]$. The distribution $p_{f_\theta}(f)$ of the random Doppler frequency $f(\theta)$ can then be obtained from the distribution of θ. By definition, $p_{f_\theta}(f)$ is proportional to the density of scatterers at Doppler frequency f. Hence, $S_{r_I}(f)$ is also proportional to this density, and we can characterize the PSD from the pdf $p_{f_\theta}(f)$. For this characterization, in Figure 3.7 we plot $f_D(\theta) = f_D \cos(\theta) = v/\lambda \cos(\theta)$ along with a dotted line straight-line segment approximation $\underline{f}_D(\theta)$ to $f_D(\theta)$. On the right in this figure we plot the PSD $S_{r_i}(f)$ along with a dotted

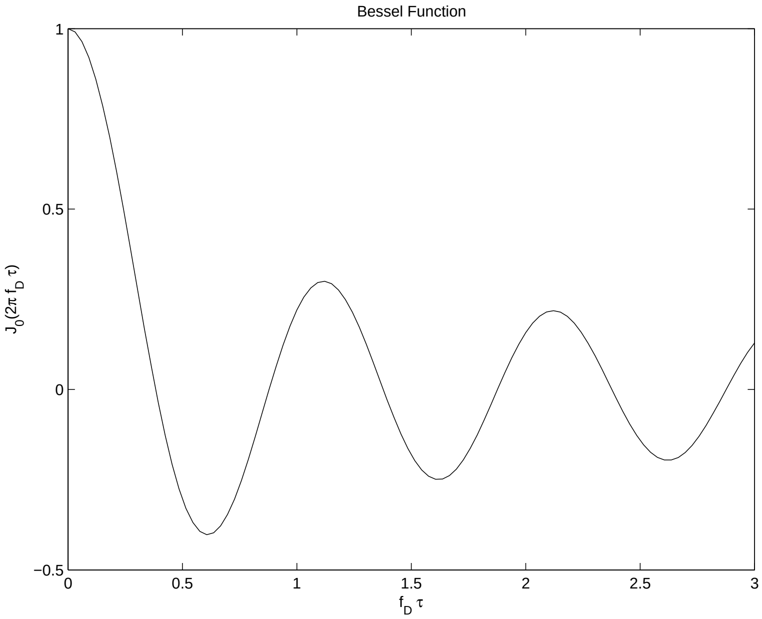


Figure 3.5: Bessel Function versus $f_d\tau$

line straight line segment approximation to it $\underline{S}_{r_i}(f)$, which corresponds to the Doppler approximation $\underline{f}_D(\theta)$. We see that $\cos(\theta) \approx \pm 1$ for a relatively large range of θ values. Thus, multipath components with angles of arrival in this range of values have Doppler frequency $f_D(\theta) \approx \pm f_D$, so the power associated with all of these multipath components will add together in the PSD at $f \approx f_D$. This is shown in our approximation by the fact that the segments where $\underline{f}_D(\theta) = \pm f_D$ on the left lead to delta functions at $\pm f_D$ in the pdf approximation $\underline{S}_{r_i}(f)$ on the right. The segments where $\underline{f}_D(\theta)$ has uniform slope on the left lead to the flat part of $\underline{S}_{r_i}(f)$ on the right, since there is one multipath component contributing power at each angular increment. Formulas for the autocorrelation and PSD in nonuniform scattering, corresponding to more typical microcell and indoor environments, can be found in [5, Chapter 1], [11, Chapter 2].

The PSD is useful in constructing simulations for the fading process. A common method for simulating the envelope of a narrowband fading process is to pass two independent white Gaussian noise sources with PSD $N_0/2$ through lowpass filters with frequency response $H(f)$ that satisfies

$$S_{r_I}(f) = S_{r_Q}(f) = \frac{N_0}{2}|H(f)|^2. \tag{3.30}$$

The filter outputs then correspond to the in-phase and quadrature components of the narrowband fading process with PSDs $S_{r_I}(f)$ and $S_{r_Q}(f)$. A similar procedure using discrete filters can be used to generate discrete fading processes. Most communication simulation packages (e.g. Matlab, COSSAP) have standard modules that simulate narrowband fading based on this method. More details on this simulation method, as well as alternative methods, can be found in [11, 6, 7].

We have now completed our model for the three characteristics of power versus distance exhibited in narrowband wireless channels. These characteristics are illustrated in Figure 3.8, adding narrowband fading to the path loss and shadowing models developed in Chapter 2. In this figure we see the decrease in signal power due to path loss decreasing as d^γ with γ the path loss exponent, the more rapid variations due to shadowing which change on the order of the decorrelation distance X_c, and the very rapid variations due to multipath fading which change on the order of half the signal wavelength. If we blow up a small segment of this figure over distances where path loss

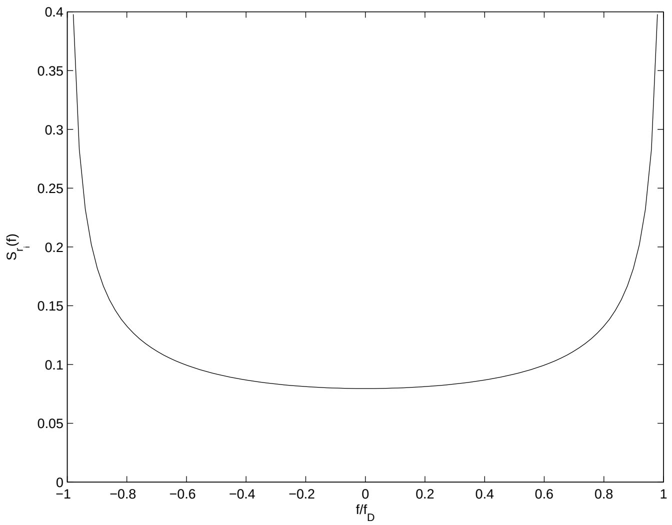


Figure 3.6: In-Phase and Quadrature PSD: $S_{r_I}(f) = S_{r_Q}(f)$

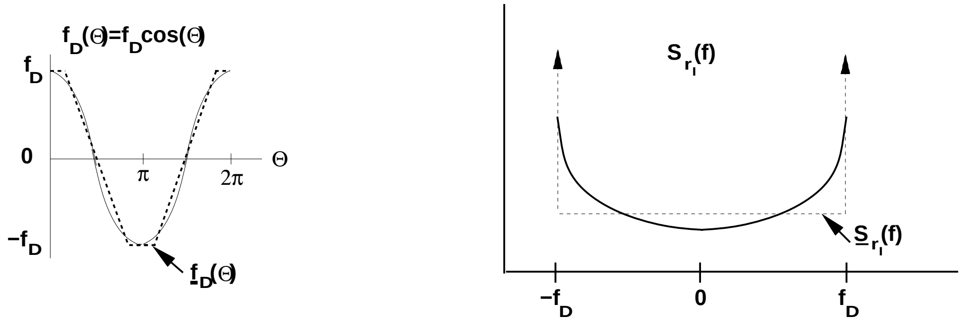


Figure 3.7: Cosine and PSD Approximation by Straight Line Segments

and shadowing are constant we obtain Figure 3.9, where we show dB fluctuation in received power versus linear distance $d = vt$ (not log distance). In this figure the average received power P_r is normalized to 0 dBm. A mobile receiver traveling at fixed velocity v would experience the received power variations over time illustrated in this figure.

3.2.2 Envelope and Power Distributions

For any two Gaussian random variables X and Y, both with mean zero and equal variance σ^2, it can be shown that $Z = \sqrt{X^2 + Y^2}$ is Rayleigh-distributed and Z^2 is exponentially distributed. We saw above that for $\phi_n(t)$ uniformly distributed, r_I and r_Q are both zero-mean Gaussian random variables. If we assume a variance of σ^2 for both in-phase and quadrature components then the signal envelope

$$z(t) = |r(t)| = \sqrt{r_I^2(t) + r_Q^2(t)} \tag{3.31}$$

is Rayleigh-distributed with distribution

$$p_Z(z) = \frac{2z}{P_r}\text{exp}[-z^2/P_r] = \frac{z}{\sigma^2}\text{exp}[-z^2/(2\sigma^2)], \quad x \geq 0, \tag{3.32}$$

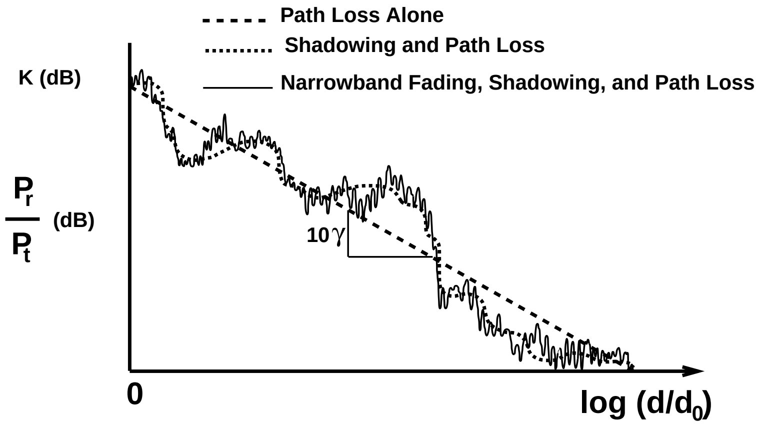


Figure 3.8: Combined Path Loss, Shadowing, and Narrowband Fading.

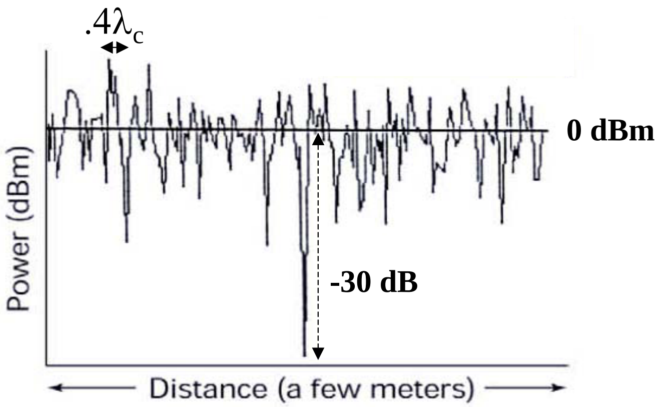


Figure 3.9: Narrowband Fading.

where $P_r = \sum_n E[\alpha_n^2] = 2\sigma^2$ is the average received signal power of the signal, i.e. the received power based on path loss and shadowing alone.

We obtain the power distribution by making the change of variables $z^2(t) = |r(t)|^2$ in (3.32) to obtain

$$p_{Z^2}(x) = \frac{1}{P_r} e^{-x/P_r} = \frac{1}{2\sigma^2} e^{-x/(2\sigma^2)}, \quad x \geq 0. \tag{3.33}$$

Thus, the received signal power is exponentially distributed with mean $2\sigma^2$. The complex lowpass equivalent signal for $r(t)$ is given by $r_{LP}(t) = r_I(t) + jr_Q(t)$ which has phase $\theta = \arctan(r_Q(t)/r_I(t))$. For $r_I(t)$ and $r_Q(t)$ uncorrelated Gaussian random variables we can show that θ is uniformly distributed and independent of $|r_{LP}|$. So $r(t)$ has a Rayleigh-distributed amplitude and uniform phase, and the two are mutually independent.

Example 3.2: Consider a channel with Rayleigh fading and average received power $P_r = 20$ dBm. Find the probability that the received power is below 10 dBm.
Solution. We have $P_r = 20$ dBm =100 mW. We want to find the probability that $Z^2 < 10$ dBm =10 mW. Thus

$$p(Z^2 < 10) = \int_0^{10} \frac{1}{100} e^{-x/100} dx = .095.$$

If the channel has a fixed LOS component then $r_I(t)$ and $r_Q(t)$ are not zero-mean. In this case the received signal equals the superposition of a complex Gaussian component and a LOS component. The signal envelope in this case can be shown to have a Rician distribution [9], given by

$$p_Z(z) = \frac{z}{\sigma^2} \text{exp} \left[\frac{-(z^2 + s^2)}{2\sigma^2} \right] I_0 \left(\frac{zs}{\sigma^2} \right), \quad z \geq 0, \tag{3.34}$$

where $2\sigma^2 = \sum_{n,n \neq 0} \text{E}[\alpha_n^2]$ is the average power in the non-LOS multipath components and $s^2 = \alpha_0^2$ is the power in the LOS component. The function I_0 is the modified Bessel function of 0th order. The average received power in the Rician fading is given by

$$P_r = \int_0^\infty z^2 p_Z(z) dx = s^2 + 2\sigma^2. \tag{3.35}$$

The Rician distribution is often described in terms of a fading parameter K, defined by

$$K = \frac{s^2}{2\sigma^2}. \tag{3.36}$$

Thus, K is the ratio of the power in the LOS component to the power in the other (non-LOS) multipath components. For $K = 0$ we have Rayleigh fading, and for $K = \infty$ we have no fading, i.e. a channel with no multipath and only a LOS component. The fading parameter K is therefore a measure of the severity of the fading: a small K implies severe fading, a large K implies more mild fading. Making the substitution $s^2 = KP/(K + 1)$ and $2\sigma^2 = P/(K + 1)$ we can write the Rician distribution in terms of K and P_r as

$$p_Z(z) = \frac{2z(K+1)}{P_r} \text{exp} \left[-K - \frac{(K+1)z^2}{P_r} \right] I_0 \left(2z\sqrt{\frac{K(K+1)}{P_r}} \right), \quad z \geq 0. \tag{3.37}$$

Both the Rayleigh and Rician distributions can be obtained by using mathematics to capture the underlying physical properties of the channel models [1, 9]. However, some experimental data does not fit well into either of these distributions. Thus, a more general fading distribution was developed whose parameters can be adjusted to fit a variety of empirical measurements. This distribution is called the Nakagami fading distribution, and is given by

$$p_Z(z) = \frac{2m^m z^{2m-1}}{\Gamma(m) P_r^m} \exp\left[\frac{-mz^2}{P_r}\right], \quad m \geq .5, \tag{3.38}$$

where P_r is the average received power and $\Gamma(\cdot)$ is the Gamma function. The Nakagami distribution is parameterized by P_r and the fading parameter m. For $m = 1$ the distribution in (3.38) reduces to Rayleigh fading. For $m = (K+1)^2/(2K+1)$ the distribution in (3.38) is approximately Rician fading with parameter K. For $m = \infty$ there is no fading: P_r is a constant. Thus, the Nakagami distribution can model Rayleigh and Rician distributions, as well as more general ones. Note that some empirical measurements support values of the m parameter less than one, in which case the Nakagami fading causes more severe performance degradation than Rayleigh fading. The power distribution for Nakagami fading, obtained by a change of variables, is given by

$$p_{Z^2}(x) = \left(\frac{m}{P_r}\right)^m \frac{x^{m-1}}{\Gamma(m)} \exp\left(\frac{-mx}{P_r}\right). \tag{3.39}$$

3.2.3 Level Crossing Rate and Average Fade Duration

The envelope level crossing rate L_Z is defined as the expected rate (in crossings per second) at which the signal envelope crosses the level Z in the downward direction. Obtaining L_Z requires the joint distribution of the signal envelope $z = |r|$ and its derivative with respect to time $\dot{z}$, $p(z, \dot{z})$. We now derive L_Z based on this joint distribution.

Consider the fading process shown in Figure 3.10. The expected amount of time the signal envelope spends in the interval $(Z, Z + dz)$ with envelope slope in the range $[\dot{z}, \dot{z} + d\dot{z}]$ over time duration dt is $A = p(Z, \dot{z}) dz d\dot{z} dt$. The time required to cross from Z to $Z + dz$ once for a given envelope slope $\dot{z}$ is $B = dz/\dot{z}$. The ratio $A/B = \dot{z} p(Z, \dot{z}) d\dot{z} dt$ is the expected number of crossings of the envelope z within the interval $(Z, Z + dz)$ for a given envelope slope $\dot{z}$ over time duration dt. The expected number of crossings of the envelope level Z for slopes between $\dot{z}$ and $\dot{z} + d\dot{z}$ in a time interval $[0, T]$ in the downward direction is thus

$$\int_0^T \dot{z} p(Z, \dot{z}) d\dot{z} dt = \dot{z} p(Z, \dot{z}) d\dot{z} T. \tag{3.40}$$

So the expected number of crossings of the envelope level Z with negative slope over the interval $[0, T]$ is

$$N_Z = T \int_{-\infty}^0 \dot{z} p(Z, \dot{z}) d\dot{z}. \tag{3.41}$$

Finally, the expected number of crossings of the envelope level Z per second, i.e. the level crossing rate, is

$$L_Z = \frac{N_Z}{T} = \int -\infty^0 \dot{z} p(Z, \dot{z}) d\dot{z}. \tag{3.42}$$

Note that this is a general result that applies for any random process.

The joint pdf of z and $\dot{z}$ for Rician fading was derived in [9] and can also be found in [11]. The level crossing rate for Rician fading is then obtained by using this pdf in (3.42), and is given by

$$L_Z = \sqrt{2\pi(K+1)} f_D \rho e^{-K-(K+1)\rho^2} I_0(2\rho\sqrt{K(K+1)}), \tag{3.43}$$

where $\rho = Z/\sqrt{P_r}$. It is easily shown that the rate at which the received signal power crosses a threshold value γ_0 obeys the same formula (3.43) with $\rho = \sqrt{\gamma_0/P_r}$. For Rayleigh fading ($K = 0$) the level crossing rate simplifies to

$$L_Z = \sqrt{2\pi} f_D \rho e^{-\rho^2}, \tag{3.44}$$

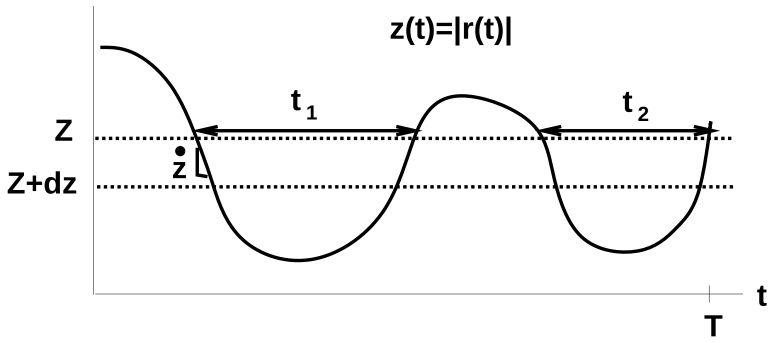


Figure 3.10: Level Crossing Rate and Fade Duration for Fading Process.

where $\rho = Z/\sqrt{P_r}$.

We define the average signal fade duration as the average time that the signal envelope stays below a given target level Z. This target level is often obtained from the signal amplitude or power level required for a given performance metric like bit error rate. Let t_i denote the duration of the ith fade below level Z over a time interval $[0, T]$, as illustrated in Figure 3.10. Thus t_i equals the length of time that the signal envelope stays below Z on its ith crossing. Since $z(t)$ is stationary and ergodic, for T sufficiently large we have

$$p(z(t) < Z) = \frac{1}{T}\sum_i t_i. \tag{3.45}$$

Thus, for T sufficiently large the average fade duration is

$$\bar{t}_Z = \frac{1}{TL_Z}\sum_{i=1}^{L_Z T} t_i \approx \frac{p(z(t) < Z)}{L_Z}. \tag{3.46}$$

Using the Rayleigh distribution for $p(z(t) < Z)$ yields

$$\bar{t}_Z = \frac{e^{\rho^2} - 1}{\rho f_D \sqrt{2\pi}} \tag{3.47}$$

with $\rho = Z/\sqrt{P_r}$. Note that (3.47) is the average fade duration for the signal envelope (amplitude) level with Z the target amplitude and $\sqrt{P_r}$ the average envelope level. By a change of variables it is easily shown that (3.47) also yields the average fade duration for the signal power level with $\rho = \sqrt{P_0/P_r}$, where P_0 is the target power level and P_r is the average power level. Note that average fade duration decreases with Doppler, since as a channel changes more quickly it remains below a given fade level for a shorter period of time. The average fade duration also generally increases with ρ for $\rho >> 1$. That is because as the target level increases relative to the average, the signal is more likely to be below the target. The average fade duration for Rician fading is more difficult to compute, it can be found in [11, Chapter 1.4].

The average fade duration indicates the number of bits or symbols affected by a deep fade. Specifically, consider an uncoded system with bit time T_b. Suppose the probability of bit error is high when $z < Z$. Then if $T_b \approx \bar{t}_Z$, the system will likely experience single error events, where bits that are received in error have the previous and subsequent bits received correctly (since $z > Z$ for these bits). On the other hand, if $T_b << \bar{t}_Z$ then many subsequent bits are received with $z < Z$, so large bursts of errors are likely. Finally, if $T_b >> \bar{t}_Z$ the fading is averaged out over a bit time in the demodulator, so the fading can be neglected. These issues will be explored in more detail in Chapter 8, when we consider coding and interleaving.

Example 3.3:

Consider a voice system with acceptable BER when the received signal power is at or above half its average value. If the BER is below its acceptable level for more than 120 ms, users will turn off their phone. Find the range of Doppler values in a Rayleigh fading channel such that the average time duration when users have unacceptable voice quality is less than $t = 60$ ms.

Solution: The target received signal value is half the average, so $P_0 = .5P_r$ and thus $\rho = \sqrt{.5}$. We require

$$\bar{t}_Z = \frac{e^{.5} - 1}{f_D\sqrt{\pi}} \leq t = .060$$

and thus $f_D \geq (e-1)/(.060\sqrt{2\pi}) = 6.1$ Hz.

3.2.4 Finite State Markov Channels

The complex mathematical characterization of flat fading described in the previous subsections can be difficult to incorporate into wireless performance analysis such as the packet error probability. Therefore, simpler models that capture the main features of flat fading channels are needed for these analytical calculations. One such model is a finite state Markov channel (FSMC). In this model fading is approximated as a discrete-time Markov process with time discretized to a given interval T (typically the symbol period). Specifically, the set of all possible fading gains is modeled as a set of finite channel states. The channel varies over these states at each interval T according to a set of Markov transition probabilities. FSMCs have been used to approximate both mathematical and experimental fading models, including satellite channels [13], indoor channels [14], Rayleigh fading channels [15, 19], Ricean fading channels [20], and Nakagami-m fading channels [17]. They have also been used for system design and system performance analysis in [18, 19]. First-order FSMC models have been shown to be deficient in computing performance analysis, so higher order models are generally used. The FSMC models for fading typically model amplitude variations only, although there has been some work on FSMC models for phase in fading [21] or phase-noisy channels [22].

A detailed FSMC model for Rayleigh fading was developed in [15]. In this model the time-varying SNR associated with the Rayleigh fading, γ, lies in the range $0 \leq \gamma \leq \infty$. The FSMC model discretizes this fading range into regions so that the jth region R_j is defined as $R_j = \gamma : A_j \leq \gamma < A_{j+1}$, where the region boundaries $\{A_j\}$ and the total number of fade regions are parameters of the model. This model assumes that γ stays within the same region over time interval T and can only transition to the same region or adjacent regions at time $T + 1$. Thus, given that the channel is in state R_j at time T, at the next time interval the channel can only transition to R_{j-1}, R_j, or R_{j+1}, a reasonable assumption when $f_D T$ is small. Under this assumption the transition probabilities between regions are derived in [15] as

$$p_{j,j+1} = \frac{N_{j+1}T_s}{\pi_j}, \quad p_{j,j-1} = \frac{N_j T_s}{\pi_j}, \quad p_{j,j} = 1 - p_{j,j+1} - p_{j,j-1}, \tag{3.48}$$

where N_j is the level-crossing rate at A_j and π_j is the steady-state distribution corresponding to the jth region: $\pi_j = p(\gamma \in R_j) = p(A_j \leq \gamma < A_{j+1})$.

3.3 Wideband Fading Models

When the signal is not narrowband we get another form of distortion due to the multipath delay spread. In this case a short transmitted pulse of duration T will result in a received signal that is of duration $T + T_m$, where T_m is the multipath delay spread. Thus, the duration of the received signal may be significantly increased. This is illustrated in Figure 3.11. In this figure, a pulse of width T is transmitted over a multipath channel. As discussed in Chapter 5, linear modulation consists of a train of pulses where each pulse carries information in its amplitude and/or phase corresponding to a data bit or symbol[5]. If the multipath delay spread $T_m << T$ then the multipath components are received roughly on top of one another, as shown on the upper right of the figure. The resulting constructive and destructive interference causes narrowband fading of the pulse, but there is little time-spreading of the pulse and therefore little interference with a subsequently transmitted pulse. On the other hand, if the multipath delay spread $T_m >> T$, then each of the different multipath components can be resolved, as shown in the lower right of the figure. However, these multipath components interfere with subsequently transmitted pulses. This effect is called intersymbol interference (ISI).

There are several techniques to mitigate the distortion due to multipath delay spread, including equalization, multicarrier modulation, and spread spectrum, which are discussed in Chapters 11-13. ISI migitation is not necessary if $T >> T_m$, but this can place significant constraints on data rate. Multicarrier modulation and spread spectrum actually change the characteristics of the transmitted signal to mostly avoid intersymbol interference, however they still experience multipath distortion due to frequency-selective fading, which is described in Section 3.3.2.

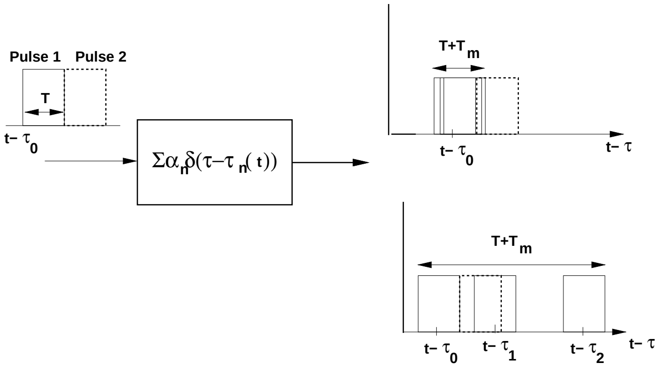


Figure 3.11: Multipath Resolution.

The difference between wideband and narrowband fading models is that as the transmit signal bandwidth B increases so that $T_m \approx B^{-1}$, the approximation $u(t - \tau_n(t)) \approx u(t)$ is no longer valid. Thus, the received signal is a sum of copies of the original signal, where each copy is delayed in time by τ_n and shifted in phase by $\phi_n(t)$. The signal copies will combine destructively when their phase terms differ significantly, and will distort the direct path signal when $u(t - \tau_n)$ differs from $u(t)$.

Although the approximation in (3.11) no longer applies when the signal bandwidth is large relative to the inverse of the multipath delay spread, if the number of multipath components is large and the phase of each component is uniformly distributed then the received signal will still be a zero-mean complex Gaussian process with a Rayleigh-distributed envelope. However, wideband fading differs from narrowband fading in terms of the resolution of the different multipath components. Specifically, for narrowband signals, the multipath components have a time resolution that is less than the inverse of the signal bandwidth, so the multipath components characterized

[5] Linear modulation typically uses nonsquare pulse shapes for bandwidth efficiency, as discussed in Chapter 5.4

in Equation (3.6) combine at the receiver to yield the original transmitted signal with amplitude and phase characterized by random processes. These random processes are characterized by their autocorrelation or PSD, and their instantaneous distributions, as discussed in Section 3.2. However, with wideband signals, the received signal experiences distortion due to the delay spread of the different multipath components, so the received signal can no longer be characterized by just the amplitude and phase random processes. The effect of multipath on wideband signals must therefore take into account both the multipath delay spread and the time-variations associated with the channel.

The starting point for characterizing wideband channels is the equivalent lowpass time-varying channel impulse response $c(\tau, t)$. Let us first assume that $c(\tau, t)$ is a continuous[6] deterministic function of τ and t. Recall that τ represents the impulse response associated with a given multipath delay, while t represents time variations. We can take the Fourier transform of $c(\tau, t)$ with respect to t as

$$S_c(\tau, \rho) = \int_{-\infty}^{\infty} c(\tau, t) e^{-j2\pi\rho t} dt. \tag{3.49}$$

We call $S_c(\tau, \rho)$ the **deterministic scattering function** of the lowpass equivalent channel impulse response $c(\tau, t)$. Since it is the Fourier transform of $c(\tau, t)$ with respect to the time variation parameter t, the deterministic scattering function $S_c(\tau, \rho)$ captures the Doppler characteristics of the channel via the frequency parameter ρ.

In general the time-varying channel impulse response $c(\tau, t)$ given by (3.6) is random instead of deterministic due to the random amplitudes, phases, and delays of the random number of multipath components. In this case we must characterize it statistically or via measurements. As long as the number of multipath components is large, we can invoke the Central Limit Theorem to assume that $c(\tau, t)$ is a complex Gaussian process, so its statistical characterization is fully known from the mean, autocorrelation, and cross-correlation of its in-phase and quadrature components. As in the narrowband case, we assume that the phase of each multipath component is uniformly distributed. Thus, the in-phase and quadrature components of $c(\tau, t)$ are independent Gaussian processes with the same autocorrelation, a mean of zero, and a cross-correlation of zero. The same statistics hold for the in-phase and quadrature components if the channel contains only a small number of multipath rays as long as each ray has a Rayleigh-distributed amplitude and uniform phase. Note that this model does not hold when the channel has a dominant LOS component.

The statistical characterization of $c(\tau, t)$ is thus determined by its **autocorrelation function**, defined as

$$A_c(\tau_1, \tau_2; t, \Delta t) = E[c^*(\tau_1; t) c(\tau_2; t + \Delta t)]. \tag{3.50}$$

Most channels in practice are wide-sense stationary (WSS), such that the joint statistics of a channel measured at two different times t and $t + \Delta t$ depends only on the time difference Δt. For wide-sense stationary channels, the autocorrelation of the corresponding bandpass channel $h(\tau, t) = \Re\{c(\tau, t)e^{j2\pi f_c t}\}$ can be obtained [16] from $A_c(\tau_1, \tau_2; t, \Delta t)$ as[7] $A_h(\tau_1, \tau_2; t, \Delta t) = .5\Re\{A_c(\tau_1, \tau_2; t, \Delta t)e^{j2\pi f_c \Delta t}\}$. We will assume that our channel model is WSS, in which case the autocorrelation becomes indepedent of t:

$$A_c(\tau_1, \tau_2; \Delta t) = E[c^*(\tau_1; t) c(\tau_2; t + \Delta t)]. \tag{3.51}$$

Moreover, in practice the channel response associated with a given multipath component of delay τ_1 is uncorrelated with the response associated with a multipath component at a different delay $\tau_2 \neq \tau_1$, since the two components are caused by different scatterers. We say that such a channel has uncorrelated scattering (US). We abbreviate

[6] The wideband channel characterizations in this section can also be done for discrete-time channels that are discrete with respect to τ by changing integrals to sums and Fourier transforms to discrete Fourier transforms.

[7] It is easily shown that the autocorrelation of the passband channel response $h(\tau, t)$ is given by $E[h(\tau_1, t)h(\tau_2, t + \Delta t)] = .5\Re\{A_c(\tau_1, \tau_2; t, \Delta t)e^{j2\pi f_c \Delta t}\} + .5\Re\{\hat{A}_c(\tau_1, \tau_2; t, \Delta t)e^{j2\pi f_c(2t+\Delta t)}\}$, where $\hat{A}_c(\tau_1, \tau_2; t, \Delta t) = E[c(\tau_1; t)c(\tau_2; t + \Delta t)]$. However, if $c(\tau, t)$ is WSS then $\hat{A}_c(\tau_1, \tau_2; t, \Delta t) = 0$, so $E[h(\tau_1, t)h(\tau_2, t + \Delta t)] = .5\Re\{A_c(\tau_1, \tau_2; t, \Delta t)e^{j2\pi f_c \Delta}\}$.

channels that are WSS with US as WSSUS channels. The WSSUS channel model was first introduced by Bello in his landmark paper [16], where he also developed two-dimensional transform relationships associated with this autocorrelation. These relationships will be discussed in Section 3.3.4. Incorporating the US property into (3.51) yields

$$E[c^*(\tau_1;t)c(\tau_2;t+\Delta t)] = A_c(\tau_1;\Delta t)\delta[\tau_1-\tau_2] \stackrel{\triangle}{=} A_c(\tau;\Delta t), \tag{3.52}$$

where $A_c(\tau;\Delta t)$ gives the average output power associated with the channel as a function of the multipath delay $\tau=\tau_1=\tau_2$ and the difference Δt in observation time. This function assumes that τ_1 and τ_2 satisfy $|\tau_1-\tau_2|>B^{-1}$, since otherwise the receiver can't resolve the two components. In this case the two components are modeled as a single combined multipath component with delay $\tau\approx\tau_1\approx\tau_2$.

The **scattering function** for random channels is defined as the Fourier transform of $A_c(\tau;\Delta t)$ with respect to the Δt parameter:

$$S_c(\tau,\rho) = \int_{-\infty}^{\infty} A_c(\tau,\Delta t)e^{-j2\pi\rho\Delta t}d\Delta t. \tag{3.53}$$

The scattering function characterizes the average output power associated with the channel as a function of the multipath delay τ and Doppler ρ. Note that we use the same notation for the deterministic scattering and random scattering functions since the function is uniquely defined depending on whether the channel impulse response is deterministic or random. A typical scattering function is shown in Figure 3.12.

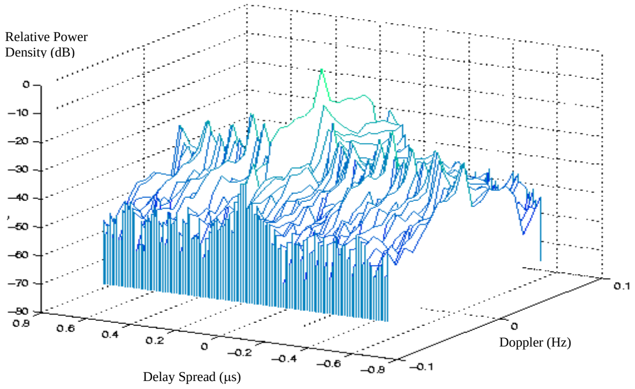


Figure 3.12: Scattering Function.

The most important characteristics of the wideband channel, including the power delay profile, coherence bandwidth, Doppler power spectrum, and coherence time, are derived from the channel autocorrelation $A_c(\tau,\Delta t)$ or scattering function $S(\tau,\rho)$. These characteristics are described in the subsequent sections.

3.3.1 Power Delay Profile

The **power delay profile** $A_c(\tau)$, also called the **multipath intensity profile**, is defined as the autocorrelation (3.52) with $\Delta t=0$: $A_c(\tau)\stackrel{\triangle}{=}A_c(\tau,0)$. The power delay profile represents the average power associated with a given multipath delay, and is easily measured empirically. The average and rms delay spread are typically defined in terms of the power delay profile $A_c(\tau)$ as

$$\mu_{T_m} = \frac{\int_0^\infty \tau A_c(\tau)d\tau}{\int_0^\infty A_c(\tau)d\tau}, \tag{3.54}$$

and

$$\sigma_{T_m} = \sqrt{\frac{\int_0^\infty (\tau - \mu_{T_m})^2 A_c(\tau) d\tau}{\int_0^\infty A_c(\tau) d\tau}}. \tag{3.55}$$

Note that if we define the pdf p_{T_m} of the random delay spread T_m in terms of $A_c(\tau)$ as

$$p_{T_m}(\tau) = \frac{A_c(\tau)}{\int_0^\infty A_c(\tau) d\tau} \tag{3.56}$$

then μ_{T_m} and σ_{T_m} are the mean and rms values of T_m, respectively, relative to this pdf. Defining the pdf of T_m by (3.56) or, equivalently, defining the mean and rms delay spread by (3.54) and (3.55), respectively, weights the delay associated with a given multipath component by its relative power, so that weak multipath components contribute less to delay spread than strong ones. In particular, multipath components below the noise floor will not significantly impact these delay spread characterizations.

The time delay T where $A_c(\tau) \approx 0$ for $\tau \geq T$ can be used to roughly characterize the delay spread of the channel, and this value is often taken to be a small integer multiple of the rms delay spread, i.e. $A_c(\tau) \approx 0$ for $\tau > 3\sigma_{T_m}$. With this approximation a linearly modulated signal with symbol period T_s experiences significant ISI if $T_s << \sigma_{T_m}$. Conversely, when $T_s >> \sigma_{T_m}$ the system experiences negligible ISI. For calculations one can assume that $T_s << \sigma_{T_m}$ implies $T_s < \sigma_{T_m}/10$ and $T_s >> \sigma_{T_m}$ implies $T_s > 10\sigma_{T_m}$. When T_s is within an order of magnitude of σ_{T_m} then there will be some ISI which may or may not significantly degrade performance, depending on the specifics of the system and channel. We will study the performance degradation due to ISI in linearly modulated systems as well as ISI mitigation methods in later chapters.

While $\mu_{T_m} \approx \sigma_{T_m}$ in many channels with a large number of scatterers, the exact relationship between μ_{T_m} and σ_{T_m} depends on the shape of $A_c(\tau)$. A channel with no LOS component and a small number of multipath components with approximately the same large delay will have $\mu_{T_m} >> \sigma_{T_m}$. In this case the large value of μ_{T_m} is a misleading metric of delay spread, since in fact all copies of the transmitted signal arrive at rougly the same time and the demodulator would synchronize to this common delay. It is typically assumed that the synchronizer locks to the multipath component at approximately the mean delay, in which case rms delay spread characterizes the time-spreading of the channel.

Example 3.4:

The power delay spectrum is often modeled as having a one-sided exponential distribution:

$$A_c(\tau) = \frac{1}{\overline{T}_m} e^{-\tau/\overline{T}_m}, \;\; \tau \geq 0.$$

Show that the average delay spread (3.54) is $\mu_{T_m} = \overline{T}_m$ and find the rms delay spread (3.55).

Solution: It is easily shown that $A_c(\tau)$ integrates to one. The average delay spread is thus given by

$$\mu_{T_m} = \frac{1}{\overline{T}_m} \int_0^\infty \tau e^{-\tau/\overline{T}_m} d\tau = \overline{T}_m.$$

$$\sigma_{T_m} = \sqrt{\frac{1}{\overline{T}_m} \int_0^\infty \tau^2 e^{-\tau/\overline{T}_m} d\tau - \mu_{T_m}^2} = 2\overline{T}_m - \overline{T}_m = \overline{T}_m.$$

Thus, the average and rms delay spread are the same for exponentially distributed power delay profiles.

Example 3.5:

Consider a wideband channel with multipath intensity profile

$$A_c(\tau) = \begin{cases} e^{-\tau/.00001} & 0 \leq \tau \leq 20 \;\; \mu\text{sec.} \\ 0 & \text{else} \end{cases} .$$

Find the mean and rms delay spreads of the channel and find the maximum symbol rate such that a linearly-modulated signal transmitted through this channel does not experience ISI.

Solution: The average delay spread is

$$\mu_{T_m} = \frac{\int_0^{20*10^{-6}} \tau e^{-\tau/.00001} d\tau}{\int_0^{20*10^{-6}} e^{-\tau/.00001} d\tau} = 6.87 \;\; \mu\text{sec.}$$

The rms delay spread is

$$\sigma_{T_m} = \sqrt{\frac{\int_0^{20*10^{-6}} (\tau - \mu_{T_m})^2 e^{-\tau} d\tau}{\int_0^{20*10^{-6}} e^{-\tau} d\tau}} = 5.25 \;\; \mu\text{sec.}$$

We see in this example that the mean delay spread is roughly equal to its rms value. To avoid ISI we require linear modulation to have a symbol period T_s that is large relative to σ_{T_m}. Taking this to mean that $T_s > 10\sigma_{T_m}$ yields a symbol period of $T_s = 52.5$ μsec or a symbol rate of $R_s = 1/T_s = 19.04$ Kilosymbols per second. This is a highly constrained symbol rate for many wireless systems. Specifically, for binary modulations where the symbol rate equals the data rate (bits per second, or bps), high-quality voice requires on the order of 32 Kbps and high-speed data requires on the order of 10-100 Mbps.

3.3.2 Coherence Bandwidth

We can also characterize the time-varying multipath channel in the frequency domain by taking the Fourier transform of $c(\tau, t)$ with respect to τ. Specifically, define the random process

$$C(f;t) = \int_{-\infty}^{\infty} c(\tau;t) e^{-j2\pi f\tau} d\tau. \tag{3.57}$$

Since $c(\tau;t)$ is a complex zero-mean Gaussian random variable in t, the Fourier transform above just represents the sum[8] of complex zero-mean Gaussian random processes, and therefore $C(f;t)$ is also a zero-mean Gaussian random process completely characterized by its autocorrelation. Since $c(\tau;t)$ is WSS, its integral $C(f;t)$ is as well. Thus, the autocorrelation of (3.57) is given by

$$A_C(f_1, f_2; \Delta t) = \mathrm{E}[C^*(f_1;t)C(f_2;t+\Delta t)]. \tag{3.58}$$

[8] We can express the integral as a limit of a discrete sum.

We can simplify $A_C(f_1, f_2; \Delta t)$ as

$$\begin{aligned} A_C(f_1, f_2; \Delta t) &= E\left[\int_{-\infty}^{\infty} c^*(\tau_1; t) e^{j2\pi f_1 \tau_1} d\tau_1 \int_{-\infty}^{\infty} c(\tau_2; t + \Delta t) e^{-j2\pi f_2 \tau_2} d\tau_2\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[c^*(\tau_1; t) c(\tau_2; t + \Delta t)] e^{j2\pi f_1 \tau_1} e^{-j2\pi f_2 \tau_2} d\tau_1 d\tau_2 \\ &= \int_{-\infty}^{\infty} A_c(\tau, \Delta t) e^{-j2\pi(f_2 - f_1)\tau} d\tau. \\ &= A_C(\Delta f; \Delta t) \end{aligned} \quad (3.59)$$

where $\Delta f = f_2 - f_1$ and the third equality follows from the WSS and US properties of $c(\tau; t)$. Thus, the autocorrelation of $C(f; t)$ in frequency depends only on the frequency difference Δf. The function $A_C(\Delta f; \Delta t)$ can be measured in practice by transmitting a pair of sinusoids through the channel that are separated in frequency by Δf and calculating their cross correlation at the receiver for the time separation Δt.

If we define $A_C(\Delta f) \overset{\triangle}{=} A_C(\Delta f; 0)$ then from (3.59),

$$A_C(\Delta f) = \int_{-\infty}^{\infty} A_c(\tau) e^{-j2\pi\Delta f \tau} d\tau. \quad (3.60)$$

So $A_C(\Delta f)$ is the Fourier transform of the power delay profile. Since $A_C(\Delta f) = \mathrm{E}[C^*(f; t)C(f + \Delta f; t]$ is an autocorrelation, the channel response is approximately independent at frequency separations Δf where $A_C(\Delta f) \approx 0$. The frequency B_c where $A_C(\Delta f) \approx 0$ for all $\Delta f > B_c$ is called the **coherence bandwidth** of the channel. By the Fourier transform relationship between $A_c(\tau)$ and $A_C(\Delta f)$, if $A_c(\tau) \approx 0$ for $\tau > T$ then $A_C(\Delta f) \approx 0$ for $\Delta f > 1/T$. Thus, the minimum frequency separation B_c for which the channel response is roughly independent is $B_c \approx 1/T$, where T is typically taken to be the rms delay spread σ_{T_m} of $A_c(\tau)$. A more general approximation is $B_c \approx k/\sigma_{T_m}$ where k depends on the shape of $A_c(\tau)$ and the precise specification of coherence bandwidth. For example, Lee has shown that $B_c \approx .02/\sigma_{T_m}$ approximates the range of frequencies over which channel correlation exceeds 0.9, while $B_c \approx .2/\sigma_{T_m}$ approximates the range of frequencies over which this correlation exceeds 0.5. [12].

In general, if we are transmitting a narrowband signal with bandwidth $B << B_c$, then fading across the entire signal bandwidth is highly correlated, i.e. the fading is roughly equal across the entire signal bandwidth. This is usually referred to as **flat fading**. On the other hand, if the signal bandwidth $B >> B_c$, then the channel amplitude values at frequencies separated by more than the coherence bandwidth are roughly independent. Thus, the channel amplitude varies widely across the signal bandwidth. In this case the channel is called **frequency-selective**. When $B \approx B_c$ then channel behavior is somewhere between flat and frequency-selective fading. Note that in linear modulation the signal bandwidth B is inversely proportional to the symbol time T_s, so flat fading corresponds to $T_s \approx 1/B >> 1/B_c \approx \sigma_{T_m}$, i.e. the case where the channel experiences negligible ISI. Frequency-selective fading corresponds to $T_s \approx 1/B << 1/B_c = \sigma_{T_m}$, i.e. the case where the linearly modulated signal experiences significant ISI. Wideband signaling formats that reduce ISI, such as multicarrier modulation and spread spectrum, still experience frequency-selective fading across their entire signal bandwidth which causes performance degradation, as will be discussed in Chapters 12 and 13, respectively.

We illustrate the power delay profile $A_c(\tau)$ and its Fourier transform $A_C(\Delta f)$ in Figure 3.13. This figure also shows two signals superimposed on $A_C(\Delta f)$, a narrowband signal with bandwidth much less than B_c and a wideband signal with bandwidth much greater than B_c. We see that the autocorrelation $A_C(\Delta f)$ is flat across the bandwidth of the narrowband signal, so this signal will experience flat fading or, equivalently, negligible ISI. The autocorrelation $A_C(\Delta f)$ goes to zero within the bandwidth of the wideband signal, which means that fading will be independent across different parts of the signal bandwidth, so fading is frequency selective and a linearly-modulated signal transmitted through this channel will experience significant ISI.

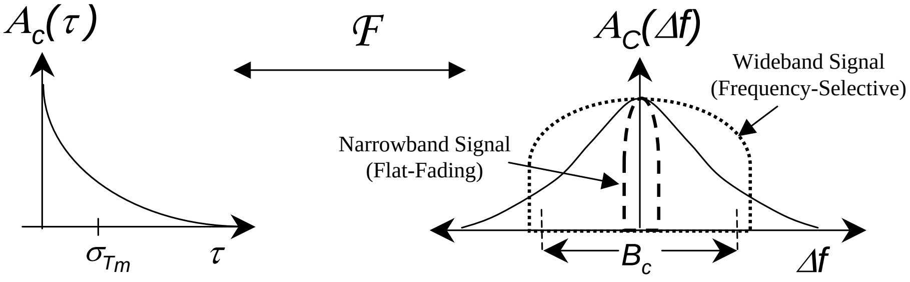


Figure 3.13: Power Delay Profile, RMS Delay Spread, and Coherence Bandwidth.

Example 3.6: In indoor channels $\sigma_{T_m} \approx 50$ ns whereas in outdoor microcells $\sigma_{T_m} \approx 30\mu$sec. Find the maximum symbol rate $R_s = 1/T_s$ for these environments such that a linearly-modulated signal transmitted through these environments experiences negligible ISI.

Solution. We assume that negligible ISI requires $T_s >> \sigma_{T_m}$, i.e. $T_s \geq 10\sigma_{T_m}$. This translates to a symbol rate $R_s = 1/T_s \leq .1/\sigma_{T_m}$. For $\sigma_{T_m} \approx 50$ ns this yields $R_s \leq 2$ Mbps and for $\sigma_{T_m} \approx 30\mu$sec this yields $R_s \leq 3.33$ Kbps. Note that indoor systems currently support up to 50 Mbps and outdoor systems up to 200 Kbps. To maintain these data rates for a linearly-modulated signal without severe performance degradation due to ISI, some form of ISI mitigation is needed. Moreover, ISI is less severe in indoor systems than in outdoor systems due to their lower delay spread values, which is why indoor systems tend to have higher data rates than outdoor systems.

3.3.3 Doppler Power Spectrum and Channel Coherence Time

The time variations of the channel which arise from transmitter or receiver motion cause a Doppler shift in the received signal. This Doppler effect can be characterized by taking the Fourier transform of $A_C(\Delta f; \Delta t)$ relative to Δt:

$$S_C(\Delta f; \rho) = \int_{-\infty}^{\infty} A_C(\Delta f; \Delta t) e^{-j2\pi\rho\Delta t} d\Delta t. \tag{3.61}$$

In order to characterize Doppler at a single frequency, we set Δf to zero and define $S_C(\rho) \stackrel{\triangle}{=} S_C(0; \rho)$. It is easily seen that

$$S_C(\rho) = \int_{-\infty}^{\infty} A_C(\Delta t) e^{-j2\pi\rho\Delta t} d\Delta t \tag{3.62}$$

where $A_C(\Delta t) \stackrel{\triangle}{=} A_C(\Delta f = 0; \Delta t)$. Note that $A_C(\Delta t)$ is an autocorrelation function defining how the channel impulse response decorrelates over time. In particular $A_C(\Delta t = T) = 0$ indicates that observations of the channel impulse response at times separated by T are uncorrelated and therefore independent, since the channel is a Gaussian random process. We define the **channel coherence time** T_c to be the range of values over which $A_C(\Delta t)$ is approximately nonzero. Thus, the time-varying channel decorrelates after approximately T_c seconds. The function $S_C(\rho)$ is called the **Doppler power spectrum** of the channel: as the Fourier transform of an autocorrelation

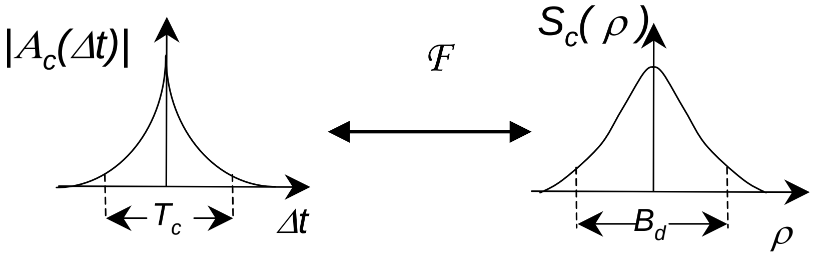


Figure 3.14: Doppler Power Spectrum, Doppler Spread, and Coherence Time.

it gives the PSD of the received signal as a function of Doppler ρ. The maximum ρ value for which $|S_C(\rho)|$ is greater than zero is called the **Doppler spread** of the channel, and is denoted by B_D. By the Fourier transform relationship between $A_C(\Delta t)$ and $S_C(\rho)$, $B_D \approx 1/T_c$. If the transmitter and reflectors are all stationary and the receiver is moving with velocity v, then $B_D \leq v/\lambda = f_D$. Recall that in the narrowband fading model samples became independent at time $\Delta t = .4/f_D$, so in general $B_D \approx k/T_c$ where k depends on the shape of $S_c(\rho)$. We illustrate the Doppler power spectrum $S_C(\rho)$ and its inverse Fourier transform $A_C(\Delta_t)$ in Figure 3.14.

Example 3.7:
For a channel with Doppler spread $B_d = 80$ Hz, what time separation is required in samples of the received signal such that the samples are approximately independent.

Solution: The coherence time of the channel is $T_c \approx 1/B_d = 1/80$, so samples spaced 12.5 ms apart are approximately uncorrelated and thus, given the Gaussian properties of the underlying random process, these samples are approximately independent.

3.3.4 Transforms for Autocorrelation and Scattering Functions

From (3.61) we see that the scattering function $S_c(\tau;\rho)$ defined in (3.53) is the inverse Fourier transform of $S_C(\Delta f;\rho)$ in the Δf variable. Furthermore $S_c(\tau;\rho)$ and $A_C(\Delta f;\Delta t)$ are related by the double Fourier transform

$$S_c(\tau;\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A_C(\Delta f;\Delta t)e^{-j2\pi\rho\Delta t}e^{j2\pi\tau\Delta f}d\Delta t d\Delta f. \tag{3.63}$$

The relationships among the four functions $A_C(\Delta f;\Delta t)$, $A_c(\tau;\Delta t)$, $S_C(\Delta f;\rho)$, and $S_c(\tau;\rho)$ are shown in Figure 3.15

Empirical measurements of the scattering function for a given channel are often used to approximate empirically the channel's delay spread, coherence bandwidth, Doppler spread, and coherence time. The delay spread for a channel with empirical scattering function $S_c(\tau;\rho)$ is obtained by computing the empirical power delay profile $A_c(\tau)$ from $A_c(\tau,\Delta t) = \mathcal{F}_\rho^{-1}[S_c(\tau;\rho)]$ with $\Delta t = 0$ and then computing the mean and rms delay spread from this power delay profile. The coherence bandwidth can then be approximated as $B_c \approx 1/\sigma_{T_m}$. Similarly, the Doppler

$$\begin{tikzcd}
 & A_c(\tau,\Delta t) \arrow[dl, shift right, "\mathcal{F}_\tau"'] \arrow[dr, shift left, "\mathcal{F}_{\Delta t}"] & \\
A_C(\Delta f,\Delta t) \arrow[ur, shift right, "\mathcal{F}^{-1}_{\Delta f}"'] \arrow[dr, shift left, "\mathcal{F}_{\Delta t}"] & & S_c(\tau,\rho) \arrow[ul, shift left, "\mathcal{F}^{-1}_{\rho}"] \arrow[dl, shift right, "\mathcal{F}_\tau"'] \\
 & S_C(\Delta f,\rho) \arrow[ul, shift left, "\mathcal{F}^{-1}_{\rho}"] \arrow[ur, shift right, "\mathcal{F}^{-1}_{\Delta f}"'] &
\end{tikzcd}$$

Figure 3.15: Fourier Transform Relationships

spread B_D is approximated as the range of ρ values over which $S(0;\rho)$ is roughly nonzero, with the coherence time $T_c \approx 1/B_D$.

3.4 Discrete-Time Model

Often the time-varying impulse response channel model is too complex for simple analysis. In this case a discrete-time approximation for the wideband multipath model can be used. This discrete-time model, developed by Turin in [3], is especially useful in the study of spread spectrum systems and RAKE receivers, which is covered in Chapter 13. This discrete-time model is based on a physical propagation environment consisting of a composition of isolated point scatterers, as shown in Figure 3.16. In this model, the multipath components are assumed to form subpath clusters: incoming paths on a given subpath with approximate delay τ_n are combined, and incoming paths on different subpath clusters with delays r_n and r_m where $|r_n - r_m| > 1/B$ can be resolved, where B denotes the signal bandwidth.

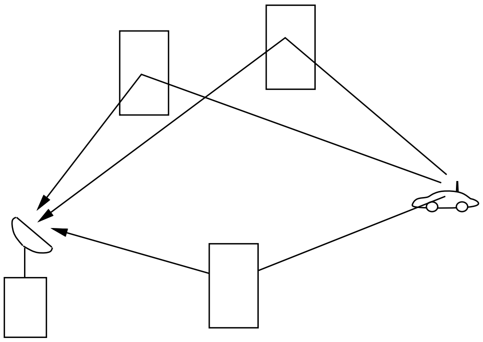


Figure 3.16: Point Scatterer Channel Model

The channel model of (3.6) is modified to include a fixed number $N+1$ of these subpath clusters as

$$c(\tau;t) = \sum_{n=0}^{N} \alpha_n(t) e^{-j\phi_n(t)} \delta(\tau - \tau_n(t)). \tag{3.64}$$

The statistics of the received signal for a given t are thus given by the statistics of $\{\tau_n\}_0^N$, $\{\alpha_n\}_0^N$, and $\{\phi_n\}_0^N$. The model can be further simplified using a discrete time approximation as follows: For a fixed t, the time axis is divided into M equal intervals of duration T such that $MT \geq \sigma_{T_m}$, where σ_{T_m} is the rms delay spread of the channel, which is derived empirically. The subpaths are restricted to lie in one of the M time interval bins, as shown in Figure 3.17. The multipath spread of this discrete model is MT, and the resolution between paths is T. This resolution is based on the transmitted signal bandwidth: $T \approx 1/B$. The statistics for the nth bin are that r_n, $1 \leq n \leq M$, is a binary indicator of the existence of a multipath component in the nth bin: so r_n is one if there is a multipath component in the nth bin and zero otherwise. If $r_n = 1$ then (a_n, θ_n), the amplitude and phase corresponding to this multipath component, follow an empirically determined distribution. This distribution is obtained by sample averages of (a_n, θ_n) for each n at different locations in the propagation environment. The empirical distribution of (a_n, θ_n) and (a_m, θ_m), $n \neq m$, is generally different, it may correspond to the same family of fading but with different parameters (e.g. Ricean fading with different K factors), or it may correspond to different fading distributions altogether (e.g. Rayleigh fading for the nth bin, Nakagami fading for the mth bin).

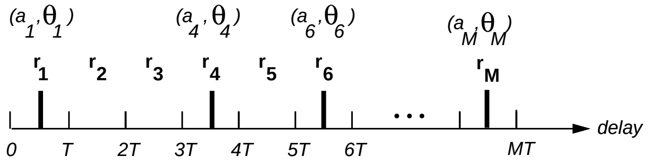


Figure 3.17: Discrete Time Approximation

This completes the statistical model for the discrete time approximation for a single snapshot. A sequence of profiles will model the signal over time as the channel impulse response changes, e.g. the impulse response seen by a receiver moving at some nonzero velocity through a city. Thus, the model must include both the first order statistics of $(\tau_n, \alpha_n, \phi_n)$ for each profile (equivalently, each t), but also the temporal and spatial correlations (assumed Markov) between them. More details on the model and the empirically derived distributions for N and for $(\tau_n, \alpha_n, \phi n)$ can be found in [3].

3.5 Space-Time Channel Models

Multiple antennas at the transmitter and/or receiver are becoming very common in wireless systems, due to their diversity and capacity benefits. Systems with multiple antennas require channel models that characterize both spatial (angle of arrival) and temporal characteristics of the channel. A typical model assumes the channel is composed of several scattering centers which generate the multipath [23, 24]. The location of the scattering centers relative to the receiver dictate the angle of arrival (AOA) of the corresponding multipath components. Models can be either two dimensional or three dimensional.

Consider a two-dimensional multipath environment where the receiver or transmitter has an antenna array with M elements. The time-varying impulse response model (3.6) can be extended to incorporate AOA for the array as follows.

$$c(\tau, t) = \sum_{n=0}^{N(t)} \alpha_n(t) e^{-j\phi_n(t)} \overline{a}(\theta_n(t)) \delta(\tau - \tau_n(t)), \tag{3.65}$$

where $\phi_n(t)$ corresponds to the phase shift at the origin of the array and $\overline{a}(\theta_n(t))$ is the array response vector given by

$$\overline{a}(\theta_n(t)) = [e^{-j\psi_{n,1}}, \ldots, e^{-j\psi_{n,M}}]^T, \tag{3.66}$$

where $\psi_{n,i} = [x_i \cos\theta_n(t) + y_i \sin\theta_n(t)]2\pi/\lambda$ for (x_i, y_i) the antenna location relative to the origin and $\theta_n(t)$ the AOA of the multipath relative to the origin of the antenna array. Assume the AOA is stationary and identically distributed for all multipath components and denote this random AOA by θ. Let $A(\theta)$ denote the average received signal power as a function of θ. Then we define the mean and rms angular spread in terms of this power profile as

$$\mu_\theta = \frac{\int_{-\pi}^{\pi} \theta A(\theta) d\theta}{\int_{-\pi}^{\pi} A(\theta) d\theta}, \tag{3.67}$$

and

$$\sigma_\theta = \sqrt{\frac{\int_{-\pi}^{\pi} (\theta - \mu_\theta)^2 A(\theta) d\theta}{\int_{-\pi}^{\pi} A(\theta) d\theta}}, \tag{3.68}$$

We say that two signals received at AOAs separated by $1/\sigma_\theta$ are roughly uncorrelated. More details on the power distribution relative to the AOA for different propagation environments along with the corresponding correlations across antenna elements can be found in [24]

Extending the two dimensional models to three dimensions requires characterizing the elevation AOAs for multipath as well as the azimuth angles. Different models for such 3-D channels have been proposed in [25, 26, 27]. In [23] the Jakes model is extended to produce spatio-temporal characteristics using the ideas of [25, 26, 27]. Several other papers on spatio-temporal modeling can be found in [29].

Bibliography

[1] R.S. Kennedy. *Fading Dispersive Communication Channels*. New York: Wiley, 1969.

[2] D.C. Cox. "910 MHz urban mobile radio propagation: Multipath characteristics in New York City," *IEEE Trans. Commun.*, Vol. COM-21, No. 11, pp. 1188–1194, Nov. 1973.

[3] G.L. Turin. "Introduction to spread spectrum antimultipath techniques and their application to urban digital radio," *IEEE Proceedings*, Vol. 68, No. 3, pp. 328–353, March 1980.

[4] R.H. Clarke, "A statistical theory of mobile radio reception," *Bell Syst. Tech. J.*, pp. 957-1000, July-Aug. 1968.

[5] W.C. Jakes, Jr., *Microwave Mobile Communications*. New York: Wiley, 1974.

[6] T.S. Rappaport, *Wireless Communications - Principles and Practice,* 2nd Edition, Prentice Hall, 2001.

[7] M. Pätzold, *Mobile fading channels: Modeling, analysis, and simulation*, Wiley, 2002.

[8] M.K. Simon and M.-Sl. Alouini, *Digital Communication over Fading Channels,* New York: Wiley, 2000.

[9] S.O. Rice, "Mathematical analysis of random noise," *Bell System Tech. J.*, Vol. 23, No. 7, pp. 282–333, July 1944, and Vol. 24, No. 1, pp. 46–156, Jan. 1945.

[10] J.G. Proakis, *Digital Communications*, 3rd Ed., New York: McGraw-Hill, 1995.

[11] G.L. Stuber, *Principles of Mobile Communications*, Kluwer Academic Publishers, 2nd Ed., 2001.

[12] W.C.Y. Lee, *Mobile Cellular Telecommunications Systems*, New York: Mcgraw Hill, 1989.

[13] F. Babich, G. Lombardi, and E. Valentinuzzi, "Variable order Markov modeling for LEO mobile satellite channels," *Electronic Letters*, pp. 621–623, April 1999.

[14] A.M. Chen and R.R. Rao, "On tractable wireless channel models," *Proc. International Symp. on Pers., Indoor, and Mobile Radio Comm.*, pp. 825–830, Sept. 1998.

[15] H.S. Wang and N. Moayeri, "Finite-state Markov channel - A useful model for radio communication channels," *IEEE Trans. Vehic. Technol.*, pp. 163–171, Feb. 1995.

[16] P.A. Bello, "Characterization of randomly time-variant linear channels," *IEEE Trans. Comm. Syst.*, pp. 360–393, Dec. 1963.

[17] Y. L. Guan and L. F. Turner, "Generalised FSMC model for radio channels with correlated fading," *IEE Proc. Commun.*, pp. 133–137, April 1999.

[18] M. Chu and W. Stark,“Effect of mobile velocity on communications in fading channels,” *IEEE Trans. Vehic. Technol.*, Vol 49, No. 1, pp. 202–210, Jan. 2000.

[19] C.C. Tan and N.C. Beaulieu, “On first-order Markov modeling for the Rayleigh fading channel,” *IEEE Trans. Commun.*, Vol. 48, No. 12, pp. 2032–2040, Dec. 2000.

[20] C. Pimentel and I.F. Blake, “”Modeling burst channels using partitioned Fritchman’s Markov models, *IEEE Trans. Vehic. Technol.*, pp. 885–899, Aug. 1998.

[21] C. Komninakis and R. D. Wesel, ”Pilot-aided joint data and channel estimation in flat correlated fading,” *Proc. of IEEE Globecom Conf. (Comm. Theory Symp.)*, pp. 2534–2539, Nov. 1999.

[22] M. Peleg, S. Shamai (Shitz), and S. Galan, “Iterative decoding for coded noncoherent MPSK communications over phase-noisy AWGN channels,” *IEE Proceedings - Communications*, Vol. 147, pp. 87–95, April 2000.

[23] Y. Mohasseb and M.P. Fitz, “A 3-D spatio-temporal simulation model for wireless channels,” *IEEE J. Select. Areas Commun.* pp. 1193–1203, Aug. 2002.

[24] R. Ertel, P. Cardieri, K.W. Sowerby, T. Rappaport, and J. H. Reed, “Overview of spatial channel models for antenna array communication systems,” *IEEE Pers. Commun. Magazine*, pp. 10–22, Feb. 1998.

[25] T. Aulin, “A modified model for fading signal at the mobile radio channel,” *IEEE Trans. Vehic. Technol.*, pp. 182–202, Aug. 1979.

[26] J.D. Parsons and M.D.Turkmani, “Characterization of mobile radio signals: model description.” *Proc. Inst. Elect. Eng.* pt. 1, pp. 459–556, Dec. 1991.

[27] J.D. Parsons and M.D.Turkmani, “Characterization of mobile radio signals: base station crosscorrelation.” *Proc. Inst. Elect. Eng.* pt. 2, pp. 459–556, Dec. 1991.

[28] D. Parsons, *The Mobile Radio Propagation Channel*. New York: Wiley, 1994.

[29] L.G. Greenstein, J.B. Andersen, H.L. Bertoni, S. Kozono, and D.G. Michelson, (Eds.), *IEEE Journal Select. Areas Commun.* Special Issue on Channel and Propagation Modeling for Wireless Systems Design, Aug. 2002.

Chapter 3 Problems

1. Consider a two-path channel consisting of a direct ray plus a ground-reflected ray where the transmitter is a fixed base station at height h and the receiver is mounted on a truck also at height h. The truck starts next to the base station and moves away at velocity v. Assume signal attenuation on each path follows a free-space path loss model. Find the time-varying channel impulse at the receiver for transmitter-receiver separation $d = vt$ sufficiently large such that the length of the reflected path can be approximated by $r + r' \approx d + 2h^2/d$.

2. Find a formula for the multipath delay spread T_m for a two-path channel model. Find a simplified formula when the transmitter-receiver separation is relatively large. Compute T_m for $h_t = 10$m, $h_r = 4$m, and $d = 100$m.

3. Consider a time-invariant indoor wireless channel with LOS component at delay 23 nsec, a multipath component at delay 48 nsec, and another multipath component at delay 67 nsec. Find the delay spread assuming the demodulator synchronizes to the LOS component. Repeat assuming that the demodulator synchronizes to the first multipath component.

4. Show that the minimum value of $f_c \tau_n$ for a system at $f_c = 1$ GHz with a fixed transmitter and a receiver separated by more than 10 m from the transmitter is much greater than 1.

5. Prove that for X and Y independent zero-mean Gaussian random variables with variance σ^2, the distribution of $Z = \sqrt{X^2 + Y^2}$ is Rayleigh-distributed and the distribution of Z^2 is exponentially-distributed.

6. Assume a Rayleigh fading channel with the average signal power $2\sigma^2 = -80$ dBm. What is the power outage probability of this channel relative to the threshold $P_o = -95$ dBm? How about $P_o = -90$ dBm?

7. Assume an application that requires a power outage probability of .01 for the threshold $P_o = -80$ dBm, For Rayleigh fading, what value of the average signal power is required?

8. Assume a Rician fading channel with $2\sigma^2 = -80$ dBm and a target power of $P_o = -80$ dBm. Find the outage probability assuming that the LOS component has average power $s^2 = -80$ dBm.

9. This problem illustrates that the tails of the Ricean distribution can be quite different than its Nakagami approximation. Plot the CDF of the Ricean distribution for $K = 1, 5, 10$ and the corresponding Nakagami distribution with $m = (K+1)^2/(2K+1)$. In general, does the Ricean distribution or its Nakagami approximation have a larger outage probability $p(\gamma < x)$ for x large?

10. In order to improve the performance of cellular systems, multiple base stations can receive the signal transmitted from a given mobile unit and combine these multiple signals either by selecting the strongest one or summing the signals together, perhaps with some optimized weights. This typically increases SNR and reduces the effects of shadowing. Combining of signals received from multiple base stations is called *macrodiversity*, and in this problem we explore the benefits of this technique. Diversity will be covered in more detail in Chapter 7.

    Consider a mobile at the midpoint between two base stations in a cellular network. The received signals (in dBW) from the base stations are given by

    $$P_{r,1} = W + Z_1,$$

    $$P_{r,2} = W + Z_2,$$

    where $Z_{1,2}$ are $\mathcal{N}(0, \sigma^2)$ random variables. We define outage with macrodiversity to be the event that both $P_{r,1}$ and $P_{r,2}$ fall below a threshould T.

(a) Interpret the terms W, Z_1, Z_2 in $P_{r,1}$ and $P_{r,2}$.

(b) If Z_1 and Z_2 are independent, show that the outage probability is given by

$$P_{out} = [Q(\Delta/\sigma)]^2,$$

where $\Delta = W - T$ is the fade margin at the mobile's location.

(c) Now suppose Z_1 and Z_2 are correlated in the following way:

$$Z_1 = a\,Y_1 + b\,Y,$$

$$Z_2 = a\,Y_2 + b\,Y,$$

where Y, Y_1, Y_2 are independent $\mathcal{N}(0, \sigma^2)$ random variables, and a, b are such that $a^2 + b^2 = 1$. Show that

$$P_{out} = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \left[Q\left(\frac{\Delta + by\sigma}{|a|\sigma} \right) \right]^2 e^{-y^2/2} dy.$$

(d) Compare the outage probabilities of (b) and (c) for the special case of $a = b = 1/\sqrt{2}$, $\sigma = 8$ and $\Delta = 5$ (this will require a numerical integration).

11. The goal of this problem is to develop a Rayleigh fading simulator for a mobile communications channel using the method based on filtering Gaussian processes based on the in-phase and quadrature PSDs described in 3.2.1. In this problem you must do the following:

    (a) Develop simulation code to generate a signal with Rayleigh fading amplitude over time. Your sample rate should be at least 1000 samples/sec, the average received envelope should be 1, and your simulation should be parameterized by the Doppler frequency f_D. Matlab is the easiest way to generate this simulation, but any code is fine.

    (b) Write a description of your simulation that clearly explains how your code generates the fading envelope using a block diagram and any necessary equations.

    (c) Turn in your well-commented code.

    (d) Provide plots of received amplitude (dB) vs. time for $f_D = 1, 10, 100$ Hz. over 2 seconds.

12. For a Rayleigh fading channel with average power $P_r = 30$dB, compute the average fade duration for target fade values $P_0 = 0$ dB, $P_0 = 15$ dB, and $P_0 = 30$dB.

13. Derive a formula for the average length of time a Rayleigh fading process with average power P_r stays **above** a given target fade value P_0. Evaluate this average length of time for $P_r = 20$ dB, $P_0 = 25$ dB, and $f_D = 50$ Hz.

14. Assume a Rayleigh fading channel with average power $P_r = 10$ dB and Doppler $f_D = 80$ Hz. We would like to approximate the channel using a finite state Markov model with eight states. The regions R_j corresponds to $R_1 = \gamma : -\infty \leq \gamma \leq -10$dB, $R_2 = \gamma : -10$dB $\leq \gamma \leq 0$dB, $R_3 = \gamma : 0$dB $\leq \gamma \leq 5$dB, $R_4 = \gamma : 5$dB $\leq \gamma \leq 10$dB, $R_5 = \gamma : 10$dB $\leq \gamma \leq 15$dB, $R_6 = \gamma : 15$dB $\leq \gamma \leq 20$dB, $R_7 = \gamma : 20$dB $\leq \gamma \leq 30$dB, $R_8 = \gamma : 30$dB $\leq \gamma \leq \infty$. Find the transition probabilties between each region for this model.

15. Consider the following channel scattering function obtained by sending a 900 MHz sinusoidal input into the channel:

$$S(\tau,\rho) = \begin{cases} \alpha_1\delta(\tau) & \rho = 70\text{Hz.} \\ \alpha_2\delta(\tau - .022\mu\text{sec}) & \rho = 49.5\text{Hz.} \\ 0 & \text{else} \end{cases}$$

where α_1 and α_2 are determined by path loss, shadowing, and multipath fading. Clearly this scattering function corresponds to a 2-ray model. Assume the transmitter and receiver used to send and receive the sinusoid are located 8 meters above the ground.

(a) Find the distance and velocity between the transmitter and receiver.

(b) For the distance computed in part (a), is the path loss as a function of distance proportional to d^{-2} or d^{-4}? *Hint: use the fact that the channel is based on a 2-ray model.*

(c) Does a 30 KHz voice signal transmitted over this channel experience flat or frequency-selective fading?

16. Consider a wideband channel characterized by the autocorrelation function

$$A_c(\tau,\Delta t) = \begin{cases} \text{sinc}(W\Delta t) & 0 \le \tau \le 10\mu\text{sec.} \\ 0 & \text{else} \end{cases},$$

where $W = 100$Hz and $sinc(x) = \sin(\pi x)/(\pi x)$.

(a) Does this channel correspond to an indoor channel or an outdoor channel, and why?

(b) Sketch the scattering function of this channel.

(c) Compute the channel's average delay spread, rms delay spread, and Doppler spread.

(d) Over approximately what range of data rates will a signal transmitted over this channel exhibit frequency-selective fading?

(e) Would you expect this channel to exhibit Rayleigh or Ricean fading statistics, and why?

(f) Assuming that the channel exhibits Rayleigh fading, what is the average length of time that the signal power is continuously below its average value.

(g) Assume a system with narrowband binary modulation sent over this channel. Your system has error correction coding that can correct two simultaneous bit errors. Assume also that you always make an error if the received signal power is below its average value, and never make an error if this power is at or above its average value. If the channel is Rayleigh fading then what is the maximum data rate that can be sent over this channel with error-free transmission, making the approximation that the fade duration never exceeds twice its average value.

17. Let a scattering function $S(\tau,\rho)$ be nonzero over $0 \le \tau \le .1$ ms and $-.1 \le \rho \le .1$ Hz. Assume that the power of the scattering function is approximately uniform over the range where it is nonzero.

(a) What are the multipath spread and the doppler spread of the channel?

(b) Suppose you input to this channel two identical sinusoids separated in time by Δt. What is the minimum value of Δf for which the channel response to the first sinusoid is approximately independent of the channel response to the second sinusoid.

(c) For two sinusoidal inputs to the channel $u_1(t) = \sin 2\pi f t$ and $u_2(t) = \sin 2\pi f(t + \Delta t)$, what is the minimum value of Δt for which the channel response to $u_1(t)$ is approximately independent of the channel response to $u_2(t)$.

(d) Will this channel exhibit flat fading or frequency-selective fading for a typical voice channel with a 3 KHz bandwidth? How about for a cellular channel with a 30 KHz bandwidth?

Chapter 4

Capacity of Wireless Channels

The growing demand for wireless communication makes it important to determine the capacity limits of these channels. These capacity limits dictate the maximum data rates that can be transmitted over wireless channels with asymptotically small error probability, assuming no constraints on delay or complexity of the encoder and decoder. Channel capacity was pioneered by Claude Shannon in the late 1940s, using a mathematical theory of communication based on the notion of mutual information between the input and output of a channel [1, 2, 3]. Shannon defined capacity as the mutual information maximized over all possible input distributions. The significance of this mathematical construct was Shannon's coding theorem and converse, which proved that a code did exist that could achieve a data rate close to capacity with negligible probability of error, and that any data rate higher than capacity could not be achieved without an error probability bounded away from zero. Shannon's ideas were quite revolutionary at the time, given the high data rates he predicted were possible on telephone channels and the notion that coding could reduce error probability without reducing data rate or causing bandwidth expansion. In time sophisticated modulation and coding technology validated Shannon's theory such that on telephone lines today, we achieve data rates very close to Shannon capacity with very low probability of error. These sophisticated modulation and coding strategies are treated in Chapters 5 and 8, respectively.

In this chapter we examine the capacity of a single-user wireless channel where the transmitter and/or receiver have a single antenna. Capacity of single-user systems where the transmitter and receiver have multiple antennas is treated in Chapter 10 and capacity of multiuser systems is treated in Chapter 14. We will discuss capacity for channels that are both time-invariant and time-varying. We first look at the well-known formula for capacity of a time-invariant AWGN channel. We next consider capacity of time-varying flat-fading channels. Unlike in the AWGN case, capacity of a flat-fading channel is not given by a single formula, since capacity depends on what is known about the time-varying channel at the transmitter and/or receiver. Moreover, for different channel information assumptions, there are different definitions of channel capacity, depending on whether capacity characterizes the maximum rate averaged over all fading states or the maximum constant rate that can be maintained in all fading states (with or without some probability of outage).

We will consider flat-fading channel capacity where only the fading distribution is known at the transmitter and receiver. Capacity under this assumption is typically very difficult to determine, and is only known in a few special cases. Next we consider capacity when the channel fade level is known at the receiver only (via receiver estimation) or that the channel fade level is known at both the transmitter and the receiver (via receiver estimation and transmitter feedback). We will see that the fading channel capacity with channel side information at both the transmitter and receiver is achieved when the transmitter adapts its power, data rate, and coding scheme to the channel variation. The optimal power allocation in this case is a "water-filling" in time, where power and data rate are increased when channel conditions are favorable and decreased when channel conditions are not favorable.

We will also treat capacity of frequency-selective fading channels. For time-invariant frequency-selective channels the capacity is known and is achieved with an optimal power allocation that water-fills over frequency instead of time. The capacity of a time-varying frequency-selective fading channel is unknown in general. However, this channel can be approximated as a set of independent parallel flat-fading channels, whose capacity is the sum of capacities on each channel with power optimally allocated among the channels. The capacity of this channel is known and is obtained with an optimal power allocation that water-fills over both time and frequency.

We will consider only discrete-time systems in this chapter. Most continuous-time systems can be converted to discrete-time systems via sampling, and then the same capacity results hold. However, care must be taken in choosing the appropriate sampling rate for this conversion, since time variations in the channel may increase the sampling rate required to preserve channel capacity [4].

4.1 Capacity in AWGN

Consider a discrete-time additive white Gaussian noise (AWGN) channel with channel input/output relationship $y[i] = x[i] + n[i]$, where $x[i]$ is the channel input at time i, $y[i]$ is the corresponding channel output, and $n[i]$ is a white Gaussian noise random process. Assume a channel bandwidth B and transmit power P. The channel SNR, the power in $x[i]$ divided by the power in $n[i]$, is constant and given by $\gamma = P/(N_0 B)$, where N_0 is the power spectral density of the noise. The capacity of this channel is given by Shannon's well-known formula [1]:

$$C = B \log_2(1+\gamma), \tag{4.1}$$

where the capacity units are bits/second (bps). Shannon's coding theorem proves that a code exists that achieves data rates arbitrarily close to capacity with arbitrarily small probability of bit error. The converse theorem shows that any code with rate $R > C$ has a probability of error bounded away from zero. The theorems are proved using the concept of mutual information between the input and output of a channel. For a memoryless time-invariant channel with random input x and random output y, the channel's **mutual information** is defined as

$$I(X;Y) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{4.2}$$

where the sum is taken over all possible input and output pairs $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ for $\mathcal{X}$ and $\mathcal{Y}$ the input and output alphabets. The log function is typically with respect to base 2, in which case the units of mutual information are bits per second. Mutual information can also be written in terms of the **entropy** in the channel output y and conditional output $y|x$ as $I(X;Y) = H(Y) - H(Y|X)$, where $H(Y) = -\sum_{y\in\mathcal{Y}} p(y)\log p(y)$ and $H(Y|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(y|x)$. Shannon proved that channel capacity equals the mutual information of the channel maximized over all possible input distributions:

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{4.3}$$

For the AWGN channel, the maximizing input distribution is Gaussian, which results in the channel capacity given by (4.1). For channels with memory, mutual information and channel capacity are defined relative to input and output sequences x^n and y^n. More details on channel capacity, mutual information, and the coding theorem and converse can be found in [2, 5, 6].

The proofs of the coding theorem and converse place no constraints on the complexity or delay of the communication system. Therefore, Shannon capacity is generally used as an upper bound on the data rates that can be achieved under real system constraints. At the time that Shannon developed his theory of information, data rates over standard telephone lines were on the order of 100 bps. Thus, it was believed that Shannon capacity, which predicted speeds of roughly 30 Kbps over the same telephone lines, was not a very useful bound for real systems. However, breakthroughs in hardware, modulation, and coding techniques have brought commercial modems of today very close to the speeds predicted by Shannon in the 1950s. In fact, modems can exceed this 30 Kbps Shannon limit on some telephone channels, but that is because transmission lines today are of better quality than in Shannon's day and thus have a higher received power than that used in Shannon's initial calculation. On AWGN radio channels, turbo codes have come within a fraction of a dB of the Shannon capacity limit [7].

Wireless channels typically exhibit flat or frequency-selective fading. In the next two sections we consider capacity of flat-fading and frequency-selective fading channels under different assumptions regarding what is known about the channel.

Example 4.1: Consider a wireless channel where power falloff with distance follows the formula $P_r(d) = P_t(d_0/d)^3$ for $d_0 = 10$ m. Assume the channel has bandwidth $B = 30$ KHz and AWGN with noise power spectral density of $N_0 = 10^{-9}$ W/Hz. For a transmit power of 1 W, find the capacity of this channel for a transmit-receive distance of 100 m and 1 Km.

Solution: The received SNR is $\gamma = P_r(d)/(N_0B) = .1^3/(10^{-9} \times 30 \times 10^3) = 33 = 15$ dB for $d = 100$ m and $\gamma = .01^3/(10^{-9} \times 30 \times 10^3) = .033 = -15$ dB for $d = 1000$ m. The corresponding capacities are $C = B\log_2(1+\gamma) = 30000\log_2(1+33) = 152.6$ Kbps for $d = 100$ m and $C = 30000\log_2(1+.033) = 1.4$ Kbps for $d = 1000$ m. Note the significant decrease in capacity at farther distances, due to the path loss exponent of 3, which greatly reduces received power as distance increases.

4.2 Capacity of Flat-Fading Channels

4.2.1 Channel and System Model

We assume a discrete-time channel with stationary and ergodic time-varying gain $\sqrt{g[i]}, 0 \leq g[i]$, and AWGN $n[i]$, as shown in Figure 4.1. The channel power gain $g[i]$ follows a given distribution $p(g)$, e.g. for Rayleigh fading $p(g)$ is exponential. We assume that $g[i]$ is independent of the channel input. The channel gain $g[i]$ can change at each time i, either as an i.i.d. process or with some correlation over time. In a **block fading channel** $g[i]$ is constant over some blocklength T after which time $g[i]$ changes to a new independent value based on the distribution $p(g)$. Let $\overline{P}$ denote the average transmit signal power, $N_0/2$ denote the noise power spectral density of $n[i]$, and B denote the received signal bandwidth. The instantaneous received signal-to-noise ratio (SNR) is then $\gamma[i] = \overline{P}g[i]/(N_0B)$, $0 \leq \gamma[i] < \infty$, and its expected value over all time is $\overline{\gamma} = \overline{P}\overline{g}/(N_0B)$. Since $\overline{P}/(N_0B)$ is a constant, the distribution of $g[i]$ determines the distribution of $\gamma[i]$ and vice versa.

The system model is also shown in Figure 4.1, where an input message $\mathbf{w}$ is sent from the transmitter to the receiver. The message is encoded into the codeword $\mathbf{x}$, which is transmitted over the time-varying channel as $x[i]$ at time i. The channel gain $g[i]$, also called the **channel side information** (CSI), changes during the transmission of the codeword.

The capacity of this channel depends on what is known about $g[i]$ at the transmitter and receiver. We will consider three different scenarios regarding this knowledge:

1. **Channel Distribution Information (CDI):** The distribution of $g[i]$ is known to the transmitter and receiver.

2. **Receiver CSI:** The value of $g[i]$ is known at the receiver at time i, and both the transmitter and receiver know the distribution of $g[i]$.

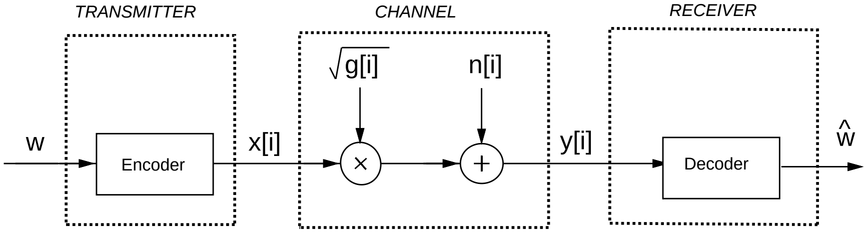


Figure 4.1: Flat-Fading Channel and System Model.

3. **Transmitter and Receiver CSI:** The value of $g[i]$ is known at the transmitter and receiver at time i, and both the transmitter and receiver know the distribution of $g[i]$.

Transmitter and receiver CSI allow the transmitter to adapt both its power and rate to the channel gain at time i, and leads to the highest capacity of the three scenarios. Note that since the instantaneous SNR $\gamma[i]$ is just $g[i]$ multipled by the constant $\overline{P}/(N_0B)$, known CSI or CDI about $g[i]$ yields the same information about $\gamma[i]$. Capacity for time-varying channels under assumptions other than these three are discussed in [8, 9].

4.2.2 Channel Distribution Information (CDI) Known

We first consider the case where the channel gain distribution $p(g)$ or, equivalently, the distribution of SNR $p(\gamma)$ is known to the transmitter and receiver. For i.i.d. fading the capacity is given by (4.3), but solving for the capacity-achieving input distribution, i.e. the distribution achieving the maximum in (4.3), can be quite complicated depending on the fading distribution. Moreover, fading correlation introduces channel memory, in which case the capacity-achieving input distribution is found by optimizing over input blocks, which makes finding the solution even more difficult. For these reasons, finding the capacity-achieving input distribution and corresponding capacity of fading channels under CDI remains an open problem for almost all channel distributions.

The capacity-achieving input distribution and corresponding fading channel capacity under CDI is known for two specific models of interest: i.i.d. Rayleigh fading channels and FSMCs. In i.i.d. Rayleigh fading the channel power gain is exponential and changes independently with each channel use. The optimal input distribution for this channel was shown in [10] to be discrete with a finite number of mass points, one of which is located at zero. This optimal distribution and its corresponding capacity must be found numerically. The lack of closed-form solutions for capacity or the optimal input distribution is somewhat surprising given the fact that the fading follows the most common fading distribution and has no correlation structure. For flat-fading channels that are not necessarily Rayleigh or i.i.d. upper and lower bounds on capacity have been determined in [11], and these bounds are tight at high SNRs.

FSMCs to approximate Rayleigh fading channels was discussed in Chapter 3.2.4. This model approximates the fading correlation as a Markov process. While the Markov nature of the fading dictates that the fading at a given time depends only on fading at the previous time sample, it turns out that the receiver must decode all past channel outputs jointly with the current output for optimal (i.e. capacity-achieving) decoding. This significantly complicates capacity analysis. The capacity of FSMCs has been derived for i.i.d. inputs in [13, 14] and for general inputs in [15]. Capacity of the FSMC depends on the limiting distribution of the channel conditioned on all past inputs and outputs, which can be computed recursively. As with the i.i.d. Rayleigh fading channel, the complexity of the capacity analysis along with the final result for this relatively simple fading model is very high, indicating the difficulty of obtaining the capacity and related design insights into channels when only CDI is available.

4.2.3 Channel Side Information at Receiver

We now consider the case where the CSI $g[i]$ is known at the receiver at time i. Equivalently, $\gamma[i]$ is known at the receiver at time i. We also assume that both the transmitter and receiver know the distribution of $g[i]$. In this case there are two channel capacity definitions that are relevant to system design: Shannon capacity, also called **ergodic capacity**, and **capacity with outage**. As for the AWGN channel, Shannon capacity defines the maximum data rate that can be sent over the channel with asymptotically small error probability. Note that for Shannon capacity the rate transmitted over the channel is constant: the transmitter cannot adapt its transmission strategy relative to the CSI. Thus, poor channel states typically reduce Shannon capacity since the transmission strategy must incorporate the effect of these poor states. An alternate capacity definition for fading channels with receiver CSI is capacity with outage. Capacity with outage is defined as the maximum rate that can be transmitted over a channel with some outage probability corresponding to the probability that the transmission cannot be decoded with negligible error probability. The basic premise of capacity with outage is that a high data rate can be sent over the channel and decoded correctly except when the channel is in deep fading. By allowing the system to lose some data in the event of deep fades, a higher data rate can be maintained than if all data must be received correctly regardless of the fading state, as is the case for Shannon capacity. The probability of outage characterizes the probability of data loss or, equivalently, of deep fading.

Shannon (Ergodic) Capacity

Shannon capacity of a fading channel with receiver CSI for an average power constraint $\overline{P}$ can be obtained from results in [16] as

$$C = \int_0^\infty B\log_2(1+\gamma)p(\gamma)d\gamma. \tag{4.4}$$

Note that this formula is a probabilistic average, i.e. Shannon capacity is equal to Shannon capacity for an AWGN channel with SNR γ, given by $B\log_2(1+\gamma)$, averaged over the distribution of γ. That is why Shannon capacity is also called ergodic capacity. However, care must be taken in interpreting (4.4) as an average. In particular, it is incorrect to interpret (4.4) to mean that this average capacity is achieved by maintaining a capacity $B\log_2(1+\gamma)$ when the instantaneous SNR is γ, since only the receiver knows the instantaneous SNR $\gamma[i]$, and therefore the data rate transmitted over the channel is constant, regardless of γ. Note, also, the capacity-achieving code must be sufficiently long so that a received codeword is affected by all possible fading states. This can result in significant delay.

By Jensen's inequality,

$$\mathbf{E}[B\log_2(1+\gamma)] = \int B\log_2(1+\gamma)p(\gamma)d\gamma \leq B\log_2(1+\mathbf{E}[\gamma]) = B\log_2(1+\overline{\gamma}), \tag{4.5}$$

where $\overline{\gamma}$ is the average SNR on the channel. Thus we see that the Shannon capacity of a fading channel with receiver CSI only is less than the Shannon capacity of an AWGN channel with the same average SNR. In other words, fading reduces Shannon capacity when only the receiver has CSI. Moreover, without transmitter CSI, the code design must incorporate the channel correlation statistics, and the complexity of the maximum likelihood decoder will be proportional to the channel decorrelation time. In addition, if the receiver CSI is not perfect, capacity can be significantly decreased [20].

Example 4.2: Consider a flat-fading channel with i.i.d. channel gain $g[i]$ which can take on three possible values: $g_1 = .05$ with probability $p_1 = .1$, $g_2 = .5$ with probability $p_2 = .5$, and $g_3 = 1$ with probability $p_3 = .4$. The transmit power is 10 mW, the noise spectral density is $N_0 = 10^{-9}$ W/Hz, and the channel bandwidth is 30 KHz. Assume the receiver has knowledge of the instantaneous value of $g[i]$ but the transmitter does not. Find the

Shannon capacity of this channel and compare with the capacity of an AWGN channel with the same average SNR.

Solution: The channel has 3 possible received SNRs, $\gamma_1 = P_t g_1/(N_0 B) = .01 * (.05^2)/(30000 * 10^{-9}) = .8333 = -.79$ dB, $\gamma_2 = P_t g_2/(N_0 B) = .01 \times (.5^2)/(30000 * 10^{-9}) = 83.333 = 19.2$ dB, and $\gamma_3 = P_t g_3/(N_0 B) = .01/(30000 * 10^{-9}) = 333.33 = 25$ dB. The probabilities associated with each of these SNR values is $p(\gamma_1) = .1$, $p(\gamma_2) = .5$, and $p(\gamma_3) = .4$. Thus, the Shannon capacity is given by

$$C = \sum_i B \log_2(1+\gamma_i) p(\gamma_i) = 30000(.1 \log_2(1.8333) + .5 \log_2(84.333) + .4 \log_2(334.33)) = 199.26 \text{ Kbps.}$$

The average SNR for this channel is $\overline{\gamma} = .1(.8333) + .5(83.33) + .4(333.33) = 175.08 = 22.43$ dB. The capacity of an AWGN channel with this SNR is $C = B \log_2(1 + 175.08) = 223.8$ Kbps. Note that this rate is about 25 Kbps larger than that of the flat-fading channel with receiver CSI and the same average SNR.

Capacity with Outage

Capacity with outage applies to slowly-varying channels, where the instantaneous SNR γ is constant over a large number of transmissions (a transmission burst) and then changes to a new value based on the fading distribution. With this model, if the channel has received SNR γ during a burst then data can be sent over the channel at rate $B \log_2(1+\gamma)$ with negligible probability of error[1]. Since the transmitter does not know the SNR value γ, it must fix a transmission rate independent of the instantaneous received SNR.

Capacity with outage allows bits sent over a given transmission burst to be decoded at the end of the burst with some probability that these bits will be decoded incorrectly. Specifically, the transmitter fixes a minimum received SNR γ_{min} and encodes for a data rate $C = B \log_2(1+\gamma_{min})$. The data is correctly received if the instantaneous received SNR is greater than or equal to γ_{min} [17, 18]. If the received SNR is below γ_{min} then the bits received over that transmission burst cannot be decoded correctly with probability approaching one, and the receiver declares an outage. The probability of outage is thus $p_{out} = p(\gamma < \gamma_{min})$. The average rate correctly received over many transmission bursts is $C_o = (1 - p_{out}) B \log_2(1+\gamma_{min})$ since data is only correctly received on $1 - p_{out}$ transmissions. The value of γ_{min} is a design parameter based on the acceptable outage probability. Capacity with outage is typically characterized by a plot of capacity versus outage, as shown in Figure 4.2. In this figure we plot the normalized capacity $C/B = \log_2(1+\gamma_{min})$ as a function of outage probability $p_{out} = p(\gamma < \gamma_{min})$ for a Rayleigh fading channel (γ exponential) with $\overline{\gamma} = 20$ dB. We see that capacity approaches zero for small outage probability, due to the requirement to correctly decode bits transmitted under severe fading, and increases dramatically as outage probability increases. Note, however, that these high capacity values for large outage probabilities have higher probability of incorrect data reception. The average rate correctly received can be maximized by finding the γ_{min} or, equivalently, the p_{out}, that maximizes C_o.

Example 4.3: Assume the same channel as in the previous example, with a bandwidth of 30 KHz and three possible received SNRs: $\gamma_1 = .8333$ with $p(\gamma_1) = .1$, $\gamma_2 = 83.33$ with $p(\gamma_2) = .5$, and $\gamma_3 = 333.33$ with $p(\gamma_3) = .4$. Find the capacity versus outage for this channel, and find the average rate correctly received for outage probabilities $p_{out} < .1$, $p_{out} = .1$ and $p_{out} = .6$.

[1] The assumption of constant fading over a large number of transmissions is needed since codes that achieve capacity require very large blocklengths.

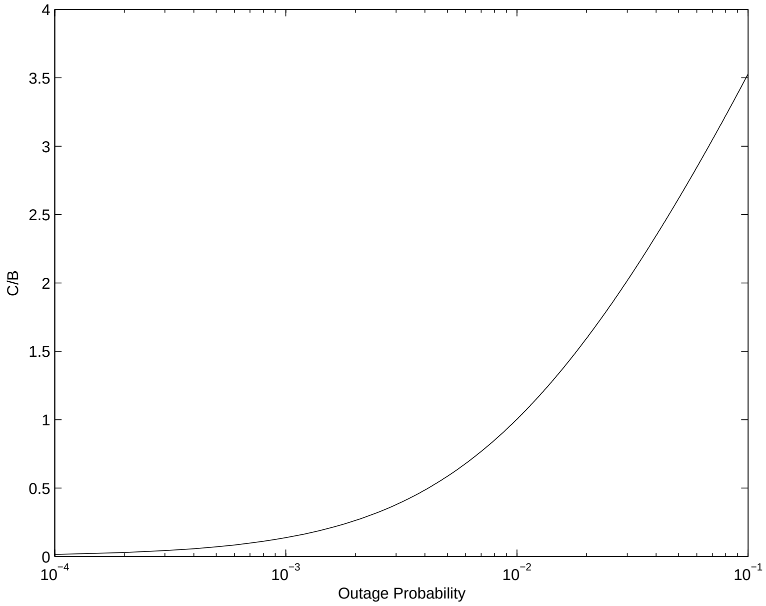


Figure 4.2: Normalized Capacity (C/B) versus Outage Probability.

Solution: For time-varying channels with discrete SNR values the capacity versus outage is a staircase function. Specifically, for $p_{out} < .1$ we must decode correctly in all channel states. The minimum received SNR for p_{out} in this range of values is that of the weakest channel: $\gamma_{min} = \gamma_1$, and the corresponding capacity is $C = B\log_2(1+\gamma_{min}) = 30000\log_2(1.833) = 26.23$ Kbps. For $.1 \leq p_{out} < .6$ we can decode incorrectly when the channel is in the weakest state only. Then $\gamma_{min} = \gamma_2$ and the corresponding capacity is $C = B\log_2(1+\gamma_{min}) = 30000\log_2(84.33) = 191.94$ Kbps. For $.6 \leq p_{out} < 1$ we can decode incorrectly if the channel has received SNR γ_1 or γ_2. Then $\gamma_{min} = \gamma_3$ and the corresponding capacity is $C = B\log_2(1+\gamma_{min}) = 30000\log_2(334.33) = 251.55$ Kbps. Thus, capacity versus outage has $C = 26.23$ Kbps for $p_{out} < .1$, $C = 191.94$ Kbps for $.1 \leq p_{out} < .6$, and $C = 251.55$ Kbps for $.6 \leq p_{out} < 1$.

For $p_{out} < .1$ data transmitted at rates close to capacity $C = 26.23$ Kbps are always correctly received since the channel can always support this data rate. For $p_{out} = .1$ we transmit at rates close to $C = 191.94$ Kbps, but we can only correctly decode these data when the channel SNR is γ_2 or γ_3, so the rate correctly received is $(1-.1)191940 = 172.75$ Kbps. For $p_{out} = .6$ we transmit at rates close to $C = 251.55$ Kbps but we can only correctly decode these data when the channel SNR is γ_3, so the rate correctly received is $(1-.6)251550 = 125.78$ Kbps. It is likely that a good engineering design for this channel would send data at a rate close to 191.94 Kbps, since it would only be received incorrectly at most 10% of this time and the data rate would be almost an order of magnitude higher than sending at a rate commensurate with the worst-case channel capacity. However, 10% retransmission probability is too high for some applications, in which case the system would be designed for the 26.23 Kbps data rate with no retransmissions. Design issues regarding acceptable retransmission probability will be discussed in Chapter 14.

4.2.4 Channel Side Information at Transmitter and Receiver

When both the transmitter and receiver have CSI, the transmitter can adapt its transmission strategy relative to this CSI, as shown in Figure 4.3. In this case there is no notion of capacity versus outage where the transmitter sends bits that cannot be decoded, since the transmitter knows the channel and thus will not send bits unless they can be decoded correctly. In this section we will derive Shannon capacity assuming optimal power and rate adaptation relative to the CSI, as well as introduce alternate capacity definitions and their power and rate adaptation strategies.

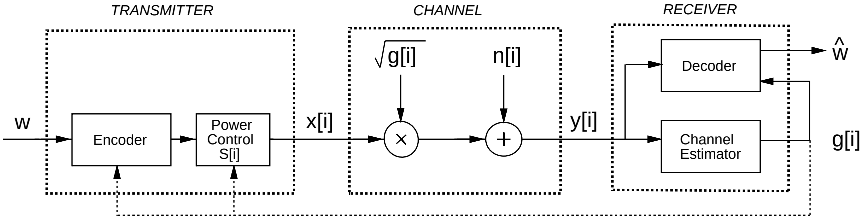


Figure 4.3: System Model with Transmitter and Receiver CSI.

Shannon Capacity

We now consider the Shannon capacity when the channel power gain $g[i]$ is known to both the transmitter and receiver at time i. The Shannon capacity of a time-varying channel with side information about the channel state at both the transmitter and receiver was originally considered by Wolfowitz for the following model. Let $s[i]$ be a stationary and ergodic stochastic process representing the channel state, which takes values on a finite set $\mathcal{S}$ of discrete memoryless channels. Let C_s denote the capacity of a particular channel $s \in \mathcal{S}$, and $p(s)$ denote the probability, or fraction of time, that the channel is in state s. The capacity of this time-varying channel is then given by Theorem 4.6.1 of [19]:

$$C = \sum_{s \in \mathcal{S}} C_s p(s). \tag{4.6}$$

We now apply this formula to the system model in Figure 4.1. We know the capacity of an AWGN channel with average received SNR γ is $C_\gamma = B \log_2(1+\gamma)$. Let $p(\gamma) = p(\gamma[i] = \gamma)$ denote the probability distribution of the received SNR. From (4.6) the capacity of the fading channel with transmitter and receiver side information is thus[2]

$$C = \int_0^\infty C_\gamma p(\gamma) d\gamma = \int_0^\infty B \log_2(1+\gamma) p(\gamma) d\gamma. \tag{4.7}$$

We see that without power adaptation, (4.4) and (4.7) are the same, so transmitter side information does not increase capacity unless power is also adapted.

Let us now allow the transmit power $P(\gamma)$ to vary with γ, subject to an average power constraint $\overline{P}$:

$$\int_0^\infty P(\gamma) p(\gamma) d\gamma \leq \overline{P}. \tag{4.8}$$

With this additional constraint, we cannot apply (4.7) directly to obtain the capacity. However, we expect that the capacity with this average power constraint will be the average capacity given by (4.7) with the power optimally

[2] Wolfowitz's result was for γ ranging over a finite set, but it can be extended to infinite sets [21].

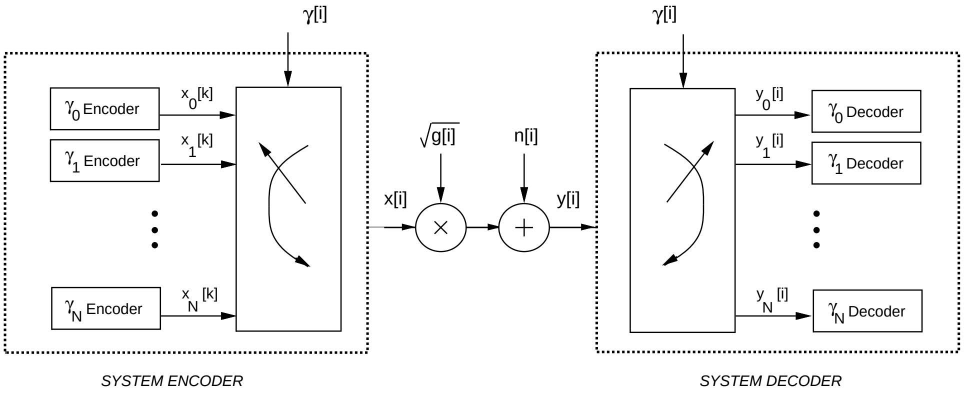


Figure 4.4: Multiplexed Coding and Decoding.

distributed over time. This motivates defining the fading channel capacity with average power constraint (4.8) as

$$C = \max_{P(\gamma):\int P(\gamma)p(\gamma)d\gamma=\overline{P}} \int_0^\infty B\log_2\left(1+\frac{P(\gamma)\gamma}{\overline{P}}\right)p(\gamma)d\gamma. \tag{4.9}$$

It is proved in [21] that the capacity given in (4.9) can be achieved, and any rate larger than this capacity has probability of error bounded away from zero. The main idea behind the proof is a "time diversity" system with multiplexed input and demultiplexed output, as shown in Figure 4.4. Specifically, we first quantize the range of fading values to a finite set $\{\gamma_j : 1 \leq j \leq N\}$. For each γ_j, we design an encoder/decoder pair for an AWGN channel with SNR γ_j. The input x_j for encoder γ_j has average power $P(\gamma_j)$ and data rate $R_j = C_j$, where C_j is the capacity of a time-invariant AWGN channel with received SNR $P(\gamma_j)\gamma_j/\overline{P}$. These encoder/decoder pairs correspond to a set of input and output ports associated with each γ_j. When $\gamma[i] \approx \gamma_j$, the corresponding pair of ports are connected through the channel. The codewords associated with each γ_j are thus multiplexed together for transmission, and demultiplexed at the channel output. This effectively reduces the time-varying channel to a set of time-invariant channels in parallel, where the jth channel only operates when $\gamma[i] \approx \gamma_j$. The average rate on the channel is just the sum of rates associated with each of the γ_j channels weighted by $p(\gamma_j)$, the percentage of time that the channel SNR equals γ_j. This yields the average capacity formula (4.9).

To find the optimal power allocation $P(\gamma)$, we form the Lagrangian

$$J(P(\gamma)) = \int_0^\infty B\log_2\left(1+\frac{\gamma P(\gamma)}{\overline{P}}\right)p(\gamma)d\gamma - \lambda\int_0^\infty P(\gamma)p(\gamma)d\gamma. \tag{4.10}$$

Next we differentiate the Lagrangian and set the derivative equal to zero:

$$\frac{\partial J(P(\gamma))}{\partial P(\gamma)} = \left[\left(\frac{B/\ln(2)}{1+\gamma P(\gamma)/\overline{P}}\right)\frac{\gamma}{\overline{P}} - \lambda\right]p(\gamma) = 0. \tag{4.11}$$

Solving for $P(\gamma)$ with the constraint that $P(\gamma) > 0$ yields the optimal power adaptation that maximizes (4.9) as

$$\frac{P(\gamma)}{\overline{P}} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma} & \gamma \geq \gamma_0 \\ 0 & \gamma < \gamma_0 \end{cases} \tag{4.12}$$

for some "cutoff" value γ_0. If $\gamma[i]$ is below this cutoff then no data is transmitted over the ith time interval, so the channel is only used at time i if $\gamma_0 \leq \gamma[i] < \infty$. Substituting (4.12) into (4.9) then yields the capacity formula:

$$C = \int_{\gamma_0}^\infty B\log_2\left(\frac{\gamma}{\gamma_0}\right)p(\gamma)d\gamma. \tag{4.13}$$

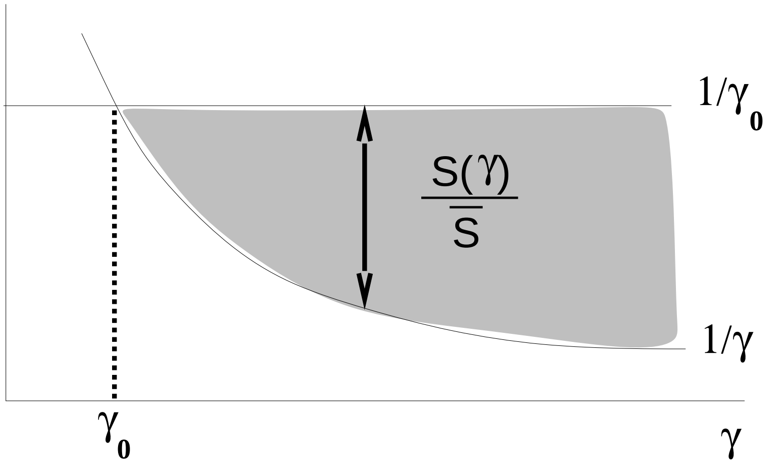


Figure 4.5: Optimal Power Allocation: Water-Filling.

The multiplexing nature of the capacity-achieving coding strategy indicates that (4.13) is achieved with a time-varying data rate, where the rate corresponding to instantaneous SNR γ is $B\log_2(\gamma/\gamma_0)$. Since γ_0 is constant, this means that as the instantaneous SNR increases, the data rate sent over the channel for that instantaneous SNR also increases. Note that this multiplexing strategy is not the only way to achieve capacity (4.13): it can also be achieved by adapting the transmit power and sending at a fixed rate [22]. We will see in Section 4.2.6 that for Rayleigh fading this capacity can exceed that of an AWGN channel with the same average power, in contrast to the case of receiver CSI only, where fading always decreases capacity.

Note that the optimal power allocation policy (4.12) only depends on the fading distribution $p(\gamma)$ through the cutoff value γ_0. This cutoff value is found from the power constraint. Specifically, by rearranging the power constraint (4.8) and replacing the inequality with equality (since using the maximum available power will always be optimal) yields the power constraint

$$\int_0^\infty \frac{P(\gamma)}{\overline{P}} p(\gamma)d\gamma = 1. \tag{4.14}$$

Now substituting the optimal power adaptation (4.12) into this expression yields that the cutoff value γ_0 must satisfy

$$\int_{\gamma_0}^\infty \left(\frac{1}{\gamma_0} - \frac{1}{\gamma}\right) p(\gamma)d\gamma = 1. \tag{4.15}$$

Note that this expression only depends on the distribution $p(\gamma)$. The value for γ_0 cannot be solved for in closed form for typical continuous pdfs $p(\gamma)$ and thus must be found numerically [23].

Since γ is time-varying, the maximizing power adaptation policy of (4.12) is a “water-filling” formula in time, as illustrated in Figure 4.5. This curve shows how much power is allocated to the channel for instantaneous SNR $\gamma(t) = \gamma$. The water-filling terminology refers to the fact that the line $1/\gamma$ sketches out the bottom of a bowl, and power is poured into the bowl to a constant water level of $1/\gamma_0$. The amount of power allocated for a given γ equals $1/\gamma_0 - 1/\gamma$, the amount of water between the bottom of the bowl ($1/\gamma$) and the constant water line ($1/\gamma_0$). The intuition behind water-filling is to take advantage of good channel conditions: when channel conditions are good (γ large) more power and a higher data rate is sent over the channel. As channel quality degrades (γ small) less power and rate are sent over the channel. If the instantaneous channel SNR falls below the cutoff value, the channel is not used. Adaptive modulation and coding techniques that follow this same principle were developed in [24, 25] and are discussed in Chapter 9.

Note that the multiplexing argument sketching how capacity (4.9) is achieved applies to any power adaptation policy, i.e. for any power adaptation policy $P(\gamma)$ with average power $\overline{P}$ the capacity

$$C = \int_0^\infty B \log_2 \left(1 + \frac{P(\gamma)\gamma}{\overline{P}}\right) p(\gamma) d\gamma. \tag{4.16}$$

can be achieved with arbitrarily small error probability. Of course this capacity cannot exceed (4.9), where power adaptation is optimized to maximize capacity. However, there are scenarios where a suboptimal power adaptation policy might have desirable properties that outweigh capacity maximization. In the next two sections we discuss two such suboptimal policies, which result in constant data rate systems, in contrast to the variable-rate transmission policy that achieves the capacity in (4.9).

Example 4.4: Assume the same channel as in the previous example, with a bandwidth of 30 KHz and three possible received SNRs: $\gamma_1 = .8333$ with $p(\gamma_1) = .1$, $\gamma_2 = 83.33$ with $p(\gamma_2) = .5$, and $\gamma_3 = 333.33$ with $p(\gamma_3) = .4$. Find the ergodic capacity of this channel assuming both transmitter and receiver have instantaneous CSI.

Solution: We know the optimal power allocation is water-filling, and we need to find the cutoff value γ_0 that satisfies the discrete version of (4.15) given by

$$\sum_{\gamma_i \geq \gamma_0} \left(\frac{1}{\gamma_0} - \frac{1}{\gamma_i}\right) p(\gamma_i) = 1. \tag{4.17}$$

We first assume that all channel states are used to obtain γ_0, i.e. assume $\gamma_0 \leq \min_i \gamma_i$, and see if the resulting cutoff value is below that of the weakest channel. If not then we have an inconsistency, and must redo the calculation assuming at least one of the channel states is not used. Applying (4.17) to our channel model yields

$$\sum_{i=1}^{3} \frac{p(\gamma_i)}{\gamma_0} - \sum_{i=1}^{3} \frac{p(\gamma_i)}{\gamma_i} = 1 \Rightarrow \frac{1}{\gamma_0} = 1 + \sum_{i=1}^{3} \frac{p(\gamma_i)}{\gamma_i} = 1 + \left(\frac{.1}{.8333} + \frac{.5}{83.33} + \frac{.4}{333.33}\right) = 1.13$$

Solving for γ_0 yields $\gamma_0 = 1/1.13 = .89 > .8333 = \gamma_1$. Since this value of γ_0 is greater than the SNR in the weakest channel, it is inconsistent as the channel should only be used for SNRs above the cutoff value. Therefore, we now redo the calculation assuming that the weakest state is not used. Then (4.17) becomes

$$\sum_{i=2}^{3} \frac{p(\gamma_i)}{\gamma_0} - \sum_{i=2}^{3} \frac{p(\gamma_i)}{\gamma_i} = 1 \Rightarrow \frac{.9}{\gamma_0} = 1 + \sum_{i=2}^{3} \frac{p(\gamma_i)}{\gamma_i} = 1 + \left(\frac{.5}{83.33} + \frac{.4}{333.33}\right) = 1.0072$$

Solving for γ_0 yields $\gamma_0 = .89$. So by assuming the weakest channel with SNR γ_1 is not used, we obtain a consistent value for γ_0 with $\gamma_1 < \gamma_0 \leq \gamma_2$. The capacity of the channel then becomes

$$C = \sum_{i=2}^{3} B \log_2(\gamma_i/\gamma_0) p(\gamma_i) = 30000(.5 \log_2(83.33/.89) + .4 \log_2(333.33/.89)) = 200.82 \text{ Kbps}.$$

Comparing with the results of the previous example we see that this rate is only slightly higher than for the case of receiver CSI only, and is still significantly below that of an AWGN channel with the same average SNR. That is because the average SNR for this channel is relatively high: for low SNR channels capacity in flat-fading can exceed that of the AWGN channel with the same SNR by taking advantage of the rare times when the channel is in a very good state.

Zero-Outage Capacity and Channel Inversion

We now consider a suboptimal transmitter adaptation scheme where the transmitter uses the CSI to maintain a constant received power, i.e., it inverts the channel fading. The channel then appears to the encoder and decoder as a time-invariant AWGN channel. This power adaptation, called **channel inversion**, is given by $P(\gamma)/\overline{P} = \sigma/\gamma$, where σ equals the constant received SNR that can be maintained with the transmit power constraint (4.8). The constant σ thus satisfies $\int \frac{\sigma}{\gamma} p(\gamma) d\gamma = 1$, so $\sigma = 1/\mathbf{E}[1/\gamma]$.

Fading channel capacity with channel inversion is just the capacity of an AWGN channel with SNR σ:

$$C = B \log_2 [1 + \sigma] = B \log_2 \left[1 + \frac{1}{\mathbf{E}[1/\gamma]}\right]. \tag{4.18}$$

The capacity-achieving transmission strategy for this capacity uses a fixed-rate encoder and decoder designed for an AWGN channel with SNR σ. This has the advantage of maintaining a fixed data rate over the channel regardless of channel conditions. For this reason the channel capacity given in (4.18) is called **zero-outage capacity**, since the data rate is fixed under all channel conditions and there is no channel outage. Note that there exist practical coding techniques that achieve near-capacity data rates on AWGN channels, so the zero-outage capacity can be approximately achieved in practice.

Zero-outage capacity can exhibit a large data rate reduction relative to Shannon capacity in extreme fading environments. For example, in Rayleigh fading $\mathbf{E}[1/\gamma]$ is infinite, and thus the zero-outage capacity given by (4.18) is zero. Channel inversion is common in spread spectrum systems with near-far interference imbalances [26]. It is also the simplest scheme to implement, since the encoder and decoder are designed for an AWGN channel, independent of the fading statistics.

Example 4.5: Assume the same channel as in the previous example, with a bandwidth of 30 KHz and three possible received SNRs: $\gamma_1 = .8333$ with $p(\gamma_1) = .1$, $\gamma_2 = 83.33$ with $p(\gamma_2) = .5$, and $\gamma_3 = 333.33$ with $p(\gamma_3) = .4$. Assuming transmitter and receiver CSI, find the zero-outage capacity of this channel.

Solution: The zero-outage capacity is $C = B \log_2 [1 + \sigma]$, where $\sigma = 1/\mathbf{E}[1/\gamma]$. Since

$$\mathbf{E}[1/\gamma] = \frac{.1}{.8333} + \frac{.5}{83.33} + \frac{.4}{333.33} = .1272,$$

we have $C = 30000 \log_2 (1 + 1/.1272) = 9443$ Kbps. Note that this is less than half of the Shannon capacity with optimal water-filling adaptation.

Outage Capacity and Truncated Channel Inversion

The reason zero-outage capacity may be significantly smaller than Shannon capacity on a fading channel is the requirement to maintain a constant data rate in all fading states. By suspending transmission in particularly bad fading states (outage channel states), we can maintain a higher constant data rate in the other states and thereby significantly increase capacity. The **outage capacity** is defined as the maximum data rate that can be maintained in all nonoutage channel states times the probability of nonoutage. Outage capacity is achieved with a **truncated channel inversion** policy for power adaptation that only compensates for fading above a certain cutoff fade depth γ_0:

$$\frac{P(\gamma)}{\overline{P}} = \begin{cases} \frac{\sigma}{\gamma} & \gamma \geq \gamma_0 \\ 0 & \gamma < \gamma_0 \end{cases}, \tag{4.19}$$

where γ_0 is based on the outage probability: $p_{out} = p(\gamma < \gamma_0)$. Since the channel is only used when $\gamma \geq \gamma_0$, the power constraint (4.8) yields $\sigma = 1/\mathbf{E}_{\gamma_0}[1/\gamma]$, where

$$\mathbf{E}_{\gamma_0}[1/\gamma] \stackrel{\triangle}{=} \int_{\gamma_0}^{\infty} \frac{1}{\gamma} p(\gamma) d\gamma. \tag{4.20}$$

The outage capacity associated with a given outage probability p_{out} and corresponding cutoff γ_0 is given by

$$C(p_{out}) = B \log_2 \left(1 + \frac{1}{\mathbf{E}_{\gamma_0}[1/\gamma]}\right) p(\gamma \geq \gamma_0). \tag{4.21}$$

We can also obtain the **maximum outage capacity** by maximizing outage capacity over all possible γ_0:

$$C = \max_{\gamma_0} B \log_2 \left(1 + \frac{1}{\mathbf{E}_{\gamma_0}[1/\gamma]}\right) p(\gamma \geq \gamma_0). \tag{4.22}$$

This maximum outage capacity will still be less than Shannon capacity (4.13) since truncated channel inversion is a suboptimal transmission strategy. However, the transmit and receive strategies associated with inversion or truncated inversion may be easier to implement or have lower complexity than the water-filling schemes associated with Shannon capacity.

Example 4.6: Assume the same channel as in the previous example, with a bandwidth of 30 KHz and three possible received SNRs: $\gamma_1 = .8333$ with $p(\gamma_1) = .1$, $\gamma_2 = 83.33$ with $p(\gamma_2) = .5$, and $\gamma_3 = 333.33$ with $p(\gamma_3) = .4$. Find the outage capacity of this channel and associated outage probabilities for cutoff values $\gamma_0 = .84$ and $\gamma_0 = 83.4$. Which of these cutoff values yields a larger outage capacity?

Solution: For $\gamma_0 = .84$ we use the channel when the SNR is γ_2 or γ_3, so $\mathbf{E}_{\gamma_0}[1/\gamma] = \sum_{i=2}^{3} p(\gamma_i)/\gamma_i = .5/83.33 + .4/333.33 = .0072$. The outage capacity is $C = B\log_2(1 + 1/\mathbf{E}_{\gamma_0}[1/\gamma])p(\gamma \geq \gamma_0) = 30000 \log_2(1 + 138.88) * .9 = 192.457$. For $\gamma_0 = 83.34$ we use the channel when the SNR is γ_3 only, so $\mathbf{E}_{\gamma_0}[1/\gamma] = p(\gamma_3)/\gamma_3 = .4/333.33 = .0012$. The capacity is $C = B\log_2(1 + 1/\mathbf{E}_{\gamma_0}[1/\gamma])p(\gamma \geq \gamma_0) = 30000 \log_2(1 + 833.33) * .4 = 116.45$ Kbps. The outage capacity is larger when the channel is used for SNRs γ_2 and γ_3. Even though the SNR γ_3 is significantly larger than γ_2, the fact that this SNR only occurs 40% of the time makes it inefficient to only use the channel in this best state.

4.2.5 Capacity with Receiver Diversity

Receiver diversity is a well-known technique to improve the performance of wireless communications in fading channels. The main advantage of receiver diversity is that it mitigates the fluctuations due to fading so that the channel appears more like an AWGN channel. More details on receiver diversity and its performance will be given in Chapter 7. Since receiver diversity mitigates the impact of fading, an interesting question is whether it also increases the capacity of a fading channel. The capacity calculation under diversity combining first requires that the distribution of the received SNR $p(\gamma)$ under the given diversity combining technique be obtained. Once this distribution is known it can be substituted into any of the capacity formulas above to obtain the capacity under diversity combining. The specific capacity formula used depends on the assumptions about channel side information, e.g. for the case of perfect transmitter and receiver CSI the formula (4.13) would be used. Capacity under both maximal ratio and selection combining diversity for these different capacity formulas was computed in [23]. It was found that, as expected, the capacity with perfect transmitter and receiver CSI is bigger than with receiver CSI only, which in turn is bigger than with channel inversion. The performance gap of these different formulas decreases as the number of antenna branches increases. This trend is expected, since a large number of antenna branches makes the channel look like AWGN, for which all of the different capacity formulas have roughly the same performance.

Recently there has been much research activity on systems with multiple antennas at both the transmitter and the receiver. The excitement in this area stems from the breakthrough results in [28, 27, 29] indicating that the capacity of a fading channel with multiple inputs and outputs (a MIMO channel) is M times larger then the channel capacity without multiple antennas, where $M = \min(M_t, M_r)$ for M_t the number of transmit antennas and M_r the number of receive antennas. We will discuss capacity of multiple antenna systems in Chapter 10.

4.2.6 Capacity Comparisons

In this section we compare capacity with transmitter and receiver CSI for different power allocation policies along with the capacity under receiver CSI only. Figures 4.6, 4.7, and 4.8 show plots of the different capacities (4.4), 4.9), (4.18), and (4.22) as a function of average received SNR for log-normal fading (σ=8 dB standard deviation), Rayleigh fading, and Nakagami fading (with Nakagami parameter $m = 2$). Nakagami fading with $m = 2$ is roughly equivalent to Rayleigh fading with two-antenna receiver diversity. The capacity in AWGN for the same average power is also shown for comparison. Note that the capacity in log-normal fading is plotted relative to average dB SNR (μ_{dB}), not average SNR in dB ($10\log_{10}\mu$): the relation between these values, as given by (**??**) in Chapter 2, is $10\log_{10}\mu = \mu_{dB} + \sigma_{dB}^2 \ln(10)/20$.

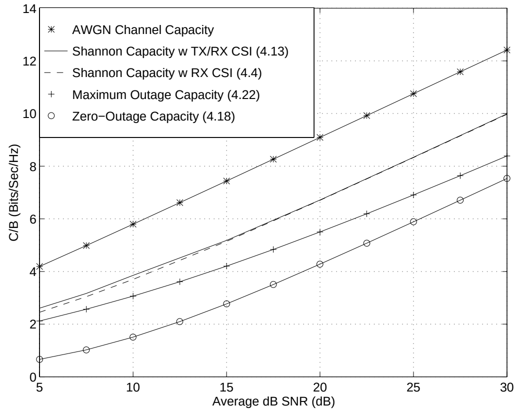


Figure 4.6: Capacity in Log-Normal Shadowing.

Several observations in this comparison are worth noting. First, we see in the figure that the capacity of the AWGN channel is larger than that of the fading channel for all cases. However, at low SNRs the AWGN and fading channel with transmitter and receiver CSI have almost the same capacity. In fact, at low SNRs (below 0 dB), capacity of the fading channel with transmitter and receiver CSI is larger than the corresponding AWGN channel capacity. That is because the AWGN channel always has the same low SNR, thereby limiting it capacity. A fading channel with this same low average SNR will occasionally have a high SNR, since the distribution has infinite range. Thus, if all power and rate is transmitted over the channel during these very infrequent high SNR values, the capacity will be larger than on the AWGN channel with the same low average SNR.

The severity of the fading is indicated by the Nakagami parameter m, where $m = 1$ for Rayleigh fading and $m = \infty$ for an AWGN channel without fading. Thus, comparing Figures 4.7 and 4.8 we see that, as the severity

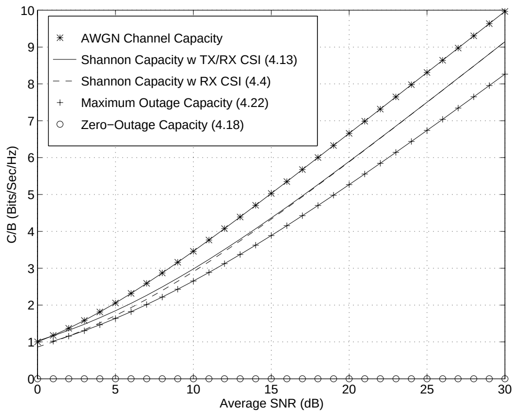


Figure 4.7: Capacity in Rayleigh Fading.

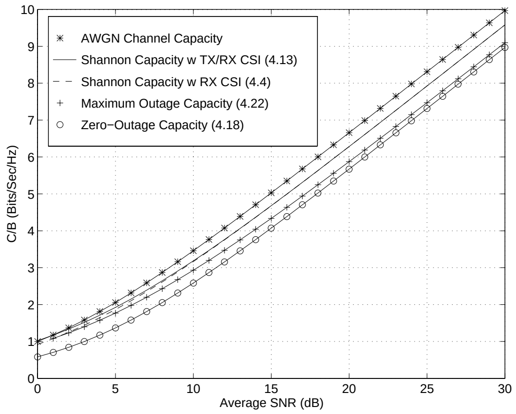


Figure 4.8: Capacity in Nakagami Fading ($m = 2$).

of the fading decreases (Rayleigh to Nakagami with $m = 2$), the capacity difference between the various adaptive policies also decreases, and their respective capacities approach that of the AWGN channel.

The difference between the capacity curves under transmitter and receiver CSI (4.9) and receiver CSI only (4.4) are negligible in all cases. Recalling that capacity under receiver CSI only (4.4) and under transmitter and receiver CSI without power adaptation (4.7) are the same, this implies that when the transmission rate is adapted relative to the channel, adapting the power as well yields a negligible capacity gain. It also indicates that transmitter adaptation yields a negligible capacity gain relative to using only receiver side information. In severe fading conditions (Rayleigh and log-normal fading), maximum outage capacity exhibits a 1-5 dB rate penalty and zero-outage capacity yields a very large capacity loss relative to Shannon capacity. However, under mild fading conditions (Nakagami with $m = 2$) the Shannon, maximum outage, and zero-outage capacities are within 3 dB of each other and within 4 dB of the AWGN channel capacity. These differences will further decrease as the fading diminishes ($m \to \infty$ for Nakagami fading).

We can view these results as a tradeoff between capacity and complexity. The adaptive policy with transmitter and receiver side information requires more complexity in the transmitter (and it typically also requires a feedback path between the receiver and transmitter to obtain the side information). However, the decoder in the receiver is relatively simple. The nonadaptive policy has a relatively simple transmission scheme, but its code design must use the channel correlation statistics (often unknown), and the decoder complexity is proportional to the channel decorrelation time. The channel inversion and truncated inversion policies use codes designed for AWGN channels, and are therefore the least complex to implement, but in severe fading conditions they exhibit large capacity losses relative to the other techniques.

In general, Shannon capacity analysis does not show how to design adaptive or nonadaptive techniques for real systems. Achievable rates for adaptive trellis-coded MQAM have been investigated in [25], where a simple 4-state trellis code combined with adaptive six-constellation MQAM modulation was shown to achieve rates within 7 dB of the Shannon capacity (4.9) in Figures 4.6 and 4.7. More complex codes further close the gap to the Shannon limit of fading channels with transmitter adaptation.

4.3 Capacity of Frequency-Selective Fading Channels

In this section we consider the Shannon capacity of frequency-selective fading channels. We first consider the capacity of a time-invariant frequency-selective fading channel. This capacity analysis is similar to that of a flat-fading channel with the time axis replaced by the frequency axis. Next we discuss the capacity of time-varying frequency-selective fading channels.

4.3.1 Time-Invariant Channels

Consider a time-invariant channel with frequency response $H(f)$, as shown in Figure 4.9. Assume a total transmit power constraint P. When the channel is time-invariant it is typically assumed that $H(f)$ is known at both the transmitter and receiver: capacity of time-invariant channels under different assumptions of this channel knowledge are discussed in [18].

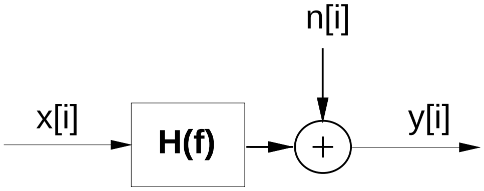


Figure 4.9: Time-Invariant Frequency-Selective Fading Channel.

Let us first assume that $H(f)$ is block-fading, so that frequency is divided into subchannels of bandwidth B, where $H(f) = H_j$ is constant over each block, as shown in Figure 4.10. The frequency-selective fading channel thus consists of a set of AWGN channels in parallel with SNR $|H_j|^2 P_j/(N_0 B)$ on the jth channel, where P_j is the power allocated to the jth channel in this parallel set, subject to the power constraint $\sum_j P_j \leq P$.

The capacity of this parallel set of channels is the sum of rates associated with each channel with power optimally allocated over all channels [5, 6]

$$C = \sum_{\max P_j : \sum_j P_j \leq P} B \log_2 \left(1 + \frac{|H_j|^2 P_j}{N_0 B} \right). \tag{4.23}$$

Note that this is similar to the capacity and optimal power allocation for a flat-fading channel, with power and rate changing over frequency in a deterministic way rather than over time in a probabilistic way. The optimal power allocation is found via the same Lagrangian technique used in the flat-fading case, which leads to the water-filling power allocation

$$\frac{P_j}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_j} & \gamma_j \geq \gamma_0 \\ 0 & \gamma_j < \gamma_0 \end{cases} \tag{4.24}$$

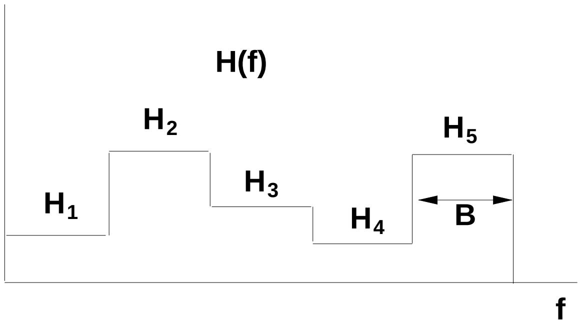


Figure 4.10: Block Frequency-Selective Fading

for some cutoff value γ_0, where $\gamma_j = |H_j|^2 P/(N_0 B)$ is the SNR associated with the jth channel assuming it is allocated the entire power budget. This optimal power allocation is illustrated in Figure 4.11. The cutoff value is obtained by substituting the power adaptation formula into the power constraint, so γ_0 must satisfy

$$\sum_j \left(\frac{1}{\gamma_0} - \frac{1}{\gamma_j} \right) = 1. \tag{4.25}$$

The capacity then becomes

$$C = \sum_{j:\gamma_j \geq \gamma_0} B \log_2(\gamma_j/\gamma_0). \tag{4.26}$$

This capacity is achieved by sending at different rates and powers over each subchannel. Multicarrier modulation uses the same technique in adaptive loading, as discussed in more detail in Chapter 12.

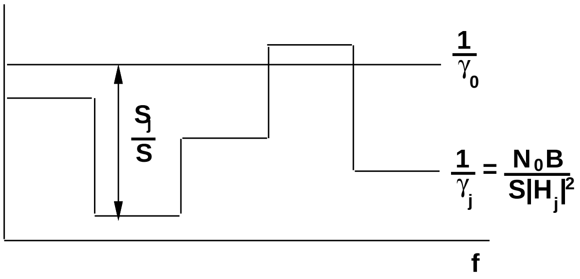


Figure 4.11: Water-Filling in Block Frequency-Selective Fading

When $H(f)$ is continuous the capacity under power constraint P is similar to the case of the block-fading channel, with some mathematical intricacies needed to show that the channel capacity is given by

$$C = \max_{P(f): \int P(f) df \leq P} \int \log_2 \left(1 + \frac{|H(f)|^2 P(f)}{N_0} \right) df. \tag{4.27}$$

The equation inside the integral can be thought of as the incremental capacity associated with a given frequency f over the bandwidth df with power allocation $P(f)$ and channel gain $|H(f)|^2$. This result is formally proven using a Karhunen-Loeve expansion of the channel $h(t)$ to create an equivalent set of parallel independent channels [5, Chapter 8.5]. An alternate proof decomposes the channel into a parallel set using the discrete Fourier transform (DFT) [12]: the same premise is used in the discrete implementation of multicarrier modulation described in Chapter 12.4.

The power allocation over frequency, $P(f)$, that maximizes (4.27) is found via the Lagrangian technique. The resulting optimal power allocation is water-filling over frequency:

$$\frac{P(f)}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma(f)} & \gamma(f) \geq \gamma_0 \\ 0 & \gamma(f) < \gamma_0 \end{cases} \tag{4.28}$$

This results in channel capacity

$$C = \int_{f:\gamma(f)\geq\gamma_0} \log_2(\gamma(f)/\gamma_0) df. \tag{4.29}$$

Example 4.7: Consider a time-invariant frequency-selective block fading channel consisting of three subchannels of bandwidth $B = 1$ MHz. The frequency response associated with each channel is $H_1 = 1$, $H_2 = 2$ and $H_3 = 3$. The transmit power constraint is $P = 10$ mW and the noise PSD is $N_0 = 10^{-9}$ W/Hz. Find the Shannon capacity of this channel and the optimal power allocation that achieves this capacity.

Solution: We first first find $\gamma_j = |H_j|^2 P/(N_b)$ for each subchannel, yielding $\gamma_1 = 10$, $\gamma_2 = 40$ and $\gamma_3 = 90$. The cutoff γ_0 must satisfy (4.25). Assuming all subchannels are allocated power, this yields

$$\frac{3}{\gamma_0} = 1 + \sum_j \frac{1}{\gamma_j} = 1.14 \Rightarrow \gamma_0 = 2.64 < \gamma_j \, \forall j.$$

Since the cutoff γ_0 is less than γ_j for all j, our assumption that all subchannels are allocated power is consistent, so this is the correct cutoff value. The corresponding capacity is $C = \sum_{j=1}^{3} B \log_2(\gamma_j/\gamma_0) = 1000000(\log_2(10/2.64) + \log_2(40/2.64) + \log_2(90/2.64)) = 10.93$ Mbps.

4.3.2 Time-Varying Channels

The time-varying frequency-selective fading channel is similar to the model shown in Figure 4.9, except that $H(f) = H(f, i)$, i.e. the channel varies over both frequency and time. It is difficult to determine the capacity of time-varying frequency-selective fading channels, even when the instantaneous channel $H(f, i)$ is known perfectly at the transmitter and receiver, due to the random effects of self-interference (ISI). In the case of transmitter and receiver side information, the optimal adaptation scheme must consider the effect of the channel on the past sequence of transmitted bits, and how the ISI resulting from these bits will affect future transmissions [30]. The capacity of time-varying frequency-selective fading channels is in general unknown, however upper and lower bounds and limiting formulas exist [30, 31].

We can approximate channel capacity in time-varying frequency-selective fading by taking the channel bandwidth B of interest and divide it up into subchannels the size of the channel coherence bandwidth B_c, as shown in Figure 4.12. We then assume that each of the resulting subchannels is independent, time-varying, and flat-fading with $H(f, i) = H_j[i]$ on the jth subchannel.

Under this assumption, we obtain the capacity for each of these flat-fading subchannels based on the average power $\overline{P}_j$ that we allocate to each subchannel, subject to a total power constraint $\overline{P}$. Since the channels are independent, the total channel capacity is just equal to the sum of capacities on the individual narrowband flat-fading channels subject to the total average power constraint, averaged over both time and frequency:

$$C = \max_{\{\overline{P}_j\}:\sum_j \overline{P}_j \leq \overline{P}} \sum_j C_j(\overline{P}_j), \tag{4.30}$$

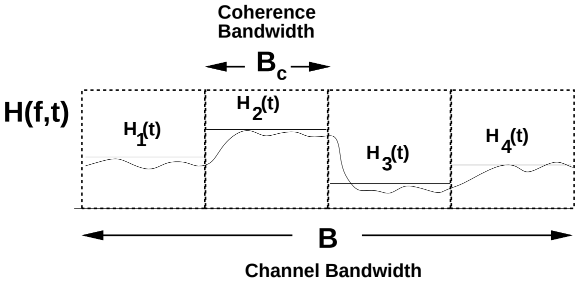


Figure 4.12: Channel Division in Frequency-Selective Fading

where $C_j(\overline{P}_j)$ is the capacity of the flat-fading subchannel with average power $\overline{P}_j$ and bandwidth B_c given by (4.13), (4.4), (4.18), or (4.22) for Shannon capacity under different side information and power allocation policies. We can also define $C_j(\overline{S}_j)$ as a capacity versus outage if only the receiver has side information.

We will focus on Shannon capacity assuming perfect transmitter and receiver channel CSI, since this upper-bounds capacity under any other side information assumptions or suboptimal power allocation strategies. We know that if we fix the average power per subchannel, the optimal power adaptation follows a water-filling formula. We also expect that the optimal average power to be allocated to each subchannel should also follow a water-filling, where more average power is allocated to better subchannels. Thus we expect that the optimal power allocation is a two-dimensional water-filling in both time and frequency. We now obtain this optimal two-dimensional water-filling and the corresponding Shannon capacity.

Define $\gamma_j[i] = |H_j[i]|^2\overline{P}/(N_0B)$ to be the instantaneous SNR on the jth subchannel at time i assuming the total power $\overline{P}$ is allocated to that time and frequency. We allow the power $P_j(\gamma_j)$ to vary with $\gamma_j[i]$. The Shannon capacity with perfect transmitter and receiver CSI is given by optimizing power adaptation relative to both time (represented by $\gamma_j[i] = \gamma_j$) and frequency (represented by the subchannel index j):

$$C = \max_{P_j(\gamma_j):\sum_j \int_0^\infty P_j(\gamma_j)p(\gamma_j)d\gamma_j \leq \overline{P}} \sum_j \int_0^\infty B_c \log_2\left(1+\frac{P_j(\gamma_j)\gamma_j}{\overline{P}}\right)p(\gamma_j)d\gamma_j. \tag{4.31}$$

To find the optimal power allocation $P_j(\gamma_j)$, we form the Lagrangian

$$J(P_j(\gamma_j)) = \sum_j \int_0^\infty B_c \log_2\left(1+\frac{P_j(\gamma_j)\gamma_j}{\overline{P}}\right)p(\gamma_j)d\gamma_j - \lambda\sum_j \int_0^\infty P_j(\gamma_j)p(\gamma_j)d\gamma_j. \tag{4.32}$$

Note that (4.32) is similar to the Lagrangian for the flat-fading channel (4.10) except that the dimension of frequency has been added by summing over the subchannels. Differentiating the Lagrangian and setting this derivative equal to zero eliminates all terms except the given subchannel and associated SNR:

$$\frac{\partial J(P_j(\gamma_j))}{\partial P_j(\gamma_j)} = \left[\left(\frac{B/\ln(2)}{1+\gamma_j P(\gamma_j)/\overline{P}}\right)\frac{\gamma_j}{\overline{P}} - \lambda\right]p(\gamma_j) = 0. \tag{4.33}$$

Solving for $P_j(\gamma_j)$ yields the same water-filling as the flat fading case:

$$\frac{P_j(\gamma_j)}{\overline{P}} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_j} & \gamma_j \geq \gamma_0 \\ 0 & \gamma_j < \gamma_0 \end{cases}, \tag{4.34}$$

where the cutoff value γ_0 is obtained from the total power constraint over both time and frequency:

$$\sum_j \int_0^\infty P_j(\gamma)p_j(\gamma)d\gamma_j = \overline{P}. \tag{4.35}$$

Thus, the optimal power allocation (4.34) is a two-dimensional waterfilling with a common cutoff value γ_0. Dividing the constraint (4.35) by $\overline{P}$ and substituting in the optimal power allocation (4.34), we get that γ_0 must satisfy

$$\sum_j \int_{\gamma_0}^{\infty} \left(\frac{1}{\gamma_0} - \frac{1}{\gamma_j} \right) p(\gamma_j) d\gamma_j = 1. \tag{4.36}$$

It is interesting to note that in the two-dimensional water-filling the cutoff value for all subchannels is the same. This implies that even if the fading distribution or average fade power on the subchannels is different, all subchannels suspend transmission when the instantaneous SNR falls below the common cutoff value γ_0. Substituting the optimal power allocation (4.35) into the capacity expression (4.31) yields

$$C = \sum_j \int_{\gamma_0}^{\infty} B_c \log_2 \left(\frac{\gamma_j}{\gamma_0} \right) p(\gamma_j) d\gamma_j. \tag{4.37}$$

Bibliography

[1] C. E. Shannon *A Mathematical Theory of Communication*. *Bell Sys. Tech. Journal*, pp. 379–423, 623–656, 1948.

[2] C. E. Shannon *Communications in the presence of noise*. *Proc. IRE*, pp. 10-21, 1949.

[3] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: Univ. Illinois Press, 1949.

[4] M. Medard, “The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel,” *IEEE Trans. Inform. Theory*, pp. 933-946, May 2000.

[5] R.G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[6] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] C. Heegard and S.B. Wicker, *Turbo Coding*. Kluwer Academic Publishers, 1999.

[8] I. Csiszár and J. Kórner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. New York: Academic Press, 1981.

[9] I. Csiszár and P. Narayan, “The capacity of the Arbitrarily Varying Channel,” *IEEE Trans. Inform. Theory*, Vol. 37, No. 1, pp. 18–26, Jan. 1991.

[10] I.C. Abou-Faycal, M.D. Trott, and S. Shamai, “The capacity of discrete-time memoryless Rayleigh fading channels,” *IEEE Trans. Inform. Theory*, pp. 1290–1301, May 2001.

[11] A. Lapidoth and S. M. Moser, “Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels,” *IEEE Trans. Inform. Theory*, pp. 2426-2467, Oct. 2003.

[12] W. Hirt and J.L. Massey, “Capacity of the discrete-time Gaussian channel with intersymbol interference,” *IEEE Trans. Inform. Theory*, Vol. 34, No. 3, pp. 380-388, May 1988

[13] A.J. Goldsmith and P.P. Varaiya, “Capacity, mutual information, and coding for finite-state Markov channels,” *IEEE Trans. Inform. Theory*. pp. 868–886, May 1996.

[14] M. Mushkin and I. Bar-David, “Capacity and coding for the Gilbert-Elliot channel,” *IEEE Trans. Inform. Theory*, Vol. IT-35, No. 6, pp. 1277–1290, Nov. 1989.

[15] T. Holliday, A. Goldsmith, and P. Glynn, “Capacity of Finite State Markov Channels with general inputs,” *Proc. IEEE Intl. Symp. Inform. Theory*, pg. 289, July 2003. Also submitted to *IEEE Trans. Inform. Theory*.

[16] R.J. McEliece and W. E. Stark, “Channels with block interference,” *IEEE Trans. Inform. Theory*, Vol IT-30, No. 1, pp. 44-53, Jan. 1984.

[17] G.J. Foschini, D. Chizhik, M. Gans, C. Papadias, and R.A. Valenzuela, “Analysis and performance of some basic space-time architectures,” newblock *IEEE J. Select. Areas Commun.*, pp. 303–320, April 2003.

[18] W.L. Root and P.P. Varaiya, “Capacity of classes of Gaussian channels,” *SIAM J. Appl. Math*, pp. 1350-1393, Nov. 1968.

[19] J. Wolfowitz, *Coding Theorems of Information Theory*. 2nd Ed. New York: Springer-Verlag, 1964.

[20] A. Lapidoth and S. Shamai, “Fading channels: how perfect need “perfect side information” be?” *IEEE Trans. Inform. Theory*, pp. 1118-1134, Nov. 1997.

[21] A.J. Goldsmith and P.P. Varaiya, “Capacity of fading channels with channel side information,” *IEEE Trans. Inform. Theory*, pp. 1986-1992, Nov. 1997.

[22] G. Caire and S. Shamai, “On the capacity of some channels with channel state information,” *IEEE Trans. Inform. Theory*, pp. 2007–2019, Sept. 1999.

[23] M.S. Alouini and A. J. Goldsmith, “Capacity of Rayleigh fading channels under different adaptive transmission and diversity combining techniques,” *IEEE Transactions on Vehicular Technology,* pp. 1165–1181, July 1999.

[24] S.-G. Chua and A.J. Goldsmith, “Variable-rate variable-power MQAM for fading channels,” *IEEE Trans. on Communications*, pp. 1218-1230, Oct. 1997.

[25] S.-G. Chua and A.J. Goldsmith, “Adaptive coded modulation,” *IEEE Trans. on Communications*, pp. 595-602, May 1998.

[26] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr., and C. E. Wheatley III, “On the capacity of a cellular CDMA system,” *IEEE Trans. Vehic. Technol.*, Vol. VT-40, No. 2, pp. 303–312, May 1991.

[27] E. Teletar, “Capacity of multi-antenna Gaussian channels,” AT&T Bell Labs Internal Tech. Memo, June 1995.

[28] G. Foschini, “Layered space-time architecture for wireless communication in a fading environment when using multiple antennas,” *Bell Labs Technical Journal*, pp. 41-59, Autumn 1996.

[29] G. Foschini and M. Gans, “On limits of wireless communication in a fading environment when using multiple antennas,” *Wireless Personal Communications*, pp. 311-335, March 1998.

[30] A. Goldsmith and M Medard, “Capacity of time-varying channels with channel side information,” *IEEE Intl. Symp. Inform. Theory*, pg. 372, Oct. 1996. Also to appear: *IEEE Trans. Inform. Theory*.

[31] S. Diggavi, “Analysis of multicarrier transmission in time-varying channels,” *Proc. IEEE Intl. Conf. Commun.* pp. 1191–1195, June 1997.

Chapter 4 Problems

1. Capacity in AWGN is given by $C = B\log_2(1 + S/(N_0B))$. Find capacity in the limit of infinite bandwidth $B \to \infty$ as a function of S.

2. Consider an AWGN channel with bandwidth 50 MHz, received power 10 mW, and noise PSD $N_0 = 2 \times 10^{-9}$W/Hz. How much does capacity increase by doubling the received power? How much does capacity increase by doubling the channel bandwidth?

3. Consider two users simultaneously transmitting to a single receiver in an AWGN channel. This is a typical scenario in a cellular system with multiple users sending signals to a base station. Assume the users have equal received power of 10 mW and total noise at the receiver in the bandwidth of interest of 0.1 mW. The channel bandwidth for each user is 20 MHz.

   (a) Suppose that the receiver decodes user 1's signal first. In this decoding, user 2's signal acts as noise (assume it has the same statistics as AWGN). What is the capacity of user 1's channel with this additional interference noise?

   (b) Suppose that after decoding user 1's signal, the decoder re-encodes it and subtracts it out of the received signal. Then in the decoding of user 2's signal, there is no interference from user 1's signal. What then is the Shannon capacity of user 2's channel?

   Note: We will see in Chapter 14 that the decoding strategy of successively subtracting out decoded signals is optimal for achieving Shannon capacity of a multiuser channel with independent transmitters sending to one receiver.

4. Consider a flat-fading channel of bandwidth 20 MHz where for a fixed transmit power $\overline{S}$, the received SNR is one of six values: $\gamma_1 = 20$ dB, $\gamma_2 = 15$ dB, $\gamma_3 = 10$ dB, $\gamma_4 = 5$ dB, and $\gamma_5 = 0$ dB and $\gamma_6 = -5$ dB. The probability associated with each state is $p_1 = p_6 = .1$, $p_2 = p_4 = .15$, $p_3 = p_5 = .25$. Assume only the receiver has CSI.

   (a) Find the Shannon capacity of this channel.

   (b) Plot the capacity versus outage for $0 \leq p_{out} < 1$ and find the maximum average rate that can be correctly received (maximum C_o).

5. Consider a flat-fading channel where for a fixed transmit power $\overline{S}$, the received SNR is one of four values: $\gamma_1 = 30$ dB, $\gamma_2 = 20$ dB, $\gamma_3 = 10$ dB, and $\gamma_4 = 0$ dB. The probability associated with each state is $p_1 = .2$, $p_2 = .3$, $p_3 = .3$, and $p_4 = .2$. Assume both transmitter and receiver have CSI.

   (a) Find the optimal power control policy $S(i)/\overline{S}$ for this channel and its corresponding Shannon capacity per unit Hertz (C/B).

   (b) Find the channel inversion power control policy for this channel and associated zero-outage capacity per unit bandwidth.

   (c) Find the truncated channel inversion power control policy for this channel and associated outage capacity per unit bandwidth for 3 different outage probabilities: $p_{out} = .1$, $p_{out} = .01$, and p_{out} (and the associated cutoff γ_0) equal to the value that achieves maximum outage capacity.

6. Consider a cellular system where the power falloff with distance follows the formula $P_r(d) = P_t(d_0/d)^\alpha$, where $d_0 = 100$m and α is a random variable. The distribution for α is $p(\alpha = 2) = .4$, $p(\alpha = 2.5) = .3$, $p(\alpha = 3) = .2$, and $p(\alpha = 4) = .1$ Assume a receiver at a distance $d = 1000$ m from the transmitter, with

an average transmit power constraint of $P_t = 100$ mW and a receiver noise power of .1 mW. Assume both transmitter and receiver have CSI.

(a) Compute the distribution of the received SNR.

(b) Derive the optimal power control policy for this channel and its corresponding Shannon capacity per unit Hertz (C/B).

(c) Determine the zero-outage capacity per unit bandwidth of this channel.

(d) Determine the maximum outage capacity per unit bandwidth of this channel.

7. Assume a Rayleigh fading channel, where the transmitter and receiver have CSI and the distribution of the fading SNR $p(\gamma)$ is exponential with mean $\overline{\gamma} = 10$dB. Assume a channel bandwidth of 10 MHz.

(a) Find the cutoff value γ_0 and the corresponding power adaptation that achieves Shannon capacity on this channel.

(b) Compute the Shannon capacity of this channel.

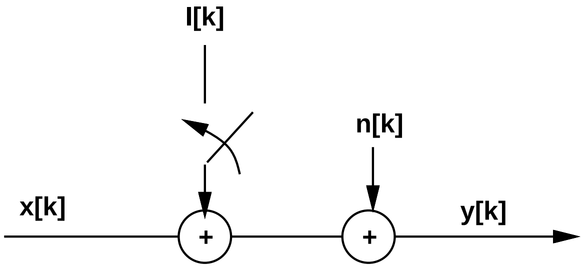
(c) Compare your answer in part (b) with the channel capacity in AWGN with the same average SNR.

(d) Compare your answer in part (b) with the Shannon capacity where only the receiver knows $\gamma[i]$.

(e) Compare your answer in part (b) with the zero-outage capacity and outage capacity with outage probability .05.

(f) Repeat parts b, c, and d (i.e. obtain the Shannon capacity with perfect transmitter and receiver side information, in AWGN for the same average power, and with just receiver side information) for the same fading distribution but with mean $\overline{\gamma} = -5$dB. Describe the circumstances under which a fading channel has higher capacity than an AWGN channel with the same average SNR and explain why this behaivor occurs.

8. Time-Varying Interference: This problem illustrates the capacity gains that can be obtained from interference estimation, and how a malicious jammer can wreak havoc on link performance. Consider the following interference channel.



The channel has a combination of AWGN $n[k]$ and interference $I[k]$. We model $I[k]$ as AWGN. The interferer is on (i.e. the switch is down) with probability .25 and off (i.e. the switch is up) with probability .75. The average transmit power is 10 mW, the noise spectral density is 10^{-8} W/Hz, the channel bandwidth B is 10 KHz (receiver noise power is N_oB), and the interference power (when on) is 9 mW.

(a) What is the Shannon capacity of the channel if neither transmitter nor receiver know when the interferer is on?

(b) What is the capacity of the channel if both transmitter and receiver know when the interferer is on?

(c) Suppose now that the interferer is a malicious jammer with perfect knowledge of $x[k]$ (so the interferer is no longer modeled as AWGN). Assume that neither transmitter nor receiver have knowledge of the jammer behavior. Assume also that the jammer is always on and has an average transmit power of 10 mW. What strategy should the jammer use to minimize the SNR of the received signal?

9. Consider the malicious interferer from the previous problem. Suppose that the transmitter knows the interference signal perfectly. Consider two possible transmit strategies under this scenario: the transmitter can ignore the interference and use all its power for sending its signal, or it can use some of its power to cancel out the interferer (i.e. transmit the negative of the interference signal). In the first approach the interferer will degrade capacity by increasing the noise, and in the second strategy the interferer also degrades capacity since the transmitter sacrifices some power to cancel out the interference. Which strategy results in higher capacity? *Note: there is a third strategy, where the encoder actually exploits the structure of the interference in its encoding. This strategy is called dirty paper coding, and is used to achieve Shannon capacity on broadcast channels with multiple antennas.*

10. Show using Lagrangian techniques that the optimal power allocation to maximize the capacity of a time-invariant block fading channel is given by the water filling formula in (4.24).

11. Consider a time-invariant block fading channel with frequency response

$$H(f) = \begin{cases} 1 & f_c - 20\text{MHz} \le f < f_c - 10\text{MHz} \\ .5 & f_c - 10\text{MHz} \le f < f_c \\ 2 & f_c \le f < f_c + 10\text{MHz} \\ .25 & f_c + 10\text{MHz} \le f < f_c + 20\text{MHz} \\ 0 & \text{else} \end{cases}$$

For a transmit power of 10mW and a noise power spectral density of $.001\mu$W per Hertz, find the optimal power allocation and corresponding Shannon capacity of this channel.

12. Show that the optimal power allocation to maximize the capacity of a time-invariant frequency selective fading channel is given by the water filling formula in (4.28).

13. Consider a frequency-selective fading channel with total bandwidth 12 MHz and coherence bandwidth $B_c = 4$ MHz. Divide the total bandwidth into 3 subchannels of bandwidth B_c , and assume that each subchannel is a Rayleigh flat-fading channel with independent fading on each subchannel. Assume the subchannels have average gains $\mathbf{E}[|H_1(t)|^2] = 1$, $\mathbf{E}[|H_2(t)|^2] = .5$, and $\mathbf{E}[|H_3(t)|^2] = .125$. Assume a total transmit power of 30 mW, and a receiver noise spectral density of $.001\mu$W per Hertz.

(a) Find the optimal two-dimensional water-filling power adaptation for this channel and the corresponding Shannon capacity, assuming both transmitter and receiver know the instantaneous value of $H_j(t), j = 1, 2, 3$.

(b) Compare the capacity of part (a) with that obtained by allocating an equal average power of 10 mW to each subchannel and then water-filling on each subchannel relative to this power allocation.

Chapter 6

Performance of Digital Modulation over Wireless Channels

We now consider the performance of the digital modulation techniques discussed in the previous chapter when used over AWGN channels and channels with flat-fading. There are two performance criteria of interest: the probability of error, defined relative to either symbol or bit errors, and the outage probability, defined as the probability that the instantaneous signal-to-noise ratio falls below a given threshold. Flat-fading can cause a dramatic increase in either the average bit-error-rate or the signal outage probability. Wireless channels may also exhibit frequency selective fading and Doppler shift. Frequency-selective fading gives rise to intersymbol interference (ISI), which causes an irreducible error floor in the received signal. Doppler causes spectral broadening, which leads to adjacent channel interference (typically small at reasonable user velocities), and also to an irreducible error floor in signals with differential phase encoding (e.g. DPSK), since the phase reference of the previous symbol partially decorrelates over a symbol time. This chapter describes the impact on digital modulation performance of noise, flat-fading, frequency-selective fading, and Doppler.

6.1 AWGN Channels

In this section we define the signal-to-noise power ratio (SNR) and its relation to energy-per-bit (E_b) and energy-per-symbol (E_s). We then examine the error probability on AWGN channels for different modulation techniques as parameterized by these energy metrics. Our analysis uses the signal space concepts of Chapter **??**.

6.1.1 Signal-to-Noise Power Ratio and Bit/Symbol Energy

In an AWGN channel the modulated signal $s(t) = \Re\{u(t)e^{j2\pi f_c t}\}$ has noise $n(t)$ added to it prior to reception. The noise $n(t)$ is a white Gaussian random process with mean zero and power spectral density $N_0/2$. The received signal is thus $r(t) = s(t) + n(t)$.

We define the received signal-to-noise power ratio (SNR) as the ratio of the received signal power P_r to the power of the noise within the bandwidth of the transmitted signal $s(t)$. The received power P_r is determined by the transmitted power and the path loss, shadowing, and multipath fading, as described in Chapters 2-3. The noise power is determined by the bandwidth of the transmitted signal and the spectral properties of $n(t)$. Specifically, if the bandwidth of the complex envelope $u(t)$ of $s(t)$ is B then the bandwidth of the transmitted signal $s(t)$ is $2B$. Since the noise $n(t)$ has uniform power spectral density $N_0/2$, the total noise power within the bandwidth $2B$ is

$N = N_0/2 \times 2B = N_0B$. So the received SNR is given by

$$\text{SNR} = \frac{P_r}{N_0B}.$$

In systems with interference, we often use the received signal-to-interference-plus-noise power ratio (SINR) in place of SNR for calculating error probability. If the interference statistics approximate those of Gaussian noise then this is a reasonable approximation. The received SINR is given by

$$\text{SINR} = \frac{P_r}{N_0B + P_I},$$

where P_I is the average power of the interference.

The SNR is often expressed in terms of the signal energy per bit E_b or per symbol E_s as

$$\text{SNR} = \frac{P_r}{N_0B} = \frac{E_s}{N_0BT_s} = \frac{E_b}{N_0BT_b}, \tag{6.1}$$

where T_s is the symbol time and T_b is the bit time (for binary modulation $T_s = T_b$ and $E_s = E_b$). For data pulses with $T_s = 1/B$, e.g. raised cosine pulses with $\beta = 1$, we have SNR $= E_s/N_0$ for multilevel signaling and SNR $= E_b/N_0$ for binary signaling. For general pulses, $T_s = k/B$ for some constant k, in which case $k \cdot \text{SNR} = E_s/N_0$.

The quantities $\gamma_s = E_s/N_0$ and $\gamma_b = E_b/N_0$ are sometimes called the SNR per symbol and the SNR per bit, respectively. For performance specification, we are interested in the bit error probability P_b as a function of γ_b. However, for M-aray signaling (e.g. MPAM and MPSK), the bit error probability depends on both the symbol error probability and the mapping of bits to symbols. Thus, we typically compute the symbol error probability P_s as a function of γ_s based on the signal space concepts of Chapter **??** and then obtain P_b as a function of γ_b using an exact or approximate conversion. The approximate conversion typically assumes that the symbol energy is divided equally among all bits, and that Gray encoding is used so that at reasonable SNRs, one symbol error corresponds to exactly one bit error. These assumptions for M-aray signaling lead to the approximations

$$\gamma_b \approx \frac{\gamma_s}{\log_2 M} \tag{6.2}$$

and

$$P_b \approx \frac{P_s}{\log_2 M}. \tag{6.3}$$

6.1.2 Error Probability for BPSK and QPSK

We first consider BPSK modulation with coherent detection and perfect recovery of the carrier frequency and phase. With binary modulation each symbol corresponds to one bit, so the symbol and bit error rates are the same. The transmitted signal is $s_1(t) = Ag(t)\cos(2\pi f_c t)$ to sent a 0 bit and $s_2(t) = -Ag(t)\cos(2\pi f_c t)$ to send a 1 bit. From (**??**) we have that the probability of error is

$$P_b = Q\left(\frac{d_{min}}{\sqrt{2N_0}}\right). \tag{6.4}$$

From Chapter 5, $d_{min} = ||s_1 - s_0|| = ||A - (-A)|| = 2A$. Let us now relate A to the energy-per-bit. We have

$$E_b = \int_0^{T_b} s_1^2(t)dt = \int_0^{T_b} s_2^2(t)dt = \int_0^{T_b} A^2g^2(t)\cos^2(2\pi f_c t)dt = A^2 \tag{6.5}$$

from (**??**). Thus, the signal constellation for BPSK in terms of energy-per-bit is given by $\mathbf{s}_0 = \sqrt{E_b}$ and $\mathbf{s}_1 = -\sqrt{E_b}$. This yields the minimum distance $d_{min} = 2A = 2\sqrt{E_b}$. Substituting this into (6.4) yields

$$P_b = Q\left(\frac{2\sqrt{E_b}}{\sqrt{2N_0}}\right) = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) = Q(\sqrt{2\gamma_b}). \tag{6.6}$$

QPSK modulation consists of BPSK modulation on both the in-phase and quadrature components of the signal. With perfect phase and carrier recovery, the received signal components corresponding to each of these branches are orthogonal. Therefore, the bit error probability on each branch is the same as for BPSK: $P_b = Q(\sqrt{2\gamma_b})$. The symbol error probability equals the probability that either branch has a bit error:

$$P_s = 1 - [1 - Q(\sqrt{2\gamma_b})]^2 \tag{6.7}$$

Since the symbol energy is split between the in-phase and quadrature branches, we have $\gamma_s = 2\gamma_b$. Substituting this into (6.7) yields P_s is terms of γ_s as

$$P_s = 1 - [1 - Q(\sqrt{\gamma_s})]^2. \tag{6.8}$$

From Section **??**, the union bound (**??**) on P_s for QPSK is

$$P_s \leq 2Q(A/\sqrt{N_0}) + Q(\sqrt{2}A/\sqrt{N_0}). \tag{6.9}$$

Writing this in terms of $\gamma_s = 2\gamma_b = A^2/N_0$ yields

$$P_s \leq 2Q(\sqrt{\gamma_s}) + Q(\sqrt{2\gamma_s}) \leq eQ(\sqrt{\gamma_s}). \tag{6.10}$$

The closed form bound (**??**) becomes

$$P_s \leq \frac{3}{\sqrt{\pi}}\text{exp}\left[\frac{-.5A^2}{N_0}\right] = \frac{3}{\sqrt{\pi}}\text{exp}[-\gamma_s/2]. \tag{6.11}$$

Using the fact that the minimum distance between constellation points is $d_{min} = \sqrt{2A^2}$, we get the nearest neighbor approximation

$$P_s \approx 2Q\left(\sqrt{\frac{A^2}{N_0}}\right) = 2Q\left(\sqrt{\gamma_s/2}\right). \tag{6.12}$$

Note that with Gray encoding, we can approximate P_b from P_s by $P_b \approx P_s/2$, since we have 2 bits per symbol.

Example 6.1:
Find the bit error probability P_b and symbol error probability P_s of QPSK assuming $\gamma_b = 7$ dB. Compare the exact P_b with the approximation $P_b = P_s/2$ based on the assumption of Gray coding. Finally, compute P_s based on the nearest-neighbor bound using $\gamma_s = 2\gamma_b$, and compare with the exact P_s.

Solution: We have $\gamma_b = 10^{7/10} = 5.012$, so

$$P_b = Q(\sqrt{2\gamma_b}) = Q(\sqrt{10.024}) = 7.726 * 10^{-4}.$$

The exact symbol error probability P_s is

$$P_s = 1 - [1 - Q(\sqrt{2\gamma_b})]^2 = 1 - [1 - Q(\sqrt{10.02})]^2 = 1.545 * 10^{-3}.$$

The bit-error-probability approximation assuming Gray coding yields $P_b \approx P_s/2 = 7.723 * 10^{-4}$, which is quite close to the exact P_s. The nearest neighbor approximation to P_s yields

$$P_s \approx 2Q(\sqrt{\gamma_s}) = 2Q(\sqrt{10.024}) = 1.545 \times 10^{-3},$$

which matches well with the exact P_s.

6.1.3 Error Probability for MPSK

The signal constellation for MPSK has $s_{i1} = A\cos[\frac{2\pi(i-1)}{M}]$ and $s_{i2} = A\sin[\frac{2\pi(i-1)}{M}]$ for $i = 1,\ldots,M$. The symbol energy is $E_s = A^2$, so $\gamma_s = A^2/N_0$. From (**??**), for the received vector $\mathbf{x} = re^{j\theta}$ represented in polar coordinates, an error occurs if the ith signal constellation point is transmitted and $\theta \notin (2\pi(i-1-.5)/M, 2\pi(i-1+.5)/M)$. The joint distribution of r and θ can be obtained through a bivariate transformation of the noise n_1 and n_2 on the in-phase and quadrature branches [4, Chapter 5.4], which yields

$$p(r,\theta) = \frac{r}{\pi N_0}\text{exp}\left[-\frac{1}{N_0}\left(r^2 - 2\sqrt{2E_s}r\cos\theta + 2E_s\right)\right]. \tag{6.13}$$

Since the error probability depends only on the distribution of θ, we can integrate out the dependence on r, yielding

$$p(\theta) = \int_0^\infty p(r,\theta)dr = \frac{1}{\pi}e^{-2\gamma_s\sin^2(\theta)}\int_0^\infty z\text{exp}\left[\left(z - \sqrt{2\gamma_s}\cos(\theta)\right)^2\right]dz. \tag{6.14}$$

By symmetry, the probability of error is the same for each constellation point. Thus, we can obtain P_s from the probability of error assuming the constellation point $\mathbf{s}_1 = (A, 0)$ is transmitted, which is

$$P_s = 1 - \int_{-\pi/M}^{\pi/M} p(\theta)d\theta = 1 - \int_{-\pi/M}^{\pi/M}\frac{1}{\pi}e^{-2\gamma_s\sin^2(\theta)}\int_0^\infty z\text{exp}\left[-\left(z - \sqrt{2\gamma_s}\cos(\theta)\right)^2\right]dz. \tag{6.15}$$

A closed-form solution to this integral does not exist for $M > 4$, and hence the exact value of P_s must be computed numerically.

Each point in the MPSK constellation has two nearest neighbors at distance $d_{min} = 2A\sin(\pi/M)$. Thus, the nearest neighbor approximation (**??**) to P_s is given by

$$P_s \approx 2Q(\sqrt{2}A/\sqrt{N_0} \times \sin(\pi/M)) = 2Q(\sqrt{2\gamma_s}\sin(\pi/M)). \tag{6.16}$$

As shown in the prior example for QPSK, this nearest neighbor approximation can differ from the exact value of P_s by more than an order of magnitude. However, it is much simpler to compute than the numerical integration of (6.15) that is required to obtain the exact P_s. A tighter approximation for P_s can be obtained by approximating $p(\theta)$ as

$$p(\theta) \approx \sqrt{\gamma_s}\pi\cos(\theta)e^{-\gamma_s\sin^2(\theta)}. \tag{6.17}$$

Using this approximation in the left hand side of (6.15) yields

$$P_s \approx 2Q\left(\sqrt{2\gamma_s}\sin(\pi/M)\right). \tag{6.18}$$

Example 6.2:

Compare the probability of bit error for 8PSK and 16PSK assuming $\gamma_b = 15$ dB and using the P_s approximation given in (6.18) along with the approximations (6.3) and (6.2).
Solution: From (6.2)we have that for 8PSK, $\gamma_s = (\log_2 8) \cdot 10^{15/10} = 94.87$. Substituting this into (6.18) yields

$$P_s \approx 2Q\left(\sqrt{189.74}\sin(\pi/8)\right) = 1.355 \cdot 10^{-7}.$$

and using (6.3) we get $P_b = P_s/3 = 4.52 \cdot 10^{-8}$. For 16PSK we have $\gamma_s = (\log_2 16) \cdot 10^{15/10} = 126.49$. Substituting this into (6.18) yields

$$P_s \approx 2Q\left(\sqrt{252.98}\sin(\pi/16)\right) = 1.916 \cdot 10^{-3},$$

and using (6.3) we get $P_b = P_s/4 = 4.79 \cdot 10^{-4}$. Note that P_b is much larger for 16PSK than for 8PSK for the same γ_b. This result is expected, since 16PSK packs more bits per symbol into a given constellation, so for a fixed energy-per-bit the minimum distance between constellation points will be smaller.

The error probability derivation for MPSK assumes that the carrier phase is perfectly known at the receiver. Under phase estimation error, the distribution of $p(\theta)$ used to obtain P_s must incorporate the distribution of the phase rotation associated with carrier phase offset. This distribution is typically a function of the carrier phase estimation technique and the SNR. The impact of phase estimation error on coherent modulation is studied in [1, Appendix C] [2, Chapter 4.3.2][9, 10]. These works indicate that, as expected, significant phase offset leads to an irreducible bit error probability. Moreover, nonbinary signalling is more sensitive than BPSK to phase offset due to the resulting cross-coupling between the in-phase and quadrature signal components. The impact of phase estimation error can be especially severe in fast fading, where the channel phase changes rapidly due to constructive and destructive multipath interference. Even with differential modulation, phase changes over and between symbol times can produce irreducible errors [11]. Timing errors can also degrade performance: analysis of timing errors in MPSK performance can be found in [2, Chapter 4.3.3][12].

6.1.4 Error Probability for MPAM and MQAM

The constellation for MPAM is $A_i = (2i - 1 - M)d, i = 1, 2, \ldots, M$. Each of the $M - 2$ inner constellation points of this constellation have two nearest neighbors at distance $2d$. The probability of making an error when sending one of these inner constellation points is just the probability that the noise exceeds d in either direction: $P_s(\mathbf{s}_i) = p(|\mathbf{n}| > d), i = 2, \ldots, M - 1$. For the outer constellation points there is only one nearest neighbor, so an error occurs if the noise exceeds d in one direction only: $P_s(\mathbf{s}_i) = p(\mathbf{n} > d) = .5p(|\mathbf{n}| > d), i = 1, M$. The probability of error is thus

$$P_s = \frac{1}{M}\sum_{i=1}^{M} P_s(\mathbf{s}_i) = \frac{M-2}{M}2Q\left(\sqrt{\frac{2d^2}{N_0}}\right) + \frac{2}{M}Q\left(\sqrt{\frac{2d^2}{N_0}}\right) = \frac{2(M-1)}{M}Q\left(\sqrt{\frac{2d^2}{N_0}}\right). \tag{6.19}$$

From (**??**) the average energy per symbol for MPAM is

$$\overline{E}_s = \frac{1}{M}\sum_{i=1}^{M} A_i^2 = \frac{1}{M}\sum_{i=1}^{M}(2i-1-M)^2 d^2 = \frac{1}{3}(M^2-1)d^2. \tag{6.20}$$

Thus we can write P_s in terms of the average energy $\overline{E}_s$ as

$$P_s = \frac{2(M-1)}{M} Q\left(\sqrt{\frac{6\overline{\gamma}_s}{M^2-1}}\right). \tag{6.21}$$

Consider now MQAM modulation with a square signal constellation of size $M = L^2$. This system can be viewed as two MPAM systems with signal constellations of size L transmitted over the in-phase and quadrature signal components, each with half the energy of the original MQAM system. The constellation points in the in-phase and quadrature branches take values $A_i = (2i - 1 - L)d, i = 1, 2, \ldots, L$. The symbol error probability for each branch of the MQAM system is thus given by (6.21) with M replaced by $L = \sqrt{M}$ and $\overline{\gamma}_s$ equal to the average energy per symbol in the MQAM constellation:

$$P_s = \frac{2(\sqrt{M}-1)}{\sqrt{M}} Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right). \tag{6.22}$$

Note that $\overline{\gamma}_s$ is multiplied by a factor of 3 in (6.22) instead of the factor of 6 in (6.21) since the MQAM constellation splits its total average energy $\overline{\gamma}_s$ between its in-phase and quadrature branches. The probability of symbol error for the MQAM system is then

$$P_s = 1 - \left(1 - \frac{2(\sqrt{M}-1)}{\sqrt{M}} Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right)\right)^2. \tag{6.23}$$

The nearest neighbor approximation to probability of symbol error depends on whether the constellation point is an inner or outer point. If we average the nearest neighbor approximation over all inner and outer points, we obtain the MPAM probability of error associated with each branch:

$$P_s \approx \frac{2(\sqrt{M}-1)}{\sqrt{M}} Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right). \tag{6.24}$$

For nonrectangular constellations, it is relatively straightforward to show that the probability of symbol error is upper bounded as

$$P_s \leq 1 - \left[1 - 2Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right)\right]^2 \leq 4Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right). \tag{6.25}$$

The nearest neighbor approximation for nonrectangular constellations is

$$P_s \approx M_{d_{min}} Q\left(\frac{d_{min}}{\sqrt{2N_0}}\right), \tag{6.26}$$

where $M_{d_{min}}$ is the largest number of nearest neighbors for any constellation point in the constellation and d_{min} is the minimum distance in the constellation.

Example 6.3:
For 16QAM with $\gamma_b = 15$ dB ($\gamma_s = \log_2 M \cdot \gamma_b$), compare the exact probability of symbol error (6.23) with the nearest neighbor approximation (6.24), and with the symbol error probability for 16PSK with the same γ_b that was obtained in the previous example.

Solution: The average symbol energy $\gamma_s = 4 \cdot 10^{1.5} = 126.49$. The exact P_s is then given by

$$P_s = 1 - \left(1 - \frac{2(4-1)}{4} Q\left(\sqrt{\frac{3 \cdot 126.49}{15}}\right)\right)^2 = 7.37 \cdot 10^{-7}.$$

The nearest neighbor approximation is given by

$$P_s \approx \frac{2(4-1)}{4} Q\left(\sqrt{\frac{3 \cdot 126.49}{15}}\right) = 3.68 \cdot 10^{-7},$$

which differs by roughly a factor of 2 from the exact value. The symbol error probability for 16PSK in the previous example is $P_s \approx 1.916 \cdot 10^{-3}$, which is roughly four orders of magnitude larger than the exact P_s for 16QAM. The larger P_s for MPSK versus MQAM with the same M and same γ_b is due to the fact that MQAM uses both amplitude and phase to encode data, whereas MPSK uses just the phase. Thus, for the same energy per symbol or bit, MQAM makes more efficient use of energy and thus has better performance.

The MQAM demodulator requires both amplitude and phase estimates of the channel so that the decision regions used in detection to estimate the transmitted bit are not skewed in amplitude or phase. The analysis of the performance degradation due to phase estimation error is similar to the case of MPSK discussed above. The channel amplitude is used to scale the decision regions to correspond to the transmitted symbol: this scaling is called Automatic Gain Control (AGC). If the channel gain is estimated in error then the AGC improperly scales the received signal, which can lead to incorrect demodulation even in the absence of noise. The channel gain is typically obtained using pilot symbols to estimate the channel gain at the receiver. However, pilot symbols do not lead to perfect channel estimates, and the estimation error can lead to bit errors. More details on the impact of amplitude and phase estimation errors on the performance of MQAM modulation can be found in [15, Chapter 10.3][16].

6.1.5 Error Probability for FSK and CPFSK

Let us first consider the error probability of traditional binary FSK with the coherent demodulator of Figure **??**. Since demodulation is coherent, we can neglect any phase offset in the carrier signals. The transmitted signal is defined by

$$s_i(t) = A\sqrt{2}T_b \cos(2\pi f_i t), i = 1, 2. \tag{6.27}$$

So $E_b = A^2$ and $\gamma_b = A^2/N_0$. The input to the decision device is

$$\mathbf{z} = \mathbf{x}_1 - \mathbf{x}_2. \tag{6.28}$$

The device outputs a 1 bit if $\mathbf{z} > 0$ and a 0 bit if $\mathbf{z} \leq 0$. Let us assume that $s_1(t)$ is transmitted, then

$$\mathbf{z}|1 = A + n_1 - n_2. \tag{6.29}$$

An error occurs if $\mathbf{z} = A + n_1 - n_2 \leq 0$. On the other hand, if $s_2(t)$ is transmitted, then

$$\mathbf{z}|0 = n_1 - A - n_2, \tag{6.30}$$

and an error occurs if $\mathbf{z} = n_1 - A - n_2 > 0$. For n_1 and n_2 independent white Gaussian random variables with mean zero and variance $N_0/2$, their difference is a white Gaussian random variable with mean zero and variance equal to the sum of variances $N_0/2 + N_0/2 = N_0$. Then for equally likely bit transmissions,

$$P_b = .5p(A + n_1 - n_2 \leq 0) + .5p(n_1 - A - n_2 > 0) = Q(A/\sqrt{N_0}) = Q(\sqrt{\gamma_b}). \tag{6.31}$$

The derivation of P_s for coherent M-FSK with $M > 2$ is more complex and does not lead to a closed-form solution [Equation 4.92][2]. The probability of symbol error for noncoherent M-FSK is derived in [19, Chapter 8.1] as

$$P_s = \sum_{m=1}^{M} (-1)^{m+1} \left(\begin{array}{c} M-1 \\ m \end{array} \right) \frac{1}{m+1} \exp\left[\frac{-m\gamma_s}{m+1} \right]. \tag{6.32}$$

The error probability of CPFSK depends on whether the detector is coherent or noncoherent, and also whether it uses symbol-by-symbol detection or sequence estimation. Analysis of error probability for CPFSK is complex since the memory in the modulation requires error probability analysis over multiple symbols. The formulas for error probability can also become quite complex. Detailed derivations of error probability for these different CPFSK structures can be found in [1, Chapter 5.3]. As with linear modulations, FSK performance degrades under frequency and timing errors. A detailed analysis of the impact of such errors on FSK performance can be found in [2, Chapter 5.2][13, 14].

6.1.6 Error Probability Approximation for Coherent Modulations

Many of the approximations or exact values for P_s derived above for coherent modulation are in the following form:

$$P_s(\gamma_s) \approx \alpha_M \, Q\left(\sqrt{\beta_M \gamma_s}\right), \tag{6.33}$$

where α_M and β_M depend on the type of approximation and the modulation type. In particular, the nearest neighbor approximation has this form, where α_M is the number of nearest neighbors to a constellation at the minimum distance, and β_M is a constant that relates minimum distance to average symbol energy. In Table 6.1 we summarize the specific values of α_M and β_M for common P_s expressions for PSK, QAM, and FSK modulations based on the derivations in the prior sections.

Performance specifications are generally more concerned with the bit error probability P_b as a function of the bit energy γ_b. To convert from P_s to P_b and from γ_s to γ_b, we use the approximations (6.3) and (6.2), which assume Gray encoding and high SNR. Using these approximations in (6.33) yields a simple formula for P_b as a function of γ_b:

$$P_b(\gamma_b) = \hat{\alpha}_M \, Q\left(\sqrt{\hat{\beta}_M \gamma_b}\right), \tag{6.34}$$

where $\hat{\alpha}_M = \alpha_M / \log_2 M$ and $\hat{\beta}_M = (\log_2 M)\beta_M$ for α_M and β_M in (6.33). This conversion is used below to obtain P_b versus γ_b from the general form of P_s versus γ_s in (6.33).

6.1.7 Error Probability for Differential Modulation

The probability of error for differential modulation is based on the phase difference associated with the phase comparator input of Figure **??**. Specifically, the phase comparator extracts the phase of

$$\mathbf{z}(k)\mathbf{z}^*(k-1) = A^2 e^{j(\theta(k)-\theta(k-1))} + Ae^{j(\theta(k)+\phi_0)} n^*(k-1) + Ae^{-j(\theta(k-1)+\phi_0)} n(k) + n(k)n^*(k-1) \tag{6.35}$$

Modulation	$P_s(\gamma_s)$	$P_b(\gamma_b)$
BFSK:		$P_b = Q\left(\sqrt{\gamma_b}\right)$
BPSK:		$P_b = Q\left(\sqrt{2\gamma_b}\right)$
QPSK,4QAM:	$P_s \approx 2\,Q\left(\sqrt{\gamma_s}\right)$	$P_b \approx Q\left(\sqrt{2\gamma_b}\right)$
MPAM:	$P_s \approx \frac{2(M-1)}{M} Q\left(\sqrt{\frac{6\overline{\gamma}_s}{M^2-1}}\right)$	$P_b \approx \frac{2(M-1)}{M\log_2 M} Q\left(\sqrt{\frac{6\overline{\gamma}_b \log_2 M}{(M^2-1)}}\right)$
MPSK:	$P_s \approx 2Q\left(\sqrt{2\gamma_s}\sin(\pi/M)\right)$	$P_b \approx \frac{2}{\log_2 M} Q\left(\sqrt{2\gamma_b \log_2 M}\sin(\pi/M)\right)$
Rectangular MQAM:	$P_s \approx \frac{4(\sqrt{M}-1)}{\sqrt{M}} Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right)$	$P_b \approx \frac{4(\sqrt{M}-1)}{\sqrt{M}\log_2 M} Q\left(\sqrt{\frac{3\overline{\gamma}_b \log_2 M}{(M-1)}}\right)$
Nonrectangular MQAM:	$P_s \approx 4Q\left(\sqrt{\frac{3\overline{\gamma}_s}{M-1}}\right)$	$P_b \approx \frac{4}{\log_2 M} Q\left(\sqrt{\frac{3\overline{\gamma}_b \log_2 M}{(M-1)}}\right)$

Table 6.1: Approximate Symbol and Bit Error Probabilities for Coherent Modulations

to determine the transmitted symbol. Due to symmetry, we can assume a given phase difference to compute the error probability. Assuming a phase difference of zero, $\theta(k) - \theta(k-1) = 0$, yields

$$\mathbf{z}(k)\mathbf{z}^*(k-1) = A^2 + Ae^{j(\theta(k)+\phi_0)}n^*(k-1) + Ae^{-j(\theta(k-1)+\phi_0)}n(k) + n(k)n^*(k-1). \tag{6.36}$$

Next we define new random variables $\tilde{n}(k) = n(k)e^{-j(\theta(k-1)+\phi_0)}$ and $\tilde{n}(k-1) = n(k-1)e^{-j(\theta(k)+\phi_0)}$, which have the same statistics as $n(k)$ and $n(k-1)$. Then we have

$$\mathbf{z}(k)\mathbf{z}^*(k-1) = A^2 + A(\tilde{n}^*(k-1) + \tilde{n}(k)) + \tilde{n}(k)\tilde{n}^*(k-1). \tag{6.37}$$

There are three terms in (6.37): the first term with the desired phase difference of zero, and the second and third terms, which contribute noise. At reasonable SNRs the third noise term is much smaller than the second, so we neglect it. Dividing the remaining terms by A yields

$$\tilde{z} = A + \Re\{\tilde{n}^*(k-1) + \tilde{n}(k)\} + j\Im\{\tilde{n}^*(k-1) + \tilde{n}(k)\}. \tag{6.38}$$

Let us define $x = \Re\{\tilde{z}\}$ and $y = \Im\{\tilde{z}\}$. The phase of $\tilde{z}$ is thus given by

$$\theta_{\tilde{z}} = \tan^{-1}\frac{y}{x}. \tag{6.39}$$

Given that the phase difference was zero, and error occurs if $|\theta_{\tilde{z}}| \geq \pi/M$. Determining $p(|\theta_{\tilde{z}}| \geq \pi/M)$ is identical to the case of coherent PSK, except that from (6.38) we see that we have two noise terms instead of one, and therefore the noise power is twice that of the coherent case. This will lead to a performance of differential modulation that is roughly 3 dB worse than that of coherent modulation.

In DPSK modulation we need only consider the in-phase branch of Figure **??** to make a decision, so we set $x = \Re\{\tilde{z}\}$ in our analysis. In particular, assuming a zero is transmitted, if $x = A + \Re\{\tilde{n}^*(k-1) + \tilde{n}(k)\} < 0$ then a decision error is made. This probability can be obtained by finding the characteristic or moment-generating function for x, taking the inverse Laplace transform to get the distribution of x, and then integrating over the decision region $x < 0$. This technique is very general and can be applied to a wide variety of different modulation and detection types in both AWGN and fading [19, Chapter 1.1]: we will use it later to compute the average probability of symbol error for linear modulations in fading both with and without diversity. In DPSK the characteristic function for x is obtained using the general quadratic form of complex Gaussian random variables [1, Appendix B][18, Appendix B], and the resulting bit error probability is given by

$$P_b = \frac{1}{2}e^{-\gamma_b}. \tag{6.40}$$

For DQPSK the characteristic function for $\tilde{z}$ is obtained in [1, Appendix C], which yields the bit error probability

$$P_b \approx \int_b^\infty x \exp\left(\frac{-(a^2+x^2)}{2}\right) \mathrm{I}_0(ax)dx - \frac{1}{2}\exp\left(\frac{-(a^2+b^2)}{2}\right) \mathrm{I}_0(ab), \tag{6.41}$$

where $a \approx .765\sqrt{\gamma_b}$ and $b \approx 1.85\sqrt{\gamma_b}$.

6.2 Alternate Q Function Representation

In (6.33) we saw that P_s for many coherent modulation techniques in AWGN is approximated in terms of the Gaussian Q function. Recall that $Q(z)$ is defined as the probability that a Gaussian random variable x with mean zero and variance one exceeds the value z, i.e.

$$Q(z) = p(x \geq z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \tag{6.42}$$

The Q function is not that easy to work with since the argument z is in the lower limit of the integrand, the integrand has infinite range, and the exponential function in the integral doesn't lead to a closed form solution.

In 1991 an alternate representation of the Q function was obtained by Craig [5]. The alternate form is given by

$$Q(z) = \frac{1}{\pi}\int_0^{\pi/2} \exp\left[\frac{-z^2}{2\sin^2\phi}\right] d\phi \;\; z > 0. \tag{6.43}$$

This representation can also be deduced from the work of Weinstein [6] or Pawula *et al.* [7]. Note that in this alternate form, the integrand is over a finite range that is independent of the function argument z, and the integral is Gaussian with respect to z. These features will prove important in using the alternate representation to derive average error probability in fading.

Craig's motivation for deriving the alternate representation was to simplify the probability of error calculation for AWGN channels. In particular, we can write the probability of bit error for BPSK using the alternate form as

$$P_b = Q(\sqrt{2\gamma_b}) = \frac{1}{\pi}\int_0^{\pi/2} \exp\left[\frac{-\gamma_b}{\sin^2\phi}\right] d\phi. \tag{6.44}$$

Similarly, the alternate representation can be used to obtain a simple *exact* formula for P_s of MPSK in AWGN as [5]

$$P_s = \frac{1}{\pi}\int_0^{(M-1)\pi/M} \exp\left[\frac{-g_{psk}\gamma_s}{\sin^2\phi}\right] d\phi, \tag{6.45}$$

where $g_{psk} = \sin^2(\pi/M)$. Note that this formula does not correspond to the general form $\alpha_M Q(\sqrt{\beta_M\gamma_s})$, since the general form is an approximation while (6.45) is exact. Note also that (6.45) is obtained via a finite range integral of simple trigonometric functions that is easily computed via a numerical computer package or calculator.

6.3 Fading

In AWGN the probability of symbol error depends on the received SNR or, equivalently, on γ_s. In a fading environment the received signal power varies randomly over distance or time due to shadowing and/or multipath fading. Thus, in fading γ_s is a random variables with distribution $p_{\gamma_s}(\gamma)$, and therefore $P_s(\gamma_s)$ is also random. The performance metric when γ_s is random depends on the rate of change of the fading. There are three different performance criteria that can be used to characterize the random variable P_s:

- The outage probability, P_{out}, defined as the probability that γ_s falls below a given value corresponding to the maximum allowable P_s.
- The average error probability, $\overline{P_s}$, averaged over the distribution of γ_s.
- Combined average error probability and outage, defined as the average error probability that can be achieved some percentage of time or some percentage of spatial locations.

The average probability of symbol error applies when the signal fading is on the order of a symbol time ($T_s \approx T_c$), so that the signal fade level is constant over roughly one symbol time. Since many error correction coding techniques can recover from a few bit errors, and end-to-end performance is typically not seriously degraded by a few simultaneous bit errors, the average error probability is a reasonably good figure of merit for the channel quality under these conditions.

However, if the signal power is changing slowly ($T_s << T_c$), then a deep fade will affect many simultaneous symbols. Thus, fading may lead to large error bursts, which cannot be corrected for with coding of reasonable complexity. Therefore, these error bursts can seriously degrade end-to-end performance. In this case acceptable performance cannot be guaranteed over all time or, equivalently, throughout a cell, without drastically increasing transmit power. Under these circumstances, an outage probability is specified so that the channel is deemed unusable for some fraction of time or space. Outage and average error probability are often combined when the channel is modeled as a combination of fast and slow fading, e.g. log-normal shadowing with fast Rayleigh fading.

Note that when $T_c << T_s$, the fading will be averaged out by the matched filter in the demodulator. Thus, for very fast fading, performance is the same as in AWGN.

6.3.1 Outage Probability

The outage probability relative to γ_0 is defined as

$$P_{out} = p(\gamma_s < \gamma_0) = \int_0^{\gamma_0} p_{\gamma_s}(\gamma)d\gamma, \tag{6.46}$$

where γ_0 typically specifies the minimum SNR required for acceptable performance. For example, if we consider digitized voice, $P_b = 10^{-3}$ is an acceptable error rate since it generally can't be detected by the human ear. Thus, for a BPSK signal in Rayleigh fading, $\gamma_b < 7$ dB would be declared an outage, so we set $\gamma_0 = 7$ dB.

In Rayleigh fading the outage probability becomes

$$P_{out} = \int_0^{\gamma_0} \frac{1}{\overline{\gamma}_s} e^{-\gamma_s/\overline{\gamma}_s} d\gamma_s = 1 - e^{-\gamma_0/\overline{\gamma}_s}. \tag{6.47}$$

Inverting this formula shows that for a given outage probability, the required average SNR $\overline{\gamma}_s$ is

$$\overline{\gamma}_s = \frac{\gamma_0}{-\ln(1 - P_{out})}. \tag{6.48}$$

In dB this means that $10\log\gamma_s$ must exceed the target $10\log\gamma_0$ by $F_d = -10\log[-\ln(1 - P_{out})]$ to maintain acceptable performance more than $100 * (1 - P_{out})$ percent of the time. The quantity F_d is typically called the **dB fade margin**.

Example 6.4: Determine the required $\overline{\gamma}_b$ for BPSK modulation in slow Rayleigh fading such that 95% of the time (or in space), $P_b(\gamma_b) < 10^{-4}$.

Solution: For BPSK modulation in AWGN the target BER is obtained at 8.5 dB, i.e. for $P_b(\gamma_b) = Q(\sqrt{2\gamma_b})$, $P_b(10^{.85}) = 10^{-4}$. Thus, $\gamma_0 = 8.5$ dB. Since we want $P_{out} = p(\gamma_b < \gamma_0) = .05$ we have

$$\overline{\gamma}_b = \frac{\gamma_0}{-\ln(1-P_{out})} = \frac{10^{.85}}{-\ln(1-.05)} = 21.4\,\text{dB}. \tag{6.49}$$

6.3.2 Average Probability of Error

The average probability of error is used as a performance metric when $T_s \approx T_c$. Thus, we can assume that γ_s is roughly constant over a symbol time. Then the averaged probability of error is computed by integrating the error probability in AWGN over the fading distribution:

$$\overline{P}_s = \int_0^\infty P_s(\gamma)p_{\gamma_s}(\gamma)d\gamma, \tag{6.50}$$

where $P_s(\gamma)$ is the probability of symbol error in AWGN with SNR γ, which can be approximated by the expressions in Table 6.1. For a given distribution of the fading amplitude r (i.e. Rayleigh, Rician, log-normal, etc.), we compute $p_{\gamma_s}(\gamma)$ by making the change of variable

$$p_{\gamma_s}(\gamma)d\gamma = p(r)dr. \tag{6.51}$$

For example, in Rayleigh fading the received signal amplitude r has the Rayleigh distribution

$$p(r) = \frac{r}{\sigma^2}e^{-r^2/2\sigma^2}, \quad r \geq 0, \tag{6.52}$$

and the signal power is exponentially distributed with mean $2\sigma^2$. The SNR per symbol for a given amplitude r is

$$\gamma = \frac{r^2T_s}{2\sigma_n^2}, \tag{6.53}$$

where $\sigma_n^2 = N_0/2$ is the PSD of the noise in the in-phase and quadrature branches. Differentiating both sides of this expression yields

$$d\gamma = \frac{rT_s}{\sigma_n^2}dr. \tag{6.54}$$

Substituting (6.53) and (6.54) into (6.52) and then (6.51) yields

$$p_{\gamma_s}(\gamma) = \frac{\sigma_n^2}{\sigma^2T_s}e^{-\gamma\sigma_n^2/\sigma^2T_s}. \tag{6.55}$$

Since the average SNR per symbol $\overline{\gamma}_s$ is just σ^2T_s/σ_n^2, we can rewrite (6.55) as

$$p_{\gamma_s}(\gamma) = \frac{1}{\overline{\gamma}_s}e^{-\gamma/\overline{\gamma}_s}, \tag{6.56}$$

which is exponential. For binary signaling this reduces to

$$p_{\gamma_b}(\gamma) = \frac{1}{\overline{\gamma}_b}e^{-\gamma/\overline{\gamma}_b}, \tag{6.57}$$

Integrating (6.6) over the distribution (6.57) yields the following average probability of error for BPSK in Rayleigh fading.

$$\text{BPSK:} \qquad \overline{P}_b = \frac{1}{2}\left[1 - \sqrt{\frac{\overline{\gamma}_b}{1+\overline{\gamma}_b}}\right] \approx \frac{1}{4\overline{\gamma}_b}, \tag{6.58}$$

where the approximation holds for large $\overline{\gamma}_b$. A similar integration of (6.31) over (6.57) yields the average probability of error for binary FSK in Rayleigh fading as

$$\text{Binary FSK:} \qquad \overline{P}_b = \frac{1}{2}\left[1 - \sqrt{\frac{\overline{\gamma}_b}{2+\overline{\gamma}_b}}\right] \approx \frac{1}{4\overline{\gamma}_b}. \tag{6.59}$$

Thus, the performance of BPSK and binary FSK converge at high SNRs. For noncoherent modulation, if we assume the channel phase is relatively constant over a symbol time, then we obtain the probability of error by again integrating the error probability in AWGN over the fading distribution. For DPSK this yields

$$\text{DPSK:} \qquad \overline{P}_b = \frac{1}{2(1+\overline{\gamma}_b)} \approx \frac{1}{2\overline{\gamma}_b}, \tag{6.60}$$

where again the approximation holds for large $\overline{\gamma}_b$. Note that in the limit of large $\overline{\gamma}_b$, there is an approximate 3 dB power penalty in using DPSK instead of BPSK. This was also observed in AWGN, and is the power penalty of differential detection. In practice the power penalty is somewhat smaller, since DPSK can correct for slow phase changes introduced in the channel or receiver, which are not taken into account in these error calculations.

If we use the general approximation $P_s \approx \alpha_M Q(\sqrt{\beta_M \gamma_s})$ then the average probability of symbol error in Rayleigh fading can be approximated as

$$\overline{P}_s \approx \int_0^\infty \alpha_M Q(\sqrt{\beta_M \gamma}) \cdot \frac{1}{\overline{\gamma}_s} e^{-\gamma/\overline{\gamma}_s} d\gamma_s. = \frac{\alpha_m}{2}\left[1 - \sqrt{\frac{.5\beta_M\overline{\gamma}_s}{1+.5\beta_M\overline{\gamma}_s}}\right] \approx \frac{\alpha_M}{2\beta_M\overline{\gamma}_s}, \tag{6.61}$$

where the last approximation is in the limit of high SNR.

It is interesting to compare bit error probability of the different modulation schemes in AWGN and fading. For binary PSK, FSK, and DPSK, the bit error probability in AWGN decreases exponentially with increasing γ_b. However, in fading the bit error probability for all the modulation types decreases just linearly with increasing $\overline{\gamma}_b$. Similar behavior occurs for nonbinary modulation. Thus, the power necessary to maintain a given P_b, particularly for small values, is much higher in fading channels than in AWGN channels. For example, in Figure 6.1 we plot the error probability of BPSK in AWGN and in flat Rayleigh fading. We see that it requires approximately 8 dB SNR to maintain a 10^{-3} bit error rate in AWGN while it requires approximately 24 dB SNR to maintain the same error rate in fading. A similar plot for the error probabilities of MQAM, based on the approximations (6.24) and (6.61), is shown in Figure 6.2. From these figures it is clear that to maintain low power requires some technique to remove the effects of fading. We will discuss some of these techniques, including diversity combining, spread spectrum, and RAKE receivers, in later chapters.

Rayleigh fading is one of the worst-case fading scenarios. In Figure 6.3 we show the average bit error probability of BPSK in Nakagami fading for different values of the Nakagami-m parameter. We see that as m increases, the fading decreases, and the average bit error probability converges to that of an AWGN channel.

6.3.3 Moment Generating Function Approach to Average Error Probability

The **moment generating function** (MGF) for a nonnegative random variable γ with pdf $p_\gamma(\gamma), \gamma \geq 0$, is defined as

$$\mathcal{M}_\gamma(s) = \int_0^\infty p_\gamma(\gamma) e^{s\gamma} d\gamma. \tag{6.62}$$

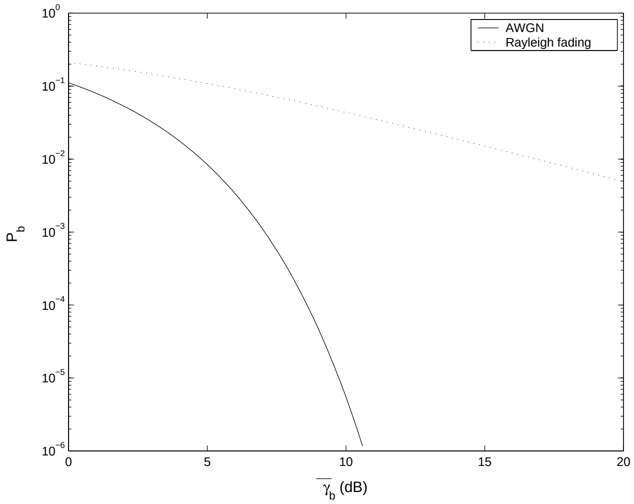


Figure 6.1: Average P_b for BPSK in Rayleigh Fading and AWGN.

Note that this function is just the Laplace transform of the pdf $p_\gamma(\gamma)$ with the argument reversed in sign: $\mathcal{L}[p_\gamma(\gamma)] = \mathcal{M}_\gamma(-s)$. Thus, the MGF for most fading distributions of interest can be computed either in closed-form using classical Laplace transforms or through numerical integration. In particular, the MGF for common multipath fading distributions are as follows [19, Chapter 5.1].

Rayleigh:

$$\mathcal{M}_{\gamma_s}(s) = (1 - s\overline{\gamma}_s)^{-1}. \tag{6.63}$$

Ricean with factor K:

$$\mathcal{M}_{\gamma_s}(s) = \frac{1+K}{1+K-s\overline{\gamma}_s} \exp\left[\frac{Ks\overline{\gamma}_s}{1+K-s\overline{\gamma}_s}\right]. \tag{6.64}$$

Nakagami-m:

$$\mathcal{M}_{\gamma_s}(s) = \left(1 - \frac{s\overline{\gamma}_s}{m}\right)^{-m}. \tag{6.65}$$

As indicated by its name, the moments $E[\gamma^n]$ of γ can be obtained from $\mathcal{M}_\gamma(s)$ as

$$E[\gamma^n] = \frac{\partial^n}{\partial s^n} \left[\mathcal{M}_{\gamma_s}(s)\right]\big|_{s=0}. \tag{6.66}$$

The MGF is a very useful tool in performance analysis of modulation in fading both with and without diversity. In this section we discuss how it can be used to simplify performance analysis of average probability of symbol error in fading. In the next chapter we will see that it also greatly simplifies analysis in fading channels with diversity.

The basic premise of the MGF approach for computing average error probability in fading is to express the probability of error P_s in AWGN for the modulation of interest either as an exponential function of γ_s,

$$P_s = c_1 \exp[-c_2\gamma_s] \tag{6.67}$$

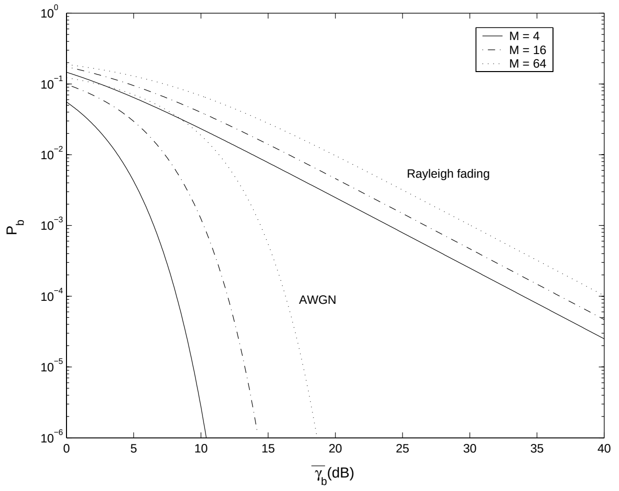


Figure 6.2: Average P_b for MQAM in Rayleigh Fading and AWGN.

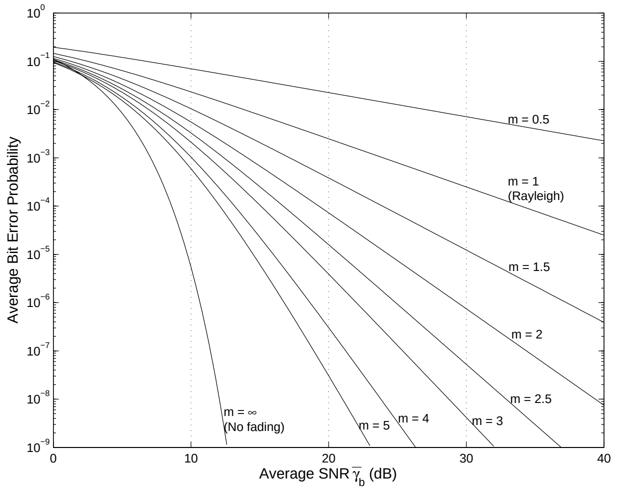


Figure 6.3: Average P_b for BPSK in Nakagami Fading.

for constants c_1 and c_2, or as a finite range integral of such an exponential function:

$$P_s = \int_A^B c_1 \exp[-c_2(x)\gamma]dx, \tag{6.68}$$

where the constant $c_2(x)$ may depend on the integrand but the SNR γ does not and is not in the limits of integration either. These forms allow the average probability of error to be expressed in terms of the MGF for the fading distribution. Specifically, if $P_s = \alpha \exp[-\beta\gamma_s]$, then

$$\overline{P}_s = \int_0^\infty c_1 \exp[-c_2\gamma] p_{\gamma_s}(\gamma) d\gamma = c_1 \mathcal{M}_{\gamma_s}(-c_2). \tag{6.69}$$

Since DPSK is in this form with $c_1 = 1/2$ and $c_2 = 1$, we see that the average probability of bit error for DPSK in any type of fading is

$$\overline{P}_b = \frac{1}{2}\mathcal{M}_{\gamma_s}(-1), \tag{6.70}$$

where $\mathcal{M}_{\gamma_s}(s)$ is the MGF of the fading distribution. For example, using $\mathcal{M}_{\gamma_s}(s)$ for Rayleigh fading given by (6.63) with $s = -1$ yields $\overline{P}_b = [2(1+\overline{\gamma}_b)]^{-1}$, which is the same as we obtained in (6.60). If P_s is in the integral form of (6.68) then

$$\overline{P}_s = \int_0^\infty \int_A^B c_1 \exp[-c_2(x)\gamma] dx p_{\gamma_s}(\gamma) d\gamma = c_1 \int_A^B \left[\int_0^\infty \exp[-c_2(x)\gamma] p_{\gamma_s}(\gamma) d\gamma\right] dx = c_1 \int_A^B \mathcal{M}_{\gamma_s}(-c_2(x)) dx. \tag{6.71}$$

In this latter case, the average probability of symbol error is a single finite-range integral of the MGF of the fading distribution, which can typically be found in closed form or easily evaluated numerically.

Let us now apply the MGF approach to specific modulations and fading distributions. In (6.33) we gave a general expression for P_s of coherent modulation in AWGN in terms of the Gaussian Q function. We now make a slight change of notation in (6.33) setting $\alpha = \alpha_M$ and $g = .5\beta_M$ to get

$$P_s(\gamma_s) = \alpha Q(\sqrt{2g\gamma_s}), \tag{6.72}$$

where α and g are constants that depend on the modulation. The notation change is to obtain the error probability as an exact MGF, as we now show.

Using the alternate Q function representation (6.43), we get that

$$P_s = \frac{\alpha}{\pi} \int_0^{\pi/2} \exp\left[\frac{-g\gamma}{\sin^2\phi}\right] d\phi, \tag{6.73}$$

which is in the desired form (6.68). Thus, the average error probability in fading for modulations with $P_s = \alpha Q(\sqrt{2g\gamma_s})$ in AWGN is given by

$$\begin{aligned}\overline{P}_s &= \frac{\alpha}{\pi} \int_0^\infty \int_0^{\pi/2} \exp\left[\frac{-g\gamma}{\sin^2\phi}\right] d\phi p_{\gamma_s}(\gamma) d\gamma \\ &= \frac{\alpha}{\pi} \int_0^{\pi/2} \left[\int_0^\infty \exp\left[\frac{-g\gamma}{\sin^2\phi}\right] p_{\gamma_s}(\gamma) d\gamma\right] d\phi = \frac{\alpha}{\pi} \int_0^{\pi/2} \mathcal{M}_{\gamma_s}\left(\frac{-g}{\sin^2\phi}\right) d\phi,\end{aligned} \tag{6.74}$$

where $\mathcal{M}_{\gamma_s}(s)$ is the MGF associated with the pdf $p_{\gamma_s}(\gamma)$ as defined by (6.62). Recall that Table 6.1 approximates the error probability in AWGN for many modulations of interest as $P_s \approx \alpha Q(\sqrt{2g\gamma_s})$, and thus (6.74) gives an approximation for the average error probability of these modulations in fading. Moreover, the exact average probability of symbol error for coherent MPSK can be obtained in a form similar to (6.74) by noting that Craig's formula for P_s of MPSK in AWGN given by (6.45) is in the desired form (6.68). Thus, the exact average probability of error for MPSK becomes

$$\begin{aligned}\overline{P}_s &= \int_0^\infty \frac{1}{\pi} \int_0^{(M-1)\pi/M} \exp\left[\frac{-g\gamma_s}{\sin^2\phi}\right] d\phi p_{\gamma_s}(\gamma) d\gamma \\ &= \frac{1}{\pi} \int_0^{\frac{(M-1)\pi}{M}} \left[\int_0^\infty \exp\left[\frac{-g\gamma_s}{\sin^2\phi}\right] p_{\gamma_s}(\gamma) d\gamma\right] d\phi = \frac{1}{\pi} \int_0^{\frac{(M-1)\pi}{M}} \mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right) d\phi,\end{aligned} \tag{6.75}$$

where $g = \sin^2(\pi/M)$ depends on the size of the MPSK constellation. The MGF $\mathcal{M}_{\gamma_s}(s)$ for Rayleigh, Rician, and Nakagami-m distributions were given, respectively, by (6.63), (6.64), and (6.65) above. Substituting $s = -g/\sin^2\phi$ in these expressions yields

Rayleigh:

$$\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right) = \left(1 + \frac{g\,\overline{\gamma}_s}{\sin^2\phi}\right)^{-1}. \tag{6.76}$$

Ricean with factor K:

$$\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right)=\frac{(1+K)\ \sin^2\phi}{(1+K)\sin^2\phi+g\,\overline{\gamma}_s}\ \exp\left(-\frac{K\,g\,\overline{\gamma}_s}{(1+K)\sin^2\phi+g\,\overline{\gamma}_s}\right). \tag{6.77}$$

Nakagami-m:

$$\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right)=\left(1+\frac{g\,\overline{\gamma}_s}{m\ \sin^2\phi}\right)^{-m}. \tag{6.78}$$

All of these functions are simple trigonometrics and are therefore easy to integrate over the finite range in (6.74) or (6.75).

Example 6.5: Use the MGF technique to find an expression for the average probability of error for BPSK modulation in Nakagami fading.

Solution: We use the fact that for an AWGN channel BPSK has $P_b=Q(\sqrt{2\gamma_b})$, so $\alpha=1$ and $g=1$ in (6.72). The moment generating function for Nakagami-m fading is given by (6.78), and substituting this into (6.74) with $\alpha=g=1$ yields

$$\overline{P}_b=\frac{1}{\pi}\int_0^{\pi/2}\left(1+\frac{\overline{\gamma}_b}{m\sin^2\phi}\right)^{-m}d\phi.$$

From (6.23) we see that the exact probability of symbol error for MQAM in AWGN contains both the Q function and its square. Fortunately, an alternate form of $Q^2(z)$ derived in [8] allows us to apply the same techniques used above for MPSK to MQAM modulation. Specifically, an alternate representation of $Q^2(z)$ is derived in [8] as

$$Q^2(z)=\frac{1}{\pi}\int_0^{\pi/4}\exp\left[\frac{-z^2}{2\sin^2\phi}\right]d\phi. \tag{6.79}$$

Note that this is identical to the alternate representation for $Q(z)$ given in (6.43) except that the upper limit of the integral is $\pi/4$ instead of $\pi/2$. Thus we can write (6.23) in terms of the alternate representations for $Q(z)$ and $Q^2(z)$ as

$$P_s(\gamma_s)=\frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\exp\left(-\frac{g\gamma_s}{\sin^2\phi}\right)d\phi-\frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\exp\left(-\frac{g\gamma_s}{\sin^2\phi}\right)d\phi, \tag{6.80}$$

where $g=1.5/(M-1)$ is a function of the size of the MQAM constellation. Then the average probability of symbol error in fading becomes

$$\begin{aligned}\overline{P}_s&=\int_0^\infty P_s(\gamma)p_{\gamma_s}(\gamma)d\gamma\\&=\frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\int_0^\infty\exp\left(-\frac{g\gamma}{\sin^2\phi}\right)p_{\gamma_s}(\gamma)d\gamma d\phi-\frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\int_0^\infty\exp\left(-\frac{g\gamma}{\sin^2\phi}\right)p_{\gamma_s}(\gamma)d\gamma d\phi\\&=\frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)\int_0^{\pi/2}\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right)d\phi-\frac{4}{\pi}\left(1-\frac{1}{\sqrt{M}}\right)^2\int_0^{\pi/4}\mathcal{M}_{\gamma_s}\left(-\frac{g}{\sin^2\phi}\right)d\phi. \qquad (6.81)\end{aligned}$$

Thus, the exact average probability of symbol error is obtained via two finite-range integrals of the MGF of the fading distribution, which can typically be found in closed form or easily evaluated numerically.

The MGF approach can also be applied to noncoherent and differential modulations. For example, consider noncoherent M-FSK, with P_s in AWGN given by (6.32), which is a finite sum of the desired form (6.67). Thus, in fading, the average symbol error probability of noncoherent M-FSK is given by

$$\begin{aligned}\overline{P}_s &= \int_0^\infty \sum_{m=1}^{M} (-1)^{m+1} \left(\begin{array}{c} M-1 \\ m \end{array} \right) \frac{1}{m+1} \exp\left[\frac{-m\gamma}{m+1}\right] p_{\gamma_s}(\gamma) d\gamma \\ &= \sum_{m=1}^{M} (-1)^{m+1} \left(\begin{array}{c} M-1 \\ m \end{array} \right) \frac{1}{m+1} \left[\int_0^\infty \exp\left[\frac{-m\gamma}{m+1}\right] p_{\gamma_s}(\gamma) d\gamma\right] \\ &= \sum_{m=1}^{M} (-1)^{m+1} \left(\begin{array}{c} M-1 \\ m \end{array} \right) \frac{1}{m+1} \mathcal{M}_{\gamma_s}\left(-\frac{m}{m+1}\right). \qquad (6.82)\end{aligned}$$

Finally, for differential MPSK, it can be shown [11] that the average probability of symbol error is given by

$$P_s = \frac{\sqrt{g_{psk}}}{2\pi} \int_{-\pi/2}^{\pi/2} \frac{\exp[-\gamma_s(1-\sqrt{1-g_{psk}}\cos\theta)]}{1-\sqrt{1-g_{psk}}\cos\theta} d\theta \qquad (6.83)$$

for $g_{psk} = \sin^2(\pi/M)$, which is in the desired form (6.68). Thus we can express the average probability of symbol error in terms of the MGF of the fading distribution as

$$\overline{P}_s = \frac{\sqrt{g_{psk}}}{2\pi} \int_{-\pi/2}^{\pi/2} \frac{\mathcal{M}_{\gamma_s}\left(-(1-\sqrt{1-g_{psk}}\cos\theta)\right)}{1-\sqrt{1-g_{psk}}\cos\theta} d\theta. \qquad (6.84)$$

A more extensive discussion of the MGF technique for finding average probability of symbol error for different modulations and fading distributions can be found in [19, Chapter 8.2].

6.3.4 Combined Outage and Average Error Probability

When the fading environment is a superposition of both fast and slow fading, i.e. log-normal shadowing and Rayleigh fading, a common performance metric is combined outage and average error probability, where outage occurs when the slow fading falls below some target value and the average performance in nonoutage is obtained by averaging over the fast fading. We use the following notation:

- Let $\overline{\overline{\gamma}}_s$ denote the average SNR per symbol for a fixed path loss with averaging over fast fading and shadowing.
- Let $\overline{\gamma}_s$ denote the (random) SNR per symbol for a fixed path loss and random shadowing but averaged over fast fading. Its average value is $\overline{\overline{\gamma}}_s$.
- Let γ_s denote the random SNR due to fixed path loss, shadowing, and multipath.

With this notation we can specify an average error probability $\overline{P}_s$ with some probability $1 - P_{out}$. An outage is declared when the received SNR per symbol due to shadowing and path loss alone, $\overline{\gamma}_s$, falls below a given target value $\overline{\gamma}_{s_0}$. When not in outage ($\overline{\gamma}_s \geq \overline{\gamma}_{s_0}$), the average probability of error is obtained by averaging over the distribution of the fast fading conditioned on the mean SNR:

$$\overline{P}_s = \int_0^\infty P_s(\gamma_s) p(\gamma_s|\overline{\gamma}_s) d\gamma_s. \qquad (6.85)$$

The criterion used to determine the outage target $\overline{\gamma}_{s_0}$ is typically based on a given maximum average probability of error, i.e. $\overline{P}_s \leq \overline{P}_{s_0}$, where the target $\overline{\gamma}_{s_0}$ must then satisfy

$$\overline{P}_{s_0} = \int_0^\infty P_s(\gamma_s) p(\gamma_s|\overline{\gamma}_{s_0}) d\gamma_s. \tag{6.86}$$

Clearly whenever $\overline{\gamma}_s > \overline{\gamma}_{s_0}$, the average error probability will be below the target value.

Example 6.6:

Consider BPSK modulation in a channel with both log-normal shadowing ($\sigma = 8$ dB) and Rayleigh fading. The desired maximum average error probability is $\overline{P}_{b_0} = 10^{-4}$, which requires $\overline{\gamma}_{b_0} = 34$ dB. Determine the value of $\overline{\overline{\gamma}}_b$ that will insure $\overline{P}_b \leq 10^{-4}$ with probability $1 - P_{out} = .95$.

Solution: We must find $\overline{\overline{\gamma_b}}$, the average of γ_b in both the fast and slow fading, such that $p(\overline{\gamma_b} > \gamma_{b_0}) = 1 - P_{out}$. For log-normal shadowing we compute this as:

$$p(\overline{\gamma_b} > 34) = p\left(\frac{\overline{\gamma_b} - \overline{\overline{\gamma_b}}}{\sigma} \geq \frac{34 - \overline{\overline{\gamma_b}}}{\sigma}\right) = Q\left(\frac{34 - \overline{\overline{\gamma_b}}}{\sigma}\right) = 1 - P_{out}, \tag{6.87}$$

since $(\overline{\gamma_b} - \overline{\overline{\gamma_b}})/\sigma$ is a Gauss-distributed random variable with mean zero and standard deviation one. Thus, the value of $\overline{\overline{\gamma_b}}$ is obtained by substituting the values of P_{out} and σ in (6.87) and using a table of Q functions or an inversion program, which yields $(34 - \overline{\overline{\gamma_b}})/8 = -1.6$ or $\overline{\overline{\gamma}}_b = 46.8$ dB.

6.4 Doppler Spread

Doppler spread results in an irreducible error floor for modulation techniques using differential detection. This is due to the fact that in differential modulation the signal phase associated with one symbol is used as a phase reference for the next symbol. If the channel phase decorrelates over a symbol, then the phase reference becomes extremely noisy, leading to a high symbol error rate that is independent of received signal power. The phase correlation between symbols and therefore the degradation in performance are functions of the Doppler frequency $f_D = v/\lambda$ and the symbol time T_s.

The first analysis of the irreducible error floor due to Doppler was done by Bello and Nelin in [17]. In that work analytical expressions for the irreducible error floor of noncoherent FSK and DPSK due to Doppler are determined for a Gaussian Doppler power spectrum. However, these expressions are not in closed-form, so must be evaluated numerically. Closed-form expressions for the bit error probability of DPSK in fast Rician fading, where the channel decorrelates over a bit time, can be obtained using the MGF technique, with the MGF obtained based on the general quadratic form of complex Gaussian random variables [18, Appendix B] [1, Appendix B]. A different approach utilizing alternate forms of the Marcum Q function can also be used [19, Chapter 8.2.5]. The resulting average bit error probability for DPSK is

$$\overline{P}_b = \frac{1}{2}\left[\frac{1 + K + \overline{\gamma}_b(1 - \rho_C)}{1 + K + \overline{\gamma}_b}\right] \exp\left(-\frac{K\overline{\gamma}_b}{1 + K + \overline{\gamma}_b}\right), \tag{6.88}$$

where ρ_C is the channel correlation coefficient after a bit time T_b, K is the fading parameter of the Rician distribution, and $\overline{\gamma}_b$ is the average SNR per bit. For Rayleigh fading ($K = 0$) this simplifies to

$$\overline{P}_b = \frac{1}{2}\left[\frac{1 + \overline{\gamma}_b(1 - \rho_C)}{1 + \overline{\gamma}_b}\right]. \tag{6.89}$$

Letting $\overline{\gamma}_b \to \infty$ in (6.88) yields the irreducible error floor:

$$\text{DPSK:} \quad \overline{P}_{floor} = \frac{(1-\rho_C)e^{-K}}{2}. \tag{6.90}$$

A similar approach is used in [20] to bound the bit error probability of DQPSK in fast Rician fading as

$$P_b \leq \frac{1}{2}\left[1 - \sqrt{\frac{(\rho_C\overline{\gamma}_s/\sqrt{2})^2}{(\overline{\gamma}_s+1)^2 - (\rho_C\overline{\gamma}_s/\sqrt{2})^2}}\right] \exp\left[-\frac{(2-\sqrt{2})K\overline{\gamma}_s/2}{(\overline{\gamma}_s+1)-(\rho_C\overline{\gamma}_s/\sqrt{2})}\right], \tag{6.91}$$

where K is as before, ρ_C is the signal correlation coefficient after a symbol time T_s, and $\overline{\gamma}_s$ is the average SNR per symbol. Letting $\overline{\gamma}_s \to \infty$ yields the irreducible error floor:

$$\text{DQPSK:} \quad \overline{P}_{floor} = \frac{1}{2}\left[1 - \sqrt{\frac{(\rho_C/\sqrt{2})^2}{1-(\rho_C/\sqrt{2})^2}}\right] \exp\left[-\frac{(2-\sqrt{2})(K/2)}{1-\rho_C/\sqrt{2}}\right]. \tag{6.92}$$

As discussed in Chapter 3.2.1, the channel correlation $A_C(t)$ over time t equals the inverse Fourier transform of the Doppler power spectrum $S_C(f)$ as a function of Doppler frequency f. The correlation coefficient is thus $\rho_C = A_C(T)/A_C(0)$ evaluated at $T = T_s$ for DQPSK or $T = T_b$ for DPSK. Table 6.2, from [21], gives the value of ρ_C for several different Doppler power spectra models, where B_D is the Doppler spread of the channel. Assuming the uniform scattering model ($\rho_C = J_0(2\pi f_D T_b)$) and Rayleigh fading ($K = 0$) in (6.90) yields an irreducible error for DPSK of

$$P_{floor} = \frac{1 - J_0(2\pi f_D T_b)}{2} \approx .5(\pi f_D T_b)^2, \tag{6.93}$$

where $B_D = f_D = v/\lambda$ is the maximum Doppler in the channel. Note that in this expression, the error floor decreases with data rate $R = 1/T_b$. This is true in general for irreducible error floors of differential modulation due to Doppler, since the channel has less time to decorrelated between transmitted symbols. This phenomenon is one of the few instances in digital communications where performance improves as data rate increases.

Type	Doppler Power Spectrum $S_C(f)$	$\rho_C = A_C(T)/A_C(0)$
Rectangular	$\frac{S_0}{2B_D}, \lvert f \rvert < B_D$	sinc$(2B_DT)$
Gaussian	$\frac{S_0}{\sqrt{\pi}B_D}e^{-f^2/B_D^2}$	$e^{-(\pi B_D T)^2}$
Uniform Scattering	$\frac{S_0}{\pi\sqrt{B_D^2-f^2}}, \lvert f \rvert < B_D$	$J_0(2\pi B_D T)$
1st Order Butterworth	$\frac{S_0 B_D}{\pi(f^2+B_D^2})$	$e^{-2\pi B_D T}$

Table 6.2: Correlation Coefficients for Different Doppler Power Spectra Models.

A plot of (6.88), the error probability of DPSK in fast Rician fading, for uniform scattering ($\rho_C = J_0(2\pi f_D T_b)$) and different values of $f_D T_b$ is shown in Figure 6.4. We see from this figure that the error floor starts to dominate at $\overline{\gamma}_b$ = 15 dB in Rayleigh fading ($K = 0$), and as K increases the value of $\overline{\gamma}_b$ where the error floor dominates also increases. We also see that increasing the data rate $R_b = 1/T_b$ by an order of magnitude decreases the error floor by roughly two orders of magnitude.

Example 6.7:

Assume a Rayleigh fading channel with uniform scattering and a maximum Doppler of $f_D = 80$ Hertz. For what approximate range of data rates will the irreducible error floor of DPSK be below 10^{-4}.

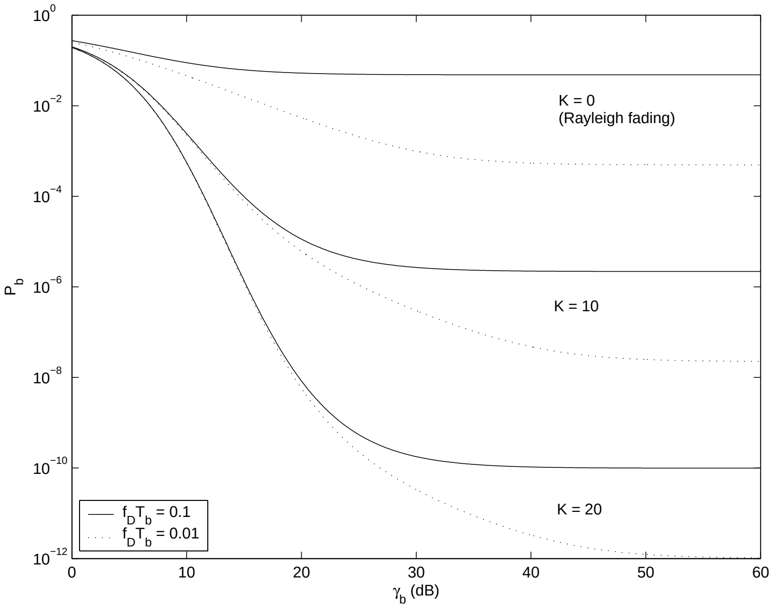


Figure 6.4: Average P_b for DPSK in Fast Rician Fading with Uniform Scattering.

Solution: We have $P_{floor} \approx .5(\pi f_D T_b)^2 < 10^{-4}$. Solving for T_b with $f_D = 80$ Hz, we get

$$T_b < \frac{\sqrt{2 \cdot 10^{-4}}}{\pi \cdot 80} = 5.63 \cdot 10^{-5},$$

which yields $R > 17.77$ Kbps.

Deriving analytical expressions for the irreducible error floor becomes intractable with more complex modulations, in which case simulations are often used. In particular, simulatons of the irreducible error floor for $\pi/4$ DQPSK with square root raised cosine filtering have been conducted since this modulation is used in the IS-54 TDMA standard [22, 23]. These simulation results indicate error floors between 10^{-3} and 10^{-4}. As expected, in these simulations the error floor increases with vehicle speed, since at higher vehicle speeds the channel decorrelates more over a symbol time.

6.5 Intersymbol Interference

Frequency-selective fading gives rise to ISI, where the received symbol over a given symbol period experiences interference from other symbols that have been delayed by multipath. Since increasing signal power also increases the power of the ISI, this interference gives rise to an irreducible error floor that is independent of signal power. The irreducible error floor is difficult to analyze, since it depends on the ISI characteristics and the modulation format, and the ISI characteristics depend on the characteristics of the channel and the sequence of transmitted symbols.

The first extensive analysis of ISI degradation to symbol error probability was done by Bello and Nelin [24]. In that work analytical expressions for the irreducible error floor of coherent FSK and noncoherent DPSK are determined assuming a Gaussian delay profile for the channel. To simplify the analysis, only ISI associated with adjacent symbols was taken into account. Even with this simplification, the expressions are very complex and must be approximated for evaluation. The irreducible error floor can also be evaluated analytically based on the worst-case sequence of transmitted symbols or it can be averaged over all possible symbol sequences [25, Chapter 8.2]. These expressions are also complex to evaluate due to their dependence on the channel and symbol sequence characteristics. An approximation to symbol error probability with ISI can be obtained by treating the ISI as uncorrelated white Gaussian noise [28]. Then the SNR becomes

$$\hat{\gamma}_s = \frac{P_r}{N_0 B + I}, \tag{6.94}$$

where I is the power associated with the ISI. In a static channel the resulting probability of symbol error will be $P_s(\hat{\gamma}_s)$ where P_s is the probability of symbol error in AWGN. If both the transmitted signal and the ISI experience flat-fading, then $\hat{\gamma}_s$ will be a random variable with a distribution $p(\hat{\gamma}_s)$, and the average symbol error probability is then $\overline{P}_s = \int P_s(\hat{\gamma}_s)p(\hat{\gamma}_s)d\gamma_s$. Note that $\hat{\gamma}_s$ is the ratio of two random variables: the received power P_r and the ISI power I, and thus the resulting distribution $p(\hat{\gamma}_s)$ may be hard to obtain and is not in closed form.

Irreducible error floors due to ISI are often obtained by simulation, which can easily incorporate different channel models, modulation formats, and symbol sequence characteristics [26, 28, 27, 22, 23]. The most extensive simulations for determining irreducible error floor due to ISI were done by Chuang in [26]. In this work BPSK, DPSK, QPSK, OQPSK and MSK modulations were simulated for different pulse shapes and for channels with different power delay profiles, including a Gaussian, exponential, equal-amplitude two-ray, and empirical power delay profile. The results of [26] indicate that the irreducible error floor is more sensitive to the rms delay spread of the channel than to the shape of its power delay profile. Moreover, pulse shaping can significantly impact the error floor: in the raised cosine pulses discussed in Chapter **??**, increasing β from zero to one can reduce the error floor by over an order of magnitude. An example of Chuang's simulation results is shown in Figure 6.5. This figure plots the irreducible bit error rate as a function of normalized rms delay spread $d = \sigma_{T_m}/T_s$ for BPSK, QPSK, OQPSK, and MSK modulation assuming a static channel with a Gaussian power delay profile. We see from this figure that for all modulations, we can approximately bound the irreducible error floor as $P_{floor} \leq d^2$ for $.02 \leq d \leq .1$. Other simulation results support this bound as well [28]. This bound imposes severe constraints on data rate even when symbol error probabilities on the order of 10^{-2} are acceptable. For example, the rms delay spread in a typical urban environment is approximately $\sigma_{T_m} = 2.5\mu$sec. To keep $\sigma_{T_m} < .1T_s$ requires that the data rate not exceed 40 Kbaud, which generally isn't enough for high-speed data applications. In rural environments, where multipath is not attenuated to the same degree as in cities, $\sigma_{T_m} \approx 25\mu$sec, which reduces the maximum data rate to 4 Kbaud.

Example 6.8:
Using the approximation $P_{floor} \leq (\sigma_{T_m}/T_s)^2$, find the maximum data rate that can be transmitted through a channel with delay spread $\sigma_{T_m} = 3\mu$ sec using either BPSK or QPSK modulation such that the probability of bit error P_b is less than 10^{-3}.

Solution: For BPSK, we set $P_{floor} = (\sigma_{T_m}/T_b)^2$, so we require $T_b \geq \sigma_{T_m}/\sqrt{P_{floor}} = 94.87\mu$secs, which leads to a data rate of $R = 1/T_b = 10.54$ Kbps. For QPSK, the same calculation yields $T_s \geq \sigma_{T_m}/\sqrt{P_{floor}} = 94.87\mu$secs. Since there are 2 bits per symbol, this leads to a data rate of $R = 2/T_s = 21.01$ Kbps. This indicates that for a given data rate, QPSK is more robust to ISI than BPSK, due to that fact that its symbol time is slower. This result is also true using the more accurate error floors associated with Figure 6.5 rather than the bound in this example.

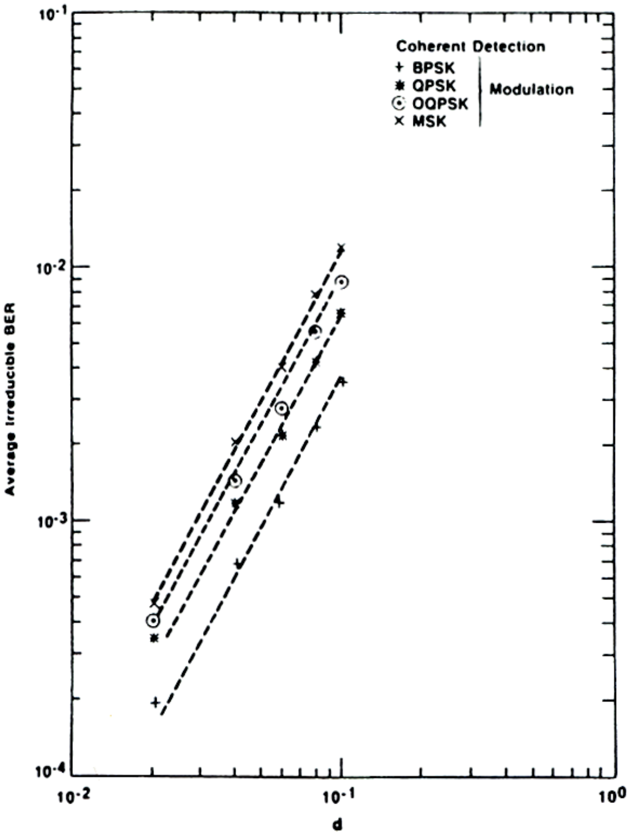


Figure 6.5: Irreducible error versus normalized rms delay spread $d = \sigma_{T_m}/T_s$ for Gaussian power delay profile (from [26] ©IEEE).

Bibliography

[1] J.G. Proakis, *Digital Communications*. 3rd Ed. New York: McGraw-Hill, 1995.

[2] M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection,* Prentice Hall: 1995.

[3] S. Haykin, *An Introduction to Analog and Digital Communications*. New York: Wiley, 1989.

[4] G. L. Stuber, *Principles of Mobile Communications*, Kluwer Academic Publishers, 1996.

[5] J. Craig, "New, simple and exact result for calculating the probability of error for two-dimensional signal constellations," Proc. Milcom 1991.

[6] F. S. Weinstein, "Simplified relationships for the probability distribution of the phase of a sine wave in narrow-band normal noise," *IEEE Trans. on Inform. Theory*, pp. 658–661, Sept. 1974.

[7] R. F. Pawula, "A new formula for MDPSK symbol error probability," *IEEE Commun. Letters*, pp. 271–272, Oct. 1998.

[8] M.K. Simon and D. Divsalar, "Some new twists to problems involving the Gaussian probability integral," *IEEE Trans. Commun.*, pp. 200-210, Feb. 1998.

[9] S. Rhodes, "Effect of noisy phase reference on coherent detection of offset-QPSK signals," *IEEE Trans. Commun.*, Vol 22, No. 8, pp. 1046–1055, Aug. 1974.

[10] N. R. Sollenberger and J. C.-I. Chuang, "Low-overhead symbol timing and carrier recovery for portable TDMA radio systems," *IEEE Trans. Commun.*, Vol 39, No. 10, pp. 1886–1892, Oct. 1990.

[11] R.F. Pawula, "on M-ary DPSK transmission over terrestrial and satellite channels," *IEEE Trans. Commun.*, Vol 32, No. 7, pp. 754–761, July 1984.

[12] W. Cowley and L. Sabel, "The performance of two symbol timing recovery algorithms for PSK demodulators," *IEEE Trans. Commun.*, Vol 42, No. 6, pp. 2345–2355, June 1994.

[13] S. Hinedi, M. Simon, and D. Raphaeli, "The performance of noncoherent orthogonal M-FSK in the presence of timing and frequency errors," *IEEE Trans. Commun.*, Vol 43, No. 2-4, pp. 922–933, Feb./March/April 1995.

[14] E. Grayver and B. Daneshrad, A low-power all-digital FSK receiver for deep space applications," *IEEE Trans. Commun.*, Vol 49, No. 5, pp. 911–921, May 2001.

[15] W.T. Webb and L. Hanzo, *Modern Quadrature Amplitude Modulation*, IEEE/Pentech Press, 1994.

[16] X. Tang, M.-S. Alouini, and A. Goldsmith, "Effects of channel estimation error on M-QAM BER performance in Rayleigh fading," *IEEE Trans. Commun.*, Vol 47, No. 12, pp. 1856–1864, Dec. 1999.

[17] P. A. Bello and B.D. Nelin, "The influence of fading spectrum on the bit error probabilities of incoherent and differentially coherent matched filter receivers," *IEEE Trans. Commun. Syst.*, Vol. 10, No. 2, pp. 160–168, June 1962.

[18] M. Schwartz, W.R. Bennett, and S. Stein, *Communication Systems and Techniques*, New York: McGraw Hill 1966, reprinted by Wily-IEEE Press, 1995.

[19] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels A Unified Approach to Performance Analysis*, Wiley 2000.

[20] P. Y. Kam, "Tight bounds on the bit-error probabilities of 2DPSK and 4DPSK in nonselective Rician fading," *IEEE Trans. Commun.*, pp. 860–862, July 1998.

[21] P. Y. Kam, "Bit error probabilities of MDPSK over the nonselective Rayleigh fading channel with diversity reception," *IEEE Trans. Commun.*, pp. 220–224, Feb. 1991.

[22] V. Fung, R.S. Rappaport, and B. Thoma, "Bit error simulation for $\pi/4$ DQPSK mobile radio communication using two-ray and measurement based impulse response models," *IEEE J. Select. Areas Commun.*, Vol. 11, No. 3, pp. 393–405, April 1993.

[23] S. Chennakeshu and G. J. Saulnier, "Differential detection of $\pi/4$-shifted-DQPSK for digital cellular radio," *IEEE Trans. Vehic. Technol.*, Vol. 42, No. 1, Feb. 1993.

[24] P. A. Bello and B.D. Nelin, "The effects of frequency selective fading on the binary error probabilities of incoherent and differentially coherent matched filter receivers," *IEEE Trans. Commun. Syst.*, Vol 11, pp. 170–186, June 1963.

[25] M. B. Pursley, *Introduction to Digital Communications*, Prentice Hall, 2005.

[26] J. Chuang, "The effects of time delay spread on portable radio communications channels with digital modulation," *IEEE J. Selected Areas Commun.*, Vol. SAC-5, No. 5, pp. 879–889, June 1987.

[27] C. Liu and K. Feher, "Bit error rate performance fo $\pi/4$ DQPSK in a frequency selective fast Rayleigh fading channel," *IEEE Trans. Vehic. Technol.*, Vol. 40, No. 3, pp. 558–568, Aug. 1991.

[28] S. Gurunathan and K. Feher, "Multipath simulation models for mobile radio channels," *Proc. IEEE Vehic. Technol. Conf.* pp. 131–134, May 1992.

Chapter 6 Problems

1. Consider a system in which data is transferred at a rate of 100 bits/sec over the channel.

   (a) Find the symbol duration if we use sinc pulse for signalling and the channel bandwidth is 10 kHz.

   (b) If the received SNR is 10 dB. Find the SNR per symbol and the SNR per bit if 4-QAM is used.

   (c) Find the SNR per symbol and the SNR per bit for 16-QAM and compare with these metrics for 4-QAM.

2. Consider BPSK modulation where the apriori probability of 0 and 1 is not the same. Specifically $p[s_n = 0]$ = 0.3 and $p[s_n = 1]$ = 0.7.

   (a) Find the probability of bit error P_b in AWGN assuming we encode a **1** as $s_1(t) = A\cos(2\pi f_c t)$ and a **0** as amplitude $s_2(t) = -A\cos(2\pi f_c t)$, and the receiver structure is as shown in Figure **??**.

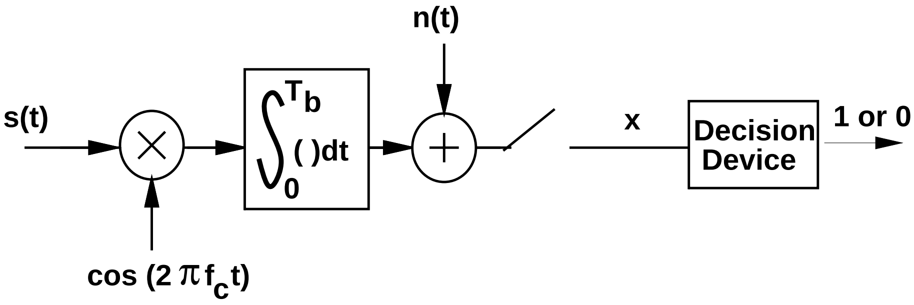
   (b) Suppose you can change the threshold value in the receiver of Figure **??**. Find the threshold value that yields equal error probability regardless of which bit is transmitted, i.e. the threshold value that yields $p(\hat{m} = 0|m = 1)p(m = 1) = p(\hat{m} = 1|m = 0)p(m = 0)$.

   (c) Now suppose we change the modulation so that $s_1(t) = A\cos(2\pi f_c t)$ and $s_2(t) = -B\cos(2\pi f_c t$. Find A and B so that the receiver of Figure **??** with threshold at zero has $p(\hat{m} = 0|m = 1)p(m = 1) = p(\hat{m} = 1|m = 0)p(m = 0)$.

   (d) Compute and compare the expression for P_b in parts (a), (b) and (c) assuming $E_b/N_0 = 10$ dB. For which system is p_b minimized?

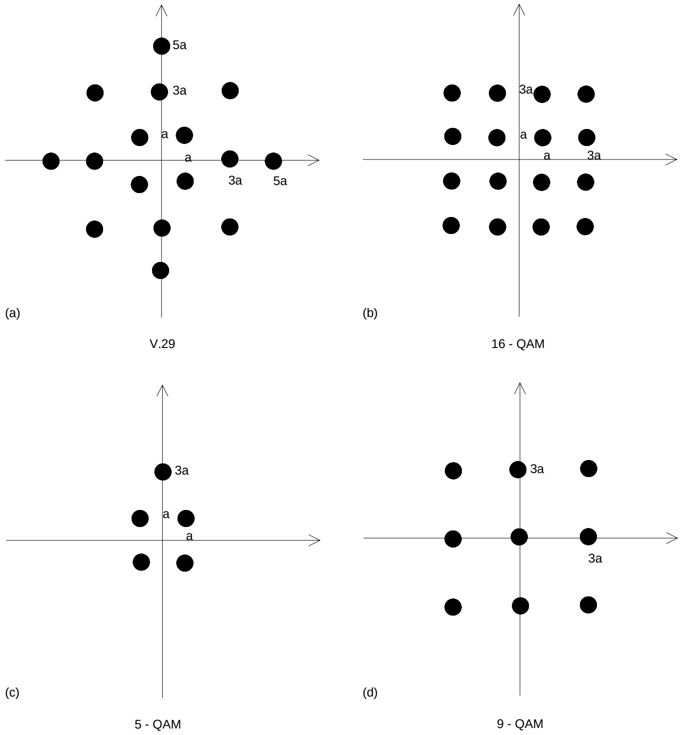
3. Consider a BPSK receiver where the demodulator has a phase offset of ϕ relative to the transmitted signal, so for a transmitted signal $s(t) = \pm g(t)\cos(2\pi f_c t)$, the carrier in the demodulator of Figure **??** is $\cos(2\pi f_c t + \phi)$. Determine the threshold level in the threshold device of Figure **??** that minimizes probability of bit error, and find this minimum error probability.

4. Assume a BPSK demodulator where the receiver noise is added after the integrator, as shown in the figure below. The decision device outputs a "1" if its input $\mathbf{x}$ has $\Re\mathbf{x} \geq 0$, and a "0" otherwise. Suppose the tone jammer $n(t) = 1.1e^{j\theta}$, where $p(\theta = n\pi/3) = 1/6$ for $n = 0, 1, 2, 3, 4, 5$. What is the probability of making a decision error in the decision device, i.e. outputting the wrong demodulated bit, assuming $A_c = \sqrt{2/T_b}$ and that information bits corresponding to a "1" ($s(t) = A_c\cos(2\pi f_c t)$) or a "0" ($s(t) = -A_c\cos(2\pi f_c t)$) are equally likely.



5. Find an approximation to P_s for the following signal constellations:

6. Plot the exact symbol error probability and the approximation from Table 6.1 of 16QAM with $0 \leq \gamma_s \leq 30$ dB. Does the error in the approximation increase or decrease with γ_s and why?



7. Plot the symbol error probability P_s for QPSK using the approximation in Table 6.1 and Craig's exact result for $0 \leq \gamma_s \leq 30$ dB. Does the error in the approximation increase or decrease with γ_s and why?

8. In this problem we derive an algebraic proof of the alternate representation of the Q-function (6.43) from its original representation (6.42). We will work with the complementary error function (erfc) for simplicity and make the conversion at the end. The erfc(x) function is traditionally defined by

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt. \tag{6.95}$$

The alternate representation is of this, corresonding to the alternate representation of the Q-function (6.43) is

$$\text{erfc}(x) = \frac{2}{\pi} \int_0^{\pi/2} e^{-x^2/\sin^2\theta} d\theta. \tag{6.96}$$

(a) Consider the integral

$$I_x(a) \triangleq \int_0^\infty \frac{e^{-at^2}}{x^2+t^2} dt. \tag{6.97}$$

Show that $I_x(a)$ satisfies the following differential equation:

$$x^2 I_x(a) - \frac{\partial I_x(a)}{\partial a} = \frac{1}{2}\sqrt{\frac{\pi}{a}}. \tag{6.98}$$

(b) Solve the differential equation (6.98) and deduce that

$$I_x(a) \triangleq \int_0^\infty \frac{e^{-at^2}}{x^2+t^2} dt = \frac{\pi}{2x} e^{ax^2} \text{erfc}(x\sqrt{a}). \tag{6.99}$$

`Hint:` $I_x(a)$ is a function in two variables x and a. However, since all our manipulations deal with a only, you can assume x to be a constant while solving the differential equation.

(c) Setting $a = 1$ in (6.99) and making a suitable change of variables in the LHS of (6.99), derive the alternate representation of the erfc function :

$$\text{erfc}(x) = \frac{2}{\pi} \int_0^{\pi/2} e^{-x^2/\sin^2\theta} d\theta$$

(d) Convert this alternate representation of the erfc function to the alternate representation of the Q function.

9. Consider a communication system which uses BPSK signalling with average signal power of 100 Watts and the noise power at the receiver is 4 Watts. Can this system be used for transmission of data? Can it be used for voice? Now consider there is fading with an average SNR $\bar{\gamma}_b$ = 20 dB. How does your answer to the above question change?

10. Consider a cellular system at 900 MHz with a transmission rate of 64 Kbps and multipath fading. Explain which performance metric, average probability of error or outage probability, is more appropriate and why for user speeds of 1 mph, 10 mph, and 100 mph.

11. Derive the expression for the moment generating function for Rayleigh fading.

12. This problem illustrates why satellite systems that must compensate for shadow fading are going bankrupt. Consider a LEO satellite system orbiting 500 Km above the earth. Assume the signal follows a free space path loss model with no multipath fading or shadowing. The transmitted signal has a carrier frequency of 900 MHz and a bandwidth of 10 KHz. The handheld receivers have noise spectral density of 10^{-16} (total noise power is N_oB) mW/Hz. Assume nondirectional antennas (0 dB gain) at both the transmitter and receiver. Suppose the satellite must support users in a circular cell on the earth of radius 100 Km at a BER of 10^{-6}.

    (a) For DPSK modulation what transmit power is needed such that all users in the cell meet the 10^{-6} BER target.

    (b) Repeat part (a) assuming that the channel also experiences log normal shadowing with $\sigma = 8$ dB, and that users in a cell must have $P_b = 10^{-6}$ (for each bit) with probability 0.9.

13. In this problem we explore the power penalty involved in going to higher level signal modulations, i.e. from BPSK to 16PSK.

    (a) Find the minimum distance between constellation points in 16PSK modulation as a function of signal energy E_s.

    (b) Find α_M and β_M such that the symbol error probability of 16PSK in AWGN is approximately

    $$P_s \approx \alpha_M Q\left(\sqrt{\beta_M \gamma_s}\right).$$

    (c) Using your expression in part (b), find an approximation for the average symbol error probability of 16PSK in Rayleigh fading in terms of $\overline{\gamma}_s$.

    (d) Convert the expressions for average symbol error probability of 16PSK in Rayleigh fading to expressions for average bit error probability assuming Gray coding.

    (e) Find the approximate value of $\overline{\gamma}_b$ required to obtain a BER of 10^{-3} in Rayleigh fading for BPSK and 16PSK. What is the power penalty in going to the higher level signal constellation at this BER?

14. Find a closed-form expression for the average probability of error for DPSK modulation in Nakagami-m fading evalute for $m = 4$ and $\overline{\gamma}_b = 10$ dB.

15. The Nakagami distribution is parameterized by m, which ranges from $m = .5$ to $m = \infty$. The m parameter measures the ratio of LOS signal power to multipath power, so $m = 1$ corresponds to Rayleigh fading, $m = \infty$ corresponds to an AWGN channel with no fading, and $m = .5$ corresponds to fading that results in performance that is worse than with a Rayleigh distribution. In this problem we explore the impact of the parameter m on the performance of BPSK modulation in Nakagami fading.

    Plot the average bit error $\overline{P_b}$ of BPSK modulation in Nakagami fading with average SNR ranging from 0 to 20dB for m parameters $m = 1$ (Rayleigh), $m = 2$, and $m = 4$ (The Moment Generating Function technique of Section 6.3.3 should be used to obtain the average error probability). At an average SNR of 10 dB, what is the difference in average BER?

16. Assume a cellular system with log-normal shadowing plus Rayleigh fading. The signal modulation is DPSK. The service provider has determined that it can deal with an outage probability of .01, i.e. 1 in 100 customers are unhappy at any given time. In nonoutage the voice BER requirement is $\overline{P}_b = 10^{-3}$. Assume a noise power spectral density of $N_o = 10^{-16}$ mW/Hz, a signal bandwidth of 30 KHz, a carrier frequency of 900 MHz, free space path loss propagation with nondirectional antennas, and shadowing standard deviation of $\sigma = 6$ dB. Find the maximum cell size that can achieve this performance if the transmit power at the mobiles is limited to 100 mW.

17. Consider a cellular system with circular cells with radius equal to 100 meters. Assume propagation follows the simplified path loss model with $K = 1$, $d_0 = 1$ m, and $\gamma = 3$. Assume the signal experiences log-normal shadowing on top of path loss with $\sigma_{\psi_{dB}} = 4$ as well as Rayleigh fading. The transmit power at the base station is $P_t = 100$ mW, the system bandwidth is $B = 30$ KHz, and the noise PSD is $N_0 = 10^{-14}$ W/Hz. Assuming BPSK modulation, we want to find the cell coverage area (percentage of locations in the cell) where users have average P_b less than 10^{-3}.

    (a) Find the received power due to path loss at the cell boundary.

    (b) Find the minimum average received power (due to path loss and shadowing) such that with Rayleigh fading about this average, a BPSK modulated signal with this average received power at a given cell location has $\overline{P}_b < 10^{-4}$.

    (c) Given the propagation model for this system (simplified path loss, shadowing, and Rayleigh fading), find the percentage of locations in the cell where under BPSK modulation, $\overline{P}_b < 10^{-4}$.

18. In this problem we derive the probability of bit error for DPSK in fast Rayleigh fading. By symmetry, the probability of error is the same for transmitting a zero bit or a one bit. Let us assume that over time kT_b a zero bit is transmitted, so the transmitted symbol at time kT_b is the same as at time $k-1$: $\mathbf{s}(k) = \mathbf{s}(k-1)$. In fast fading the corresponding received symbols are $\mathbf{z}(k-1) = g_{k-1}\mathbf{s}(k-1) + n(k-1)$ and $\mathbf{z}(k) = g_k\mathbf{s}(k-1) + n(k)$, where g_{k-1} and g_k are the fading channel gains associated with transmissions over times $(k-1)T_b$ and kT_b.

    a) Show that the decision variable input to the phase comparator of Figure **??** to extract the phase difference is $\mathbf{z}(k)\mathbf{z}^*(k-1) = g_k g_{k-1}^* + g_k\mathbf{s}(k-1)n_{k-1}^* + g_{k-1}^* s_{k-1}^* n_k + n_k n_{k-1}^*$.

    Assuming a reasonable SNR, the last term $n_k n_{k-1}^*$ of this expression can be neglected. Neglecting this term and defining $\tilde{n}_k = s_{k-1}^* n_k$ and $\tilde{n}_{k-1} = s_{k-1}^* n_{k-1}$, we get a new random variable $\tilde{z} = g_k g_{k-1}^* + g_k \tilde{n}_{k-1}^* + g_{k-1}^* \tilde{n}_k$. Given that a zero bit was transmitted over time kT_b, an error is made if $x = \Re\{\tilde{z}\} < 0$, so we must determine the distribution of x. The characteristic function for x is the 2-sided Laplace transform of the pdf of x:

$$\Phi_X(s) = \int_{-\infty}^{\infty} p_X(s) e^{-sx} dx = E[e^{-sx}].$$

This function will have a left plane pole p_1 and a right plane pole p_2, so can be written as

$$\Phi_X(s) = \frac{p_1 p_2}{(s-p_1)(s-p_2)}.$$

The left plane pole p_1 corresponds to the pdf $p_X(x)$ for $x \geq 0$ and the right plane pole corresponds to the pdf $p_X(x)$ for $x < 0$

b) Show through partial fraction expansion that $\Phi_X(s)$ can be written as

$$\Phi_X(s) = \frac{p_1 p_2}{(p_1 - p_2)} \frac{1}{(s-p_1)} + \frac{p_1 p_2}{(p_2 - p_1)} \frac{1}{(s-p_2)}.$$

An error is made if $x = \Re\{\tilde{z}\} < 0$, so we need only consider the pdf $p_X(x)$ for $x < 0$ corresponding to the second term of $\Phi_X(s)$ in part b).

c) Show that the inverse Laplace transform of the second term of $\Phi_X(s)$ from part b) is

$$p_X(x) = \frac{p_1 p_2}{p_2 - p_1} e^{p_2 x}, \quad x < 0.$$

d) Use part c) to show that $P_b = -p_1/(p_2 - p_1)$.

In $x = \Re\{\tilde{z}\} = \Re\{g_k g_{k-1}^* + g_k \tilde{n}_{k-1}^* + g_{k-1}^* \tilde{n}_k.\}$ the channel gains g_k and g_{k-1} and noises $\tilde{n}_k$ and $\tilde{n}_{k-1}$ are complex Gaussian random variables. Thus, the poles p_1 and p_2 in $p_X(x)$ are derived using the general quadratic form of complex Gaussian random variables [1, Appendix B][18, Appendix B] as

$$p_1 = \frac{-1}{2(\overline{\gamma}_b[1+\rho_c])] + N_0)},$$

and

$$p_2 = \frac{1}{2(\overline{\gamma}_b[1-\rho_c])] + N_0)},$$

for ρ_C the correlation coefficient of the channel over the bit time T_b.

e) Find a general expression for P_b in fast Rayleigh fading using these values of p_1 and p_2 in the P_e expression from part d).

f) Show that this reduces to the average probability of error $\overline{P}_b = \frac{1}{2(1+\overline{\gamma}_b)}$ for a slowly fading channel that does not decorrelate over a bit time.

19. Plot the bit error probability for DPSK in fast Rayleigh fading for $\overline{\gamma}_b$ ranging from 0 to 60 dB and $\rho_C = J_0(2\pi B_D T)$ with $B_D T = .01, .001,$ and $.0001$. For each value of $B_d T$, at approximately what value of $\overline{\gamma}_b$ does the error floor dominate the error probability/

20. Find the irreducible error floor for DQPSK modulation due to Doppler, assuming a Gaussian Doppler power spectrum with $B_D = 80$ Hz and Rician fading with $K = 2$.

21. Consider a wireless channel with an average delay spread of 100 nsec and a doppler spread of 80 Hz. Given the error floors due to doppler and ISI, for DQPSK modulation in Rayleigh fading and uniform scattering, approximately what range of data rates can be transmitted over this channel with a BER less than 10^{-4}.

22. Using the error floors of Figure 6.5, find the maximum data rate that can be transmitted through a channel with delay spread $\sigma_{T_m} = 3\mu$ sec using BPSK, QPSK, or MSK modulation such that the probability of bit error P_b is less than 10^{-3}.

Chapter 10

Multiple Antennas and Space-Time Communications

In this chapter we consider systems with multiple antennas at the transmitter and receiver, which are commonly referred to as multiple input multiple output (MIMO) systems. The multiple antennas can be used to increase data rates through multiplexing or to improve performance through diversity. We have already seen diversity in Chapter **??**. In MIMO systems the transmit and receive antennas can both be used for diversity gain. Multiplexing is obtained by exploiting the structure of the channel gain matrix to obtain independent signalling paths that can be used to send independent data. Indeed, the initial excitement about MIMO was sparked by the pioneering work of Winters [1], Foschini [2], Gans [3], and Telatar [4][5] predicting remarkable spectral efficiencies for wireless systems with multiple transmit and receive antennas. These spectral efficiency gains often require accurate knowledge of the channel at the receiver, and sometimes at the transmitter as well. In addition to spectral efficiency gains, ISI and interference from other users can be reduced using smart antenna techniques. The cost of the performance enhancements obtained through MIMO techniques is the added cost of deploying multiple antennas, the space and power requirements of these extra antennas (especially on small handheld units), and the added complexity required for multi-dimensional signal processing. In this chapter we examine these different uses for multiple antennas and find their performance advantages. The mathematics in this chapter uses several key results from matrix theory: Appendix C provides a brief overview of these results.

10.1 Narrowband MIMO Model

In this section we consider a narrowband MIMO channel. A narrowband point-to-point communication system of M_t transmit and M_r receive antennas is shown in Figure 10.1 This system can be represented by the following discrete time model:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{M_r} \end{bmatrix} = \begin{bmatrix} h_{11} & \cdots & h_{1M_t} \\ \vdots & \ddots & \vdots \\ h_{M_r1} & \cdots & h_{M_rM_t} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{M_t} \end{bmatrix} + \begin{bmatrix} n_1 \\ \vdots \\ n_{M_r} \end{bmatrix}$$

or simply as $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$. Here $\mathbf{x}$ represents the M_t-dimensional transmitted symbol, $\mathbf{n}$ is the M_r-dimensional noise vector, and $\mathbf{H}$ is the $M_r \times M_t$ matrix of channel gains h_{ij} representing the gain from transmit antenna j to receive antenna i. We assume a channel bandwidth of B and complex Gaussian noise with zero mean and covariance matrix $\sigma_n^2\mathbf{I}_{M_r}$, where typically $\sigma_n^2 = N_0B$. For simplicity, given a transmit power constraint P we will assume an equivalent model with a noise power of unity and transmit power $P/\sigma_n^2 = \rho$, where ρ can be interpreted

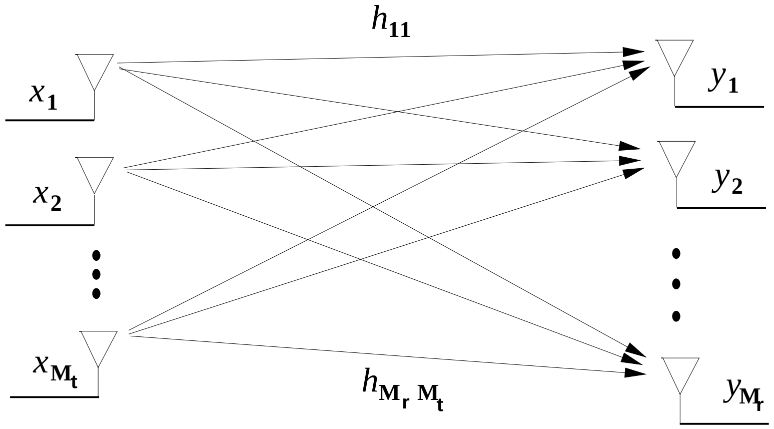


Figure 10.1: MIMO Systems.

as the average SNR per receive antenna under unity channel gain. This power constraint implies that the input symbols satisfy

$$\sum_{i=1}^{M_t} \mathrm{E}[x_i x_i^*] = \rho, \tag{10.1}$$

or, equivalently, that Tr$(\mathbf{R_x}) = \rho$, where Tr$(\mathbf{R_x})$ is the trace of the input covariance matrix $\mathbf{R_x} = E[\mathbf{x}\mathbf{x}^{\mathbf{T}}]$.

Different assumptions can be made about the knowledge of the channel gain matrix $\mathbf{H}$ at the transmitter and receiver, referred to as channel side information at the transmitter (CSIT) and channel side information at the receiver (CSIR), respectively. For a static channel CSIR is typically assumed, since the channel gains can be obtained fairly easily by sending a pilot sequence for channel estimation. More details on estimation techniques for MIMO channels can be found in [10, Chapter 3.9]. If a feedback path is available then CSIR from the receiver can be sent back to the transmitter to provide CSIT: CSIT may also be available in time-division duplexing systems without a feedback path by exploiting reciprocal properties of propagation. When the channel is not known at either the transmitter or receiver then some distribution on the channel gain matrix must be assumed. The most common model for this distribution is a zero-mean spatially white (ZMSW) model, where the entries of $\mathbf{H}$ are assumed to be i.i.d. zero mean, unit variance, complex circularly symmetric Gaussian random variables[1]. We adopt this model unless stated otherwise. Alternatively, these entries may be complex circularly symmetric Gaussian random variables with a non-zero mean or with a covariance matrix not equal to the identity matrix. In general, different assumptions about CSI and about the distribution of the $\mathbf{H}$ entries lead to different channel capacities and different approaches to space-time signalling.

Optimal decoding of the received signal requires ML demodulation. If the symbols modulated onto each of the M_t transmit antennas are chosen from an alphabet of size $|\mathcal{X}|$, then because of the cross-coupling between transmitted symbols at the receiver antennas, ML demodulation requires an exhaustive search over all $|\mathcal{X}|^{M_t}$ possible input vector of M_t symbols. For general channel matrices, when the transmitter does not know H this complexity cannot be reduced further. This decoding complexity is typically prohibitive for even a small number of transmit antennas. However, decoding complexity is significantly reduced if the channel is known at the transmitter,

[1] A complex Gaussian vector $\mathbf{x}$ is circularly symmetric if

$$E[(\mathbf{x} - E[\mathbf{x}])((\mathbf{x} - \mathbf{E}[\mathbf{x}])^{\mathbf{H}}] = \mathbf{.5}\begin{bmatrix} \Re\{Q\} & -\Im\{Q\} \\ \Im\{Q\} & \Re\{Q\} \end{bmatrix}$$

for some Hermitian non-negative definite matrix $\mathbf{Q}$

as shown in Section 10.2.

10.2 Parallel Decomposition of the MIMO Channel

We have seen in Chapter **??** that multiple antennas at the transmitter or receiver can be used for diversity gain. When *both* the transmitter and receiver have multiple antennas, there is another mechanism for performance gain called **multiplexing gain**. The multiplexing gain of a MIMO system results from the fact that a MIMO channel can be decomposed into a number R of parallel independent channels. By multiplexing independent data onto these independent channels, we get an R-fold increase in data rate in comparison to a system with just one antenna at the transmitter and receiver. This increased data rate is called the multiplexing gain. In this section we describe how to obtain independent channels from a MIMO system.

Consider a MIMO channel with $M_r \times M_t$ channel gain matrix $\mathbf{H}$ known to both the transmitter and the receiver. Let R_H denote the rank of $\mathbf{H}$. From matrix theory, for any matrix $\mathbf{H}$ we can obtain its singular value decomposition (SVD) as

$$\mathbf{H} = \mathbf{U\Sigma V^H}, \tag{10.2}$$

where the $M_r \times M_r$ matrix $\mathbf{U}$ and the $M_t \times M_t$ matrix $\mathbf{V}$ are unitary matrices[2] and $\mathbf{\Sigma}$ is an $M_r \times M_t$ diagonal matrix of singular values $\{\sigma_i\}$ of $\mathbf{H}$. These singular values have the property that $\sigma_i = \sqrt{\lambda_i}$ for λ_i the ith eigenvalue of HH^H, and R_H of these singular values are nonzero, where R_H is the rank of the matrix $\mathbf{H}$. Since R_H cannot exceed the number of columns or rows of $\mathbf{H}$, $R_H \leq \min(M_t, M_r)$. If $\mathbf{H}$ is full rank, which is sometimes referred to as a **rich scattering environment**, then $R_H = \min(M_t, M_r)$. Other environments may lead to a low rank $\mathbf{H}$: a channel with high correlation among the gains in $\mathbf{H}$ may have rank 1.

The parallel decomposition of the channel is obtained by defining a transformation on the channel input and output $\mathbf{x}$ and $\mathbf{y}$ through **transmit precoding** and **receiver shaping**. In transmit precoding the input to the antennas $\mathbf{x}$ is generated through a linear transformation on input vector $\tilde{\mathbf{x}}$ as $\mathbf{x} = \mathbf{V^H}\tilde{\mathbf{x}}$. Receiver shaping performs a similar operation at the receiver by multiplying the channel output $\mathbf{y}$ with U^H, as shown in Figure 10.2.

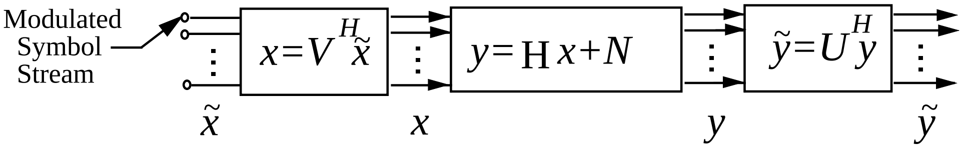


Figure 10.2: Transmit Precoding and Receiver Shaping.

The transmit precoding and receiver shaping transform the MIMO channel into R_H parallel single-input single-output (SISO) channels with input $\tilde{\mathbf{x}}$ and output $\tilde{\mathbf{y}}$, since from the SVD, we have that

$$\begin{aligned}
\tilde{\mathbf{y}} &= \mathbf{U}^H(\mathbf{Hx}+\mathbf{n}) \\
&= \mathbf{U}^H(\mathbf{U\Sigma V}^H\mathbf{x}+\mathbf{n}) \\
&= \mathbf{U}^H(\mathbf{U\Sigma V}^H\mathbf{V}\tilde{\mathbf{x}}+\mathbf{n}) \\
&= \mathbf{U}^H\mathbf{U\Sigma V}^H\mathbf{V}\tilde{\mathbf{x}}+\mathbf{U}^H\mathbf{n} \\
&= \Sigma\tilde{\mathbf{x}}+\tilde{\mathbf{n}},
\end{aligned}$$

where $\tilde{\mathbf{n}} = \mathbf{U}^H\mathbf{n}$ and Σ is the diagonal matrix of singular values of $\mathbf{H}$ with σ_i on the ith diagonal. Note that multiplication by a unitary matrix does not change the distribution of the noise, i.e. $\mathbf{n}$ and $\tilde{\mathbf{n}}$ are identically

[2] U and V unitary imply $\mathbf{UU^H} = \mathbf{I_{M_r}}$ and $\mathbf{V^HV} = \mathbf{I_{M_t}}$.

distributed. Thus, the transmit precoding and receiver shaping transform the MIMO channel into R_H parallel independent channels where the ith channel has input $\tilde{x}_i$, output $\tilde{y}_i$, noise $\tilde{n}_i$, and channel gain σ_i. Note that the σ_is are related since they are all functions of $\mathbf{H}$, but since the resulting parallel channels do not interfere with each other, we say that the channels with these gains are independent, linked only through the total power constraint. This parallel decomposition is shown in Figure 10.3. Since the parallel channels do not interfere with each other, the optimal ML demodulation complexity is linear in R_H, the number of independent paths that need to be decoded. Moreover, by sending independent data across each of the parallel channels, the MIMO channel can support R_H times the data rate of a system with just one transmit and receive antenna, leading to a multiplexing gain of R_H. Note, however, that the performance on each of the channels will depend on its gain σ_i. The next section will more precisely characterize the multiplexing gain associated with the Shannon capacity of the MIMO channel.

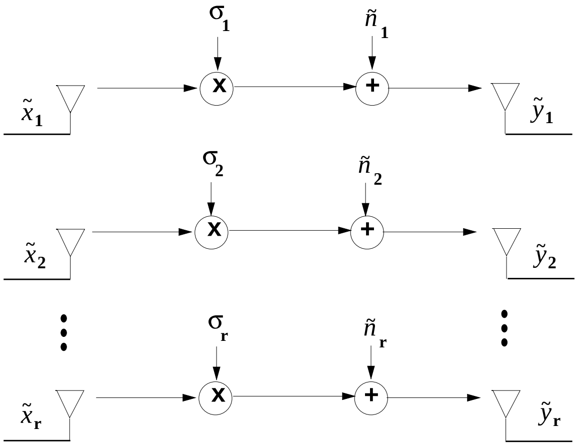


Figure 10.3: Parallel Decomposition of the MIMO Channel.

Example 10.1: Find the equivalent parallel channel model for a MIMO channel with channel gain matrix

$$\mathbf{H} = \begin{bmatrix} .1 & .3 & .7 \\ .5 & .4 & .1 \\ .2 & .6 & .8 \end{bmatrix} \tag{10.3}$$

Solution: The SVD of $\mathbf{H}$ is given by

$$\mathbf{H} = \begin{bmatrix} -0.555 & .3764 & -.7418 \\ -.3338 & -.9176 & -.2158 \\ -.7619 & 0.1278 & .6349 \end{bmatrix} \begin{bmatrix} 1.3333 & 0 & 0 \\ 0 & .5129 & 0 \\ 0 & 0 & .0965 \end{bmatrix} \begin{bmatrix} -.2811 & -.7713 & -.5710 \\ -.5679 & -.3459 & .7469 \\ -.7736 & .5342 & -.3408 \end{bmatrix}. \tag{10.4}$$

Thus, since there are 3 nonzero singular values, $R_H = 3$, leading to three parallel channels, with channel gains $\sigma_1 = 1.3333$, and $\sigma_2 = .5129$, and $\sigma_3 = .0965$, respectively. Note that the channels have diminishing gain, with a very small gain on the third channel. Hence, this last channel will either have a high error probability or a low capacity.

10.3 MIMO Channel Capacity

This section focuses on the Shannon capacity of a MIMO channel, which equals the maximum data rate that can be transmitted over the channel with arbitrarily small error probability. Capacity versus outage defines the maximum rate that can be transmitted over the channel with some nonzero outage probability. Channel capacity depends on what is known about the channel gain matrix or its distribution at the transmitter and/or receiver. Throughout this section it is assumed that the receiver has knowledge of the channel matrix $\mathbf{H}$, since for static channels a good estimate of $\mathbf{H}$ can be obtained fairly easily. First the static channel capacity will be given, which forms the basis for the subsequent section on capacity of fading channels.

10.3.1 Static Channels

The capacity of a MIMO channel is an extension of the mutual information formula for a SISO channel given by (4.3) in Chapter 4 to a matrix channel. Specifically, the capacity is given in terms of the mutual information between the channel input vector $\mathbf{x}$ and output vector $\mathbf{y}$ as

$$C = \max_{p(\mathbf{x})} I(\mathbf{X};\mathbf{Y}) = \max_{p(\mathbf{x})} \left[H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})\right], \tag{10.5}$$

for $H(\mathbf{Y})$ and $H(\mathbf{Y}|\mathbf{X})$ the entropy in $\mathbf{y}$ and $\mathbf{y}|\mathbf{x}$, as defined in Chapter 4.1[3]. The definition of entropy yields that $H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{N})$, the entropy in the noise. Since this noise $\mathbf{n}$ has fixed entropy independent of the channel input, maximizing mutual information is equivalent to maximizing the entropy in $\mathbf{y}$.

The mutual information of $\mathbf{y}$ depends on its covariance matrix, which for the narrowband MIMO model is given by

$$\mathbf{R_y} = E[\mathbf{y}\mathbf{y}^H] = \mathbf{H}\mathbf{R_x}\mathbf{H}^H + \mathbf{I}_{M_r}, \tag{10.6}$$

where $\mathbf{R_x}$ is the covariance of the MIMO channel input. It turns out that for all random vectors with a given covariance matrix $\mathbf{R_y}$, the entropy of $\mathbf{y}$ is maximized when $\mathbf{y}$ is a zero-mean circularly-symmetric complex Gaussian (ZMCSCG) random vector [5]. But $\mathbf{y}$ is only ZMCSCG if the input $\mathbf{x}$ is ZMCSCG, and therefore this is the optimal distribution on $\mathbf{x}$. This yields $H(\mathbf{y}) = \mathbf{B}\log_\mathbf{2}\det[\pi\mathbf{e}\mathbf{R_y}]$ and $H(\mathbf{n}) = \mathbf{B}\log_\mathbf{2}\det[\pi\mathbf{e}\mathbf{I}_{\mathbf{M_r}}]$, resulting in the mutual information

$$I(\mathbf{X};\mathbf{Y}) = B\log_2\det\left[\mathbf{I}_{M_r} + \mathbf{H}\mathbf{R_x}\mathbf{H}^H\right]. \tag{10.7}$$

This formula was derived in [3, 5] for the mutual information of a multiantenna system, and also appeared in earlier works on MIMO systems [6, 7] and matrix models for ISI channels [8, 9].

The MIMO capacity is achieved by maximizing the mutual information (10.7) over all input covariance matrices $\mathbf{R_x}$ satisfying the power constraint:

$$C = \max_{\mathbf{R_x}:\mathbf{Tr}(\mathbf{R_x})=\rho} B\log_2\det\left[\mathbf{I}_{M_r} + \mathbf{H}\mathbf{R_x}\mathbf{H}^H\right], \tag{10.8}$$

where $\det[\mathbf{A}]$ denotes the determinant of the matrix $\mathbf{A}$. Clearly the optimization relative to $\mathbf{R_x}$ will depend on whether or not $\mathbf{H}$ is known at the transmitter. We now consider this maximizing under different assumptions about transmitter CSI.

Channel Known at Transmitter: Waterfilling

The MIMO decomposition described in Section 10.2 allows a simple characterization of the MIMO channel capacity for a fixed channel matrix $\mathbf{H}$ known at the transmitter and receiver. Specifically, the capacity equals the sum

[3] Entropy was defined in Chapter 4.1 for scalar random variables, but the definition is identical for random vectors

of capacities on each of the independent parallel channels with the transmit power optimally allocated between these channels. This optimization of transmit power across the independent channels results from optimizing the input covariance matrix to maximize the capacity formula (10.8). Substituting the matrix SVD (10.2) into (10.8) and using properties of unitary matrices we get the MIMO capacity with CSIT and CSIR as

$$C = \max_{\rho_i : \sum_i \rho_i \leq \rho} \sum_i B \log_2 \left(1 + \sigma_i^2 \rho_i\right). \tag{10.9}$$

Since $\rho = P/\sigma_n^2$, the capacity (10.9) can also be expressed in terms of the power allocation P_i to the ith parallel channel as

$$C = \max_{P_i : \sum_i P_i \leq P} \sum_i B \log_2 \left(1 + \frac{\sigma_i^2 P_i}{\sigma_n^2}\right) = \max_{P_i : \sum_i P_i \leq P} \sum_i B \log_2 \left(1 + \frac{P_i \gamma_i}{P}\right) \tag{10.10}$$

where $\rho_i = P_i/\sigma_n^2$ and $\gamma_i = \sigma_i^2 P/\sigma_n^2$ is the SNR associated with the ith channel at full power. This capacity formula is the same as in the case of flat fading (4.9) or in frequency-selective fading (4.23). Solving the optimization leads to a water-filling power allocation for the MIMO channel:

$$\frac{P_i}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_i} & \gamma_i \geq \gamma_0 \\ 0 & \gamma_i < \gamma_0 \end{cases} \tag{10.11}$$

for some cutoff value γ_0. The resulting capacity is then

$$C = \sum_{i : \gamma_i \geq \gamma_0} B \log(\gamma_i/\gamma_0). \tag{10.12}$$

Example 10.2: Find the capacity and optimal power allocation for the MIMO channel given in the previous example, assuming $\rho = P/\sigma_n^2 = 10$ dB and $B = 1$ Hz.

Solution: From the previous example, the singular values of the channel are $\sigma_1 = 1.3333$, $\sigma_2 = 0.5129$, and $\sigma_3 = 0.0965$. Since $\gamma_i = 10\sigma_i^2$, this yields $\gamma_1 = 17.77$, $\gamma_2 = 2.63$, and $\gamma_3 = .087$. Assuming that power is allocated to all three parallel channels, the power constraint yields

$$\sum_{i=1}^{3} \left(\frac{1}{\gamma_0} - \frac{1}{\gamma_i}\right) = 1 \rightarrow \frac{3}{\gamma_0} = 1 + \sum_{i=1}^{3} \frac{1}{\gamma_i} = 12.974.$$

Solving for γ_0 yields $\gamma_0 = .231$, which is inconsistent since $\gamma_3 = .087 < \gamma_0 = .231$. Thus, the third channel is not allocated any power. Then the power constraint yields

$$\sum_{i=1}^{2} \left(\frac{1}{\gamma_0} - \frac{1}{\gamma_i}\right) = 1 \rightarrow \frac{2}{\gamma_0} = 1 + \sum_{i=1}^{2} \frac{1}{\gamma_i} = 1.436.$$

Solving for γ_0 for this case yields $\gamma_0 = 1.392 < \gamma_2$, so this is the correct cutoff value. Then $P_i = 1/1.392 - 1/\gamma_i$, so $P_1 = .662$ and $P_2 = .338$. The capacity is given by $C = \log_2(\gamma_1/\gamma_0) + \log_2(\gamma_2/\gamma_0) = 4.59$.

Capacity under perfect CSIT and CSIR can also be defined on channels where there is a single antenna at the transmitter and multiple receive antennas (single-input multiple-output or SIMO) or multiple transmit antennas and a single receive antenna (multiple-input single-output or MISO). These channels can only obtain diversity gain from the multiple antennas. When both transmitter and receiver know the channel the capacity equals that of a SISO channel with the signal transmitted or received over the multiple antennas coherently combined to maximize the channel SNR, as in MRC. This results in capacity $C = B\log_2(1+\rho\mathbf{hc})$, with the channel matrix $\mathbf{H}$ reduced to a vector $\mathbf{h}$ of channel gains, the optimal weight vector $\mathbf{c} = \mathbf{h}^*/||\mathbf{h}||$, and $\rho = P/\sigma_n^2$.

Channel Unknown at Transmitter: Uniform Power Allocation

Suppose now that the receiver knows the channel but the transmitter does not. Without channel information, the transmitter cannot optimize its power allocation or input covariance structure across antennas. If the distribution of $\mathbf{H}$ follows the ZMSW channel gain model, there is no bias in terms of the mean or covariance of $\mathbf{H}$. Thus, it seems intuitive that the best strategy should be to allocate equal power to each transmit antenna, resulting in an input covariance matrix equal to the scaled identity matrix: $\mathbf{R_x} = \frac{\rho}{M_t}\mathbf{I}_{M_t}$. It is shown in [4] that under these assumptions this input covariance matrix indeed maximizes the mutual information of the channel. For an M_t-transmit, M_r-receive antenna system, this yields mutual information given by

$$I = B\log_2 \det[\mathbf{I}_{M_r} + \frac{\rho}{M_t}\mathbf{H}\mathbf{H}^H].$$

Using the SVD of $\mathbf{H}$, we can express this as

$$I = \sum_{i=1}^{R_H} B\log_2\left(1+\frac{\gamma_i}{M_t}\right), \qquad (10.13)$$

where $\gamma_i = \sigma_i^2\rho = \sigma_i^2 P/\sigma_n^2$ and R_H is the number of nonzero singular values of $\mathbf{H}$.

The mutual information of the MIMO channel (10.13) depends on the specific realization of the matrix $\mathbf{H}$, in particular its singular values $\{\sigma_i\}$. The average mutual information of a random matrix $\mathbf{H}$, averaged over the matrix distribution, depends on the probability distribution of the singular values of $\mathbf{H}$ [5, 13, 11]. In fading channels the transmitter can transmit at a rate equal to this average mutual information and insure correct reception of the data, as discussed in the next section. But for a static channel, if the transmitter does not know the channel realization or, more precisely, the channel's average mutual information then it does not know at what rate to transmit such that the data will be received correctly. In this case the appropriate capacity definition is capacity with outage. In capacity with outage the transmitter fixes a transmission rate C, and the outage probability associated with C is the probability that the transmitted data will not be received correctly or, equivalently, the probability that the channel $\mathbf{H}$ has mutual information less than C. This probability is given by

$$p_{out} = p\left(\mathbf{H} : B\log_2\det\left[\mathbf{I}_{M_r} + \frac{\rho}{M_t}\mathbf{H}\mathbf{H}^H\right] < C\right). \qquad (10.14)$$

As the number of transmit and receive antennas grows large, random matrix theory provides a central limit theorem for the distribution of the singular values of $\mathbf{H}$ [14], resulting in a constant mutual information for all channel realizations. These results were applied to obtain MIMO channel capacity with uncorrelated fading in [15, 16, 17, 18] and with correlated fading in [19, 20, 12]. As an example of this limiting distribution, note that for fixed M_r, under the ZMSW model the law of large numbers implies that

$$\lim_{M_t\to\infty} \frac{1}{M_t}\mathbf{H}\mathbf{H}^H = \mathbf{I}_{M_r}. \qquad (10.15)$$

Substituting this into (10.13) yields that the mutual information in the asymptotic limit of large M_t becomes a constant equal to $C = M_r B\log_2(1+\rho)$. Defining $M = \min(M_t, M_r)$, this implies that as M grows large, the

MIMO channel capacity in the absence of CSIT approaches $C = MB\log_2(1+\rho)$, and hence grows linearly in M. Moreover, this linear growth of capacity with M in the asymptotic limit of large M is observed even for a small number of antennas [20]. Similarly, as SNR grows large, capacity also grows linearly with $M = \min(M_t, M_r)$ for any M_t and M_r [2]. These results are the main reason for the widespread appeal of MIMO techniques: even if the channel realization is not known at the transmitter, the capacity of MIMO channels still grows linearly with the minimum number of transmit and receiver antennas, as long as the channel can be accurately estimated at the receiver. Thus, MIMO channels can provide very high data rates without requiring increased signal power or bandwidth. Note, however, that at very low SNRs transmit antennas are not beneficial: capacity only scales with the number of receive antennas indepedent of the number of transmit antennas. The reason is that at these low SNRs, the MIMO system is just trying to collect energy rather than exploit all available dimensions, so all energy is concentrated into one of the available transmit antenna to achieve capacity [4].

While lack of CSIT does not affect the growth rate of capacity relative to M, at least for a large number of antennas, it does complicate demodulation. Specifically, without CSIT the transmission scheme cannot convert the MIMO channel into non-interfering SISO channels. Recall that the decoding complexity is exponential in the number of independent symbols transmitted over the multiple transmit antennas, and this number equals the rank of the input covariance matrix.

The above analysis under perfect CSIR and no CSIT assumes that the channel gain matrix has a ZMSW distribution, i.e. it has mean zero and covariance matrix equal to the identity matrix. When the channel has nonzero mean or a non-identity covariance matrix, there is a spatial bias in the channel that should be exploited by the optimal transmission strategy, so equal power allocation across antennas is no longer optimal [23, 24, 25]. Results in [25, 26] indicate that when the channel has a dominant mean or covariance direction, **beamforming**, described in Section 10.4, achieves channel capacity. This is a fortuitous situation due to the simplicity of beamforming.

10.3.2 Fading Channels

Suppose now that the channel gain matrix experiences flat-fading, so the gains h_{ij} vary with time. As in the case of the static channel, the capacity depends on what is known about the channel matrix at the transmitter and receiver. With perfect CSIR and CSIT the transmitter can adapt to the channel fading and its capacity equals the average over all channel matrix realizations with optimal power allocation. With CSIR and no CSIT outage capacity is used to characterize the outage probability associated with any given channel rate. These different characterizations are described in more detail in the following sections.

Channel Known at Transmitter: Water-Filling

With CSIT and CSIR the transmitter optimizes its transmission strategy for each fading channel realization as in the case of a static channel. The capacity is then just the average of capacities associated with each channel realization, given by (10.8), with power optimally allocated. This average capacity is called the ergodic capacity of the channel. There are two possibilities for allocating power under ergodic capacity. A short-term power constraint assumes that the power associated with each channel realization must equal the average power constraint P. In this case the ergodic capacity becomes

$$C = \mathrm{E}_{\mathbf{H}}\left[\max_{\mathbf{R_x}:\mathrm{Tr}(\mathbf{R_x})=\rho} B\log_2\det\left[\mathbf{I_{M_r}} + \mathbf{H}\mathbf{R_x}\mathbf{H^H}\right]\right] = \mathrm{E}_{\mathbf{H}}\left[\max_{P_i:\sum_i P_i \le P}\sum_i B\log_2\left(1+\frac{P_i\gamma_i}{P}\right)\right]. \quad (10.16)$$

A less restrictive constraint is a long-term power constraint, where we can use different powers for different channel realizations subject to the average power constraint over all channel realizations. The ergodic capacity under this assumption is given by

$$C = \max_{\rho_H : E[\rho_H] = \rho} \mathrm{E}_{\mathbf{H}} \left[\max_{\mathbf{R_x}:\mathrm{Tr}(\mathbf{R_x}) = \rho_H} B \log_2 \det \left[\mathbf{I_{M_r}} + \mathbf{H R_x H^H} \right] \right] \tag{10.17}$$

The short-term power constraint gives rise to a water-filling in space across the antennas, whereas the long-term power constraint allows for a two-dimensional water-filling across both space and time, similar to the frequency-time water-filling associated with the capacity of a time-varying frequency-selective fading channel.

Channel Unknown at Transmitter: Ergodic Capacity and Capacity with Outage

Consider now a time-varying channel with random matrix $\mathbf{H}$ known at the receiver but not the transmitter. The transmitter assumes a ZMSW distribution for $\mathbf{H}$. The two relevant capacity definitions in this case are ergodic capacity and capacity with outage. Ergodic capacity defines the maximum rate, averaged over all channel realizations, that can be transmitted over the channel for a transmission strategy based only on the distribution of $\mathbf{H}$. This leads to the transmitter optimization problem - i.e., finding the optimum input covariance matrix to maximize ergodic capacity subject to the transmit power constraint. Mathematically, the problem is to characterize the optimum $\mathbf{R_x}$ to maximize

$$C = \max_{\mathbf{R_x}:\mathrm{Tr}(\mathbf{R_x}) = \rho} \mathrm{E}_{\mathbf{H}} \left[B \log_2 \det \left[\mathbf{I}_{M_r} + \mathbf{H R_x H^H} \right] \right], \tag{10.18}$$

where the expectation is with respect to the distribution on the channel matrix $\mathbf{H}$, which for the ZMSW model is i.i.d. zero-mean circularly symmetric unit variance.

As in the case of scalar channels, the optimum input covariance matrix that maximizes ergodic capacity for the ZMSW model is the scaled identity matrix $\mathbf{R_x} = \frac{\rho}{M_t} \mathbf{I}_{M_t}$, i.e. the transmit power is divided equally among all the transmit antennas and independent symbols are sent over the different antennas. Thus the ergodic capacity is given by:

$$C = \mathrm{E}_{\mathbf{H}} \left[B \log_2 \det \left[\mathbf{I}_{M_r} + \frac{\rho}{M_t} \mathbf{H H^H} \right] \right]. \tag{10.19}$$

Since the capacity of the static channel grows as $M = \min(M_T, M_R)$ for M large, this will also be true of the ergodic capacity since it just averages the static channel capacity. Expressions for the growth rate constant can be found in [4] [27]. When the channel is not ZMSW, capacity depends on the distribution of the singular values for the random channel matrix: these distributions and the resulting ergodic capacity in this more general setting are studied in in [13].

The ergodic capacity of a 4×4 MIMO system with i.i.d. complex Gaussian channel gains is shown in Figure 10.4. This figure shows capacity with both transmitter and receiver CSI and with receiver CSI only. There is little difference between the two, and this difference decreases with SNR, which is also the case for a SISO channel. Comparing the capacity of this channel to that of a SISO fading channel shown in Figure 4.7, we see that the MIMO ergodic capacity is 4 times larger than the SISO ergodic capacity, which is as expected since $\min(M_t, M_r) = 4$.

When the channel gain matrix is unknown at the transmitter and the entries are complex Gaussian but not i.i.d. then the channel mean or covariance matrix can be used at the transmitter to increase capacity. The basic idea is to allocate power according to the mean or covariance. This channel model is sometimes referred to as mean or covariance feedback. This model assumes perfect receiver CSI, and the impact of correlated fading depends on what is known at the transmitter: if the transmitter knows the channel realization or doesn't know the realization or the correlation structure than antenna correlation decreases capacity relative to i.i.d. fading. However, if the

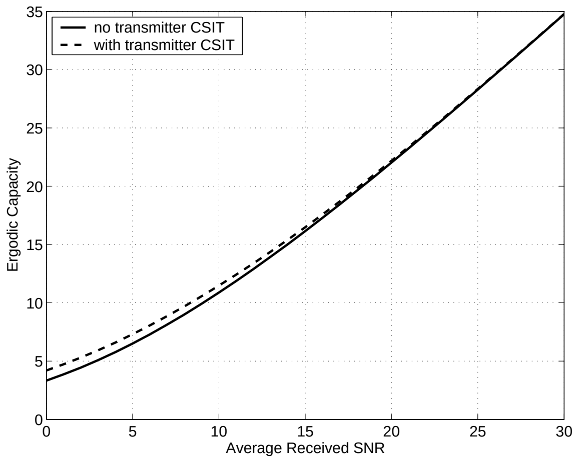


Figure 10.4: Ergodic Capacity of 4×4 MIMO Channel.

transmitter knows the correlation structure than capacity is increased relative to i.i.d. fading. Details on capacity under these different conditions can be found in [28, 25, 26].

Capacity with outage is defined similar to the definition for static channels described in Section 10.3.1, although now capacity with outage applies to a slowly-varying channel where the channel matrix $\mathbf{H}$ is constant over a relatively long transmission time, then changes to a new value. As in the static channel case, the channel realization and corresponding channel capacity is not known at the transmitter, yet the transmitter must still fix a transmission rate to send data over the channel. For any choice of this rate C, there will be an outage probability associated with C, which defines the probability that the transmitted data will not be received correctly. The outage probability is the same as in the static case, given by (10.14). The outage capacity can sometimes be improved by not allocating power to one or more of the transmit antennas, especially when the outage probability is high. [4]. This is because outage capacity depends on the tail of the probability distribution. With fewer antennas, less averaging takes place and the spread of the tail increases.

The capacity with outage of a 4×4 MIMO system with i.i.d. complex Gaussian channel gains is shown in Figure 10.5 for outage of 1% and 10%. We see that the difference in outage capacity for these two outage probabilities increases with SNR. This can be explained from the distribution curves for capacity shown in Figure 10.6. These curves show that at low SNRs, the distribution is very steep, so that the capacity with outage at 1% is very close to that at 10% outage. At higher SNRs the curves become less steep, leading to more of a capacity difference at different outage probabilities.

No CSI at the Transmitter or Receiver

When there is no CSI at either the transmitter or receiver, the linear growth in capacity as a function of the number of transmit and receive antennas disappears, and in some cases adding additional antennas provides negligible capacity gain. Moreover, channel capacity becomes heavily dependent on the underlying channel model, which makes it difficult to make generalizations about capacity growth. For an i.i.d. block fading channel it is shown in [33] that increasing the number of transmit antennas by more than the duration of the block does not increase capacity. Thus, there is no data rate increase beyond a certain number of transmit antennas. However, when fading is correlated, additional transmit antennas do increase capacity [29]. These results were extended in [34] to explicitly characterize capacity and the capacity-achieving transmission strategy for this model in the high SNR regime. Similar results were obtained for a block-Markov fading model in [35]. However, a general analysis in [36] indicates that these results are highly dependent on the structure of the fading process; when this structure is removed

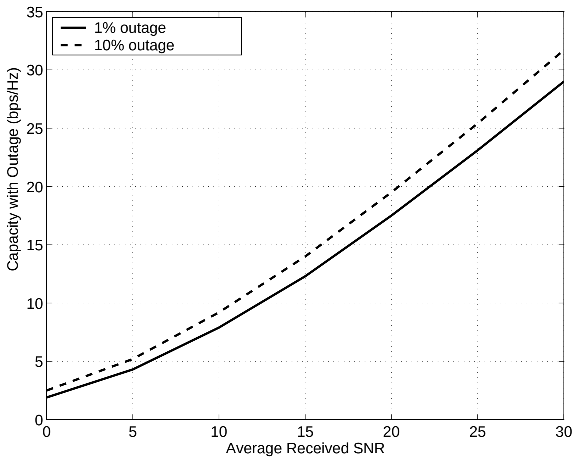


Figure 10.5: Capacity with Outage of a 4×4 MIMO Channel.

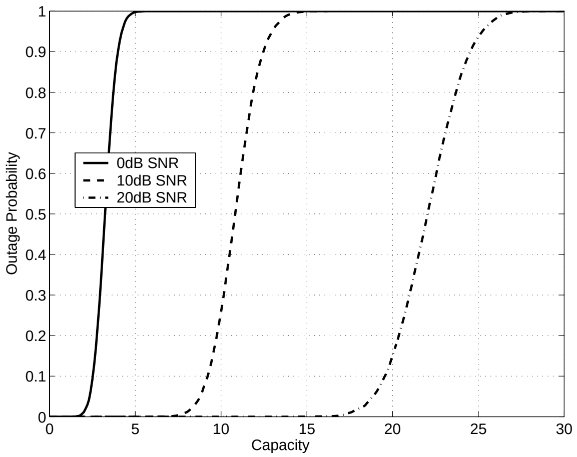


Figure 10.6: Outage Probability Distribution of a 4×4 MIMO Channel.

and a general fading process is considered, in the high SNR regime capacity grows only doubly logarithmically with SNR, and the number of antennas adds at most a constant factor to this growth term. In other words, there is no multiplexing gain associated with multiple antennas when there is no transmitter or receiver CSI.

10.4 MIMO Diversity Gain: Beamforming

The multiple antennas at the transmitter and receiver can be used to obtain diversity gain instead of capacity gain. In this setting, the same symbol, weighted by a complex scale factor, is sent over each transmit antenna, so that the input covariance matrix has unit rank. This scheme is also referred to as **MIMO beamforming**[4]. A beamforming strategy corresponds to the precoding and receiver matrices described in Section 10.2 being just column vectors: $\mathbf{V} = \mathbf{v}$ and $\mathbf{U} = \mathbf{u}$, as shown in Figure 10.7. As indicated in the figure, the transmit symbol x is sent over the ith antenna with weight v_i. On the receive side, the signal received on the ith antenna is weighted by u_i. Both transmit

[4] Unfortunately, beamforming is also used in the smart antenna context of Section 10.8 to describe adjustment of the transmit or receive antenna directivity in a given direction.

and receive weight vectors are normalized so that $||u|| = ||v|| = 1$. The resulting received signal is given by

$$y = \mathbf{u}^*\mathbf{H}\mathbf{v}x + \mathbf{u}^*\mathbf{n}, \tag{10.20}$$

where if $\mathbf{n} = (n_1, \ldots, n_{M_r})$ has i.i.d. elements, the statistics of $\mathbf{u}^*\mathbf{n}$ are the same as the statistics for each of these elements.

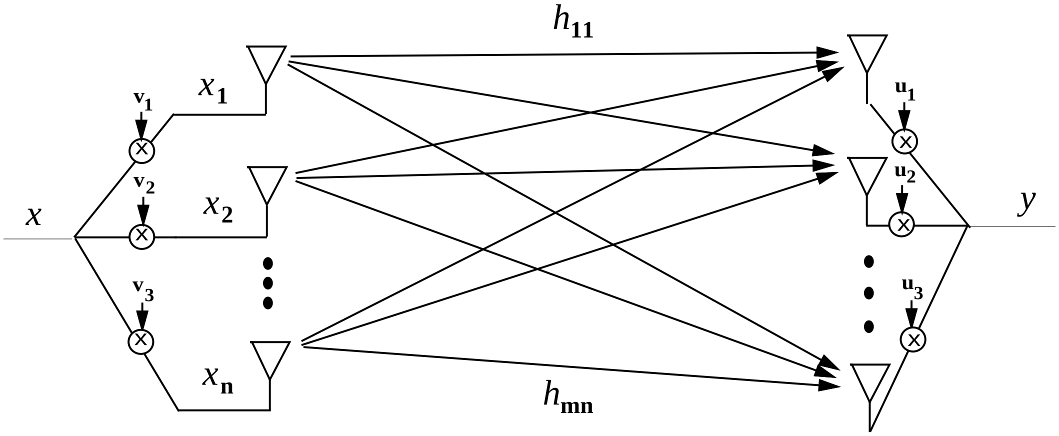


Figure 10.7: MIMO Channel with Beamforming.

Beamforming provides diversity gain by coherent combining of the multiple signal paths. Channel knowledge at the receiver is typically assumed since this is required for coherent combining. The diversity gain then depends on whether or not the channel is known at the transmitter. When the channel matrix $\mathbf{H}$ is known, the received SNR is optimized by choosing $\mathbf{u}$ and $\mathbf{v}$ as the principal left and right singular vectors of the channel matrix $\mathbf{H}$. The corresponding received SNR can be shown to equal $\gamma = \lambda_{max}\rho$, where λ_{max} is the largest eigenvalue of the **Wishart matrix** $\mathbf{W} = \mathbf{H}\mathbf{H}^H$ [21]. The resulting capacity is $C = B\log_2(1 + \lambda_{max}\rho)$, corresponding to the capacity of a SISO channel with channel power gain λ_{max}. When the channel is not known at the transmitter, the transmit antenna weights are all equal, so the received SNR equals $\gamma = ||\mathbf{H}\mathbf{u}^*||$, where $\mathbf{u}$ is chosen to maximize γ. Clearly the lack of transmitter CSI will result in a lower SNR and capacity than with optimal transmit weighting. While beamforming has a reduced capacity relative to optimizing the transmit precoding and receiver shaping matrices, the optimal demodulation complexity with beamforming is of the order of $|\mathcal{X}|$ instead of $|\mathcal{X}|^{\mathcal{R}_\mathcal{H}}$. An even simpler strategy is to use MRC at either the transmitter or receiver and antenna selection on the other end: this was analyzed in [22].

Example 10.3: Consider a MIMO channel with gain matrix

$$\mathbf{H} = \begin{bmatrix} .7 & .9 & .8 \\ .3 & .8 & .2 \\ .1 & .3 & .9 \end{bmatrix}$$

Find the capacity of this channel under beamforming assuming channel knowledge at the transmitter and receiver, $B = 100$ KHz, and $\rho = 10$ dB.

Solution The Wishart matrix for $\mathbf{H}$ is

$$\mathbf{W} = \mathbf{H}\mathbf{H}^H = \begin{bmatrix} 1.94 & 1.09 & 1.06 \\ 1.09 & .77 & .45 \\ 1.06 & .45 & .91 \end{bmatrix}$$

and the largest eigenvalue of this matrix is $\lambda_{max} = 3.17$. Thus, $C = B\log_2(1 + \lambda_{max}\rho) = 10^5 \log_2(1 + 31.7) =$ 503 Kbps.

10.5 Diversity/Multiplexing Tradeoffs

The previous sections suggest two mechanisms for utilizing multiple antennas to improve wireless system performance. One option is to obtain capacity gain by decomposing the MIMO channel into parallel channels and multiplexing different data streams onto these channels. This capacity gain is also referred to as a **multiplexing gain**. However, the SNR associated with each of these channels depends on the singular values of the channel matrix. In capacity analysis this is taken into account by assigning a relatively low rate to these channels. However, practical signaling strategies for these channels will typically have poor performance, unless powerful channel coding techniques are employed. Alternatively, beamforming can be used, where the channel gains are coherently combined to obtain a very robust channel with high diversity gain. It is not necessary to use the antennas purely for multiplexing or diversity. Some of the space-time dimensions can be used for diversity gain, and the remaining dimensions used for multiplexing gain. This gives rise to a fundamental design question in MIMO systems: should the antennas be used for diversity gain, multiplexing gain, or both?

The diversity/multiplexing tradeoff or, more generally, the tradeoff between data rate, probability of error, and complexity for MIMO systems has been extensively studied in the literature, from both a theoretical perspective and in terms of practical space-time code designs [50, 37, 38, 42]. This work has primarily focused on block fading channels with receiver CSI only since when both transmitter and receiver know the channel the tradeoff is relatively straightforward: antenna subsets can first be grouped for diversity gain and then the multiplexing gain corresponds to the new channel with reduced dimension due to the grouping. For the block fading model with receiver CSI only, as the blocklength grows asymptotically large, full diversity gain and full multiplexing gain (in terms of capacity with outage) can be obtained simultaneously with reasonable complexity by encoding diagonally across antennas [51, 52, 2]. An example of this type of encoding is D-BLAST, described in Section 10.6.4. For finite blocklengths it is not possible to achieve full diversity and full multiplexing gain simultaneously, in which case there is a tradeoff between these gains. A simple characterization of this tradeoff is given in [37] for block fading channels with blocklength $T \geq M_t + M_r - 1$ in the limit of asymptotically high SNR. In this analysis a transmission scheme is said to achieve multiplexing gain r and diversity gain d if the data rate (bps) per unit Hertz $R(\text{SNR})$ and probability of error $P_e(\text{SNR})$ as functions of SNR satisfy

$$\lim_{\log_2 \text{SNR}\to\infty} \frac{R(\text{SNR})}{\log_2 \text{SNR}} = r, \tag{10.21}$$

and

$$\lim_{\log \text{SNR}\to\infty} \frac{\log P_e(\text{SNR})}{\log \text{SNR}} = -d, \tag{10.22}$$

where the log in (10.22) can be in any base[5]. For each r the optimal diversity gain $d_{opt}(r)$ is the maximum the diversity gain that can be achieved by any scheme. It is shown in [37] that if the fading blocklength exceeds the total number of antennas at the transmitter and receiver, then

$$d_{opt}(r) = (M_t - r)(M_r - r),\ \ 0 \leq r \leq \min(M_t, M_r). \tag{10.23}$$

[5] The base of the log cancels out of the expression since (10.22) is the ratio of two logs with the same base.

The function (10.23) is plotted in Fig. 10.8. Recall that in Chapter 7 we found that transmitter or receiver diversity with M antennas resulted in an error probability proportional to SNR^{-M}. The formula (10.23) implies that in a MIMO system, if we use all transmit *and* receive antennas for diversity, we get an error probability proportional to $SNR^{-M_tM_r}$ and that, moreover, we can use some of these antennas to increase data rate at the expense of diversity gain.

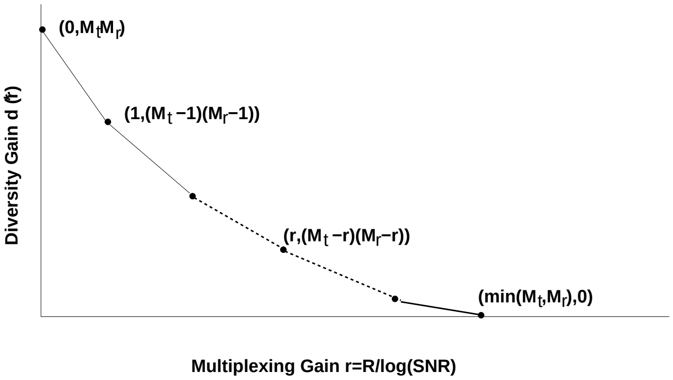


Figure 10.8: Diversity-Multiplexing Tradeoff for High SNR Block Fading.

It is also possible to adapt the diversity and multiplexing gains relative to channel conditions. Specifically, in poor channel states more antennas can be used for diversity gain, whereas in good states more antennas can be used for multiplexing. Adaptive techniques that change antenna use to trade off diversity and multiplexing based on channel conditions have been investigated in [39, 40, 41].

Example 10.4: Let the multiplexing and diversity parameters r and d be as defined in (10.21) and (10.22). Suppose that r and d approximately satisfy the diversity/multiplexing tradeoff $d_{opt}(r) = (M_t - r)(M_r - r)$ at any large finite SNR. For an $M_t = M_r = 8$ MIMO system with an SNR of 15 dB, if we require a data rate per unit Hertz of $R = 15$ bps, what is the maximum diversity gain the system can provide?

Solution: With SNR=15 dB, to get $R = 15$ we require $r\log_2(10^{1.5}) = 15$ which implies $r = 3.01$. Thus, three of the antennas are used for multiplexing and the remaining five for diversity. The maximum diversity gain is then $d_{opt}(r) = (M_t - r)(M_r - r) = (8 - 3)(8 - 3) = 25$.

10.6 Space-Time Modulation and Coding

Since a MIMO channel has input-output relationship $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, the symbol transmitted over the channel each symbol time is a vector rather than a scalar, as in traditional modulation for the SISO channel. Moreover, when the signal design extends over both space (via the multiple antennas) and time (via multiple symbol times), it is typically referred to as a **space-time code**.

Most space-time codes, including all codes discussed in this section, are designed for quasi-static channels where the channel is constant over a block of T symbol times, and the channel is assumed unknown at the transmitter. Under this model the channel input and output become matrices, with dimensions corresponding to space (antennas) and time. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ denote the $M_t \times T$ channel input matrix with ith column $\mathbf{x}_i$ equal to the vector channel input over the ith transmission time. Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ denote the $M_r \times T$ channel output matrix with ith column $\mathbf{y}_i$ equal to the vector channel output over the ith transmission time, and let $\mathbf{N} = [\mathbf{n}_1, \ldots, \mathbf{n}_T]$ denote the $M_r \times T$ noise matrix with ith column $\mathbf{n}_i$ equal to the receiver noise vector on the ith transmission time. With this matrix representation the input-output relationship over all T blocks becomes

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N}. \tag{10.24}$$

10.6.1 ML Detection and Pairwise Error Probability

Assume a space-time code where the receiver has knowledge of the channel matrix $\mathbf{H}$. Under ML detection it can be shown using similar techniques as in the scalar (Chapter 5) or vector (Chapter 8) case that given received matrix $\mathbf{Y}$, the ML transmit matrix $\hat{\mathbf{X}}$ satisfies

$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X} \in \mathcal{X}^{M_t \times T}} ||\mathbf{Y} - \mathbf{H}\mathbf{X}||_F^2 = \arg \min_{\mathbf{X} \in \mathcal{X}^{M_t \times T}} \sum_{i=1}^{T} ||\mathbf{y}_i - \mathbf{H}\mathbf{x}_i||_F^2, \tag{10.25}$$

where $||A||_F$ denotes the Frobenius norm[6] of the matrix A and the minimization is taken over all possible space-time input matrices $\mathcal{X}^T$. The pairwise error probability for mistaking a transmit matrix $\mathbf{X}$ for another matrix $\hat{\mathbf{X}}$, denoted as $p(\hat{\mathbf{X}} \to \mathbf{X})$, depends only on the distance between the two matrices after transmission through the channel and the noise power, i.e.

$$p(\hat{\mathbf{X}} \to \mathbf{X}) = Q\left(\sqrt{\frac{||\mathbf{H}(\mathbf{X} - \hat{\mathbf{X}})||_F^2}{2\sigma_n^2}}\right). \tag{10.26}$$

Let $\mathbf{D}_{\mathbf{X}} = \mathbf{X} - \hat{\mathbf{X}}$ denote the difference matrix between $\mathbf{X}$ and $\hat{\mathbf{X}}$. Applying the Chernoff bound to (10.26) yields

$$p(\hat{\mathbf{X}} \to \mathbf{X}) \leq \exp\left(-\frac{||\mathbf{H}\mathbf{D}_{\mathbf{X}}||_F^2}{4\sigma_n^2}\right). \tag{10.27}$$

Let $\mathbf{h}_i$ denote the ith row of $\mathbf{H}, i = 1, \ldots, M_r$. Then

$$||\mathbf{H}\mathbf{D}_{\mathbf{X}}||_F^2 = \sum_{i=1}^{M_r} \mathbf{h}_i D_X D_X^H \mathbf{h}_i^H. \tag{10.28}$$

Let $\mathcal{H} = \text{vec}(H^T)^T$ where vec($\mathbf{A}$) is defined as the vector that results from stacking the columns of matrix $\mathbf{A}$ on top of each other to form a vector[7]. So $\mathcal{H}^T$ is a vector of length $M_r M_t$. Also define $\mathcal{D}_X = I_{M_r} \otimes \mathbf{D}_{\mathbf{X}}$, where $\otimes$ denotes the Kronecker product. With these definitions,

$$||\mathbf{H}\mathbf{D}_{\mathbf{X}}||_F^2 = ||\mathcal{D}_X^H \mathcal{H}^H \mathcal{H} \mathcal{D}_X||_F^2. \tag{10.29}$$

Substituting (10.29) into (10.27) and taking the expectation relative to all possible channel realizations yields

$$p(\mathbf{X} \to \hat{\mathbf{X}}) \leq \left(\frac{1}{\det\left[\mathbf{I}_{M_t M_r} + E\left(\mathcal{D}_X^H \mathcal{H}^H \mathcal{H} \mathcal{D}_X\right)\right]}\right)^{M_r}. \tag{10.30}$$

[6] The Frobenious norm of a matrix is the square root of the sum of the square of its elements.

[7] So for the $M \times N$ matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$, where $\mathbf{a}_i$ is a vector of length M, vec(A) $= [\mathbf{a}_1^T, \ldots, \mathbf{a}_N^T]^T$ is a vector of length MN.

Suppose that the channel matrix $\mathbf{H}$ is random and spatially white, so that its entries are i.i.d. zero-mean unit variance complex Gaussian random variables. Then taking the expectation yields

$$p(\mathbf{X} \to \hat{\mathbf{X}}) \leq \left(\frac{1}{\det \left[\mathbf{I}_{M_t} + \boldsymbol{\Delta} \right]} \right)^{M_r}, \tag{10.31}$$

where $\boldsymbol{\Delta} = \mathbf{D}_X \mathbf{D}_X^H$. This simplifies to

$$p(\mathbf{X} \to \hat{\mathbf{X}}) \leq \prod_{k=1}^{N_\Delta} \left(\frac{1}{1 + \gamma \lambda_k(\boldsymbol{\Delta})/(4M_t)} \right)^{M_r}, \tag{10.32}$$

where $\gamma = E_s/\sigma_n^2$ is the SNR per input symbol $\mathbf{x}$ or, equivalently, γ/M_t is the SNR per antenna and $\lambda_k(\boldsymbol{\Delta})$ is the kth nonzero eigenvalue of $\boldsymbol{\Delta}$, $k = 1, \ldots, N_\Delta$, where N_Δ is the rank of $\boldsymbol{\Delta}$. In the high SNR regime, i.e. for $\gamma >> 1$, this simplifies to

$$p(\mathbf{X} \to \hat{\mathbf{X}}) \leq \frac{1}{\left(\prod_{k=1}^{N_\Delta} \lambda_k(\boldsymbol{\Delta}) \right)^{M_r}} \left(\frac{\gamma}{4M_t}^{-N_\Delta M_r} \right). \tag{10.33}$$

This equation gives rise to the main criteria for design of space-time codes, described in the next section.

10.6.2 Rank and Determinant Criterion

The pairwise error probability in (10.33) indicates that the probability of error decreases as γ^{-d} for $d = N_\Delta M_r$. Thus, $N_\Delta M_r$ is the diversity gain of the space-time code. The maximum diversity gain possible through coherent combining of M_t transmit and M_r receive antennas is $M_t M_r$. Thus, to obtain this maximum diversity gain, the space-time code must be designed such that the $M_t \times M_t$ difference matrix $\boldsymbol{\Delta}$ between any two code words has full rank equal to M_t. This design criterion is referred to as the **rank criterion**.

The coding gain associated with the pairwise error probability in (10.33) depends on the first term $\left(\prod_{k=1}^{N_\Delta} \lambda_k(\boldsymbol{\Delta}) \right)^{M_r}$. Thus, a high coding gain is achieved by maximizing the minimum of the determinant of $\boldsymbol{\Delta}$ over all input matrix pairs $\mathbf{X}$ and $\hat{\mathbf{X}}$. This criterion is referred to as the **determinant criterion**.

The rank and determinant criteria were first developed in [43, 50, 44]. These criteria are based on the pairwise error probability associated with different transmit signal matrices, rather than the binary domain of traditional codes, and hence often require computer searches to find good codes [45, 46]. A general binary rank criteria was developed in [47] to provide a better construction method for space-time codes.

10.6.3 Space-Time Trellis and Block Codes

The rank and determinant criteria have been primarily applied to the design of space-time trellis codes (STTCs). STTCs are an extension of conventional trellis codes to MIMO systems [10, 44]. They are described using a trellis and decoded using ML sequence estimation via the Viterbi algorithm. STTCs can extract excellent diversity and coding gain, but the complexity of decoding increases exponentially with the diversity level and transmission rate [48]. Space-time block codes (STBCs) are an alternative space-time code that can also extract excellent diversity and coding gain with linear receiver complexity. Interest in STBCs were initiated by the Alamouti code described in Section **??**, which obtains full diversity order with linear receiver processing for a two-antenna transmit system. This scheme was generalized in [49] to STBCs that achieve full diversity order with an arbitrary number of transmit antennas. However, while these codes achieve full diversity order, they do not provide coding gain, and thus have inferior performance to STTCs, which achieve both full diversity gain as well as coding gain. Added coding gain for both STTCs and STBCs can be achieved by concatenating these codes either in serial or in parallel with an outer channel code to form a turbo code [30, 32]. The linear complexity of the STBC designs in [49] result from making the codes orthogonal along each dimension of the code matrix. A similar design premise is used in [53] to design **unitary space-time modulation** schemes for block fading channels when neither the transmitter nor the receiver have channel CSI. More comprehensive treatments of space-time coding can be found in [10, 54, 55, 48] and the references therein.

10.6.4 Spatial Multiplexing and BLAST Architectures

The basic premise of spatial multiplexing is to send M_t independent symbols per symbol period using the dimensions of space and time. In order to get full diversity order an encoded bit stream must be transmitted over all M_t transmit antennas. This can be done through a serial encoding, illustrated in Figure 10.10. With serial encoding the bit stream is temporally encoded over the channel blocklength T, interleaved, and mapped to a constellation point, then demultiplexed onto the different antennas. If each codeword is sufficiently long, it can be transmitted over all M_t transmit antennas and received by all M_r receive antennas, resulting in full diversity gain. However, the codeword length T required to achieve this full diversity is M_tM_r, and decoding complexity grows exponentially with this codeword length. This high level of complexity makes serial encoding impractical.

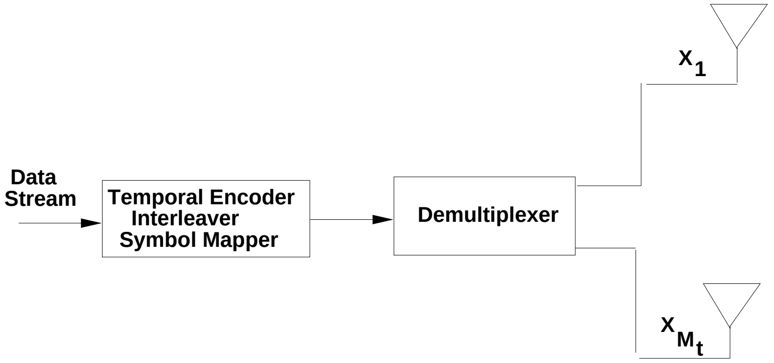


Figure 10.9: Spatial Multiplexing with Serial Encoding.

A simpler method to achieve spatial multiplexing, pioneered at Bell Laboratories as one of the Bell Labs Layered Space Time (BLAST) architectures for MIMO channels [2], is parallel encoding, illustrated in Figure 10.10. With parallel encoding the data stream is demultiplexed into M_t independent streams. Each of the resulting substreams is passed through a SISO temporal encoder with blocklenth T, interleaved, mapped to a signal constellation point, and transmitted over its corresponding transmit antenna. This process can be considered to be the encoding of the serial data into a vertical vector, and hence is also referred to as vertical encoding or V-BLAST [56]. Vertical encoding can achieve at most a diversity order of M_r, since each coded symbol is transmitted from one antenna and received by M_r antennas. This system has a simple encoding complexity that is linear in the number of antennas. However, optimal decoding still requires joint detection of the codewords from each of the transmit antennas, since all transmitted symbols are received by all the receive antennas. It was shown in [57] that the receiver complexity can be significantly reduced through the use of symbol interference cancellation, as shown in Figure 10.11. The symbol interference cancellation, which exploits the synchronicity of the symbols transmitted from each antenna, works as follows. First the M_t transmitted symbols are ordered in terms of their received SNR. An estimate of the received symbol with the highest SNR is made while treating all other symbols as noise. This estimated symbol is subtracted out, and the symbol with the next highest SNR estimated while treating the remaining symbols as noise. This process repeats until all M_t transmitted symbols have been estimated. After cancelling out interfering symbols, the coded substream associated with each transmit antenna can be individually decoded, resulting in a receiver complexity that is linear in the number of transmit antennas. In fact, coding is not even needed with this architecture, and data rates of 20-40 bps/Hz with reasonable error rates were reported in [56] using uncoded V-BLAST.

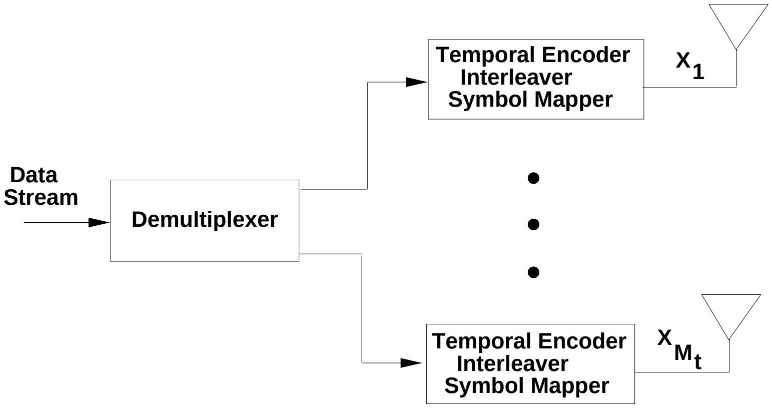

Data Stream
Demultiplexer
Temporal Encoder Interleaver Symbol Mapper
X_1
Temporal Encoder Interleaver Symbol Mapper
X_{M_t}

Figure 10.10: Spatial Multiplexing with Parallel Encoding: VBLAST.

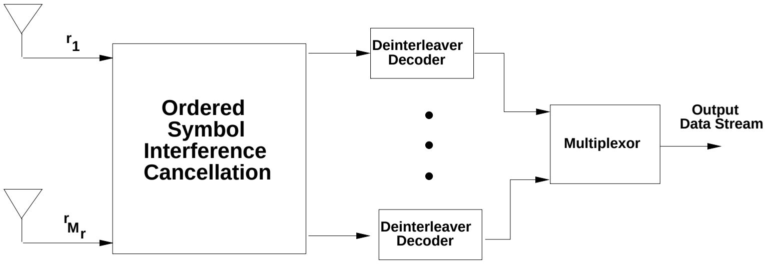


Figure 10.11: VBLAST Receiver with Linear Complexity.

The simplicity of parallel encoding and the diversity benefits of serial encoding can be obtained using a creative combination of the two techniques called diagonal encoding or D-BLAST [2], illustrated in Figure 10.12. In D-BLAST, the data stream is first horizontally encoded. However, rather than transmitting the independent codewords on separate antennas, the codeword symbols are rotated across antennas, so that a codeword is spread over all M_t antennas. The operation of the stream rotation is shown in Figure 10.13. Suppose the ith encoder generates the codeword $\mathbf{x}_i = x_{i1}, \ldots, x_{iM_t}$. The stream rotator transmits each coded symbol on a different antenna, so x_{i1} is sent on antenna 1, x_{i2} is sent on antenna 2, and so forth. If the code blocklength T exceeds M_t then the rotation begins again on the 1st atnenna. As a result, the codeword is spread across all spatial dimensions. Transmission schemes based on D-BLAST can achieve the full M_tM_r diversity gain if the temporal coding with stream rotation is capacity-achieving (Gaussian code books with infinite block size T) [10, Chapter 6.3.5]. Moreover, the D-BLAST system can achieve the maximum capacity with outage if the wasted space-time dimensions along the diagonals are neglected [10, Chapter 12.4.1]. Receiver complexity is also linear in the number of transmit antennas, since the receiver decodes each diagonal code independently. However, this simplicity comes as a price, as the efficiency loss of the wasted space-time dimensions illustrated in Figure 10.12 can be large if the frame size is not appropriately chosen.

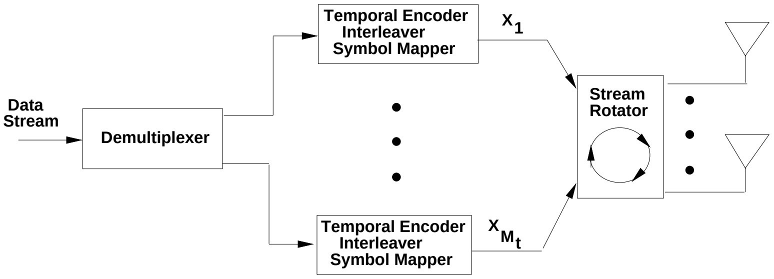


Figure 10.12: Diagonal Encoding with Stream Rotation.

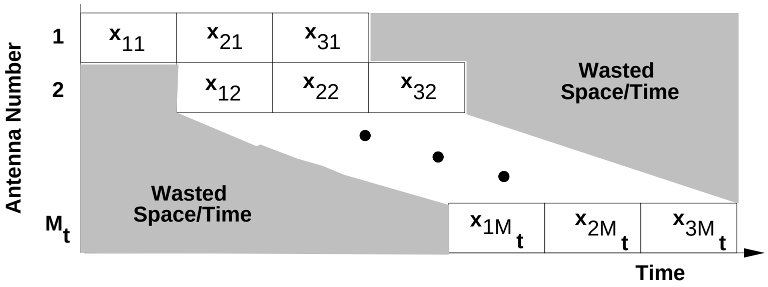


Figure 10.13: Stream Rotation.

10.7 Frequency-Selective MIMO Channels

When the MIMO channel bandwidth is large relative to the channel's multipath delay spread, the channel suffers from ISI, similar to the case of SISO channels. There are two approaches to dealing with ISI in MIMO channels. A channel equalizer can be used to mitigate the effects of ISI. However, the equalizer is much more complex in MIMO channels since the channel must be equalized over both space and time. Moreover, when the equalizer is used in conjuction with a space-time code, the nonlinear and noncausal nature of the code further complicates the equalizer design. In some cases the structure of the code can be used to convert the MIMO equalization problem to a SISO problem for which well-established SISO equalizer designs can be used [58, 59, 60].

An alternative to equalization in frequency-selective fading is multicarrier modulation or orthogonal frequency division multiplexing (OFDM). OFDM techniques for SISO channels are described in Chapter 12: the main premise is to convert the wideband channel into a set of narrowband subchannels that only exhibit flat-fading. Applying OFDM to MIMO channels results in a set of narrowband MIMO channels, and the space-time modulation and coding techniques described above for a single MIMO channel are applied to the parallel set. MIMO frequency-selective fading channels exhibit diversity across space, time, and frequency, so ideally all three dimensions should fully exploited in the signaling scheme.

10.8 Smart Antennas

We have seen that multiple antennas at the transmitter and/or receiver can provide diversity gain as well as increased data rates through space-time signal processing. Alternatively, sectorization or phased array techniques can be used to provide directional antenna gain at the transmit or receive antenna array. This directionality can increase the signaling range, reduce delay-spread (ISI) and flat-fading, and suppress interference between users. In particular, interference typically arrives at the receiver from different directions. Thus, directional antennas can exploit these differences to null or attenuate interference arriving from given directions, thereby increasing system capacity. The reflected multipath components of the transmitted signal also arrive at the receiver from different directions, and can also be attenuated, thereby reducing ISI and flat-fading. The benefits of directionality that can be obtained with multiple antennas must be weighed against their potential diversity or multiplexing benefits, giving rise to a multiplexing/diversity/directionality tradeoff analysis. Whether it is best to use the multiple antennas to increase data rates through multiplexing, increase robustness to fading through diversity, or reduce ISI and interference through directionality is a complex tradeoff decision that depends on the overall system design.

The most common directive antennas are sectorized or phased (directional) antenna arrays, and the gain patterns for these antennas along with an omnidirectional antenna gain pattern are shown in Figure 10.14. Sectorized antennas are designed to provide high gain across a range of signal arrival angles. Sectorization is commonly used at cellular system base stations to cut down on interference: if different sectors are assigned different frequencies or timeslots, then only users within a sector interfere with each other, thereby reducing the average interference by a factor equal to the number of sectors. For example, Figure 10.14 shows a sectorized antenna with a 120^o beamwidths. A base station could divide its 360^o angular range into three sectors to be covered by three 120^o sectroized antennas, in which case the interference in each sector is reduced by a factor of 3 relative to an omnidirectional base station antenna. The price paid for reduced interference in cellular systems via sectorization is the need for handoff between sectors.

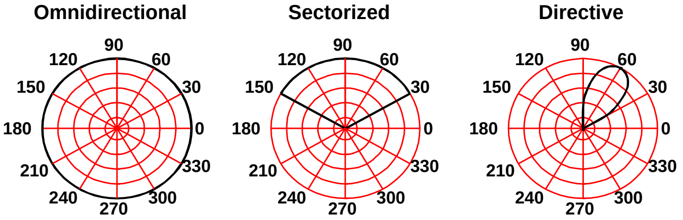


Figure 10.14: Antenna Gains for Omnidirectional, Sectorized, and Directive Antennas.

Directional antennas typically use antenna arrays coupled with phased array techniques to provide directional gain, which can be tightly contolled with sufficiently many antenna elements. Phased array techniques work by adapting the phase of each antenna element in the array, which changes the angular locations of the antenna beams (angles with large gain) and nulls (angles with small gain). For an antenna array with N antennas, N nulls can be formed to significantly reduce the received power of N separate interferers. If there are $N_I < N$ interferers, then the N_I interferers can be cancelled out using N_I antennas in a phased array, and the remaining $N - N_I$ antennas can be used for diversity gain. Note that directional antennas must know the angular location of the desired and interfering signals to provide high or low gains in the appropriate directions. Tracking of user locations can be a significant impediment in highly mobile systems, which is why cellular base stations use sectorization instead of directional antennas.

The complexity of antenna array processing along with the required real estate of an antenna array make the use of smart antennas in small, lightweight, low-power handheld devices unlikely in the near future. However base stations and access points already use antenna arrays in many cases. More details on the technology behind smart antennas and their use in wireless systems can be found in [61].

Bibliography

[1] J. Winters, “On the capacity of radio communication systems with diversity in a rayleigh fading environment,” *IEEE J. Sel. Areas Commun.*, vol. 5, pp. 871–878, June 1987.

[2] G. J. Foschini, “Layered space-time architecture for wireless communication in fading environments when using multi-element antennas,” *Bell Labs Techn. J.*, pp. 41–59, Autumn 1996.

[3] G. J. Foschini and M. Gans, “On limits of wireless communications in a fading environment when using multiple antennas,” *Wireless Pers. Commun.*, vol. 6, pp. 311–355, March 1998.

[4] E. Telatar, “Capacity of multi-antenna gaussian channels,” *AT&T-Bell Labs Internal Memo.*, pp. 585–595, June 1995.

[5] E. Telatar, “Capacity of multi-antenna Gaussian channels,” *European Trans. on Telecomm. ETT*, vol. 10, pp. 585–596, Nov. 1999.

[6] L.H. Brandenburg and A.D. Wyner, “Capacity of the Gaussian channel with memory: the multivariate case,” *Bell System Tech. J.*, Vol. 53, No. 5, pp. 745-778, May-June 1974.

[7] J. Salz and A.D. Wyner, “On data transmission over cross coupled multi-input, multi-output linear channels with applications to mobile radio,” *AT&T MEMO*, 1990.

[8] B. Tsybakov, “The capacity of a memoryless Gaussian vector channel,” *Problems of Information Transmission*, Vol. 1, No. 1, pp.18-29, 1965.

[9] J.L. Holsinger, “Digital communication over fixed time-continuous channels with memory, with special application to telephone channels,” *MIT Res. Lab Elect. Tech. Rep. 430*, 1964.

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

[11] “Random matrix theory and wireless communications,” *Found. Trends Commun. Inform. Theory,* Vol. 1, No. 1, 2004.

[12] A. Lozano, A.M. Tulino, and S. Verdú, “Multiple-antenna capacity in the low-power regime,” *IEEE Trans. Inform. Theory.,* Vol. 49, No. 10, pp. 2527-2544, Oct. 2003.

[13] H. Shin and J.H. Lee, “Capacity of multiple-antenna fading channels: spatial fading correlation, double scattering, and keyhole,” *IEEE Trans. Inform. Theory.*, Vol. 49, No. 10, pp. 2636-2647, Oct. 2003.

[14] V. L. Girko, “A refinement of the central limit theorem for random determinants,” *Theory Probab. Applic.*, Vol. 42, No. 1, pp. 121-129, 1998.

[15] A. Grant, “Rayleigh fading multiple-antenna channels,” *EURASIP J. Appl. Signal Processing (Special Issue on Space-Time Coding (Part I))*, Vol. 2002, No. 3, pp. 316-329, Mar. 2002.

[16] S. Verdú and S. Shamai (Shitz), “Spectral efficiency of CDMA with random spreading,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 622-640, Mar. 1999.

[17] P. J. Smith and M. Shafi, “On a Gaussian approximation to the capacity of wireless MIMO systems,” *Proc. IEEE Int. Conf. Communications (ICC02)*, New York, Apr. 2002, pp. 406-410.

[18] Z.Wang and G. B. Giannakis, “Outage mutual information of space-time MIMOchannels,” *Proc. 40th Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2002, pp. 885-894.

[19] C.-N. Chuah, D. N. C. Tse, J. M. Kahn, and R. A. Valenzuela, “Capacity scaling in MIMO wireless systems under correlated fading,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 637-650, Mar. 2002.

[20] A.L. Moustakas, S.H.Simon, A.M. Sengupta, “MIMO capacity through correlated channels in the presence of correlated interferers and noise: a (not so) large N analysis,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 2545 - 2561, Oct. 2003.

[21] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal Processing Advances in Wireless and Mobile Communications: Trends in Single- and Multi-user Systems*. New York: Prentice Hall PTR, 2001.

[22] A. Molisch, M. Win, and J. H. Winters, “Reduced-complexity transmit/receive-diversity systems,” *IEEE Trans. Signal Proc.*, vol. 51, pp. 2729–2738, November 2003.

[23] A. Narula, M. Lopez, M. Trott, and G. Wornell, “Efficient use of side information in multiple-antenna data transmission over fading channels,” *IEEE J. Select. Areas Commun.*, pp. 1423–1436, Oct. 1998.

[24] E. Visotsky and U. Madhow, “Space-time transmit precoding with imperfect feedback,” *Proc. Intl. Symp. Inform. Theory*, pp. 357–366, June 2000.

[25] S. Jafar and A. Goldsmith, “Transmitter optimization and optimality of beamforming for multiple antenna systems,” *IEEE Trans. Wireless Comm.*, vol. 3, pp. 1165–1175, July 2004.

[26] E. Jorswieck and H. Boche, “Channel capacity and capacity-range of beamforming in MIMO wireless systems under correlated fading with covariance feedback,” *IEEE Trans. Wireless Comm.*, vol. 3, pp. 1543–1553, Sept. 2004.

[27] B. H. T. M. V. Tarokh., “Multiple-antenna channel hardening and its implications for rate feedback and scheduling,” *IEEE Trans. Info. Theory*, vol. 50, pp. 1893–1909, Sept. 2004.

[28] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” *IEEE J. Select. Areas Comm.*, vol. 21, pp. 684–701, June 2003.

[29] S. A. Jafar and A. J. Goldsmith, “Multiple-antenna capacity in correlated Rayleigh fading with channel covariance information”, *IEEE Trans. Wireless Commun.* 2005.

[30] Y. Liu, M.P. Fitz, and O.Y. Takeshita, “Full-rate space-time codes,” *IEEE J. Select. Areas Commun.* Vol. 19, No. 5, pp. 969-980, May 2001.

[31] K.R. Narayanan, “Turno decoding of concatenated space-time codes,” *Proc. Aller. Conf. Commun., Contr., Comp.*, Sept. 1999.

[32] V. Gulati and K.R. Narayanan, “Concatenated codes for fading channels based on recurvisve space-time trellis codes,” *IEEE Trans. Wireless Commun.*, Vol. 2, No. 1, pp. 118-128, Jan. 2003.

[33] T. Marzetta and B. Hochwald, “Capacity of a mobile multiple-antenna communication link in ra yleigh flat fading,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan 1999.

[34] L. Zheng and D. N. Tse, “Communication on the grassmann manifold: A geometric approach to the non-coherent multi-antenna channel,,” *IEEE Trans. Inform. Theory*, vol. 48, pp. 359–383, Feb. 2002.

[35] R. Etkin and D. Tse, “Degrees of freedom in underspread MIMO fading channels,” *Proc. Intl. Symp. Inform. Theory*, p. 323, July 2003.

[36] A. Lapidoth and S. Moser, “On the fading number of multi-antenna systems over flat fading channels with memory and incomplete side information,” *Proc. Intl. Symp. Inform. Theory*, p. 478, July 2002.

[37] L. Zheng and D. N. Tse, “Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 1073–1096, May 2003.

[38] H. Gamal, G. Caire, and M. Damon, “Lattice coding and decoding achieve the optimal diversity-multiplexing tradeoff of MIMO channels,” *IEEE Trans. Inform. Theory*, vol. 50, pp. 968–985, June 2004.

[39] R. W. Heath, Jr.and A. J. Paulraj, “Switching between multiplexing and diversity based on constellation distance,” *Proc. Allerton Conf. Comm. Control and Comp.*, Sept. 30 - Oct. 2, 2000.

[40] R. W. Heath Jr. and D. J. Love, “Multi-mode Antenna Selection for Spatial Multiplexing with Linear Receivers,” *IEEE Trans. on Signal Processing*, 2005.

[41] V. Jungnickel, T. Haustein, V. Pohl, C. Von Helmolt, “Link adaptation in a multi-antenna system,” *Proc. IEEE Vehic. Tech. Conf.*, pp. 862 - 866, April 2003

[42] H. Yao and G. Wornell, “Structured space-time block codes with optimal diversity-multiplexing tradoeff and minimum delay,” in *Proc. IEEE Global Telecomm. Conf*, pp. 1941–1945, Dec. 2003.

[43] J.-C. Guey, M. P. Fitz, M. Bell, and W.-Y. Kuo, “Signal design for transmitter diversity wireless communication systems over rayleigh fading channels,” *IEEE Trans. Commun.*, vol. 47, pp. 527–537, April 1999.

[44] V. Tarokh, A. Naguib, N. Seshadri, and A. Calderbank, “Space-time codes for high data rate wireless communication: performance criteria in the presence of channel estimation errors, mobility, and multiple paths,” *IEEE Trans. Commun.*, vol. 47, pp. 199–207, Feb. 1999.

[45] S. Baro, G. Bauch, and A. Hansman, “Improved codes for space-time trellis coded modulation,” *IEEE Commun. Letts.*, vol. 4, pp. 20–22, Jan 2000.

[46] J. Grimm, M. Fitz, and J. Korgmeier, “Further results in space-time coding for rayleigh fading,” in *Proc. Allerton Conf. Commun. Contrl. Comput.*, pp. 1941–1945, Sept. 1998.

[47] H. Gamal and A. Hammons, “On the design of algebraic space-time codes for MIMO block-fading channels,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 151–163, Jan 2003.

[48] A. Naguib, N. Seshadri, and A. Calderbank, “Increasing data rate over wireless channels,” *IEEE Sign. Proc. Magazine*, vol. 17, pp. 76–92, May 2000.

[49] V. Tarokh, H. Jafarkhani, and A. Calderbank, “Space-time block codes from orthogonal designs,” *IEEE Trans. Inform. Theory.*, vol. 45, pp. 1456–1467, July 1999.

[50] V. Tarokh, N. Seshadri, and A. Calderbank, “Space-time codes for high data rate wireless communication: performance criterion and code construction,” *IEEE Trans. Inform. Theory.*, Vol. 44, No. 2, pp. 744-765, March 1998.

[51] H. El Gamal and M.O. Damen, “Universal space-time coding,” *IEEE Trans. Inform. Theory.*, Vol. 49, No. 5, pp. 1097-1119, May 2003.

[52] M.O. Damen, H. El Gamal, and N. C. Beaulieu, “Linear threaded algebraic space-time constellations,” *IEEE Trans. Inform. Theory.*, Vol. 49, No. 10, pp. 2372-2388, Oct. 2003.

[53] B. Hochwald and T. Marzetta, “Unitary space-time modulation for multiple-antenna communications in rayleigh flat fading,” *IEEE Trans. Info. Theory*, vol. 46, pp. 543–564, March 2000.

[54] E. Larsson and P. Stoica, *Space-Time Block Coding for Wireless Communications*. Cambridge, England: Cambridge University Press, 2003.

[55] D. Gesbert, M. Shafi, D.-S. Shiu, P. Smith, and A. Naguib, “From theory to practice: an overview of MIMO space-time coded wireless systems,” *IEEE J. Select. Areas Commun.*, pp. 281–302, April 2003.

[56] 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 *Proc. URSI Intl. Symp. Sign. Syst. Electr.*, pp. 295–300, Oct. 1998.

[57] G. Foschini, G. Golden, R. Valenzuela, and P. Wolniansky, “Simplified processing for high spectral efficiency wireless communication employing multi-element arrays,” *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1841–1852, Nov. 1999.

[58] C. Fragouli, N. Al-Dhahir, and S. Diggavi, “Pre-filtered space-time m-bcjr equalizer for frequency selective channels,” *IEEE. Trans. Commun.*, vol. 50, pp. 742–753, May 2002.

[59] A. Naguib, “Equalization of transmit diversity space-time coded signals,” in *Proc. IEEE Global Telecomm. Conf*, pp. 1077–1082, Dec. 2000.

[60] G. Bauch and A. Naguib, “Map equalization of space-time coded signals over frequency selective channels,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, pp. 261–265, Sept. 1999.

[61] J. Winters, “Smart antennas for wireless systems,” *IEEE Pers. Comm. Mag.*, vol. 5, pp. 23–27, Feb. 1998.

Chapter 10 Problems

1. Matrix identities are commonly used in the analysis of MIMO channels. Prove the following matrix identities.

   (a) Given an $M \times N$ matrix $\mathbf{A}$ show that the matrix $\mathbf{A}\mathbf{A}^H$ is Hermitian. What does this reveal about the eigendecomposition of $\mathbf{A}\mathbf{A}^H$?

   (b) Show that $\mathbf{A}\mathbf{A}^H$ is positive semidefinite.

   (c) Show that $\mathbf{I}_M + \mathbf{A}\mathbf{A}^{\mathbf{H}}$ is Hermitian positive definite.

   (d) Show that $\det[\mathbf{I}_M + \mathbf{A}\mathbf{A}^{\mathbf{H}}] = \det[\mathbf{I}_{\mathbf{N}} + \mathbf{A}^{\mathbf{H}}\mathbf{A}]$.

2. Find the SVD of the following matrix

$$\mathbf{H} = \begin{bmatrix} .7 & .6 & .2 & .4 \\ .1 & .5 & .9 & .2 \\ .3 & .6 & .9 & .1 \end{bmatrix}$$

3. Find a 3×3 channel matrix $\mathbf{H}$ with 2 nonzero singular

4. Consider the 4×4 MIMO channels given below. What is the maximum multiplexing gain of each, i.e., how many independent scalar data streams can be supported reliably?

$$\mathbf{H}_1 = \begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}$$

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}.$$

5. The capacity of a static MIMO channel with only receiver CSI is given by $C = \sum_{i=1}^{R_H} \log_2 \left(1 + \frac{\lambda_i \rho}{M_t}\right)$. Show that if the sum of singular values is bounded, this expression is maximized when all R_H singular values are equal.

6. Consider a MIMO system with the following channel matrix:

$$H = \begin{bmatrix} .1 & .3 & .4 \\ .3 & .2 & .2 \\ .1 & .3 & .7 \end{bmatrix} = \begin{bmatrix} -.5196 & -.0252 & -.8541 \\ -.3460 & -.9077 & .2372 \\ -.7812 & .4188 & .4629 \end{bmatrix} \begin{bmatrix} .9719 & 0 & 0 \\ 0 & .2619 & 0 \\ 0 & 0 & .0825 \end{bmatrix} \begin{bmatrix} -.2406 & -.4727 & -.8477 \\ -.8894 & -.2423 & .3876 \\ .3886 & -.8472 & .3621 \end{bmatrix}.$$

   Note that H is written in terms of its singular value decomposition (SVD) $H = U\Lambda V$.

   (a) Check if $H = U\Lambda V$. You will see that the matrices U, Λ, and V do not have sufficiently large precision so that $U\Lambda V$ is only approximately equal to H. This indicates the sensitivity of the SVD, in particular the matrix Λ, to small errors in the estimate of the channel matrix H.

(b) Based on the singular value decomposition $H = U\Lambda V$, find an equivalent MIMO system consisting of three independent channels. Find the transmit precoding filter and the receiver shaping filter necessary to transform the original system into the equivalent system.

(c) Find the optimal power allocation $P_i, i = 1, 2, 3$ across the three channels found in part (b), and the corresponding total capacity of the equivalent system, assuming $\overline{P}/\sigma_n^2 = 20$ dB and the system bandwidth $B = 100$ KHz.

(d) Compare the capacity in part (c) to that when the channel is unknown at the transmitter, so equal power is allocated to each antenna.

7. Show using properties of the SVD that for the MIMO channel known at the transmitter and receiver, the general capacity expression

$$C = \max_{\mathbf{R_x}:\mathbf{Tr}(\mathbf{R_x})=\rho} B \log_2 \det \left[\mathbf{I}_{M_r} + \mathbf{H}\mathbf{R_x}\mathbf{H}^H\right].$$

reduces to

$$C = \max_{\rho_i:\sum_i \rho_i \le \rho} \sum_i B \log_2 \left(1 + \lambda_i \rho_i\right),$$

for singular values $\{\sqrt{\lambda_i}\}$ and SNR ρ.

8. For the 4×4 MIMO channels given below, find their capacity per unit Hz assuming both transmitter and receiver know the channel, for channel SNR $\rho = 10$ dB.

$$\mathbf{H}_1 = \begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}$$

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}.$$

9. Assume a ZMCSCG MIMO system with channel matrix $\mathbf{H}$ corresponding to $M_t = M_r = M$ transmit and receive antennas. Show using the law of large numbers that

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{H}\mathbf{H}^H = \mathbf{I}_M.$$

Then use this to show that

$$\lim_{M\to\infty} B \log_2 \det[\mathbf{I}_M + \frac{\rho}{M}\mathbf{H}\mathbf{H}^H] = MB\log_2(1+\rho).$$

10. Plot the ergodic capacities per unit Hz for a ZMCSCG MIMO channel with SNR $0 \le \rho \le 30$ dB for the following MIMO dimensions:

(a) $M_t = M_r = 1$

(b) $M_t = 2$, $M_r = 1$

(c) $M_t = M_r = 2$

(d) $M_t = 2$, $M_r = 3$

(e) $M_t = M_r = 3$

Verify that at high SNRs, capacity grows linearly as $M = \min(M_t, M_r)$.

11. Plot the outage capacities per unit Hz for an outage probability of 1% for a ZMCSCG MIMO channel with SNR $0 \leq \rho \leq 30$ dB for the following MIMO dimensions:

    (a) $M_t = M_r = 1$

    (b) $M_t = 2$, $M_r = 1$

    (c) $M_t = M_r = 2$

    (d) $M_t = 2$, $M_r = 3$

    (e) $M_t = M_r = 3$

    Verify that at high SNRs, capacity grows linearly as $M = \min(M_t, M_r)$.

12. Show that if the noise vector $\mathbf{n} = (n_1, \ldots, n_{M_r})$ has i.i.d. elements then, for $||u|| = 1$, the statistics of $\mathbf{u}^*\mathbf{n}$ are the same as the statistics for each of these elements.

13. Consider a MIMO system where the channel gain matrix H is known at the transmitter and receiver. Show that if transmit and receive antennas are used for diversity, the optimal weights at the transmitter and receiver lead to an SNR of $\gamma = \lambda_{max}\rho$, where ρ is the largest eigenvalue of HH^H.

14. Consider a channel with channel matrix

$$H = \begin{bmatrix} .1 & .5 & .9 \\ .3 & .2 & .6 \\ .1 & .3 & .7 \end{bmatrix}.$$

    Assuming $\rho = 10$ dB, find the output SNR when beamforming is used on the channel with equal weights on each transmit antenna and optimal weighting at the receiver. Compare with the SNR under beamforming with optimal weights at both the transmitter and receiver.

15. Consider an 8×4 MIMO system. Assume a coding scheme that can achieve the rate/diversity tradeoff $d(r) = (M_t - r)(M_r - r)$.

    (a) What is the maximum multiplexing rate for this channel given a required $P_e = \rho^{-d} \leq 10^{-3}$, assuming $\rho = 10$ dB?

    (b) Given the r in part (a), what is the resulting P_e?

16. Find the capacity of a SIMO channel with channel gain vector $\mathbf{h} = [.1 \ .4 \ .75 \ .9]$, optimal receiver weighting, and $\rho = 10$ dB.

17. Consider a 2x2 MIMO system with channel gain matrix $\mathbf{H}$ given by

$$\mathbf{H} = \begin{bmatrix} .3 & .5 \\ .7 & .2 \end{bmatrix}.$$

    Assume $\mathbf{H}$ is known at both the transmitter and receiver, and that there is a total transmit power of $P = 10$ mW across the two transmit antennas, AWGN with power $N_0 = 10^{-9}$ W/Hz at each receive antenna, and bandwidth $B = 100$ KHz.

(a) Find the SVD for $\mathbf{H}$.

(b) Find the capacity of this channel.

(c) Assuming transmit precoding and receiver shaping is used to transform this channel into two parallel independent channels with a total power constraint P. Find the maximum data rate that can be transmitted over this parallel set assuming MQAM modulation on each channel with optimal power adaptation across the channels subject to power constraint P. Assume a target BER of 10^{-3} on each channel, the BER is bounded by $\leq .2e^{-1.5\gamma/(M-1)}$, and the constellation size of the MQAM is unrestricted.

(d) Suppose now that the antennas at the transmitter and receiver are all used for diversity with optimal weighting at the transmitter and receiver to maximize the SNR of the combiner output. Find the SNR of the combiner output, and the BER of a BPSK modulated signal transmitted over this diversity system. Compare the data rate and BER of this BPSK signaling with diversity (assuming $B = 1/T_b$) to the rate and BER from part (b).

(e) Comment on the diversity/multiplexing tradeoffs between the systems in parts (b) and (c).

18. Consider an $M \times M$ MIMO channel with ZMCSCG channel gains.

    (a) Plot the ergodic capacity per unit Hz of this channel for $M = 1$ and $M = 4$ with $0 \leq \rho \leq 20$ dB assuming both transmitter and receiver have channel CSI.

    (b) Repeat part (a) assuming only the receiver has transmitter CSI.

19. Find the outage capacity for a 4×4 MIMO channel with ZMCSCG elements at 10% outage for $\rho = 10$ dB.

20. Plot the CDF of capacity for a $M \times M$ MIMO channel with $\rho = 10$ dB assuming no transmitter knowledge for $M = 4, 6, 8$. What happens as M increases? What are the implications of this behavior in a practical system design?

Chapter 12

Multicarrier Modulation

The basic idea of multicarrier modulation is to divide the transmitted bitstream into many different substreams and send these over many different subchannels. Typically the subchannels are orthogonal under ideal propagation conditions. The data rate on each of the subchannels is much less than the total data rate, and the corresponding subchannel bandwidth is much less than the total system bandwidth. The number of substreams is chosen to insure that each subchannel has a bandwidth less than the coherence bandwidth of the channel, so the subchannels experience relatively flat fading. Thus, the ISI on each subchannel is small. The subchannels in multicarrier modulation need not be contiguous, so a large continuous block of spectrum is not needed for high rate multicarrier communications. Moreover, multicarrier modulation is efficiently implemented digitally. In this discrete implementation, called orthogonal frequency division multiplexing (OFDM), the ISI can be completely eliminated through the use of a cyclic prefix.

Multicarrier modulation is currently used in many wireless systems. However, it is not a new technique: it was first used for military HF radios in the late 1950's and early 1960's. Starting around 1990 [1], multicarrier modulation has been used in many diverse wired and wireless applications, including digital audio and video broadcasting in Europe [3], digital subscriber lines (DSL) using discrete multitone [5, 12], and the most recent generation of wireless LANs [26, 28]. There are also a number of newly emerging uses for multicarrier techniques, including fixed wireless broadband services [27, 14], mobile wireless broadband known as FLASH-OFDM [13], and even for ultrawideband radios, where multiband OFDM is one of the two competing proposals for the IEEE 802.15 ultrawideband standard. Multicarrier modulation is also a candidate for the air interface in next generation cellular systems [18, 32].

The multicarrier technique can be implemented in multiple ways, including vector coding [17] and OFDM [7], all of which are discussed in this chapter. These techniques have subtle differences, but are all based on the same premise of breaking a wideband channel into multiple parallel narrowband channels by means of an orthogonal channel partition.

There is some debate as to whether multicarrier or single carrier modulation is better for ISI channels with delay spreads on the order of the symbol time. It is claimed in [3] that for some mobile radio applications, single carrier with equalization has roughly the same performance as multicarrier modulation with channel coding, frequency-domain interleaving, and weighted maximum-likelihood decoding. Adaptive loading was not taken into account in [3], which has the potential to significantly improve multicarrier performance [8]. But there are other problems with multicarrier modulation that impair its performance, most significantly frequency offset and timing jitter, which degrade the orthogonality of the subchannels. In addition, the peak-to-average power ratio of multicarrier is significantly higher than that of single carrier systems, which is a serious problem when nonlinear amplifiers are used. Tradeoffs between multicarrier and single carrier block transmission systems with respect to these impairments are discussed in [9].

Despite these challenges, multicarrier techniques are common in high data rate wireless systems with moderate to large delay spread, as they have significant advantages over time-domain equalization. In particular, the number of taps required for an equalizer with good performance in a high data rate system is typically large. Thus, these equalizers are highly complex. Moreover, it is difficult to maintain accurate weights for a large number of equalizer taps in a rapidly varying channel. For these reasons, most emerging high rate wireless systems use either multicarrier modulation or spread spectrum instead of equalization to compensate for ISI.

12.1 Data Transmission using Multiple Carriers

The simplest form of multicarrier modulation divides the data stream into multiple substreams to be transmitted over different orthogonal subchannels centered at different subcarrier frequencies. The number of substreams is chosen to make the symbol time on each substream much greater than the delay spread of the channel or, equivalently, to make the substream bandwidth less than the channel coherence bandwidth. This insures that the substreams will not experience significant ISI.

Consider a linearly-modulated system with data rate R and passband bandwidth B. The coherence bandwidth for the channel is assumed to be $B_c < B$, so the signal experiences frequency-selective fading. The basic premise of multicarrier modulation is to break this wideband system into N linearly-modulated subsystems in parallel, each with subchannel bandwidth $B_N = B/N$ and data rate $R_N \approx R/N$. For N sufficiently large, the subchannel bandwidth $B_N = B/N << B_c$, which insures relatively flat fading on each subchannel. This can also be seen in the time domain: the symbol time T_N of the modulated signal in each subchannel is proportional to the subchannel bandwidth $1/B_N$. So $B_N << B_c$ implies that $T_N \approx 1/B_N >> 1/B_c \approx T_m$, where T_m denotes the delay spread of the channel. Thus, if N is sufficiently large, the symbol time is much bigger than the delay spread, so each subchannel experiences little ISI degradation.

Figure 12.1 illustrates a multicarrier transmitter[1]. The bit stream is divided into N substreams via a serial-to-parallel converter. The nth substream is linearly-modulated (typically via QAM or PSK) relative to the subcarrier frequency f_n and occupies passband bandwidth B_N. We assume coherent demodulation of the subcarriers so the subcarrier phase is neglected in our analysis. If we assume raised cosine pulses for $g(t)$ we get a symbol time $T_N = (1+\beta)/B_N$ for each substream, where β is the rolloff factor of the pulse shape. The modulated signals associated with all the subchannels are summed together to form the transmitted signal, given as

$$s(t) = \sum_{i=0}^{N-1} s_i g(t) \cos(2\pi f_i t + \phi_i), \tag{12.1}$$

where s_i is the complex symbol associated with the ith subcarrier and ϕ_i is the phase offset of the ith carrier. For nonoverlapping subchannels we set $f_i = f_0 + i(B_N), i = 0, \ldots, N-1$. The substreams then occupy orthogonal subchannels with passband bandwidth B_N, yielding a total passband bandwidth $NB_N = B$ and data rate $NR_N \approx R$. Thus, this form of multicarrier modulation does not change the data rate or signal bandwidth relative to the original system, but it almost completely eliminates ISI for $B_N << B_c$.

The receiver for this multicarrier modulation is shown in Figure 12.2. Each substream is passed through a narrowband filter to remove the other substreams, demodulated, and combined via a parallel-to-serial converter to form the original data stream. Note that the ith subchannel will be affected by flat fading corresponding to a channel gain $\alpha_i = |H(f_i)|$.

Although this simple type of multicarrier modulation is easy to understand, it has several significant shortcomings. First, in a realistic implementation, subchannels will occupy a larger bandwidth than under ideal raised

[1] In practice the complex symbol s_i would have its real part transmitted over the in-phase signaling branch and its imaginary part transmitted over the quadrature signaling branch. For simplicity we illustrate multicarrier based on sending a complex symbol along the in-phase signaling branch.

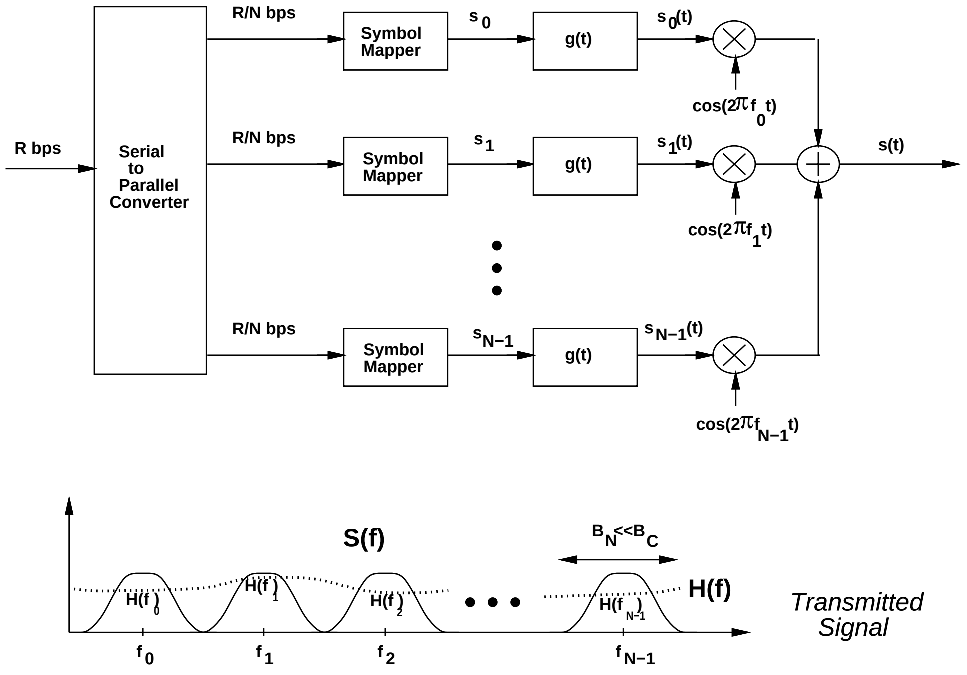


Figure 12.1: Multicarrier Transmitter.

cosine pulse shaping since the pulse shape must be time-limited. Let ϵ/T_N denote the additional bandwidth required due to time-limiting of these pulse shapes. The subchannels must then be separated by $(1+\beta+\epsilon)/T_N$, and since the multicarrier system has N subchannels, the bandwidth penalty for time limiting is $\epsilon N/T_N$. In particular, the total required bandwidth for nonoverlapping subchannels is

$$B = \frac{N(1+\beta+\epsilon)}{T_N}. \tag{12.2}$$

Thus, this form of multicarrier modulation can be spectrally inefficient. Additionally, near-ideal (and hence, expensive) low pass filters will be required to maintain the orthogonality of the subcarriers at the receiver. Perhaps most importantly, this scheme requires N independent modulators and demodulators, which entails significant expense, size, and power consumption. The next section presents a modulation method that allows subcarriers to overlap and removes the need for tight filtering. Section 12.4 presents the discrete implementation of multicarrier modulation, which eliminates the need for multiple modulators and demodulators.

Example 12.1: Consider a multicarrier system with a total passband bandwidth of 1 MHz. Suppose the system operates in a city with channel delay spread $T_m = 20\mu$s. How many subchannels are needed to obtain approximately flat-fading in each subchannel.

Solution: The channel coherence bandwidth is $B_c = 1/T_m = 1/.00002 = 50$ KHz. To insure flat-fading on each subchannel, we take $B_N = B/N = .1B_c << B_c$. Thus, $N = B/.1B_c = 1000000/5000 = 200$ subchannels are needed to insure flat-fading on each subchannel. In discrete implementations of multicarrier N must be a power of two for the DFT and IDFT operations, in which case $N = 256$ for this set of parameters.

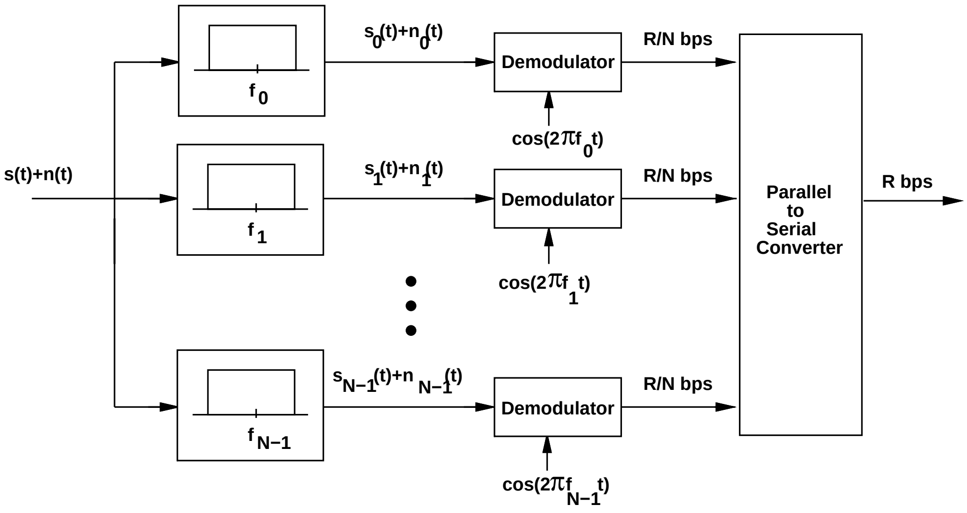


Figure 12.2: Multicarrier Receiver.

Example 12.2: Consider a multicarrier system with $T_N = .2$ ms: $T_N >> T_m$ for T_m the channel delay spread, so each subchannel experiences minimal ISI. Assume the system has $N = 128$ subchannels. If raised cosine pulses with $\beta = 1$ are used, and the additional bandwidth due to time limiting required to insure minimal power outside the signal bandwidth is $\epsilon/T_N = .1$, then what is the total bandwidth of the system?

Solution: From (12.2),

$$B = \frac{N(1+\beta+\epsilon)}{T_N} = \frac{128(1+1+.1)}{.0002} = 1.344 \text{ MHz}.$$

We will see in the next section that the bandwidth requirements for this system can be substantially reduced by overlapping subchannels.

12.2 Multicarrier Modulation with Overlapping Subchannels

We can improve on the spectral efficiency of multicarrier modulation by overlapping the subchannels. The subcarriers must still be orthogonal so that they can be separated out by the demodulator in the receiver. The subcarriers $\{\cos(2\pi(f_0 + i/T_N) + \phi_i), i = 0, 1, 2 \ldots\}$ form a set of (approximately) orthogonal basis functions on the interval $[0, T_N]$ for any set of subcarrier phase offsets $\{\phi_i\}$ since

$$\begin{aligned}
&\int_0^{T_N} \cos(2\pi(f_0 + i/T_N)t + \phi_i)\cos(2\pi(f_0 + j/T_N)t + \phi_j)dt \\
&= \int_0^{T_N} .5\cos(2\pi(i-j)t/T_N + \phi_i - \phi_j)dt + \int_0^{T_N} .5\cos(2\pi(2f_0 + i + j)t/T_n + \phi_i + \phi_j)dt \quad (12.3)\\
&\approx \int_0^{T_N} .5\cos(2\pi(i-j)t/T_N + \phi_i - \phi_j)dt \\
&= .5T_N\delta(i-j),
\end{aligned}$$

where the approximation follows from that fact that the second integral in (12.3) is approximately zero for $f_0 T_N >> 1$. Moreover, it is easily shown that no set of subcarriers with a smaller frequency separation forms an orthogonal set on $[0, T_N]$ for arbitrary subcarrier phase offsets. This implies that the minimum frequency separation required for subcarriers to remain orthogonal over the symbol interval $[0, T_N]$ is $1/T_N$. Since the carriers are orthogonal, from Chapter 5.1 the set of functions $\{g(t)\cos(2\pi(f_0 + i/T_N)t + \phi_i), i = 0, 1, \ldots N-1\}$ also form a set of (approximately) orthonormal basis functions for appropriately chosen baseband pulse shapes $g(t)$: the family of raised cosine pulses are a common choice for this pulse shape. Given this orthonormal basis set, even if the subchannels overlap, the modulated signals transmitted in each subchannel can be separated out in the receiver, as we now show.

Consider a multicarrier system where each subchannel is modulated using raised cosine pulse shapes with rolloff factor β. The passband bandwidth of each subchannel is then $B_N = (1+\beta)/T_N$. The ith subcarrier frequency is set to $(f_0 + i/T_N)$, $i = 0, 1 \ldots N-1$ for some f_0, so the subcarriers are separated by $1/T_N$. However, the passband bandwidth of each subchannel is $B_N = (1+\beta)/T_N > 1/T_N$ for $\beta > 0$, so the subchannels overlap. Excess bandwidth due to time windowing will increase the subcarrier bandwidth by an additional ϵ/T_N. However, β and ϵ do not affect the total system bandwidth due to the subchannel overlap except in the first and last subchannels, as illustrated in Figure 12.3. The total system bandwidth with overlapping subchannels is given by

$$B = \frac{N + \beta + \epsilon}{T_N} \approx \frac{N}{T_N}, \tag{12.4}$$

where the approximation holds for N large. Thus, with N large, the impact of β and ϵ on the total system bandwidth is negligible, in contrast to the required bandwidth $B = N(1+\beta+\epsilon)/T_N$ when the subchannels do not overlap.

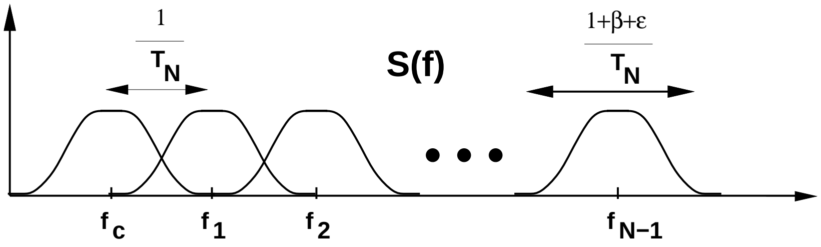


Figure 12.3: Multicarrier with Overlapping Subcarriers.

Example 12.3: Compare the required bandwidth of a multicarrier system with overlapping subchannels versus nonoverlapping subchannels using the same parameters as in Example 12.2.

Solution In the prior example $T_N = .2$ ms, $N = 128$, $\beta = 1$, and $\epsilon = .1$ With overlapping subchannels, from (12.4),

$$B = \frac{N + \beta + \epsilon}{T_N} = \frac{128 + 1 + .1}{.0002} = 645.5 \text{ KHz} \approx B/T_N = 640 \text{ KHz}.$$

By comparison, in the prior example the required bandwidth with nonoverlapping subchannels was shown to be 1.344 MHz, more than double the required bandwidth when the subchannels overlap.

Clearly, in order to separate out overlapping subcarriers, a different receiver structure is needed than the one shown in Figure 12.2. In particular, overlapping subchannels are demodulated with the receiver structure shown in Figure 12.4, which demodulates the appropriate symbol without interference from overlapping subchannels. Specifically, if the effect of the channel $h(t)$ and noise $n(t)$ are neglected then for received signal $s(t)$ given by (12.1), the input to each symbol demapper in Figure 12.4 is

$$\begin{aligned}
\hat{s}_i &= \int_0^{T_N} \left(\sum_{j=0}^{N-1} s_j g(t) \cos(2\pi f_j t + \phi_j) \right) g(t) \cos(2\pi f_i t + \phi_i) dt \\
&= \sum_{j=0}^{N-1} s_j \int_0^{T_N} g^2(t) \cos(2\pi(f_0 + j/T_N)t + \phi_j) \cos(2\pi(f_0 + i/T_N)t + \phi_i) dt \\
&= \sum_{j=0}^{N-1} s_j \delta(j - i) \qquad (12.5) \\
&= s_i, \qquad (12.6)
\end{aligned}$$

where (12.5) follows from the fact that the functions $\{g(t)cos(2\pi f_j t + \phi_j)\}$ form a set of orthonormal basis functions on $[0, T_N]$. If the channel and noise effects are included, the symbol in the ith subchannel is scaled by the channel gain $\alpha_i = H(f_i)$ and corrupted by the noise sample, so $\hat{s}_i = \alpha_i s_i + n_i$, where n_i is AWGN with power $N_0 B_N$. This multicarrier system makes much more efficient use of bandwidth than in systems with nonoverlapping subcarriers. However, since the subcarriers overlap, their orthogonality is compromised by timing and frequency offset. These effects, even when relatively small, can significantly degrade performance, as they cause subchannels to interfere with each other. These effects are discussed in more detail in Section 12.5.2.

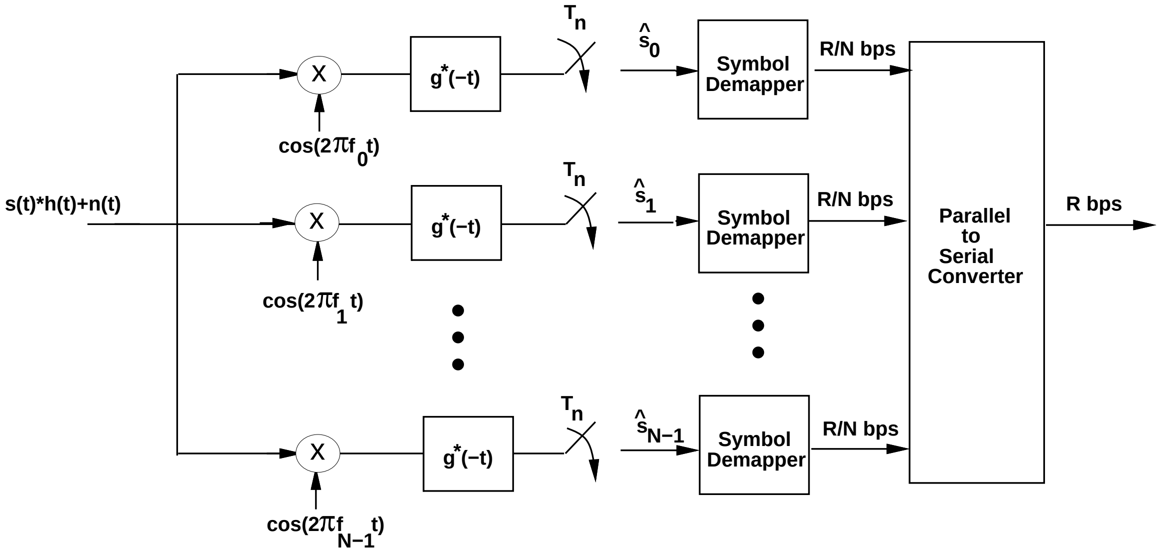


Figure 12.4: Multicarrier Receiver for Overlapping Subcarriers.

12.3 Mitigation of Subcarrier Fading

The advantage of multicarrier modulation is that each subchannel is relatively narrowband, which mitigates the effect of delay spread. However, each subchannel experiences flat-fading, which can cause large BERs on some of the subchannels. In particular, if the transmit power on subcarrier i is P_i, and the fading on that subcarrier is α_i, then the received SNR is $\gamma_i = \alpha_i^2 P_i/(N_0 B_N)$, where B_N is the bandwidth of each subchannel. If α_i is small then the received SNR on the ith subchannel is quite low, which can lead to a high BER on that subchannel. Moreover, in wireless channels the α_i's will vary over time according to a given fading distribution, resulting in the same performance degradation associated with flat fading for single carrier systems discussed in Chapter 6. Since flat fading can seriously degrade performance in each subchannel, it is important to compensate for flat fading in the subchannels. There are several techniques for doing this, including coding with interleaving over time and frequency, frequency equalization, precoding, and adaptive loading, all described in subsequent sections. Coding with interleaving is the most common, and has been adopted as part of the European standards for digital audio and video broadcasting [3, 4]. Moreover, in rapidly changing channels it is difficult to estimate the channel at the receiver and feed this information back to the transmitter. Without channel information at the transmitter, precoding and adaptive loading cannot be done, so only coding with interleaving is effective at fading mitigation.

12.3.1 Coding with Interleaving over Time and Frequency

The basic idea in coding with interleaving over time and frequency is to encode data bits into codewords, interleave the resulting coded bits over both time and frequency, and then transmit the coded bits over different subchannels such that the coded bits within a given codeword all experience independent fading [19]. If most of the subchannels have a high SNR, the codeword will have most coded bits received correctly, and the errors associated with the few bad subchannels can be corrected. Coding across subchannels basically exploits the frequency diversity inherent to a multicarrier system to correct for errors. This technique only works well if there is sufficient frequency diversity across the total system bandwidth. If the coherence bandwidth of the channel is large, then the fading across subchannels will be highly correlated, which will significantly reduce the effect of coding. Most coding for OFDM assumes channel information in the decoder. Channel estimates are typically obtained by a two dimensional pilot symbol transmission over both time and frequency [20].

Note that coding with frequency/time interleaving takes advantage of the fact that the data on all the subcarriers is associated with the same user, and can therefore be jointly processed. The other techniques for fading mitigation discussed in subsequent sections are all basically flat fading compensation techniques, which apply equally to multicarrier systems as well as narrowband flat fading single carrier systems [3, 2].

12.3.2 Frequency Equalization

In frequency equalization the flat fading α_i on the ith subchannel is basically inverted in the receiver [3]. Specifically, the received signal is multiplied by $1/\alpha_i$, which gives a resultant signal power $\alpha_i^2 P_i/\alpha_i^2 = P_i$. While this removes the impact of flat fading on the signal, it enhances the noise. Specifically, the incoming noise signal is also multiplied by $1/\alpha_i$, so the noise power becomes $N_0 B_N/\alpha_i^2$ and the resultant SNR on the ith subchannel after frequency equalization is the same as before equalization. Therefore, frequency equalization does not really change the performance degradation associated with subcarrier flat fading.

12.3.3 Precoding

Precoding uses the same idea as frequency equalization, except that the fading is inverted at the transmitter instead of the receiver [21]. This technique requires that the transmitter have knowledge of the subchannel flat fading gains $\alpha_i, i = 0, \ldots, N-1$, which must be obtained through estimation [22]. In this case, if the desired received signal power in the ith subchannel is P_i, and the channel introduces a flat-fading gain α_i in the ith subchannel, then under precoding the power transmitted in the ith subchannel is P_i/α_i^2. The subchannel signal is corrupted by flat-fading with gain α_i, so the received signal power is $P_i \alpha_i^2/\alpha_i^2 = P_i$, as desired. Note that the channel inversion takes place at the transmitter instead of the receiver, so the noise power remains as $N_0 B_N$. Precoding is quite common on wireline multicarrier systems like HDSL. There are two main problems with precoding in a wireless setting. First, precoding is basically channel inversion, and we know from Section 6.3.5 that inversion is not power-efficient in fading channels. In fact, an infinite amount of power is needed to do channel inversion on a Rayleigh fading channel. The other problem with precoding is the need for accurate channel estimates at the transmitter, which are difficult to obtain in a rapidly fading channel.

12.3.4 Adaptive Loading

Adaptive loading is based on the adaptive modulation techniques discussed in Chapter **??**. It is commonly used on slowly changing channels like digital subscriber lines [8], where channel estimates at the transmitter can be obtained fairly easily. The basic idea is to vary the data rate and power assigned to each subchannel relative to that subchannel gain. As in the case of precoding, this requires knowledge of the subchannel fading $\{\alpha_i, i = 0, \ldots, N-1\}$ at the transmitter. In adaptive loading power and rate on each subchannel is adapted to maximize the total rate of the system using adaptive modulation such as variable-rate variable-power MQAM.

Before investigating adaptive modulation, let us consider the capacity of the multicarrier system with N independent subchannels of bandwidth B_N and subchannel gain $\{\alpha_i, i = 0, \ldots, N-1\}$. Assuming a total power constraint P, this capacity is given by[2]:

$$C = \max_{P_i : \sum P_i = P} \sum_{i=0}^{N-1} B_N \log\left(1 + \frac{\alpha_i^2 P_i}{N_0 B_N}\right). \tag{12.7}$$

The power allocation P_i that maximizes this expression is a water-filling over frequency given by Equation (4.24):

$$\frac{P_i}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_i} & \gamma_i \geq \gamma_0 \\ 0 & \gamma_i < \gamma_0 \end{cases} \tag{12.8}$$

for some cutoff value γ_0, where $\gamma_i = \alpha_i^2 P/(N_0 B_N)$. The cutoff value is obtained by substituting the power adaptation formula into the power constraint. The capacity then becomes

$$C = \sum_{i:\gamma_i \geq \gamma_0} B_N \log(\gamma_i/\gamma_0). \tag{12.9}$$

Applying the variable-rate variable-power MQAM modulation scheme described in Chapter **??** to the subchannels, the total data rate is given by

$$R = B_N \sum_{i=1}^{N} \log(1 + K\gamma_i P_i/P), \tag{12.10}$$

where $K = -1.5/\ln(5P_b)$ for P_b is the desired target BER in each subchannel. Optimizing this expression relative to the P_i's yields the optimal power allocation

$$\frac{P_i}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma_K} & \gamma_i \geq \gamma_K \\ 0 & \gamma_i < \gamma_K \end{cases} \tag{12.11}$$

[2] As discussed in Chapter 4.3.1, this summation is the exact capacity when the α_is are independent. However, in order for the α_is to be independent, the subchannels must be separated by the coherence bandwidth of the channel, which would imply that the subchannels are no longer flat fading. Since the subchannels are designed to be flat fading, the subchannel gains $\{\alpha_i, i = 1, \ldots, N\}$ will be correlated, in which case the capacity obtained by summing over the capacity in each subchannel is an upper bound on the true capacity. We will take this bound to be the actual capacity, since in practice the bound is quite tight.

and corresponding data rate

$$R = B_N \sum_{i:\gamma_i \geq \gamma_K} \log(\gamma_i/\gamma_K), \tag{12.12}$$

where γ_K is a cutoff fade depth dictated by the power constraint P and K.

12.4 Discrete Implementation of Multicarrier

Although multicarrier modulation was invented in the 1950's, its requirement for separate modulators and demodulators on each subchannel was far too complex for most system implementations at the time. However, the development of simple and cheap implementations of the discrete Fourier transform (DFT) and the inverse DFT (IDFT) twenty years later, combined with the realization that multicarrier modulation can be implemented with these algorithms, ignited its widespread use. In this section, after first reviewing the basic properties of the DFT, we illustrate OFDM, which implements multicarrier modulation using the DFT and IDFT.

12.4.1 The DFT and its Properties

Let $x[n], 0 \leq n \leq N-1$, denote a discrete time sequence. The N-point DFT of $x[n]$ is defined as [11]

$$\text{DFT}\{x[n]\} = X[i] \triangleq \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi ni}{N}}, \;\; 0 \leq i \leq N-1. \tag{12.13}$$

The DFT is the discrete-time equivalent to the continuous-time Fourier transform, as $X[i]$ characterizes the frequency content of the time samples $x[n]$ associated with the original signal $x(t)$. Both the continuous-time Fourier transform and the DFT are based on the fact that complex exponentials are eigenfunctions for any linear system. The sequence $x[n]$ can be recovered from its DFT using the IDFT:

$$\text{IDFT}\{X[i]\} = x[n] \triangleq \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} X[i] e^{j\frac{2\pi ni}{N}}, 0 \leq n \leq N-1. \tag{12.14}$$

The DFT and its inverse are typically performed in hardware using the fast Fourier transform (FFT) and inverse FFT (IFFT).

When an input data stream $x[n]$ is sent through a linear time-invariant discrete-time channel $h[n]$, the output $y[n]$ is the discrete-time convolution of the input and the channel impulse response:

$$y[n] = h[n] * x[n] = x[n] * h[n] = \sum_k h[k]x[n-k]. \tag{12.15}$$

The N-point **circular convolution** of $x[n]$ and $h[n]$ is defined as

$$y[n] = x[n] \otimes h[n] = h[n] \otimes x[n] = \sum_k h[k]x[n-k]_N, \tag{12.16}$$

where $[n-k]_N$ denotes $[n-k]$ modulo N. In other words, $x[n-k]_N$ is a periodic version of $x[n-k]$ with period N. It is easily verified that $y[n]$ given by (12.16) is also periodic with period N. From the definition of the DFT, circular convolution in time leads to multiplication in frequency:

$$\text{DFT}\{y[n] = x[n] \otimes h[n]\} = X[i]H[i], \; 0 \leq i \leq N-1. \tag{12.17}$$

By (12.17), if the channel and input are circularly convoluted then if $h[n]$ is known at the receiver, the original data sequence $x[n]$ can be recovered by taking the IDFT of $Y[i]/H[i], 0 \leq i \leq N-1$. Unfortunately, the channel output is not a circular convolution but a linear convolution. However, the linear convolution between the channel input and impulse response can be turned into a circular convolution by adding a special prefix to the input called a **cyclic prefix**, described in the next section.

12.4.2 The Cyclic Prefix

Consider a channel input sequence $x[n] = x[0], \ldots, x[N-1]$ of length N and a discrete-time channel with finite impulse response (FIR) $h[n] = h[0], \ldots, h[\mu]$ of length $\mu + 1 = T_m/T_s$, where T_m is the channel delay spread and T_s the sampling time associated with the discrete time sequence. The cyclic prefix for $x[n]$ is defined as $\{x[N-\mu], \ldots, x[N-1]\}$: it consists of the last μ values of the $x[n]$ sequence. For each input sequence of length N, these last μ samples are appended to the beginning of the sequence. This yields a new sequence $\tilde{x}[n], -\mu \leq n \leq N-1$, of length $N+\mu$, where $\tilde{x}[-\mu], \ldots, \tilde{x}[N-1] = x[N-\mu], \ldots, x[N-1], x[0], \ldots, x[N-1]$, as shown in Figure 12.5. Note that with this definition, $\tilde{x}[n] = x[n]_N$ for $-\mu \leq n \leq N-1$, which implies that $\tilde{x}[n-k] = x[n-k]_N$ for $-\mu \leq n-k \leq N-1$.

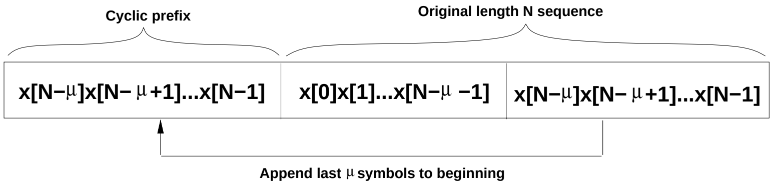


Figure 12.5: Cyclic Prefix of Length μ.

Suppose $\tilde{x}[n]$ is input to a discrete-time channel with impulse response $h[n]$. The channel output $y[n], 0 \leq n \leq N-1$ is then

$$\begin{aligned} y[n] &= \tilde{x}[n] * h[n] \\ &= \sum_{k=0}^{\mu-1} h[k]\tilde{x}[n-k] \\ &= \sum_{k=0}^{\mu-1} h[k]x[n-k]_N \\ &= x[n] \otimes h[n], \end{aligned} \tag{12.18}$$

where the third equality follows from the fact that for $0 \leq k \leq \mu-1$, $\tilde{x}[n-k] = x[n-k]_N$ for $0 \leq n \leq N-1$. Thus, by appending a cyclic prefix to the channel input, the linear convolution associated with the channel impulse response $y[n]$ for $0 \leq n \leq N-1$ becomes a circular convolution. Taking the DFT of the channel output in the absense of noise then yields

$$Y[i] = \mathrm{DFT}\{y[n] = x[n] \otimes h[n]\} = X[i]H[i], \;\; 0 \leq i \leq N-1, \tag{12.19}$$

and the input sequence $x[n], 0 \leq n \leq N-1$, can be recovered from the channel output $y[n], 0 \leq n \leq N-1$, for known $h[n]$ by

$$x[n] = \mathrm{IDFT}\{Y[i]/H[i]\} = \mathrm{IDFT}\{\mathrm{DFT}\{y[n]\}/\mathrm{DFT}\{h[n]\}\}. \tag{12.20}$$

Note that $y[n], -\mu \leq n \leq N-1$, has length $N+\mu$, yet from (12.20) the first μ samples $y[-\mu], \ldots, y[-1]$ are not needed to recover $x[n], 0 \leq n \leq N-1$, due to the redundancy associated with the cyclic prefix. Moreover, if we assume that the input $x[n]$ is divided into data blocks of size N with a cyclic prefix appended to each block to form $\tilde{x}[n]$, then the first μ samples of $y[n] = h[n] * \tilde{x}[n]$ in a given block are corrupted by ISI associated with the last μ samples of $x[n]$ in the prior block, as illustrated in Figure 12.6. The cyclic prefix serves to eliminate ISI between the data blocks since the first μ samples of the channel output affected by this ISI can be discarded without any loss relative to the original information sequence. In continuous time this is equivalent to using a guard band of duration T_m (the channel delay spread) after every block of N symbols of duration NT_s to eliminate the ISI between these data blocks.

The benefits of adding a cyclic prefix come at a cost. Since μ symbols are added to the input data blocks, there is an overhead of μ/N, resulting in a data rate reduction of $N/(\mu + N)$. The transmit power associated with sending the cyclic prefix is also wasted since this prefix consists of redundant data. It is clear from Figure 12.6 that any prefix of length μ appended to input blocks of size N eliminates ISI between data blocks if the first μ samples of the block are discarded. In particular, the prefix can consist of all zero symbols, in which case although the data rate is still reduced by $N/(N+\mu)$, no power is used in transmitting the prefix. Tradeoffs associated with the cyclic prefix versus this all-zero prefix, which is a form of vector coding, are discussed in Section 12.9.

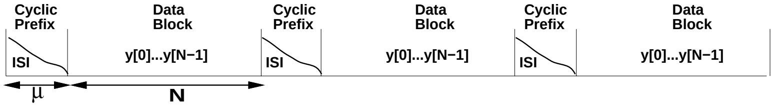


Figure 12.6: ISI Between Data Blocks in Channel Output.

The above analysis motivates the design of OFDM. In OFDM the input data is divided into blocks of size N referred to as an **OFDM symbol**. A cyclic prefix is added to each OFDM symbol to induce circular convolution of the input and channel impulse response. At the receiver, the output samples affected by ISI between OFDM symbols are removed. The DFT of the remaining samples are used to recover the original input sequence. The details of this OFDM system design are given in the next section.

Example 12.4: Consider an OFDM system with total passband bandwidth $B = 1$ MHz assuming $\beta = \epsilon = 0$. A single carrier system would have symbol time $T_s = 1/B = 1\mu$s. The channel has a maximum delay spread of $T_m = 5$ μsec, so with $T_s = 1$ μsec and $T_m = 5$ μsec there would clearly be severe ISI. Assume an OFDM system with MQAM modulation applied to each subchannel. To keep the overhead small, the OFDM system uses $N = 128$ subcarriers to mitigate ISI. So $T_N = NT_s = 128$ μsec. The length of the cyclic prefix is set to $\mu = 8 > T_m/T_s$ to insure no ISI between OFDM symbols. For these parameters, find the subchannel bandwidth, the total transmission time associated with each OFDM symbol, the overhead of the cyclic prefix, and the data rate of the system assuming $M = 16$.

Solution: The subchannel bandwidth $B_N = 1/T_N = 7.812$ KHz, so $B_N << B_c = 1/T_m = 200$ KHz, insuring negligible ISI. The total transmission time for each OFDM symbol is $T = T_N + \mu T_s = 128 + 8 = 136$ μs. The overhead associated with the cyclic prefix is $8/136$ which is roughly 5.9%. The system transmits $\log_2 16 = 4$ bits/subcarrier every T seconds, so the data rate is $128 \times 4/136 \times 10^{-6} = 3.76$ Mbps, which is slightly less than $4B$ due to the cyclic prefix overhead.

12.4.3 Orthogonal Frequency Division Multiplexing (OFDM)

The OFDM implementation of multicarrier modulation is shown in Figure 12.7. The input data stream is modulated by a QAM modulator, resulting in a complex symbol stream $X[0], X[1], \ldots, X[N-1]$. This symbol stream is passed through a serial-to-parallel converter, whose output is a set of N parallel QAM symbols $X[0], \ldots, X[N-1]$ corresponding to the symbols transmitted over each of the subcarriers. Thus, the N symbols output from the serial-to-parallel converter are the discrete frequency components of the OFDM modulator output $s(t)$. In order to generate $s(t)$, these frequency components are converted into time samples by performing an inverse DFT on these N symbols, which is efficiently implemented using the IFFT algorithm. The IFFT yields the OFDM symbol consisting of the sequence $x[n] = x[0], \ldots, x[N-1]$ of length N, where

$$x[n] = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} X[i] e^{j2\pi ni/N}, \;\; 0 \leq n \leq N-1. \tag{12.21}$$

This sequence corresponds to samples of the multicarrier signal: i.e. the multicarrier signal consists of linearly-modulated subchannels, and the right hand side of (12.21) corresponds to samples of a sum of QAM symbols $X[i]$ each modulated by carrier frequency $e^{j2\pi it/T_N}, i = 0, \ldots, N-1$. The cyclic prefix is then added to the OFDM symbol, and the resulting time samples $\tilde{x}[n] = \tilde{x}[-\mu], \ldots, \tilde{x}[N-1] = x[N-\mu], \ldots, x[0], \ldots, x[N-1]$ are ordered by the parallel-to-serial converter and passed through a D/A converter, resulting in the baseband OFDM signal $\tilde{x}(t)$, which is then upconverted to frequency f_0.

The transmitted signal is filtered by the channel impulse response $h(t)$ and corrupted by additive noise, so that the received signal is $y(t) = \tilde{x}(t) * h(t) + n(t)$. This signal is downconverted to baseband and filtered to remove the high frequency components. The A/D converter samples the resulting signal to obtain $y[n] = \tilde{x}[n] * h[n] + \nu[n], -\mu \leq n \leq N-1$. The prefix of $y[n]$ consisting of the first μ samples is then removed. This results in N time samples whose DFT in the absence of noise is $Y[i] = H[i]X[i]$. These time samples are serial-to-parallel converted and passed through an FFT. This results in scaled versions of the original symbols $H[i]X[i]$, where $H[i] = H(f_i)$ is the flat-fading channel gain associated with the ith subchannel. The FFT output is parallel-to-serial converted and passed through a QAM demodulator to recover the original data.

The OFDM system effectively decomposes the wideband channel into a set of narrowband orthogonal subchannels with a different QAM symbol sent over each subchannel. Knowledge of the channel gains $H[i], i = 0, \ldots, N-1$ is not needed for this decomposition, in the same way that a continuous time channel with frequency response $H(f)$ can be divided into orthogonal subchannels without knowledge of $H(f)$ by splitting the total signal bandwidth into nonoverlapping subbands. The demodulator can use the channel gains to recover the original QAM symbols by dividing out these gains: $X[i] = Y[i]/H[i]$. This process is called frequency equalization. However, as discussed in Section 12.3.2 for continuous-time OFDM, frequency equalization leads to noise enhancement, since the noise in the ith subchannel is also scaled by $1/H[i]$. Hence, while the effect of flat fading on $X[i]$ is removed by this equalization, its received SNR is unchanged. Precoding, adaptive loading, and coding across subchannels, as discussed in Section 12.3, are better approaches to mitigate the effects of flat fading across subcarriers. An alternative to using the cyclic prefix is to use a prefix consisting of all zero symbols. In this case the OFDM symbol consisting of $x[n], 0 \leq n \leq N-1$ is preceeded by μ null samples, as illustrated in Figure 12.8. At the receiver the "tail" of the ISI associated with the end of a given OFDM symbol is added back in to the beginning of the symbol, which recreates the effect of a cyclic prefix, so the rest of the OFDM system functions as usual. This zero prefix reduces the transmit power relative to a cyclic prefix by $\frac{N}{\mu+N}$, since the prefix does not require any transmit power. However, the noise from the received tail is added back into the beginning of the symbol, which increases the noise power by $\frac{N+\mu}{N}$. Thus, the difference in SNR is not significant for the two prefixes.

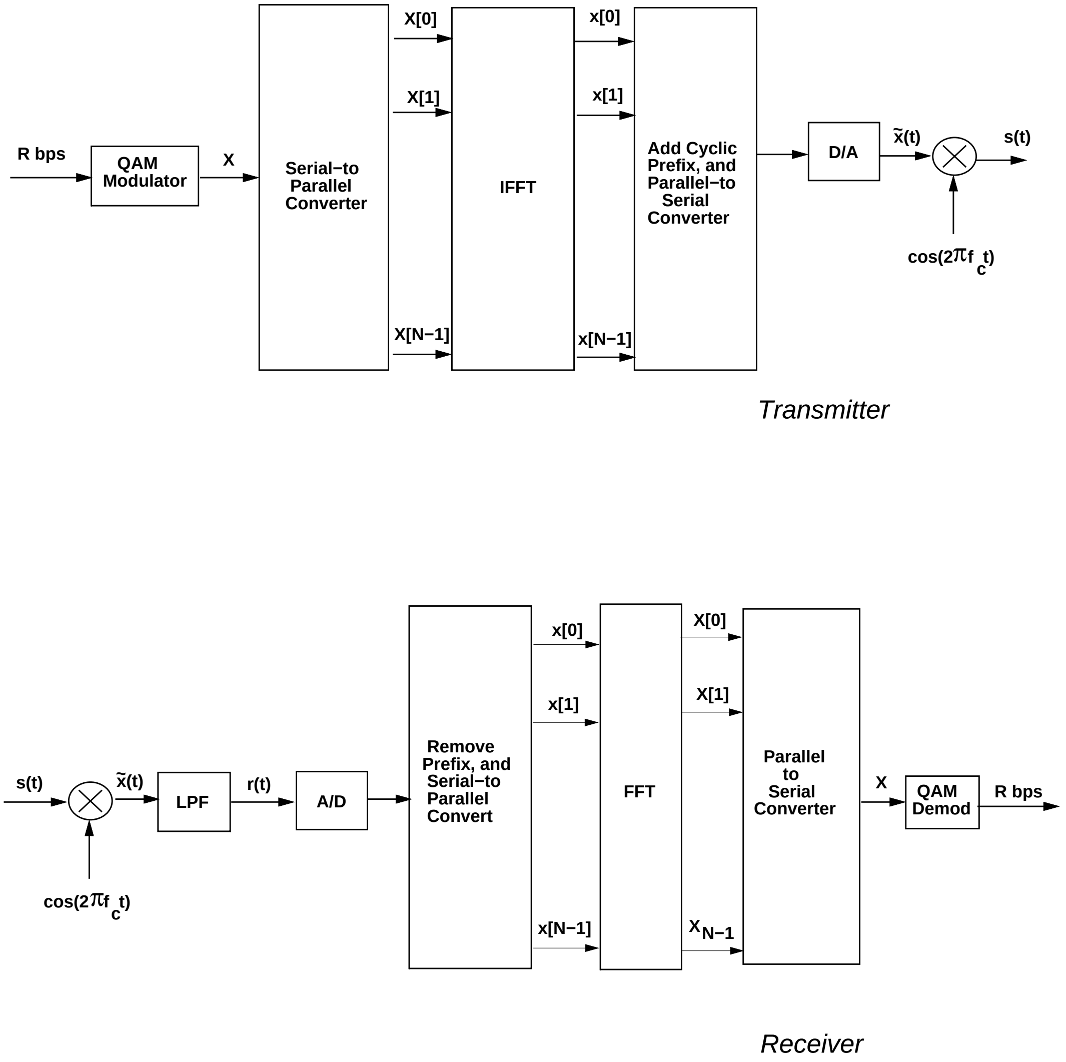


Figure 12.7: OFDM with IFFT/FFT Implementation.

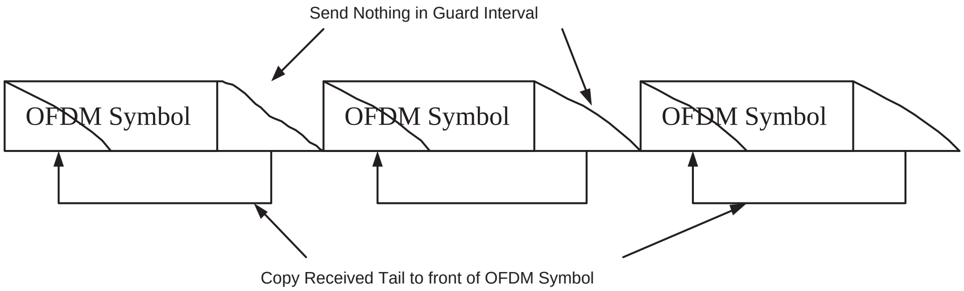

Send Nothing in Guard Interval
OFDM Symbol
OFDM Symbol
OFDM Symbol
Copy Received Tail to front of OFDM Symbol

Figure 12.8: Creating a Circular Channel with an All-Zero Prefix.

12.4.4 Matrix Representation of OFDM

An alternate analysis for OFDM is based on a matrix representation of the system. Consider a discrete-time channel with FIR $h[n], 0 \leq n \leq \mu$, input $\tilde{x}[n]$, noise $\nu[n]$, and output $y[n] = \tilde{x}[n] * h[n] + \nu[n]$. Denote the nth element of these sequences as $h_n = h[n]$, $\tilde{x}_n = \tilde{x}[n]$, $\nu_n = \nu[n]$, and $y_n = y[n]$. With this notation the channel output sequence can be written in matrix form as

$$\begin{bmatrix} y_{N-1} \\ y_{N-2} \\ \vdots \\ y_0 \end{bmatrix} = \begin{bmatrix} h_0 & h_1 & \dots & h_\mu & 0 & \dots & 0 \\ 0 & h_0 & \dots & h_{\mu-1} & h_\mu & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & h_0 & \dots & h_{\mu-1} & h_\mu \end{bmatrix} \begin{bmatrix} x_{N-1} \\ \vdots \\ x_0 \\ x_{-1} \\ \vdots \\ x_{-\mu} \end{bmatrix} + \begin{bmatrix} \nu_{N-1} \\ \nu_{N-2} \\ \vdots \\ \nu_0 \end{bmatrix}, \tag{12.22}$$

which can be written more compactly as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \nu. \tag{12.23}$$

The received symbols $y_{-1} \dots y_{-\mu}$ are discarded since they are affected by ISI in the prior data block, and they are not needed to recover the input. The last μ symbols of $x[n]$ correspond to the cyclic prefix: $x_{-1} = x_{N-1}$, $x_{-2} = x_{N-2}, \dots\ x_{-\mu} = x_{N-\mu}$. From this it can be shown that the matrix representation (12.22) is equivalent to the following representation:

$$\begin{bmatrix} y_{N-1} \\ y_{N-2} \\ \vdots \\ \vdots \\ \vdots \\ y_0 \end{bmatrix} = \begin{bmatrix} h_0 & h_1 & \dots & h_\mu & 0 & \dots & 0 \\ 0 & h_0 & \dots & h_{\mu-1} & h_\mu & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & h_0 & \dots & h_{\mu-1} & h_\mu \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ h_2 & h_3 & \dots & h_{\mu-2} & \dots & h_0 & h_1 \\ h_1 & h_2 & \dots & h_{\mu-1} & \dots & 0 & h_0 \end{bmatrix} \begin{bmatrix} x_{N-1} \\ x_{N-2} \\ \vdots \\ \vdots \\ x_0 \end{bmatrix} + \begin{bmatrix} \nu_{N-1} \\ \nu_{N-2} \\ \vdots \\ \vdots \\ \vdots \\ \nu_0 \end{bmatrix}, \tag{12.24}$$

which can be written more compactly as

$$\mathbf{y} = \tilde{\mathbf{H}}\mathbf{x} + \nu. \tag{12.25}$$

This equivalent model shows that the inserted cyclic prefix allows the channel to be modelled as a circulant convolution matrix $\tilde{\mathbf{H}}$ over the N samples of interest. The matrix $\tilde{\mathbf{H}}$ is $N \times N$, so it has an eigenvalue decomposition

$$\tilde{\mathbf{H}} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^H, \tag{12.26}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues of $\tilde{\mathbf{H}}$ and $\mathbf{M}^H$ is a unitary matrix whose rows comprise the eigenvectors of $\tilde{\mathbf{H}}$.

It is straightforward to show that the DFT operation on $x[n]$ can be represented by the matrix multiplication

$$\mathbf{X} = \mathbf{Q}\mathbf{x},$$

where $\mathbf{X} = (X[0], \ldots, X[N-1])^T$, $\mathbf{x} = (x[0], \ldots, x[N-1])^T$, and $\mathbf{Q}$ is an $N \times N$ matrix given by

$$Q = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & W_N & W_N^2 & \ldots & W_N^{N-1} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \ldots & W_N^{(N-1)^2} \end{bmatrix}, \tag{12.27}$$

for $W_N = e^{-j\frac{2\pi}{N}}$. Since

$$Q^{-1} = Q^H, \tag{12.28}$$

the IDFT can be similarly represented as

$$\mathbf{x} = Q^{-1}\mathbf{X} = Q^H\mathbf{X}. \tag{12.29}$$

Let $\mathbf{v}$ be an eigenvector of $\mathbf{H}$ with eigenvalue λ. Then

$$\lambda\mathbf{v} = \mathbf{H}\mathbf{v},$$

The unitary matrix $\mathbf{M}^H$ has rows that are the eigenvectors of $\mathbf{H}$, i.e. $\lambda_i\mathbf{m}_i^T = \mathbf{H}\mathbf{m}_i^T$ for $i = 0, 1, \ldots, N-1$, where $\mathbf{m}_i$ denotes the ith row of $\mathbf{M}^H$. It can also be shown by induction that the rows of the DFT matrix $\mathbf{Q}$ are eigenvectors of $\tilde{\mathbf{H}}$, which implies that $\mathbf{Q} = \mathbf{M}^H$ and $\mathbf{Q}^H = \mathbf{M}$. Thus we have that

$$\begin{aligned} \mathbf{Y} &= \mathbf{Q}\mathbf{y} \\ &= \mathbf{Q}[\tilde{\mathbf{H}}\mathbf{x} + \nu] \\ &= \mathbf{Q}[\tilde{\mathbf{H}}\mathbf{Q}^H\mathbf{X} + \nu] \\ &= \mathbf{Q}[\mathbf{M}\Lambda\mathbf{M}^H\mathbf{Q}^H\mathbf{X} + \nu] && (12.30) \\ &= \mathbf{Q}\mathbf{M}\Lambda\mathbf{M}^H\mathbf{Q}^H\mathbf{X} + \mathbf{Q}\nu \\ &= \mathbf{M}^H\mathbf{M}\Lambda\mathbf{M}^H\mathbf{M}\mathbf{X} + \mathbf{Q}\nu \\ &= \Lambda\mathbf{X} + \nu_{\mathbf{Q}} && (12.31) \end{aligned}$$

where since $\mathbf{Q}$ is unitary, $\nu_{\mathbf{Q}} = \mathbf{Q}\nu$ has the same noise autocorrelation matrix as ν, and hence is still generally white and Gaussian, with unchanged noise power. Thus, this matrix analysis also shows that by adding a cyclic prefix and using the IDFT/DFT, OFDM decomposes an ISI channel into N orthogonal subchannels and knowledge of the channel matrix $\mathbf{H}$ is not needed for this decomposition.

The matrix representation is also useful in analyzing OFDM systems with multiple antennas. As discussed in Chapter 10, a MIMO channel is typically represented by an $M_r \times M_t$ matrix, where M_t is the number of transmit antennas and M_r the number of receive antennas. Thus, an OFDM-MIMO channel with N subchannels, M_t transmit antennas, M_r receive antennas, and a channel FIR of duration μ can be represented as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \nu, \tag{12.32}$$

where $\mathbf{y}$ is a vector of dimension $M_rN \times 1$ corresponding to N output time samples at each of the M_r antennas, H is a $NM_r \times (N+\mu)M_t$ matrix corresponding to the N flat-fading subchannel gains on each transmit-receive antenna pair, and $\mathbf{x}$ is a vector of dimension $M_t(N+\mu)\times 1$ corresponding to N input time samples with appended cyclic prefix of length μ at each of the M_t transmit antennas. The matrix is in the same form as in the case of OFDM without multiple antennas, so the same design and analysis applies: with MIMO-OFDM the ISI is removed by breaking the wideband channel into many narrowband subchannels. Each subchannel experiences flat fading, so can be treated as a flat-fading MIMO channel. The capacity of this channel is obtained by applying the same matrix analysis as for standard MIMO to the augmented channel with MIMO and OFDM [16]. In discrete implementations the input associated with each transmit antenna is broken into blocks of size N with a cyclic prefix appended to convert linear convolution to circular and eliminate ISI between input blocks. More details can be found in [24].

12.4.5 Vector Coding

In OFDM the $N\times N$ circular convolution channel matrix $\tilde{\mathbf{H}}$ is decomposed using its eigenvalues and eigenvectors. Vector coding (VC) is a similar technique whereby the original $N\times(N+\mu)$ channel matrix $\mathbf{H}$ from (12.23) is decomposed using an SVD, which can be applied to a matrix of any dimension. The SVD decomposition does not require a cyclic prefix to make the subchannels orthogonal, so it is more efficient than OFDM in terms of energy. However, it is more complex, and requires knowledge of the channel impulse response for the decomposition, in contrast to OFDM, which does not require channel knowledge for its decomposition.

The singular value decomposition of $\mathbf{H}$ can be written as

$$\mathbf{H}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{12.33}$$

where $\mathbf{U}$ is $N\times N$ unitary, $\mathbf{V}$ is $(N+\mu)\times(N+\mu)$ unitary, and $\mathbf{\Sigma}$ is a diagonal matrix whose ith element σ_i is the ith singular value of $\mathbf{H}$. The singular values of $\mathbf{H}$ are related to the eigenvalues of $\mathbf{H}\mathbf{H}^H$ by $\sigma_i=\sqrt{\lambda_i}$ for λ_i the ith eigenvalue of the matrix $\mathbf{H}\mathbf{H}^H$. Because $\mathbf{H}$ is a block-diagonal convolutional matrix, $\mathrm{rank}(\mathbf{H})=N$, i.e. $\sigma_i\neq 0\;\forall\, i$.

In vector coding, as in OFDM, input data symbols are grouped into vectors of N symbols. Let X_i denote the symbol to be transmitted over the ith subchannel and $\mathbf{X}=(X_0,\ldots,X_{N-1})$ denote a vector of these symbols. Each of the data symbols X_i are multiplied by a column of $\mathbf{V}$ in parallel to form a vector, and then added together. At the receiver, the received vector $\mathbf{Y}$ is multiplied by each row of $\mathbf{U}^H$ to yield N output symbols, Y_i, $i=0,1,...,N-1$. This process is illustrated in Figure 12.9, where the multiplication with $\mathbf{V}$ and $\mathbf{U}^H$ performs a similar function as the transmit precoding and receiver shaping in MIMO systems.

Mathematically, it can be seen that the filtered transmit and received vectors are

$$\mathbf{x}=\mathbf{V}\mathbf{X}$$

and

$$\mathbf{Y}=\mathbf{U}^H\mathbf{y}. \tag{12.34}$$

As a result, it can be shown through simple linear algebra that the filtered received vector $\mathbf{Y}$ is ISI-free, since

$$\begin{aligned}\mathbf{Y} &= \mathbf{U}^H\mathbf{y}\\ &= \mathbf{U}^H(\mathbf{H}\mathbf{x}+\nu)\\ &= \mathbf{U}^H(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H)\mathbf{V}\mathbf{X}+\mathbf{U}^H\nu\\ &= \Sigma\mathbf{X}+\mathbf{U}^H\nu.\end{aligned} \tag{12.35}$$

Hence, each element of $\mathbf{X}$ is effectively passed through a scalar channel without ISI, where the scalar gain of subchannel i is the ith singular value of $\mathbf{H}$. Additionally, the new noise vector $\tilde{\nu}=\mathbf{U}^H\nu$ has unchanged noise

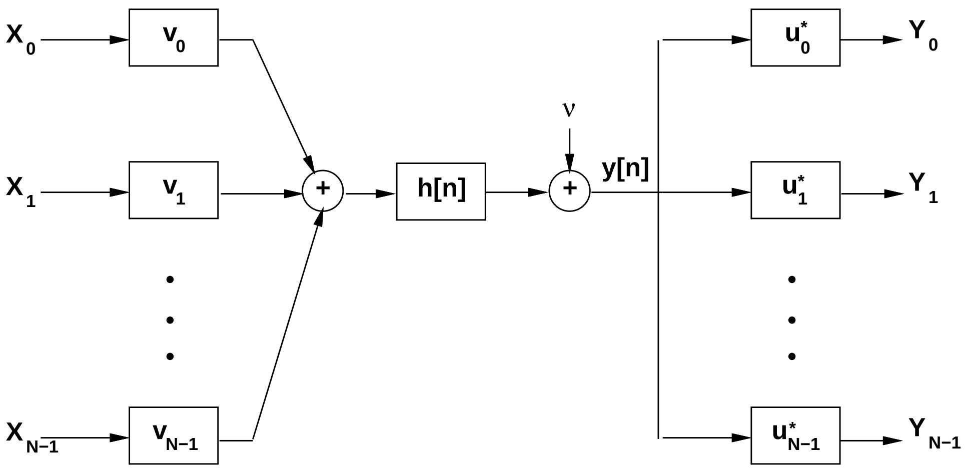


Figure 12.9: Vector Coding.

variance, since **U** is unitary. The resulting received vector is thus

$$\begin{bmatrix} Y_{N-1} \\ Y_{N-2} \\ \vdots \\ Y_0 \end{bmatrix} = \begin{bmatrix} \sigma_1 X_{N-1} \\ \sigma_2 X_{N-2} \\ \vdots \\ \sigma_N X_0 \end{bmatrix} + \begin{bmatrix} \tilde{\nu}_{N-1} \\ \tilde{\nu}_{N-2} \\ \vdots \\ \tilde{\nu}_0 \end{bmatrix} \tag{12.36}$$

From this analysis we see from (12.22) that the matrix **H** is obtained by appending μ extra symbols to each block of N data symbols, which are called **vector codewords**. However, in constrast to OFDM, the SVD decomposition does not require these extra symbols to have any particular form, they are just inserted to eliminate ISI between blocks. In particular, these symbols need not be a cyclic prefix, nor must the "tail" be added back in if the prefix is all zeros. In practice the extra symbols are set to zero to save transmit power, thereby forming a guardband or null prefix between the vector codeword (VC) symbols, as shown in Figure 12.10.

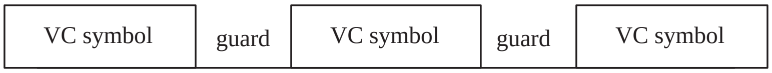


Figure 12.10: Guard Interval (Null Prefix) in Vector Coding

Vector coding has been proven using information and estimation theory to be the optimal partition of the N-dimensional channel **H**. Thus, the capacity of any other channel partitioning scheme will be upper bounded by vector coding. Despite its theoretical optimality and ability to create ISI-free channels with relatively small overhead and no wasted transmit power, there are a number of important practical problems with vector coding. The two most important problems are:

1. **Complexity.** With vector coding, like in simple multichannel modulation, the complexity still scales quickly with N, the number of subcarriers. As seen from Figure 12.9, N transmit precoding and N receive shaping filters are required to implement vector coding. Furthermore, the complexity of finding the SVD of the $N \times (N + \mu)$ matrix **H** increases rapidly with N.

2. **SVD and Channel Knowledge**. In order to orthogonally partition the channel, the SVD of the channel matrix **H** must be computed. In particular, the precoding filter matrix must be known at the transmitter.

This means that every time the channel changes, a new SVD must be computed, and the results conveyed to the transmitter. Generally, the computational complexity of the SVD and the delay incurred in getting the channel information back to the transmitter is prohibitive in wireless systems. Since OFDM can perform this decomposition without channel knowledge, OFDM is the method of choice for discrete multicarrier modulation in wireless applications.

Example 12.5: Consider a simple two-tap discrete-time channel (i.e. $\mu = 1$) described as:

$$H(z) = 1 + 0.9z^{-1}$$

Since $\mu = T_m/T_s = 1$, with $N = 8$ we insure $B_N \approx 1/(NT_s) << B_c \approx 1/T_c$. Find the system matrix representation (12.23) and the singular values of the associated channel matrix $\mathbf{H}$.

Solution: The representation (12.23) for $H(z) = 1 + 0.9z^{-1}$ and $N = 8$ is given by

$$\begin{bmatrix} y_7 \\ y_6 \\ \vdots \\ y_0 \end{bmatrix} = \begin{bmatrix} 1 & 0.9 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & 0.9 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 & 1 & 0.9 \end{bmatrix} \begin{bmatrix} x_7 \\ x_6 \\ \vdots \\ x_{-1} \end{bmatrix} + \begin{bmatrix} n_7 \\ n_6 \\ \vdots \\ n_0 \end{bmatrix} \tag{12.37}$$

The singular values of the matrix $\mathbf{H}$ in (12.37) can be found by a standard computer package (e.g. Matlab) as

$$\Sigma = \text{diag}(1.87, 1.78, 1.65, 1.46, 1.22, 0.95, 0.66, 0.34)$$

The precoding and shaping matrices $\mathbf{U}$ and $\mathbf{V}$ are also easily found. Given $\mathbf{U}$, $\mathbf{V}$, and Σ, this communication is ISI-free, with the symbols X_0, X_1, ... , X_{L-1} being multiplied by the corresponding singular values as in (12.36).

12.5 Challenges in Multicarrier Systems

12.5.1 Peak to Average Power Ratio

The peak to average power ratio (PAR) is a very important attribute of a communication system. A low PAR allows the transmit power amplifier to operate efficiently, whereas a high PAR forces the transmit power amplifier to have a large *backoff* in order to ensure linear amplification of the signal. This is demonstrated in Figure 12.11 showing a typical power amplifier response. Operation in the linear region of this response is generally required to avoid signal distortion, so the peak value is constrained to be in this region. Clearly it would be desirable to have the average and peak values be as close together as possible in order to have the power amplifier operate at the maximum efficiency. Additionally, a high PAR requires high resolution for the receiver A/D convertor, since the dynamic range of the signal is much larger for high PAR signals. High resolution A/D conversion places a complexity and power burden on the receiver front end.

The PAR of a continuous-time signal is given by

$$\text{PAR} \triangleq \frac{\max_t |x(t)|^2}{E_t[|x(t)|^2]}. \tag{12.38}$$

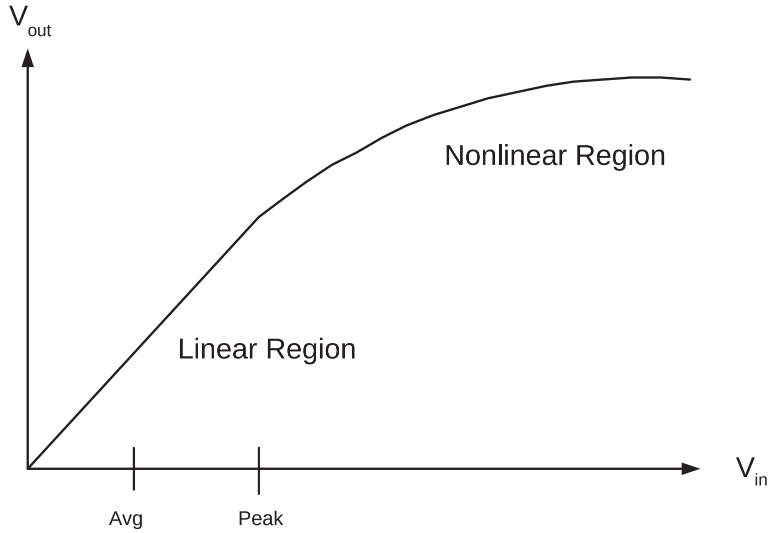


Figure 12.11: A typical power amplifier response.

and for a discrete-time signal it is given by

$$\text{PAR} \triangleq \frac{\max_n |x[n]|^2}{E_n[|x[n]|^2]}. \tag{12.39}$$

Any constant amplitude signal, e.g. a square wave, has $\text{PAR} = 0$ dB. A sine wave has $\text{PAR} = 3$ dB since $\max[\sin^2(t/T)] = 1$ and

$$E[\sin^2(t/T)] = \int_0^T \sin^2(t/T)dt = .5,$$

so PAR=1/.5=2.

In general PAR should be measured with respect to the continuous time signal using (12.38), since the input to the amplifier is an analog signal. The PAR given by (12.38) is sensitive to the pulse shape $g(t)$ used in the modulation, and does not generally lead to simple analytical formulas [41]. For illustration we will focus on the PAR associated with the discrete-time signal, since it lends itself to a simple characterization. However, care must be taken in interpreting these results, since by not taking into account the pulse shape $g(t)$ they can be quite inaccurate.

Consider the time domain samples output from the IFFT:

$$x[n] = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} X[i] e^{j\frac{2\pi i n}{N}}, \;\; 0 \leq n \leq N-1. \tag{12.40}$$

If N is large, the Central Limit Theorem is applicable, and $x[n]$ are zero-mean complex Gaussian random variables since the real and imaginary parts are summed. The Gaussian approximation for IFFT outputs is generally quite accurate for a reasonably large number of subcarriers ($N \geq 64$). For $x[n]$ complex Gaussian, the envelope of the OFDM signal is Rayleigh distributed with variance σ_n^2, and the phase of the signal is uniform. Since the Rayleigh distribution has infinite support, the peak value of the signal will exceed any given value with nonzero probability. It can then be shown that the probability that the PAR given by (12.39) exceeds a threshold $P_0 = \sigma_0^2/\sigma_n^2$ is given by [40]

$$p(\text{PAR} \geq P_0) = 1 - (1 - e^{-P_0})^N. \tag{12.41}$$

Let us now investigate how PAR grows with the number of subcarriers. Consider N Gaussian i.i.d. random variables x_n, $0 \leq n \leq N-1$ with mean zero and unit power. The average signal power $E_n[|x[n]|^2]$ is then

$$\begin{aligned} E\left[\frac{1}{\sqrt{N}}\left|x_0+x_1+\cdots+x_{N-1})\right|^2\right] &= \frac{1}{N}E\left|x_0+x_1+\cdots+x_{N-1}\right|^2 \\ &= \frac{E|x_0^2|+E|x_1^2|+\cdots+E|x_{N-1}^2|}{N} \\ &= 1 \end{aligned} \quad (12.42)$$

The maximum value occurs when all the x_is add coherently, in which case

$$\max\left[\frac{1}{\sqrt{N}}\left|x_0+x_1+\cdots+x_{N-1}\right|\right]^2 = \left|\frac{N}{\sqrt{N}}\right|^2 = N. \quad (12.43)$$

Hence the maximum PAR is N for N subcarriers. In practice full coherent addition of all N symbols is highly improbable, so the observed PAR is typically less than N, usually by many dB. Nevertheless, PAR increases approximately linearly with the number of subcarriers. So, although it is desirable to have N as large as possible in order to keep the overhead associated with the cyclic prefix down, a large PAR is an important penalty that must be paid for large N.

There are a number of ways to reduce or tolerate the PAR of OFDM signals, including clipping the OFDM signal above some threshold, peak cancellation with a complementary signal, allowing non-linear distortion from the power amplifier (and correction for it), and special coding techniques [31]. A good summary of some of these techniques can be found in [38].

12.5.2 Frequency and Timing Offset

OFDM modulation encodes the data symbols X_i onto orthogonal subchannels, where orthogonality is assured by the subcarrier separation $\Delta f = 1/T_N$. The subchannels may overlap in the frequency domain, as shown in Figure 12.12 for a rectangular pulse shape in time (sinc function in frequency). In practice, the frequency separation of the subcarriers is imperfect: so Δf is not exactly equal to $1/T_N$. This is generally caused by mismatched oscillators, Doppler frequency shifts, or timing synchronization errors. For example, if the carrier frequency oscillator is accurate to 0.1 parts per million (ppm), the frequency offset $\Delta f_\epsilon \approx (f_0)(0.1 \times 10^{-6})$. If $f_0 = 5$ GHz, the carrier frequency for 802.11a WLANs, then $\Delta f_\epsilon = 500$ Hz, which will degrade the orthogonality of the subchannels, since now the received samples of the FFT will contain interference from adjacent subchannels. We'll now analyze this intercarrier interference (ICI).

The signal corresponding to subcarrier i can be simply expressed for the case of rectangular pulse shapes (suppressing the data symbol and the carrier frequency) as

$$x_i(t) = e^{j\frac{2\pi i t}{T_N}}. \quad (12.44)$$

An interfering subchannel signal can be written as

$$x_{i+m}(t) = e^{j\frac{2\pi(i+m)t}{T_N}}. \quad (12.45)$$

If the signal is demoduled with a frequency offset of δ/T_n then this interference becomes

$$x_{i+m}(t) = e^{j\frac{2\pi(i+m+\delta)t}{T_N}}. \quad (12.46)$$

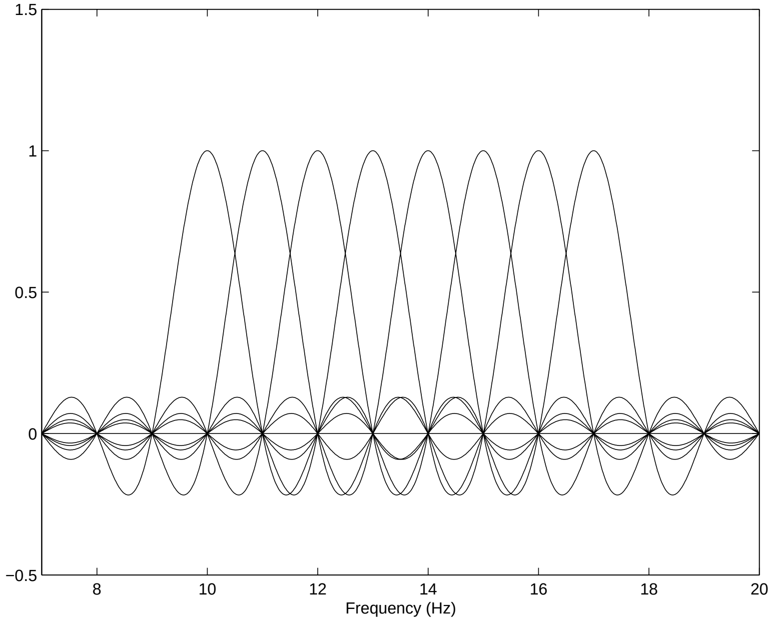


Figure 12.12: OFDM Overlapping Subcarriers: Rectangular Pulses, $f_0 = 10$ Hz and $\Delta f = 1$ Hz.

The ICI between subchannel signals x_i and x_{i+m} is simply the inner product between them:

$$I_m = \int_0^{T_N} x_i(t)x_{i+m}(t)dt = \frac{T_N\left(1 - e^{-j2\pi(\delta+m)}\right)}{j2\pi(m+\delta)}. \tag{12.47}$$

It can be seen that in the above expression, $\delta = 0 \Rightarrow I_m = 0$, as expected. The total ICI power on subcarrier i is then

$$ICI_i = \sum_{m \neq i} |I_m|^2 \approx C_0(T_N\delta)^2, \tag{12.48}$$

where C_0 is some constant. Several important trends can be observed from this simple approximation. First, as T_N increases, the subcarriers grow narrower and hence more closely spaced, which then results in more ICI. Second, the ICI predictably grows with the frequency offset δ, and the growth is about quadratic. Another interesting observation is that (12.48) does not appear to be directly effected by N. But picking N large generally forces T_N to also be large, which then causes the subcarriers to be closer together. Along with the larger PAR that comes with large N, the increased ICI is another reason to pick N as low as possible, given that the overhead budget can be met. In order to further reduce the ICI for a given choice of N, non-rectangular windows can also be used [30, 33].

The effects from timing offset are generally less than those from the frequency offset, as long as a full N sample OFDM symbol is used at the receiver, without interference from the previous or subsequent OFDM symbols (this is ensured by taking the cyclic prefix length $\mu >> \sigma_{T_m}/T_s$, where σ_{T_m} is the channel's rms delay spread). It can be shown that the ICI power on subcarrier i due to a receiver timing offset τ can be approximated as $2(\tau/T_N)^2$. Since usually $\tau \ll T_N$, this effect is typically negligible.

12.6 Case Study: The IEEE 802.11a Wireless LAN Standard

The IEEE 802.11a Wireless LAN standard, which occupies 20 MHz of bandwidth in the 5 GHz unlicensed band, is based on OFDM [26]. The IEEE 802.11g standard is virtually identical, but operates in the smaller and more crowded 2.4 GHz unlicensed ISM band [28]. In this section we study the properties of this OFDM design and discuss some of the design choices.

In 802.11a, $N = 64$ subcarriers are generated, although only 48 are actually used for data transmission, with the outer 12 zeroed in order to reduce adjacent channel interference, and 4 used as pilot symbols for channel estimation. The cyclic prefix consists of $\mu = 16$ samples, so the total number of samples associated with each OFDM symbol, including both data samples and the cyclic prefix, is 80. The transmitter gets periodic feedback from the receiver about the packet error rate, which it uses to pick an appropriate error correction code and modulation technique. The same code and modulation must be used for *all* the subcarriers at any given time. The error correction code is a convolutional code with one of three possible coding rates: $r = \frac{1}{2}$, $\frac{2}{3}$, or $\frac{3}{4}$. The modulation types that can be used on the subchannels are BPSK, QPSK, 16QAM, or 64QAM.

Since the bandwidth B (and sampling rate $1/T_s$) is 20 MHz, and there are 64 subcarriers evenly spaced over that bandwidth, the subcarrier bandwidth is:

$$B_N = \frac{20\text{ MHz}}{64} = 312.5\text{ KHz}.$$

Since $\mu = 16$ and $1/T_s = 20\text{MHz}$, the maximum delay spread for which ISI is removed is

$$T_m < \mu T_s = \frac{16}{20\text{MHz}} = 0.8\ \mu\text{sec},$$

which corresponds to delay spread in an indoor environment. Including both the OFDM symbol and cyclic prefix, there are 80=64+16 samples per OFDM symbol time, so the symbol time per subchannel is

$$T_N = 80T_s = \frac{80}{20 \times 10^6} = 4\ \mu\text{s}$$

The data rate per subchannel is $\log_2 M/T_N$. Thus, the minimum data rate for this system, corresponding to BPSK (1 bit/symbol), an $r = \frac{1}{2}$ code, and taking into account that only 48 subcarriers actually carry usable data, is given by

$$\begin{aligned} R_{min} &= 48\text{ subcarriers} \times \frac{1/2\text{ bit}}{\text{codedbit}} \times \frac{1\text{ coded bit}}{\text{subcarrier symbol}} \times \frac{1\text{ subcarrier symbol}}{4 \times 10^{-6}\text{ seconds}} \\ &= 6\text{ Mbps} \end{aligned} \tag{12.49}$$

The maximum data rate that can be transmitted is

$$R_{max} = 48\text{ subcarriers} \times \frac{3/4\text{ bit}}{\text{coded bit}} \times \frac{6\text{ coded bits}}{\text{subcarrier symbol}} \frac{1\text{ subcarrier symbol}}{4 \times 10^{-6}\text{ seconds}} = 54\text{ Mbps}. \tag{12.50}$$

Naturally, a wide range of data rates between these two extremes is possible.

Example 12.6: Find the data rate of an 802.11a system assuming 16QAM modulation and rate 2/3 coding.

Solution: With 16QAM modulation each subcarrier transmits $\log_2(16) = 4$ coded bits per subcarrier symbol and there are a total of 48 subcarriers used for data transmission. With a rate 2/3 code, each coded bit relays 2/3 of an information bit per T_N seconds. Thus, the data rate is given by

$$R_{max} = 48\text{ subcarriers} \times \frac{2/3\text{ bit}}{\text{coded bit}} \frac{4\text{ coded bits}}{\text{subcarrier symbol}} \frac{1\text{ subcarrier symbol}}{4 \times 10^{-6}\text{seconds}} = 32\text{ Mbps}. \tag{12.51}$$

Bibliography

[1] J. Bingham, “Multicarrier modulation for data transmission: an idea whose time has come,” *IEEE Commun. Mag.* Vol. 28, No. 5, pp. 5-14, May 1990.

[2] L.J. Cimini, B. Daneshrad. N.R. Sollenberger, “Clustered OFDM with transmitter diversity and coding,” *Proc. Glob. Telecommun. Conf.*, pp. 703 - 707, Nov. 1996.

[3] H. Sari, G. Karam, and I. Jeanclaude, “Transmission techniques for digital terrestrial TV broadcasting,” *IEEE Commun. Mag.* Vol. 33, No. 2, pp. 100-109, Feb. 1995.

[4] R.K. Jurgen, “Broadcasting with digital audio,” *IEEE Spectrum*, pp. 52-59, March 1996 Pages:52 - 59

[5] J.S. Chow, J.C. Tu, and J.M. Cioffi, “A discrete multitone transceiver system for HDSL applications,” *IEEE J. Select. Areas. Commun.*, Vol. 9, No. 6, pp. 895–908, Aug. 1991.

[6] I. Kalet and N. Zervos, “Optimized decision feedback equalization versus optimized orthogonal frequency division multiplexing for high-speed data transmission over the local cable network,” *Proc. of ICC’89,* pp. 1080–1085, Sept. 1989.

[7] L.J. Cimini, “Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing,” *IEEE Trans. Inform. Theory,* Vol. 33, No. 7, pp. 665–675, July 1985.

[8] P.S. Chow, J.M. Cioffi, and John A.C. Bingham, “A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels,” *IEEE Trans. Commun.*, Vol. 43, No. 2/3/4, Feb.-Apr. 1995.

[9] Z. Wang, X. Ma, and G.B. Giannakis, “OFDM or single-carrier block transmissions?,” *IEEE Trans. Commun.*, Vol. 52 , No. 3, pp. 380-394, March 2004.

[10] J. M. Cioffi. *Digital Communications, Chapter 4: Multichannel Modulation*. Unpublished course notes, available at http://www.stanford.edu/class/ee379c/.

[11] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-Time Signal Processing*, 2nd. Ed., New York, 1999.

[12] J. M. Cioffi. A multicarrier primer. Stanford University/Amati T1E1 contribution, I1E1.4/91-157, Nov. 1991.

[13] M. Corson, R. Laroia, A. O’Neill, V. Park, and G. Tsirtsis. “A new paradigm for IP-based cellular networks,”. *IT Professional,* 3(6):20–29, November-December 2001.

[14] C. Eklund, R. B. Marks, K. L. Stanwood, and S. Wang, “IEEE Standard 802.16: A technical overview of the WirelessMAN 326 air interface for broadband wireless access, *IEEE Commun. Mag.*, pp. 98–107, June 2002.

[15] S. Hara and R. Prasad. “Overview of multicarrier CDMA,” *IEEE Commun. Mag.*, Vol. 35, pp. 126–33, Dec. 1997.

[16] L.H. Brandenburg and A.D. Wyner, “Capacity of the Gaussian channel with memory: the multivariate case,” *Bell System Tech. J.*, Vol. 53, No. 5, pp. 745-778, May-June 1974.

[17] S. Kasturia, J. Aslanis, and J. Cioffi. Vector coding for partial response channels. *IEEE Trans. on Info. Theory*, Vol. 36, pp. 741-762, July 1990.

[18] W. Lu. “4G mobile research in asia,” *IEEE Commun. Mag.*, pp. 104-106, Mar. 2003.

[19] S. Kaider, “Performance of multi-carrier CDM and COFDM in fading channels,” *Proc. Global Telecommun. Conf.*, pp. 847 - 851, Dec. 1999.

[20] P. Hoeher, S. Kaiser, and P. Robertson, “Two-dimensional pilot-symbol-aided channel estimation by Wiener filtering,” *Proc. IEEE Int. Conf. Acous., Speech, Sign. Proc. (ICASSP)*, pp. 1845 - 1848, April 1997.

[21] A. Scaglione, G.B. Giannakis, and S. Barbarossa, “Redundant filterbank precoders and equalizers. I. Unification and optimal designs, *IEEE Trans. Sign. Proc*, Vol. 47, No. 7, pp. 1988 - 2006, July 1999.

[22] A. Scaglione, G.B. Giannakis, and S. Barbarossa, “Redundant filterbank precoders and equalizers. II: Blind channel estimation, synchronization, and direct equalization, *IEEE Trans. Sign. Proc*, Vol. 47, No. 7, pp. 2007-2022, July 1999.

[23] R.G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[24] G.L. Stuber, J.R. Barry, S.W. McLaughlin, Y. Li, M.A. Ingram, T.G. Pratt, “Broadband MIMO-OFDM wireless communications,” *Proc. IEEE*, Vol. 92, No. 2, pp. 271-294, Feb. 2004.

[25] A. R. S. Bahai and B. R. Saltzberg, *Multi-Carrier Digital Communications - Theory and Applications of OFDM*, Kluwer Academic Publisher: Plenum Press, 1999.

[26] IEEE 802.11a-1999: High-speed physical layer in the 5 GHz band, 1999.

[27] IEEE 802.16a-2001 IEEE recommended practice for local and metropolitan area networks, 2001.

[28] IEEE 802.11g-2003: Further Higher-Speed Physical Layer Extension in the 2.4 GHz Band, 2003.

[29] T. H. Meng, B. McFarland, D. Su, and J. Thomson. “Design and implementation of an all-CMOS 802.11a wireless LAN chipset, *IEEE Commun. Mag.*, Vol. 41, pp. 160-168, Aug. 2003.

[30] C. Muschallik. Improving an OFDM reception using an adaptive nyquist windowing. *IEEE Trans. Consumer Electron.*, 42(3):259–69, Aug. 1996.

[31] K. G. Paterson and V. Tarokh. On the existence and construction of good codes with low peak-to-average power ratios. *IEEE Trans. on Info. Theory*, 46(6):1974–87, Sept. 2000.

[32] T. S. Rappaport, A. Annamalai, R. M. Buehrer, and W. H. Tranter. “ Wireless communications: Past events and a future perspective,” *IEEE Commun. Mag.*, pp. 148–61, May 2002.

[33] A. Redfern. “Receiver window design for multicarrier communication systems,” *IEEE J. Select. Areas Commun.*, Vol. 20, pp. 1029–36, June 2002.

[34] W. Rhee and J. M. Cioffi. “Increase in capacity of multiuser OFDM system using dynamic subchannel allocation,” In *Proc., IEEE Vehic. Technol.Conf.*, pp. 1085-1089, May 2000.

[35] H. Sampath, S. Talwar, J. Tellado, V. Erceg, and A. Paulraj. A fourth-generation MIMO-OFDM broadband wireless system: design, performance, and field trial results. *IEEE Communications Magazine*, 40(9):143–9, Sept. 2002.

[36] T. M. Schmidl and D. C. Cox. Robust frequency and timing synchronization for OFDM. *IEEE Trans. on Communications*, 45(12):1613 – 21, Dec. 1997.

[37] Z. Shen, J. Andrews, and B. Evans. “Optimal power allocation for multiuser OFDM,” *Proc. IEEE Glob. Commun. Conf.*, Dec. 2003.

[38] J. Tellado. *Multicarrier Modulation with low PAR: Applications to DSL and wireless*. Kluwer Academic Publishers, Boston, 2000.

[39] C. Wong, R. Cheng, K. Letaief, and R. Murch. “Multiuser OFDM with adaptive subcarrier, bit, and power allocation,” *IEEE J. Select. Areas Commun.*, Vol. 17, pp. 1747-1758, Oct. 1999.

[40] D.J.G Mestdagh, P.M.P. Spruyt, “A method to reduce the probability of clipping in DMT-based transceivers,” *IEEE Trans. Commun.*, Vol. 44, pp. 1234 - 1238, Oct. 1996.

[41] H. Ochiai and H. Imai, “On the distribution of the peak-to-average power ratio in OFDM signals,” *IEEE Trans. Commun.*, vol. 49, pp. 282-289, Feb. 2001.

Chapter 12 Problems

1. Show that the minimum separation for subcarriers $\{\cos(2\pi j/T_N + \phi_j), j = 1, 2 \ldots\}$ to form a set of orthonormal basis functions on the interval $[0, T_n]$ is $1/T_N$ for any initial phase ϕ_j. Show that if $\phi_j = 0 \forall j$ then this carrier separation can be reduced by half.

2. Consider an OFDM system operating in a channel with coherence bandwidth $B_c = 10$ KHz.

   (a) Find a subchannel symbol time $T_N = 1/B_N = 10T_m$, assuming $T_m = 1/B_c$. This should insure flat-fading on the suchannels.

   (b) Assume the system has $N = 128$ subchannels. If raised cosine pulses with $\beta = 1.5$ are used, and the required additional bandwidth due to time limiting to insure minimal power outside the signal bandwidth is $\epsilon = .1$, what is the total bandwidth of the system?

   (c) Find the total required bandwidth of the system using overlapping carriers separated by $1/T_N$, and compare with your answer in part (c).

3. Show from the definition of the DFT that circular convolution of discrete-time sequences leads to multiplication of their DFTs.

4. Consider a high-speed data signal with bandwidth .5 MHz and a data rate of .5 Mbps. The signal is transmitted over a wireless channel with a delay spread of 10 μsec.

   (a) If multicarrier modulation with nonoverlapping subchannels is used to mitigate the effects of ISI, approximately how many subcarriers are needed? What is the data rate and symbol time on each subcarrier? (We do not need to eliminate the ISI completely. So $T_s = T_m$ is enough)

   Assume for the remainder of the problem that the average received SNR (γ_s) on the nth subcarrier is $1000/n$ (linear units) and that each subcarrier experiences flat Rayleigh fading (so ISI is completely eliminated).

   (b) Suppose BPSK modulation is used for each subcarrier. If a repetition code is used across all subcarriers (i.e. a copy of each bit is sent over each subcarrier) then what is the BER after majority decoding? What is the data rate of the system?

   (c) Suppose you use adaptive loading (i.e. use different constellations on each subcarrier) such that the average BER on each subcarrier does not exceed 10^{-3} (this is averaged over the fading distribution, do not assume that the TX and RX adapt power or rate to the instantaneous fade values). Find the MQAM constellation that can be transmitted over each subcarrier while meeting this average BER target. What is the total data rate of the system with adaptive loading?

5. Consider a multicarrier modulation transmission scheme with three nonoverlapping subchannels spaced 200 KHz apart (from carrier to carrier) with subchannel baseband bandwidth of 100 KHz.

   (a) For what values of the channel coherence bandwidth will the subchannels of your multicarrier scheme exhibit flat-fading (approximately no ISI)? For what values of the channel coherence bandwidth will the subcarriers of your multicarrier scheme exhibit independent fading? If the subcarriers exhibit correlated fading, what impact will this have on coding across subchannels?

   (b) Suppose you have a total transmit power $P = 300$ mW, and the noise power in each subchannel is 1 mW. With equal power of 100 mW transmitted on each subchannel, the received SNR on each subchannel is $\gamma_1 = 11$ dB, $\gamma_2 = 14$ dB, and $\gamma_3 = 18$ dB. Assume the subchannels do not experience fading, so these SNRs are constant. For these received SNRs find the maximum signal constellation size for MQAM that can be transmitted over each subchannel for a target BER of 10^{-3}. Assume the MQAM constellation is restricted to be a power of 2 and use the BER bound BER $\leq .2e^{-1.5\gamma/(M-1)}$ for your calculations. What is the corresponding total data rate of the multicarrier signal, assuming a symbol rate on each subchannel of $T_s = 1/B$, where B is the baseband subchannel bandwidth?

(c) For the subchannel SNRs given in part (b), suppose we want to use precoding to equalize the received SNR in each subchannel and then send the same signal constellation over each subchannel. What size signal constellation is needed to achieve the same data rate as in part (b)? What transmit power would be needed on each subchannel to achieve the required received SNR for this constellation with a 10^{-3} BER target? How much must the total transmit power be increased over the 300 mW transmit power in part (b)?

6. Consider a channel with impulse response

$$h(t) = \alpha_0\delta(t) + \alpha_1\delta(t-T_1) + \alpha_2\delta(t-T_2).$$

Assume that $T_1 = 10$ μsecs and $T_2 = 20$ μsecs. You want to design a multicarrier system for the channel, with subchannel bandwidth $B_N = B_c/2$. If raised cosine pulses with $\beta = 1$ are used, and the subcarriers are separated by the minimum bandwidth necessary to remain orthogonal, then what is the total bandwidth occupied by a multicarrier system with 8 subcarriers? Assuming a constant SNR on each subchannel of 20 dB, what is the maximum constellation size for MQAM modulation that can be sent over each subchannel with a target BER of 10^{-3}, assuming M is restricted to be a power of 2. Also find the corresponding total data rate of the system.

7. Show that the matrix representations and (12.22) and (12.24) for the DMT system with a cyclic prefix appended to the input are equivalent.

8. Show that the DFT operation on $x[n]$ can be represented by the matrix multiplication $X[i] = \mathbf{Q}x[n]$ where

$$Q = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & W_N & W_N^2 & \dots & W_N^{N-1} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \dots & W_N^{(N-1)^2} \end{bmatrix}, \qquad (12.52)$$

for $W_N = e^{-j\frac{2\pi}{N}}$.

9. This problem shows that the rows of the DFT matrix $\mathbf{Q}$ are eigenvectors of $\mathbf{H}$.

(a) Show that the first row of $\mathbf{Q}$ is an eigenvector of $\mathbf{H}$ with eigenvalue $\lambda_0 = \sum_{i=0}^{\mu} h_i$.

(b) Show that the second row of $\mathbf{Q}$ is an eigenvector of $\mathbf{H}$ with eigenvalue $\lambda_1 = \sum_{i=0}^{\mu} h_i W_N^i$.

(c) Argue by induction that similar relations hold for all rows of $\mathbf{Q}$.

10. Show that appending the all-zero prefix to an OFDM symbol and then adding in the tail of the received sequence, as shown in Figure 12.8, results in the same received sequence as with a cyclic prefix.

11. Show that the two matrix representations of the DMT given by (12.22) and (12.24), are equivalent.

12. Consider a discrete-time FIR channel with $h[n] = .7 + .5\delta[n-1] + .3\delta[n-3]$. Consider an OFDM system with $N = 8$ subchannels.

(a) Find the matrix $\mathbf{H}$ corresponding to the matrix representation of DMT $\mathbf{y} = \mathbf{H}\mathbf{x} + \nu$ given in (12.23).

(b) Find the circulant convolution matrix $\mathbf{H}$ corresponding to the matrix representation in (12.25), as well as its eigenvalue decomposition $\mathbf{H} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^H$.

(c) What are the flat-fading channel gains associated with each subchannel in the representation of part (b)?

13. Consider a five-tap discrete-time channel

$$H(z) = 1 + 0.6z^{-1} + .7z^{-2} + .3z^{-3} + .2z^{-4}$$

Assume this channel model characterizes the maximum delay spread of the channel. Assume a VC system is used over this channel with $N = 256$ carriers.

(a) What value of μ is needed for the prefix to eliminate ISI between VC symbols. What is the overhead associated with this μ.

(b) Find the system matrix representation (12.23) and the singular values of the associated channel matrix $\mathbf{H}$.

(c) Find the transmit precoding and shaping matrices, V and U^H, required to orthogonalize the subchannels.

14. Suppose the 4 subchannels in 802.11a used for pilot estimation could be used for data transmission by taking advantage of blind estimation techniques. What maximum and minimum data rates could be achieved by including these extra subchannels, assuming the same modulation and coding formats are available.

15. Find the data rate of an 802.11a system assuming half the available 48 subchannels use BPSK with a rate 1/2 channel code and the others use 64QAM with a rate 3/4 channel code.

16. Find the PAR of a raised cosine pulse with $\beta = 0, 1, 2$. Which pulse shape has the lowest PAR? Is this pulse shape more or less sensitive to timing errors?

17. Find the constant C_0 associated with intercarrier interference in (12.48).

Chapter 14

Multiuser Systems

In multiuser systems the system resources must be divided among multiple users. This chapter develops techniques to allocate resources among multiple users, as well as the fundamental capacity limits of multiuser systems. We know from Chapter **??** that signals of bandwidth B and time duration T occupy a signal space of dimension $2BT$. In order to support multiple users, the signal space dimensions of a multiuser system must be allocated to the different users [1]. Allocation of signaling dimensions to specific users is called multiple access[2]. Multiple access methods perform differently in different multiuser channels, and we will apply these methods to the two basic multiuser channels, downlink channels and uplink channels. Because signaling dimensions can be allocated to different users in an infinite number of different ways, multiuser channel capacity is defined by a **rate region** rather than a single number. This region describes all user rates that can be simultaneously supported by the channel with arbitrarily small error probability. We will discuss multiuser channel capacity regions for both the uplink and the downlink. We also consider random access techniques, whereby signaling dimensions are only allocated to active users, as well as power control, which insures that users maintain the SINR required for acceptable performance. The performance benefits of multiuser diversity, which exploits the time-varying nature of the user's channels, is also described. We conclude with a discussion of the performance gains and signaling techniques associated with multiple antennas in multiuser systems.

14.1 Multiuser Channels: The Uplink and Downlink

A multiuser channel refers to any channel that must be shared among multiple users. There are two different types of multiuser channels: the **uplink** channel and the **downlink** channel, which are illustrated in Figure 14.1. A downlink, also called a broadcast channel or forward channel, has one transmitter sending to many receivers. Since the signals transmitted to all users originate from the downlink transmitter, the transmitted signal $s(t) = \sum_{k=1}^{K} s_k(t)$, with total power P and bandwidth B, is the sum of signals transmitted to all K users. Thus, the total signaling dimensions and power of the transmitted signal must be divided among the different users. Synchronization of the different users is relatively easy in the downlink since all signals originate from the same transmitter, although multipath in the channel can corrupt this synchronization. Another important characteristic of the downlink is that both signal and interference are distorted by the same channel. In particular, user k's signal $s_k(t)$ and all interfering signals $s_j(t), j \neq k$ pass through user k's channel $h_k(t)$ to arrive at user k's receiver. This is a fundamental difference between the uplink and the downlink, since in the uplink signals from different users are distorted by different

[1] Allocation of signaling dimensions through either multiple access or random access is performed by the Medium Access Control layer in the Open Systems Interconnection (OSI) network model [1, Chapter 1.3].

[2] The dimensions allocated to the different users need not be orthogonal, as in the superposition coding technique discussed in Section 14.5.

channels. Examples of wireless downlinks include all radio and television broadcasting, the transmission link from a satellite to multiple ground stations, and the transmission link from a base station to the mobile terminals in a cellular system.

An uplink channel, also called a multiple access channel[3] or reverse channel, has many transmitters sending signals to one receiver, where each signal must be within the the total system bandwidth B. However, in contrast to the downlink, in the uplink each user has an individual power constraint P_k associated with its transmitted signal $s_k(t)$. In addition, since the signals are sent from different transmitters, these transmitters must coordinate if signal synchronization is required. Figure 14.1 also indicates that the signals of the different users in the uplink travel through different channels, so even if the transmitted powers P_k are the same, the received powers associated with the different users will be different if their channel gains are different. Examples of wireless uplinks include laptop wireless LAN cards transmitting to a wireless LAN access point, transmissions from ground stations to a satellite, and transmissions from mobile terminals to a base station in cellular systems.

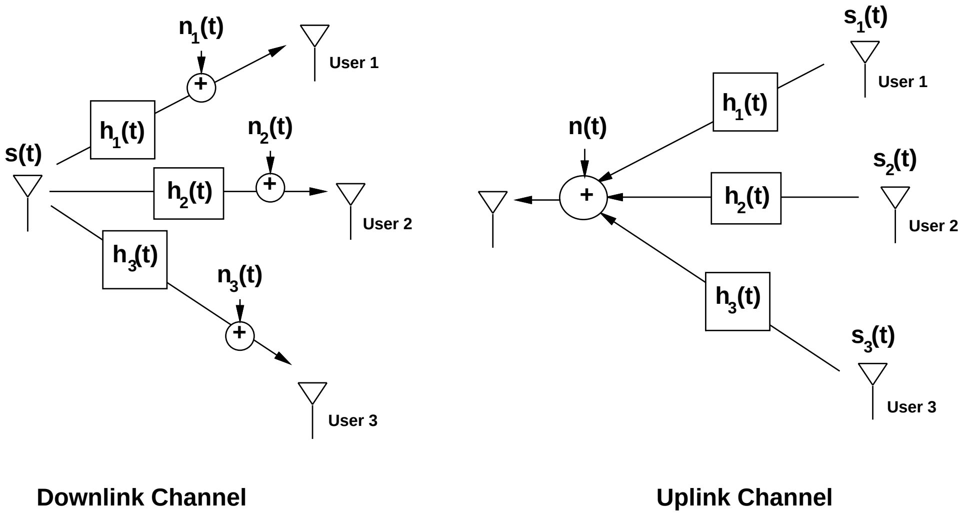


Figure 14.1: Downlink and Uplink Channels.

Most communication systems are bi-directional, and hence consist of both uplinks and downlinks. The radio transceiver that sends to users over a downlink channel and receives from these users over an uplink channel is often refered to as an access point or base station. It is generally not possible for radios to receive and transmit on the same frequency band due to the interference that results. Thus, bi-directional systems must separate the uplink and downlink channels into orthogonal signaling dimensions, typically using time or frequency dimensions. This separation is called **duplexing**. In particular, time-division duplexing (TDD) assigns orthogonal timeslots to a given user for receiving from an access point and transmitting to the access point, and frequency-division duplexing (FDD) assigns separate frequency bands for transmitting to and receiving from the access point. An advantage of TDD is that bi-directional channels are typically symmetrical in their channel gains, so channel measurements made in one direction can be used to estimate the channel in the other direction. This is not necessarily the case for FDD in frequency-selective fading: if the frequencies assigned to each direction are separated by more than the coherence bandwidth associated with the channel multipath, then these channels will exhibit independent fading.

[3] Note that multiple access techniques must be applied to both multiple access channels, i.e. uplinks, as well as to downlinks

14.2 Multiple Access

Efficient allocation of signaling dimensions between users is a key design aspect of both uplink and downlink channels, since bandwidth is usually scarce and/or very expensive. When dedicated channels are allocated to users it is often called **multiple access**[4]. Applications with continuous transmission and delay constraints, such as voice or video, typically require dedicated channels for good performance to insure their transmission is not interrupted. Dedicated channels are obtained from the system signal space using a channelization method such as time-division, frequency-division, code-division, or hybrid combinations of these techniques. Allocation of signaling dimensions for users with bursty transmissions generally use some form of random channel allocation which does not guarantee channel access. Bandwidth sharing using random channel allocation is called random multiple access or simply **random access**, which will be described in Section 14.3. In general, the choice of whether to use multiple access or random access, and which specific multiple or random access technique to apply, will depend on the system applications, the traffic characteristics of the users in the system, the performance requirements, and the characteristics of the channel and other interfering systems operating in the same bandwidth.

Multiple access techniques divide up the total signaling dimensions into channels and then assign these channels to different users. The most common methods to divide up the signal space are along the time, frequency, and/or code axes. The different user channels are then created by an orthogonal or non-orthogonal division along these axes: time-division multiple access (TDMA) and frequency-division multiple access (FDMA) are orthogonal channelization methods whereas code-division multiple access (CDMA) can be orthogonal or non-orthogonal, depending on the code design. Directional antennas, often obtained through antenna array processing, add an additional angular dimension which can also be used to channelize the signal space: this technique is called space-division multiple access (SDMA). The performance of different multiple access methods depends on whether they are applied to an uplink or downlink, and their specific characteristics. TDMA, FDMA, and orthogonal CDMA are all equivalent in the sense that they orthogonally divide up the signaling dimensions, and they therefore create the same number of orthogonal channels. In particular, given a signal space of dimension $2BT$, N orthogonal channels of dimension $2BT/N$ can be created, regardless of the channelization method. As a result, all multiple access techniques that divide the signal space orthogonally have the same channel capacity in AWGN, as will be discussed in Sections 14.5-14.6. However, channel impairments such as flat and frequency-selective fading affect these techniques in different ways, which lead to different channel capacities and different performance in practice.

14.2.1 Frequency-Division Multiple Access (FDMA)

In FDMA the system signaling dimensions are divided along the frequency axis into nonoverlapping channels, and each user is assigned a different frequency channel, as shown in Figure 14.2. The channels often have guard bands between them to compensate for imperfect filters, adjacent channel interference, and spectral spreading due to Doppler. If the channels are sufficiently narrowband then even if the total system bandwidth is large, the individual channels will not experience frequency-selective fading. Transmission is continuous over time, which can complicate overhead functions such as channel estimation since these functions must be performed simultaneously and in the same bandwidth as data transmission. FDMA also requires frequency-agile radios that can tune to the different carriers associated with the different channels. It is difficult to assign multiple channels to the same user under FDMA, since this requires the radios to simultaneously demodulate signals received over multiple frequency channels. FDMA is the most common multiple access option for analog communication systems, where transmission is continuous, and serves as the basis for the AMPS and ETACS analog cellular phone standards [2, Chapter 11.1]. Multiple access in OFDM systems, called OFDMA, implements FDMA by assigning different subcarriers to different users.

[4] An uplink channel is also referred to as a multiple access channel, however multiple access techniques are needed for both uplinks and downlinks.

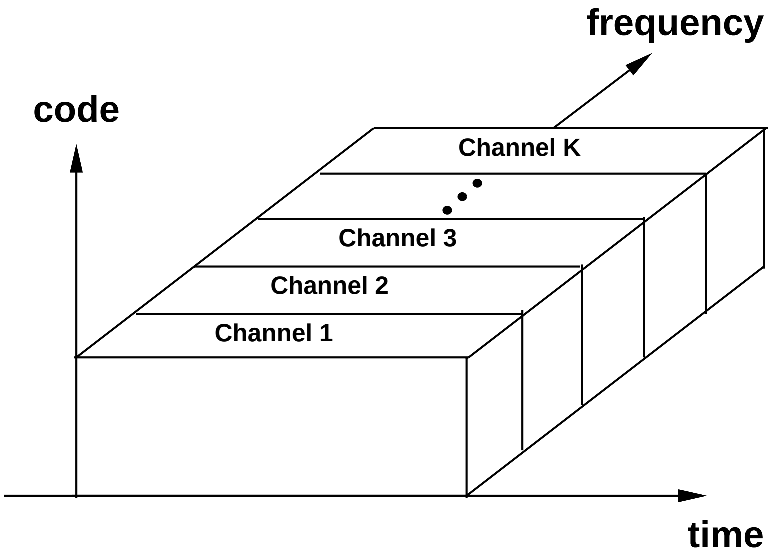


Figure 14.2: Frequency-Division Multiple Access.

Example 14.1: First-generation analog systems were allocated a total bandwidth of $B = 25$ MHz for uplink channels and another $B = 25$ MHz for downlink channels. This bandwidth allocation was split between two operators in every region, so each operator had 12.5 MHz for both their uplink and downlink channels. Each user was assigned $B_c = 30$ KHz of spectrum for its analog voice signal, corresponding to 24 KHz for the FM modulated signal and 3 KHz guardbands on each side. The total uplink and downlink bandwidths also requred guard bands of $B_g = 10$ KHz on each side to mitigate interference to and from adjacent systems. Find the total number of analog voice users that could be supported in the total 25 MHz of bandwidth allocated to the uplink and the downlink. Also consider a more efficient digital system with high-level modulation so that only 10 KHz channels are required for a digital voice signal with tighter filtering such that only 5 KHz guard bands are required on the band edges. How many users can be supported in the same 25 MHz of spectrum for this more efficient digital system?

Solution: For either the uplink or the downlink, with guard bands on each side of the voice channel, each user requires a total bandwidth of $B_c + 2B_g$. Thus, the total number of users that can be supported in the total uplink or downlink bandwidth $B = 25$ Khz is

$$N = \frac{B - 2B_g}{B_c} = \frac{25 \times 10^6 - 2 \times 10 \times 10^3}{30 \times 10^3} = 832,$$

or 416 users per operator. Indeed, first-generation analog systems could support 832 users in each cell. The digital system has

$$N = \frac{B - 2B_g}{B_c} = \frac{25 \times 10^6 - 2 \times 5 \times 10^3}{10 \times 10^3} = 2599$$

users that can be supported in each cell, almost a three-fold increase over the analog system. The increase is primarily due to the bandwidth savings of the high-level digital modulation, which can accommodate a voice signal in one third the bandwidth of the analog voice signal.

14.2.2 Time-Division Multiple Access (TDMA)

In TDMA the system dimensions are divided along the time axis into nonoverlapping channels, and each user is assigned a different cyclically-repeating timeslot, as shown in Figure 14.3. These TDMA channels occupy the entire system bandwidth, which is typically wideband, so some form of ISI mitigation is required. The cyclically-repeating timeslots imply that transmission is not continuous for any user. Therefore, digital transmission techniques which allow for buffering are required. The fact that transmission is not continuous simplifies overhead functions such as channel estimation, since these functions can be done during the timeslots occupied by other users. TDMA also has the advantage that it is simple to assign multiple channels to a single user by simply assigning him multiple timeslots.

A major difficulty of TDMA, at least for uplink channels, is the requirement for synchronization among the different users. Specifically, in a downlink channel all signals originate from the same transmitter and pass through the same channel to any given receiver. Thus, for flat-fading channels, if users transmit on orthogonal timeslots the received signal will maintain this orthogonality. However, in the uplink channel the users transmit over different channels with different respective delays. To maintain orthogonal timeslots in the received signals, the different uplink transmitters must synchronize such that *after* transmission through their respective channels, the received signals are orthogonal in time. This synchronization is typically coordinated by the base station or access point, and can entail significant overhead. Multipath can also destroy time-division orthogonality in both uplinks and downlinks if the multipath delays are a significant fraction of a timeslot. TDMA channels therefore often have guard bands between them to compensate for synchronization errors and multipath. Another difficulty of TDMA is that with cyclically repeating timeslots the channel characteristics change on each cycle. Thus, receiver functions that require channel estimates, like equalization, must re-estimate the channel on each cycle. When transmission is continuous, the channel can be tracked, which is more efficient. TDMA is used in the GSM, PDC, IS-54, and IS-136 digital cellular phone standards [2, Chapter 11].

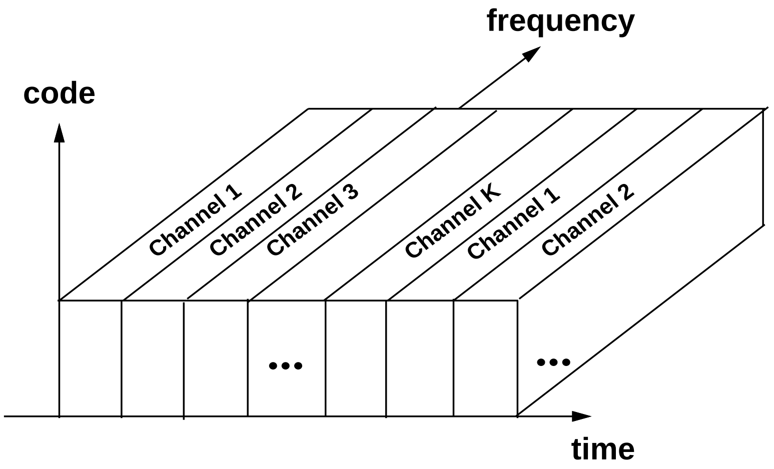


Figure 14.3: Time-Division Multiple Access.

Example 14.2: The original GSM design uses 25 MHz of bandwidth for the uplink and for the downlink, the same as AMPs. This bandwidth is divided into 125 TDMA channels of 200 KHz each. Each TDMA channel consists of 8 user timeslots: the 8 timeslots along with a preamble and trailing bits form a frame, which is cyclically repeated in time. Find the total number of users that can be supported in the GSM system and the channel bandwidth of each user. If the rms delay spread of the channel is 10 μsecs, will ISI mitigation be needed in this system?

Solution: Since there are 8 users per channel and 125 channels, the total number of users that can be supported in this system is $125 \times 8 = 1000$ users. The bandwidth of each TDMA channel is $25 \times 10^6/125 = 200$ KHz. A delay spread of 10 μsecs corresponds to a channel coherence bandwidth of $B_c \approx 100$ KHz, which is less than the TDMA channel bandwidth of 200 KHz. Thus, ISI mitigation is needed. The GSM specification includes an equalizer to compensate for ISI, but the type of equalizer is at the discretion of the designer.

14.2.3 Code-Division Multiple Access (CDMA)

In CDMA the information signals of different users are modulated by orthogonal or non-orthogonal spreading codes. The resulting spread signals simultaneously occupy the same time and bandwidth, as shown in Figure 14.4. The receiver uses the spreading code structure to separate out the different users. The most common form of CDMA is multiuser spread spectrum with either DS or FH, which are described and analyzed in Chapters **??-??**.

Downlinks typically use orthogonal spreading codes such as Walsh-Hadamard codes, although the orthogonality can be degraded by multipath. Uplinks generally use non-orthogonal codes due to the difficulty of user synchronization and the complexity of maintaining code orthogonality in uplinks with multipath [5]. One of the big advantages of non-orthogonal CDMA in uplinks is that little dynamic coordination of users in time or frequency is required, since the users can be separated by the code properties alone. In addition, since TDMA and FDMA carve up the signaling dimensions orthogonally, there is a hard limit on how many orthogonal channels can be obtained. This is also true for CDMA using orthogonal codes. However, if non-orthogonal codes are used, there is no hard limit on the number of channels that can be obtained. However, because non-orthogonal codes cause mutual interference between users, the more users that simultaneously share the system bandwidth using non-orthogonal codes, the higher the level of interference, which degrades the performance of all the users. A non-orthogonal CDMA scheme also requires power control in the uplink to compensate for the near-far effect. The near-far effect arises in the uplink because the channel gain between a user's transmitter and the receiver is different for different users. Specifically, suppose that one user is very close to his base station or access point, and another user very far away. If both users transmit at the same power level, then the interference from the close user will swamp the signal from the far user. Thus, power control is used such that the received signal power of all users is roughly the same. This form of power control, which essentially inverts any attenuation and/or fading on the channel, causes each interferer to contribute an equal amount of power, thereby eliminating the near-far effect. CDMA systems with non-orthogonal spreading codes can also use MUD to reduce interference between users. MUD provides considerable performance improvement even under perfect power control, and works even better when the power control is jointly optimized with the MUD technique [6]. We will see in Sections 14.5-14.6 that CDMA with different forms of multiuser detection achieves the Shannon capacity of both the uplink and the downlink, although the capacity-achieving transmission and reception strategies for the two channels are very different. Finally, it is simple to allocate multiple channels to one user with CDMA by assigning that user multiple codes. CDMA is used for multiple access in the IS-95 digital cellular standards, with orthgonal spreading codes on the downlink and a combination of orthogonal and non-orthogonal codes on the uplink [2, Chapter 11.4]. It is also used in the W-CDMA and CDMA2000 digital cellular standards [4, Chapter 10.5].

Example 14.3: The SIR for a CDMA uplink with non-orthogonal codes under the standard Gaussian assumption was given in (**??**) as

$$\text{SIR} = \frac{3G}{(K-1)},$$

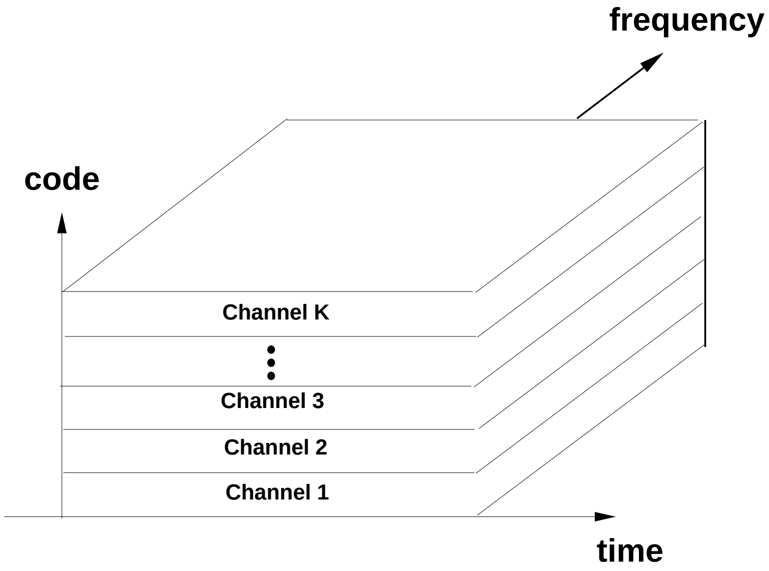


Figure 14.4: Code-Division Multiple Access.

where K is the number of users and $G \approx 128$ is the ratio of spread bandwidth to signal bandwidth. In IS-95 the uplink channel is assigned 1.25 MHz of spectrum. Thus, the bandwidth of the information signal prior to spreading is $B_s \approx 1.25 \times 10^6/128 = 9.765$ KHz. Neglecting noise, if the required SINR on a channel is 10 dB, how many users can the CDMA uplink support? How many could be supported within the same total bandwidth for an FDMA system?

Solution: To determine how many users can be supported, we invert the SIR expression to get

$$K \leq \frac{3G}{\text{SIR}} + 1 = \frac{256}{20} + 1 = 39.4,$$

and since K must be an integer, the system can support 39 users. In FDMA we have

$$K = \frac{1.25 \times 10^6}{9.765 \times 10^3} = 128,$$

so the total system bandwidth of 1.25 MHz can support 128 channels of 9.765 KHz. This calculation implies that FDMA is three times more efficient than non-orthogonal CDMA under the standard Gaussian assumption for code cross-correlation (FDMA is even more efficient under different assumptions about the code cross correlation). But in fact, IS-95 typically supports 64 users on the uplink and downlink by allowing variable voice compression rates depending on interference and channel quality and taking advantage of the fact that interference is not always present (called a voice-activity factor). While this makes CDMA less efficient than FDMA for a single cell, cellular systems have channel reuse, which can be done more efficiently in CDMA than in FDMA, as discussed in more detail in Chapter **??**.

14.2.4 Space-Division

Space-division multiple access (SDMA) uses direction (angle) as another dimension in signal space, which can be channelized and assigned to different users. This is generally done with directional antennas, as shown in Figure 14.5. Orthogonal channels can only be assigned if the angular separation between users exceeds the angular resolution of the directional antenna. If directionality is obtained using an antenna array, precise angular resolution requires a very large array, which may be impractical for the base station or access point and is certainly infeasible in small user terminals. In practice SDMA is often implemented using sectorized antenna arrays, discussed in Chapter 10.8. In these arrays the 360^o angular range is divided into N sectors. There is high directional gain in each sector and little interference between sectors. TDMA or FDMA is used to channelize users within a sector. For mobile users SDMA must adapt as user angles change or, if directionality is achieved via sectorized antennas, then a user must be handed off to a new sector when it moves out of its original sector.

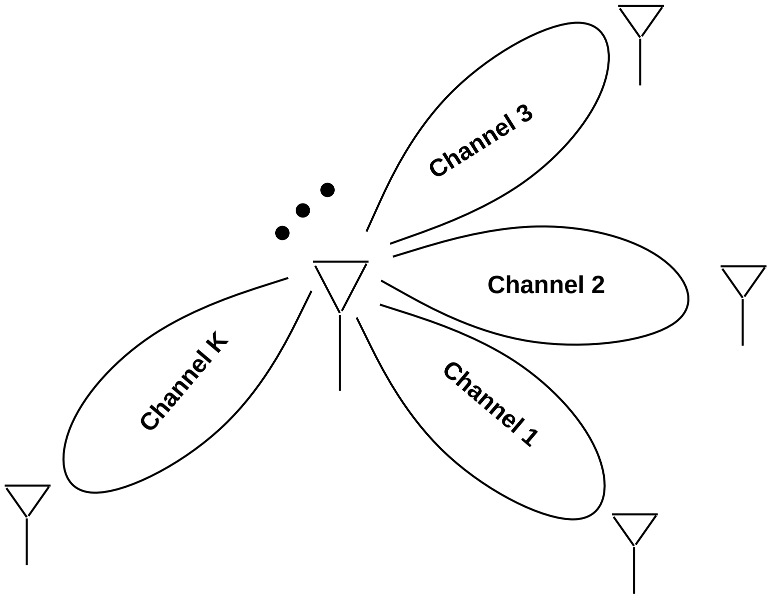


Figure 14.5: Space-Division Multiple Access.

14.2.5 Hybrid Techniques

Many systems use a combination of different multiple access schemes to allocate signaling dimensions. OFDMA can be combined with tone hopping to improve frequency diversity [9]. DSSS can be combined with FDMA to break the system bandwidth into subbands. In this hybrid method different users are assigned to different subbands with their signals spread across the subband bandwidth. Within a subband, the processing gain is smaller than it would be over the entire system bandwidth, so interference and ISI rejection is reduced. However, this technique does not require contiguous spectrum between subbands, and also allows more flexibility in spreading user signals over different size subbands depending on their requirements. Another hybrid method combines DS-CDMA with FH-CDMA so that the carrier frequency of the spread signal is hopped over the available bandwidth. This reduces the near-far effect since the interfering users change on each hop. Alternatively, TDMA and FH can be combined so that a channel with deep fading or interference is only used on periodic hops, so that the fading and interference effects can be mitigated by error correction coding. This idea is used in the GSM standard, which combines FH with its TDMA scheme to reduce the effect of strong interferers in other cells.

There has been much discussion, debate, and analysis about the relative performance of different multiple access techniques for current and future wireless systems, e.g. [8, 9, 10, 11, 12, 13]. While analysis and general conclusions can be made for simple system and channel models, it is difficult to come up with a definitive answer as to the best technique for a complex multiuser system under a range of typical operating conditions. Moreover, simplifying assumptions must be made to perform a comparative analysis or simulation study, and these assumptions can bias the results in favor of one particular scheme. As with most engineering design questions, the choice of which multiple access technique to use will depend on the system requirements and characteristics along with cost and complexity constraints.

14.3 Random Access

Multiple access techniques are primarily for continuous-time applications like voice and video, where a dedicated channel facilitates good performance. However, most data users do not require continuous transmission: data is generated at random time instances, so dedicated channel assignment can be extremely inefficient. Moreover, most systems have many more total users (active plus idle users) than can be accommodated simultaneously, so at any given time channels can only be allocated to users that need them. Random access strategies are used in such systems to efficiently assign channels to the active users.

All random access techniques are based on the premise of packetized data or **packet radio**. In packet radio user data is collected into packets of N bits, and once a packet is formed it is transmitted over the channel. Assuming a fixed channel data rate of R bps, the transmission time of a packet is $\tau = N/R$. The transmission rate R is assumed to require the entire signal bandwidth, and all users transmit their packets over this bandwidth, with no additional coding that would allow separation of simultaneously transmitted packets. Thus, if packets from different users overlap in time a **collision** occurs, in which case neither packet may be decoded successfully. Analysis of random access techniques typically assumes that collectively the users accessing the channel generate packets according to a Poisson process at a rate of λ packets per unit time, i.e. λ is the average number of packets that arrive in any time interval $[0, t]$ divided by t. Equivalently, λN is the average number of bits generated in any time inteval $[0, t]$ divided by t. For a Poisson process, the probability that the number of packet arrivals in a time period $[0, t]$, denoted as $X(t)$, is equal to some integer k is given by

$$p(X(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \tag{14.1}$$

Poisson processes are memoryless, so that the number of packet arrivals during any given time period does not affect the distribution of packet arrivals in any other time period. Note that the Poisson model is not necessarily a good model for all types of user traffic, especially Internet data, where bursty data causes correlated packet arrivals [14].

The traffic **load** on the channel given Poisson packet arrivals at rate λ and packet transmission duration τ is defined as $L = \lambda\tau$. If the channel data rate is R_p packets per second then $\tau = 1/R_p = N/R$ for R the channel data rate in bps. Note that L is unitless: it is the ratio of the packet arrival rate divided by the packet rate that can be transmitted over the channel at the channel's data rate R. If $L > 1$ then on average more packets (or bits) arrive in the system over a given time period than can be transmitted in that period, so systems with $L > 1$ are unstable. If the transmitter is informed by the receiver about packets received in error and retransmits these packets, then the packet arrival rate λ and corresponding load $L = \lambda\tau$ is computed based on arrivals of both new packets and packets that require retransmission. In this case L is referred to as the **total offered load**.

Performance of random access techniques is typically characterized by the **throughput** T of the system. The throughput, which is unitless, is defined as the ratio of the average number of packets successfully transmitted in any given time interval divided by the number of attempted transmissions in that interval. The throughput thus equals the offered load multiplied by the probability of successful packet reception, $T = Lp(\text{successful packet reception})$, where this probability is a function of the random access protocol in use as well as the channel characteristics, which can cause packet errors in the absence of collisions. If we assume that colliding packets always cause errors, then $T \leq L$, since no more than one packet can be successfully transmitted at any one time. Moreover, since a system with $L > 1$ is unstable, stable systems where colliding packets always cause errors have $T \leq L \leq 1$. Note that the throughput is independent of the channel data rate R, since the load and corresponding throughput are normalized with respect to this rate. This allows analysis of random access protocols to be generic to any underlying link design or channel capacity. For a packet radio with a link data rate of R bps, the **effective data rate** of the system is RT, since T is the fraction of packets or bits successfully transmitted at rate R. The goal of a random access method is to make T as large as possible in order to fully utilize the underlying link rates. Note that in some circumstances overlapping packets do not cause a collision. In particular, short periods of overlap between colliding packets, different channel gains on the received packets, and/or error correction coding can allow one or more packets to be successfully received even with a collision. This is called the **capture effect** [15, Chapter 4.3].

Random access techniques were pioneered by Abramson with the ALOHA protocol [16], where data is packetized and users send packets whenever they have data to send. ALOHA is very inefficient due to collisions between users, which leads to very low throughput. The throughput can be doubled by slotting time and synchronizing the users, but even then collisions lead to relatively low throughput values. Modifications to ALOHA protocols to avoid collisions and thereby increase throughput include carrier sensing, collision detection, and collision avoidance. Long bursts of packets can be scheduled to avoid collisions, but this typically takes additional overhead. In this section we will describe the various techniques for random access, their performance, and their design tradeoffs.

14.3.1 Pure ALOHA

In pure or unslotted ALOHA users transmit data packets as soon as they are formed. If we neglect the capture effect, then packets that overlap in time are assumed to be received in error, and must be retransmitted. If we also assume packets that do not collide are successfully received (i.e. there is no channel distortion or noise), then the throughput equals the offered load times the probability of no collisions: $T = Lp(\text{no collisions})$. Suppose a given user transmits a packet of duration τ during time $[0, \tau]$. Then if any other user generates a packet during time $[-\tau, \tau]$, that packet, of duration τ, will overlap with the transmitted packet, causing a collision. From (14.1), the probability that no packets are generated during the time $[-\tau, \tau]$ is given by (14.1) with $t = 2\tau$:

$$p(X(t) = 0) = e^{-2\lambda\tau} = e^{-2L}, \tag{14.2}$$

with corresponding throughput

$$T = Le^{-2L}. \tag{14.3}$$

This throughput is plotted in Figure 14.6, where we see that throughput increases with offered load up to a maximum throughput of approximately .18 for $L = .5$, after which point it decreases. In other words, the data rate is only 18% of what it would be with a single user transmitting continuously on the system. The reason for this maximum is that for small values of L there are many idle periods when no user is transmitting, so throughput is small. As L increases, the channel is utilized more but collisions also start to occur. At $L = .5$ there is the optimal balance between users generating enough packets to utilize the channel with reasonable efficiency and these packet generations colliding infrequently. Beyond $L = .5$ the collisions become more frequent, which degrades throughput below its maximum, and as L grows very large, most packets experience collisions, and throughput approaches zero.

Part of the reason for the inefficiency of pure ALOHA is the fact that users can start their packet transmissions at any time, and any partial overlap of two or more packets destroys the successful reception of all packets. By synchronizing users such that all packet transmissions are aligned in time, the partial overlap of packet transmissions can be avoided. That is the basic premise behind Slotted ALOHA.

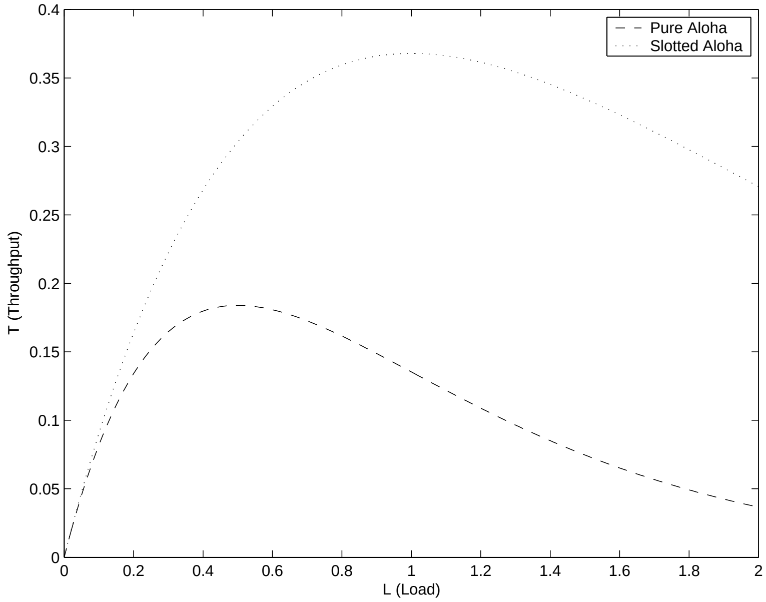


Figure 14.6: Throughput of Pure and Slotted ALOHA.

14.3.2 Slotted ALOHA

In slotted ALOHA, time is assumed to be slotted in timeslots of duration τ, and users can only start their packet transmissions at the beginning of the next timeslot after the packet has formed. Thus, there is no partial overlap of transmitted packets, which increases throughput. Specifically, a packet transmitted over the time period $[0, \tau]$ is successfully received if no other packets are transmitted during this period. This probability is obtained from (14.1) with $t = \tau$: $p(X(t) = 0) = e^{-L}$, with corresponding throughput

$$T = Le^{-L}. \tag{14.4}$$

This throughput is also plotted in Figure 14.6, where we see that throughput increases with offered load up to a maximum throughput of approximately $T = .37$ for $L = 1$, after which point it decreases. Thus, slotted ALOHA has double the maximum throughput as pure ALOHA, and achieves this maximum at a higher offered load. While this represents a marked improvement over pure ALOHA, the effective data rate is still less than 40% of the raw transmission rate. This is extremely wasteful of the limited wireless bandwidth, so more sophisticated techniques are needed to increase efficiency.

Note that slotted ALOHA requires synchronization of all nodes in the network, which can entail significant overhead. Even in a slotted system, collisions occur whenever two or more users attempt transmission in the same slot. Error control coding can result in correct detection of a packet even after a collision, but if the error correction is insufficient then the packet must be retransmitted. A study on design optimization between error correction and retransmission is described in [19].

Example 14.4: Consider a slotted ALOHA system with a transmission rate of $R = 10$ Mbps. Suppose packets consist of 1000 bits. For what packet arrival rate λ will the system achieve maximum throughput, and what is the effective data rate associated with this throughput?

Solution: The throughput T is maximized for $L = \lambda\tau = 1$, where λ is the packet arrival rate and τ is the packet duration. With a 10 Mbps transmission rate and 1000 bits/packet, $\tau = 1000/10^6 = .1$ ms. Thus, $\lambda = 1/.0001 = 10^4$ packets per second maximizes throughput. The throughput for $L = 1$ is $T = .37$, so the effective data rate is $TR = 3.7$ Mbps. Thus, the data rate is reduced by roughly a factor of 3 as compared to continuous data transmission due to the random nature of the packet arrivals and their corresponding collisions.

14.3.3 Carrier Sense Multiple Access

Collisions can be reduced by Carrier Sense Multiple Access (CSMA), where users sense the channel and delay transmission if they detect that another user is currently transmitting. To be effective, detection time and propagation delays in the system must be small [3, Chapter 4.19]. Typically a user waits to transmit a random time period after sensing a busy channel. This **random backoff** avoids multiple users simultaneously transmitting as soon as the channel is free. CSMA only works when all users can detect each other's transmissions and the propagation delays are small. Wired LANs have these characteristics, hence CSMA is part of the Ethernet protocol. However, the nature of the wireless channel may prevent a given user from detecting the signals transmitted by all other users. This gives rise to the **hidden terminal problem**, illustrated in Figure 14.7, where each node can hear its immediate neighbor but no other nodes in the network. In this figure both node 3 and node 5 wish to transmit to node 4. Suppose node 5 starts his transmission. Since node 3 is too far away to detect this transmission, he assumes that the channel is idle and begins his transmission, thereby causing a collision with node 5's transmission. Node 3 is said to be hidden from node 5 since it cannot detect node 5's transmission. ALOHA with CSMA also creates inefficiencies in channel utilization from the exposed terminal problem, also illustrated in Figure 14.7. Suppose the exposed terminal in this figure - node 2 - wishes to send a packet to node 1 at the same time node 3 is sending to node 4. When node 2 senses the channel it will detect node 3's transmission and assume the channel is busy, even though node 3 does not interfere with the reception of node 2's transmission by node 1. Thus node 2 will not transmit to node 1 even though no collision would have occurred. Exposed terminals only occur in multihop networks, so we will defer their discussion until Chapter 16.

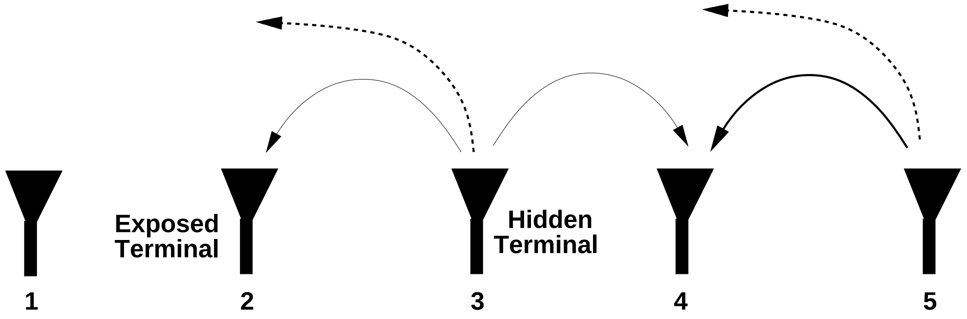


Figure 14.7: Hidden and Exposed Terminals.

The collisions introduced by hidden terminals are often avoided in wireless networks by a four-way handshake prior to transmission [20, 17]. This **collision avoidance** is done as follows. A node that wants to send a data packet will first wait for the channel to become available and then transmit a short RTS (Request To Send) packet. The potential receiver, assuming it perceives an available channel, will immediately respond with a CTS (Clear To Send) packet that authorizes the initiating node to transmit, and also informs neighboring hidden nodes (i.e., nodes that are outside the communication range of the transmitter but within the communication range of the receiver) that they will have to remain silent for the duration of the transmission. Nodes that overhear the RTS or CTS packet will refrain from transmitting over the expected packet duration. A node can only send an RTS packet if it perceives an idle channel and has not been silenced by another control packet. A node will only transmit a CTS packet if it has not been silenced by another control packet. The RTS/CTS handshake is typically coupled with random backoff to avoid all nodes transmitting as soon as the channel becomes available. In some incarnations [17, 18], including the 802.11 WLAN standard [4, Chapter 14.3], the receiver sends an ACK (Acknowledgement) packet back to the transmitter to verify when it has correctly received the packet, after which the channel again becomes available.

Another technique to avoid hidden terminals is busy tone transmission. In this strategy users first check to see whether the transmit channel is busy by listening for a "busy tone" on a separate control channel [1, Chapter 4.6]. There is typically not an actual busy tone but instead a bit is set in a predetermined field on the control channel. This scheme works well in preventing collisions when a centralized controller can be "heard" by users throughout the network. In a flat network without centralized control, more complicated measures are used to ensure that any potential interferer on the first channel can hear the busy tone on the second [21, 22]. Hybrid techniques using handshakes, busy tone transmission, and power control can also be used [22]. Collisions can also be reduced by combining DSSS with ALOHA. In this scheme each user modulates his signal with the same code, but if user transmissions are separated by more than a chip time, the interference due to a collision is reduced by the code autocorrelation [23].

14.3.4 Scheduling

Random access protocols work well with bursty traffic where there are many more users than available channels, yet these users rarely transmit. If users have long strings of packets or continuous stream data, then random access works poorly as most transmissions result in collisions. In this scenario performance can be improved by assigning channels to users in a more systematic fashion through transmission scheduling. In scheduled access the available bandwidth is channelized into multiple time, frequency, or code division channels. Each node schedules its transmission on different channels in such a way as to avoid conflicts with neighboring nodes while making the most efficient use of the available signaling dimensions.

Even with a scheduling access protocol, some form of ALOHA will still be needed since a predefined mechanism for scheduling will be, by definition of random access, unavailable at startup. ALOHA provides a means for initial contact and the establishment of some form of scheduled access for the transmission of relatively large amounts of data. A systematic approach to this initialization that also combines the benefits of random access for bursty data with scheduling for continuous data is packet reservation multiple access (PRMA) [24]. PRMA assumes a slotted system with both continuous and bursty users (e.g. voice and data users). Multiple users vie for a given time slot under a random access strategy. A successful transmission by one user in a given timeslot reserves that timeslot for all subsequent transmissions by the same user. If the user has a continuous or long transmission then after successfully capturing the channel he has a dedicated channel for the remainder of his transmission (assuming subsequent transmissions are not corrupted by the channel: this corruption causes users to lose their slots and they must then recontend for an unreserved slot, which can entail significant delay and packet dropping [25]). When this user has no more packets to transmit, the slot is returned to the pool of available slots that users attempt to capture via random access. Thus, data users with short transmissions benefit from the random access protocol assigned to unused slots, and users with continuous transmissions get scheduled periodic transmissions after successfully capturing an initial slot. A similar technique using a combined reservation and ALOHA policy is described in [26].

14.4 Power Control

Power control is applied to systems where users interfere with each other. The goal of power control is to adjust the transmit powers of all users such that the SINR of each user meets a given threshold required for acceptable performance. This threshold may be different for different users, depending on their required performance. This problem is straightforward for the downlink, where both users and interferers have the same channel gains, but is more complicated in the uplink, where the channel gains may be different. Seminal work on power control for cellular systems and ad-hoc networks was done in [30, 31, 32], and power control for the the uplink is a special case for which these results can be applied. In the uplink model, the kth transmitter has a fixed channel power gain g_k to the receiver. The quality of each link is determined by the SIR at the intended receiver. In an uplink with K interfering users we denote the SIR for the kth user as

$$\gamma_k = \frac{g_k P_k}{n + \rho \sum_{j \neq k} g_j P_j}, \quad k = 1, \ldots, K, \tag{14.5}$$

where P_k is the power of the kth transmitter, n is the receiver noise power, and ρ is interference reduction due to signal processing. For example, in a CDMA uplink the interference power is reduced by the processing gain of the code, so $\rho \approx 1/G$ for G the processing gain, whereas in TDMA $\rho = 1$.

Each link is assumed to have a minimum SIR requirement $\gamma_k^* > 0$. This constraint can be represented in matrix form with component-wise inequalities as

$$(\mathbf{I} - \mathbf{F})\mathbf{P} \geq \mathbf{u} \ \text{ with } \ \mathbf{P} > 0, \tag{14.6}$$

where $\mathbf{P} = (P_1, P_2, \ldots, P_K)^T$ is the column vector of transmitter powers,

$$\mathbf{u} = \left(\frac{n\gamma_1^*}{g_1}, \frac{n\gamma_2^*}{g_2}, \ldots, \frac{n\gamma_K^*}{g_K} \right)^T, \tag{14.7}$$

is the column vector of noise power scaled by the SIR constraints and channel gain, and $\mathbf{F}$ is a matrix with

$$F_{kj} = \begin{cases} 0, & \text{if } \ k = j \\ \frac{\gamma_k^* g_j \rho}{g_k}, & \text{if } \ k \neq j \end{cases} \tag{14.8}$$

with $k, j = 1, 2, \ldots, K$.

The matrix $\mathbf{F}$ has non-negative elements and is irreducible. Let ρ_F be the Perron-Frobenius eigenvalue of $\mathbf{F}$. This is the maximum modulus eigenvalue of $\mathbf{F}$, and for $\mathbf{F}$ irreducible this eigenvalue is simple, real, and positive. Moreover, from the Perron-Frobenius theorem and standard matrix theory [33], the following statements are equivalent:

1. $\rho_F < 1$
2. There exists a vector $\mathbf{P} > 0$ (i.e. $P_k > 0$ for all k) such that $(\mathbf{I} - \mathbf{F})\mathbf{P} \geq \mathbf{u}$
3. $(\mathbf{I} - \mathbf{F})^{-1}$ exists and is positive componentwise.

Furthermore, if any of the above conditions holds we also have that $\mathbf{P}^* = (\mathbf{I} - \mathbf{F})^{-1}\mathbf{u}$ is the Pareto optimal solution to (14.6). That is, if $\mathbf{P}$ is any other solution to (14.6) then $\mathbf{P} \geq \mathbf{P}^*$ componentwise. Hence, if the SIR requirements for all users can be met simultaneously, the best power allocation is $\mathbf{P}^*$ so as to minimize the transmit power of the users.

In [32] the authors also show that the following iterative power control algorithm converges to $\mathbf{P}^*$ when $\rho_F < 1$, and diverges to infinity otherwise. This iterative Foschini-Miljanic algorithm is given by

$$\mathbf{P}(i+1) = \mathbf{F}\mathbf{P}(i) + \mathbf{u}, \tag{14.9}$$

for $i = 1, 2, 3, \ldots$. Furthermore, the above algorithm can be simplified to a per-user version as follows. Let

$$P_k(i+1) = \frac{\gamma_k^*}{\gamma_k(i)} P_k(i), \tag{14.10}$$

for each link $k \in \{1, 2, \ldots, N\}$. Hence, each transmitter increases power when its SIR is below its target and decreases power when its SIR exceeds its target. SIR measurements or a function of them such as BER are typically made at the base station or access points, and a simple "up" or "down" command regarding transmit power can be fed back to each of the transmitters to perform the iterations. It is easy to show that (14.9) and (14.10) are pathwise equivalent and hence the per-user version of the power control algorithm also converges to $\mathbf{P}^*$. The feasible region of power vectors that achieve the SIR targets for a two-user system along with the iterative algorithms that converges to the minimum power vector in this region is illustrated in Figure 14.8. We see in this figure that the feasible region consists of all power pairs $\mathbf{P} = (P_1, P_2)$ that achieve a given pair of SIR targets, and the optimal pair $\mathbf{P}^*$ is the minimum power vector in this two-dimensional region.

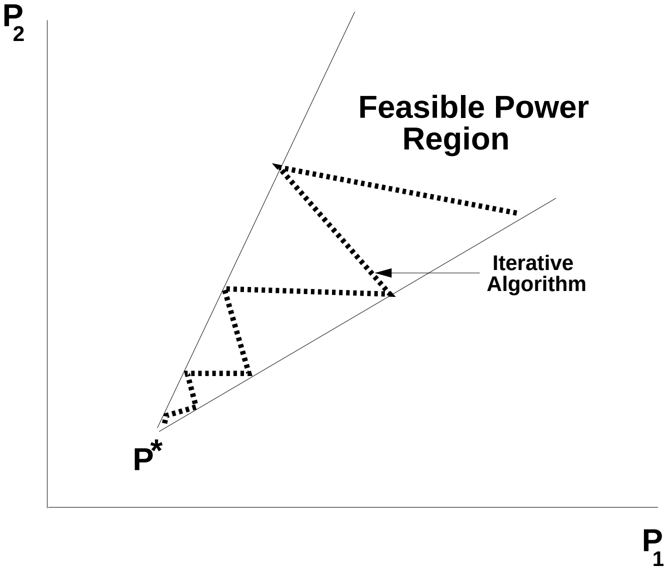


Figure 14.8: Iterative Foschini-Miljanic Algorithm.

The Foschini-Miljanic power control algorithm can also be combined with access control [28]. In this combination, access to the system is based on whether the new user causes other users to fall below their SINR targets. Specifically, when a new user requests access to the system, the base station or access point determines if a set of transmit powers exists such that he can be admitted without degrading existing users below their desired SINR threshold. If the new user cannot be accommodated in the system without violating the SINR requirements of existing users then he is denied access. If he can be accommodated then the power control algorithms of the new and existing users are set to the feasible power vector under which all users (new and existing) meet their SINR targets.

A power control strategy for multiple access that takes into account delay constraints is proposed and analyzed in [29]. This strategy optimizes the transmit power relative to both channel conditions and the delay constraints via dynamic programming. The optimal strategy exhibits three modes: very low transmit power when the channel is poor and the tolerable delay large, higher transmit power when the channel and delay are average, and very high transmit power when the delay constraint is tight. This strategy exhibits significant power savings over constant transmit power while meeting the delay constraints of the traffic.

14.5 Downlink (Broadcast) Channel Capacity

When multiple users share the same channel, the channel capacity can no longer be characterized by a single number. At the extreme, if only one user occupies all signaling dimensions in the channel then the region reduces to the single-user capacity described in Chapter 4. However, since there is an infinite number of ways to divide the channel between many users, the multiuser channel capacity is characterized by a *rate region*, where each point in the region is a vector of achievable rates that can be maintained by all the users simultaneously with arbitrarily small error probability. The union of achievable rate vectors under all multiuser transmission strategies is called the **capacity region** of the multiuser system. The channel capacity is different for uplink channels and downlink channels due to the fundamental differences between these channel models. However, the fact that downlink and uplink channels look like mirror-images of each other implies that there might be a connection between their capacities. In fact, there is a duality between these channels that allows the capacity region of either channel to be obtained from the capacity region of the other. Note that in the analysis of channel capacity the downlink is commonly refered to as the broadcast channel (BC) and the uplink is commonly refered to as the multiple access channel (MAC)[5] and we will use this terminology in our capacity discussions. In this section we describe the capacity region of the BC, Section 14.6 treats the MAC capacity region, and Section 14.7 characterizes the duality between these two channels and how it can be exploited in capacity calculations.

After first describing the AWGN BC model, we will characterize its rate region using superposition code-division (CD) with successive interference cancellation, time-division (TD), and frequency-division (FD). We then obtain the rate regions using DSSS for orthogonal and non-orthogonal codes. The BC and corresponding capacity results under fading is also treated.

We will see that capacity is achieved using superposition CD with interference cancellation. In addition, DSSS with successive interference cancellation has a capacity penalty relative to superposition coding which increases with spreading gain. Finally, spread spectrum with orthogonal CD can achieve a subset of the TD and FD capacity regions, but spread spectrum with non-orthogonal coding and no interference cancellation is inferior to all the other spectrum-sharing techniques. The capacity regions in fading depend on what is known about the fading channel at the transmitter and receiver, analogous to single-user capacity in fading.

14.5.1 Channel Model

We consider a BC consisting of one transmitter sending different data streams, also called independent information or data, to different receivers. Thus, our model is not applicable to a typical radio or TV broadcast channel, where the same data stream, also called common information or data, is received by all users. However the capacity results easily extend to include common data as described in Section 14.5.3. The capacity region of the BC characterizes the rates at which information can be conveyed to the different receivers simultaneously. We mainly focus on capacity regions for the two-user BC, since the general properties and the relative performance of the different spectrum-sharing techniques are the same for any finite number of users [35].

The two-user BC has one transmitter and two distant receivers receiving data at rate R_k, $k = 1, 2$. The channel power gain between the transmitter and kth receiver is $g_k, k = 1, 2$, and each receiver has AWGN of PSD $N_0/2$. We define the effective noise on the kth channel as $n_k = N_0/g_k$, $k = 1, 2$, and we arbitrarily assume that $n_1 \leq n_2$,

[5] MAC is also used as an abbreviation for the medium access control layer in networks[1, Chapter 1.2].

i.e. we assume the first user has a larger channel gain to its receiver than the second user. Incorporating the channel gains into the noise PSD does not change the SINR for any user, since the signal and interference on each user's channel are attenuated by the same channel gain. Thus, the BC capacity with channel gains $\{g_k\}$ is the same as the BC capacity based on the effective noises $\{n_k\}$ [41]. The fact that the channel gains or, equivalently, the effective noise of the users can be ordered makes the channel model a **degraded broadcast channel**, for which a general formula for channel capacity is known [34, Chapter 14.6]. We denote the transmitter's total average power and bandwidth by P and B, respectively.

If the transmitter allocates all the power and bandwidth to one of the users, then clearly the other user will receive a rate of zero. Therefore, the set of simultaneously achievable rates (R_1, R_2) includes the pairs $(C_1, 0)$ and $(0, C_2)$, where

$$C_k = B\log_2\left(1+\frac{P}{n_k B}\right),\ k = 1, 2, \tag{14.11}$$

is the single-user capacity in bps for an AWGN channel, as given in Chapter 4.1. These two points bound the BC capacity region. We now consider rate pairs in the interior of the region, which are achieved using more equitable methods of dividing the system resources.

14.5.2 Capacity in AWGN

In this section we compute the set of achievable rate vectors of the AWGN BC under TD, FD, and the optimal method of superposition coding, which achieves capacity. In TD, the transmit power P and bandwidth B are allocated to user 1 for a fraction τ of the total transmission time, and then to user 2 for the remainder of the transmission. This TD scheme achieves a straight line between the points C_1 and C_2, corresponding to the rate pairs

$$\mathcal{C}_{TD} = \bigcup_{\{\tau:\, 0\le\tau\le 1\}} \left(R_1 = \tau B\log_2\left(1+\frac{P}{n_1B}\right), R_2 = (1-\tau)B\log_2\left(1+\frac{P}{n_2B}\right)\right). \tag{14.12}$$

This equal-power TD achievable rate region is illustrated in Figures 14.10 and 14.11. In these figures, n_1B and n_2B differ by 3 dB and 20 dB, respectively. This dB difference, which reflects the difference in the channel gains of the two users, is a crucial parameter in comparing the achievable rates of the different spectrum-sharing techniques, as we discuss in more detail below.

If we also vary the average transmit power of each user then we can obtain a larger set of achievable rates. Let P_1 and P_2 denote the average power allocated to users 1 and 2, respectively, over their assigned time slots. The average power constraint then becomes $\tau P_1 + (1-\tau)P_2 = P$. The achievable rate region with TD and variable power allocation is then

$$\mathcal{C}_{TD,VP} = \bigcup_{\{\tau, P_1, P_2:\, 0\le\tau\le 1;\, \tau P_1+(1-\tau)P_2=P\}} \left(R_1 = \tau B\log_2\left(1+\frac{P_1}{n_1B}\right), R_2 = (1-\tau)B\log_2\left(1+\frac{P_2}{n_2B}\right)\right). \tag{14.13}$$

In FD the transmitter allocates P_k of its total power P and B_k of its total bandwidth B to user k. The power and bandwidth constraints require that $P_1 + P_2 = P$ and $B_1 + B_2 = B$. The set of achievable rates for a fixed frequency division (B_1, B_2) is thus

$$\mathcal{C}_{FFD} = \bigcup_{\{P_1, P_2:\, P_1+P_2=P\}} \left(R_1 = B_1\log_2\left(1+\frac{P_1}{n_1B_1}\right), R_2 = B_2\log_2\left(1+\frac{P_2}{n_2B_2}\right)\right). \tag{14.14}$$

It was shown by Bergmans [35] that, for n_1 strictly less than n_2 and any fixed frequency division (B_1, B_2), there exists a range of power allocations $\{P_1, P_2 : P_1 + P_2 = P\}$ whose corresponding rate pairs exceed a segment of the equal-power TD line (14.12). This superiority is illustrated in Figures 14.10 and 14.11, where we also plot the rate regions for fixed FD under two different bandwidth divisions. The superiority is difficult to distinguish in Figure 14.10, where the users have similar channel gains, but is much more apparent in Figure 14.11, where the users have a 20 dB difference in gain.

The FD achievable rate region is defined as the union of fixed FD rate regions (14.14) over all bandwidth divisions:

$$\mathcal{C}_{FD} = \bigcup_{\{P_1,P_2,B_1,B_2\colon\, P_1+P_2=P;\; B_1+B_2=B\}} \left(R_1 = B_1 \log_2\left(1+\frac{P_1}{n_1B_1}\right), R_2 = B_2\log_2\left(1+\frac{P_2}{n_2B_2}\right)\right). \tag{14.15}$$

It was shown in [35] that this achievable rate region exceeds the equal-power TD rate region (14.12). This superiority is indicated by the closure of the fixed FD regions in Figures 14.10 and 14.11, although it is difficult to see in Figure 14.10, where the users have a similar received SNR. In fact, when $n_1 = n_2$, (14.15) reduces to (14.12) [35]. Thus, optimal power and/or frequency allocation is more beneficial when the users have very disparate channel quality.

Note that the achievable rate region for TD with unequal power allocation given by (14.13) is the same as the FD achievable rate region (14.15). This is seen by letting $B_i = \tau_i B$ and $\pi_i = \tau_i P_i$ in (14.13), where $\tau_1 = \tau$ and $\tau_2 = 1-\tau$. The power constraint then becomes $\pi_1 + \pi_2 = P$. Making these substitutions in (14.13) yields

$$\mathcal{C}_{TD,VP} = \bigcup_{\{\pi_1,\pi_2\colon\, \pi_1+\pi_2=P\}} \left(R_1 = B_1\log_2\left(1+\frac{\pi_1}{n_1B_1}\right), R_2 = B_2\log_2\left(1+\frac{\pi_2}{n_2B_2}\right)\right). \tag{14.16}$$

Comparing this with (14.14) we see that with appropriate choice of P_k and τ_k, any point in the FD achievable rate region can also be achieved through TD with variable power.

Superposition coding with successive interference cancellation is a multiresolution coding technique whereby the user with the higher channel gain can distinguish the fine resolution of the received signal constellation, while the user with the worse channel can only distinguish the constellation's coarse resolution [35][34, Chapter 14.6]. An example of a two-level superposition code constellation taken from [37] is 32-QAM with embedded 4-PSK, as shown in Figure 14.9. In this example, the transmitted constellation point is one of the 32-QAM signal points chosen as follows. The data stream intended for the user with the worse channel, user 2 in our model since $n_2 > n_1$, provides 2 bits to select one of the 4-PSK superpoints. The data stream intended for the user with the better SNR provides 3 bits to select one of the 8 constellation points surrounding the selected superpoint. After transmission through the channel, the user with the better SNR can easily distinguish the quadrant in which the constellation point lies. Thus, the 4-PSK superpoint is effectively subtracted out by this user. However, the user with the worse channel cannot distinguish between the 32-QAM points around its 4-PSK superpoints. Hence, the 32-QAM modulation superimposed on the 4-PSK modulation appears as noise to this user, and this user can only decode the 4-PSK. These ideas can be easily extended to multiple users using more complex signal constellations. Since superposition coding achieves multiple rates by expanding its signal constellation, it does not require bandwidth expansion.

The two-user capacity region using superposition coding and successive interference cancellation was derived in [35] to be the set of rate pairs

$$\mathcal{C}_{BC} = \bigcup_{\{P_1,P_2\colon\, P_1+P_2=P\}} \left(R_1 = B\log_2\left(1+\frac{P_1}{n_1B}\right), R_2 = B\log_2\left(1+\frac{P_2}{n_2B+P_1}\right)\right). \tag{14.17}$$

The intuitive explanation for (14.17) is the same as for the example illustrated in Figure 14.9: Since $n_1 < n_2$, user 1 correctly receives all the data transmitted to user 2. Therefore, user 1 can decode and subtract out user 2's

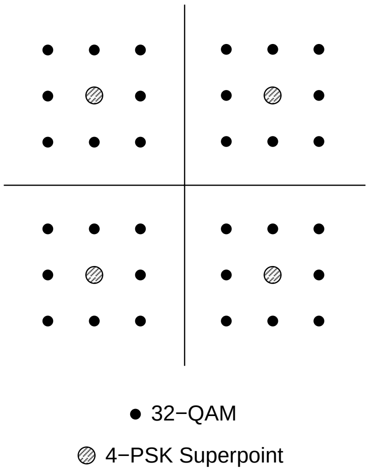


Figure 14.9: 32-QAM with embedded 4-PSK

message, then decode its own message. User 2 cannot decode the message intended for user 1, since it has a worse channel. Thus, user 1's message, with power P_1, contributes an additional noise term to user 2's received message. This message can be treated as an additional AWGN term since the capacity-achieving distributions for the signals associated with each user are Gaussian [34, Chapter 14.1][35]. This same process is used by the successive interference cancellation method for DSSS described in Chapter **??**. However, although successive interference cancellation achieves the capacity region (14.17), it is not necessarily the best method to use in practice. The capacity analysis assumes perfect signal decoding, whereas real systems exhibit some decoding error. This error leads to decision-feedback errors in the successive interference cancellation scheme. Thus, multiuser detection methods that do not suffer from this type of error may work better in practice than successive cancellation.

The rate region defined by (14.17) was shown in [36] to exceed the regions achievable through either TD or FD, when $n_1 < n_2$. Moreover, it was also shown in [36] that this is the maximum achievable set of rate pairs for any type of coding and spectrum sharing, and thus (14.17) defines the BC capacity region, hence the notation $\mathcal{C}_{BC}$. However, if the users all have the same SNR, then this capacity region collapses to the equal-power TD line (14.12). Thus, when $n_1 = n_2$, all the spectrum-sharing methods have the same rate region.

The ideas of superposition coding are easily extended to a K-user system for $K > 2$. Assume a BC with K users, each with channel gain g_k. We first order the users relative to their effective noise $n_k = .5N_0/g_k$. Based on this effective noise ordering, the superposition coding will now have K levels, where the coarsest level can be detected by the user with the largest effective noise, the next level can be detected by the user with the next largest effective noise, and so forth. Each user can remove the effects of the constellation points associated with the noisier channels of other users, but the constellation points transmitted to users with better channels appear as noise. Assuming a total power constraint P, the multiuser extension to the two-user region (14.17) is given by

$$\mathcal{C}_{BC} = \bigcup_{\{P_k: \sum_{k=1}^K P_k = P\}} \left\{ (R_1, \ldots, R_K) : R_k = B \log_2 \left(1 + \frac{P_k}{n_k B + \sum_{j=1}^K P_j \mathbf{1}[n_k > n_j]} \right) \right\}, \tag{14.18}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function.

We define the **sum-rate capacity** of a BC as the maximum sum of rates taken over all rate vectors in the capacity region:

$$C_{BCSR} = \max_{(R_1,\ldots,R_K) \in C_{BC}} \sum_{k=1}^K R_k. \tag{14.19}$$

Sum-rate capacity is a single number that defines the maximum throughput of the system regardless of fairness in terms of rate allocation between the users. It is therefore much easier to characterize than the K-dimensional capacity region, and often leads to important insights. In particular, it can be shown from (14.18) that sum-rate capacity is achieved on the AWGN BC by assigning all power P to the user with the highest channel gain or, equivalently, the lowest effective noise. Defining $n_{min} = \min_k n_k$ and $g_{max} = \max_k g_k$, this implies that the sum-rate capacity C_{BCSR} for the K-user AWGN BC is given by

$$\mathcal{C}_{BCSR} = B\log_2\left(1+\frac{P}{n_{min}B}\right) = B\log_2\left(1+\frac{g_{max}P}{N_0B}\right). \tag{14.20}$$

The sum-rate point is therefore one of the boundary points (14.11) of the capacity region, which is the same for superposition coding, TD, and FD, since all resources are assigned to a single user.

Example 14.5: Consider an AWGN BC with total transmit power $P = 10$ mW, $n_1 = 10^{-9}$ W/Hz, $n_2 = 10^{-8}$ W/Hz, and $B = 100$ KHz. Suppose user 1 requires a data rate of 300 Kbps. Find the rate that can be allocated to user 2 under fixed power TD, equal-bandwidth FD, and superposition coding.

Solution: In equal-power time division user 1 has a rate of $R_1 = \tau B\log_2\left(1+\frac{P}{n_1B}\right) = 6.644\times10^5\tau$ bps. Setting R_1 to the desired value $R_1 = 6.644\times10^5\tau = 3\times10^5$ bps and solving for τ yields $\tau = 3\times10^5/6.644\times10^5 = .452$. Then user 2 gets a rate of $R_2 = (1-\tau)B\log_2\left(1+\frac{P}{n_2B}\right) = 1.89\times10^5$ bps. In equal-bandwidth FD we require $R_1 = .5B\log_2\left(1+\frac{P_1}{.5n_1B}\right) = 3\times10^5$ bps. Solving for $P_1 = .5n_1B(2^{R_1/(.5B)}-1)$ yields $P_1 = 3.15$ mW. Setting $P_2 = P - P_1 = 6.85$ mW, we get $R_2 = .5B\log_2\left(1+\frac{P_2}{.5n_2B}\right) = 1.94\times10^5$ bps. Finally, with superposition coding we have $R_1 = B\log_2\left(1+\frac{P_1}{n_1B}\right) = 3\times10^5$. Solving for $P_1 = n_1B(2^{R_1/B}-1)$ yields $P_1 = .7$ mW. Then

$$R_2 = B\log\left(1+\frac{P-P_1}{n_2B+P_1}\right) = 2.69\times10^5 \text{ bps}.$$

Clearly superposition coding is superior to both TD and FD, as expected, although the performance of these techniques would be closer to that of superposition coding if we optimized the power allocation for TD or the bandwidth allocation for FD.

Example 14.6: Find the sum-rate capacity for the system in the prior example.

Solution: We have $P = 10$ mW, $n_1 = 10^{-9}$ W/Hz, $n_2 = 10^{-8}$ W/Hz, and $B = 100$ KHz. The minimum noise is associated with user 1, $n_{min} = 10^{-9}$. Thus, $\mathcal{C}_{BCSR} = B\log_2\left(1+\frac{P}{n_{min}B}\right) = 6.644\times10^5$ bps, and this sum-rate is achievable with TD, FD, or superposition coding, which are all equivalent for this sum-rate capacity since all resources are allocated to the first user.

CD for multiple users can also be implemented using DSSS, as discussed in Chapter **??**. In such systems the modulated data signal for each user is modulated by a unique spreading code, which increases the transmit signal bandwidth by approximately G, the processing gain of the spreading code. For orthogonal spreading codes, the cross correlation between the respective codes is zero, and these codes require a spreading gain of N to produce N orthogonal codes. For a total bandwidth constraint B, the information bandwidth of each user's signal with these spreading codes is thus limited to B/N. The two-user achievable rate region with these spreading codes is then

$$\mathcal{C}_{DS,OC} = \bigcup_{\{P_1,P_2:\, P_1+P_2=P\}} \left(R_1 = \frac{B}{2}\log_2\left(1+\frac{P_1}{n_1B/2}\right), R_2 = \frac{B}{2}\log_2\left(1+\frac{P_2}{n_2B/2}\right)\right). \tag{14.21}$$

Comparing (14.21) with (14.14) we see that CD with orthogonal coding is the same as fixed FD with the bandwidth equally divided ($B_1 = B_2 = B/2$). From (14.16), TD with unequal power allocation can also achieve all points in this rate region. Thus, orthogonal CD with Walsh-Hadamard codes achieves a subset of the TD and FD achievable rate regions. More general orthogonal codes are needed to achieve the same region as these other techniques.

We now consider DSSS with non-orthogonal spreading codes. As discussed in Chapter **??**, in these systems interference between users is attenuated by the code cross correlation. Thus, if interference is treated as noise, its power contribution to the SIR is reduced by the square of the code cross correlation. From (**??**), we will assume that spreading codes with a processing gain of G reduce the interference power by $1/G$. This is a reasonable approximation for random spreading codes, although as discussed in Chapter 14 the exact value of the interference power reduction depends on the nature of the spreading codes and other assumptions [38, 39]. Since the signal bandwidth is increased by G, the two-user BC rate region achievable through non-orthogonal DSSS and successive interference cancellation is given by

$$\mathcal{C}_{DS,SC,IC} \bigcup_{\{P_1,P_2:\, P_1+P_2=P\}} \left(R_1 = \frac{B}{G}\log_2\left(1+\frac{P_1}{n_1B/G}\right), R_2 = \frac{B}{G}\log_2\left(1+\frac{P_2}{n_2B/G+P_1/G}\right)\right). \tag{14.22}$$

By the convexity of the log function, the rate region defined by (14.22) for $G > 1$ is smaller than the rate region (14.17) obtained using superposition coding, and the degradation increases with increasing values of G. This implies that for nonorthogonal coding, the spreading gain should be minimized in order to maximize capacity.

With non-orthogonal coding and no interference cancellation, the receiver treats all signals intended for other users as noise, resulting in the achievable rate region

$$\mathcal{C}_{DS,SC} = \bigcup_{\{P_1,P_2:\, P_1+P_2=P\}} \left(R_1 = \frac{B}{G}\log_2\left(1+\frac{P_1}{n_1B/G+P_2/G}\right), R_2 = \frac{B}{G}\log_2\left(1+\frac{P_2}{n_2B/G+P_1/G}\right)\right) \tag{14.23}$$

Again using the log function convexity, $G = 1$ maximizes this rate region, and the rate region decreases as G increases. Moreover, the radius of curvature for (14.23) is given by

$$\chi = \frac{\dot{R}_1\ddot{R}_2 - \dot{R}_2\ddot{R}_1}{(\dot{R}_1^2+\dot{R}_2^2)^{3/2}}, \tag{14.24}$$

where $\dot{R}_i$ and $\ddot{R}_i$ denote, respectively, the first and second derivatives of R_i with respect to α for $P_1 = \alpha P$ and $P_2 = (1-\alpha)P$. For $G = 1$, $\chi \geq 0$. Thus, the rate region for nonorthogonal coding without interference cancellation (14.23) is bounded by a convex function with end points C_1 and C_2, as shown in Figures 14.10 and 14.11. Therefore, the achievable rate region for nonorthogonal CD without interference cancellation will lie beneath the regions for TD and FD, which are bounded by concave functions with the same endpoints.

The acheivable rate regions for equal-power TD (14.12), FD (14.14), orthogonal CD (14.21), and nonorthogonal CD with (14.17) and without (14.23) interference cancellation are illustrated in Figures 14.10 and 14.11, where the SNR between the users differs by 3 dB and 20 dB, respectively. For the calculation of (14.23) we assume CD through superposition coding with $G = 1$. Spread spectrum CD with larger values of the spreading gain will result in a smaller rate region.

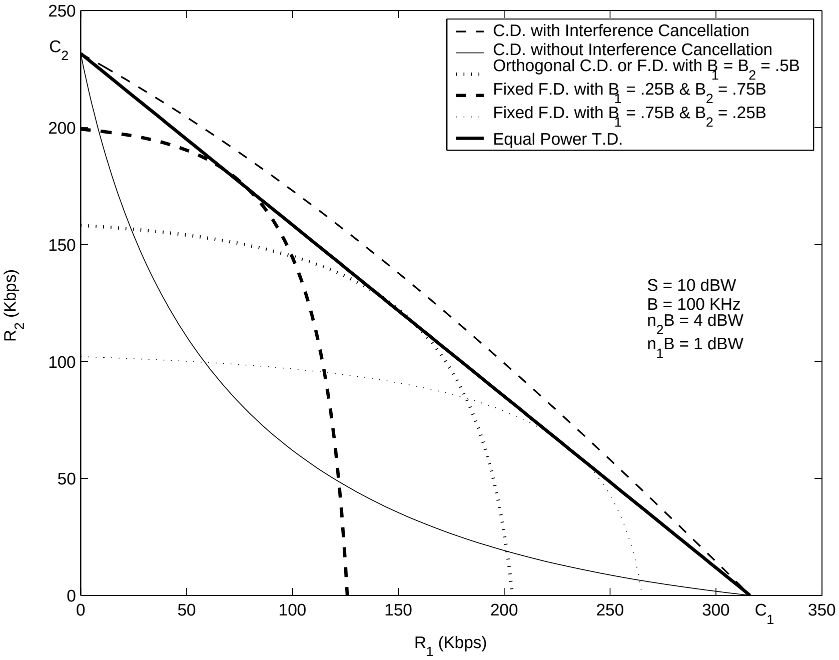


Figure 14.10: Two-User Capacity Region: 3 dB SNR Difference.

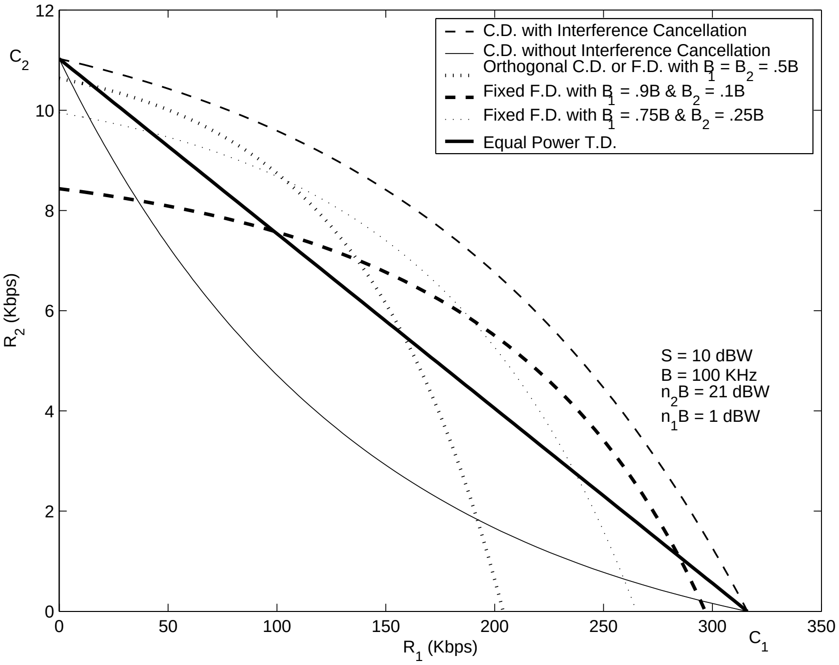


Figure 14.11: Two-User Capacity Region: 20 dB SNR Difference.

14.5.3 Common Data

In many broadcasting applications common data is sent to all users in the system. For example, television and radio stations broadcast the same data to all users, and in wireless Internet applications many users may want to download the same stock quotes and sports scores. The nature of superposition coding makes it straightforward to develop optimal broadcasting techniques for common data and to incorporate common data into the capacity region for the BC. In particular, for a two-user BC with superposition coding, the user with the better channel always receives the data intended for the user with the worse channel, along with his own data. Thus, since common data must be transmitted to both users, we can encode all common data as independent data intended for the user with the worse channel. Since the user with the better channel will also receive this data, it will be received by both users.

Under this transmission strategy, if the rate pair (R_1, R_2) is in the capacity region of the two-user BC with independent data defined by (14.17), for any $R_0 \leq R_2$ we can achieve the rate triple $(R_0, R_1, R_2 - R_0)$ for the BC with common and independent data, where R_0 is the rate of common data, R_1 is the rate of user 1's independent data, and $R_2 - R_0$ is the rate of user 2's independent data. Mathematically, this gives the three-dimensional capacity region

$$\mathcal{C}_{BC} = \bigcup_{\{P_1,P_2:P_1+P_2=P\}} \left(R_0 \leq B\log_2\left(1+\frac{P_2}{n_2B+P_1}\right), R_1 = B\log_2\left(1+\frac{P_1}{n_1B}\right), R_2 = B\log_2\left(1+\frac{P_2}{n_2B+P_1}\right) - R_0\right). \tag{14.25}$$

Example 14.7: In Example 14.5 we saw that for a broadcast channel with total transmit power $P = 10$ mW, $n_1 = 10^{-9}$ W/Hz, $n_2 = 10^{-8}$ W/Hz, and $B = 100$ KHz, the rate pair $(R_1, R_2) = (3 \times 10^5, 2.69 \times 10^5)$ is on the boundary of the capacity region. Suppose user 1 desires an independent data rate of 300 Kbps, and a common data rate of 100 Kbps is required for the system. At what rate can user 2 get independent data?

Solution: In order for $R_1 = 300$ Kbps, we require the same $P_1 = .7$ mW as in Example 14.5.2. The common information rate $R_0 = 10^5 < 2.69 \times 10^5$, so from (14.25), the independent information rate to user 2 is just $R_2 - R_0 = 2.69 \times 10^5 - 10^5 = 1.69 \times 10^5$ bps.

14.5.4 Capacity in Fading

We now consider the capacity region of BCs with fading, where the users have independent random channel gains that change over time. As described in Chapter 4.2 for single-user channels, the capacity of fading BCs depends on what is known about the channel at the transmitter and receiver. However, capacity of a BC is only known for degraded BCs, and this model requires that the channel gains are known to both the transmitter and receiver. Moreover, superposition coding cannot be used without transmitter knowledge of the channel gains, since if the transmitter does not know the relative channel gains it does not know which user can receive the coarse constellation point and which can receive the fine one. Thus we will only consider fading BCs where there is perfect channel side information (CSI) about the instantaneous channel gains at both the transmitter and at all receivers. We also assume that the channel is slowly fading so that for a given fading state, the coding strategy that achieves any point in the capacity region for the static BC with this state has sufficient time to drive the error probability close to zero before the channel gains change.[6]

As with the single-user fading channel, there are two notions of capacity for multiuser fading channels with perfect CSI: ergodic (Shannon) capacity and outage capacity. The ergodic capacity region of a BC characterizes the achievable rate vectors averaged over all fading states [40, 41] while the outage capacity region dictates the set of fixed rate vectors that can be maintained in all fading states subject to a given outage probability[42, 43, 44]. Zero-outage capacity refers to outage capacity with zero outage probability [42], i.e. the set of fixed rate vectors that can be maintained in all fading states. The ergodic capacity region, analogous to ergodic capacity for single-user systems, defines the data rate vectors that can be maintained over time without any constraints on delay. Hence, in some fading states, the data rate may be small or zero, which can be problematic for delay-constrained applications like voice or video. The outage capacity region, analogous to outage capacity in single-user systems, forces a fixed rate vector in all nonoutage fading states, which is perhaps a more appropriate capacity metric for delay-constrained applications. However, the requirement to maintain a fixed rate even in very deep fades can severely decrease the outage capacity region relative to the ergodic capacity region. In fact, the zero-outage capacity region when all users exhibit Rayleigh fading is zero for all users.

We consider a BC with AWGN and fading where a single transmitter communicates independent information to K users over bandwidth B with average transmit power $\overline{P}$. The transmitter and all receivers have a single antenna. The time-varying power gain of user k's channel at time i is $g_k[i]$. Each receiver has AWGN with PSD $N_0/2$. We define the effective time-varying noise of the kth user as [7] $n_k[i] = N_0/g_k[i]$. The **effective noise vector** at time i is defined as

$$\mathbf{n}[i] = (n_1[i], \ldots, n_K[i]). \tag{14.26}$$

We also call this the **fading state** at time i, since it characterizes the channel gains $g_k[i]$ associated with each user at time i. We will denote the kth element of this vector as $n_k[i]$ or just n_k when the time reference is clear. As with the static channel, the capacity of the fading BC can be computed based on its time-varying channel gains or its time-varying effective noise vector. The ergodic BC capacity region is defined as the set of all average rates achievable in a fading channel with arbitrarily small probability of error, where the average is taken with respect to all fading states. In [41], the ergodic capacity region and optimal power allocation scheme for the fading BC is found by decomposing the fading channel into a parallel set of static BCs, one for every possible fading state $\mathbf{n} = (N_0/g_1, \ldots, N_0/g_K)$. In each fading state, the channel can be viewed as a static AWGN BC, and time, frequency, or code division techniques can be applied to the channel in each fading state.

Since the transmitter and all receivers know $\mathbf{n}[i]$, superposition coding according to the ordering of the current effective noise vector can be used by the transmitter. Each receiver can perform successive decoding in which the users with larger effective noise are decoded and subtracted off before decoding the desired signal. Furthermore, the power transmitted to each user $P_j(\mathbf{n})$ is a function of the current fading state. Since the transmission scheme is based on superposition coding, it only remains to determine the optimal power allocation across users and over time.

We define a power policy $\mathcal{P}$ over all possible fading states as a function that maps from any fading state $\mathbf{n}$ to the transmitted power $P_k(\mathbf{n})$ for each user. Let $\mathcal{F}_{BC}$ denote the set of all power policies satisfying average power constraint $\overline{P}$:

$$\mathcal{F}_{BC} \equiv \left\{ \mathcal{P} : \ \mathbf{E}_{\mathbf{n}} \left[\sum_{k=1}^{K} P_k(\mathbf{n}) \right] \leq \overline{P} \right\}. \tag{14.27}$$

From (14.18), the capacity region assuming a constant fading state $\mathbf{n}$ with power allocation $\mathbf{P}(\mathbf{n}) = \{P_k(\mathbf{n}) : k =$

[6] More precisely, the coding strategy that achieves a point in the AWGN BC capacity region uses a block code, and the error probability of the code goes to zero with blocklength. Our slow fading assumption presumes that the channel gains stay constant long enough for the block code associated with these gains to drive the error probability close to zero.

[7] Notice that the noise vector is the instantaneous power of the noise and not the instantaneous noise sample.

$1,\ldots,K\}$ is given by

$$\mathcal{C}_{BC}(\mathbf{P}(\mathbf{n})) = \left\{ (R_1(\mathbf{P}(\mathbf{n}),\ldots,R_K(\mathbf{P}(\mathbf{n})) : R_k(\mathbf{P}(\mathbf{n})) = B\log_2\left(1+\frac{P_k(\mathbf{n})}{n_kB+\sum_{j=1}^K P_j(\mathbf{n})\mathbf{1}[n_k>n_j]}\right)\right\} \tag{14.28}$$

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

$$\mathcal{C}_{BC}(\mathcal{P}) = \{R_k : R_k \leq \mathbf{E}_{\mathbf{n}}\left[R_k(\mathbf{P}(\mathbf{n}))\right], \quad k=1,2,\ldots,K\}$$

where $R_k(\mathbf{P}(\mathbf{n}))$ is as given in (14.28). From [41], the ergodic capacity region of the BC with perfect CSI and power constraint $\overline{P}$ is:

$$\mathcal{C}_{BC}(\overline{P}) = \bigcup_{\mathcal{P}\in\mathcal{F}_{BC}} \mathcal{C}_{BC}(\mathcal{P}). \tag{14.29}$$

It is further shown in [41] that the region $\mathcal{C}_{BC}(\overline{P})$ is convex, and that the optimal power allocation scheme is an extension of water-filling with K different water-levels for a K-user system.

We can also define achievable rate vectors for TD or FD, although these will clearly lie inside the ergodic capacity region, since superposition coding outperforms both of these techniques in every fading state. The optimal form of TD adapts the power assigned to each user relative to the current fading state. Similarly, the optimal form of FD adapts the bandwidth and power assigned to each user relative to the current fading state. As described in Section 14.5.2, for each fading state varying the power in TD yields the same rates as varying the power and bandwidth in FD. Thus, the achievable rates for these two techniques averaged over all fading states are the same. Focusing on the FD region, assume a power policy $\mathcal{P}\in\mathcal{F}_{BC}$ that assigns power $P_k(\mathbf{n})$ to the kth user in fading state $\mathbf{n}$. From (14.27), a power policy $\mathcal{P}\in\mathcal{F}_{BC}$ satisfies the average power constraint. Also assume a bandwidth policy $\mathcal{B}$ that assigns bandwidth $B_k(\mathbf{n})$ to user k in state n and let $\mathcal{G}$ denote the set of all bandwidth policies satisfying the bandwidth constraint of the system:

$$\mathcal{G} \equiv \left\{\mathcal{B} : \sum_{k=1}^K B_k(\mathbf{n}) = B \;\; \forall \mathbf{n}\right\}.$$

The set of achievable rates for FD under these policies is

$$\mathcal{C}_{FD}(\mathcal{P},\mathcal{B}) = \{R_k : R_k \leq \mathbf{E}_{\mathbf{n}}\left[R_k(\mathbf{P}(\mathbf{n}),\mathcal{B})\right], \quad k=1,2,\ldots,K\}, \tag{14.30}$$

where

$$R_k(P(\mathbf{n}),\mathcal{B}) = B_k(\mathbf{n})\log_2\left(1+\frac{P_k(\mathbf{n})}{n_kB_k(\mathbf{n})}\right) \tag{14.31}$$

The set of all achievable rates under frequency division with perfect CSI subject to power constraint $\overline{P}$ and bandwidth constraint B is then

$$\mathcal{C}_{FD}(\overline{P},B) = \bigcup_{\mathcal{P}\in\mathcal{F}_{BC},\mathcal{B}\in\mathcal{G}} \mathcal{C}_{FD}(\mathcal{P},\mathcal{B}). \tag{14.32}$$

The sum-rate capacity for fading BCs is defined as the maximum sum of achievable rates, maximized over all rate vectors in the ergodic BC capacity region. Since sum-rate for the AWGN BC is maximized by transmitting only to the user with the best channel, in fading sum-rate is maximized by transmitting only to the user with the best channel in each channel state. Clearly superposition CD, TD, and FD are all equivalent in this setting, since all resources are assigned to a single user in each state. We can compute the sum-rate capacity and the optimal power allocation over time from an equivalent single-user fading channel with time-varying effective noise

$n[i] = \min_k n_k[i]$ and average power constraint $\overline{P}$. From Chapter 4.2.4 the optimal power allocation to the user with the best channel at time i is thus a water-filling in time, with cutoff value determined from the distribution of $\min_k n_k[i]$.

The ergodic capacity and achievable rate regions for fading broadcast channels under CD, TD, and FD are computed in [41] for different fading distributions, along with the optimal adaptive resource allocation strategies that achieve the boundaries of these regions. These adaptive transmission policies exploit **multiuser diversity** in that more resources (power, bandwidth, timeslots) are allocated to the users with the best channels in any given fading state. In particular, sum-rate capacity is achieved by allocating all resources in any given state to the user with the best channel. Multiuser diversity will be discussed in more detail in Section 14.8.

The zero-outage BC capacity region defines the set of rates that can be simultaneously achieved for all users in *all* fading states while meeting the average power constraint. It is the multiuser extension of zero-outage capacity defined in Chapter 4.2.4 for single-user channels. From [43], the power required to support a rate vector $\mathbf{R} = (R_1, R_2, \ldots, R_K)$ in fading state $\mathbf{n}$ is:

$$P^{min}(\mathbf{R}, \mathbf{n}) = \sum_{k=1}^{K-1} \left[2^{\sum_{j=k+1}^{K} R_{\pi(j)}/B} \left(2^{R_{\pi(k)}/B} - 1 \right) n_{\pi(k)} B \right] + \left(2^{R_{\pi(K)}/B} - 1 \right) n_{\pi(K)} B, \tag{14.33}$$

where $\pi(.)$ is the permutation such that

$$n_{\pi(1)} < n_{\pi(2)} < \cdots < n_{\pi(K)}.$$

Therefore the zero-outage capacity region is the union of all rate vectors that meet the average power constraint:

$$\mathcal{C}_{BC}^{0}(\overline{P}) = \bigcup_{\{\mathbf{R}:\mathbf{E_n}[P^{min}(\mathbf{R},\mathbf{n})] \le \overline{P}\}} \mathbf{R} = (R_1, R_2, \ldots, R_K). \tag{14.34}$$

The boundary of the zero-outage capacity region is the set of all rate vectors $\mathbf{R}$ such that the power constraint is met with equality. For the two-user BC with time-varying AWGN with powers n_1 and n_2, this boundary simplifies to the set of all (R_1, R_2) that satisfy the following equation [43]:

$$\begin{aligned} \overline{P} = p(n_1 < n_2) & \left[\mathbf{E}[n_1 | n_1 < n_2] 2^{R_2/B} (2^{R_1/B} - 1) + \mathbf{E}[n_2 | n_1 < n_2] (2^{R_2/B} - 1) \right] \quad + \\ p(n_1 \ge n_2) & \left[\mathbf{E}[n_2 | n_1 \ge n_2] 2^{R_1/B} (2^{R_2/B} - 1) + \mathbf{E}[n_1 | n_1 \ge n_2] (2^{R_1/B} - 1) \right] \end{aligned}$$

The boundary is determined solely by $\mathbf{E}[n_1|n_1 < n_2]$, $\mathbf{E}[n_2|n_1 < n_2]$, $\mathbf{E}[n_1|n_1 \ge n_2]$, and $\mathbf{E}[n_2|n_1 \ge n_2]$. This is due to the fact that the power required to achieve a rate vector is a linear function of the noise levels in each state, as seen in (14.33). The zero-outage capacity region depends on the conditional expectations of the noises as opposed to their unconditional expectations since every different ordering of noises leads to a different expression for the required power in each state, as can be seen from (14.33).

The outage capacity region of the BC is defined similarly as the zero-outage capacity region, except that users may have some nonzero probability of outage so that they can suspend transmission in some outage states. This provides additional flexibility in the system since under severe fading conditions, maintaining a fixed rate in *all* fading states can consume a great deal of power. In particular, we saw in Chapter 4.2 that for a single-user fading channel, maintaining any non-zero fixed rate in Rayleigh fading requires infinite power. By allowing some outage, power can be conserved from outage states to maintain higher rates in non-outage states. The outage capacity region is more difficult to obtain than the zero-outage capacity region, since in any given fading state the transmission strategy must determine which users to put into outage. Once the outage users are determined, the power required to maintain the remaining users is given by (14.33) for the rate vector associated with the $K' \leq K$ users that are not in outage. It is shown in [43] that this decision should be made based on a threshold policy, and the resulting outage capacity region is then obtained implicitly based on the threshold policy and the power allocation (14.33) for non-outage users.

The notions of ergodic capacity and outage capacity can also be combined. This combination results in the minimum rate capacity region [46]. A rate vector in this region characterizes the set of all average rate vectors that can be maintained, averaged over all fading states, subject to some minimum rate vector that must be maintained in all states (possible subject to some outage probability). Minimum rate capacity is useful for systems supporting a mix of delay-constrained and delay-unconstrained data. The minimum rates dictate the data rates available for the constrained data that must be maintained in all fading states, while the rates above these minimums are what is available for the unconstrained data, where these additional rates vary depending on the current fading state. The minimum rate capacity region (with zero outage probability) lies between that of the zero-outage capacity region and the ergodic capacity region: for minimum rates of zero it equals the ergodic capacity region, and for minimum rates on the boundary of the zero-outage capacity region, it cannot exceed these boundary points. This is illustrated in Figure 14.12, where we plot the ergodic, zero-outage, and minimum rate capacity region for a BC with Rician fading. We see from this figure that the ergodic capacity region is the largest, since it can adapt to the different channel states to maximize its average rate, averaged over all fading states. The zero-outage capacity region is the smallest, since it is forced to maintain a fixed rate in all states, which consumes much power when the fading is severe. The minimum rate capacity region lies between the other two, and depends on the minimum rate requirements. As the minimum rate vector that must be maintained in all fading states increases, the minimum rate capacity region approaches the zero-outage capacity region, and as this minimum rate vector decreases, the minimum rate capacity region approaches the ergodic capacity region.

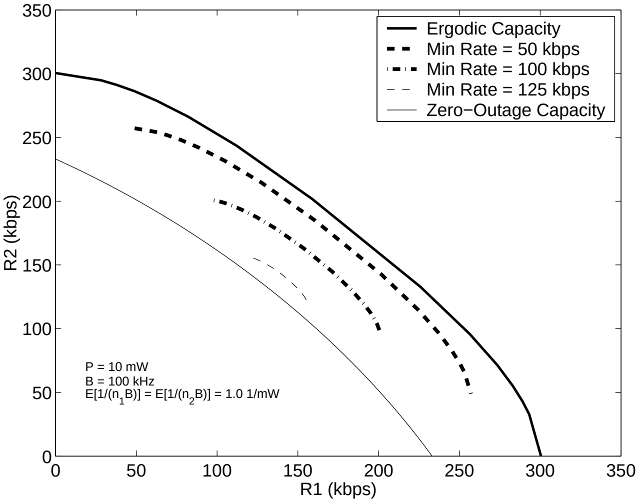


Figure 14.12: Ergodic, Zero-Outage, and Minimum Rate BC Capacity Regions (Rician fading with a K factor of 1, Average SNR = 10 dB)

14.5.5 Capacity with Multiple Antennas

We now investigate the capacity region for a BC with multiple antennas. We have seen in Chapter 10.3 that MIMO systems can provide large capacity increases for single-user systems. The same will be true of multiuser systems: in fact multiple users can exploit multiple spatial dimensions even more effectively than a single user.

Consider a K-user BC where the transmitter has M_t antennas and each receiver has M_r antennas. The $M_r \times M_t$ channel matrix $\mathbf{H}_k$ characterizes the channel gains between each antenna at the transmitter and each antenna at the kth receiver. The received signal for the kth user is then

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{n}_k, \tag{14.35}$$

where $\mathbf{x}$ is the input to the transmit antennas, and we denote its covariance matrix as $\mathbf{\Sigma}_x$. For simplicity, we normalize the bandwidth to unity[8], $B = 1$ Hz, and assume the noise vector $\mathbf{n}_k$ is a circularly symmetric complex Gaussian with $\mathbf{n}_k \sim N(0, \mathbf{I})$.

When the transmitter has more than one antenna, $M_t > 1$, the BC is no longer degraded. In other words, receivers cannot generally be ranked by their channel quality since receivers have different channel gains associated with the different antennas at the transmitter. The capacity region of the general non-degraded broadcast channels is unknown. However, an achievable region for this channel was proposed in [54, 55] which was later shown to equal the capacity region [58]. The region is based on the notion of **dirty paper coding** (DPC) [59]. The basic premise of DPC is as follows. If the transmitter (but not the receiver) has perfect, non-causal knowledge of interference to a given user, then the capacity of the channel is the same as if there was no interference or, equivalently, as if the receiver had knowledge of the interference and could subtract it out. DPC is a technique that allows non-causally known interference to be "pre-subtracted" at the transmitter but in such a way that the transmit power is not increased. A more practical (and more general) technique to perform this pre-subtraction is described in [60].

In the MIMO BC, DPC can be applied at the transmitter when choosing codewords for different users. The transmitter first picks a codeword for User 1. The transmitter then chooses a codeword for User 2 with full (non-causal) knowledge of the codeword intended for User 1. Therefore the codeword of User 1 can be pre-subtracted such that User 2 does not see the codeword intended for User 1 as interference. Similarly, the codeword for User 3 is chosen such that User 3 does not see the signals intended for Users 1 and 2 as interference. This process continues for all K users. The ordering of the users clearly matters in such a procedure, and needs to be optimized in the capacity calculation. Let $\pi(\cdot)$ denote a permutation of the user indices and $\mathbf{\Sigma} = [\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K]$ denote a set of positive semi-definite covariance matrices with Tr$(\mathbf{\Sigma}_1 + \ldots \mathbf{\Sigma}_K) \leq P$. Under DPC, if User $\pi(1)$ is encoded first, followed by User $\pi(2)$, etc., then the following rate vector is achievable:

$$\mathbf{R}(\pi, \mathbf{\Sigma}) : R_{\pi(k)} = \log \frac{|\mathbf{I} + \mathbf{H}_{\pi(k)} (\sum_{j \geq k} \mathbf{\Sigma}_{\pi(j)}) \mathbf{H}_{\pi(k)}^H|}{|\mathbf{I} + \mathbf{H}_{\pi(k)} (\sum_{j > k} \mathbf{\Sigma}_{\pi(j)}) \mathbf{H}_{\pi(k)}^H|}, \quad k = 1, \ldots, K. \tag{14.36}$$

The capacity region $\mathcal{C}$ is then the convex hull of the union of all such rates vectors over all permutations and all positive semi-definite covariance matrices satisfying the average power constraint:

$$\mathcal{C}_{\mathrm{BC}}(P, \mathbf{H}) \triangleq Co \left(\bigcup_{\pi, \mathbf{\Sigma}} \mathbf{R}(\pi, \mathbf{\Sigma}) \right) \tag{14.37}$$

where $\mathbf{R}(\pi, \mathbf{\Sigma})$ is given by (14.36). The transmitted signal is $\mathbf{x} = \mathbf{x_1} + \ldots + \mathbf{x_K}$ and the input covariance matrices are of the form $\mathbf{\Sigma}_k = \mathbb{E}[\mathbf{x_k} \mathbf{x_k}^*]$. The DPC implies that $\mathbf{x}_1, \ldots, \mathbf{x}_K$ are uncorrelated, and thus $\mathbf{\Sigma}_x = \mathbf{\Sigma}_1 + \ldots + \mathbf{\Sigma}_K \leq P$.

One important feature to notice about the rate equations defined by (14.36) is that these equations are neither a concave nor convex function of the covariance matrices. This makes finding the capacity region very difficult, because generally the entire space of covariance matrices that meet the power constraint must be searched over

[8] Capacity of unity bandwidth MIMO channels has a factor of .5 preceeding the log function for real (one-dimensional) channels with no such factor for complex (two-dimensional) channels [47, Chapter 3.1].

[54, 55]. However, as described in Section 14.7, there is a duality between the MIMO BC and the MIMO MAC that can be exploited to greatly simplify this calculation. The capacity region for a 2-user channel with $M = 2$ and $N = 1$ computed by exploiting this duality is shown in Fig. 14.13. The region is defined by the outer boundary, and the lines inside this boundary each correspond to the capacity region of a different dual MIMO MAC channel whose sum power equals the power of the MIMO BC. The union of these dual regions yields the boundary of the MIMO BC region, as will be discussed in Section 14.7.

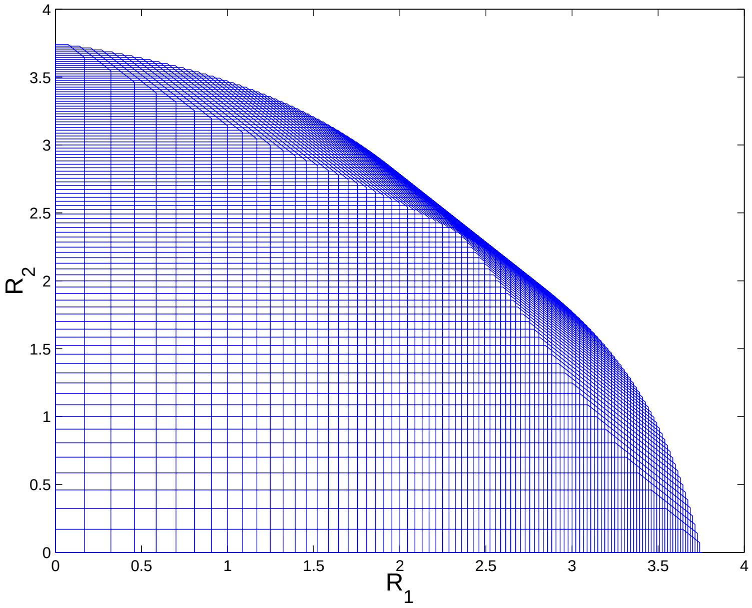


Figure 14.13: MIMO BC capacity region, $\mathbf{H}_1 = [1 \;\; 0.5]$, $\mathbf{H}_2 = [0.5 \;\; 1]$, $P = 10$

14.6 Uplink (Multiple Access) Channel Capacity

14.6.1 Capacity in AWGN

The MAC consists of K transmitters, each with power P_k, sending to a receiver over a channel with power gain g_k. We assume all transmitters and the receiver have a single antenna. The received signal is corrupted by AWGN with PSD $N_0/2$. The two-user multiaccess capacity region is the closed convex hull of all vectors (R_1, R_2) satisfying the following constraints [34]:

$$R_k \leq B \log_2 \left(1 + \frac{g_k P_k}{N_0 B}\right), k = 1, 2 \tag{14.38}$$

and

$$R_1 + R_2 \leq B \log_2 \left(1 + \frac{g_1 P_1 + g_2 P_2}{N_0 B}\right). \tag{14.39}$$

The first constraint (14.38) is just the capacity associated with each individual channel. The second constraint (14.39) indicates that the sum of rates for all users cannot exceed the capacity of a “superuser” with received power equal to the sum of received powers from all users. For K users, the region becomes

$$\mathcal{C}_{MAC} = \left\{(R_1, \ldots, R_K) : \sum_{k \in S} R_k \leq B \log_2 \left(1 + \frac{\sum_{k \in S} g_k P_k}{N_0 B}\right), \forall S \subset \{1, 2, \ldots, K\}\right\}. \tag{14.40}$$

Thus, the region (14.40) indicates that the sum of rates for any subset of the K users cannot exceed the capacity of a superuser with received power equal to the sum of received powers associated with this user subset.

The sum-rate capacity of a MAC is the maximum sum of rates $\sum_{k=1}^{K} R_k$ where the maximum is taken over all rate vectors $(R_1, \ldots, R_K)$ in the MAC capacity region. As with the sum-rate capacity of the BC, the MAC sum-rate also measures the maximum throughput of the system regardless of fairness, and is easier to characterize than the K-dimensional capacity region. It can be shown from (14.40) that sum-rate capacity is achieved on the AWGN MAC by having all users transmit at their maximum power, which yields:

$$\mathcal{C}_{MACSR} = B \log_2 \left(1 + \frac{\sum_{k=1}^{K} g_k P_k}{N_0 B} \right). \tag{14.41}$$

The intuition behind this result is that each user in the MAC has an individual power constraint, so not allowing a user to transmit at full power wastes system power. By contrast, the AWGN BC sum-rate capacity (14.20) is achieved by only transmitting to the user with the best channel. However, since all users share the power resource, no power is wasted in this case.

The MAC capacity region for two users is shown in Figure 14.14, where C_k and C_k^* are given by

$$C_k = B \log_2 \left(1 + \frac{g_k P_k}{N_0 B} \right), \; k = 1, 2, \tag{14.42}$$

$$C_1^* = B \log_2 \left(1 + \frac{g_1 P_1}{N_0 B + g_2 P_2} \right), \tag{14.43}$$

and

$$C_2^* = B \log_2 \left(1 + \frac{g_2 P_2}{N_0 B + g_1 P_1} \right). \tag{14.44}$$

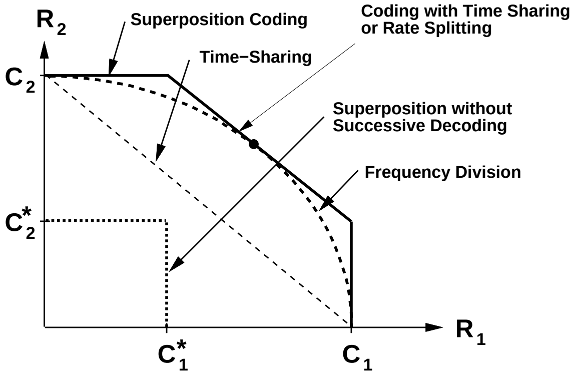


Figure 14.14: Two-User MAC Capacity Region.

The point $(C_1, 0)$ is the achievable rate vector when transmitter 1 is sending at its maximum rate and transmitter 2 is silent, and the opposite scenario achieves the rate vector $(0, C_2)$. The corner points (C_1, C_2^*) and (C_1^*, C_2) are achieved using the successive interference cancellation described above for superposition codes. Specifically, let the first user operate at the maximum data rate C_1. Then its signal will appear as noise to user 2; thus, user 2 can send data at rate C_2^* which can be decoded at the receiver with arbitrarily small error probability. If the receiver then subtracts out user 2's message from its received signal, the remaining message component is just users 1's message corrupted by noise, so rate C_1 can be achieved with arbitrarily small error probability. Hence, (C_1, C_2^*)

is an achievable rate vector. A similar argument with the user roles reversed yields the rate point (C_1^*, C_2). Time-sharing between these two strategies yields any point on the straight line connecting (C_1, C_2^*) and (C_1^*, C_2). Note that in the broadcast channel the better user must always be decoded last, whereas in the MAC decoding can be done in either order. This is a fundamental difference of the two channels.

TD between the two transmitters operating at their maximum rates, given by (14.42), yields any rate vector on the straight line connecting C_1 and C_2. With FD, the rates depend on the fraction of the total bandwidth that is allocated to each transmitter. Letting B_1 and B_2 denote the bandwidth allocated to each of the two users, we get the achievable rate region

$$\mathcal{C}_{FD} = \bigcup_{\{B_1,B_2:B_1+B_2=B\}} \left(R_1 = B_1 \log_2\left(1+\frac{g_1P_1}{N_0B_1}\right), R_2 = B_2\log_2\left(1+\frac{g_2P_2}{N_0B_2}\right)\right), \tag{14.45}$$

which is plotted in Figure 14.14. Clearly this region dominates TD, since setting $B_1 = \tau B$ and $B_2 = (1-\tau)B$ in (14.45) has $R_1 > \tau C_1$ and $R_2 > (1-\tau)C_2$. It can be shown [34] that this curve touches the capacity region boundary at one point, and this point corresponds to the rate vector that maximizes the sum-rate $R_1 + R_2$. To achieve this point, the bandwidths B_1 and B_2 must be proportional to their corresponding received powers g_1P_1 and g_2P_2.

As with the BC, we can obtain the same achievable rate region with TD as with FD by efficient use of the transmit power. If we take the constraints P_1 and P_2 to be average power constraints, then since user k only uses the channel a fraction τ_k of the time, its average power over that time fraction can be increased to P_k/τ_k. The rate region achievable through variable-power TD is then given by

$$\mathcal{C}_{TD,VP} = \bigcup_{\{\tau_1,\tau_2:\tau_1+\tau_2=1\}} \left(R_1 = \tau_1 B \log_2\left(1+\frac{g_1P_1}{N_0\tau_1B}\right), R_2 = \tau_2 B\log_2\left(1+\frac{g_2P_2}{N_0\tau_2B}\right)\right), \tag{14.46}$$

and substituting $B_k \overset{\triangle}{=} \tau_k B$ in (14.46) yields the same rate region as in (14.45).

Superposition codes without successive decoding can also be used. With this approach, each transmitter's message acts as noise to the others. Thus, the maximum achievable rate in this case cannot exceed (C_1^*, C_2^*), which is clearly dominated by FD and TD for some bandwidth or time allocations, in particular the allocation that intersects the rate region boundary.

Example 14.8: Consider a MAC channel in AWGN with transmit power $P_1 = P_2 = 100$ mW for both users, and channel gains $g_1 = .08$ for user 1 and $g_2 = .001$ for user 2. Assume the receiver noise has $N_0 = 10^{-9}$ W/Hz and the system bandwidth is $B = 100$ KHz. Find the corner points of the MAC capacity region. Also find the rate that user 1 can achieve if user 2 requires a rate of $R_2 = 100$ Kbps and of $R_2 = 50$ Kbps.

Solution: From (14.42)-(14.44) we have $C_1 = B\log_2\left(1+\frac{g_1P_1}{N_0B}\right) = 6.34\times 10^5$, $C_2 = B\log_2\left(1+\frac{g_2P_2}{N_0B}\right) = 1\times 10^5$,

$$C_1^* = B\log_2\left(1+\frac{g_1P_1}{N_0B+g_2P_2}\right) = 5.36\times 10^5,$$

and

$$C_2^* = B\log_2\left(1+\frac{g_2P_2}{N_0B+g_1P_1}\right) = 1.77\times 10^3.$$

The maximum rate for user 2 is 100 Kbps, so if he requires $R_2 = 100$ Kbps, this rate point is associated with the corner point (C_1^*, C_2) of the capacity region, so user 1 can achieve a rate of $R_1 = C_1^* = 536$ Kbps. If user 2 requires only $R_2 = 50$ Kbps then the rate point lies on the TD portion of the capacity region. In particular, timesharing as $\tau(C_1, C_2^*) + (1-\tau)(C_1^*, C_2)$, the timeshare value that yields $R_2 = 50$ Kbps satisfies $\tau C_2^* + (1-\tau)C_2 = R_2$. Solving for τ yields $\tau = (R_2 - C_2)/(C_2^* - C_2) = .51$, about halfway between the two corner points. Then user 1 can get rate $R_1 = \tau C_1 + (1-\tau)C_1^* = 5.86 \times 10^5$. This example illustrates the dramatic impact of the near-far effect in MAC channels. Even though both users have the same transmit power, the channel gain of user 2 is much less than the gain of user 1. Hence, user 2 can achieve at most a rate of 100 Kbps, whereas user 1 can achieve a rate between 536 and 634 Kbps. Moreover, the interference from user 2 does not have that much of an impact on user 1 due to the weak channel gain associated with the interference: user 1 sees data rates of $C_1 = 634$ Kbps without interference and $C_1^* = 536$ Kbps with interference, However, the interference from user 1 severely limits the data rate of user 2, decreasing it almost two orders of magnitude from $C_2 = 100$ Kbps to $C_2^* = 1.77$ Kbps.

14.6.2 Capacity in Fading

We now consider the capacity region of a MAC with AWGN and fading, where the channel gains for each user change over time. We assume all transmitters and the receiver have a single antenna and that the receiver has AWGN with PSD $N_0/2$. Each user has an individual power constraint $\overline{P}_k, k = 1, \ldots, K$. The time-varying power gain of user k's channel at time i is $g_k[i]$ and is independent of the fading of other users. We define the fading state at time i as $\mathbf{g}[i] = (g_1[i], \ldots, g_K[i])$, with the time reference dropped when the context is clear. We assume perfect CSI about the fading state at both the transmitter and receiver; the case of receiver CSI only is treated in [45, Chapter 6.3]. Like the BC and single-user channels, the fading MAC also has two notions of capacity: the ergodic capacity region that characterizes the achievable rate vectors averaged over all fading states, and the outage capacity region that characterizes the maximum rate vector that can be maintained in all states with possibly some nonzero probability of outage.

We first consider the ergodic capacity region, as derived in [40]. Define a power policy $\mathcal{P}$ as a function that maps a fading state $\mathbf{g} = (g_1, \ldots, g_K)$ to a set of powers $P_1(\mathbf{g}), \ldots, P_K(\mathbf{g})$, one for each user. Let $\mathcal{F}_{MAC}$ denote the set of all power policies satisfying the average per-user power constraint $\overline{P}_k$:

$$\mathcal{F}_{MAC} \equiv \left\{\mathcal{P} : \ \mathbf{E_g}\left[P_k(\mathbf{g})\right] \le \overline{P}_k, k = 1, \ldots, K\right\}.$$

The MAC capacity region assuming a constant fading state $\mathbf{g}$ with power allocation $P_1(\mathbf{g}), \ldots, P_K(\mathbf{g})$ is given by

$$\mathcal{C}_{MAC}(P_1(\mathbf{g}), \ldots, P_K(\mathbf{g})) = \left\{(R_1, \ldots, R_K) : \sum_{k\in S} R_k \le B\log_2\left(1 + \frac{\sum_{k\in S} g_k P_k(\mathbf{g})}{N_0 B}\right), \forall S \subset \{1, 2, \ldots, K\}\right\}. \tag{14.47}$$

The set of achievable rates averaged over all fading states under power policy $\mathcal{P}$ is given by

$$\mathcal{C}_{MAC}(\mathcal{P}) = \left\{(R_1, \ldots, R_K) : \sum_{k\in S} R_k \le \mathbf{E_g}\left[B\log_2\left(1 + \frac{\sum_{k\in S} g_k P_k(\mathbf{g})}{N_0 B}\right)\right], \forall S \subset \{1, 2, \ldots, K\}\right\}. \tag{14.48}$$

The ergodic capacity region is then the union over all power policies that satisfy the individual user power constraints:

$$\mathcal{C}_{MAC}(\overline{P}_1, \ldots, \overline{P}_K) = \bigcup_{\mathcal{P}\in\mathcal{F}_{MAC}} \mathcal{C}_{MAC}(\mathcal{P}). \tag{14.49}$$

From (14.41), (14.48), and (14.49), the sum-rate capacity of the MAC in fading reduces to

$$\mathcal{C}_{MACSR} = \max_{\mathcal{P}\in\mathcal{F}_{MAC}} \mathbf{E_g}\left[B\log_2\left(1+\frac{\sum_{k=1}^{K}\mathbf{g}_kP_k(\mathbf{g})}{N_0B}\right)\right]. \tag{14.50}$$

The maximization in (14.50) is solved using Lagrangian techniques, and the solution reveals that the optimal transmission strategy to achieve sum-rate is to only allow one user to transmit in every fading state [62]. Under this optimal policy the user that transmits in a given fading state $\mathbf{g}$ is the one with the largest *weighted* channel gain $\mathbf{g}_k/\lambda_k$, where λ_k is the Lagrange multiplier associated with the average power constraint of the kth user. This Lagrangian is a function of the user's average power constraint and fading distribution. By symmetry, if all the users have the same fading distribution and the same average power constraint, then the λ_ks are the same for all users, and the optimal policy is to allow only the user with the best channel $\mathbf{g}_k$ to transmit in fading state $\mathbf{g}$. Once it is determine which user should transmit in a given state, the power the user allocates to that state is determine via a water-filling over time. The intuition behind only allowing one user at a time to transmit is as follows. Since users can adapt their powers over time, system resources are best utilized by assigning them to the user with the best channel and allowing that user to transmit at a power commensurate with his channel quality. When users have unequal average received power this strategy is no longer optimal, since users with weak average received SNR would rarely transmit, so their individual power resources would not be utilized as effectively as they could be.

The MAC zero-outage capacity region, derived in [42], defines the set of rates that can be simultaneously achieved for all users in *all* fading states while meeting the average power constraints of each user. From (14.40), given a power policy $\mathcal{P}$ that maps fading states to user powers, the MAC capacity region in state $\mathbf{g}$ is

$$\mathcal{C}_{MAC}(\mathcal{P}) = \left\{(R_1,\ldots,R_K) : \sum_{k\in S}R_k \le B\log_2\left(1+\frac{\sum_{k\in S}g_kP_k(\mathbf{g})}{N_0B}\right), \forall S\subset\{1,2,\ldots,K\}\right\}. \tag{14.51}$$

Then under policy $\mathcal{P}$ the set of rates that can be maintained in all fading states $\mathbf{g}$ is

$$\mathcal{C}^0_{MAC}(\mathcal{P}) = \bigcap_{\mathbf{g}}\mathcal{C}_{MAC}(\mathcal{P}). \tag{14.52}$$

The zero-outage capacity region is then the union of $\mathcal{C}^0_{MAC}(\mathcal{P})$ over all power policies $\mathcal{P}$ that satisfy the user power constraints of $\mathcal{C}^0_{MAC}(\mathcal{P})$. Thus, the zero-outage MAC capacity region is given by

$$\mathcal{C}^0_{MAC}(\overline{P}_1,\ldots,\overline{P}_K) = \bigcup_{\mathcal{P}\in\mathcal{F}_{MAC}}\bigcap_{\mathbf{g}}\mathcal{C}_{MAC}(\mathcal{P}). \tag{14.53}$$

The outage capacity region of the MAC is similar to the zero-outage capacity region, except that users can suspend transmission in some outage states subject to a given nonzero probability of outage. As with the BC, the MAC outage capacity region is more difficult to obtain than the zero-outage capacity region, since in any given fading state the transmission strategy must determine which users to put into outage, the decoding order of the nonoutage users, and the power at which these nonoutage users should transmit. The MAC outage capacity region is obtained implicitly in [44] by determining whether a given rate vector $\mathbf{R}$ can be maintained in all fading states, subject to a given per-user outage probability, without violating the per-user power constraints. Ergodic and outage capacities can also be combined to obtain the minimum rate capacity region for the MAC. As with the BC, this region characterizes the set of all average rate vectors that can be maintained, averaged over all fading states, subject to some minimum rate vector that must be maintained in all states with some outage probability (possibly zero). The minimum rate capacity region for the fading MAC is derived in [48] using the duality principle that relates capacity regions of the BC and the MAC. This duality principle is described in the next section.

14.6.3 Capacity with Multiple Antennas

We now consider MAC channels with multiple antennas. We will model the channel based on symmetry between the MIMO BC on the downlink and the corresponding MIMO MAC on the uplink. As in the MIMO BC model, we normalize bandwidth to unity, $B = 1$ Hz, and assume the noise vector $\mathbf{n}$ at the MAC receiver is a circularly symmetric complex Gaussian with $\mathbf{n} \sim N(0, \mathbf{I})$. Since the channel gains on an uplink and downlink are generally symmetric, if the channel matrix of user k on the MIMO BC is given by $\mathbf{H}_k$, then the channel gains on the MIMO MAC corresponding to the uplink of the BC are given by $\mathbf{H}_k^H$. Define $\mathbf{H}^H = [\mathbf{H}_1^H \ldots \mathbf{H}_K^H]$. Then the capacity region of the Gaussian MIMO MAC where user k has channel gain matrix $\mathbf{H}_k^H$ and power P_k is given by [51, 52, 53]

$$\mathcal{C}_{\text{MAC}}((P_1, \ldots, P_K); \mathbf{H}^H) \;=\; \bigcup_{\{\mathbf{Q}_k \geq 0,\ \text{Tr}(\mathbf{Q}_k) \leq P_k \ \ \forall k\}} \left\{ \begin{array}{l} (R_1, \ldots, R_K) : \\ \sum_{k \in S} R_k \leq \log |\mathbf{I} + \sum_{k \in S} \mathbf{H}_k^H \mathbf{Q}_k \mathbf{H}_k| \ \ \forall S \subseteq \{1, \ldots, K\} \end{array} \right\} \tag{14.54}$$

This region is achieved as follows. The kth user transmits a zero-mean Gaussian with spatial covariance matrix $\mathbf{Q}_k$. Each set of covariance matrices $(\mathbf{Q}_1, \ldots, \mathbf{Q}_K)$ corresponds to a K-dimensional polyhedron (i.e. $\{(R_1, \ldots, R_K) : \sum_{k \in S} R_k \leq \frac{1}{2} \log |\mathbf{I} + \sum_{k \in S} \mathbf{H}_k^H \mathbf{Q}_k \mathbf{H}_k| \ \ \forall S \subseteq \{1, \ldots, K\}\}$), and the capacity region is equal to the union (over all covariance matrices satisfying the power constraints) of all such polyhedrons. The corner points of this pentagon can be achieved by successive decoding, in which users' signals are successively decoded and subtracted out of the received signal. Note that the capacity region (14.54) has several similarities with its single-antenna counterpart: it is defined based on the rate sum associated with subsets of users, and the corner points of the region are obtained using successive decoding.

For the two-user case, each set of covariance matrices corresponds to a pentagon, similar in form to the capacity region of the single-antenna MAC. For example, the corner point where $R_1 = \log |\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1|$ and $R_2 = \log |\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1 + \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2| - R_1 = \log |\mathbf{I} + (\mathbf{I} + \mathbf{H}_1^H \mathbf{Q}_1 \mathbf{H}_1)^{-1} \mathbf{H}_2^H \mathbf{Q}_2 \mathbf{H}_2|$ corresponds to decoding User 2 first (i.e. in the presence of interference from User 1) and decoding User 1 last (without interference from User 2).

14.7 Uplink/Downlink Duality

The downlink and uplink channels shown in Figure 14.1 appear quite similar: the downlink is almost the same as the uplink with the direction of the arrows reversed. There are three fundamental differences between the two channel models. First, in the downlink there is an additive noise term associated with each receiver, whereas in the uplink there is only one additive noise term since there is only one receiver. Another fundamental difference is that the downlink has a single power constraint associated with the transmitter, whereas the uplink has different power constraints associated with each user. Finally, on the downlink both the signal and interference associated with each user travel through the same channel, whereas on the uplink these signals travel through different channels, giving rise to the near-far effect. Despite extensive study of uplink and downlink channels individually, there has been little effort to draw connections between the two models or exploit these connections in analysis and design. In this section we will describe a duality relationship between these two channels, and show how this relationship can be used in capacity analysis and in the design of uplink and downlink transmission strategies.

We say that K-user downlink and uplink, as shown in Figure 14.1 for $K = 3$, are **duals** of each other under the following three conditions:

- The channel impulse responses $h_k(t), k = 1, \ldots, K$ in the downlink are the same as in the uplink for all k.

- Each receiver in the downlink has the same noise statistics and these statistics are the same as those of the receiver noise in the uplink.
- The power constraint P on the downlink equals the sum of individual power constraints $P_k, k = 1, \ldots, K$ on the uplink.

Despite the similarities between the downlink (BC) and uplink (MAC), their capacity regions are quite different. In particular, the two-user AWGN BC capacity region shown by the largest region in Figure 14.10 is markedly different from the two-user AWGN MAC capacity region shown in Figure 14.14. The capacity regions of dual MACs and BCs are also very different in fading under any of the fading channel capacity definitions: ergodic, outage, or minimum-rate capacity. However, despite their different shapes, the capacity regions of the dual channels are both achieved using a superposition coding strategy, and the optimal decoders for the dual channels exploit successive decoding and interference cancellation.

The duality relationship between the two channels is based on exploiting their similar encoding and decoding strategies while bridging their differences by summing the individual MAC power constraints to obtain the BC power constraint and scaling the BC gains to achieve the near-far effect of the MAC. This relationship was developed in [48], where it was used to show that the capacity region and optimal transmission strategy of either the BC or the MAC can be obtained from the capacity region and optimal transmission strategy of the dual channel. In particular, it was shown in [48] that the capacity region of the AWGN BC with power P and channel gains $\mathbf{g} = (g_1, \ldots, g_K)$ is equal to the capacity region of the dual AWGN MAC with the same channel gains, but where the MAC is subject to a sum power constraint $\sum_{k=1}^K P_k \leq P$ instead of individual power constraints $(P_1, \ldots, P_k)$. The sum power constraint in the MAC implies that the MAC transmitters draw power from a single pooled power source with total power P, and that power is allocated between the MAC transmitters such that $\sum_{k=1}^K P_k \leq P$. Mathematically, the BC capacity region can be expressed as the union of capacity regions for its dual MAC with a pooled power constraint as [48]

$$\mathcal{C}_{BC}(P, \mathbf{g}) = \bigcup_{\{(P_1,\ldots,P_K):\sum_{i=1}^K P_k = P\}} \mathcal{C}_{MAC}(P_1, \ldots, P_K; \mathbf{g}). \tag{14.55}$$

where $\mathcal{C}_{BC}(P, \mathbf{g})$ is the AWGN BC capacity region with total power constraint P and channel gains $\mathbf{g} = (g_1, \ldots, g_K)$, as given by (14.18) with $n_k = N_0/g_k$, and $\mathcal{C}_{MAC}(P_1, \ldots, P_K; \mathbf{g})$ is the AWGN MAC capacity region with individual power constraints $P_1, \ldots, P_K$ and channel gains $\mathbf{g} = (g_1, \ldots, g_K)$, as given by (14.40). This relationship is illustrated for two users in Figure 14.15 where we see the BC capacity region formed from the union of MAC capacity regions with different power allocations between MAC transmitters that sum to the total power P of the dual BC.

In addition to the capacity region relationship of (14.55), it is also shown in [48] that the optimal power allocation for the BC associated with any point on the boundary of its capacity region can be obtained from the allocation of the sum-power on the dual MAC that intersects with that point. Moreover, the decoding order of the BC for that intersection point is the reverse decoding order of this dual MAC. Thus, the optimal encoding and decoding strategy for the BC can be obtained from the optimal strategies associated with its dual MAC. This connection between optimal uplink and downlink strategies may have interesting implications for practical designs.

Duality also implies that the MAC capacity region can be obtained from that of its dual BC. This relationship is based on the notion of channel scaling. It is easily seen from (14.40) that the AWGN MAC capacity region is not affected if the kth user's channel gain g_k is scaled by power gain α as long as its power P_k is also scaled by $1/\alpha$. However, the dual BC is fundamentally changed by channel scaling since the encoding and decoding order of superposition coding on the BC is determined by the order of the channel gains. Thus, the capacity region of the BC with different channel scalings will be different, and it is shown in [48] that the MAC capacity region can

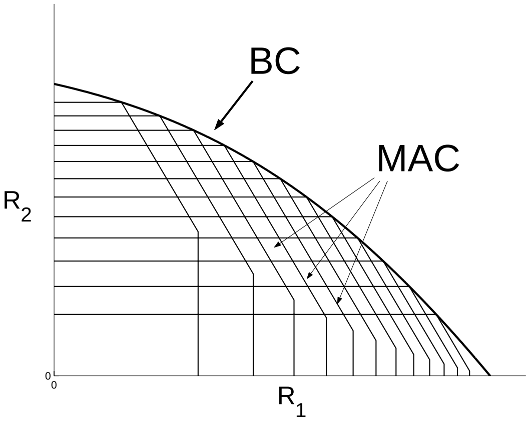


Figure 14.15: AWGN Downlink (BC) Capacity Region as a Union of Capacity Regions for the Dual Uplink (MAC)

be obtained by taking an intersection of the BC with all possible channel scalings α_k on the kth user's channel. Mathematically, we obtain the MAC capacity region from the dual BC as

$$\mathcal{C}_{MAC}(P_1, \ldots, P_K; \mathbf{g}) = \bigcap_{(\alpha_1, \ldots, \alpha_K) > 0} \mathcal{C}_{BC}\left(\sum_{k=1}^{K} P_k/\alpha_k; (\alpha_1 g_1, \ldots, \alpha_K g_K)\right). \qquad (14.56)$$

This relationship is illustrated for two users with channel gain $\mathbf{g} = (g_1, g_2)$ in Figure 14.16. This figure shows that the MAC capacity region is formed from the intersection of BC capacity regions with different channel scalings α applied to the first user.[9]. As $\alpha \to 0$, the channel gain αg_1 of the 1st user goes to zero but the total power $P = P_1/\alpha + P_2$ goes to infinity. Since user 2's channel gain doesn't change, he takes advantage of the increased power and his rate grows asymptotically large with α. The opposite happens as $\alpha \to \infty$, user 1's channel gain grows and the total power $P = P_1/\alpha + P_2 \geq P_2$, so user 1 takes advantage of his increasing channel gain to get asymptotically large rate with any portion of the total power P. All scalings between zero and infinity sketch out different BC capacity regions that intersect to form the MAC region. In particular, when $\alpha = g_2/g_1$, the channel gains of both users in the scaled BC channel are the same, and this yields the time-sharing segment of the MAC capacity region. The optimal decoding order of the MAC for a given point on its capacity region can also be obtained from the channel scaling associated with the dual scaled BC whose capacity region intersects the MAC capacity region at that point.

These duality relationships are extended in [48] to many other important channel models. In particular, duality applies to fading MACs and BCs, so that the ergodic, outage, and minimum-rate capacity regions, along with the optimal encoding and decoding strategies, for one channel can be obtained from the regions and strategies for the dual channel. MAC and BC duality also holds for parallel and frequency-selective fading channels, which defines the connection between the capacity regions of MACs and BCs with ISI [49, 50]. Another important application of duality is to multiple antenna (MIMO) MACs and BCs. In [56] the notion of duality between the BC and MAC was extended to MIMO systems such that the MIMO BC capacity region with power constraint P was shown to equal to the union of capacity regions of the dual MAC, where the union is taken over all individual power constraints

[9] It is sufficient to take the intersection for scaling over just $K-1$ users because scaling by $(\alpha_1, \ldots, \alpha_{K-1}, \alpha_K)$ is equivalent to scaling by $(\frac{\alpha_1}{\alpha_K}, \ldots, \frac{\alpha_{K-1}}{\alpha_K}, 1)$

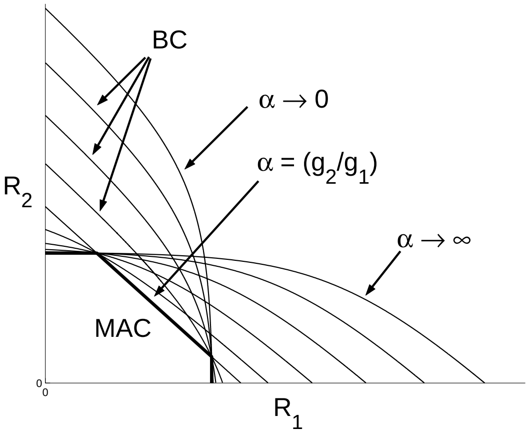


Figure 14.16: AWGN Uplink (MAC) Capacity Region as an Intersection of Capacity Regions for the Scaled Dual Downlink (BC)

that sum to P. Mathematically

$$\mathcal{C}_{\rm BC}(P,\mathbf{H}) = \bigcup_{(P_1,\ldots,P_K):\sum_{k=1}^K P_k = P} \mathcal{C}_{MAC}((P_1,\ldots,P_K);\mathbf{H}^H).$$

This duality relationship is illustrated in Figure 14.13, where the MIMO BC capacity region is defined by the outer boundary in the figure. The regions inside this boundary are the MIMO MAC capacity region under different individual user power constraints that sum to the total BC power P. Recall that the MIMO BC capacity region is extremely difficult to compute direction, since it is not concave or convex over the covariance matrices that must be optimized. However, the optimal MIMO MAC is obtained via a standard convex optimization that is easy to solve [61]. Moreover, duality not only relates the two capacity regions, but can also be used to obtain the optimal transmission strategy on the MIMO BC capacity region from a duality transformation of the optimal MIMO MAC strategy that achieves the same point. Thus, for MIMO channels, duality can not only be exploited to greatly simplify the calculations in finding the capacity region, but it also greatly simplifies finding the corresponding optimal transmission strategy.

14.8 Multiuser Diversity

Multiuser diversity takes advantage of the fact that in a system with many users whose channels fade independently, at any given time some users will have better channels than others. By transmitting only to users with the best channels at any given time, system resources are allocated to the users that can best exploit them, which leads to improved system capacity and/or performance. Multiuser diversity was first explored in [62] as a means to increase throughput and reduce error probability in uplink channels, and the same ideas can be applied to downlink channels. The multiuser diversity concept is an extension of the single-user diversity concepts described in Chapter 7. In single-user diversity systems a point-to-point link consists of multiple independent channels whose signals can be combined to improve performance. In multiuser diversity the multiple channels are associated with different users, and the system typically uses selection-diversity to select the user with the best channel in any given fading state. The multiuser diversity gain relies on disparate channels between users, so the larger the dynamic range of the fading, the higher the multiuser diversity gain. In addition, as with any diversity technique, performance improves with the number of independent channels. Thus, multiuser diversity is most effective in systems with a large number of users.

From Section 14.5, we have seen that the total throughput (sum-rate capacity) of the fading downlink is maximized by allocating the full system bandwidth to the user with the best channel in each fading state. As described in Section 14.6, a similar result holds for the fading uplink if all users have the same fading distribution and average power. If the users have different fading statistics or average powers, then the channel in any given state is allocated to the user with the best weighted channel gain, where the weight depends on the user's channel gain in the given state, his fading statistics, and his average power constraint. The notion of scheduling transmissions to users based on their channel conditions is called **opportunistic scheduling**, and numerical results in [62, 41] show that opportunistic scheduling coupled with power control can significantly increase both uplink and downlink throughput as measured by sum-rate capacity.

Opportunistic scheduling can also improve BER performance [62]. Let $\gamma_k[i], k = 1, \ldots, K$ denote the SNR for each user's channel at time i. By transmitting only to the user with the largest SNR, the system SNR at time i is $\gamma[i] = \max_k \gamma_k[i]$. It is shown in [63] that in i.i.d. Rayleigh fading this maximum SNR is roughly $\ln K$ larger than the SNR of any one user as K grows asymptotically large, leading to a multiuser diversity gain in SNR of $\ln K$. Moreover, if $P_s(\gamma)$ denotes the probability of symbol error for the user with the best channel gain at time i, then $P_s(\gamma)$ will exhibit the same diversity gains as selection-combining in a single-user system (described in Chapter **??**) as compared to the probability of error associated with any one user. As the number of users in the system increases, the probability of error approaches that of an AWGN channel without fading, analogous to increasing the number of branches in single-user selection-combining diversity.

Scheduling transmission to users with the best channel raises two problems in wireless systems: fairness and delay. If user fade levels change very slowly, then one user will occupy the system for a long period of time. The time between channel uses for any one user could be quite long, and such latency might be unacceptable for a given application. In addition, users with poor average SNRs will rarely have the best channel and therefore rarely get to transmit, which leads to unfairness in the allocation of the system resources. A solution to the fairness and delay problems in the downlink called **proportional fair scheduling** was proposed in [63]. Suppose at time i each of the K users in the downlink system can support rate $R_k[i]$ if allocated the full power and system bandwidth. Let $T_k[i]$ denote that the average throughput of the kth user at time i, averaged over a time window $[i - i_c, i]$, where the window size i_c is a parameter of the scheduler design. In the ith time slot, the scheduler then transmits to the user with the largest ratio $R_k[i]/T_k[i]$. With this scheduler, if at time i all users have experienced the same average throughput $T_k[i] = T[i]$ over the prior time window then the scheduler transmits to the user with the best channel. Suppose, however, that one user, user j, has experienced poor throughput over the prior time window so that $T_j[i] << T_k[i], j \neq k$. Then at time i user j will likely have a high ratio of $R_j[i]/T_j[i]$ and thus will be favored in the allocation of resources at time i. Assuming that at time i the user k^* has the highest ratio of $R_k[i]/T_k[i]$, the throughput on the next timeslot is updated as

$$T_k(i+1) = \begin{cases} \left(1 - \frac{1}{i_c}\right) T_k(i) + \frac{1}{i_c} R_k(i) & k = k^* \\ \left(1 - \frac{1}{i_c}\right) T_k(i) & k \neq k^* \end{cases} \tag{14.57}$$

With this scheduling scheme, users with the best channels are still allocated the channel resources when throughput between users is reasonably fair. However, if the throughput of any one user is poor, that user will be favored for resource allocation until his throughput becomes reasonably balanced with that of the other users. Clearly this scheme will have a lower throughput than allocating all resources to the user with the best channel, which maximizes throughput, and the throughput penalty will increase as the users have more disparate average channel qualities. The latency with this scheduling scheme is controlled via the time window i_c. As the window size increases the latency also increases, but system throughput increases as well since the scheduler has more flexibility in allocating resources to users. As the window size grows to the entire transmission time, the proportional fair scheduler just reduces to allocating system resources to the user with the best channel. The proportional fair scheduling algorithm is part of the standard for packet data transmission in CDMA2000 cellular systems [64] and its performance for that system is evaluated in [65]. Alternative methods for incorporating fairness and delay constraints in opportunistic scheduling have been evaluated in [66, 67], along with their performance under practical constraints such as imperfect channel estimates.

14.9 MIMO Multiuser Systems

Multiuser systems with multiple antennas at the transmitter(s) and/or receiver(s) are called MIMO multiuser systems. These multiple antennas can significantly enhance performance in multiple ways. The antennas can be used to provide diversity gain to improve BER performance. The capacity region of the multiuser channel is increased by MIMO, providing multiplexing gain. Finally, multiple antennas can provide directivity gain to spatially separate users, which reduces interference. There is typically a tradeoff between these three types of gains in MIMO multiuser systems [68].

The multiplexing gain of a MIMO multiuser system characterizes the increase in the uplink or downlink capacity region associated with adding multiple antennas. The capacity regions of MIMO multiuser channels have been extensively studied, motivated by the large capacity gains associated with single-user systems. For AWGN channels the MIMO capacity region is known for both the uplink [51] and the downlink [58]. These results can be extended to find the MIMO capacity region in fading with perfect CSI at all transmitters and receivers. Capacity results and open problems related to MIMO multiuser fading channels under other assumptions about channel CSI are described in [69].

Beamforming was discussed in Chapter 10.4 as a technique to achieve full diversity in single-user systems at the expense of some capacity loss. In multiuser systems, beamforming has less of a capacity penalty due to the multiuser diversity effect, and in fact beamforming can achieve the sum-rate capacity of the MIMO downlink in the asymptotic limit of a large number of users [71, 72].

Multiuser diversity is based on the idea that in multiuser channels the channel quality varies across users, so performance can be improved by allocating system resources at any given time to the users with the best channels. Design techniques to exploit multiuser diversity were discussed in Section 14.8 for single-antenna multiuser systems. In MIMO multiuser systems the benefits of multiuser diversity are two-fold. First, MIMO multiuser diversity provides improved channel quality since only users with the best channels are allocated system resources. In addition, MIMO multiuser diversity provides abundant directions where users have good channel gains, so that the users chosen for resource allocation in a given state not only have very good channel quality, but they also have good spatial separation, thereby limiting interference between them. This two-fold diversity benefit allows relatively simple suboptimal transmitter and receiver techniques to have near-optimal performance as the number of users increases [73, 71]. It also eliminates the requirement for multiple receive antennas in downlinks and multiple transmit antennas in uplinks to obtain large capacity gains, which simplifies mobile terminal design. In particular, the sum-rate capacity gain in MIMO BCs increases roughly linearly with the number of users and transmit antennas, independent of the number of receive antennas at each user and similarly, the sum-rate capacity gain in MIMO MACs increases roughly linearly with the number of users and receive antennas, independent of the number of transmit antennas at each user [75]. Note that multiuser diversity increases with the dynamic range and rate of the channel fading. By modulating in a controlled fashion the amplitude and phase of multiple transmit antennas, the fading rate and dynamic range can be increased, leading to higher multiuser diversity gains. This technique, called **opportunistic beamforming**, is investigated in [63].

Space-time modulation and coding techniques for MIMO multiuser systems have also been developed [76, 77, 70]. The goal of these techniques is to achieve the full range of diversity, multiplexing, and directivity tradeoffs inherent to MIMO multiuser systems. Multiuser detection techniques can also be extended to MIMO channels and provide substantial performance gains [79, 78, 80]. In wideband channels the multiuser MIMO techniques must also cope with frequency-selective fading [81, 82, 83]. Advanced transmission techniques for these wideband channels promise even more significant performance gains than in narrowband channels, since frequency-selective fading provides yet another form of diversity. The challenge for MIMO multiuser systems is to develop signaling techniques of reasonable complexity that deliver on the promised performance gains even in practical operating environments.

Bibliography

[1] D. Bertsekas and R. Gallager, Data Networks, 2nd Edition, Prentice Hall 1992.

[2] T.S. Rappaport, *Wireless Communications - Principles and Practice*, IEEE Press, 1996.

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

[4] W. Stallins, *Wireless Communications and Networks*, 2nd Ed., Prentice Hall, 2005.

[5] G. Leus, S. Zhou, and G.B. Giannakis, "Orthogonal multiple access over time- and frequency-selective channels," *IEEE Trans. Inform. Theory*, Vol. 49, pp. 1942-1950, Aug. 2003.

[6] S. Verdú, "Demodulation in the presence of multiuser interference: progress and misconceptions," *Intelligent Methods in Signal Processing and Communications,* Eds. D. Docampo, A. Figueiras, and F. Perez-Gonzalez, pp. 15-46, Birkhauser Boston, 1997.

[7] M. Gudmundson, "Generalized frequency hopping in mobile radio systems," *Proc. IEEE Vehic. Technol. Conf.*, pp. 788-791, May 1993.

[8] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr., and C. E. Wheatley III, "On the capacity of a cellular CDMA system," *IEEE Trans. Vehic. Technol.*, pp. 303–312, May 1991.

[9] B. Gundmundson, J. Sköld, and J.K. Ugland, "A comparison of CDMA and TDMA systems," *IEEE Vehic. Technol. Conf. Rec.*, pp. 732–735, May 1992.

[10] P. Jung, P.W. Baier, and A. Steil, "Advantages of CDMA and spread spectrum techniques over FDMA and TDMA in cellular mobile radio applications," *IEEE Trans. Vehic. Technol.*, pp. 357–364, Aug. 1993.

[11] J. Chuang and N. Sollenberger, "Beyond 3G: wideband wireless data access based on OFDM and dynamic packet assignment," *IEEE Commun. Mag.*, Vol. 38, pp. 78-87, July 2000.

[12] K.R. Santhi, V.K. Srivastava, G. SenthilKumaran, and A. Butare, "Goals of true broadband's wireless next wave (4G-5G)," *Proc. IEEE Vehic. Technol. Conf.*, pp. 2317 - 2321, Oct. 2003.

[13] M. Frodigh, S. Parkvall, C. Roobol, P. Johansson, and P. Larsson, "Future-generation wireless networks," *IEEE Wireless Commun. Mag.*, Vol. 8, pp. 10-17, Oct. 2001.

[14] E. Anderlind and J. Zander, "A traffic model for non-real-time data users in a wireless radio network," *IEEE Commun. Lett*, Vol. 1, pp. 37-39, March 1997.

[15] K. Pahlavan and P. Krishnamurthy, *Principles of Wireless Networks: A Unified Approach*, Prentice Hall, 2002.

[16] N. Abramson, ”The Aloha system - another alternative for computer communications,” Proc. Fall Joint Comput. Conf., AFIPS Conf,. p. 37, 1970.

[17] V. Bharghavan, A. Demers, S. Shenkar, and L. Zhang, “MACAW: A media access protocol for wireless LAN’s,” in *Proc. ACM SIGCOMM*, London, UK, Aug. 1994, vol. 1, pp. 212–225.

[18] *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard 802.11, 1997.

[19] A. Chockalingam and M. Zorzi, ”Energy consumption performance of a class of access protocols for mobile data networks,” Proc. IEEE Vehic. Technol. Conf. pp. 820-824, May 1998.

[20] P. Karn, ”MACA: A new channel access method for packet radio,” Proc. Comp. Net. Conf., pp. 134-140, Sept. 1990.

[21] Z.J. Haas, J. Deng, and S. Tabrizi, ”Collision-free medium access control scheme for ad hoc networks, Proc. Milt. Commun. Conf. (MILCOM), pp. 276-280, 1999.

[22]] S.-L. Wu, Y.-C. Tseng and J.-P. Sheu, “Intelligent Medium Access for Mobile Ad Hoc Networks with Busy Tones and Power Control”, IEEE J. Select. Areas Commun., pp. 1647- 1657, Sept. 2000.

[23] N. Abramson, “Wide-band random-access for the last mile,” *IEEE Pers. Commun. Mag.*, Vol. 3, No. 6, pp. 29–33, Dec. 1996.

[24] D.J. Goodman, R.A. Valenzuela, K.T. Gayliard, and B. Ramamurthi, “Packet reservation multiple access for local wireless communications,” *IEEE Trans. Commun.*, Vol. 37, pp. 885-890, Aug. 1989.

[25] N.B Mehta and A.J. Goldsmith, “Effect of fixed and interference-induced packet error probability on PRMA,” *IEEE Intl Conf. Commun.*, pp. 362-366, June 2000.

[26] P. Agrawal, ”Energy efficient protocols for wireless systems,” Proc. IEEE Intl. Symp. Personal, Indoor, Mobile Radio Commun., pp. 564-569, Sept. 1998.

[27] K.K. Parhi and R. Ramaswami, ”Distributed scheduling of broadcasts in a radio network,” Proc. IEEE INFOCOM, pages 497-504, March 1989.

[28] N. Bambos, S.C. Chen, and G.J. Pottie, “Channel access algorithms with active link protection for wireless communication networks with power control,” *IEEE/ACM Trans. Network.*, Vol. 8, pp. 583 - 597, Oct. 2000.

[29] S. Kandukuri and N. Bambos, ”Power controlled multiple access (PCMA) in wireless communication networks,” Proc. IEEE Infocom, pp. 386-395, March 2000.

[30] J. Zander, “Performance of optimum transmitter power control in cellular radio systems,” *IEEE Trans. Vehic. Technol.*, Vol. 41, pp. 57-62, Feb. 1992.

[31] S.A. Grandhi, R. Vijayan, and D.J. Goodman, “Distributed power control in cellular radio systems,” *IEEE Trans. Commun.*, Vol. 42, pp. 226-228, Feb.-Apr. 1994.

[32] G.J. Foschini and Z. Miljanic, “A simple distributed autonomous power control algorithm and its convergence,” *IEEE Trans. Vehic. Technol.*, Vol. 42, pp. 641 - 646, Nov. 1993.

[33] E. Seneta, “Nonnegative Matrices and Markov Chains”, New York: Springer, 1981.

[34] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[35] P.P. Bergmans and T.M. Cover, “Cooperative broadcasting,” *IEEE Trans. Inform. Theory*, Vol IT-20, No. 3, pp. 317–324, May 1974.

[36] P.P. Bergmans, “A simple converse for broadcast channels with additive white Gaussian noise,” *IEEE Trans. Inform. Theory*, Vol IT-20, No. 2, pp. 279–280, March 1974.

[37] L.-F. Wei, “Coded modulation with unequal error protection,” *IEEE Trans. Commun.*, Vol. COM-41, pp. 1439–1449, Oct. 1993.

[38] S. Verdú, *Multiuser Detection*, Cambridge University Press, 1998.

[39] R. Pickholtz, L. Milstein, and D. Schilling, “Spread spectrum for mobile communications,” *IEEE Trans. Vehic. Technol*, pp. 313-322, May 1991.

[40] D. Tse and S. Hanly, “Multiaccess fading channels–Part I:Polymatroid structure, optimal resource allocation and throughput capacities,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2796–2815, November 1998.

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

[42] S. Hanly and D. Tse, “Multiaccess fading channels–Part II:Delay-limited capacities,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2816–2831, November 1998.

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

[44] L. Li, N. Jindal, and A.J. Goldsmith, “Outage capacities and optimal power allocation for fading multiple access channels,” To appear: *IEEE Trans. Inform. Theory*, 2005.

[45] D.Tse and P. Viswanath, *Foundations of Wireless Communications*, Cambridge University Press, 2005.

[46] N. Jindal and A. J. Goldsmith, “Capacity and optimal power allocation for fading broadcast channels with minimum rates,” *IEEE Trans. Inform. Theory*, vol. 49, pp. 2895–2909, Nov. 2003.

[47] E. Larsson and P. Stoica, *Space-Time Block Coding for Wireless Communications*. Cambridge, England: Cambridge University Press, 2003.

[48] N. Jindal, S. Vishwanath, and A. J. Goldsmith, “On the duality of Gaussian multiple-access and broadcast channels,” *IEEE Trans. Inform. Theory*, Vol. 50, pp. 768-783, May 2004.

[49] R. Cheng and S. Verdú, “Gaussian multiaccess channels with ISI: capacity region and multiuser water-filling,” *IEEE Trans. Inform. Theory*, Vol. 39, pp. 773 - 785, May 1993.

[50] A. J. Goldsmith and M. Effros, “The capacity region of broadcast channels with intersymbol interference and colored Gaussian noise,” *IEEE Trans. Inform. Theory*, Vol. 47, pp. 219 - 240, Jan. 2001.

[51] S. Verdú, “Multiple-access channels with memory with and without frame synchronism,” *IEEE Trans. Info. Theory*, pp. 605-619, May 1989.

[52] E. Telatar, “Capacity of Multi-antenna Gaussian Channels,” European Trans. on Telecomm. ETT, 10(6):585-596, November 1999.

[53] W. Yu, W. Rhee, S. Boyd, J. Cioffi, "Iterative Water-filling for Vector Multipl e Access Channels", pp. 322, Proc. IEEE Int. Symp. Inf. Theory, (ISIT), Washington DC, June 24-29, 2001.

[54] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, Vol. 49, pp. 1691 - 1706, July 2003.

[55] W. Yu and J.M. Cioffi, "Trellis precoding for the broadcast channel," *Proc. Global. Telecomm. Conf.* pp. 1344-1348, Nov. 2001.

[56] S. Vishwanath, N. Jindal, and A. J. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inform. Theory*, Vol. 49, pp. 2658-2668, Oct. 2003.

[57] P. Viswanath and D.N.C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inform. Theory*, Vol. 49, pp. 1912 - 1921, Aug. 2003.

[58] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," *Proc. Intl. Symp. Inform. Theory*, pp. 174, June 2004.

[59] M. Costa. Writing on dirty paper. *IEEE Trans. Inform. Theory*, 29(3):439–441, May 1983.

[60] U. Erez, S. Shamai, and R. Zamir. Capacity and lattice strategies for cancelling known interference. In *International Symposium on Information Theory and its Applications*, pages 681–684, Nov. 2000.

[61] N. Jindal, W. Rhee, S. Vishwanath, S.A. Jafar, and A.J. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," To appear: *IEEE Trans. Inform. Theory*, 2005.

[62] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," *Proc. IEEE Intl. Conf. Commun.*, pp. 331-335, June 1995.

[63] P. Vishwanath, D.N.C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, Vol. 48, pp. 1277 - 1294, June 2002.

[64] TIA/EIA IS-856, "CDMA 2000: High rate packet data air interface specification," Std., Nov. 2000.

[65] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *Proc. IEEE Vehic. Technol. Conf.*, pp. 1854 - 1858, May 2000.

[66] X. Liu, E. K.P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Select. Areas Commun.*, Vol. 19, pp. 2053 - 2064, Oct. 2001.

[67] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, "Providing Quality of Service over a shared wireless link," *IEEE Commun. Mag.*, pp. 150 - 154, Feb. 2001.

[68] D.N.C Tse, P. Viswanath, and L. Zheng, "Diversity-multiplexing tradeoff in multiple-access channels," *IEEE Trans. Inform. Theory*, Vol. 50, pp. 1859 - 1874, Sept. 2004.

[69] A. J. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Select. Areas Commun.*, Vol. 21, pp. 684-702, June 2003.

[70] S.N. Diggavi, N. Al-Dhahir, and A.R. Calderbank, "Multiuser joint equalization and decoding of space-time codes," *Proc. IEEE Intl. Conf. Communm.*, Vol. 4, pp. 2643 - 2647, May 2003.

[71] M. Sharif and B. Hassibi, “Scaling laws of sum rate using time-sharing, DPC, and beamforming for MIMO broadcast channels,” *Proc. IEEE Intl. Symp. Inform. Theory*, pg. 175, June 2004.

[72] T. Yoo and A. J. Goldsmith, “Optimality of zero-forcing beamforming with multiuser diversity,” *Proc. IEEE Intl. Conf. Commun.*, May 2005.

[73] J. Heath, R.W., M. Airy, and A. Paulraj, “Multiuser diversity for MIMO wireless systems with linear receivers,” *Proc. Asilomar Conf. Signals, Systems and Computers*, Vol. 2, pp. 1194-1199, Nov. 2001.

[74] M. Sharif and B. Hassibi, “On the capacity of MIMO broadcast channels with partial side information,” *Proc. Asilomar Conf. Signals, Systems and Computers*, Vol. 1, pp.958–962, Nov. 2003.

[75] N. Jindal and A. Goldsmith, “DPC vs. TDMA for MIMO broadcast channels,” *IEEE Trans. Inform. Theory*, 2005.

[76] N. Al-Dhahir, C. Fragouli,, A. Stamoulis, W. Younis, R. Calderbank, “Space-time processing for broadband wireless access,” *IEEE Commun. Mag.*, Vol. 40, pp. 136-142, Sept. 2002.

[77] M. Brehler and M. K. Varanasi, “Optimum receivers and low-dimensional spreaded modulation for multiuser space-time communications,” *IEEE Trans. Inform. Theory*, Vol. 49, pp. 901-918, April 2003.

[78] S.N. Diggavi, N. Al-Dhahir and A.R. Calderbank, “On interference cancellation and high-rate space-time codes,” *Proc. IEEE Intl. Symp. Inform. Theory,*, pg. 238, June 2003.

[79] H. Dai and H.V. Poor, “Iterative space-time processing for multiuser detection in multipath CDMA channels,” *IEEE Trans. Sign. Proc.*, Vol. 50, pp. 2116 - 2127, Sept. 2002.

[80] S.J. Grant and J.K. Cavers, “System-wide capacity increase for narrowband cellular systems through multiuser detection and base station diversity arrays,” *IEEE Trans. Wireless Commun.*, Vol. 3, pp. 2072 - 2082, Nov. 2004.

[81] Z. Liu and G.B. Giannakis, “Space-time block-coded multiple access through frequency-selective fading channels,” *IEEE Trans. Commun.*, Vol. 49, pp. 1033 - 1044, June 2001.

[82] S.N. Diggavi, N. Al-Dhahir and A.R. Calderbank, “Multiuser joint equalization and decoding of space-time codes,” *Proc. IEEE Intl. Conf. Commun.*, pp. 2643 - 2647, May 2003.

[83] K.-K. Wong, R.D. Murch, and K.B. Letaief, “Performance enhancement of multiuser MIMO wireless communication systems,” *IEEE Trans. Commun.*, Vol. 50, pp. 1960 - 1970, Dec. 2002.

Chapter 14 Problems

1. Consider an FDMA system for multimedia data users. The modulation format requires 10 MHz of spectrum, and guard bands of 1 MHz are required on each side of the allocated spectrum to minimize out-of-band interference. What total bandwidth is required to support 100 simultaneous users in this system?

2. GSM systems have 25 MHz of bandwidth allocated to their uplink and downlink, divided into 125 TDMA channels, with 8 user timeslots per channel. A GSM frame consists of the 8 timeslots, preceeded by a set of preamble bits and followed by a set of trail bits. Each timeslot consists of 3 start bits at the beginning, followed by a burst of 58 data bits, then 26 equalizer training bits, another burst of 58 data bits, 3 stop bits, and a guard time corresponding to 8.25 data bits. The transmission rate is 270.833 Kbps.

   (a) Sketch the structure of a GSM frame and a timeslot within the frame.

   (b) Find the fraction of data bits within a timeslot, and the information data rate for each user.

   (c) Find the duration of a frame and the latency between timeslots assigned to a given user in a frame, neglecting the duration of the preamble and trail bits.

   (d) What is the maximum delay spread in the channel such that the guard band and stop bits prevent overlap between timeslots.

3. Consider a DS CDMA system occupying 10 MHz of spectrum. Assume an interference-limited system with a spreading gain of $G = 100$ and code cross correlation of $1/G$.

   (a) For the MAC, find a formula for the SIR of the received signal as a function of G and the number of users K. Assume that all users transmit at the same power and there is perfect power control, so all users have the same received power.

   (b) Based on your SIR formula in part (a), find the maximum number of users K that can be supported in the system, assuming BPSK modulation with a target BER of 10^{-3}. In your BER calculation you can treat interference as AWGN. How does this compare with the maximum number of users K that an FDMA system with the same total bandwidth and information signal bandwidth could support?

   (c) Modify your SIR formula in part (a) to include the effect of voice activity, defined as the percentage of time that users are talking, so interference is multiplied by this percentage. Also find the voice activity factor such that the CDMA system accommodates the same number of users as an FDMA system. Is this a reasonable value for voice activity?

4. Consider a FH CDMA system that uses FSK modulation and the same spreading and information bandwidth as the DS CDMA system in the previous problem. Thus, there are $G = 100$ frequency slots in the system, each of bandwidth 100 KHz. The hopping codes are random and uniformly distributed, so the probability that a given user occupies a given frequency slot on any hop is $.01$. As in the previous problem, noise is essentially negligible, so the probability of error on a particular hop if only one user occupies that hop is zero. Also assume perfect power control, so the received power from all users is the same.

   (a) Find an expression for the probability of bit error when m users occupy the same frequency slot.

   (b) Assume there is a total of K users in the system at any time. What is the probability that on any hop m there is more than one user occupying the same frequency?

   (c) Find an expression for the average probability of bit error as a function of K, the total number of users in the system.

5. Compute the maximum throughput T for a pure ALOHA and a slotted ALOHA random access system, along with the load L that achieves the maximum in each case.

6. Consider a pure ALOHA system with a transmission rate of $R = 10$ Mbps. Compute the load L and throughput T for the system assuming 1000 bit packets and a Poisson arrival rate of $\lambda = 10^3$ packets/sec. Also compute the effective data rate (rate of bits successfully received). What other value of load L results in the exact same throughput?

7. Consider a 3-user uplink channel with channel power gains $g_1 = 1$, $g_2 = 3$, and $g_3 = 5$ from user k to the receiver, $k = 1, 2, 3$. Assume all three users require a 10 dB SINR. The receiver noise is $n = 1$.

   (a) Confirm that the vector equation $(I - F)P \geq u$ given by (14.6) is equivalent to the SINR constraints of each user.

   (b) Determine if a feasible power vector exists for this system such that all users meet the required SINR constraints and, if so, find the optimal power vector P^* such that the desired SINRs are achieved with minimum transmit power.

8. Find the two-user broadcast channel capacity region under superposition coding for transmit power $P = 10$ mW, $B = 100$ KHz, and $N_0 = 10^{-9}$.

9. Show that the sum-rate capacity of the AWGN BC is achieved by sending all power to the user with the highest channel gain.

10. Derive a formula for the optimal power allocation on a fading broadcast channel to maximize sum-rate.

11. Find the sum-rate capacity of a two-user fading BC where the fading on each user's channel is independent. Assume each user has a received power of 10 mW and an effective noise power of 1 mW with probability .5 and 5 mW with probability .5.

12. Find the sum-rate capacity for a two-user broadcast fading channel where each user experiences Rayleigh fading. Assume an average received power of $P = 10$ mW for each user and bandwidth, $B = 100$ KHz, and $N_0 = 10^{-9}$ W/Hz.

13. Consider the set of achievable rates for a broadcast fading channel under frequency-division. Given any rate vector in $\mathcal{C}_{FD}(\mathcal{P}, \mathcal{B})$ for a given power policy ($\mathcal{P}$ and bandwidth allocation policy $\mathcal{B}$, as defined in (14.30), find the timeslot and power allocation policy that achieves the same rate vector.

14. Consider a time-varying broadcast channel with total bandwidth $B = 100$KHz. The effective noise for user 1 has pdf $n_1 = 10^{-5}$ W/Hz with probability 3/4, and the value $n_1 = 2 \times 10^{-5}$ W/Hz with probability 1/4. The effective noise for user 2 takes the value $n_2 = 10^{-5}$ W/Hz with probability 1/2, and the value $n_2 = 2 \times 10^{-5}$ W/Hz with probability 1/2. These noise densities are independent of each other over all time. The total transmit power is is $P = 10$ W.

   (a) What is the set of all possible joint noise densities and their corresponding probabilities?

   (b) Obtain the optimal power allocation between the two users and the corresponding time-varying capacity rate region using time-division. Assume user k is allocated a fixed timeslot τ_k for all time where $\tau_1 + \tau_2 = 1$ and a fixed average power P over all time, but that each user may change its power within its own timeslot, subject to the average constraint P. Find a rate point that exceeds this region assuming you don't divide power equally.

(c) Assume now fixed frequency division, where the bandwidth assigned to each user is fixed and is evenly divided between the two users: $B_1 = B_2 = B/2$. Assume also that you allocate half the power to each user within his respective bandwidth ($P_1 = P_2 = P/2$), and you can vary the power over time, subject only to the average power constraint $P/2$. What is the best rate point that can be achieved? Find a rate point that exceeds this region assuming that you don't share power and/or bandwidth equally.

(d) Is the rate point $(R_1 = 100,000, R_2 = 100,000)$ in the zero-outage capacity region of this channel?

15. Show that the K-user AWGN MAC capacity region is not affected if the kth user's channel power gain g_k is scaled by α if the kth user's transmit power P_k is also scaled by $1/\alpha$.

16. Consider a multiple access channel being shared by two users. The total system bandwidth is $B = 100$KHz. Transmit power of user 1 is $P_1 = 3m$W, while transmit power of user 2 is $P_2 = 1m$W. The receiver noise density is $.001\mu$W/Hz. You can neglect any path loss, fading, or shadowing effects.

    (a) Suppose user 1 requires a data rate of 300 Kbps to see videos. What is the maximum rate that can be assigned to user 2 under time-division? How about under superposition coding with successive interference cancellation?

    (b) Compute the rate pair (R_1, R_2) where the frequency-division rate region intersects the region achieved by code-division with successive interference cancellation ($G = 1$).

    (c) Compute the rate pair (R_1, R_2) such that $R_1 = R_2$ (i.e. where the two users get the same rate) for time division and for spread spectrum code division with and without successive interference cancellation for a spreading gain $G = 10$. *Note: To obtain this region for $G > 1$ you must use the same reasoning on the MAC as was used to obtain the BC capacity region with $G > 1$.*

17. Show that the sum-rate capacity of the AWGN MAC is achieved by having all users transmit at full power.

18. Derive the optimal power adaptation for a two-user fading MAC that achieves the sum-rate point.

19. Find the sum-rate capacity of a two-user fading MAC where the fading on each user's channel is independent. Assume each user has a received power of 10 mW and an effective noise power of 1 mW with probability .5 and 5 mW with probability .5.

20. Consider a 3-user fading downlink with bandwidth 100 KHz. Suppose that the three users all have the same fading statistics, so that their received SNR when they are allocated the full power and bandwidth are 5 dB with probability 1/3, 10 dB with probability 1/3, and 20 dB with probability 1/3. Assume a discrete time system with fading i.i.d. at each time slot.

    (a) Find the maximum throughput of this system if at each time instant the full power and bandwidth are allocated to the user with the best channel.

    (b) Simulate the throughput obtained using the proportional fair scheduling algorithm for a window size of 1, 5, and 10.

Appendix A

Representation of Bandpass Signals and Channels

Many signals in communication systems are real bandpass signals with a frequency response that occupies a narrow bandwidth $2B$ centered around a carrier frequency f_c with $2B << f_c$, as illustrated in Figure A.1. Since bandpass signals are real, their frequency response has conjugate symmetry, i.e. a bandpass signal $s(t)$ has $|S(f)| = |S(-f)|$ and $\angle S(f) = -\angle S(-f)$. However, bandpass signals are not necessarily conjugate symmetric within the signal bandwidth about the carrier frequency f_c, i.e. we may have $|S(f_c+f)| \neq |S(f_c-f)|$ or $\angle S(f_c+f) \neq -\angle S(f_c-f)$ for some $f \leq B$. This asymmetry in $|S(f)|$ is illustrated in Figure A.1. Bandpass signals result from modulation of a baseband signal by a carrier, or from filtering a deterministic or random signal with a bandpass filter. The bandwidth $2B$ of a bandpass signal is roughly equal to the range of frequencies around f_c where the signal has nonnegligible amplitude. Bandpass signals are commonly used to model transmitted and received signals in communication systems. These are real signals since the transmitter circuitry can only generate real sinusoids (not complex exponentials) and the channel just introduces an amplitude and phase change at each frequency of the real transmitted signal.

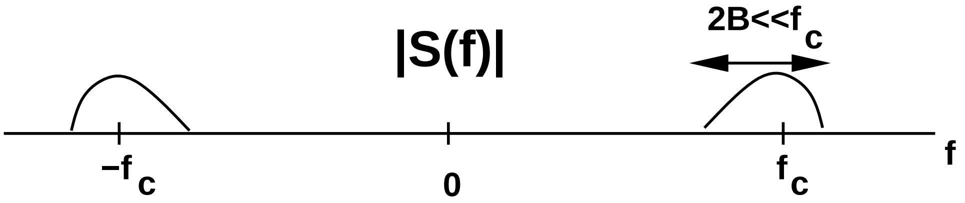


Figure A.1: Bandpass Signal $S(f)$.

We begin by representing a bandpass signal $s(t)$ at carrier frequency f_c in the following form:

$$s(t) = s_I(t)\cos(2\pi f_c t) - s_Q(t)\sin(2\pi f_c t), \tag{A.1}$$

where $s_I(t)$ and $s_Q(t)$ are real lowpass (baseband) signals of bandwidth $B << f_c$. This is a common representation for bandpass signals or noise. In fact, modulations such as MPSK and MQAM are commonly described using this representation. We call $s_I(t)$ the **in-phase component** of $s(t)$ and $s_Q(t)$ the **quadrature component** of $s(t)$. Define the complex signal $u(t) = s_I(t) + js_Q(t)$, so $s_I(t) = \Re\{u(t)\}$ and $s_Q(t) = \Im\{u(t)\}$. Then $u(t)$ is a complex lowpass signal of bandwidth B. With this definition we see that

$$s(t) = \Re\{u(t)\}\cos(2\pi f_c t) - \Im\{u(t)\}\sin(2\pi f_c t) = \Re\left\{u(t)e^{j2\pi f_c t}\right\}. \tag{A.2}$$

The representation on the right hand side of this equation is called the **complex lowpass representation** of the bandpass signal $s(t)$, and the baseband signal $u(t)$ is called the **equivalent lowpass signal** for $s(t)$ or its **complex envelope**. Note that $U(f)$ is only conjugate symmetric about $f = 0$ if $u(t)$ is real, i.e. if $s_Q(t) = 0$.

Using properties of the Fourier transform we can show that

$$S(f) = .5[U(f - f_c) + U^*(-f - f_c)]. \tag{A.3}$$

Since $s(t)$ is real, $S(f)$ is symmetric about $f = 0$. However, the lowpass signals $U(f)$ and $U^*(f)$ are not necessarily symmetric about $f = 0$, which leads to an asymmetry of $S(f)$ within the bandwidth $2B$ about the carrier frequency f_c, as shown in Figure A.1. In fact, $S(f)$ is only symmetric about the carrier frequency within this bandwidth if $u(t) = s_I(t)$, i.e. if there is no quadrature component in $u(t)$. We will see shortly that this asymmetry affects the response of bandpass channels to bandpass signals.

An alternate representation of the equivalent lowpass signal is

$$u(t) = a(t)e^{j\phi(t)}, \tag{A.4}$$

with envelope

$$a(t) = \sqrt{s_I^2(t) + s_Q^2(t)}, \tag{A.5}$$

and phase

$$\phi(t) = \tan^{-1}\left(\frac{s_Q(t)}{s_I(t)}\right). \tag{A.6}$$

With this representation

$$s(t) = \Re\left\{a(t)e^{j\phi(t)}e^{j2\pi f_c t}\right\} = a(t)\cos(2\pi f_c t + \phi(t)). \tag{A.7}$$

Let us now consider a real channel impulse response $h(t)$ with Fourier transform $H(f)$. If $h(t)$ is real then $H^*(-f) = H(f)$. In communication systems we are mainly interested in the channel frequency response $H(f)$ for $|f - f_c| < B$, since only these frequency components of $H(f)$ affect the received signal within the bandwidth of interest. A **bandpass channel** is similar to a bandpass signal: it has a real impulse response $h(t)$ with frequency response $H(f)$ centered at a carrier frequency f_c with a bandwidth of $2B << f_c$. To capture the frequency response of $H(f)$ around f_c, we develop an **equivalent lowpass channel** model similar to the equivalent lowpass signal model as follows. Since the impulse response $h(t)$ corresponding to $H(f)$ is a bandpass signal, it can be written using an equivalent lowpass representation as

$$h(t) = 2\Re\left\{h_l(t)e^{j2\pi f_c t}\right\}, \tag{A.8}$$

where the extra factor of 2 is to avoid constant factors in the $H(f)$ representation given by (A.9). We call $h_l(t)$ the **equivalent lowpass channel impulse response** for $H(f)$. From (A.2)-(A.3), the representation (A.8) implies that

$$H(f) = H_l(f - f_c) + H_l^*(-f - f_c), \tag{A.9}$$

so $H(f)$ consists of two components: $H_l(f)$ shifted up by f_c, and $H_l^*(f)$ shifted down by f_c. Note that if $H(f)$ is conjugate symmetric about the carrier frequency f_c within the bandwidth $2B$ then $h_l(t)$ will be real and its frequency response $H_l(f)$ conjugate symmetric about zero. However, in many wireless channels, such as frequency-selective fading channels, $H(f)$ is not conjugate symmetric about f_c, in which case $h_l(t)$ is complex with in-phase component $h_{l,I}(t) = \Re\{h_l(t)\}$ and quadrature component $h_{l,Q}(t) = \Im\{h_l(t)\}$. Note that if $h_l(t)$ is complex then $H_l(f)$ is not conjugate symmetric about zero.

We now use equivalent lowpass signal and channel models to study the output of a bandpass channel with a bandpass signal input. Let $s(t)$ denote the input signal with equivalent lowpass signal $u(t)$. Let $h(t)$ denote the bandpass channel impulse response with equivalent lowpass channel impulse response $h_l(t)$. The transmitted signal $s(t)$ and channel impulse response $h(t)$ are both real, so the channel output $r(t) = s(t) * h(t)$ is also real, with frequency response $R(f) = H(f)S(f)$. Since $S(f)$ is a bandpass signal, $R(f)$ will also be a bandpass signal. Therefore, it has a complex lowpass representation of

$$r(t) = \Re\left\{v(t)e^{j2\pi f_c t}\right\}. \tag{A.10}$$

We now consider the relationship between the equivalent lowpass signals corresponding to the channel input $s(t)$, channel impulse response $h(t)$, and channel output $r(t)$. We can express the frequency response of the channel output as

$$R(f) = H(f)S(f) = .5[H_l(f-f_c) + H_l^*(-f-f_c)][U(f-f_c) + U^*(-f-f_c)]. \tag{A.11}$$

For bandpass signals and channels where the bandwidth B of $u(t)$ and $h_l(t)$ is much less than the carrier frequency f_c, we have

$$H_l(f-f_c)U^*(-f-f_c) = 0$$

and

$$H_l^*(-f-f_c)U(f-f_c) = 0.$$

Thus,

$$R(f) = .5[H_l(f-f_c)U(f-f_c) + H_l^*(-f-f_c)U^*(-f-f_c)]. \tag{A.12}$$

From (A.2)-(A.3), (A.10) implies that

$$R(f) = .5[V(f-f_c) + V^*(-f-f_c)]. \tag{A.13}$$

Equating terms at positive and negative frequencies in (A.12) and (A.13), we get that

$$V(f-f_c) = H_l(f-f_c)U(f-f_c) \tag{A.14}$$

and

$$V^*(-f-f_c) = H_l^*(-f-f_c)U^*(-f-f_c) \tag{A.15}$$

or, equivalently, that

$$V(f) = H_l(f)U(f). \tag{A.16}$$

Taking the inverse Fourier transform yields that

$$v(t) = u(t) * h_l(t). \tag{A.17}$$

Thus, we can obtain the equivalent lowpass signal $v(t)$ for the received signal $r(t)$ by taking the convolution of $h_l(t)$ and $u(t)$. The received signal is therefore given by

$$r(t) = \Re\left\{(u(t) * h_l(t))e^{j2\pi f_c t}\right\}. \tag{A.18}$$

Note that $V(f) = H_l(f)U(f)$ is conjugate symmetric about $f = 0$ only if both $U(f)$ and $H_l(f)$ are. In other words, the equivalent lowpass received signal will be complex with nonzero in-phase and quadrature components if either $u(t)$ or $h_l(t)$ is complex. Moreover, if $u(t) = s_I(t)$ is real (no quadrature component) but the channel impulse response $h_l(t) = h_{l,I}(t) + jh_{l,Q}(t)$ is complex (e.g. as with frequency-selective fading) then

$$v(t) = s_I(t) * (h_{l,I}(t) + jh_{l,Q}(t)) = s_I(t) * h_{l,I}(t) + js_I(t) * h_{l,Q}(t) \tag{A.19}$$

is complex, so the received signal will have both an in-phase and a quadrature component. More generally, if $u(t) = s_I(t) + js_Q(t)$ and $h_l(t) = h_{l,I}(t) + jh_{l,Q}(t)$ then

$$v(t) = [s_I(t)+js_Q(t)]*[h_{l,I}(t)+jh_{l,Q}(t)] = [s_I(t)*h_{l,I}(t)-s_Q(t)*h_{l,Q}(t)]+j[s_I(t)*h_{l,Q}(t)+s_Q(t)*h_{l,I}(t)]. \tag{A.20}$$

So the in-phase component of $v(t)$ depends on *both* the in-phase and quadrature components of $u(t)$, and similarly for the quadrature component of $v(t)$. This creates problems in signal detection, since it causes the in-phase and quadrature parts of a modulated signal to interfere with each other in the demodulator.

The main purpose for the equivalent lowpass representations is to analyze bandpass communication systems using the equivalent lowpass models for the transmitted signal, channel impulse response, and received signal. This removes the carrier terms from the analysis, in particular the dependency of the analysis on the carrier frequency f_c.

Appendix B

Probability Theory, Random Variables, and Random Processes

B.1 Probability Theory

Probability theory provides a mathematical characterization for random events. Such events are defined on an underlying probability space $(\Omega, \mathcal{E}, p(\cdot))$. The probability space consists of a sample space Ω of possible outcomes for random events, a set of random events $\mathcal{E}$ where each $A \in \mathcal{E}$ is a subset of Ω, and a probability measure $p(\cdot)$ defined on these subsets. Thus, $\mathcal{E}$ is a set of sets, and the probability measure $p(A)$ is defined for every set $A \in \mathcal{E}$. A probability space requires that the set $\mathcal{E}$ is a σ-field. Intuitively, a set of sets $\mathcal{E}$ is a σ-field if it contains all intersections, unions, and complements of its elements.[1] More precisely, $\mathcal{E}$ is a σ-field if the set of all possible outcomes Ω is one of the sets in $\mathcal{E}$, a set $A \in \mathcal{E}$ implies that $A^c \in \mathcal{E}$, and for any sets $A_1, A_2, \ldots$ with $A_i \in \mathcal{E}$, we have $\cup_{i=1}^{\infty} A_i \in \mathcal{E}$. The set $\mathcal{E}$ must be a σ-field for the probability of intersections and unions of random events to be defined. We also require that the probability measure associated with a probability space have the following three fundamental properties:

1. $p(\Omega) = 1$.
2. $0 \leq p(A) \leq 1$ for any event $A \in \mathcal{E}$.
3. If A and B are mutually exclusive, i.e. their intersection is zero, then $p(A \cup B) = p(A) + p(B)$.

Throughout this section, we only consider sets in $\mathcal{E}$, since the probability measure is only defined on these sets.

Several important characteristics of the probability measure $p(\cdot)$ can be derived from its fundamental properties. In particular, $p(A^c) = 1 - p(A)$. Moreover, consider sets $A_1, \ldots, A_n$, where A_i and A_j, $i \neq j$, are disjoint ($A_i \cap A_j = \emptyset$). Then if $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$, we have that $\sum_{i=1}^{n} p(A_i) = 1$. We call the set $\{A_1, \ldots, A_n\}$ with these properties a *partition* of Ω. For two sets A_i and A_j that are not disjoint, $p(A_i \cup A_j) = p(A_i) + p(A_j) - p(A_i \cap A_j)$. This leads to the *union bound*, which says that for any sets $A_1, \ldots, A_n$,

$$p(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} p(A_i). \tag{B.1}$$

[1] We use the notation $A \cap B$ to denote the intersection of A and B, i.e. all elements in both A and B. The union of A and B, denoted $A \cup B$ is the set of all elements in A or B. The complement of a set $A \subset \Omega$, denoted by A^c, is defined as all elements in Ω that are not in the set A.

The occurence of one random event can affect the probability of another random event, since observing one random event indicates which subsets in $\mathcal{E}$ could have contributed to the observed outcome. To capture this effect, we define the probability of event B conditioned on the occurence of event A as $p(B|A) = p(A \cap B)/p(A)$, assuming $p(A) \neq 0$. This implies that

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A). \tag{B.2}$$

The conditional probability $p(B|A) = p(A \cap B)/p(A)$ essentially normalizes the probability of B with respect to the outcomes associated with A, since it is known that A has occured. We obtain *Bayes Rule* from (B.2) as

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \tag{B.3}$$

Independence of events is a function of the probability measure $p(\cdot)$. In particular, events A and B are independent if $p(A \cap B) = p(A)p(B)$. This implies that $p(B|A) = p(B)$ and $p(A|B) = p(A)$.

B.2 Random Variables

Random variables are defined on an underlying probability space $(\Omega, \mathcal{E}, p(\cdot))$. In particular, a random variable X is a function mapping from the sample space Ω to a subset of the real line. If X takes discrete values on the real line it is called a discrete random variable, and if it takes continuous values it is called a continuous random variable. The *cumulative distribution function* (CDF) of a random variable X is defined as $P_X(x) \stackrel{\triangle}{=} p(X \leq x)$ for some $x \in \mathcal{R}$. The CDF is derived from the underlying probability space as $p(X \leq x) = p(X^{-1}(-\infty, x))$, where $X^{-1}(\cdot)$ is the inverse mapping from the real line to a subset of Ω: $X^{-1}(-\infty, x) = \{\omega \in \Omega : X(\omega) \leq x\}$. Properties of the CDF are based on properties of the underlying probability measure. In particular, the CDF satisfies $0 \leq P_X(x) = p(X^{-1}(-\infty, x)) \leq 1$. In addition, the CDF is nondecreasing: $P_X(x_1) \leq P_X(x_2)$ for $x_1 \leq x_2$. That is because $P_X(x_2) = p(X^{-1}(-\infty, x_2)) = p(X^{-1}(-\infty, x_1)) + p(X^{-1}(x_1, x_2)) \geq p(X^{-1}(-\infty, x_1)) = P_X(x_1)$.

The *probability density function* (pdf) of a random variable X is defined as the derivative of its CDF, $p_X(x) \stackrel{\triangle}{=} \frac{d}{dx}P_X(x)$. For X a continuous random variable $p_X(x)$ is a function over the entire real line. For X a discrete random variable $p_X(x)$ is a set of delta functions at the possible values of X. The pdf, also refered to as the *probability distribution* or *distribution* of X, defines the probability that X lies in a given range of values:

$$p(x_1 < X \leq x_2) = p(X \leq x_2) - p(X \leq x_1) = P_X(x_2) - P_x(x_1) = \int_{x_1}^{x_2} p_X(x)dx. \tag{B.4}$$

Since $P_X(\infty) = 1$ and $P_X(-\infty) = 0$, the pdf integrates to 1,

$$\int_{-\infty}^{\infty} p_X(x)dx = 1. \tag{B.5}$$

Note that the subscript X is often omitted from the pdf and CDF when it is clear from the context that these functions characterize the distribution of X. In this case the pdf is written as $p(x)$ and the CDF as $P(x)$.

The *mean* or *expected value* of a random variable X is its probabalistic average, defined as

$$\mu_X = \mathbf{E}[X] \stackrel{\triangle}{=} \int_{-\infty}^{\infty} x p_X(x)dx. \tag{B.6}$$

The expectation operator $\mathbf{E}[\cdot]$ is linear and can also be applied to functions of random variables. In particular, the mean of a function of X is given by

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x)p_X(x)dx. \tag{B.7}$$

A function of particular interest is the *nth moment* of X,

$$\mathbf{E}[X^n] = \int_{-\infty}^{\infty} x^n p_X(x) dx. \tag{B.8}$$

The variance of X is defined in terms of its mean and second moment as

$$\text{Var}[X] = \sigma_X^2 \overset{\triangle}{=} \mathbf{E}[(X - \mu_X)^2] = \mathbf{E}[X^2] - \mu_X^2. \tag{B.9}$$

The variance characterizes the average squared difference between X and its mean μ_X. The standard deviation of X, σ_X, is the square root of its variance. From the linearity of the expectation operator, it is easily shown that for any constant c, $\mathbf{E}[cX] = c\mathbf{E}[X]$, $\text{Var}[cX] = c^2\text{Var}[X]$, $\mathbf{E}[X + c] = \mathbf{E}[X] + c$, and $\text{Var}[X + c] = \text{Var}[X]$. Thus, scaling a random variable by a constant scales its mean by the same constant and its variance by the constant squared. Adding a constant to a random variable shifts the mean by the same constant but doesn't affect the variance.

The distribution of a random variable X can be determined from its *characteristic function*, defined as

$$\phi_X(\nu) \overset{\triangle}{=} \mathbf{E}[e^{j\nu X}] = \int_{-\infty}^{\infty} p_X(x) e^{j\nu x} dx. \tag{B.10}$$

We see from (B.10) that the characteristic function $\phi_X(\nu)$ of $X(t)$ is the inverse Fourier transform of the distribution $p_X(x)$ evaluated at $f = \nu/(2\pi)$. Thus we can obtain $p_X(x)$ from $\phi_X(\nu)$ as

$$p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(\nu) e^{-j\nu x} dx. \tag{B.11}$$

This will become significant in finding the distribution for sums of random variables.

Let X be a random variable and $g(x)$ be a function on the real line. Let $Y = g(X)$ define another random variable. Then $P_Y(y) = \int_{x:g(x)\leq y} p_X(x)dx$. For g monotonically increasing and one-to-one this becomes $P_Y(y) = \int_{-\infty}^{g^{-1}(y)} p_X(x)dx$. For g monotonically decreasing and one-to-one this becomes $P_Y(y) = \int_{g^{-1}(y)}^{\infty} p_X(x)dx$.

We now consider joint random variables. Two random variables must share the same underlying probability space for their joint distribution to be defined. Let X and Y be two random variables defined on the same probability space $(\Omega, \mathcal{E}, p(\cdot))$. Their joint CDF is defined as $P_{XY}(x, y) \overset{\triangle}{=} p(X \leq x, Y \leq y)$. Their joint pdf (distribution) is defined as the derivative of the joint CDF:

$$p_{XY}(x, y) \overset{\triangle}{=} \frac{\partial^2 P_{XY}(x, y)}{\partial x \partial y}. \tag{B.12}$$

Thus,

$$P_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p_{XY}(v, w) dv dw. \tag{B.13}$$

For joint random variables X and Y, we can obtain the distribution of X by integrating the joint distribution with respect to Y:

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y) dy. \tag{B.14}$$

Similarly,

$$p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x, y) dx. \tag{B.15}$$

The distributions $p_X(x)$ and $p_Y(y)$ obtained in this manner are sometimes refered to as the *marginal* distributions relative to the joint distribution $p_{XY}(x,y)$. Note that the joint distribution must integrate to one:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{XY}(x,y)dxdy = 1. \tag{B.16}$$

The definitions for joint CDF and joint pdf of two random variables extend in a straightforward manner to any finite number of random variables.

As with random events, observing the value for one random variable can affect the probability of another random variable. We define the conditional distribution of the random variable Y given a realization $X = x$ of random variable X as $p_Y(y|X = x) = p_{XY}(x,y)/p_X(x)$. This implies that $p_{XY}(x,y) = p_Y(y|X = x)p_X(x)$. Independence between two random variables X and Y is a function of their joint distribution. Specifically, X and Y are independent random variables if their joint distribution $p_{XY}(x,y)$ factors into separate distributions for X and Y: $p_{XY}(x,y) = p_X(x)p_Y(y)$. For independent random variables, it is easily shown that for any functions $f(x)$ and $g(y)$, $\mathbf{E}[f(X)g(Y)] = \mathbf{E}[f(X)]\mathbf{E}[g(Y)]$.

For X and Y joint random variables with joint pdf $p_{XY}(x,y)$, we define their ijth *joint moment* as

$$\mathbf{E}[X^iY^j] \stackrel{\triangle}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^iy^jp_{XY}(x,y)dxdy. \tag{B.17}$$

The *correlation* of X and Y is defined as $\mathbf{E}[XY]$. The *covariance* of X and Y is defined as Cov$[XY] \stackrel{\triangle}{=} \mathbf{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbf{E}[XY] - \mu_X\mu_Y$. Note that the covariance and correlation of X and Y are equal if either X or Y has mean zero. The *correlation coefficient* of X and Y is defined in terms of their covariance and standard deviations as $\rho \stackrel{\triangle}{=}$ Cov$[XY]/(\sigma_X\sigma_Y)$. We say that X and Y are *uncorrelated* if their covariance is zero or, equivalently, their correlation coefficient is zero. Note that uncorrelated random variables (i.e. X and Y with Cov$[XY] = \mathbf{E}[XY] - \mu_X\mu_Y = 0$) will have a nonzero correlation ($\mathbf{E}[XY] \neq 0$) if their means are not zero. For random variables $X_1, \ldots, X_n$, we define their *covariance matrix* $\mathbf{\Sigma}$ as an $n \times n$ matrix with ijth element $\mathbf{\Sigma}_{ij} =$ Cov$[X_iX_j]$. In particular, the ith diagonal element of $\mathbf{\Sigma}$ is the variance of X_i: $\mathbf{\Sigma}_{ii} =$ Var$[X_i]$.

Consider two independent random variables X and Y. Let $Z = X + Y$ define a new random variable on the probability space $(\Omega, \mathcal{E}, p(\cdot))$. We can show directly or by using characteristic functions that the distribution of Z is the convolution of the distributions of X and Y: $p_Z(z) = p_X(x) * p_Y(y)$. Equivalently, $\phi_Z(\nu) = \phi_X(\nu)\phi_Y(\nu)$. With this distribution it can be shown that $\mathbf{E}[Z] = \mathbf{E}[X] + \mathbf{E}[Y]$, and Var$[Z] =$ Var$[X] +$ Var$[Y]$. So for sums of independent random variables, the mean of the sum is the sum of the means and the variance of the sum is the sum of the variances.

A distribution that arises frequently in the study of communication systems is the Gaussian distribution. The Gaussian distribution for a random variable X is defined in terms of its mean μ_X and variance σ_X^2 as

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X}e^{-[(x-\mu_X)^2/(2\sigma_X^2)]}. \tag{B.18}$$

The Gaussian distribution, also called the normal distribution, is denoted as $\mathcal{N}(\mu_X, \sigma_X^2)$. Note that the tail of the distribution, i.e. the value of $p_X(x)$ as x moves away from μ_X, decreases exponentially. The CDF $P_X(x) = p(X \leq x)$ for this distribution does not exist in closed form. It is defined in terms of the Gaussian Q function as

$$P_X(x) = p(X \leq x) = 1 - Q\left(\frac{x - \mu_X}{\sigma_X}\right), \tag{B.19}$$

where the Gaussian Q function, defined by

$$Q(x) \stackrel{\triangle}{=} \int_x^{\infty} \frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy, \tag{B.20}$$

is the probability that a Gaussian random variable X with mean zero and variance one is bigger than x: $Q(x) = p(X \geq x)$ for $X \sim \mathcal{N}(0,1)$. The Gaussian Q function is related to the complementary error function as $Q(x) = .5\text{erfc}(x/\sqrt{2})$. These functions are typically calculated using standard computer math packages.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote a vector of jointly Gaussian random variables. Their joint distribution is given by

$$p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det[\mathbf{\Sigma}]}} \exp\left[-.5(\mathbf{x} - \mu_{\mathbf{X}})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu_{\mathbf{X}})\right], \tag{B.21}$$

where $\mu_{\mathbf{X}} = \mathbf{E}[\mathbf{X}]^T = (\mathbf{E}[X_1], \ldots, \mathbf{E}[X_n])^T$ is the mean of $\mathbf{X}$ and $\mathbf{\Sigma}$ is the $n \times n$ covariance matrix of $\mathbf{X}$, i.e. $\mathbf{\Sigma}_{ij} = \text{Cov}[X_i X_j]$. It can be shown from (B.21) that for jointly Gaussian random variables X and Y, if $\text{Cov}[XY] = 0$ then $p_{XY}(x,y) = p_X(x)p_Y(y)$. In other words, Gaussian random variables that are uncorrelated are independent.

The underlying reason why the Gaussian distribution commonly occurs in communication system models is the Central Limit Theorem (CLT), which defines the limiting distribution for the sum of a large number of independent random variables with the same distribution. Specifically, let X_i be independent and identically distributed (i.i.d.) joint random variables. Let $Y_n = \sum_{i=1}^n X_i$ and $Z_n = (Y_n - \mu_{Y_n})/\sigma_{Y_n}$. The CLT states that the distribution of Z_n as n goes to infinity converges to a Gaussian distribution with mean zero and variance one: $\lim_{n\to\infty} p_{Z_n}(x) = \mathcal{N}(0,1)$. Thus, any random variable equal to the sum of a large number of i.i.d. random components has a distribution that is approximately Gaussian. For example, noise in a radio receiver typically consists of spurious signals generated by the various hardware components, and with a large number of i.i.d. components this noise is accurately modeled as Gauss-distributed.

Two other common distributions that arise in communication systems are the uniform distribution and the binomial distribution. A random variable X that is uniformly distributed has pdf $p_X(x) = 1/(b-a)$ for x in the interval $[a,b]$ and zero otherwise. A random phase θ is commonly modeled as uniformly-distributed on the interval $[0, 2\pi]$, which we denote as $\theta \sim \mathcal{U}[0, 2\pi]$. The binomial distribution often arises in coding analysis. Let X_i, $i = 1, \ldots, n$, be discrete random variables that take two possible values, 0 and 1. Suppose the X_i are i.i.d. with $p(X_i = 1) = p$ and $p(X_i = 0) = 1 - p$. Let $Y = \sum_{i=1}^n X_i$. Then Y is a discrete random variable that takes integer values $k = 0, 1, 2, \ldots$. The distribution of Y is the binomial distribution, given by

$$p(Y = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \tag{B.22}$$

where

$$\left(\begin{array}{c} n \\ k \end{array} \right) \stackrel{\triangle}{=} \frac{n!}{k!(n-k)!}. \tag{B.23}$$

B.3 Random Processes

A random process $X(t)$ is defined on an underlying probability space $(\Omega, \mathcal{E}, p(\cdot))$. In particular, it is a function mapping from the sample space Ω to a set of real functions $\{x_1(t), x_2(t), \ldots\}$, where each $x_i(t)$ is a possible realization of $X(t)$. Samples of $X(t)$ at times $t_0, t_1, \ldots, t_n$ are joint random variables defined on the underlying probability space. Thus, the joint CDF of samples at times $t_0, t_1, \ldots, t_n$ is given by $P_{X(t_0)X(t_1)\ldots X(t_n)}(x_0, \ldots, x_n) = p(X(t_0) \leq x_0, X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n)$. The random process $X(t)$ is fully characterized by its joint CDF $P_{X(t_0)X(t_1)\ldots X(t_n)}(x_0, \ldots, x_n)$ for all possible sets of sample times $\{t_0, t_1, \ldots, t_n\}$.

A random process $X(t)$ is stationary if for all T and all sets of sample times $\{t_0, \ldots, t_n\}$, we have that $p(X(t_0) \leq x_0, X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n) = p(X(t_0 + T) \leq x_0, X(t_1 + T) \leq x_1, \ldots, X(t_n + T) \leq x_n)$. Intuitively, a random process is stationary if time shifts do not affect its probability. Stationarity of a process is often difficult to prove since it requires checking the joint CDF of all possible sets of samples for all possible time shifts. Stationarity of a random process is often infered from the stationarity of the source generating the process.

The *mean* of a random process is defined as $E[X(t)]$. Since the mean of a stationary random process is independent of time shifts, it must be constant: $\mathbf{E}[X(t)] = \mathbf{E}[X(t-t)] = \mathbf{E}[X(0)] = \mu_X$. The autocorrelation of a random process is defined as $A_X(t, t+\tau) \triangleq \mathbf{E}[X(t)X(t+\tau)]$. The autocorrelation of $X(t)$ is also called its second moment. Since the autocorrelation of a stationary process is independent of time shifts, $A_X(t, t+\tau) = \mathbf{E}[X(t-t)X(t+\tau-t)] = \mathbf{E}[X(0)X(\tau)] \triangleq A_X(\tau)$. So for stationary processes, the autocorrelation depends only on the time difference τ between the samples $X(t)$ and $X(t+\tau)$ and not on the absolute time t. The autocorrelation of a process measures the correlation between samples of the process taken at different times.

Two random processes $X(t)$ and $Y(t)$ defined on the same underlying probability space have a joint CDF characterized by

$$\begin{aligned} & P_{X(t_0)X(t_1)\ldots X(t_n)Y(t'_0)\ldots Y(t'_m)}(x_0, \ldots, x_n, y_0, \ldots, y_m) \\ = \;& p(X(t_0) \leq x_0, \ldots, X(t_n) \leq x_n, Y(t'_0) \leq y_0, \ldots, Y(t'_m) \leq y_m) \end{aligned} \tag{B.24}$$

for all possible sets of sample times $\{t_0, t_1, \ldots, t_n\}$ and $\{t'_0, t_1, \ldots, t'_m\}$. Two random processes $X(t)$ and $Y(t)$ are independent if for all such sets we have that

$$\begin{aligned} & p_{X(t_0)X(t_1)\ldots X(t_n)Y(t'_0)\ldots Y(t'_m)}(X(t_0) \leq x_0, \ldots, X(t_n) \leq x_n, Y(t'_0) \leq y_0, \ldots, Y(t'_m) \leq y_m) \\ = \;& p_{X(t_0)X(t_1)\ldots X(t_n)}(X(t_0) \leq x_0, \ldots, X(t_n) \leq x_n) p_{Y(t'_0)\ldots Y(t'_m)}(Y(t'_0) \leq y_0, \ldots, Y(t'_m) \leq y_m) \end{aligned} \tag{B.25}$$

The cross-correlation between two random processes $X(t)$ and $Y(t)$ is defined as $A_{XY}(t, t+\tau) \triangleq \mathbf{E}[X(t)Y(t+\tau)]$. The two processes are uncorrelated if $\mathbf{E}[X(t)Y(t+\tau)] = \mathbf{E}[X(t)]\mathbf{E}[Y(t+\tau)]$ for all t and τ. As with the autocorrelation, if both $X(t)$ and $Y(t)$ are stationary, the cross-correlation is only a function of τ: $A_{XY}(t, t+\tau) = \mathbf{E}[X(t-t)Y(t+\tau-t)] = \mathbf{E}[X(0)Y(\tau)] \triangleq A_{XY}(\tau)$.

In most analysis of random processes we focus only on the first and second moments. *Wide-sense stationarity* is a notion of stationarity that only depends on the first two moments of a process, and it can also be easily verified. Specifically, a process is wide-sense stationary (WSS) if its mean is constant, $\mathbf{E}[X(t)] = \mu_X$, and its autocorrelation depends only on the time difference of the samples, $A_X(t, t+\tau) = \mathbf{E}[X(t)X(t+\tau)] = A_X(\tau)$. Stationary processes are WSS but in general WSS processes are not necessarily stationary. For WSS processes, the autocorrelation is a symmetric function of τ, since $A_X(\tau) = \mathbf{E}[X(t)X(t+\tau)] = \mathbf{E}[X(t+\tau)X(t)] = A_X(-\tau)$. Moreover, it can be shown that $A_X(\tau)$ takes its maximum value at $\tau = 0$, i.e. $|A_X(\tau)| \leq A_X(0) = \mathbf{E}[X^2(t)]$. As with stationary processes, if two processes $X(t)$ and $Y(t)$ are both WSS then their cross-correlation is independent of time shifts, and thus depends only on the time difference of the processes: $A_{XY}(t, t+\tau) = \mathbf{E}[X(0)Y(\tau)] = A_{XY}(\tau)$.

The power spectral density (PSD) of a WSS process is defined as the Fourier transform of its autocorrelation function with respect to τ:

$$S_X(f) = \int_{-\infty}^{\infty} A_X(\tau) e^{-j2\pi f\tau} d\tau. \tag{B.26}$$

The autocorrelation can be obtained from the PSD through the inverse transform:

$$A_X(\tau) = \int_{-\infty}^{\infty} S_X(f) e^{j2\pi f\tau} df. \tag{B.27}$$

The PSD takes its name from the fact that the expected power of a random process $X(t)$ is the integral of its PSD:

$$\mathbf{E}[X^2(t)] = A_X(0) = \int_{-\infty}^{\infty} S_X(f) df, \tag{B.28}$$

which follows from (B.27). Similarly, from (B.26) we get that $S_X(0) = \int_{-\infty}^{\infty} A_X(\tau)d\tau$. The symmetry of $A_X(\tau)$ can be used with (B.26) to show that $S_X(f)$ is also symmmetric, i.e. $S_X(f) = S_X(-f)$. *White noise* is defined as a zero mean WSS random process with a PSD that is constant over all frequencies. Thus, a white noise process $X(t)$ has $\mathbf{E}[X(t)] = 0$ and $S_X(f) = N_0/2$ for some constant N_0 which is typically refered to as the (one-sided) white noise PSD. By the inverse Fourier transform, the autocorrelation of white noise is given by $A_X(\tau) = (N_0/2)\delta(\tau)$. In some sense, white noise is the most random of all possible noise processes, since it decorrelates instantaneously.

Random processes are often filtered or modulated, and when the process is WSS the impact of these operations can be characterized in a simple way. In particular, if a WSS process with PSD $S_X(f)$ is passed through a linear time-invariant filter with frequency response $H(f)$, then the filter output is also a WSS process with power spectral density $|H(f)|^2 S_X(f)$. If a WSS process $X(t)$ with PSD $S_X(f)$ is multiplied by a carrier $\cos(2\pi f_c t+\theta)$ with $\theta \sim \mathcal{U}[0, 2\pi]$, the multiplication results in a WSS process $X(t)\cos(2\pi f_c t+\theta)$ with PSD $.25[S_X(f-f_c)+S_X(f+f_c)]$.

Stationarity and WSS are properties of the underlying probability space associated with a random process. We are also often interested in time-averages associated with random processes, which can be characterized by different notions of *ergodicity*. A random process $X(t)$ is *ergodic in the mean* if its time-averaged mean, defined as

$$\mu_X^{ta} = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X(t)dt, \tag{B.29}$$

is constant for all possible realizations of $X(t)$. In other words, $X(t)$ is ergodic in the mean if $\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} x_i(t)dt$ equals the same constant μ_X^{ta} for all possible realizations $x_i(t)$ of $X(t)$. Similarly, a random process $X(t)$ is *ergodic in the nth moment* if its time-averaged nth moment

$$\mu_{X^n}^{ta} = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X^n(t)dt \tag{B.30}$$

is constant for all possible realizations of $X(t)$. We can also define ergodicity of $X(t)$ relative to its time-averaged autocorrelation

$$A_X^{ta}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X(t)X(t+\tau)dt. \tag{B.31}$$

Specifically, $X(t)$ is *ergodic in autocorrelation* if $\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} x_i(t)x_i(t+\tau)dt$ equals the same value $A_X^{ta}(\tau)$ for all possible realizations $x_i(t)$ of $X(t)$. Ergodicity of the autocorrelation in higher order moments requires that the nmth order time-averaged autocorrelation

$$A_X^{ta}(n,m,\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X^n(t)X^m(t+\tau)dt \tag{B.32}$$

is constant for all realizations of $X(t)$. A process that is ergodic in all order moments and autocorrelations is called *ergodic*. Ergodicity of a process requires that its time-averaged nth moment and ijth autocorrelation, averaged over all time, be constant for all n, i, and j. This implies that the probability associated with an ergodic process is independent of time shifts, and thus the process is stationary. In other words, an ergodic process must be stationary. However, a stationary process can be either ergodic or nonergodic. Since an ergodic process is stationary,

$$\begin{aligned} \mu_X^{ta} &= \mathbf{E}[\mu_X^{ta}] \\ &= \mathbf{E}\left[\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X(t)dt\right] \\ &= \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \mathbf{E}[X(t)]dt \\ &= \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} \mu_X dt = \mu_X. \end{aligned} \tag{B.33}$$

Thus, the time-averaged mean of $X(t)$ equals its probabilistic mean. Similarly,

$$\begin{aligned} A_X^{ta}(\tau) &= \mathbf{E}[A_X^{ta}(\tau)] \\ &= \mathbf{E}\left[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} X(t)X(t+\tau)dt\right] \\ &= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\mathbf{E}[X(t)(t+\tau)]dt \\ &= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} A_X(\tau)dt = A_X(\tau), \end{aligned} \tag{B.34}$$

so the time-averaged autocorrelation of $X(t)$ equals its probabilistic autocorrelation.

B.4 Gaussian Processes

Noise processes in communication systems are commonly modeled as a Gaussian process. A random process $X(t)$ is a Gaussian process if for all values of T and all functions $g(t)$ the random variable

$$X_g = \int_0^T g(t)X(t)dt \tag{B.35}$$

has a Gaussian distribution. Since a communication receiver typically uses an integrator in signal detection, this definition implies that if the channel introduces a Gaussian noise process at the receiver input, the distribution of the random variable associated with the noise at the output of the integrator will have a Gaussian distribution. The mean of X_g is

$$\mathbf{E}[X_g] = \int_0^T g(t)\mathbf{E}[X(t)]dt \tag{B.36}$$

and the variance is

$$\text{Var}[X_g] = \int_0^T\int_0^T g(t)g(s)\mathbf{E}[X(t)X(s)]dtds - (\mathbf{E}[X_g])^2 \tag{B.37}$$

If $X(t)$ is WSS these simplify to

$$\mathbf{E}[X_g] = \int_0^T g(t)\mu_X dt \tag{B.38}$$

and

$$\text{Var}[X_g] = \int_0^T\int_0^T g(t)g(s)R_X(s-t)dtds - (\mathbf{E}[X_g])^2. \tag{B.39}$$

Several important properties of Gaussian random processes can be obtained from the definition. In particular, if a Gaussian random process is input to a linear time-invariant filter, the filter output is also a Gaussian random process. Moreover, we expect samples $X(t_i), i = 0, 1, \ldots$ of a Gaussian random process to be jointly Gaussian random variables, and indeed that follows from the definition by setting $g(t) = \delta(t - t_i)$ in (B.35). Since these samples are Gaussian random variables, if the samples are uncorrelated, they are also independent. In addition, for a WSS Gaussian processes, the distribution of X_g in (B.35) only depends on the mean and autocorrelation of the process $X(t)$. Finally, note that a random process is completely defined by the joint probability of its samples over all sets of sample times. For a Gaussian process, these samples are jointly Gaussian with their joint distribution determined by the mean and autocorrelation of the process. Thus, since the underlying probability of a Gaussian process is completely determined by its mean and autocorrelation, there are no higher moments for the process, so a WSS Gaussian process is also stationary. Similarly, a Gaussian process that is ergodic in the mean and autocorrelation is an ergodic process.

Bibliography

[1] A. Papoulis and S.U. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 2002.

[2] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd Ed., Addison-Wesley, 1994.

[3] R.M. Gray and L.D. Davisson, *Random Processes: A Mathematical Approach for Engineers*, Prentice-Hall, 1986.

[4] W. B. Davenport, Jr. and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, McGraw Hill, 1987.

[5] H. Stark and J. W. Woods, *Probability and Random Processes with Applications to Signal Processing*, 3rd Ed., Prentice Hall, 2001.

[6] R. G. Gallager, *Discrete Stochastic Processes*. Kluwer, 1996.

[7] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I and Vol II, Wiley, 1968/1971.

[8] P. Billingsley, *Probability and Measure*, 3rd. Ed., Wiley, 1995.

Appendix C

Matrix Definitions, Operations, and Properties

C.1 Matrices and Vectors

An $N \times M$ *matrix* $\mathbf{A}$ is a rectangular array of values with N rows and M columns, written as

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NM} \end{bmatrix}. \tag{C.1}$$

The ijth element (or entry) of $\mathbf{A}$, i.e. the element in the ith row and jth column, is written as $\mathbf{A}_{ij}$. In (C.1) we have $\mathbf{A}_{ij} = a_{ij}$. The matrix elements are also called *scalars* to indicate that they are single numbers. An $N \times M$ matrix is called a *square matrix* if $N = M$, a *skinny matrix* if $N > M$ and a *fat matrix* if $N < M$.

The *diagonal elements* of a square matrix are the elements along the diagonal line from the top left to the bottom right of the matrix, i.e. the elements $\mathbf{A}_{ij}$ with $i = j$. The *trace* of a square $N \times N$ matrix is the sum of its diagonal elements: $\text{Tr}[\mathbf{A}] = \sum_{i=1}^{N} \mathbf{A}_{ii}$. A square matrix is called a *diagonal matrix* if all elements that are not diagonal elements, referred to as the *off-diagonal* elements, are zero: $\mathbf{A}_{ij} = 0, j \neq i$. We denote a diagonal matrix with diagonal elements $a_1, \ldots, a_N$ as diag$[a_1, \ldots, a_N]$. The $N \times N$ identity matrix $\mathbf{I}_N$ is a diagonal matrix with $\mathbf{I}_{ii} = 1, i = 1, \ldots, N$, i.e. $\mathbf{I}_N = \text{diag}[1, \ldots, 1]$. The subscript N of $\mathbf{I}_N$ is omitted when the size is clear from the context (e.g. from the size requirements for a given operation like matrix multiplication).

A square matrix $\mathbf{A}$ is called *upper triangular* if all its elements below the diagonal are zero, i.e. $\mathbf{A}_{ij} = 0, i > j$. A *lower triangular* matrix is a square matrix where all elements above the diagonal are zero, i.e. $\mathbf{A}_{ij} = 0, i < j$. Diagonal matrices are both upper triangular and lower triangular.

Matrices can be formed from entries that are themselves matrices, as long as the dimensions are consistent. In particular, if $\mathbf{B}$ is an $N \times M_1$ matrix and $\mathbf{C}$ is an $N \times M_2$ matrix then we can form the $N \times (M_1 + M_2)$ matrix $\mathbf{A} = [\mathbf{B}\ \ \mathbf{C}]$. The ith row of this matrix is $[\mathbf{A}_{i1}\ \ldots\ \mathbf{A}_{i(M_1+M_2)}] = [\mathbf{B}_{i1}\ \ldots\ \mathbf{B}_{iM_1}\ \mathbf{C}_{i1}\ \ldots\ \mathbf{C}_{iM_2}]$. The matrix $\mathbf{A}$ formed in this way is also written as $\mathbf{A} = [\mathbf{B}|\mathbf{C}]$. If we also have a $K \times L_1$ matrix $\mathbf{D}$ and a $K \times L_2$ matrix $\mathbf{E}$ then if $M_1 + M_2 = L_1 + L_2$ we can form the $(N + K) \times (M_1 + M_2)$ matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix}. \tag{C.2}$$

The matrices $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$ are called submatrices of $\mathbf{A}$. A matrix can be composed of any number of submatrices as long as the sizes are compatible. A submatrix $\mathbf{A}'$ of $\mathbf{A}$ can also be obtained by deleting certain rows and/or columns of $\mathbf{A}$.

A matrix with only one column, i.e., with $M = 1$, is called a *column vector* or just a *vector*. The number of rows of a vector is called its dimension. For example, an N-dimensional vector $\mathbf{x}$ is given by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}. \tag{C.3}$$

The ith element of vector $\mathbf{x}$ is written as $\mathbf{x}_i$. We call an N-dimensional vector with each element equal to one a *ones vector* and denote it by $\mathbf{1}_N$. An N-dimensional vector with one element equal to one and the rest equal to zero is called a *unit vector*. In particular, the ith unit vector $\mathbf{e}^i$ has $\mathbf{e}^i_i = 1$ and $\mathbf{e}^i_j = 0, j \neq i$. A matrix with only one row, i.e., with $N = 1$, is called a row vector. The number of columns in a row vector is called its dimension, so an M-dimensional row vector $\mathbf{x}$ is given by $\mathbf{x} = [x_1 \; \ldots \; x_M]$ with ith element $\mathbf{x}_i = x_i$. The *Euclidean norm* of an N-dimensional row vector or vector, also called its *norm*, is defined as

$$||\mathbf{x}|| = \sqrt{\sum_{i=1}^{N} |\mathbf{x}_i|^2}. \tag{C.4}$$

C.2 Matrix and Vector Operations

If $\mathbf{A}$ is an $N \times M$ matrix, the *transpose* of $\mathbf{A}$, denoted $\mathbf{A}^T$, is the $M \times N$ matrix defined by $\mathbf{A}^T_{ij} = \mathbf{A}_{ji}$:

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NM} \end{bmatrix}^T = \begin{bmatrix} a_{11} & \cdots & a_{N1} \\ \vdots & \ddots & \vdots \\ a_{1M} & \cdots & a_{NM} \end{bmatrix}. \tag{C.5}$$

In other words, $\mathbf{A}^T$ is obtained by transposing the rows and columns of $\mathbf{A}$, so the ith row of $\mathbf{A}$ becomes the ith column of $\mathbf{A}^T$. The transpose of a row vector $\mathbf{x} = [\mathbf{x}_1 \; \ldots \; \mathbf{x}_N]$ yields a vector with the same elements:

$$\mathbf{x}^T = [\mathbf{x}_1 \ldots \mathbf{x}_N]^T = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}. \tag{C.6}$$

We therefore often write a column vector $\mathbf{x}$ with elements $\mathbf{x}_i$ as $\mathbf{x} = [\mathbf{x}_1 \; \ldots \; \mathbf{x}_N]^T$. Similarly, the transpose of an N-dimensional vector $\mathbf{x}$ with ith element $\mathbf{x}_i$ is the row vector $[\mathbf{x}_1 \; \ldots \; \mathbf{x}_N]$. Note that for $\mathbf{x}$ a row vector or vector, $(\mathbf{x}^T)^T = \mathbf{x}$.

The *complex conjugate* $\mathbf{A}^*$ of a matrix $\mathbf{A}$ is obtained by taking the complex conjugate of each element of $\mathbf{A}$:

$$\mathbf{A}^* = \begin{bmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NM} \end{bmatrix}^* = \begin{bmatrix} a^*_{11} & \cdots & a^*_{1M} \\ \vdots & \ddots & \vdots \\ a^*_{N1} & \cdots & a^*_{NM} \end{bmatrix}. \tag{C.7}$$

The *Hermitian* of a matrix $\mathbf{A}$, denoted as $\mathbf{A}^H$, is defined as its conjugate transpose: $\mathbf{A}^H = (\mathbf{A}^*)^T$. Note that applying the Hermitian operation twice results in the original matrix: $(\mathbf{A}^H)^H = \mathbf{A}$, so $\mathbf{A}$ is the Hermitian of $\mathbf{A}^H$. A square matrix $\mathbf{A}$ is a *Hermitian matrix* if it equals its Hermitian: $\mathbf{A} = \mathbf{A}^H$. The complex conjugate and Hermitian operators can also be applied to vectors. In particular, the complex conjugate of a vector or row vector

$\mathbf{x}$, denoted as $\mathbf{x}^*$, is obtained by taking the complex conjugate of each element of $\mathbf{x}$. The Hermitian of a vector $\mathbf{x}$, denoted as $\mathbf{x}^H$, is its conjugate transpose: $\mathbf{x}^H = (\mathbf{x}^*)^T$.

Two $N \times M$ matrices can be added together to form a new matrix of size $N \times M$. The addition is done element-by-element. In other words, if two $N \times M$ matrices $\mathbf{A}$ and $\mathbf{B}$ are added, the resulting $N \times M$ matrix $\mathbf{C} = \mathbf{A} + \mathbf{B}$ has ijth element $\mathbf{C}_{ij} = \mathbf{A}_{ij} + \mathbf{B}_{ij}$. Since matrix addition is done element-by-element, it inherits the commutative and associative properties of addition, i.e. $\mathbf{A}+\mathbf{B} = \mathbf{B}+\mathbf{A}$, and $(\mathbf{A}+\mathbf{B})+\mathbf{C} = \mathbf{A}+(\mathbf{B}+\mathbf{C})$. The transpose of a sum of matrices is the sum of the transposes of the individual matrices: $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$. Matrix subtraction is similar: for two $N \times M$ matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{C} = \mathbf{A} - \mathbf{B}$ is an $N \times M$ matrix with ijth element $\mathbf{C}_{ij} = \mathbf{A}_{ij} - \mathbf{B}_{ij}$. Two row vectors or vectors of the same dimension can be added using this definition of matrix addition since these vectors are special cases of matrices. In particular, an N-dimensional vector $\mathbf{x}$ can be added to another vector $\mathbf{y}$ of the same dimension to form the new N-dimensional vector $\mathbf{z} = \mathbf{x} + \mathbf{y}$ with ith element $\mathbf{z}_i = \mathbf{x}_i + \mathbf{y}_i$. Similarly, if $\mathbf{x}$ and $\mathbf{y}$ are row vectors of dimension N, their sum $\mathbf{z} = \mathbf{x} + \mathbf{y}$ is an N-dimensional row vector with ith element $\mathbf{z}_i = \mathbf{x}_i + \mathbf{y}_i$. However, a row vector of dimension $N > 1$ cannot be added to a vector of dimension N, since these vectors are matrices of different sizes ($1 \times N$ for the row vector, $N \times 1$ for the vector). The linear combination of vectors $\mathbf{x}$ and $\mathbf{y}$ of dimension N yields a new N-dimensional vector $\mathbf{z} = c\mathbf{x} + d\mathbf{y}$ with ith element $\mathbf{z}_i = c\mathbf{x}_i + d\mathbf{y}_i$, where c and d are arbitrary scalars. Similarly, row vectors $\mathbf{x}$ and $\mathbf{y}$ of dimension N can be linearly combined to form the N-dimensional row vector $\mathbf{z} = c\mathbf{x} + d\mathbf{y}$ with ith element $\mathbf{z}_i = c\mathbf{x}_i + d\mathbf{y}_i$ for arbitrary scalars c and d.

A matrix can be multiplied by a scalar, in which case every element of the matrix is multiplied by the scalar. Specifically, multiplication of the matrix $\mathbf{A}$ by a scalar k results in the matrix $k\mathbf{A}$ given by

$$k\mathbf{A} = k \begin{bmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NM} \end{bmatrix} = \begin{bmatrix} ka_{11} & \cdots & ka_{1M} \\ \vdots & \ddots & \vdots \\ ka_{N1} & \cdots & ka_{NM} \end{bmatrix}. \tag{C.8}$$

A row vector $\mathbf{x}$ multiplied by scalar k yields $k\mathbf{x} = [k\mathbf{x}_1 \ \ldots \ k\mathbf{x}_N]$, and a vector $\mathbf{x}$ multiplied by scalar k yields $k\mathbf{x} = [k\mathbf{x}_1 \ \ldots \ k\mathbf{x}_N]^T$.

Two matrices can be multiplied together provided they have compatible dimensions. In particular, matrices $\mathbf{A}$ and $\mathbf{B}$ can be multiplied if the number of columns of $\mathbf{A}$ equals the number of rows of $\mathbf{B}$. If $\mathbf{A}$ is an $N \times M$ matrix and $\mathbf{B}$ is a $M \times L$ matrix then their product $\mathbf{C} = \mathbf{AB}$ is an $N \times L$ matrix with ijth element $\mathbf{C}_{ij} = \sum_{k=1}^{M} \mathbf{A}_{ik}\mathbf{B}_{kj}$. Matrix multiplication is not commutative in general, i.e. in general $\mathbf{AB} \neq \mathbf{BA}$. In fact, if $\mathbf{A}$ is an $N \times M$ matrix and $\mathbf{B}$ is a $M \times L$ matrix then the product $\mathbf{BA}$ only exists if $L = N$. In this case $\mathbf{BA}$ is an $M \times M$ matrix, which may be a different size than the $N \times L$ matrix $\mathbf{AB}$. Even if $M = L = N$, so that $\mathbf{AB}$ and $\mathbf{BA}$ are the same size, they may not be equal. If $\mathbf{A}$ is a square matrix then we can multiply $\mathbf{A}$ by itself. In particular, we define $\mathbf{A}^2 = \mathbf{AA}$. Similarly $\mathbf{A}^k = \mathbf{A} \ldots \mathbf{A}$ is the product of k copies of $\mathbf{A}$. This implies that $\mathbf{A}^k\mathbf{A}^l = \mathbf{A}^{k+l}$. Multiplication of any matrix by the identity matrix of compatible size results in the same matrix, i.e. if $\mathbf{A}$ is an $N \times M$ matrix, then $\mathbf{I}_N\mathbf{A} = \mathbf{A}\mathbf{I}_M = \mathbf{A}$. The transpose of a matrix product is the product of the transpose of the individual matrices in reverse order: $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. The product of an $N \times M$ matrix $\mathbf{A}$ and its $M \times N$ Hermitian $\mathbf{A}^H$ is a square matrix. In particular, $\mathbf{AA}^H$ is an $N \times N$ square matrix while $\mathbf{A}^H\mathbf{A}$ is an $M \times M$ square matrix. The *Frobenius norm* of a matrix $\mathbf{A}$, denoted as $||\mathbf{A}||_F$, is defined as $||\mathbf{A}||_F = \sqrt{\text{Tr}[\mathbf{AA}^H]} = \sqrt{\text{Tr}[\mathbf{A}^H\mathbf{A}]} = \sum_{i=1}^{N}\sum_{j=1}^{M} |\mathbf{A}_{ij}|^2$. Matrix multiplication is associative, i.e. $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ as long as the matrix dimensions are compatible for multiplication, so the parentheses are typically omitted. Matrix multiplication is also distributive: $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ and $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.

An M-dimensional vector can be multiplied by a matrix with M columns. Specifically, if $\mathbf{A}$ is an $N \times M$ matrix and $\mathbf{x}$ is an M-dimensional vector (i.e. an $M \times 1$ matrix) then their product yields an N-dimensional vector $\mathbf{y} = \mathbf{Ax}$ with ith element $\mathbf{y}_i = \sum_{k=1}^{M} \mathbf{A}_{ik}\mathbf{x}_k$. Note that a matrix must left-multiply a vector, since the dimensions are not compatible for the product $\mathbf{xA}$. However, if $\mathbf{x}$ is an N-dimensional row vector, then $\mathbf{xA}$ is a compatible multiplication for $\mathbf{A}$ an $N \times M$ matrix, and results in the M-dimensional row vector $\mathbf{y} = \mathbf{x}\mathbf{A}$ with ith element $\mathbf{y}_i = \sum_{k=1}^{N} \mathbf{x}_k \mathbf{A}_{ki}$. An N-dimensional row vector $\mathbf{x}$ can be multiplied by an N-dimensional vector $\mathbf{y}$, which results in a scalar $z = \mathbf{x}\mathbf{y} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{y}_i$. Note that the transpose of an N-dimensional vector is an N-dimensional row vector. The *inner product* of two N-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as $< \mathbf{x}, \mathbf{y} >= \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{y}_i$.

Given a matrix $\mathbf{A}$, a subset of rows of $\mathbf{A}$ form a *linearly independent* set if any row in the subset is not equal to a linear combination of the other rows in the subset. Similarly, a subset of columns of $\mathbf{A}$ form a linearly independent set if any column in the subset is not equal to a linear combination of the other columns in the subset. The *rank* R_A of a matrix $\mathbf{A}$ is equal to the number of rows in the largest subset of linearly independent rows of $\mathbf{A}$, which can be shown to equal the number of columns in the largest subset of linearly independent columns of $\mathbf{A}$. This implies that the rank of an $N \times M$ matrix cannot exceed $\min[N, M]$. An $N \times M$ matrix $\mathbf{A}$ is *full rank* if $R_A = \min[N, M]$.

The *determinant* of a 2 × 2 matrix $\mathbf{A}$ is defined as det$[\mathbf{A}] = \mathbf{A}_{11}\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{12}$. For an $N \times N$ matrix $\mathbf{A}$ that is larger than 2 × 2, det$[\mathbf{A}]$ is defined recursively as

$$\text{det}[\mathbf{A}] = \sum_{i=1}^{N} \mathbf{A}_{ij} c_{ij} \tag{C.9}$$

for any $j : 1 \leq j \leq N$, where c_{ij} is the *co-factor* corresponding to the matrix element $\mathbf{A}_{ij}$, defined as

$$c_{ij} = (-1)^{i+j} \text{det}[\mathbf{A}'], \tag{C.10}$$

where $\mathbf{A}'$ is the submatrix of $\mathbf{A}$ obtained by deleting the ith row and jth column of $\mathbf{A}$.

If $\mathbf{A}$ is an $N \times N$ square matrix, and there is another $N \times N$ matrix $\mathbf{B}$ such that $\mathbf{BA} = \mathbf{I}_N$, then we say that $\mathbf{A}$ is *invertible* or *nonsingular*. We call $\mathbf{B}$ the *inverse* of $\mathbf{A}$, and we denote this inverse as $\mathbf{A}^{-1}$. Thus, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_N$. Moreover, for $\mathbf{A}^{-1}$ defined in this way, we also have that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_N$. Only square matrices can be invertible, and the matrix inverse is the same size as the original matrix. A square invertible matrix $\mathbf{U}$ is *unitary* if $\mathbf{U}\mathbf{U}^H = \mathbf{I}$, which implies that $\mathbf{U}^H = \mathbf{U}^{-1}$ and thus $\mathbf{U}^H\mathbf{U} = \mathbf{I}$. Not every square matrix is invertible. If a matrix is not invertible, we say it is *singular* or *noninvertible*. The inverse of an inverse matrix is the original matrix: $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$. The inverse of the product of matrices is the product of the inverses in opposite order: $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. The kth power of the inverse is $\mathbf{A}^{-k} = (\mathbf{A}^{-1})^k$.

For a diagonal matrix $\mathbf{D} = \text{diag}[d_1, \ldots, d_N]$ with $d_i \neq 0, i = 1, \ldots N$ the inverse exists and is given by $\mathbf{D}^{-1} = \text{diag}[1/d_1, \ldots, 1/d_N]$. For a general 2 × 2 matrix $\mathbf{A}$ with ijth element a_{ij}, its inverse exists if det$[\mathbf{A}] \neq 0$ and is given by

$$\mathbf{A}^{-1} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{\text{det}[\mathbf{A}]} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \tag{C.11}$$

There are more complicated formulas for the inverse of invertible matrices with size greater than 2 × 2. However, matrix inverses are usually obtained using computer math packages.

Matrix inverses are commonly used to solve systems of linear equations. In particular, consider a set of linear equations, expressed in matrix form as

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{C.12}$$

If the matrix $\mathbf{A}$ is invertible then, given $\mathbf{y}$, there is a unique vector $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ that satisfies this system of equations.

C.3 Matrix Decompositions

Given a square matrix $\mathbf{A}$, a scalar value λ for which there exists a nonzero vector $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is called an *eigenvalue* of $\mathbf{A}$. The vector $\mathbf{x}$ is called the *eigenvector* of $\mathbf{A}$ corresponding to λ. The eigenvalues of a matrix

$\mathbf{A}$ are all values of λ that satisfy the *characteristic equation* of $\mathbf{A}$, defined as det$[\mathbf{A}-\lambda\mathbf{I}]$=0. The polynomial in λ defined by det$[\mathbf{A}-\lambda\mathbf{I}]$ is called the *characteristic polynomial* of $\mathbf{A}$, so the eigenvalues of $\mathbf{A}$ are the roots of its characteristic polynomial. The characteristic polynomial of an $N\times N$ matrix has N unique roots $r_1,\ldots,r_N$, $r_i\neq r_j$ if it is of the form det$[\mathbf{A}-\lambda\mathbf{I}]=(\lambda-r_1)\ldots(\lambda-r_N)$. When the characteristic polynomial includes a term $(\lambda-r_i)^k, k>1$ we say that root r_i has *multiplicity* k. For example, if det$[\mathbf{A}-\lambda\mathbf{I}]=(\lambda-r_1)^2(\lambda-r_2)^3$ then root r_1 has multiplicity 2 and root r_2 has multiplicity 3. An $N\times N$ matrix has N eigenvalues $\lambda_1,\ldots,\lambda_N$, although they will not all be unique if any of the roots of the characteristic polynomial have multiplicity greater than 1. It can be shown that the determinant of a matrix equals the product of all its eigenvalues (i.e. an eigenvalue r_i with multiplicity k would contribute r_i^k to the product).

The eigenvalues of a Hermitian matrix are always real, although the eigenvectors can be complex. Moreover, if $\mathbf{A}$ is an $N\times N$ Hermitian matrix then it can be written in the following form:

$$\mathbf{A}=\mathbf{P}\mathbf{\Lambda}\mathbf{P}^H, \tag{C.13}$$

where $\mathbf{\Lambda}=\text{diag}[\lambda_1,\ldots,\lambda_K,0,\ldots,0]$ is an $N\times N$ diagonal matrix whose first K diagonal elements are the nonzero (real) eigenvalues of $\mathbf{A}$. We say that a matrix $\mathbf{A}$ is *positive definite* if for all nonzero vectors $\mathbf{x}$, $\mathbf{x}^H\mathbf{A}\mathbf{x}>0$. A Hermitian matrix is positive definite if and only if all its eigenvalues are positive. Similarly, we say the matrix $\mathbf{A}$ is *positive semi-definite* or *non-negative definite* if for all nonzero vectors $\mathbf{x}$, $\mathbf{x}^H\mathbf{A}\mathbf{x}\geq 0$. A Hermitian matrix is non-negative definite if and only if all of its eigenvalues are non-negative.

Suppose that $\mathbf{A}$ is an $N\times M$ matrix of rank R_A. Then there is an $N\times M$ matrix Σ and two unitary matrices $\mathbf{U}$ and $\mathbf{V}$ of size $N\times N$ and $M\times M$, respectively, such that

$$\mathbf{A}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H. \tag{C.14}$$

We call the columns of $\mathbf{V}$ the *right eigenvectors* of $\mathbf{A}$ and the columns of $\mathbf{U}$ the *left eigenvectors* of $\mathbf{A}$. The matrix $\mathbf{\Sigma}$ has a special form: all elements that are not diagonal elements are zero, so

$$\mathbf{\Sigma}_{N\times M}=\begin{bmatrix}\sigma_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_M\\ 0 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 0\end{bmatrix} \tag{C.15}$$

for $N\geq M$, and

$$\mathbf{\Sigma}_{N\times M}=\begin{bmatrix}\sigma_1 & \cdots & 0 & 0 & \ldots & 0\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma_M & 0 & \ldots & 0\end{bmatrix} \tag{C.16}$$

for $N<M$, where $\sigma_i=\sqrt{\lambda_i}$ for λ_i the ith eigenvalue of $\mathbf{A}\mathbf{A}^H$. The values of σ_i are called the *singular values* of $\mathbf{A}$, and R_A of these singular values are nonzero. The decomposition (C.14) is called the *singular value decomposition* of $\mathbf{A}$. The singular values of a matrix are always non-negative.

Let $\mathbf{A}$ be an $N\times M$ matrix where we denote its ith column as $\mathbf{A}_i$. Treating each column as a submatrix, we can write $\mathbf{A}=[\mathbf{A}_1\ \mathbf{A}_2\ \ldots\ \mathbf{A}_M]$. The vectorization of the matrix $\mathbf{A}$, denoted as vec($\mathbf{A}$), is defined as the NM-dimensional vector that results from stacking the columns $\mathbf{A}_i$, $i=1,\ldots,N$ of matrix $\mathbf{A}$ on top of each other to form a vector:

$$\text{vec}(\mathbf{A})=\begin{bmatrix}\mathbf{A}_1\\ \vdots\\ \mathbf{A}_N\end{bmatrix}=[\mathbf{A}_{11}\ \mathbf{A}_{21}\ \ldots\ \mathbf{A}_{N1}\ \mathbf{A}_{12}\ \ldots\ \mathbf{A}_{N2}\ \ldots\ \mathbf{A}_{1M}\ \ldots\ \mathbf{A}_{NM}]^T. \tag{C.17}$$

Let $\mathbf{A}$ be an $N \times M$ matrix and $\mathbf{B}$ be an $L \times K$ matrix. The *Kronecker product* of $\mathbf{A}$ and $\mathbf{B}$, denoted $\mathbf{A} \otimes \mathbf{B}$, is a $NL \times MK$ matrix defined by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B} & \cdots & \mathbf{A}_{1M}\mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{N1}\mathbf{B} & \cdots & \mathbf{A}_{NM}\mathbf{B} \end{bmatrix}. \tag{C.18}$$

Bibliography

[1] G. Strang, *Linear Algebra and its Applications*, 2nd Ed., New York: Academic Press, 1980.

[2] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[3] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1991.

[4] B. Nobel and J.W. Daniel, *Applied Linear Algebra*. Prentice-Hall, 1977.

Appendix D

Summary of Wireless Standards

This chapter summarizes the technical details associated with the two most prevalent wireless systems in operation today: cellular phones and wireless LANs. It also summarizes the specifications for three short range wireless network standards that have emerged to support a broad range of applications.

D.1 Cellular Phone Standards

D.1.1 First Generation Analog Systems

In this section we summarize cellular phone standards. We begin with the standards for first-generation (1G) analog cellular phones, whose main characteristics are summarized in Table D.1. Systems based on these standards were widely deployed in the 1980s. While many of these systems have been replaced by digital cellular systems, there are many places throughout the world where these analog systems are still in use. The best known standard is the Advanced Mobile Phone System (AMPS), developed by Bell Labs in the 1970s, and first used commercially in the US in 1983. After its US deployment, many other countries adopted it as well. AMPS has a narrowband version, narrowband AMPS (N-AMPS), with voice channels that are one third the bandwidth of regular AMPS. Japan deployed the first commercial cellular phone system in 1979 with the NTT (MCS-L1) standard based on AMPS, but at a higher frequency and with voice channels of slightly lower bandwidth. Europe also developed a similar standard to AMPS called the Total Access Communication System (TACS). TACS operates at a higher frequency and with lower bandwidth channels than AMPS. It was deployed in the U.K. and in other European coutries as well as outside Europe. The frequency range for TACS was extended in the U.K. to obtain more channels , leading to a variation called ETACS. A variation of the TACS system called JTACS was deployed in metropolitan areas of Japan in 1989 to provide higher capacity than the NTT system. JTACS operates at a slightly higher frequency than TACS and ETACS, and has a bandwidth-efficient version called NTACS, where voice channels occupy half the bandwidth of the channnels in JTACS. In addition to TACS, countries in Europe had different incompatible standards at different frequencies for analog cellular, including the Nordic Mobile Telephone (NMT) standard in Scandanavia, the Radiocom 2000 (RC2000) standard in France, and the C-450 standard in Germany and Portugal. The incompatibilities made it impossible to roam between European countries with a single analog phone, which motivated the need for one unified cellular standard and frequency allocation throughout Europe.

D.1.2 Second Generation Digital Systems

Next we consider second-generation (2G) digital cellular phone standards, whose main characteristics are summarized in Table D.2. These systems were mostly deployed in the early 1990s. Due to incompatibilities in the first-generation analog systems, in 1982 the Groupe Spécial Mobile (GSM) was formed to develop a uniform

	AMPS	TACS	NMT (450/900)	NTT	C-450	RC2000
Uplink Frequencies (MHz)	824-849	890-915	453-458/890-915	925-940[1]	450-455.74	414.8-418[2]
Downlink Frequencies (MHz)	869-894	935-960	463-468/935-960	870-885	460-465.74	424.8-428
Modulation	FM	FM	FM	FM	FM	FM
Channel Spacing (KHz)	30	25	25/12.5	25	10	12.5
Number of Channels	832	1000	180/1999	600	573	256
Multiple Access	FDMA	FDMA	FDMA	FDMA	FDMA	FDMA

Table D.1: First-Generation Analog Cellular Phone Standards

digital cellular standard for all of Europe. The TACS spectrum in the 900 MHz band was allocated for GSM operation across Europe to facilitate roaming between countries. In 1989 the GSM specification was finalized and the system was launched in 1991, although availability was limited until 1992. The GSM standard uses TDMA combined with slow frequency hopping to combat out-of-cell interference. Convolutional coding and parity check codes along with interleaving is used for error correction and detection. The standard also includes an equalizer to compensate for frequency-selective fading. The GSM standard is used in about 66 % of the world's cellphones, with more than 470 GSM operators in 172 countries supporting over a billion users. As the GSM standard became more global, the meaning of the acronym was changed to the Global System for Mobile Communications.

Although Europe got an early jump on developing 2G digital systems, the US was not far behind. In 1992 the IS-54 digital cellular standard was finalized, with commercial deployment beginning in 1994. This standard uses the same channel spacing, 30 KHz, as AMPS to facilitate the analog to digital transition for wireless operators, along with a TDMA multiple access scheme to improve handoff and control signaling over analog FDMA. The IS-54 standard, also called the North American Digital Cellular standard, was improved over time and these improvements evolved into the IS-136 standard, which subsumed the original standard. Similar to the GSM standard, the IS-136 standard uses parity check codes, convolutional codes, interleaving, and equalization.

A competing standard for 2G systems based on CDMA was proposed by Qualcomm in the early 1990s. The standard, called IS-95 or IS-95a, was finalized in 1993 and deployed commercially under the name cdmaOne in 1995. Like IS-136, IS-95 was designed to be compatible with AMPS so that the two systems could coexist in the same frequency band. In CDMA all users are superimposed on top of each other with spreading codes that can separate out the users at the receiver. Thus, channel data rate does not apply to just one user, as in TDMA systems. The channel chip rate is 1.2288 Mchips/s for a total spreading factor of 128 for both the uplink and downlink. The spreading process in IS-95 is different for the downlink (DL) and the uplink (UL), with spreading on both links accomplished through a combination of spread spectrum modulation and coding. On the downlink data is first rate 1/2 convolutionally encoded and interleaved, then modulated by one of 64 orthogonal spreading sequences (Walsh functions). Then a synchronized scrambling sequence unique to each cell is superimposed on top of the Walsh function to reduce interference between cells. The scrambling requires synchronization between base stations. Uplink spreading is accomplished using a combination of a rate 1/3 convolutional code with interleaving, modulation by an orthogonal Walsh function, and modulation by a nonorthogonal user/base station specific code. The IS-95 standard includes a parity check code for error detection, as well as power control for the reverse link to avoid the near-far problem. A 3-finger RAKE receiver is also specified to provide diversity and compensate for ISI. A form of base station diversity called soft handoff (SHO), whereby a mobile maintains a connection to both the new and old base stations during handoff and combines their signals, is also included in the standard. CDMA has some advantages over TDMA for cellular systems, including no need for frequency planning, SHO capabilities,

[1] NTT also operated in several other frequency bands around 900 MHz.

[2] RC2000 also operated in several other frequency bands around 200 MHz.

the ability to exploit voice activity to increase capacity, and no hard limit on the number of users that can be accommodated in the system. There was much debate about the relative merits of the IS-54 and IS-95 standards throughout the early 1990s, with claims that IS-95 could achieve 20 times the capacity of AMPS whereas IS-54 could only achieve 3 times this capacity. In the end, both systems turned out to achieve approximately the same capacity increase over AMPS.

The 2G digital cellular standard in Japan, called the Personal Digital Cellular (PDC) standard, was established in 1991 and deployed in 1994. It is similar to the IS-136 standard, but with 25 KHz voice channels to be compatible with the Japanese analog systems. This system operates in both the 900 MHz and 1500 MHz frequency bands.

	GSM	IS-136	IS-95 (cdmaOne)	PDC
Uplink Frequencies (MHz)	890-915	824-849	824-849	810-830,1429-1453
Downlink Frequencies (MHz)	935-960	869-894	869-894	940-960, 1477-1501
Carrier Separation (KHz)	200	30	1250	25
Number of Channels	1000	2500	~ 2500	3000
Modulation	GMSK	$\pi/4$ DQPSK	BPSK/QPSK	$\pi/4$ DQPSK
Compressed Speech Rate (Kbps)	13	7.95	1.2-9.6 (Variable)	6.7
Channel Data Rate (Kbps)	270.833	48.6	(1.2288 Mchips/s)	42
Data Code Rate	$1/2$	$1/2$	1/2 (DL), 1/3 (UL)	$1/2$
ISI Reduction/Diversity	Equalizer	Equalizer	RAKE, SHO	Equalizer
Multiple Access	TDMA/Slow FH	TDMA	CDMA	TDMA

Table D.2: Second-Generation Digital Cellular Phone Standards

D.1.3 Evolution of 2G Systems

In the late 1990s 2G systems evolved in two directions: they were ported to higher frequencies as more cellular bandwidth became available in Europe and the US, and they were modified to support data services in addition to voice. Specifically, in 1994 the FCC began auctioning spectrum in the Personal Communication Systems (PCS) band at 1.9 GHz for cellular systems. Operators purchasing spectrum in this band could adopt any standard. Different operators chose different standards, so GSM, IS-136, and IS-95 were all deployed at 1900 MHz in different parts of the country, making nationwide roaming with a single phone difficult. In fact, many of the initial digital cellphones included an analog AMPS mode in case the digital system was not available. GSM systems operating in the PCS band are sometimes refered to as PCS 1900 systems. The IS-136 and IS-95 (cdmaOne) standards translated to the PCS band go by the same names. Europe allocated additional cellular spectrum in the 1.8 GHz band. The standard for this frequency band, called GSM 1800 or DCS 1800 (for Digital Cellular System), uses GSM as the core standard with some modifications to allow overlays of macrocells and microcells. Note that second-generation cordless phones such as DECT, the Personal Access Communications System (PACS), and the Personal Handyphone System (PHS) also operate in the 1.9 GHz frequency band, but these systems are mostly within buildings supporting private branch exchange (PBX) services.

Once digital cellular became available, operators began incorporating data services in addition to voice. The 2G systems with added data capabilities are sometimes refered to as 2.5G systems. The enhancements to 2G systems made to support data services are summarized in Table D.3. GSM systems followed several different upgrade paths to provide data services. The simplest, called High Speed Circuit Switched Data (HSCSD), allows up to 4 consecutive timeslots to be assigned to a single user, thereby providing a maximum transmission rate of up to 57.6 Kbps. Circuit switching is quite inefficient for data, so a more complex enhancement provides for packet-switched data layered on top of the circuit-switched voice. This enhancement is refered to as General Packet Radio Service (GPRS). A maximum data rate of 171.2 Kbps is possible with GPRS when all 8 timeslots of a GSM frame are allocated to a single user. The data rates of GPRS are further enhanced through variable-rate modulation and coding, refered to as Enhanced Data rates for GSM Evolution (EDGE). EDGE provides data rates up to 384 Kbps with a bit rate of 48-69.2 Kbps per timeslot. GPRS and EDGE are compatible with IS-136 as well as GSM, and thus provide a convergent upgrade path for both of these systems.

The IS-95 standard was modified to provide data services by assigning multiple orthogonal Walsh functions to a single user. A maximum of 8 functions can be assigned, leading to a maximum data rate of 115.2 Kbps, although in practice only about 64 Kbps is achieved. This evolution is refered to as the IS-95b standard.

2G Standard	GSM	GSM/IS-136		IS-95
2.5G Enhancement	HSCSD	GPRS	EDGE	IS-95b
Technique	Aggregate Timeslots	Aggregate Timeslots with Packet Switching	GPRS with Variable Mod./Cod.	Aggregate Walsh Functions
Data Rates: Max/Actual	57.6/14.4-57.6 Kbps	140.8/56 Kbps	384/200 Kbps	115/64 Kbps

Table D.3: 2G Enhancements to Support 2.5G Data Capabilities

D.1.4 Third Generation Systems

The fragmentation of standards and frequency bands associated with 2G systems led the International Telecommunications Union (ITU) in the late 1990s to formulate a plan for a single global frequency band and standard for third-generation (3G) digital cellular systems. The standard was named the International Mobile Telephone 2000 (IMT-2000) standard with a desired system rollout in the 2000 timeframe. In addition to voice services, IMT-2000 was to provide Mbps data rates for demanding applications such as broadband Internet access, interactive gaming, and high quality audio and video entertainment. Agreement on a single standard did not materialize, with most countries supporting one of two competing standards: cdma2000 (backward compatible with cdmaOne) supported by the Third Generation Partnership Project 2 (3GPP2) and wideband CDMA (W-CDMA, backward compatible with GSM and IS-136) supported by the Third Generation Partnership Project 1 (3GPP1). The main characteristics of these two 3G standards are summarized in Table D.4. Both standards use CDMA with power control and RAKE receivers, but the chip rates and other specification details are different. In particular, cdma2000 and W-CDMA are not compatible standards, so a phone must be dual-mode to operate with both systems. A third 3G standard, TD-SCDMA, is under consideration in China but is unlikely to be adopted elsewhere. The key difference between TD-SCDMA and the other 3G standards is its use of TDD instead of FDD for uplink/downlink signaling.

The cdma2000 standard builds on cdmaOne to provide an evolutionary path to 3G. The core of the cdma2000 standard is refered to cdma2000 1X or cdma2000 1XRTT, indicating that the radio transmission technology (RTT) operates in one pair of 1.25 MHz radio channels, and is thus backwards compatible with cdmaOne systems. The cdma2000 1X system doubles the voice capacity of cdmaOne systems and provides high-speed data services with projected peak rates of around 300 Kbps, with actual rates of around 144 Kbps. There are two evolutions of this core technology to provide high data rates (HDR) above 1 Mbps: these evolutions are refered to as cdma2000 1XEV. The first phase of evolution, cdma2000 1XEV-DO (Data Only), enhances the cdmaOne system using a separate 1.25 MHz dedicated high-speed data channel that supports downlink data rates up to 3 Mbps and uplink data rates up to 1.8 Mbps for an averaged combined rate of 2.4 Mbps. The second phase of the evolution, cdma2000 1XEV-DV (Data and Voice), is projected to support up to 4.8 Mbps data rates as well as legacy 1X voice users, 1XRTT data users, and 1XEV-DO data users, all within the same radio channel. Another proposed enhancement to cdma2000 is to aggregate three 1.25 MHz channel into one 3.75 MHz channel. This aggregation is refered to as cdma2000 3X, and its exact specifications are still under development.

W-CDMA is the primary competing 3G standard to cdma2000. It has been selected as the 3G successor to GSM, and in this context is refered to as the Universal Mobile Telecommunications System (UMTS). W-CDMA is also used in the Japanese FOMA and J-Phone 3G systems. These different systems share the W-CDMA link layer protocol (air interface) but have different protocols for other aspects of the system such as routing and speech compression. W-CDMA supports peak rates of up to 2.4 Mbps, with typical rates anticipated in the 384 Kbps range. W-CDMA uses 5 MHz channels, in contrast to the 1.25 MHz channels of cdma2000. An enhancement to W-CDMA called High Speed Data Packet Access (HSDPC) provides data rates of around 9 Mbps, and this may be the precursor to 4th-generation systems. The main characteristics of the 3G cellular standards are summarized in Table D.4.

3G Standard	cdma2000				W-CDMA		
Subclass	1X	1XEV-DO	1XEV-DV	3X	UMTS	FOMA	J-Phone
Channel Bandwidth (MHz)	1.25	1.25		3.75	5		
Chip Rate (Mchips/s)	1.2288			3.6864	3.84		
Peak Data Rate (Mbps)	.144	2.4	4.8	5-8	2.4 (8-10 with HSDPA)		
Modulation	QPSK (DL), BPSK (UL)						
Coding	Convolutional (low rate), Turbo (high rate)						
Power Control	800 Hz				1500 Hz		

Table D.4: Third-Generation Digital Cellular Phone Standards

D.2 Wireless Local Area Networks

Wireless local area networks (WLANs) are built around the family of IEEE 802.11 standards. The main characteristics of this standards family are summarized in Table D.5. The baseline 802.11 standard, released in 1997, occupies 83.5 MHz of bandwidth in the unlicensed 2.4 GHz frequency band. It specifies PSK modulation with FHSS or DSSS. Data rates up to 2 Mbps are supported, with CSMA/CA used for random access. The baseline standard was expanded in 1999 to create the 802.11b standard, operating in the same 2.4 GHz band using only DSSS. This standard uses variable-rate modulation and coding, with BPSK or QPSK for modulation and channel coding via either Barker sequences or Complementary Code Keying (CCK). This leads to a maximum channel rate of 11 Mbps, with a maximum user data rate of around 1.6 Mbps. The transmission range is approximately 100 m. The network architecture in 802.11b is specified as either star or peer-to-peer, although the peer-to-peer feature is not typically used. This standard has been widely deployed and used, with manufacturers integrating 802.11b wireless LAN cards into many laptop computers.

The 802.11a standard was finalized in 1999 as an extension to 802.11 to improve on the 802.11b data rates. The 802.11a standard occupies 300 MHz of spectrum in the 5 GHz NII band. In fact, the 300 MHz of bandwidth is segmented into three 100 MHz subbands: a lower band from 5.15-5.25 GHz, a middle band from 5.25-5.35 GHz, and an upper band from 5.725-5.825 GHz. Channels are spaced 20 MHz apart, except on the outer edges of the lower and middle bands, where they are spaced 30 MHz apart. Three maximum transmit power levels are specified: 40 mW for the lower band, 200 mW for the middle band, and 800 mW for the upper band. These restrictions imply that the lower band is mostly just suitable for indoor applications, the middle band for indoor and outdoor, and the high band for outdoor. Variable-rate modulation and coding is used on each channel: the modulation varies over BPSK, QPSK, 16QAM, and 64QAM, and the convolutional code rate varies over 1/2, 2/3, and 3/4. This leads to a maximum data rate per channel of 54 Mbps. For indoor systems, the 5 GHz carrier coupled with the power restriction in the lower band reduces the range of 802.11a relative to 802.11b, and also makes it more difficult for the signal to penetrate walls and other obstructions. 802.11a uses orthogonal frequency division multiplexing (OFDM) multiple access instead of FHSS or DSSS, and in that sense diverges from the original 802.11 standard.

The 802.11g standard, finalized in 2003, attempts to combine the best of 802.11a and 802.11b, with data rates of up to 54 Mbps in the 2.5 GHz band for greater range. The standard is backwards compatible with 802.11b so that 802.11g access points will work with 802.11b wireless network adapters and vice versa. However, 802.11g uses the OFDM, modulation, and coding schemes of 802.11a. Both access points and wireless LAN cards are available with all three standards to avoid incompatibilities. The 802.11a/b/g family of standards are collectively refered to as Wi-Fi, for wireless fidelity. Extending these standards to frequency allocations in countries other than the US falls under the 802.11d standard. There are several other standards in the 802.11 family that are under development: these are summarized in Table D.6.

A potential competitor to the 802.11 standards as well as cellular systems is the emerging IEEE 802.16 standard called WiMAX. This standard promises broadband wireless access with data rates on the order of 40 Mbps for fixed users and 15 Mbps for mobile users, with a range of several kilometers. Details of the specification are still being worked out.

	802.11	802.11a	802.11b	802.11g
Bandwidth (MHz)	300	83.5	83.5	83.5
Frequency Range (GHz)	2.4-2.4835	5.15-5.25 (lower) 5.25-5.35 (middle) 5.725-5.825 (upper)	2.4-2.4835	2.4-2.4835
Number of Channels	3	12 (4 per subband)	3	3
Modulation	BPSK,QPSK DSSS,FHSS	BPSK, QPSK, MQAM OFDM	BPSK,QPSK DSSS	BPSK, QPSK, MQAM OFDM
Coding		Conv. (rate 1/2,2/3,3/4)	Barker, CCK	Conv. (rate 1/2,2/3,3/4)
Max. Data Rate (Mbps)	1.2	54	11	54
Range (m)		27-30 (lower band)	75-100	30
Random Access	CSMA/CA			

Table D.5: 802.11 Wireless LAN Link Layer Standards

D.3 Wireless Short-Distance Networking Standards

This last section summarizes the main characteristics of Zigbee, Bluetooth, and UWB, which have emerged to support a wide range of short distance wireless network applications. These specifications are designed to be compliant with the IEEE 802.15 standards, a family of IEEE standards for short distance wireless networking called Wireless Personal Area Networks (WPANs). Bluetooth operates in the 2.4 GHz unlicensed band, Zigbee operates in the same band as well as in the 800 MHz and 900 MHz unlicensed bands, and UWB operates across a broad range of frequencies in an underlay to existing systems. Zigbee and Bluetooth include link, MAC, and higher layer protocols specifications, whereas UWB specificies just the link layer protocol. Table D.7 summarizes the main characteristics of Zigbee (2.4 GHz band only), Bluetooth, and UWB.

Zigbee consists of link and MAC layer protocols that are compliant with the IEEE 802.15.4 standard, as well as higher layer protocols for ad-hoc networking (mesh, star, or tree topologies), power management, and security. Zigbee supports data rates up to 250 Kbps with PSK modulation and DSSS. Zigbee generally targets applications

Standard	Scope
802.11e	Provides Quality of Service (QoS) at the MAC layer
802.11f	Roaming protocol across multivendor access points
802.11h	Adds frequency and power management features to 802.11a to make it more compatible with European operation
802.11i	Enhances security and authetication mechanisms
802.11j	Modifies 802.11a link layer to meet Japanese requirements
802.11k	Provides an interface to higher layers for radio and network measurements which can be used for radio resource management.
802.11m	Maintenance of 802.11 standard (technical/editorial corrections)
802.11n	MIMO link enhancements to enable higher throughput

Table D.6: IEEE 802.11 Ongoing Standards Work

requiring relatively low data rates, low duty cycles, and large networks. Power efficiency is key, with the goal of nodes operating for months or years on a single battery charge.

In contrast to Zigbee, Bluetooth provides up to 1 Mbps data rate, including three guaranteed low latency voice channels, using GFSK modulation and FHSS. Bluetooth normally transmits at a power of 1 mW with a transmission range of 10 m, although this can be extended to 100 m by increasing the transmit power to 100 mW. Networks are formed in subnet clusters (piconets) of up to 8 nodes, with one node acting as a master and the rest as slaves. TD is used for channel access, with the master node coordinating the FH sequence and synchronization with the slave nodes. Extended networks, or scatternets, can be formed when one node is part of multiple piconets. However, forming large networks through this approach is difficult due to the synchronization requirements of FHSS. Portions of the Bluetooth standard were formally adopted by the IEEE as its 802.15.1 standard.

UWB has significantly higher data rates, up to 100 Mbps, than either Zigbee or Bluetooth. It also occupies significantly more bandwidth, and has stringest power restrictions to prevent it from interfering with primary band users. Thus, it is only suitable for short-range indoor applications. UWB only defines a link layer technology, so it requires a compatible MAC protocol as well as higher layer protocols to become part of a wireless network standard. The modulation is BPSK or QPSK, with competing camps recommending either OFDM or DSSS overlayed on the data modulation. UWB is likely to become the link layer technology for the IEEE 802.15.3 standard, a family of standards for wireless networks supporting imaging and multimedia applications.

	Zigbee (802.15.4)	Bluetooth (802.15.1)	UWB (802.15.3 proposal)
Frequency Range (GHz)	2.4-2.4835	2.4-2.4835 GHz	3.1-10.6
Bandwidth (MHz)	83.5	83.5	7500
Modulation	BPSK,OQPSK DSSS	GFSK FHSS	BPSK,QPSK OFDM or DSSS
Max. Data Rate (Mbps)	.25	1	100
Range (m)	30	10	10
Power Consumption (mW)	5-20	40-100	80-150 mW
Access	CSMA/CA (optional TD)	TD	Undefined
Networking	Mesh/Star/Tree	Subnet Clusters (8 nodes)	Undefined

Table D.7: Short-Range Wireless Network Standards

Bibliography

[1] T. S. Rappaport. *Wireless Communications: Principles and Practice*, 2nd ed. Prentice Hall, 2002.

[2] W. Stallings, *Wireless Communications and Networks*, 2nd Ed., Prentice Hall, 2005.

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

[4] J.D. Vriendt, P. Lainé, C. Lerouge, and X. Xu, “Mobile network evolution: a revolution on the move,” *IEEE Comm. Mag.*, pp. 104-111, April 2002.

[5] I. Poole, “What exactly is . . . ZigBee?,” *IEEE Commun. Eng.*, pp. 44-45, Aug.-Sept. 2004

[6] D. Porcino and W. Hirt, “Ultra-wideband radio technology: potential and challenges ahead,” *IEEE Commun. Mag.*, Vol. 41, pp. 66 - 74, July 2003